THE STABILIZING EFFECT OF GOSSIP IN DECENTRALIZED LANGUAGE MODEL PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

As language-model pre-training scales beyond a single connected cluster, every-step global synchronization becomes increasingly expensive. Gossip-based decentralized training replaces global all-reduce with local updates and peer-to-peer parameter averaging. We investigate how this change affects training dynamics, beyond its communication benefits. Across the reported model families and optimizer settings, gossip-based decentralized training maintains comparatively stable validation-loss trajectories in large-learning-rate regimes where centralized training becomes highly oscillatory or moves toward high loss. To understand this phenomenon, we compare the training dynamics of decentralized and centralized SGD in deep linear networks. We prove that gossip SGD can achieve strictly lower expected full-Hessian sharpness at the worker-averaged iterate than centralized SGD. Our analysis further establishes conditions under which gossip SGD improves stability to directional parameter perturbations at large learning rates, quantified by reduced multistep sensitivity. This provides a theoretical counterpart to the stability advantage observed in our language-model experiments.

1 INTRODUCTION

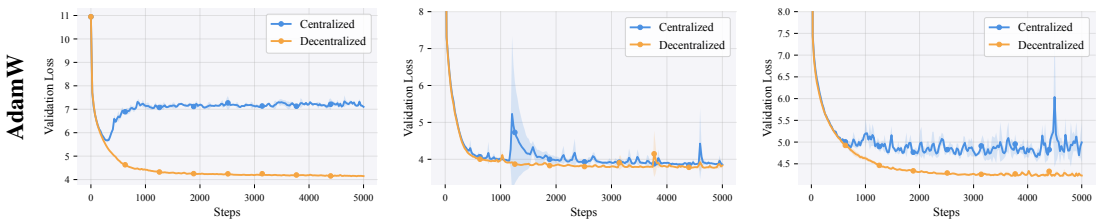
Frontier language-model pre-training may span multiple datacenters (Gemini Team, 2024), where every-step global synchronization becomes a bottleneck. Decentralized training replaces global all-reduce with local computation and peer-to-peer gossip (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009; Koloskova et al., 2020). Replicas take local optimizer steps and average parameters with peers (Boyd et al., 2006; de Vos et al., 2023), relaxing all-to-all communication for cross-datacenter and fault-tolerant scaling.

Gossip is usually analyzed as a communication-efficient approximation to centralized training. Partially synchronized replicas introduce consensus error arising from differences among worker parameters (Boyd et al., 2006), which decentralized-SGD convergence bounds control as an additional error term (Koloskova et al., 2020; Even et al., 2024). Yet these parameter deviations can also change optimization dynamics, as recent vision results suggest (Zhu et al., 2026). We ask whether they affect language-model pre-training stability beyond communication efficiency.

Training instability can waste pre-training computation through loss spikes, divergent runs, and costly recovery such as rollback, learning-rate changes, or data filtering (Chowdhery et al., 2023; Zhang et al., 2022; Takase et al., 2023; Huang et al., 2025). We use increased learning rates as controlled stability probes, following evidence that large-scale Transformer instabilities can be reproduced in smaller runs by pushing the learning rate high (Wortsman et al., 2023). In our experiments, gossip keeps validation loss comparatively stable in regimes where centralized training becomes highly oscillatory or moves toward high loss. Figure 1 shows this stability contrast across model families trained with AdamW. Further experiments test whether the contrast persists under SGD with momentum (SGDM) and across model sizes.

These observations motivate a closer look at loss geometry. For gradient descent near a fixed point, a perturbation along a Hessian eigenvector with the largest positive eigenvalue λ has linearized multiplier $1 - \eta\lambda$. Contraction along this direction requires $|1 - \eta\lambda| < 1$, equivalently $0 < \eta\lambda < 2$ (Polyak, 1987; Varga, 2000; Hairer et al., 1993). Lower top-Hessian sharpness therefore leaves more room below this upper stability boundary, near which neural-network training often evolves (Cohen et al., 2021; Wang et al., 2022). We complement validation-loss trajectories with scalable blockwise

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) Llama-style 0.1B (b) Qwen3.5-0.8B (c) GPT-2-style 1.5B

Figure 1: Validation-loss stability on FineWeb across model families and sizes with AdamW at peak learning rate $\eta = 5 \times 10^{-2}$. Across the shown panels, centralized training moves toward a high-loss or unstable trajectory, whereas decentralized training keeps validation loss comparatively stable. Full experimental settings are provided in Appendix C.1.

curvature diagnostics and ask theoretically how gossip-induced deviations change the top eigenvalue of the Hessian.

Our contributions.

- **Empirical phenomenon.** We demonstrate a stability advantage of gossip in language-model pre-training: decentralized training maintains comparatively stable validation-loss trajectories at large learning rates where centralized training becomes highly oscillatory or moves toward high loss. Controlled experiments vary the learning rate and model size, covering GPT-2-style models from 124M to 1.5B parameters. We also observe this stability advantage in Llama-style and Qwen3.5 models and with the Muon optimizer.
- **Theoretical contributions.** To understand this stability advantage, we establish three complementary results in deep linear networks. We prove a strict finite-step reduction in expected full-Hessian sharpness at the worker-averaged model relative to centralized SGD, with an explicit positive lower bound on the reduction (Theorem 1). We derive a local slow-time law for progress toward lower-curvature regions, quantifying how the mixing eigenvalues determine its rate (Theorem 2). We further prove improved stability to directional parameter perturbations relative to centralized SGD at large learning rates under suitable conditions, quantified by reduced multistep sensitivity (Theorem 3), providing a theoretical counterpart to our observed stability advantage.

2 Related Work

Training Instability in Large Models. Training stability is essential for reliable large-scale language-model pre-training: loss spikes or divergence can require checkpoint rollback, learning-rate reduction, or restart (Zhang et al., 2022; Chowdhery et al., 2023; DeepSeek-AI, 2026). Because these failures are difficult to study directly at full scale, small-scale proxy studies reproduce Transformer instabilities seen in large runs by increasing the learning rate, enabling controlled studies of large-step training (Wortsman et al., 2023). Recent analyses link these failures to optimizer instability (Molybog et al., 2023) and curvature growth measured by scalable sharpness diagnostics (Kalra et al., 2026; Ramasinghe et al., 2026). Mitigation strategies include spike-aware clipping and optimizer interventions (Huang et al., 2025; Ren et al., 2026), as well as architecture warm-up (Ramasinghe et al., 2026). We ask a complementary question: can gossip itself alter language-model pre-training stability?

Decentralized Training. Decentralized training is usually studied as a way to relax global synchronization, with workers alternating local optimizer steps and peer-to-peer parameter mixing. Classical distributed-optimization work established this setting (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009), and decentralized parallel SGD connected it to modern deep learning (Lian et al., 2017). Recent work studies convergence under changing topologies and communication schedules (Koloskova et al., 2020; Even et al., 2024), alongside topology design and consensus control (Ying et al., 2021; Kong et al., 2021). For large models, Factored Gossip DiLoCo separates synchronization into non-blocking and blocking mixing steps (Hewa Koneputugodage et al., 2026), while GASLoC combines local optimizer steps with sparse randomized gossip for language-model pre-training (Cagnasso et al., 2026). Beyond communication efficiency, decentralized training has been studied for its effects

Algorithm 1 Decentralized Training

input Initial local models $\{\theta^D_{i,0}\}_{i\in\mathcal{V}}$, number of steps T, mixing matrix distributions $\{\mathcal{P}_t\}_{t=0}^{T-1}$

for $t = 0, \dots, T-1$ **do**
  Sample the shared mixing matrix $\mathcal{W}_t \sim \mathcal{P}_t$
  for all $i \in \mathcal{V}$ in parallel **do**
    Sample a local minibatch $\mathcal{B}_{i,t}$ from $\mathcal{D}$
    $\theta^D_{i,t+1/2} \leftarrow \mathit{Optimizer}(\theta^D_{i,t}, \mathcal{B}_{i,t})$ ▷ Local optimizer step
    Send $\theta^D_{i,t+1/2}$ to out-neighbors and receive $\{\theta^D_{l,t+1/2}\}_{l\in\mathcal{N}_{\text{in}}(i)\setminus\{i\}}$
    $\theta^D_{i,t+1} \leftarrow \sum_{l\in\mathcal{N}_{\text{in}}(i)} (\mathcal{W}_t)_{i,l}\theta^D_{l,t+1/2}$ ▷ Neighbor averaging
  end for
end for

on generalization and flatness (Zhang et al., 2021; Vogels et al., 2023; Zhu et al., 2022; Cao et al., 2024; Zhu et al., 2026). Zhu et al. (2023) establish an asymptotic equivalence between decentralized SGD and average-direction sharpness-aware minimization (SAM), revealing implicit sharpness regularization. Wang & Johansson (2026) use adaptive consensus to sustain parameter deviations across workers and show how mixing and curvature determine their stationary variance in a local linear model. Dimlioglu et al. (2026) use parameter deviations in Local SGD to estimate a dominant Hessian subspace. For the deep-linear setting studied here, we prove an explicit positive lower bound on the finite-step reduction in expected full-Hessian sharpness of the worker-averaged model relative to centralized SGD. Gu et al. (2023b) derive slow dynamics for periodically averaged Local SGD, quantifying how the interval between global averages changes the noise-induced drift. For gradient steps followed by peer-to-peer mixing, we derive a local slow-time law in the small-noise regime that quantifies how the mixing spectrum, learning rate, and local curvature determine the additional rate of motion toward lower sharpness along the fitted manifold. Takezawa et al. (2026) also derive bounds that depend on the full mixing spectrum, but characterize optimization convergence rather than this local motion toward lower sharpness.

3 Notations and Preliminaries

3.1 Decentralized Training

Consider a connected communication graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ contains m workers and $\mathcal{E}$ denotes the peer-to-peer links. We study the homogeneous-population setting: each worker maintains a local model $\theta^D_i \in \mathbb{R}^d$ and samples from the same distribution $\mathcal{D}$; superscripts D and C distinguish decentralized and centralized parameters. This setting matches data-parallel large language model (LLM) pre-training, where shuffled shards approximate a shared corpus distribution rather than client-specific populations. Training targets the common population risk (Koloskova et al., 2020):

$$\min_{\theta\in\mathbb{R}^d} \left[\mathcal{L}(\theta) \triangleq \mathbb{E}_{\xi\sim\mathcal{D}}\mathcal{L}(\theta;\xi)\right]. \tag{1}$$

Thus each worker's population risk is $\mathcal{L}_i(\theta) = \mathcal{L}(\theta)$; any differences among worker parameters arise from optimization and communication, not population heterogeneity.

In practice, each worker uses stochastic gradients from a finite local dataset while communicating only with peers (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009). At step t, communication is represented by a weighted mixing matrix $\mathcal{W}_t \in [0,1]^{m\times m}$ sampled from a distribution $\mathcal{P}_t$. The coefficient $(\mathcal{W}_t)_{i,l}$ specifies how worker l contributes to worker i during averaging, with $\sum_l (\mathcal{W}_t)_{i,l} = 1$. A generic decentralized training loop alternates a local optimizer step with neighbor averaging, as shown in Algorithm 1. With communication every H local steps, $\mathcal{W}_t = I$ on steps without communication. We include worker i itself in its averaging neighborhood $\mathcal{N}_{\text{in}}(i)$ and set $(\mathcal{W}_t)_{i,l} = 0$ outside this set. For minibatch SGD, write $\mathcal{B}_{i,t} = \{Z_{i,t,j}\}_{j=1}^K$ for the K samples drawn by worker i at step t. The local update is $\theta^D_{i,t+1/2} = \theta^D_{i,t} - \eta\widehat{g}_{i,t}(\theta^D_{i,t})$, where $\widehat{g}_{i,t}(\theta) = K^{-1}\sum_{j=1}^K \nabla\mathcal{L}(\theta; Z_{i,t,j})$. The centralized SGD comparator uses the same mK samples and learning rate, evaluating all gradients at its own parameter: $\theta^C_{t+1} = \theta^C_t - \eta m^{-1}\sum_i \widehat{g}_{i,t}(\theta^C_t)$. Following Zhu et al. (2026), the main evaluation reports $\mathcal{L}_{\text{val}}(\bar{\theta}^D_t)$, where $\mathcal{L}_{\text{val}}$ is the loss averaged over validation samples and $\bar{\theta}^D_t \triangleq \frac{1}{m}\sum_{i\in\mathcal{V}}\theta^D_{i,t}$. We refer to $\theta^D_{i,t} - \bar{\theta}^D_t$ as worker i's parameter deviation from the mean.

3.2 Sharpness and Its Online Estimation

To analyze stability through local loss geometry, we use Hessian sharpness, namely the top eigenvalue of the Hessian $H = \nabla^2 \mathcal{L}(\theta)$. This quantity captures the sharpest local direction, or worst-case local curvature, and is the Hessian-based notion most directly tied to large-step stability in catapult and Edge-of-Stability analyses (Lewkowycz et al., 2020; Cohen et al., 2021). For plain GD in a positive-curvature local quadratic model, smaller Hessian sharpness leaves more room before the upper boundary $\eta\lambda_{\max}(H) < 2$. The derivation and its scope are given in Appendix C.1.4.

For Transformer models, we report a blockwise power-iteration Rayleigh diagnostic. Let $S_{\mathrm{sharp},t}$ denote the samples used for curvature measurement and $\mathcal{L}_{S_{\mathrm{sharp},t}}$ their empirical loss. We define $H_t \triangleq \nabla^2 \mathcal{L}_{S_{\mathrm{sharp},t}}(\theta_t)$, where θ_t denotes the evaluated centralized model or the worker-averaged decentralized model. Let $\mathcal{B}$ denote the set of Transformer blocks, and let $H_{t,\ell}$ be the principal block of H_t indexed by the parameters of block $\ell \in \mathcal{B}$.

Following Ramasinghe et al. (2026), we compute the Rayleigh diagnostic using Hessian-vector products and warm-started power iteration at selected checkpoints:

$$H_t v = \nabla_\theta \big(\nabla \mathcal{L}_{S_{\mathrm{sharp},t}}(\theta)^\top v\big)\big|_{\theta=\theta_t}, \qquad y^{(r+1)} \leftarrow \frac{H_t y^{(r)}}{\|H_t y^{(r)}\|_2}, \qquad \hat{\lambda}_t \leftarrow \langle y^{(r+1)}, H_t y^{(r+1)} \rangle.$$

For the blockwise diagnostic, we apply this iteration to each principal block $H_{t,\ell}$ and retain the largest returned Rayleigh quotient. These blockwise diagnostics measure curvature along the iterated directions. Warm-start details and the conditions for identifying the returned Rayleigh quotients with Hessian eigenvalues are given in Appendix C.1.4.

4 Empirical Observations

Experimental setup. All models are randomly initialized and trained from scratch on FineWeb (Penedo et al., 2024). The primary model suite consists of GPT-2-style decoder-only Transformers (Radford et al., 2019) with the standard settings from the modded-nanogpt repository (Jordan, 2024). Additional Llama-style and Qwen3.5 models and architecture details are described in Appendix C.1.1. All models are trained for $5{,}000$ steps with a global batch size of 0.5M tokens. We evaluate two representative optimizer classes: preconditioned adaptive AdamW (Loshchilov & Hutter, 2019) and non-preconditioned SGD with momentum (SGDM) (Sutskever et al., 2013). Optimizer-specific hyperparameters are provided in Appendix C.1.2. Unless stated otherwise, curves average three random seeds and shaded bands report one standard deviation.

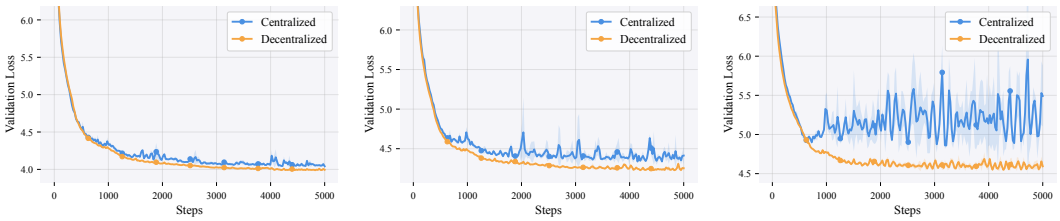
Centralized Training uses distributed data parallelism (DDP), which all-reduces gradients at every optimizer step to maintain a shared parameter trajectory; *Decentralized Training* uses gossip, alternating local updates with peer-to-peer parameter averaging. We evaluate decentralized validation loss and sharpness diagnostics on the averaged model $\bar{\theta}_t^D$ defined in Subsection 3.1 (protocols in Table 1).

4.1 Stability Under Large Learning Rates

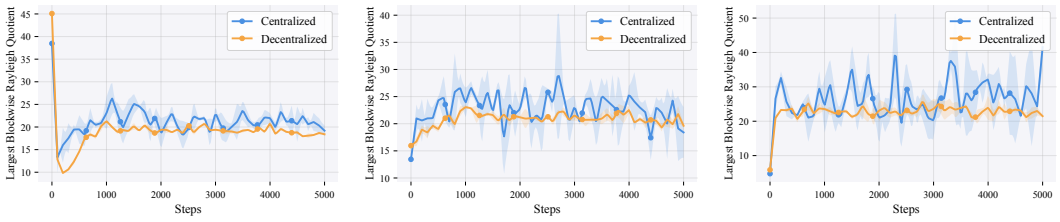
Large learning rates provide a controlled operational probe of stability in small-scale Transformer runs. Prior small-proxy work shows that instabilities observed only at large scale can be reproduced in smaller runs by increasing the learning rate (Wortsman et al., 2023), and large-step dynamics can enter qualitatively different regimes such as the catapult phase (Lewkowycz et al., 2020). Stability over a wider matched learning-rate range is therefore trajectory-level evidence, but it does not by itself identify a local Hessian or optimizer-state stability edge.

We instantiate this probe on a 124M-parameter GPT-2-style model trained with AdamW. We use large peak learning rates of $\{3, 5, 8\} \times 10^{-2}$, while fixing the data, model, global batch size, training horizon, and schedule shape. For each setting, we track validation loss together with the blockwise power-iteration Rayleigh diagnostic described in Subsection 3.2; details are in Appendix C.1.4.

Figure 2 shows that increasing the AdamW learning rate exposes a stability difference between the two protocols. At $\eta = 3 \times 10^{-2}$, both remain stable with similar validation loss; at $\eta = 5 \times 10^{-2}$ and $\eta = 8 \times 10^{-2}$, centralized training becomes noisy and enters a high-loss trajectory, whereas decentralized training keeps the averaged model on a smoother low-loss trajectory. In the same

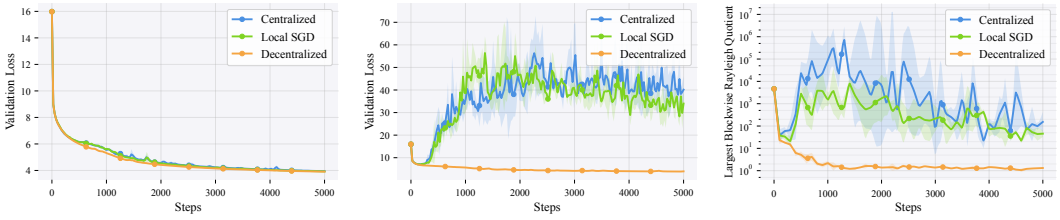


(a) Peak learning rate $\eta = 3 \times 10^{-2}$ (b) Peak learning rate $\eta = 5 \times 10^{-2}$ (c) Peak learning rate $\eta = 8 \times 10^{-2}$



(d) Peak learning rate $\eta = 3 \times 10^{-2}$ (e) Peak learning rate $\eta = 5 \times 10^{-2}$ (f) Peak learning rate $\eta = 8 \times 10^{-2}$

Figure 2: Validation-loss stability under increasing AdamW learning rates, with blockwise power-iteration Rayleigh diagnostics, on a 124M-parameter GPT-2-style model. **Top row:** validation loss over steps. **Bottom row:** the reported blockwise diagnostic at matched steps. As the learning rate increases, centralized runs increasingly overshoot or move to high-loss trajectories and reach larger diagnostic values, whereas gossip maintains a stable loss trajectory with smaller diagnostic values.



(a) Validation loss, $\eta = 0.5$ (b) Validation loss, $\eta = 1.5$ (c) Rayleigh diagnostic, $\eta = 1.5$

Figure 3: Validation-loss stability and a blockwise power-iteration Rayleigh diagnostic under large SGDM peak learning rates, with a Local SGD control, on a 124M-parameter GPT-2-style model. Panels (a) and (b) show validation loss at $\eta = 0.5$ and $\eta = 1.5$; panel (c) shows the reported blockwise diagnostic at $\eta = 1.5$. Gossip maintains lower validation loss and a smaller diagnostic value.

measured runs, the blockwise power-iteration Rayleigh diagnostic is smaller along the decentralized trajectory. At an AdamW peak learning rate of 5×10^{-2}, this stability advantage also persists across GPT-2-style models from 124M to 1.5B parameters (Appendix C.3). With Muon, as the base learning rate increases, centralized training becomes less stable while gossip maintains a more stable validation-loss trajectory (Figure 8).

4.2 Stability Under Non-Adaptive Optimization

We next study whether this stability contrast persists with SGDM, whose momentum update has no adaptive preconditioning, on the same 124M-parameter GPT-2-style model.

We use a Local-SGD-style control to examine the role of global parameter averaging in the observed separation. Model averaging can itself improve training performance, as in checkpoint merging (Tian et al., 2026). In the SGDM setting, Local SGD uses horizon $H = 1$, so workers take one local update before global parameter averaging; this preserves repeated global model averaging without gossip. We compare its validation-loss and Rayleigh-diagnostic trajectories with those of centralized training and gossip; protocol details are given in Table 1.

Figure 3 shows that the AdamW validation-loss pattern has a close non-adaptive counterpart. At $\eta = 0.5$, all three methods reduce validation loss, but centralized training and Local SGD are noisier than gossip, which generally has a lower blockwise Rayleigh diagnostic later in training (Figure 7e). At $\eta = 1.5$, both controls move to high-loss trajectories while decentralized training remains near the low-loss regime, and the reported blockwise diagnostic is smaller for decentralized training. Since Local SGD preserves repeated global averaging without gossip, global averaging alone is not sufficient to reproduce these measured trajectories. The SGDM results therefore show that this stability difference also occurs without adaptive preconditioning. Full results are in Appendix C.2.

Across the reported large-learning-rate experiments, gossip-based training maintains comparatively stable validation-loss trajectories where centralized training becomes highly oscillatory or moves toward high loss. Together with the smaller blockwise Rayleigh diagnostics in the measured runs, this pattern motivates a geometric interpretation of gossip's effect on training.

5 Why Decentralized Training Reduces Sharpness: Insights from Deep Linear Networks

To investigate a geometric mechanism behind the stability differences observed in the language-model experiments of Section 4, we study deep linear networks, a widely used model for analyzing multilayer dynamics (Saxe et al., 2014; Kawaguchi, 2016; Arora et al., 2018; Ghosh et al., 2025). Although linear in their inputs, these networks give rise to nonconvex optimization problems over their layer weights. We study how gossip changes sharpness, weight dynamics, and stability to perturbations along the initial prediction gradient.

We analyze Algorithm 1 with minibatch SGD, a constant learning rate $\eta > 0$, and a fixed nonnegative symmetric doubly stochastic matrix $\mathcal{W}$, where $\mathcal{W} \neq J := \mathbf{1}\mathbf{1}^\top/m$ and $\|\mathcal{W} - J\|_2 < 1$. All $m \geq 2$ workers and the centralized comparator share an initialization; each worker draws $K \geq 1$ iid samples per step, independently across workers and steps. Let $\theta^D_{i,t}$ denote worker i's parameter after step t and mixing, and define $\bar{\theta}^D_t = m^{-1}\sum_{i=1}^m \theta^D_{i,t}$. The centralized parameter θ^C_t uses the same mK samples and learning rate, with gradients evaluated at its own state, as defined in Subsection 3.1.

Our theory uses $f_\theta(x) = xg(\theta)$, where $g(\theta) = W_D \cdots W_1 \in \mathbb{R}$ is a bias-free deep linear network of fixed depth $D \geq 2$ and arbitrary finite hidden widths, with every matrix entry trainable. With label-noise amplitude σ, labels are $y = px + \sigma\zeta$: $p \neq 0$, x takes -1 and $+1$ with equal probability, and ζ is independent of x, bounded, centered, and has unit variance. For the squared sample loss $\ell(\theta; x, y) = \frac{1}{2}(f_\theta(x) - y)^2$, let $L_\sigma(\theta) = \mathbb{E}_{x,\zeta}[\ell(\theta; x, px + \sigma\zeta)]$. At a conditional-mean fit θ_*, $f_{\theta_*}(x) = \mathbb{E}[y \mid x] = px$, equivalently $g(\theta_*) = p$. Write $h_* = \lambda_{\max}(\nabla^2 L_\sigma(\theta_*)) = \|\nabla g(\theta_*)\|^2$.

5.1 Sharpness reduction and its dynamics

We ask whether decentralization lowers sharpness, and how this reduction develops during training. The first theorem gives a strict expected full-Hessian sharpness advantage for the worker-averaged model over centralized SGD, at each fixed training horizon. We then characterize the weight dynamics along the set of fitted weights and derive how the eigenvalues of the mixing matrix control their rate.

Theorem 1 (Finite-step sharpness reduction; informal)**.** *For every fixed horizon $N \geq 2$, decentralized SGD has strictly lower expected full-Hessian sharpness at the worker mean than centralized SGD throughout one full-parameter neighborhood of any conditional-mean fit with $0 < \eta h_* < 2$, for sufficiently small positive label noise, with a constant $c_* > 0$:*

$$\Delta_N(\theta_0, \sigma) := \mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\bar{\theta}^D_N)) - \mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\theta^C_N)) \leq -\frac{c_*}{4}\sigma^2 < 0. \tag{2}$$

The complete statement and proof are in Theorem E.1.

The result requires no balancing, alignment, or full-rank assumption. The initialization radius ε_* and coefficient c_* are independent of N; the admissible noise threshold σ_N may depend on N. An example is in Subsection E.6.

Figure 4 compares the finite-step sharpness gap with its second-order approximation at different noise amplitudes and shows how the gap changes with network depth and mixing strength. The protocol is given in Subsection H.3.

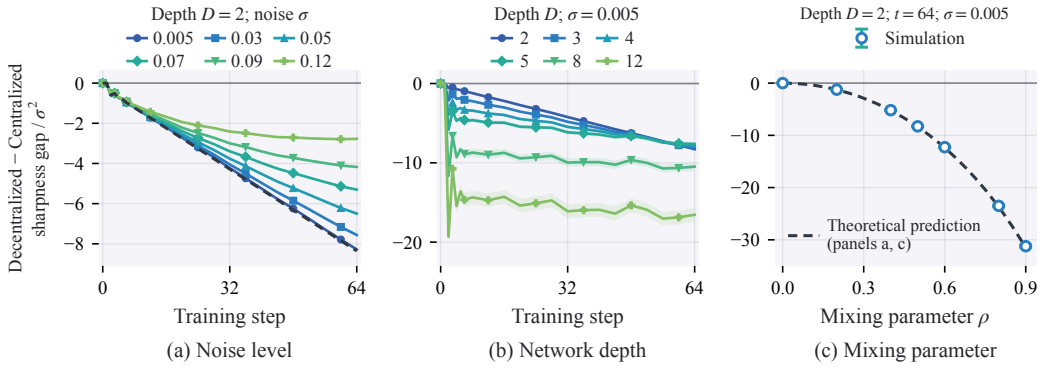


Figure 4: Finite-step sharpness reduction in matrix deep linear networks. Estimated mean decentralized–centralized full-Hessian sharpness gaps, divided by σ^2, for width-two networks initialized 0.01 from a fit, with $\eta h_* = 1.5$. (a) Noise levels at $D = 2$; (b) network depths at $\sigma = .005$; (c) mixing parameter ρ in $\mathcal{W} = J + \rho(I - J)$, at $D = 2$, $\sigma = .005$, and step 64. Larger ρ means weaker averaging. Dashed lines are second-order predictions; bands and bars are pointwise ± 1.96 paired standard errors over 4,096 noise paths. Panels have separate vertical scales.

From parameter deviations to weight dynamics. We next ask how parameter deviations change motion among fitted weights and how communication controls its rate. Earlier analyses relate decentralized SGD to implicit sharpness regularization (Zhu et al., 2023) and quantify enhanced curvature reduction for Local SGD (Gu et al., 2023b). For the update studied here, which takes a gradient step before mixing, we derive the weight dynamics and their explicit dependence on the mixing spectrum. Let $\mathcal{M} = \{\theta : g(\theta) = p\}$ be the fitted manifold. Weights on this manifold share predictions and population loss but can differ in full-Hessian sharpness $h = \|\nabla g\|^2$.

At a common exact fit, label noise induces first-order worker parameter deviations perpendicular to the fitted manifold $\mathcal{M}$. Their linear contributions to the averaged population gradient cancel. The leading quadratic correction adds an extra expected mean update. Its normal component changes the prediction to first order, while its tangential component preserves the prediction to first order and acts along $-\operatorname{grad}_{\mathcal{M}} h$, the negative gradient of sharpness along $\mathcal{M}$ (Subsection D.7.1). The finite-step sharpness comparison includes both normal and tangential responses. The next theorem characterizes the accumulated tangential dynamics in the rescaled time $s = N\sigma^2$, called slow time: a fixed s corresponds to more SGD steps as the noise amplitude decreases. The rates are determined by the mixing spectrum—the eigenvalues of $\mathcal{W}$.

Theorem 2 (Mixing controls slow sharpness dynamics; informal)**.** *In the small-noise limit, the mixing spectrum controls motion toward lower sharpness along a compact regular fitted patch (a portion of $\mathcal{M}$ where $\nabla g \neq 0$) with $0 < h_- \leq h \leq h_+$ and $\eta h_+ < 2$. For a common initialization in a sufficiently small neighborhood whose zero-noise trajectory approaches an interior fit $y^\dagger$, the expected full-Hessian sharpnesses of the original SGD iterates at $N = \lfloor s/\sigma^2 \rfloor$ converge as $\sigma \to 0$ to $h(y_C(s))$ and $h(y_D(s))$, uniformly on fixed local windows $s \in [s_-, S]$ with $0 < s_- < S \leq S_0$. The limiting fitted weights obey*

$$\dot{y}_A = -b_A(h(y_A)) \operatorname{grad}_{\mathcal{M}} h(y_A), \qquad y_A(0) = y^\dagger, \quad A \in \{C, D\}, \tag{3}$$

Here $q_j = \lambda_j(\mathcal{W})^2$, $j = 2, \ldots, m$, where the nonconsensus eigenvalues correspond to directions orthogonal to $\mathbf{1}$, describing differences among workers. The coefficients are

$$b_C(h) = \frac{\eta^2}{2mK(2 - \eta h)}, \qquad b_D(h) = b_C(h) + \frac{\eta^3 h}{2mK} \sum_{j=2}^{m} \frac{q_j}{1 - q_j(1 - \eta h)^2}. \tag{4}$$

If $\operatorname{grad}_{\mathcal{M}} h(y^\dagger) \neq 0$, then $h(y_D(s)) < h(y_C(s))$ on a local interval $0 < s \leq S_0$; the actual iterates retain this strict expected full-Hessian sharpness ordering on each fixed local slow-time window bounded away from zero for sufficiently small positive noise. The complete conditions and time quantifiers are in Theorem F.2.

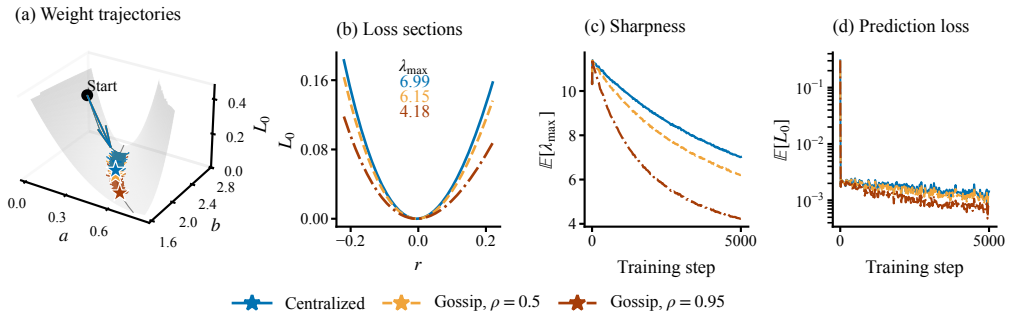


Figure 5: Weight trajectories and curvature in a three-layer linear network. (a) Centralized SGD and gossip SGD with mixing parameters $\rho = .5, .95$ (larger ρ means weaker averaging), in coordinates $a = w_1$, $b = \sqrt{2}w_2$; dots and stars mark initialization and step 5,000. (b) Prediction loss $L_0 = \frac{1}{2}(g-1)^2$ versus signed distance r along the unit top eigenvector of the full Hessian at each starred point. Panels (a,b) show one preselected noise realization. (c,d) Mean full-Hessian sharpness and prediction loss over 128 paired noise streams, evaluated at centralized weights or the gossip worker mean. See Subsection H.1 for the protocol.

At a common fitted position, $b_D(h) > b_C(h)$. Thus both limiting flows follow the same sharpness-descent curve, with the decentralized flow farther along at equal positive rescaled times when the curve is nonconstant.

At a common fitted position and fixed learning rate, increasing $|\lambda_j(\mathcal{W})|$, $j = 2, \ldots, m$, increases the corresponding contribution to $b_D(h) - b_C(h)$. Thus, the mixing eigenvalues quantify how gossip changes the rate of motion toward lower sharpness. Remark 1 details this update-order dependence.

Weight trajectories in the loss landscape. Figure 5 illustrates finite-noise weight trajectories and curvature differences in the three-layer network $g(w) = w_1 w_2 w_3$ with target $p = 1$. All three scalar weights are trainable without parameter tying. The updates preserve $w_2 = w_3$ from the common initialization, so the trajectories lie on an exact two-dimensional loss slice. In the displayed realization, all methods enter a low-loss region near the fitted valley. Gossip reaches positions farther along the valley, with lower sharpness at step 5,000 and the larger separation at weaker mixing. The ensemble curves also show lower sharpness than centralized SGD at low prediction loss. Subsection H.1 gives the protocol for this finite-time illustration and extends it to depths 6, 9, and 12 in Figure 10.

With multiple prediction directions, parameter deviations can increase sharpness; Subsection D.11 gives a bounded-noise counterexample. With full averaging $\mathcal{W} = J$, the two algorithms coincide.

5.2 Perturbation Sensitivity

We next compare the two methods' stability to a common parameter perturbation, measured by the expected squared parameter response over subsequent training steps. For a positive Hessian eigenvalue λ, the linearized update has multiplier $1 - \eta\lambda$: lowering λ weakens contraction when $\eta\lambda < 1$, but can reduce overshoot when $1 < \eta\lambda < 2$. We therefore compare responses over a finite window along a specified direction, under additional geometry and history conditions.

Multistep perturbation response. Choose $q_0 = \nabla g(\theta_0)/\|\nabla g(\theta_0)\|$, the unit direction that changes the initial prediction most strongly to first order, and keep this direction fixed. At an exact fitted initialization, it spans the Hessian's nonzero eigenspace. After n training steps, inject δq_0 into every worker and the centralized model at their respective states, then continue the original sampled updates with no reset or further injection. Let $Y^D_{i,n+k}$ and Y^C_{n+k} be weight derivatives with respect to δ at zero, starting from $Y^D_{i,n} = Y^C_n = q_0$. Define

$$\mathcal{G}^{q_0}_{n,k}(\sigma) = \mathbb{E}\left[\|\bar{Y}^D_{n+k}\|^2 - \|Y^C_{n+k}\|^2\right], \qquad \mathcal{E}^{q_0}_{n,T}(\sigma) = \sum_{k=1}^{T} \mathcal{G}^{q_0}_{n,k}(\sigma), \tag{5}$$

where $\bar{Y}^D = m^{-1} \sum_i Y_i^D$ and T counts steps after injection. Negative gaps mean smaller expected squared parameter responses to this single perturbation.

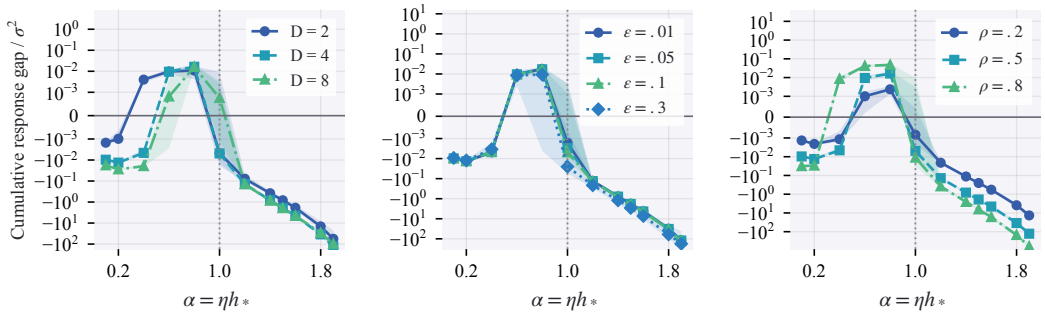
Let $\widetilde{\mathfrak{m}}_n$ denote the margin combining fitted geometry with the pre-injection second-moment history of the centered first-order parameter responses, computed from $\theta_*, \mathcal{W}, \eta, K, n$ by Equation (G.1) in Subsection G.1.

Theorem 3 (Sensitivity reduction along the initial prediction gradient; informal)**.** *Decentralized SGD has a smaller expected squared response than centralized SGD to the fixed initial prediction-gradient perturbation, both at each post-injection step and cumulatively. This holds for any fixed injection time $n \geq 2$ and finite window $T \geq 1$ when $1 < \eta h_* < 2$, $\widetilde{\mathfrak{m}}_n > 0$, initialization is sufficiently close to the fit, and positive label noise is sufficiently small:*

$$\mathcal{G}_{n,k}^{q_0}(\sigma) < 0 \quad (1 \leq k \leq T), \qquad \mathcal{E}_{n,T}^{q_0}(\sigma) < 0.$$

The complete statement is in Theorem G.1.

The conclusion is directional: a perturbation that rotates a hidden representation can have a positive cumulative response gap even when the gap along the initial prediction gradient is negative (Subsection G.2).



(a) Depth ($\varepsilon = .1$, $\rho = .5$) (b) Initialization ($D = 4$, $\rho = .5$) (c) Mixing ($D = 4$, $\varepsilon = .1$)

Figure 6: Sensitivity along the fixed initial prediction-gradient direction. Panels plot 16-step cumulative response gaps divided by σ^2, after one injection at $n = 4$, against $\alpha = \eta h_*$. Panels vary (a) depth D, (b) initialization distance $\varepsilon = \|\theta_0 - \theta_*\|$, and (c) mixing parameter ρ in $\mathcal{W} = \rho I + (1 - \rho)J$. Negative values favor decentralization. Bands are approximate simultaneous 95% group-bootstrap intervals; vertical axes use a signed log scale. See Subsection H.2 for fixed settings and the protocol.

Figure 6 shows a positive gap at intermediate learning rates and a negative gap at larger rates across the tested depths and initialization distances, with a smaller low-rate reduction. Larger initialization distances increase uncertainty; weaker mixing strengthens the reduction at larger learning rates in these experiments. These directional responses complement the stability separation in Subsection 4.1.

6 Conclusion

We study how gossip-based decentralized training affects the stability of language-model pre-training. Our experiments show that gossip sustains stable training at large learning rates where centralized training becomes highly oscillatory or moves toward high loss. To investigate the mechanism, we turn to deep linear networks, a classical theoretical model for studying optimization dynamics. Our analysis establishes conditions under which gossip achieves lower sharpness and reduced directional multistep sensitivity than centralized SGD. These results provide a theoretical basis for understanding the stabilizing effect of gossip.

Impact Statement

This work studies training stability in gossip-based decentralized language-model training. Our findings inform the design of distributed training methods that jointly prioritize communication efficiency and stable optimization.

AI USE STATEMENT

We used large language models to help identify related work and additional references. For the theoretical analysis, these tools helped develop the multilayer deep linear network setting and assisted with developing, writing, and checking the proofs of Theorems 1 and 2. We also used them to draft and revise portions of the manuscript, correct grammatical and typographical errors, and improve clarity and readability. The authors take responsibility for the final content, including the references, mathematical claims, and proofs.

ETHICS STATEMENT

Our research focuses on the optimization and theoretical properties of gossip-based decentralized language-model training. The work does not involve human-subject studies or the collection of new sensitive data. Our language-model experiments use the publicly available FineWeb corpus. The findings contribute to understanding training stability in decentralized learning; responsible downstream use of models trained with these methods should account for privacy, bias, and potential misuse.

REPRODUCIBILITY STATEMENT

Section 5 presents our theoretical results and their assumptions, with detailed derivations and proofs in Appendices D–G. Appendix C.1 documents the language-model experimental setup, including model architectures, optimizer settings, training and communication protocols, computational resources, and curvature measurements. Appendix H describes the protocols for the deep-linear-network experiments.

REFERENCES

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 244–253. PMLR, 2018. URL https://proceedings.mlr.press/v80/arora18a.html.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Pietro Cagnasso, Eugene Belilovsky, and Edouard Oyallon. Unifying local communications and local updates for LLM pretraining, 2026. URL https://arxiv.org/abs/2606.11081v1.

Ying Cao, Zhaoxian Wu, Kun Yuan, and Ali H. Sayed. On the trade-off between flatness and optimization in distributed learning. *arXiv preprint arXiv:2406.20006*, 2024. URL https://arxiv.org/abs/2406.20006.

Zachary Charles, Gabriel Teston, Lucio Dery, Keith Rush, Nova Fallen, Zachary Garrett, Arthur Szlam, and Arthur Douillard. Communication-efficient language model training scales reliably and robustly: Scaling laws for diloco, 2025. URL https://arxiv.org/abs/2503.09799.

Yiming Chen, Kun Yuan, Yingya Zhang, Pan Pan, Yinghui Xu, and Wotao Yin. Accelerating gossip sgd with periodic global averaging, 2021. URL https://arxiv.org/abs/2105.09080.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Woojin Chung, Jiwoo Hong, Na Min An, James Thorne, and Se-Young Yun. Stable language model pre-training by reducing embedding variability. In *Empirical Methods in Natural Language Processing*, pp. 10852–10863, 2024.

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. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows, 2024. URL https://arxiv.org/abs/2410.24206.

Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers, 2021. URL https://arxiv.org/abs/2106.06530v2. Version 2.

Martijn de Vos, Sadegh Farhadkhani, Rachid Guerraoui, Anne-Marie Kermarrec, Rafael Pires, and Rishi Sharma. Epidemic learning: Boosting decentralized learning with randomized communication. In *Advances in Neural Information Processing Systems*, 2023.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. Technical report, 2026. URL https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro/resolve/main/DeepSeek_V4.pdf.

Tolga Dimlioglu, Kristi Topollai, and Anna Choromanska. Worker disagreement reveals sharp directions in local SGD, 2026. URL https://arxiv.org/abs/2605.27739v1.

Arthur Douillard, Qixuan Feng, Andrei A. Rusu, Rachita Chhaparia, Yani Donchev, Adhiguna Kuncoro, Marc'Aurelio Ranzato, Arthur Szlam, and Jiajun Shen. DiLoCo: Distributed low-communication training of language models. *arXiv preprint arXiv:2311.08105*, 2024. URL https://arxiv.org/abs/2311.08105.

Arthur Douillard, Yanislav Donchev, Keith Rush, Satyen Kale, Zachary Charles, Zachary Garrett, Gabriel Teston, Dave Lacey, Ross McIlroy, Jiajun Shen, Alexandre Rame, Arthur Szlam, Marc'Aurelio Ranzato, and Paul Barham. Streaming diloco with overlapping communication: Towards a distributed free lunch, 2025. URL https://arxiv.org/abs/2501.18512.

Arthur Douillard, Keith Rush, Yani Donchev, Zachary Charles, Nova Fallen, Ayush Dubey, Ionel Gog, Josef Dean, Blake Woodworth, Zachary Garrett, Nate Keating, Jenny Bishop, Henry Prior, Edouard Yvinec, Arthur Szlam, Marc'Aurelio Ranzato, and Jeff Dean. Decoupled DiLoCo for resilient distributed pre-training. *arXiv preprint arXiv:2604.21428*, 2026. URL https://arxiv.org/abs/2604.21428.

Mathieu Even, Anastasia Koloskova, and Laurent Massoulié. Asynchronous SGD on graphs: A unified framework for asynchronous decentralized and federated optimization. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep linear networks beyond the edge of stability. *arXiv preprint arXiv:2502.20531*, 2025. URL https://arxiv.org/abs/2502.20531.

Xinran Gu, Kaifeng Lyu, Sanjeev Arora, Jingzhao Zhang, and Longbo Huang. A quadratic synchronization rule for distributed deep learning, 2023a. URL https://arxiv.org/abs/2310.14423.

Xinran Gu, Kaifeng Lyu, Longbo Huang, and Sanjeev Arora. Why (and when) does local SGD generalize better than SGD? In *International Conference on Learning Representations*, 2023b. URL https://openreview.net/forum?id=svCcui6Drl.

Mert Gurbuzbalaban, Yuanhan Hu, Umut Simsekli, Kun Yuan, and Lingjiong Zhu. Heavy-tail phenomenon in decentralized SGD. *arXiv preprint arXiv:2205.06689*, 2022. URL https://arxiv.org/abs/2205.06689.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 1993. ISBN 9783540788621. doi: 10.1007/978-3-540-78862-1.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://arxiv.org/abs/1512.03385.

Chamin Hewa Koneputugodage, Thalaiyasingam Ajanthan, Sameera Ramasinghe, Hadi Mohaghegh Dolatabadi, Shamane Siriwardhana, Gil Avraham, Violetta Shevchenko, Karol Pajak, James Snewin, and Alexander Long. Factored gossip DiLoCo: Reducing blocking communication in DiLoCo, 2026. URL https://arxiv.org/abs/2606.22768.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024. URL https://arxiv.org/abs/2404.06395.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.06842.

Sami Jaghouar, Jack Min Ong, Manveer Basra, Fares Obeid, Jannik Straube, Michael Keiblinger, Elie Bakouch, Lucas Atkins, Maziyar Panahi, Charles Goddard, et al. INTELLECT-1 technical report. *arXiv preprint arXiv:2412.01152*, 2024. URL https://arxiv.org/abs/2412.01152.

Keller Jordan. modded-nanogpt. https://github.com/kellerjordan/modded-nanogpt, 2024. URL https://github.com/KellerJordan/modded-nanogpt.

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

Dayal Singh Kalra, Jean-Christophe Gagnon-Audet, Andrey Gromov, Ishita Mediratta, Kelvin Niu, Alexander H. Miller, and Michael Shvartsman. A scalable measure of loss landscape curvature for analyzing the training dynamics of LLMs. *arXiv preprint arXiv:2601.16979*, 2026. URL https://arxiv.org/abs/2601.16979.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/f2fc990265c712c49d51a18a32b39f0c-Abstract.html.

Jari Kolehmainen, Nikolay Blagoev, John Donaghy, Oğuzhan Ersoy, and Christopher Nies. NoLoCo: No-all-reduce low communication training method for large models. *arXiv preprint arXiv:2506.10911*, 2025. URL https://arxiv.org/abs/2506.10911.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, 2020.

Lingjing Kong, Tao Lin, Anastasia Koloskova, Martin Jaggi, and Sebastian U. Stich. Consensus control for decentralized deep learning. In *International Conference on Machine Learning*, 2021.

Abhay Kumar, Louis Owen, Nilabhra Roy Chowdhury, and Fabian Gura. Zclip: Adaptive spike mitigation for llm pre-training, 2025. URL https://arxiv.org/abs/2504.02507.

Batiste Le Bars, Aurelien Bellet, Marc Tommasi, Erick Lavoie, and Anne-Marie Kermarrec. Refined convergence and topology learning for decentralized sgd with heterogeneous data, 2022. URL https://arxiv.org/abs/2204.04452.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism, 2020. URL https://arxiv.org/abs/2003.02218.

Shuaipeng Li, Penghao Zhao, Hailin Zhang, Xingwu Sun, Hao Wu, Dian Jiao, Weiyan Wang, Chengjun Liu, Zheng Fang, Jinbao Xue, Yangyu Tao, Bin Cui, and Di Wang. Surge phenomenon in optimal learning rate and batch size scaling. In *Advances in Neural Information Processing Systems*, 2024.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? – a mathematical framework, 2022. URL https://arxiv.org/abs/2110.06914v4. Version 4.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f75526659f31040afeb61cb7133e4e6d-Abstract.html.

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1eyO1BFPr.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3xHDeA8Noi.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. URL https://arxiv.org/abs/2502.16982.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Kairong Luo, Zhenbo Sun, Haodong Wen, Xinyu Shi, Jiarui Cui, Chenyi Dang, Kaifeng Lyu, and Wenguang Chen. How learning rate decay wastes your best data in curriculum-based LLM pretraining. *arXiv preprint arXiv:2511.18903*, 2025. URL https://arxiv.org/abs/2511.18903.

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

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1273–1282. PMLR, 2017.

Igor Molybog, Peter Albert, Moya Chen, Zachary DeVito, David Esiobu, Naman Goyal, Punit Singh Koura, Sharan Narang, Andrew Poulton, Ruan Silva, Binh Tang, Diana Liskovich, Puxin Xu, Yuchen Zhang, Melanie Kambadur, Stephen Roller, and Susan Zhang. A theory on adam instability in large-scale machine learning, 2023. URL https://arxiv.org/abs/2304.09871.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=2M9CUnYnBA.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoE: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024. URL https://arxiv.org/abs/2409.02060.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

Kosuke Nishida, Kyosuke Nishida, and Kuniko Saito. Initialization of large language models via reparameterization to mitigate loss spikes, 2024. URL https://arxiv.org/abs/2410.05052.

Yazan Obeidi, Amir Sarfi, Joel Lidin, Paul Janson, and Eugene Belilovsky. Heterogeneous low-bandwidth pre-training of llms, 2026. URL https://arxiv.org/abs/2601.02360.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb.

Bowen Peng, Lizhang Chen, Baiyu Su, Jeffrey Quesnelle, Diederik P. Kingma, and Qiang Liu. Demo: Decoupled momentum optimization, 2024. URL https://arxiv.org/abs/2411.19870.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software, Publications Division, New York, 1987. ISBN 0911575146.

Ji Qi, WenPeng Zhu, Li Li, Ming Wu, YingJun Wu, Wu He, Xun Gao, Jason Zeng, and Michael Heinrich. Dilocox: A low-communication large-scale training framework for decentralized cluster, 2025. URL https://arxiv.org/abs/2506.21263.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Sameera Ramasinghe, Thalaiyasingam Ajanthan, Gil Avraham, Yan Zuo, and Alexander Long. Protocol models: Scaling decentralized training with communication-efficient model parallelism. *arXiv preprint arXiv:2506.01260*, 2025. URL https://arxiv.org/abs/2506.01260.

Sameera Ramasinghe, Thalaiyasingam Ajanthan, Hadi Mohaghegh Dolatabadi, Chamin P. Hewa Koneputugodage, Gil Avraham, Violetta Shevchenko, Yan Zuo, Karol Pajak, and Alexander Long. Taming curvature: Architecture warm-up for stable transformer training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DuNf2vPTTK.

Lianhai Ren, Yucheng Ding, Xiao Liu, Qianxiao Li, Peng Cheng, and Yeyun Gong. MSign: An optimizer preventing training instability in large language models via stable rank restoration. *arXiv preprint arXiv:2602.01734*, 2026.

Max Ryabinin, Tim Dettmers, Michael Diskin, and Alexander Borzunov. SWARM parallelism: Training large models can be surprisingly communication-efficient. In *International Conference on Machine Learning*, pp. 29416–29440, 2023.

Amir Sarfi, Benjamin Therien, Joel Lidin, and Eugene Belilovsky. Communication efficient llm pre-training with sparseloco, 2025. URL https://arxiv.org/abs/2508.15706.

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. URL https://arxiv.org/abs/1312.6120.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pp. 1139–1147, 2013.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more: Stabilizing the pre-training of large language models. *arXiv preprint arXiv:2312.16903*, 2023.

Yuki Takezawa, Ryoma Sato, Han Bao, Kenta Niwa, and Makoto Yamada. Beyond exponential graph: Communication-efficient topologies for decentralized learning via finite-time convergence. In *Advances in Neural Information Processing Systems*, 2023.

Yuki Takezawa, Anastasia Koloskova, and Sebastian U. Stich. Improved convergence analysis of topology dependence in decentralized SGD, 2026. URL https://arxiv.org/abs/2606.09154v1.

Changxin Tian, Jiapeng Wang, Qian Zhao, Kunlong Chen, Jia Liu, Ziqi Liu, Jiaxin Mao, Xin Zhao, Zhiqiang Zhang, and Jun Zhou. WSM: Decay-free learning rate schedule via checkpoint merging for LLM pre-training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HhThhjKyfw.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9): 803–812, 1986.

Richard S. Varga. *Matrix Iterative Analysis*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2000. ISBN 9783642051562. doi: 10.1007/978-3-642-05156-2.

Thijs Vogels, Lie He, Anastasia Koloskova, Tao Lin, Sai Praneeth Karimireddy, Sebastian U. Stich, and Martin Jaggi. Relaysum for decentralized deep learning on heterogeneous data, 2021. URL https://arxiv.org/abs/2110.04175.

Thijs Vogels, Hadrien Hendrikx, and Martin Jaggi. Beyond spectral gap: The role of the topology in decentralized learning. *Journal of Machine Learning Research*, 24(355):1–31, 2023.

Guoxia Wang, Shuai Li, Congliang Chen, Jinle Zeng, Jiabin Yang, Dianhai Yu, Yanjun Ma, and Li Shen. Adagc: Improving training stability for large language model pretraining, 2025. URL https://arxiv.org/abs/2502.11034.

Jue Wang, Yucheng Lu, Binhang Yuan, Beidi Chen, Percy Liang, Christopher De Sa, Christopher Ré, and Ce Zhang. CocktailSGD: Fine-tuning foundation models over 500Mbps networks. In *International Conference on Machine Learning*, pp. 36058–36076, 2023.

Zesen Wang and Mikael Johansson. Decentralized SGD with controlled disagreement finds flatter minima, 2026. URL https://arxiv.org/abs/2602.02899.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*, 2024. URL https://arxiv.org/abs/2410.05192.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale Transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023. URL https://arxiv.org/abs/2309.14322.

Tongle Wu and Ying Sun. Implicit regularization of decentralized gradient descent for sparse regression. In *Advances in Neural Information Processing Systems*, 2024.

Chenghao Yang and Xuezhe Ma. Improving stability of fine-tuning pretrained language models via component-wise gradient norm clipping. In *Empirical Methods in Natural Language Processing*, pp. 4854–4859, 2022.

Kazuki Yano, Shun Kiyono, Sosuke Kobayashi, Sho Takase, and Jun Suzuki. Pre-training LLM without learning rate decay enhances supervised fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=JnebU2QLdH.

Bicheng Ying, Kun Yuan, Yiming Chen, Hanbin Hu, Pan Pan, and Wotao Yin. Exponential graph is provably efficient for decentralized deep training. In *Advances in Neural Information Processing Systems*, 2021.

Binhang Yuan, Yongjun He, Jared Quincy Davis, Tianyi Zhang, Tri Dao, Beidi Chen, Percy Liang, Christopher Ré, and Ce Zhang. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 2022.

Shahryar Zehtabi, Dong-Jun Han, Rohit Parasnis, Seyyedali Hosseinalipour, and Christopher G. Brinton. Decentralized sporadic federated learning: A unified algorithmic framework with convergence guarantees. In *International Conference on Learning Representations*, 2025.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse, 2023. URL https://arxiv.org/abs/2303.06296.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.

Wei Zhang, Mingrui Liu, Yu Feng, Xiaodong Cui, Brian Kingsbury, and Yuhai Tu. Loss landscape dependent self-adjusting learning rates in decentralized stochastic gradient descent. *arXiv preprint arXiv:2112.01433*, 2021. URL https://arxiv.org/abs/2112.01433.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/07ea874e9e4f71ec6680a3574a485a36-Abstract-Conference.html.

Zhanpeng Zhou, Mingze Wang, Yuchen Mao, Bingrui Li, and Junchi Yan. Sharpness-aware minimization efficiently selects flatter minima late in training, 2024. URL https://arxiv.org/abs/2410.10373.

Tongtian Zhu, Fengxiang He, Lan Zhang, Zhengyang Niu, Mingli Song, and Dacheng Tao. Topology-aware generalization of decentralized sgd, 2022. URL https://arxiv.org/abs/2206.12680.

Tongtian Zhu, Fengxiang He, Kaixuan Chen, Mingli Song, and Dacheng Tao. Decentralized SGD and average-direction SAM are asymptotically equivalent. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Tongtian Zhu, Tianyu Zhang, Mingze Wang, Zhanpeng Zhou, and Can Wang. On the surprising effectiveness of a single global merging in decentralized learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zrFnwRHuQo.

A Discussion and Limitations

Empirical scope. The language-model experiments report early-training validation-loss behavior under the stated model, optimizer, learning-rate, and communication settings. They do not establish an advantage for every training regime or final model quality. The blockwise power-iteration Rayleigh diagnostic is a scalable observation of local curvature, not a certified measurement of the largest eigenvalue of the full Hessian analyzed by the theorem.

Mathematical scope. Theorem 1 covers a full parameter neighborhood of a nonzero conditional-mean fit in bias-free, scalar-input/output deep linear networks of arbitrary fixed finite hidden widths. The initialization radius is independent of the horizon. The noise threshold is uniform over the ball at each horizon, but may decrease with time. The proof requires matched minibatch SGD, common worker/comparator initialization, bounded independent label noise, and fixed symmetric mixing. It does not require balanced layers, full-rank matrices, convexity, or contraction in every parameter direction. Larger prescribed noise and initialization far from fitting remain outside this local guarantee.

Interpretation. The extra mean update supplies a sufficient mechanism for lower expected full-Hessian sharpness at the worker mean for decentralized SGD and at the centralized parameters for centralized SGD, not a theorem of improved loss or generalization. The two-prediction counterexample shows that the sign can reverse beyond the structural setting. Theorem 3 proves a finite-window sensitivity advantage in arbitrary-depth matrix networks under a margin combining the fitted-point geometry with the pre-injection second-moment history of the centered first-order parameter responses and $1 < \eta h_* < 2$. Its initialization radius can depend on the response window, whereas that of Theorem 1 is horizon-independent. The noise thresholds in both results can depend on the horizon. A hidden-rotation counterexample rules out a general all-direction extension. The appendix retains the all-direction guarantee for balanced scalar-layer networks initialized at a fit. Fixed-noise infinite-time ordering and general multiple-prediction settings remain open. The synthetic full-matrix experiments test the local mechanism at finite noise; they do not establish that randomly initialized language models enter this regime.

B Additional Background and Related Work

B.1 Parallel, Federated, and Decentralized Training

Server-based local-update methods. Local-update methods reduce communication by letting workers perform multiple optimizer steps before synchronization. In data-center parallel training, Local SGD and post-local-SGD variants periodically average worker parameters to reduce all-reduce frequency and can improve generalization relative to very large minibatches (Lin et al., 2020). Federated averaging applies the same local-update-and-aggregate pattern to non-IID client data, but it relies on a central server that maintains the global model (McMahan et al., 2017). Recent language-model systems adapt this template to pre-training: DiLoCo trains learners locally for many inner steps before an outer synchronization (Douillard et al., 2024), and Decoupled DiLoCo further breaks lock-step execution by allowing independent learners to asynchronously send parameter fragments to a central synchronizer (Douillard et al., 2026). Recent variants further study scaling behavior, overlap between communication and computation, sparse or decoupled updates, and heterogeneous low-bandwidth pre-training (Charles et al., 2025; Douillard et al., 2025; Sarfi et al., 2025; Qi et al., 2025; Obeidi et al., 2026). These methods are close to our setting because they combine local drift with model averaging, but their synchronization remains global or server-mediated; our Local SGD control in Figure 7 therefore isolates repeated global model averaging from the gossip topology studied in the main text. DeMo reduces communication by exchanging compressed local momentum (Peng et al., 2024).

Fully decentralized training. Fully decentralized learning removes the central server and replaces global aggregation with peer-to-peer communication over a graph. Foundational decentralized optimization and D-PSGD analyses study how consensus error, graph connectivity, and stochastic gradients affect convergence (Tsitsiklis et al., 1986; Nedic & Ozdaglar, 2009; Lian et al., 2017; Koloskova et al., 2020). Recent analyses characterize how convergence depends on changing

topologies, local steps, asynchrony, and sporadic communication (Koloskova et al., 2020; Even et al., 2024; Zehtabi et al., 2025). Algorithmic work studies topology design and consensus control (Ying et al., 2021; Kong et al., 2021; Takezawa et al., 2023; de Vos et al., 2023), distinguishing spatial communication allocation, such as topology design, from temporal communication allocation, such as deciding when and how frequently agents synchronize (Ying et al., 2021; Takezawa et al., 2023; Kong et al., 2021; Zhu et al., 2026). Other decentralized-training work studies how periodic global averaging, synchronization schedules, relay-style momentum, and topology learning affect convergence under sparse communication or heterogeneous data (Chen et al., 2021; Gu et al., 2023a; Vogels et al., 2021; Le Bars et al., 2022). Representative large-model systems such as DT-FM (Yuan et al., 2022), SWARM Parallelism (Ryabinin et al., 2023), CocktailSGD (Wang et al., 2023), INTELLECT-1 (Jaghouar et al., 2024), Protocol Models (Ramasinghe et al., 2025), and NoLoCo (Kolehmainen et al., 2025) show that decentralized or low-communication protocols can be practical under bandwidth-constrained or heterogeneous infrastructure. Related analyses also study heavy-tailed behavior and implicit regularization in decentralized optimization (Gurbuzbalaban et al., 2022; Wu & Sun, 2024). These topics are relevant to decentralized optimization bias, but are distinct from the training-stability focus of this paper. Our work uses the same decentralized-training lens, but the target is different: rather than optimizing communication budget or final generalization, we ask whether gossip changes the training stability and sharpness profile of language-model pre-training.

B.2 Sharpness and Edge of Stability

Transformer training stability. Loss spikes have been linked to sudden gradient-norm growth and large sub-layer Jacobian spectral norms (Takase et al., 2023), as well as attention entropy collapse (Zhai et al., 2023). Prior work has proposed trajectory-level interventions for stabilizing Transformer training, including reducing embedding variability, adaptive or spike-aware clipping, reparameterized initialization, and batch-size or learning-rate scaling rules (Chung et al., 2024; Wang et al., 2025; Kumar et al., 2025; Nishida et al., 2024; McCandlish et al., 2018; Li et al., 2024). Yang & Ma (2022) propose component-wise gradient norm clipping to stabilize Transformer fine-tuning. These methods modify the centralized training recipe, whereas our study isolates whether decentralized gossip changes the stability profile of the averaged trajectory.

Sharpness-based large-step dynamics. Sharpness provides a common language for relating large learning rates, instability, and flatness-seeking training dynamics. Edge-of-stability studies show that neural-network training often approaches the raw-Hessian stability boundary (Cohen et al., 2021), with progressive sharpening shaping the trajectory before that boundary is reached (Wang et al., 2022); recent work further refines this picture (Cohen et al., 2024). Sharpness-aware minimization (SAM) directly optimizes a neighborhood-robust objective (Foret et al., 2021; Zhou et al., 2024). Decentralized SGD has been related asymptotically to an average-direction SAM update, suggesting that parameter deviations can alter the averaged trajectory (Zhu et al., 2023); related work connects topology and parameter deviations across workers to flatness, generalization, and controlled training signals (Zhang et al., 2021; Vogels et al., 2023; Cao et al., 2024; Zhu et al., 2026; Wang & Johansson, 2026). Our analysis isolates a finite-time, small-noise mechanism for the raw full-Hessian top eigenvalue under fixed gossip. A single nonzero curvature direction aligned with actual sample noise makes the leading sign nonnegative; full-parameter scalar-input/output deep linear networks make it strictly favorable at a common conditional-mean fit. Our near-fit theorem extends this ordering to a full neighborhood of that fit. Multiple prediction directions can reverse the result.

C Additional Experiments

C.1 Experimental Setup

Computational resources. All experiments are run on a GPU cluster equipped with 80 GB NVIDIA A100 GPUs, using a PyTorch implementation. Distributed runs use multiple GPUs. For representative $5{,}000$-step runs, the approximate per-run compute is 5.33 GPU-hours for GPT-2 Small (124M parameters), 24 GPU-hours for GPT-2 Large (774M parameters), 52 GPU-hours for GPT-2 XL (1.5B parameters), and 32 GPU-hours for Qwen3.5-0.8B.

Data and tokenization. We train all language models from scratch on FineWeb (Penedo et al., 2024) using the standard GPT-2 byte-level BPE tokenizer with a vocabulary size of 50,257. All training methods use the same tokenizer and a context length of 1,024 tokens. FineWeb is released under the Open Data Commons Attribution (ODC-By) 1.0 license, and the GPT-2 code and tokenizer assets are released under a modified MIT license.

C.1.1 Model Configurations

Model suite. The primary model suite consists of GPT-2-style decoder-only Transformers (Radford et al., 2019) at 124M–1.5B parameters, using the standard settings from the modded-nanogpt repository (Jordan, 2024). The modded-nanogpt codebase is released under the MIT license. We also include a repository-defined, randomly initialized LLaMA-style 0.1B (93M-parameter) decoder-only Transformer implemented with Hugging Face `LlamaForCausalLM`, rather than an official Meta LLaMA checkpoint or model size. This model follows the LLaMA-family module design (Touvron et al., 2023a) and uses 8 layers, hidden size 512, 16 attention heads of dimension 64, a 2,048-dimensional feed-forward block, context length 1,024, vocabulary size 50,304, and untied embeddings. We additionally train Qwen3.5-0.8B (Qwen Team, 2026), instantiated from the official Hugging Face configuration with its default random initialization and without loading pretrained weights.

Sparse model. We use OLMoE-small, a reduced-size configuration derived from the official OLMoE architecture (Muennighoff et al., 2024), as a sparse MoE stress test for the training-stability setting and initialize it from scratch. The profile uses 12 Transformer layers, hidden size 768, 6 attention heads, 64 experts, and top-8 routing. It has about 1.47B total parameters and about 276M active parameters per token. Because these randomly initialized OLMoE runs are especially prone to non-finite forward losses under the large-step regimes studied here, the OLMoE training loop uses the non-finite update guard described below.

C.1.2 Optimizer Hyperparameters

Common settings. All models use warmup-stable-only (WSO) learning-rate scheduling: a linear warmup for $1{,}000$ steps followed by a constant learning-rate plateau, corresponding to warmup-stable-decay (WSD) without the terminal decay phase (Hu et al., 2024; Wen et al., 2024; Yano et al., 2026). This non-decaying schedule is consistent with large-scale language-model recipes that retain a non-zero terminal learning rate (Brown et al., 2020; Touvron et al., 2023b), as well as recent evidence that overly small late-stage learning rates can waste high-quality data during continued training (Luo et al., 2025). Weight decay is set to $\lambda = 0.1$ for AdamW and $\lambda = 0.0$ for SGDM and Muon. The zero weight decay for Muon follows the original Muon formulation (Jordan et al., 2024). Gradient clipping uses a standard global norm of 1.0. All models are trained for $5{,}000$ steps with a global batch size of 0.5M tokens. The stability differences studied here emerge within the first $5{,}000$ steps. This horizon is therefore sufficient to compare the two protocols in terms of training stability. All experimental curves are averaged over three random seeds; solid lines report the mean and shaded bands report one standard deviation across seeds.

Optimizer-specific settings.

- **AdamW**: $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$.
- **SGDM**: momentum 0.95.
- **Muon**: momentum 0.95; Newton–Schulz iterations $k = 5$ for 2D parameters; the 1D AdamW parameter group uses $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$.

The AdamW beta coefficients, used both by the standalone AdamW optimizer and by the AdamW-updated parameter group inside Muon, follow the standard Transformer pre-training convention used in GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023a), and also match the AdamW baseline in Sophia's language-model pre-training study (Liu et al., 2024); we keep them fixed across all AdamW and Muon runs. For Muon, the momentum 0.95 and zero weight decay follow the original Muon formulation (Jordan et al., 2024). Muon applies the orthogonalized update to 2D weight matrices, while 1D or otherwise non-matrix parameters (e.g., biases and normalization scales) are optimized by the AdamW parameter group above. For Muon's shape-dependent learning-rate

scaling, we follow the frontier large-scale LLM-training setting used in Kimi/Moonshot's report (Liu et al., 2025): a reported learning rate η denotes the base learning rate before the matrix-update scaling. For a Muon-updated matrix of shape $A \times B$, the main Muon update uses the effective factor $\eta \cdot 0.2\sqrt{\max(A, B)}$, following their rule for matching the root-mean-square magnitude of Muon's matrix update to the AdamW update scale in stable distributed training. For SGDM, we take the conventional CNN/ImageNet default, momentum 0.9 (He et al., 2016), as a reference point and compare momentum values $\{0, 0.9, 0.95, 0.99\}$ on the 124M GPT-2-style model. We select 0.95 based on validation performance and keep it fixed across all SGDM runs; using a momentum above the CNN default is also consistent with recent language-model optimizer studies, which report that higher momentum helps SGD remain competitive with adaptive optimizers (Zhao et al., 2025).

C.1.3 Training Protocol

Protocol legends. We use short plot legends in both the main text and appendix, and reserve the full protocol names for Table 1. *Centralized* denotes centralized all-reduce implemented by DDP. *Local SGD* denotes a Local-SGD-style control that performs periodic global averaging of model parameters after H local steps. *Decentralized* denotes decentralized pairwise gossip, where each communication round averages model parameters over paired workers. This naming keeps figure legends compact while preserving the synchronization semantics needed to interpret the control experiments.

Table 1: Plot legends, full protocol names, and synchronization semantics. Local SGD is a control for separating local-update effects from communication-topology effects.

Plot legend	Full protocol name	Synced quantity	Sync scope	Sync frequency
Centralized	Centralized all-reduce (DDP)	gradients	global	every step
Local SGD	Local-SGD-style periodic global averaging	model parameters	global	every H steps
Decentralized	Decentralized pairwise gossip	model parameters	local pairs	every H steps

Data partitioning and evaluation. All runs labeled *Decentralized* use the same FineWeb training distribution as the Centralized baseline. Each training replica receives minibatches from the shared pre-training corpus, so the comparison isolates communication and local-update effects rather than data heterogeneity. For evaluation and sharpness tracking under the Decentralized protocol, we report metrics on the averaged model $\bar{\theta}_t^D$ defined in Subsection 3.1.

Local-update horizon. The Decentralized protocol alternates local optimizer updates with parameter gossip. We compare local-update horizons $H \in \{1, 5, 10, 25\}$ and use the most stable setting for each optimizer, yielding $H = 5$ for AdamW and $H = 1$ for SGDM. The Muon experiments also use $H = 1$. This small-horizon choice follows the post-local-SGD recommendation to use a small number of local steps (Lin et al., 2020).

Communication topology. Following the original DiLoCo 8-worker distributed language-model training setup (Douillard et al., 2024), we run experiments with the Decentralized protocol using eight training replicas. At each communication round, we randomly match the eight replicas into four disjoint pairs, and each pair averages its parameters. We do not tune over communication graph families; more structured topologies are an important direction for future work. The stability benefits of the Decentralized protocol in the main experiments therefore appear under a standard random-matching topology rather than a topology selected specifically for this phenomenon.

Non-finite update guard for OLMoE. For OLMoE runs, we add a non-finite update guard to prevent numerically invalid micro-batches from corrupting training. After each micro-batch forward pass, we check the scaled training loss for `NaN` or `Inf` before accumulating it or launching backpropagation. If any rank observes a non-finite loss, a distributed bad-flag reduction marks the entire optimizer step as skipped: any already accumulated gradients are cleared, no backward pass is performed for the offending loss, and no optimizer update is applied. The remaining micro-batches in the step are still advanced only to preserve data and token accounting. If the forward loss is finite but the backward pass produces non-finite gradients, the existing post-backward gradient guard also skips the optimizer step before any parameter update. Consequently, skipped steps do not modify

model parameters; forward-loss skips are omitted from the training-loss curve and recorded separately through skip diagnostics.

C.1.4 Sharpness Measurement Details

Quadratic-GD stability boundary. The connection between Hessian sharpness and GD stability becomes explicit in the classical optimization literature, where the relation is exact for quadratic GD dynamics and remains a useful local stability diagnostic in large-step regimes (Polyak, 1987; Varga, 2000; Hairer et al., 1993). In this picture, a gradient step obeys

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t).$$

A first-order Taylor expansion of the next gradient around θ_t then gives

$$\nabla \mathcal{L}(\theta_{t+1}) \approx \nabla \mathcal{L}(\theta_t) - \eta H(\theta_t) \nabla \mathcal{L}(\theta_t) = (I - \eta H(\theta_t)) \nabla \mathcal{L}(\theta_t), \qquad H(\theta) \triangleq \nabla^2 \mathcal{L}(\theta).$$

The local linearized dynamics are therefore governed by the matrix $I - \eta H(\theta_t)$. Since $H(\theta_t)$ is symmetric in this local quadratic picture, stability requires

$$\rho(I - \eta H(\theta_t)) < 1 \quad \Longleftrightarrow \quad -1 < 1 - \eta \lambda_i(H(\theta_t)) < 1 \; \forall i \qquad \Longrightarrow \qquad \eta \lambda_{\max}(H(\theta_t)) < 2. \tag{C.1}$$

Here $\rho(\cdot)$ denotes spectral radius. This derivation identifies the positive-curvature upper GD boundary $\eta \lambda_{\max}(H(\theta_t)) < 2$. For plain GD in a positive-curvature local quadratic model, smaller Hessian sharpness leaves more room before this upper edge; it is not sufficient for contraction in an indefinite Hessian and is not the exact stability operator for the experimental optimizers. Edge-of-Stability analyses still make raw curvature a useful diagnostic for large-step behavior (Cohen et al., 2021).

Measurement definitions. At a measurement checkpoint, let $S_{\text{sharp},t}$ denote the samples used for curvature measurement and $\mathcal{L}_{S_{\text{sharp},t}}$ their empirical loss. We define $H_t \triangleq \nabla^2 \mathcal{L}_{S_{\text{sharp},t}}(\bar{\theta}_t^D)$ and $\lambda_t \triangleq \lambda_{\max}(H_t)$; for centralized training, replace $\bar{\theta}_t^D$ with θ_t^C. In Subsection 3.2, the evaluation point is written uniformly as θ_t: $\theta_t = \bar{\theta}_t^D$ for decentralized training and $\theta_t = \theta_t^C$ for centralized training.

Let $\mathcal{B}$ denote the set of Transformer blocks, and let $H_{t,\ell}$ be the principal block of H_t indexed by the parameters of block $\ell \in \mathcal{B}$. We can define the max-block Hessian eigenvalue as $\lambda_t^{\text{block}} \triangleq \max_{\ell \in \mathcal{B}} \lambda_{\max}(H_{t,\ell})$. The mathematical quantity λ_t^{block} captures the largest block-level curvature. The plots report the largest Rayleigh quotient returned across Transformer blocks by warm-started power iteration.

Hessian-vector products. We compute the reported blockwise power-iteration Rayleigh diagnostics with Hessian-vector products, avoiding explicit Hessian materialization. Using the curvature sample set $S_{\text{sharp},t}$ and Hessian H_t defined above, for any vector v and measured step t,

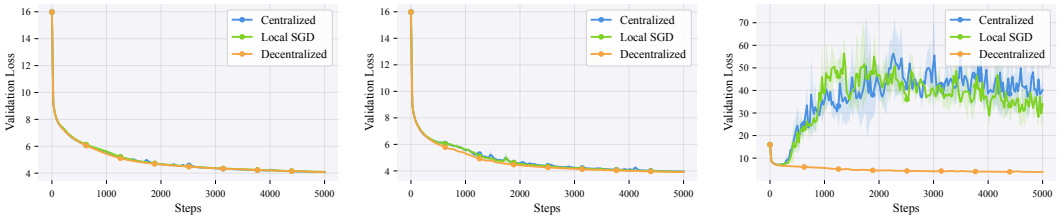
$$H_t v = \nabla_\theta \left(\nabla \mathcal{L}_{S_{\text{sharp},t}}(\theta)^\top v \right) \big|_{\theta = \theta_t} = \frac{d}{d\epsilon} \nabla \mathcal{L}_{S_{\text{sharp},t}}(\theta_t + \epsilon v) \bigg|_{\epsilon = 0}.$$

Thus the eigensolver only requires the standard Hessian-vector product primitive. For block-level diagnostics, we apply the same estimator to each principal block $H_{t,\ell}$ and then take the maximum over Transformer blocks.

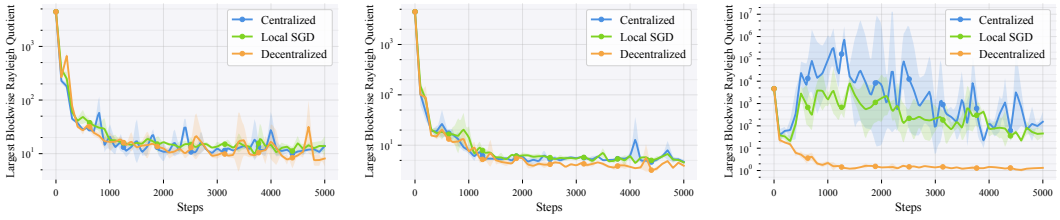
Warm-started power iteration. Starting from a unit vector $y^{(0)}$, normalized power iteration applies

$$z^{(r+1)} \leftarrow H_t y^{(r)}, \qquad y^{(r+1)} \leftarrow \frac{z^{(r+1)}}{\|z^{(r+1)}\|_2}, \qquad \hat{\lambda}_t^{(r+1)} \leftarrow \left\langle y^{(r+1)}, H_t y^{(r+1)} \right\rangle.$$

At each subsequent measurement checkpoint, the eigensolver is warm-started with $y^{(0)} = \hat{u}_{t_{\text{prev}}}$, where t_{prev} is the previous measured training step and $\hat{u}_{t_{\text{prev}}}$ is its retained final unit vector. The output $(\hat{\lambda}_t, \hat{u}_t)$ is carried forward to the next checkpoint. This is the raw-Hessian case of the preconditioned tracker in Ramasinghe et al. (2026) with the preconditioning matrix set to the identity, which is appropriate here because the globally averaged decentralized model does not have a single associated Adam preconditioner. Plain power iteration targets the eigenvalue of largest magnitude, not necessarily the largest eigenvalue. Thus the identifications $\hat{\lambda}_t \approx \lambda_t$ and $\hat{\lambda}_t^{\text{block}} \approx \lambda_t^{\text{block}}$ require either an eigensolver targeting the largest eigenvalue, a justified spectral shift, or checkpoint-level verification that the largest positive eigenvalue dominates in magnitude; without such a check, we report the computed Rayleigh quotients without identifying them with the largest eigenvalues.



(a) Learning rate $\eta = 0.25$ (b) Learning rate $\eta = 0.5$ (c) Learning rate $\eta = 1.5$



(d) Learning rate $\eta = 0.25$ (e) Learning rate $\eta = 0.5$ (f) Learning rate $\eta = 1.5$

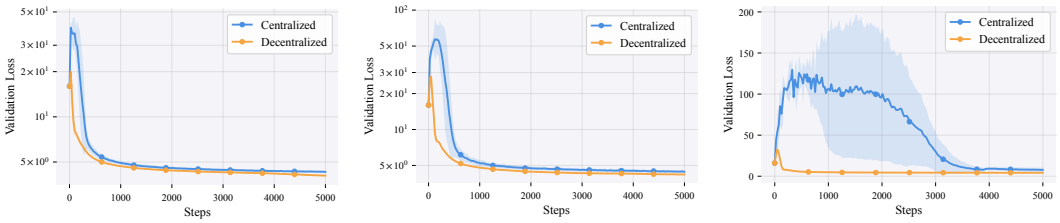
Figure 7: Validation-loss stability under increasing SGDM learning rates, with blockwise power-iteration Rayleigh diagnostics and a Local SGD control, on a 124M-parameter GPT-2-style model. **Top row:** validation loss over training steps. **Bottom row:** the reported blockwise diagnostic on a log scale at matched steps. Local SGD provides periodic global parameter averaging without decentralized gossip. The methods are similarly stable at $\eta = 0.25$; as the learning rate increases, decentralized training maintains lower validation loss and smaller diagnostic values than the controls.

C.2 Additional Results

SGDM reproduces the AdamW stability trend. Figure 7 shows that SGDM exhibits a learning-rate trend similar to AdamW in Figure 2. The Centralized and Decentralized protocols have nearly overlapping validation-loss trajectories at $\eta = 0.25$, and both remain stable. As the peak learning rate increases to $\eta = 0.5$ and then $\eta = 1.5$, centralized training becomes progressively less stable: its validation loss shows larger fluctuations at $\eta = 0.5$ and rises to a high-loss trajectory at $\eta = 1.5$. In contrast, decentralized training maintains a lower validation-loss trajectory at the same learning rates. The blockwise power-iteration Rayleigh traces follow the same pattern and are smaller and less volatile along the gossip-averaged trajectory.

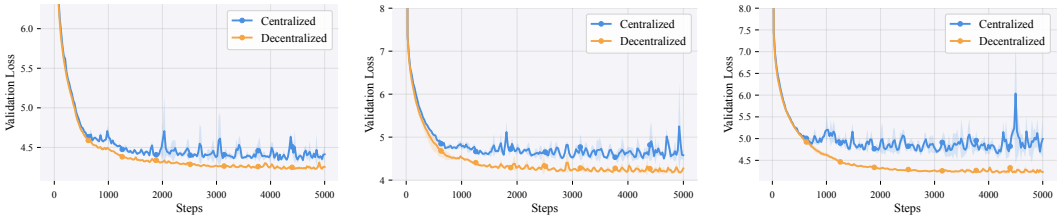
Global averaging alone does not account for the stability gain. Since the curvature diagnostic is evaluated at the globally averaged model rather than at individual local copies (Subsection 3.1), a natural confound is that weight averaging itself produces the measured separation. Checkpoint merging can emulate learning-rate decay through model averaging (Tian et al., 2026), and exponential moving averages can reduce noise and require less learning-rate decay (Morales-Brotons et al., 2024). The Local SGD control in Figure 7 directly controls for repeated global parameter averaging. With horizon $H = 1$, it preserves such averaging without multi-step local drift (Gu et al., 2023b), yet remains less stable and reaches larger Rayleigh-diagnostic values than the Decentralized protocol at large learning rates. Thus global averaging alone is not sufficient to reproduce the observed trajectories in this experiment.

Muon reproduces the AdamW stability trend. Figure 8 extends the learning-rate comparison to Muon on the 124M-parameter GPT-2-style model. The reported Muon learning rates are much larger in absolute value than the AdamW and SGDM learning rates because we follow the Kimi/Moonshot Muon setting: η denotes the base learning rate before the shape-dependent matrix-update scaling inside Muon (Liu et al., 2025). The 1D AdamW parameter group inside Muon uses a learning rate of 5×10^{-2} across these runs. These values should therefore be interpreted in Muon's own learning-rate parameterization rather than as AdamW-equivalent step sizes. Under this Muon parameterization, the qualitative trend matches AdamW and SGDM: as the learning rate is increased, centralized training becomes less stable, whereas decentralized training keeps the averaged model on a more stable validation-loss trajectory.



(a) Learning rate $\eta = 5$ (b) Learning rate $\eta = 2.5 \times 10^1$ (c) Learning rate $\eta = 1 \times 10^2$

Figure 8: Validation-loss stability under increasing reported Muon learning rates on a 124M-parameter GPT-2-style model. Each panel shows validation loss over training steps. As the reported Muon learning rate increases, centralized training becomes unstable, whereas decentralized training keeps the averaged model's validation loss comparatively stable.



(a) 124M (b) 774M (c) 1.5B

Figure 9: Validation-loss stability under increasing model size at a fixed AdamW peak learning rate $\eta = 5 \times 10^{-2}$. Each panel shows validation loss over training steps for a GPT-2-style model, with model size increasing from 124M to 1.5B parameters. As the model size increases, centralized training becomes more brittle, whereas decentralized training keeps validation loss stable.

C.3 Stability Under Larger Model Size

Larger language-model pre-training runs are more exposed to instability under aggressive optimization (Chowdhery et al., 2023; Molybog et al., 2023), making model scale a natural next axis for evaluating the stability gap observed in Subsection 4.1. After isolating a large-learning-rate gap on the 124M model, we ask whether gossip-based decentralized training continues to preserve its relative stability advantage when the same optimization regime is used to train larger models. To isolate model size, we fix AdamW, the data, global batch size, training horizon, schedule shape, and a high peak learning rate of $\eta = 5 \times 10^{-2}$, and gradually increase only the GPT-2-style model size from 124M to 1.5B parameters. This keeps all runs at the first large-step setting in Subsection 4.1 where the two protocols clearly separate, while changing the scale of the optimization problem.

Figure 9 shows that the large-step stability gap is not confined to the 124M model. At the same peak learning rate, the averaged three-seed traces become increasingly brittle for centralized training as the model size grows, with larger models moving toward high-loss or unstable trajectories. In contrast, decentralized training keeps the averaged model on comparatively controlled loss trajectories across the same tested model sizes. Thus, the stability advantage identified by the learning-rate probe persists when the controlled change is model capacity rather than the learning rate.

D Complete Proofs of the Structural Sharpness Result

D.1 Sharpness reduction in deep linear networks at fitted initialization

At a common fit, the sharpness comparison extends to arbitrary finite hidden widths without layer balance, and holds from step two throughout $0 < \eta h_* < 2$. This fitted case supplies the explicit response coefficient and normal–tangential decomposition used to prove Theorem 1. Exact fitting gives a closed response formula; Section E controls its perturbation over a full initialization neighborhood.

Theorem D.1 (Sharpness reduction in deep linear networks at fitted initialization)**.** *Consider Algorithm 1 with SGD and its centralized comparator under the mixing and sampling conditions of Section 5. Let $f_\theta(x) = xg(\theta)$, where $g(\theta) = W_D \cdots W_1 \in \mathbb{R}$ is a bias-free deep linear network with scalar input and output, fixed depth $D \geq 2$, and arbitrary finite hidden widths. Train every matrix entry using half-squared loss on samples with $x \in \{-1, 1\}$ and $y = px + \sigma\zeta$, where $p \neq 0$ and ζ is independent of x, bounded, centered, and has variance one. All models start at the same parameter point θ_* with $g(\theta_*) = p$. Set $h_* = \|\nabla g(\theta_*)\|^2$ and $L_\sigma = \mathbb{E}\ell$. For each fixed $N \geq 2$ and fixed $0 < \eta h_* < 2$, there exists $\sigma_0 > 0$ such that, for every $0 < \sigma \leq \sigma_0$,*

$$\mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\bar{\theta}_N^D)) < \mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\theta_N^C)).$$

No balancing or full-rank assumption is required. The quantitative reduction and explicit noise threshold are given in Theorem D.2 and Subsection D.9.

An exact fit here means a fit to the conditional mean px, not interpolation of noisy labels. At this point, $\nabla^2 L_\sigma(\theta_*) = \nabla g(\theta_*)\nabla g(\theta_*)^\top$, so $h_* = \|\nabla g(\theta_*)\|^2$ is its sole positive eigenvalue and hence the initial sharpness. Thus ηh_* measures the learning rate relative to the initial curvature; h_* is fixed at initialization, not recomputed along the trajectory. For the scalar chain $g = \prod_{\ell=1}^D w_\ell$ initialized with every $w_\ell = 1$, each partial derivative equals one and $h_* = D$. In general, h_* depends on the factorization, not just depth.

The proof does not require convexity or contraction in every parameter direction. The noise threshold depends on the fixed initialization, learning rate, graph, and horizon; the claim is not uniform over training time.

D.2 Matched updates and structural assumptions

We express the SGD specialization of Algorithm 1 through a centered stochastic-gradient field, under the mixing and sampling conditions of Section 5. The formulas below retain the same update order and compare with centralized SGD, which evaluates gradients at its own parameters. Here $\xi(\theta; Z)$ denotes a noise field, not the data sample Z; in the deep linear label-noise model, its amplitude δ equals σ. Let $L : \mathbb{R}^d \to \mathbb{R}$ be a common population loss and $\lambda(\theta) = \lambda_{\max}(\nabla^2 L(\theta))$ its top-Hessian sharpness. Worker i uses the minibatch field $\xi_{i,t}(\theta) = K^{-1}\sum_{k=1}^K \xi(\theta; Z_{i,t,k})$, with $\mathbb{E}\xi(\theta; Z) = 0$ at every state. Samples are independent across workers, steps, and batch entries. With fixed learning rate $\eta > 0$, the updates are

$$\theta_{i,t+1}^D = \sum_{j=1}^m \mathcal{W}_{ij}[\theta_{j,t}^D - \eta\{\nabla L(\theta_{j,t}^D) + \delta\xi_{j,t}(\theta_{j,t}^D)\}], \tag{D.1}$$

$$\theta_{t+1}^C = \theta_t^C - \eta\left\{\nabla L(\theta_t^C) + \frac{\delta}{m}\sum_{i=1}^m \xi_{i,t}(\theta_t^C)\right\}. \tag{D.2}$$

The centralized comparator evaluates the same sampled functions at its own state, using mK samples per step. It does not reuse gradients evaluated at worker states. Write $\bar{\theta}_t^D = m^{-1}\sum_i \theta_{i,t}^D$ and $J = \mathbf{1}\mathbf{1}^\top/m$.

Assumption 1 (One curvature direction and aligned stochastic gradients)**.** *All models start at θ_*, where*

$$\nabla L(\theta_*) = 0, \qquad H_* = \nabla^2 L(\theta_*) = h_* nn^\top, \qquad \mathbb{E}[\xi(\theta_*; Z)\xi(\theta_*; Z)^\top] = s_\xi nn^\top,$$

with $h_*, s_\xi > 0$ and $\|n\| = 1$. The graph matrix $\mathcal{W}$ is nonnegative, symmetric, and doubly stochastic, with $m \geq 2$, $\mathcal{W} \neq J$, and $\|\mathcal{W} - J\|_2 < 1$. On an open neighborhood of a closed ball about θ_*, L is

C^5. The sample fields and their parameter derivatives are jointly measurable. On one probability-one event, the fields are C^3 and their derivatives of orders zero through three satisfy deterministic bounds uniformly over the ball. The sample distribution is independent of δ.

These assumptions have an architectural explanation. If every sample loss depends on parameters through one common scalar $g(\theta)$, then at a regular critical prediction value, both the Hessian and the gradient covariance are multiples of $\nabla g\, \nabla g^\top$; Subsection D.8 gives the precise conditions. This is stronger than having a scalar network output: gradients at different inputs need not generally be collinear. The theorem itself needs only the local conditions in Assumption 1.

D.3 A POSITIVE FINITE-TIME SHARPNESS REDUCTION

Define the sharpness gradient $a = \nabla\lambda(\theta_*)$, its normal component $a_n = n^\top a$, and its remaining component $a_\perp = (I - nn^\top)a$. Let $c = 1 - \eta h_*$ and let $g_j = \lambda_j(\mathcal{W})^2$, $j = 2, \dots, m$, be the squared nonconsensus graph eigenvalues. The notation g_j is unrelated to the prediction function $g(\theta)$.

Theorem D.2 (Structural reduction of expected sharpness)**.** *Under Assumption 1, suppose $0 < \eta h_* < 2$ and fix $N \geq 2$. There are explicit constants $\delta_N > 0$ and $M_N < \infty$ such that the actual updates remain in a local ball and*

$$\Delta_N(\delta) := \mathbb{E}\lambda(\bar{\theta}_N^D) - \mathbb{E}\lambda(\theta_N^C) = -\frac{\delta^2}{2}\mathcal{C}_N + R_N(\delta), \qquad |R_N(\delta)| \leq M_N|\delta|^3 \tag{D.3}$$

for $|\delta| \leq \delta_N$. The coefficient is

$$d_t = \frac{\eta^2 s_\xi}{mK}\sum_{j=2}^{m} g_j \sum_{\ell=0}^{t-1}(g_j c^2)^\ell, \qquad d_0 = 0, \tag{D.4}$$

$$\mathcal{C}_N = \eta\sum_{t=0}^{N-1} d_t\left[a_n^2 c^{N-1-t} + \|a_\perp\|^2\right]. \tag{D.5}$$

Moreover,

$$\mathcal{C}_N \geq \eta d_1\big[(1-|c|)a_n^2 + (N-1)\|a_\perp\|^2\big], \qquad \mathcal{C}_N > 0 \iff a \neq 0. \tag{D.6}$$

If $a \neq 0$ and $0 < |\delta| \leq \min\{\delta_N, \mathcal{C}_N/(4M_N)\}$, then $\Delta_N(\delta) \leq -\delta^2\mathcal{C}_N/4 < 0$. The second threshold is interpreted as infinity if $M_N = 0$.

This appendix proves the finite-horizon statements for the actual updates in Eqs. (D.1) and (D.2). The proof has four steps: control the trajectories and Taylor remainder, derive the mean response with state-dependent sample gradients, determine its sign from the structural covariance, and verify the architecture and failure example.

D.4 NOTATION AND LOCAL BOUNDS

Fix a finite horizon N. Let $\mathcal{F}_t$ contain all samples drawn before step t; the current fields are independent of $\mathcal{F}_t$. All parameters start at the deterministic point θ_*. Tensor norms below are induced Euclidean multilinear norms. For $T = D^3L(\theta_*)$, define the vector $T[u, v]$ by $w^\top T[u, v] = D^3L(\theta_*)[w, u, v]$. For a symmetric matrix D, the vector contraction is $(T : D)_k = \sum_{p,q}\partial_{kpq}L(\theta_*)D_{pq}$. The squared graph eigenvalues g_j used below are distinct from the scalar prediction map $g(\theta)$.

Choose a closed ball $\mathcal{B}_R$ centered at θ_* inside the regularity neighborhood. Shrink R so that

$$\sup_{\mathcal{B}_R}\|\nabla^2 L(\theta) - H_*\|_2 \leq h_*/4. \tag{D.7}$$

For dimension at least two, Weyl's inequality gives a top eigengap of at least $h_*/2$ throughout this ball, since the Hessian spectrum at θ_* is $(h_*, 0, \dots, 0)$. In dimension one no separation condition is needed. The simple top eigenvalue $\lambda(\theta)$ is C^3 on a neighborhood of the ball, by the smooth dependence of a simple eigenvalue on a C^3 symmetric matrix. Its derivative bounds are finite:

$$A_j = \sup_{\mathcal{B}_R}\|D^j\lambda\|, \quad j = 1, 2, 3.$$

Write $L_j = \sup_{\mathcal{B}_R} \|D^j L\|$ for $j = 2, 3, 4$. Let M_j bound the sample fields and their derivatives of order $j = 0, 1, 2, 3$. The same bounds apply to every minibatch field by averaging. Because there are only finitely many sampled functions at this horizon, all bounds hold simultaneously on a common probability-one event. These assumptions provide deterministic bounds; a merely integrable local envelope would not itself control rare exits or sharpness evaluations outside the ball.

D.5 An explicit noise threshold and Taylor remainder

Set $b = 1 + \eta L_2$ and $S_t(b) = \sum_{k=0}^{t-1} b^k$, with $S_0 = 0$. Define

$$\delta_N = \min\left\{1, \frac{R}{2\eta M_0 S_N(b)}\right\}. \tag{D.8}$$

Here $M_0 > 0$ because $s_\xi > 0$. Let E_t be the maximum of the distances to θ_* of all worker states and the centralized state. As long as step-t states are in the ball,

$$\|\nabla L(\theta)\| \le L_2\|\theta - \theta_*\|, \qquad E_{t+1} \le bE_t + \eta|\delta|M_0.$$

Nonnegative row weights summing to one cannot increase the maximum worker norm; the averaged centralized field has norm at most M_0. Consequently induction from $E_0 = 0$ gives

$$E_t \le \eta|\delta|M_0 S_t(b) \le R/2, \qquad t \le N, \quad |\delta| \le \delta_N. \tag{D.9}$$

Each induction step is justified by the preceding bound, so no assumption of nonexit has been inserted into the argument. This controls the original algorithm without projection, truncation, or replacement of the sampled field. Convexity of the ball also places the averaged worker state in the ball.

For each fixed sample realization the paths are three times differentiable in δ. To bound their derivatives uniformly on $[-\delta_N, \delta_N]$, define

$$B_* = b + \eta\delta_N M_1, \qquad C_2 = \eta(L_3 + \delta_N M_2), \qquad C_3 = \eta(L_4 + \delta_N M_3).$$

Start $U_0 = V_0 = Q_0 = 0$ and set

$$\begin{aligned} U_{t+1} &= B_* U_t + \eta M_0, \\ V_{t+1} &= B_* V_t + C_2 U_t^2 + 2\eta M_1 U_t, \\ Q_{t+1} &= B_* Q_t + 3C_2 U_t V_t + C_3 U_t^3 + 3\eta M_1 V_t + 3\eta M_2 U_t^2. \end{aligned} \tag{D.10}$$

These bound respectively the first, second, and third amplitude derivatives of every worker, the centralized state, and the averaged state. Indeed, for the unmixed map $F(\theta, \delta) = \theta - \eta\nabla L(\theta) - \eta\delta\xi(\theta)$,

$$\begin{aligned} &\|F_\theta\| \le B_*, && \|F_{\theta\theta}\| \le C_2, && \|F_{\theta\theta\theta}\| \le C_3, \\ &\|F_\delta\| \le \eta M_0, && \|F_{\theta\delta}\| \le \eta M_1, && \|F_{\theta\theta\delta}\| \le \eta M_2. \end{aligned}$$

There are no second derivatives in δ holding θ fixed. The ordinary chain rule gives the first two recursions. Its third derivative has terms $F_\theta\theta''' + 3F_{\theta\theta}[\theta', \theta''] + F_{\theta\theta\theta}[\theta', \theta', \theta'] + 3F_{\theta\delta}\theta'' + 3F_{\theta\theta\delta}[\theta', \theta']$, giving the third recursion. Mixing and sample averaging preserve these bounds.

Composing with λ bounds the third derivative of sharpness with respect to δ by $A_1 Q_N + 3A_2 U_N V_N + A_3 U_N^3$. Define

$$M_N = \frac{A_1 Q_N + 3A_2 U_N V_N + A_3 U_N^3}{3}. \tag{D.11}$$

Taylor's theorem applied separately to the two sharpness values gives a bound $M_N|\delta|^3$ for the difference of their remainders: each contributes a factor $1/6$. The deterministic derivative bounds also justify interchanging expectation and amplitude derivatives by dominated convergence. Bounds on the random field derivatives similarly justify differentiating its all-state centering condition.

D.6 A finite-time response identity for centered sample gradients

We prove this step in greater generality than rank-one curvature. It applies to any stationary reference point with a simple top Hessian eigenvalue, provided the same local regularity, independence, and boundedness conditions hold. A sufficiently small ball preserves its initial eigengap; the preceding

estimates then apply with its local derivative bounds. No inverse Hessian or full-space contraction is required.

Write $H = \nabla^2 L(\theta_*)$, $A = I - \eta H$, $T = D^3 L(\theta_*)$, $a = \nabla\lambda(\theta_*)$, and

$$z_{i,t} = \partial_\delta \theta^D_{i,t}(0), \quad y_{i,t} = \partial^2_\delta \theta^D_{i,t}(0), \quad z^C_t = \partial_\delta \theta^C_t(0), \quad y^C_t = \partial^2_\delta \theta^C_t(0).$$

All zero-noise paths remain at θ_*. Let $\xi^0_{i,t} = \xi_{i,t}(\theta_*)$ and $J_{i,t} = D\xi_{i,t}(\theta_*)$. Differentiating the actual update maps gives

$$z_{i,t+1} = \sum_j \mathcal{W}_{ij}(Az_{j,t} - \eta\xi^0_{j,t}), \tag{D.12}$$

$$y_{i,t+1} = \sum_j \mathcal{W}_{ij}\{Ay_{j,t} - \eta T[z_{j,t}, z_{j,t}] - 2\eta J_{j,t} z_{j,t}\}, \tag{D.13}$$

$$z^C_{t+1} = Az^C_t - \frac{\eta}{m}\sum_i \xi^0_{i,t},$$

$$y^C_{t+1} = Ay^C_t - \eta T[z^C_t, z^C_t] - \frac{2\eta}{m}\sum_i J_{i,t} z^C_t.$$

Initial derivatives are zero. Double stochasticity therefore proves $\bar{z}_t := m^{-1}\sum_i z_{i,t} = z^C_t$ pathwise at every step.

Let $u_{i,t} = z_{i,t} - \bar{z}_t$ denote the centered first-order parameter response of worker i, and let $e_t = \bar{y}_t - y^C_t$ be the second-order mean-response difference. Since $\sum_i u_{i,t} = 0$, the mixed terms in the quadratic expansion of $T[z_{i,t}, z_{i,t}]$ vanish after averaging. Thus

$$e_{t+1} = Ae_t - \frac{\eta}{m}\sum_i T[u_{i,t}, u_{i,t}] - \frac{2\eta}{m}\sum_i J_{i,t} u_{i,t}. \tag{D.14}$$

The last term is generally nonzero on a sample path. However, $u_{i,t}$ is $\mathcal{F}_t$-measurable, and independent fresh samples and differentiated centering imply $\mathbb{E}[J_{i,t} \mid \mathcal{F}_t] = 0$. Consequently $\mathbb{E}[J_{i,t} u_{i,t} \mid \mathcal{F}_t] = 0$. Define

$$D_t = \frac{1}{m}\mathbb{E}\sum_i u_{i,t} u_{i,t}^\top, \qquad r_t = \mathbb{E}e_t.$$

Taking expectations in Equation (D.14) yields

$$r_{t+1} = Ar_t - \eta(T : D_t), \qquad r_N = -\eta\sum_{t=0}^{N-1} A^{N-1-t}(T : D_t). \tag{D.15}$$

In particular, correlation between the value and derivative of a fresh sample field does not create an extra term here: its derivative multiplies the centered first-order parameter response $u_{i,t}$, which is determined strictly before that sample is drawn.

At the terminal states, the second derivative of sharpness with respect to δ is

$$\partial^2_\delta \lambda(\bar{\theta}^D_N)\big|_0 = a^\top \bar{y}_N + \bar{z}_N^\top D^2\lambda(\theta_*)\bar{z}_N.$$

The centralized expression has y^C_N, z^C_N. Because $\bar{z}_N = z^C_N$, the first-order parameter responses agree pathwise, so the quadratic terms in the second derivative of sharpness also agree and cancel in the difference. Together with the Taylor bound, this proves the following identity.

Lemma D.3 (Mean-gradient response at a stationary reference point)**.** *Under the local conditions of this subsection,*

$$\Delta_N(\delta) = -\frac{\delta^2}{2}\eta\sum_{t=0}^{N-1} a^\top A^{N-1-t}(T : D_t) + R_N(\delta), \qquad |R_N(\delta)| \le M_N|\delta|^3. \tag{D.16}$$

The bounds in Eqs. (D.8) and (D.11) are valid on a ball with a simple top eigenvalue.

This identity alone does not fix the sign. The next step uses the structural assumptions to do so.

D.7 Covariance of centered first-order parameter responses

Return to $H = h_* nn^\top$ and single-sample covariance $S = s_\xi nn^\top$. Centering implies $\mathbb{E}\|(I - nn^\top)\xi(\theta_*; Z)\|^2 = 0$, so the noise at θ_* lies along n almost surely. The linear variations in Equation (D.12) therefore remain along n.

Choose an orthonormal graph eigenbasis $q_1 = \mathbf{1}/\sqrt{m}, q_2, \ldots, q_m$. For $Z_t = [z_{1,t}, \ldots, z_{m,t}]$ and $\Xi_t^0 = [\xi_{1,t}^0, \ldots, \xi_{m,t}^0]$,

$$Z_{t+1} q_j = \lambda_j(\mathcal{W})(A Z_t q_j - \eta \Xi_t^0 q_j).$$

Independence across workers and batch entries gives $\mathrm{Cov}(\Xi_t^0 q_j) = S/K$. The current forcing is independent of Z_t. With $c = 1 - \eta h_*$ and $g_j = \lambda_j(\mathcal{W})^2$, the first-order parameter response in each nonconsensus graph mode consequently has covariance

$$\mathbb{E}[(Z_t q_j)(Z_t q_j)^\top] = \frac{\eta^2 s_\xi}{K} g_j \sum_{\ell=0}^{t-1} (g_j c^2)^\ell \, nn^\top.$$

Parseval's identity in worker space gives

$$D_t = \frac{1}{m} \sum_{j=2}^{m} \mathbb{E}[(Z_t q_j)(Z_t q_j)^\top] = d_t nn^\top,$$

with precisely Equation (D.4). Cross-mode independence is not needed. Since $\mathcal{W} \neq J$, at least one $g_j > 0$, so $d_1 = \eta^2 s_\xi (mK)^{-1} \sum_{j=2}^m g_j > 0$.

For a simple eigenvalue, differentiation of $H(\theta)n(\theta) = \lambda(\theta)n(\theta)$ with $\|n(\theta)\| = 1$ gives

$$a^\top w = n^\top DH(\theta_*)[w]n = w^\top T[n, n] \quad \text{for all } w.$$

Symmetry of third derivatives therefore proves $T[n, n] = a$. It follows that $T : D_t = d_t a$ and

$$a^\top A^k (T : D_t) = d_t \{a_n^2 c^k + \|a_\perp\|^2\}.$$

Substitution into Equation (D.16) proves Eqs. (D.3) and (D.5).

To determine the sign even when $c < 0$, substitute the finite sum for d_t and reverse the order of summation:

$$\mathcal{C}_N = \frac{\eta^3 s_\xi}{mK} \sum_{j=2}^{m} \sum_{\ell=0}^{N-2} g_j^{\ell+1} c^{2\ell} \left[a_n^2 \frac{1 - c^{N-1-\ell}}{1 - c} + (N - 1 - \ell)\|a_\perp\|^2 \right]. \tag{D.17}$$

For $-1 < c < 1$ and integer $q \geq 1$, the sum $S_q(c) = (1 - c^q)/(1 - c)$ is positive and at least $1 - |c|$. For $c \geq 0$ it is at least one. If $c = -r$, $0 < r < 1$, pairing consecutive terms gives $(1-r)(1+r^2+\cdots)$ for even q, while odd q adds a nonnegative final term; the bound follows (and $q = 1$ is immediate). Every term of Equation (D.17) is nonnegative. Retaining $\ell = 0$ proves Equation (D.6). If $a = 0$ the coefficient vanishes; if $a \neq 0$ the lower bound is strictly positive. Finally, the cubic remainder bound gives $M_N|\delta|^3 \leq \delta^2 \mathcal{C}_N/4$, completing the proof of Theorem D.2.

Degeneracies and quantifiers. For $N = 1$, averaging the first update gives exact pathwise equality with the centralized state. If $\mathcal{W} = J$, equality holds inductively at every step. If the sample noise vanishes at θ_* almost surely, both chains remain at θ_* almost surely. If $a = 0$, only the second-order coefficient is zero; higher-order gaps can remain. For a fixed window $2 \leq t \leq N$, the minimum of the corresponding positive noise thresholds yields simultaneous strict inequalities. These are finite collections of bounds, not a limit uniform in time.

The step-size interval is sufficient, not necessary. At $N = 2$, $\mathcal{C}_2 = \eta d_1 \|a\|^2$ is positive for every $\eta > 0$ whenever $a \neq 0$, with a suitable local noise threshold. Outside $0 < \eta h_* < 2$, later signs can reverse: set $L(u, v) = \frac{1}{2}(uv - \frac{1}{2})^2$, initialize $u = v = 1/\sqrt{2}$, take $\eta = 3$, two workers, and $\mathcal{W} = J + \sqrt{0.1}(I - J)$. Then $h_* = 1$, $c = -2$, $a_\perp = 0$, and the normal factor at $N = 3$ is $c + 1 + 0.1c^2 = -3/5 < 0$. Actual bounded label noise as in Theorem D.1 gives a positive sharpness gap for sufficiently small amplitude by Lemma D.3. This counterexample does not make the stable interval a necessary finite-horizon boundary.

D.7.1 TWO COMPONENTS OF THE SHARPNESS-REDUCING RESPONSE

The positive coefficient above separates into a prediction-changing response and a first-order prediction-preserving response. For the fitted deep linear model, $n = \nu = \nabla g(\theta_*)/\|\nabla g(\theta_*)\|$ and $H = h_*\nu\nu^\top$. Thus $Dg(\theta_*)[z] = 0$ exactly when $z \perp \nu$, while $A\nu = c\nu$ and $Az = z$ for such z, where $c = 1 - \eta h_*$. These are statements about local prediction changes and linearized propagation, not exact invariances of finite parameter displacements.

On the fitted manifold $\mathcal{M}$, the top Hessian eigenvalue agrees with $h(\theta) = \|\nabla g(\theta)\|^2$, so their tangential derivatives agree. Thus, for $a = \nabla\lambda_{\max}(\nabla^2 L_\sigma)(\theta_*)$,

$$(I - \nu\nu^\top)a = \operatorname{grad}_{\mathcal{M}} h(\theta_*) = (I - \nu\nu^\top)\nabla h(\theta_*),$$

where I is the identity on weight space. The extra expected mean update along $-a$ therefore has a tangential component along the negative sharpness gradient on $\mathcal{M}$.

Write $a_n = \nu^\top a$, $a_\perp = (I - \nu\nu^\top)a$, and decompose the second noise-amplitude response as $r_t = r_{t,n}\nu + r_{t,\perp}$. Since $T : D_t = d_t a$, projecting Equation (D.15) gives the exact coefficient recursions

$$\begin{aligned} r_{t+1,n} &= c\, r_{t,n} - \eta d_t a_n, \\ r_{t+1,\perp} &= r_{t,\perp} - \eta d_t a_\perp, \qquad r_{0,n} = 0, \quad r_{0,\perp} = 0. \end{aligned} \tag{D.18}$$

Consequently $\mathcal{C}_N = \mathcal{C}_{N,n} + \mathcal{C}_{N,\perp}$, with

$$\mathcal{C}_{N,n} = \eta a_n^2 \sum_{t<N} d_t c^{N-1-t}, \qquad \mathcal{C}_{N,\perp} = \eta\|a_\perp\|^2 \sum_{t<N} d_t. \tag{D.19}$$

Each contributes $-\delta^2\mathcal{C}/2$ to the leading sharpness gap. The perpendicular sum is nonnegative term by term: there is no linear restoring force, so a sustained source accumulates. For the parallel sum, $c < 0$ allows alternating signs in individual terms, but the geometric-sum argument in Equation (D.17) proves

$$\mathcal{C}_{N,n} \geq \eta d_1(1 - |c|)a_n^2, \qquad \mathcal{C}_{N,\perp} \geq \eta d_1(N-1)\|a_\perp\|^2 \quad (N \geq 2).$$

Under the stated nondegeneracy and step-size assumptions, each nonzero component therefore yields a strictly favorable second-order coefficient. Neither component must be present individually; their sum is positive whenever $a \neq 0$. The finite-horizon remainder bound in Equation (D.3) converts this coefficient statement into the actual small-noise sharpness comparison.

Illustrating accumulation without a first-order prediction change. The second route is visible even in a two-layer network with hidden width two. Take $W_{1,*} = (1,1)^\top$ and $W_{2,*} = (0.6, 0.4)$, so the initial prediction is $g(\theta_*) = 1$. Unlike the balanced uvx example, this factorization has a nonzero component $a_\perp = (I - \nu\nu^\top)a$ of the sharpness gradient perpendicular to the prediction gradient. Independent samples continually create worker parameter deviations from their mean, while mixing and the gradient dynamics attenuate them. The leading second moment of these deviations approaches a finite level, leaving an approximately constant extra mean update. Its perpendicular component is not pulled back by the fitted-point linearization: it changes the prediction only at higher order and therefore encounters no linear restoring force. Successive contributions can accumulate, rather than cancel.

More precisely, d_t is the leading-order second-moment coefficient of the worker parameter deviations (Equation (D.4)), and r_t is the second-order response of the decentralized–centralized mean-parameter difference, so the leading expected difference is $\sigma^2 r_t/2$. The perpendicular recursion in Equation (D.18) yields

$$\frac{\sigma^2}{2} a_\perp^\top r_t^\perp = -\frac{\eta\sigma^2}{2}\|a_\perp\|^2 \sum_{s=0}^{t-1} d_s.$$

This expression is this direction's contribution to the leading sharpness gap. Once d_t is nearly constant, it is approximately linear in time. The balanced uvx example isolates the prediction-changing route; the present example additionally exposes the accumulating route. Prediction preservation and the absence of restoring force are first-order statements at the initial fit, not exact invariances of finite displacements or guarantees of indefinite linear decrease.

Connection to near-fit initialization. Away from a fit, $H_t = v_t v_t^\top + (g(b_t) - p)D^2 g(b_t)$ includes residual curvature, and the prediction direction changes with b_t. The two components can then couple: a direction that preserves prediction to first order need not have zero curvature. Section E controls these departures using the moving response and comparisons to fitted references, retaining the original history. It establishes positivity of the total reduction coefficient in the stated neighborhood, rather than separate favorable signs for the two components throughout a nonfitted trajectory.

D.8 A SINGLE PREDICTION COORDINATE GENERATES THE STRUCTURE

Suppose a sample loss has the form $\ell(g(\theta); Z)$ and $L(\theta) = \mathcal{L}(g(\theta))$, with $\mathcal{L}(q) = \mathbb{E}\ell(q; Z)$. Assume the required differentiation under expectation, local regularity, and bounded-field conditions. At $g_* = g(\theta_*)$, suppose

$$\mathcal{L}'(g_*) = 0, \qquad \ell_2 = \mathcal{L}''(g_*) > 0, \qquad v = \nabla g(\theta_*) \neq 0, \qquad V = \mathrm{Var}(\partial_q \ell(g_*; Z)) > 0.$$

The chain rule gives

$$\nabla L = \mathcal{L}'(g)\nabla g, \qquad \nabla^2 L = \mathcal{L}''(g)\nabla g \nabla g^\top + \mathcal{L}'(g)\nabla^2 g.$$

At θ_*, the centered sample gradient is $\partial_q \ell(g_*; Z)v$, hence $H_* = \ell_2 v v^\top$ and $S = V v v^\top$. Thus $n = v/\|v\|$, $h_* = \ell_2 \|v\|^2$, and $s_\xi = V\|v\|^2$.

For completeness, write $Q = D^2 g(\theta_*)$. Differentiating the full Hessian and contracting with n on both sides gives

$$a = \mathcal{L}'''(g_*)\|v\|^2 v + \ell_2\{2Qv + (n^\top Q n)v\}. \tag{D.20}$$

The last term comes from differentiating $\mathcal{L}'(g)Q$, which vanishes at θ_* but does not have a vanishing derivative there. Dropping it would differentiate only the Gauss–Newton term, not the raw Hessian. This construction explains the alignment assumption without assuming a favorable sign for a trace coefficient. It does not guarantee $a \neq 0$ for every prediction map.

D.9 PROOF FOR FULL-PARAMETER DEEP LINEAR NETWORKS

Let all matrix entries be coordinates of θ and let $g = W_D \cdots W_1$ be scalar. For half-squared loss and the data model in Theorem D.1, $x^2 = 1$ and independence imply

$$L_\sigma(\theta) = \tfrac{1}{2}(g(\theta) - p)^2 + \tfrac{1}{2}\sigma^2, \qquad \nabla \ell(\theta; x, \zeta) = (g - p)\nabla g - \sigma x \zeta \nabla g. \tag{D.21}$$

All parameter derivatives of L_σ are independent of σ. Use the fixed unit field $\xi(\theta; x, \zeta) = -x\zeta\nabla g(\theta)$ and set $\delta = \sigma$. This is the actual sample gradient, not an auxiliary additive perturbation. At θ_*, the covariance of the unit-amplitude noise field is $vv^\top$, so $s_\xi = h_* = \|v\|^2$: the label amplitude is not inserted a second time into the covariance coefficient.

The polynomial g is homogeneous of degree D in all its trained parameters together. Euler's identity and its derivative at a factorization of p give

$$\theta_*^\top v = Dp, \qquad Q\theta_* = (D - 1)v. \tag{D.22}$$

Since $p \neq 0$, $v \neq 0$. By Equation (D.20), the squared loss has

$$a = 2Qv + (n^\top Q n)v.$$

If $a = 0$, taking its inner product with n gives $3\|v\|(n^\top Q n) = 0$. Thus $n^\top Q n = 0$ and $Qv = 0$. But symmetry of Q and Equation (D.22) imply

$$\theta_*^\top Q v = v^\top Q \theta_* = (D - 1)\|v\|^2 > 0,$$

a contradiction. Hence $a \neq 0$ for every such factorization. Polynomials and bounded labels have all the required deterministic derivative bounds on each fixed compact ball. The top eigenvalue is simple at the fitted reference. Therefore all hypotheses and strictness conditions of Theorem D.2 hold, proving Theorem D.1.

No step uses equal layer norms, full-rank matrices, square layers, or active use of every hidden unit. The result applies at arbitrary fixed finite depth and widths, but its constants need not be uniform in these quantities or across unbounded factorizations. For Rademacher ζ and uniform Rademacher x, the data distribution can equivalently be the four fixed equally weighted samples $(x, y) = (\pm 1, p(\pm 1) \pm \sigma)$, sampled with replacement. Thus the realization does not require adding new label noise to the dataset at each training step. When $D = 1$, g is linear in its parameters, the squared-loss Hessian is constant, and no strict sharpness reduction of this type is available.

D.10 Exact calculations for the two-factor illustration

For $p = 1$ and $\theta_* = (1, 1)$, $v = (1, 1)$, $h_* = 2$, $n = (1, 1)/\sqrt{2}$, and

$$Q = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad a = (3, 3), \qquad a_n^2 = 18, \qquad a_\perp = 0.$$

The full Hessian is

$$\nabla^2 L_\sigma(u, v) = \begin{pmatrix} v^2 & 2uv - 1 \\ 2uv - 1 & u^2 \end{pmatrix}.$$

At $u = v = w$ its eigenvalues are $3w^2 - 1$ and $1 - w^2$, so the plotted raw sharpness always uses their maximum, not a selected branch.

Take two workers, $K = 1$, and $\mathcal{W} = J + \rho(I - J)$ with $0 < \rho < 1$. Then $s_\xi = 2$, $g_2 = \rho^2$, and $c = 1 - 2\eta$. For $0 < \eta < 1$, Equation (D.17) becomes

$$\Delta_N(\sigma) = -9\eta^3\rho^2\sigma^2 \sum_{\ell=0}^{N-2} (\rho^2 c^2)^\ell \frac{1 - c^{N-1-\ell}}{1 - c} + O_N(|\sigma|^3). \tag{D.23}$$

In particular, $\Delta_2 = -9\eta^3\rho^2\sigma^2 + O(|\sigma|^3)$.

For the following exact illustration, take independent Rademacher x and ζ, so the effective sample signs $\varepsilon_i = x_i\zeta_i$ are Rademacher. After the first update, the worker mean and half-difference are

$$b = 1 + \eta\sigma(\varepsilon_1 + \varepsilon_2)/2, \qquad s = \eta\sigma\rho(\varepsilon_1 - \varepsilon_2)/2.$$

The centralized state equals b. The population-gradient part of the next mean update differs by $-3\eta b s^2$. The fresh-field difference is $\eta\sigma s(\varepsilon_{1,\text{new}} - \varepsilon_{2,\text{new}})/2$, whose conditional expectation is zero. Thus $\mathbb{E}s^2 = \eta^2\sigma^2\rho^2/2$ and $\mathbb{E}(bs^2) = \mathbb{E}s^2$, recovering the leading second-step sharpness gap. This calculation explains the extra drift but does not replace the general propagation proof. For general iid centered unit-variance effective noise, the equality instead reads $\mathbb{E}(bs^2) = \eta^2\sigma^2\rho^2/2 + \eta^3\sigma^3\rho^2\mathbb{E}\varepsilon_i^3/4$. The additional cubic term does not change the second-order result.

For the archived two-factor numerical check, every current state is branched into the four equally probable sign pairs at each step. Worker gradients are evaluated before mixing; the centralized gradient uses the same two signs at its own state. Averaging all branches gives the exact finite-support expectation in floating-point arithmetic. The enumeration uses $\eta = 0.75$, $\rho = 0.5$, $N = 8$, and $\sigma \in \{0.05, 0.02, 0.005\}$. No smoothing, burn-in, trajectory filtering, or Monte Carlo confidence interval is used. The recorded normalized diagnostic divides the actual gap by σ^2 and compares it with the leading coefficient in Equation (D.23); agreement is a numerical check, not an independent proof.

D.11 Two prediction directions can reverse the result

Consider all three trained parameters $\theta = (z, u, v)$, predictor

$$f_\theta = \begin{pmatrix} \sqrt{2}(1 + z/4)u \\ (1 - 3z/2)v \end{pmatrix},$$

and labels $\delta(\zeta_1, \zeta_2)$ with independent bounded centered unit-variance coordinates. The half-squared loss has population gradient equal to that of

$$L_0 = (1 + z/4)^2u^2 + \tfrac{1}{2}(1 - 3z/2)^2v^2,$$

and centered unit field $\xi = -Df_\theta^\top\zeta$. It is centered at every state and satisfies the bounded local regularity conditions. At $\theta_* = 0$,

$$H = \operatorname{diag}(0, 2, 1), \qquad S = H, \qquad a = e_z, \qquad T_{zuu} = 1, \qquad T_{zvv} = -3.$$

The top eigenvalue is simple. The formula for a follows by differentiating the top u-curvature; all other components vanish at the origin.

Take $\eta = 1/2$ and any graph allowed in Assumption 1. Define $q = \eta^2 (mK)^{-1} \sum_{j=2}^{m} g_j > 0$. The u linear multiplier is zero, the v multiplier is $1/2$, and the source has no z component. The same modal covariance recursion used above yields, for $t \geq 1$,

$$(D_t)_{uu} = 2q, \qquad (D_t)_{vv} = \frac{\eta^2}{mK} \sum_{j=2}^{m} g_j \sum_{\ell=0}^{t-1} (g_j/4)^{\ell} \geq q.$$

Since $a^\top A^k = e_z^\top$,

$$a^\top A^{N-1-t} (T : D_t) = (D_t)_{uu} - 3(D_t)_{vv} \leq -q, \qquad 1 \leq t < N.$$

Therefore the coefficient in Equation (D.16) obeys $\mathcal{C}_N \leq -\eta q (N-1) < 0$. The local no-exit and Taylor proof applies here with a ball preserving the Hessian eigengap $2 - 1 = 1$ at the reference point. Taking a sufficiently small nonzero δ makes $\Delta_N(\delta) > 0$ for every fixed $N \geq 2$. This is an actual bounded-label-noise comparison, not a freely prescribed covariance with no sampling realization. The source is even proportional to the Hessian, but competing curvature directions change sharpness with opposite signs. Rank-one alignment in the positive result cannot therefore be replaced merely by covariance–Hessian proportionality.

E A HORIZON-INDEPENDENT NEAR-FIT NEIGHBORHOOD

This appendix gives the complete statement and proof of Theorem 1, including a finite-arithmetic construction of the initialization radius and a finite-horizon noise threshold. The fitted reference is arbitrary: all hidden matrix entries remain trainable, and no balance or rank assumption is introduced. We use h_* for the reference-fit sharpness throughout.

Theorem E.1 (Finite-step sharpness reduction in deep linear networks)**.** *Consider the decentralized and centralized SGD updates defined in Section 5, under the mixing and sampling conditions stated there. Let $f_\theta(x) = xg(\theta)$, where $g(\theta) = W_D \cdots W_1 \in \mathbb{R}$ is a bias-free deep linear network of fixed depth $D \geq 2$ and arbitrary finite hidden widths; train every matrix entry independently. The labels satisfy $y = px + \sigma\zeta$, where $p \neq 0$, x takes -1 and $+1$ with equal probability (a Rademacher input), and ζ is independent of x, bounded, centered, and has unit variance. Define the sample loss $\ell(\theta; x, y) = \frac{1}{2}(f_\theta(x) - y)^2$ and population loss $L_\sigma(\theta) = \mathbb{E}_{x,\zeta}[\ell(\theta; x, px + \sigma\zeta)]$. Here σ is the label-noise amplitude. Fix any reference θ_* with $g(\theta_*) = p$, which fits the conditional mean of the labels rather than each noisy label, and define its sharpness $h_* = \lambda_{\max}(\nabla^2 L_\sigma(\theta_*)) = \|\nabla g(\theta_*)\|^2$. For every $0 < \eta h_* < 2$, there are explicitly constructible constants $\varepsilon_* > 0$ and $c_* > 0$, independent of the horizon and noise amplitude, such that for every integer $N \geq 2$ there is $\sigma_N > 0$ for which every common initialization $\|\theta_0 - \theta_*\| \leq \varepsilon_*$ and every $0 < \sigma \leq \sigma_N$ satisfy*

$$\Delta_N(\theta_0, \sigma) := \mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\bar{\theta}_N^D)) - \mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\theta_N^C)) \leq -\frac{c_*}{4}\sigma^2 < 0.$$

Why the same neighborhood works at every horizon. The zero-noise trajectory from a nearby initialization approaches a nearby fit within a time-independent neighborhood. Worker-to-mean parameter deviations create an extra mean response whose history remains controlled relative to a fitted reference: the prediction-changing component has a restoring force, while the first-order prediction-preserving component can accumulate. A uniform late-time bound combines these effects; a finite-prefix bound covers all preceding steps without changing that late-time bound. This joins the two parts into a single initialization ball, retaining the response generated from the original starting point rather than restarting training at a fit.

E.1 THE MOVING-BASELINE RESPONSE

Let $g(\theta) = W_D \cdots W_1$, $g(\theta_*) = p \neq 0$, and $L(\theta) = \frac{1}{2}(g(\theta) - p)^2$. The population loss is $L_\sigma = L + \sigma^2/2$, so its Hessian does not depend explicitly on σ. Throughout, tensor norms are induced Euclidean multilinear norms, and matrix covariance errors use the nuclear norm $\|\cdot\|_*$. Set

$$v = \nabla g(\theta_*), \quad Q_* = D^2 g(\theta_*), \quad h_* = \|v\|^2, \quad \nu = v/\sqrt{h_*},$$
$$a = \nabla\lambda(\theta_*), \qquad \lambda(\theta) = \lambda_{\max}(D^2 L(\theta)).$$

Homogeneity gives $\theta_*^\top v = Dp$ and $Q_*\theta_* = (D-1)v$, hence $h_* > 0$. At the fit,

$$a = 2Q_* v + (\nu^\top Q_* \nu)v \neq 0. \tag{E.1}$$

Indeed, $a = 0$ first implies $\nu^\top Q_* \nu = 0$ by projection onto ν, and then $Q_* v = 0$, contradicting $\theta_*^\top Q_* v = (D-1)h_* > 0$. More quantitatively, $Q_* v = \frac{1}{2}(I - \nu\nu^\top/3)a$ implies $\|a\| \geq 2(D-1)h_*/\|\theta_*\|$. These identities apply at every nearby fit as well.

Fix an arbitrary common initialization θ_0. Let $b_0 = \theta_0$ and $b_{t+1} = b_t - \eta\nabla L(b_t)$ denote the zero-noise trajectory. Write $v_t = \nabla g(b_t)$, $Q_t = D^2 g(b_t)$, $e_t = g(b_t) - p$, $H_t = v_t v_t^\top + e_t Q_t$, $A_t = I - \eta H_t$, and $T_t = D^3 L(b_t)$. For the nonconsensus eigenvalues of $\mathcal{W}$, set

$$q_j = \lambda_j(\mathcal{W})^2 \quad (j = 2, \ldots, m), \qquad q_{\max} = \max_{j>1} q_j, \qquad Q_{\mathcal{W}} = \sum_{j>1} q_j.$$

The graph assumptions give $0 < q_{\max} < 1$ and $Q_{\mathcal{W}} > 0$.

For completeness, differentiate the actual sampled updates with respect to σ at zero. The effective worker noise is $z_{i,t} = K^{-1}\sum_{k=1}^K x_{i,t,k}\zeta_{i,t,k}$, with conditional mean zero and variance $1/K$. The first derivatives obey the linear update with Jacobian A_t and forcing $\eta z_{i,t} v_t$, followed by $\mathcal{W}$. Their worker mean equals the centralized first derivative pathwise. In an orthonormal graph eigenbasis,

let $D_{j,t}$ be $1/m$ times the covariance of the jth first-derivative mode, and set $D_t = \sum_{j>1} D_{j,t}$. Let $r_t = \mathbb{E}\,\partial_\sigma^2(\bar\theta_t^D - \theta_t^C)|_{\sigma=0}$. Independence of fresh samples eliminates the conditional expectation of the state-dependent noise derivative. The remaining recursions are

$$D_{j,t+1} = q_j\left(A_t D_{j,t} A_t^\top + \frac{\eta^2}{mK} v_t v_t^\top\right), \qquad D_{j,0} = 0, \tag{E.2}$$

$$r_{t+1} = A_t r_t - \eta T_t : D_t, \qquad r_0 = 0. \tag{E.3}$$

The common first derivative also cancels the quadratic term in the second-order expansion of sharpness in the difference of sharpnesses. Whenever the top eigenvalue is simple at the noise-free iterate b_N,

$$\Delta_N(\theta_0, \sigma) = -\frac{\sigma^2}{2}\mathcal{C}_N(\theta_0) + O_N(\sigma^4), \qquad \mathcal{C}_N(\theta_0) = -\nabla\lambda(b_N)^\top r_N. \tag{E.4}$$

The even remainder follows because $x\zeta$ is symmetric, including when ζ itself is not. The constants will be uniform over a compact initialization ball.

E.2 Explicit local bounds and an all-time deterministic neighborhood

Choose derivative bounds at θ_* $G_j \ge \|D^j g(\theta_*)\|$, with $G_1 = \sqrt{h_*}$ and $G_j = 0$ for $j > D$. Define $L_1 = 0$ and, for $2 \le k \le 2D$, $L_k = \frac{1}{2}\sum_{j=1}^{k-1}\binom{k}{j} G_j G_{k-j}$. They bound $D^k L(\theta_*)$ because the residual vanishes there. The following finite polynomials bound derivatives and their variation in a radius-r ball:

$$\begin{aligned}
\mathcal{E}(r) &= \sum_{j=1}^{D} G_j r^j/j!, & \mathcal{V}(r) &= \sum_{j=1}^{D-1} G_{j+1} r^j/j!,\\
G_j(r) &= \sum_{k=0}^{D-j} G_{j+k} r^k/k!, & \ell_j(r) &= \sum_{k=0}^{2D-j} L_{j+k} r^k/k!,\\
\mathcal{H}(r) &= \sum_{j=1}^{2D-2} L_{j+2} r^j/j!, & L_a(r) &= \ell_4(r) + \frac{2\ell_3(r)^2}{h_* - 2\mathcal{H}(r)}.
\end{aligned} \tag{E.5}$$

Terms with derivative order beyond the polynomial degree are zero. The bound L_a controls the Lipschitz constant of $\nabla\lambda$: Weyl's inequality gives a top-eigenvalue gap of at least $h_* - 2\mathcal{H}(r)$, and eigenvector differentiation gives the displayed denominator. If a convenient bound $G_1 > \sqrt{h_*}$ is used, the same formulas remain valid with that upper bound in positive derivative majorants and the exact $\sqrt{h_*}$ in lower bounds.

Choose radii $0 < e < R$ and define

$$\begin{aligned}
&E = \mathcal{E}(e), \quad V = G_1(R), \quad Q = G_2(R), \quad h_R^- = (\sqrt{h_*} - \mathcal{V}(R))^2, \quad h_R^+ = V^2,\\
&\rho = \max\{|1 - \eta h_R^-|, |1 - \eta h_R^+|\} + \eta^2 Q V^2 E/2, \quad s = \eta Q E,\\
&P = \exp\{s/(1-\rho)\}, \qquad r_f = e\sqrt{P}, \qquad B = \frac{\eta V E}{1-\rho}, \qquad \beta = q_{\max}(1+s)^2.
\end{aligned} \tag{E.6}$$

Here ρ is a deterministic residual bound, not the mixing parameter in the figures. The required static checks are

$$\sqrt{h_*} > \mathcal{V}(R), \quad \eta V^2 < 2, \quad h_* - 2\mathcal{H}(R) > 0, \quad \rho < 1, \quad r_f < R, \quad \beta < 1. \tag{E.7}$$

Lemma E.2 (Uniform deterministic attraction)**.** *Under Equation (E.7), every $\|\theta_0 - \theta_*\| \le e$ satisfies*

$$|e_t| \le E\rho^t, \qquad \|b_t - \theta_*\| \le z_t := e\exp\left\{\frac{s(1-\rho^t)}{2(1-\rho)}\right\} \le r_f. \tag{E.8}$$

There is a fitted endpoint θ_∞ with $\|b_t - \theta_\infty\| \le B\rho^t$. Moreover, $\|A_t\| \le 1 + s\rho^t$ and $\|A_{t-1}\cdots A_u\| \le P$.

Proof. For $d = b_t - \theta_*$, Taylor's formula gives $|d^\top v_t - e_t| \le Q\|d\|^2/2$. For every partial step $u\eta$, $0 \le u \le 1$,

$$\|d - u\eta e_t v_t\|^2 \le (1 + u\eta Q|e_t|)\|d\|^2 - u\eta(2 - u\eta V^2)e_t^2 \le (1 + u\eta Q|e_t|)\|d\|^2.$$

Thus the whole update segment remains in the proof ball under the induction bound. Taylor expansion of the residual on that segment gives

$$|e_{t+1}| \le \left(|1 - \eta\|v_t\|^2| + \eta^2 QV^2|e_t|/2\right)|e_t| \le \rho|e_t|.$$

Joint induction and $1 + x \le e^x$ yield Equation (E.8). The increments satisfy $\|b_{t+1} - b_t\| \le \eta V E\rho^t$ and are summable, proving endpoint convergence and its stated bound. Finally, $H_t = v_t v_t^\top + e_t Q_t$ and $\eta\|v_t\|^2 < 2$ imply $\|A_t\| \le 1 + \eta Q|e_t|$; multiplying these bounds proves the propagator estimate. No contraction in every parameter direction is assumed. □

E.3 A UNIFORM TAIL BOUND THAT RETAINS THE ORIGINAL HISTORY

Set

$$h_- = (\sqrt{h_*} - \mathcal{V}(r_f))^2, \quad h_+ = G_1(r_f)^2, \quad \bar{c} = \max\{|1 - \eta h_-|, |1 - \eta h_+|\}, \quad \kappa = 1 - \bar{c} > 0,$$

and

$$a_+ = \|a\| + L_a(r_f)r_f, \qquad a_- = \max\left\{\|a\| - L_a(r_f)r_f,\ \frac{2(D-1)h_-}{\|\theta_*\| + r_f}\right\} > 0.$$

They uniformly bound the curvature and sharpness gradient at the unknown fitted endpoint. Define

$$d_+ = \frac{\eta^2 h_+ Q_{\mathcal{W}}}{mK(1 - q_{\max}\bar{c}^2)}, \qquad j_- = \frac{\eta^2 h_- Q_{\mathcal{W}}}{mK}, \qquad A = \eta\ell_3(R)B,$$

$$F = q_{\max}d_+(2\bar{c}A + A^2) + \frac{2\eta^2 Q_{\mathcal{W}} VQB}{mK},$$

$$R_\Delta = P\left[\frac{A\eta d_+ a_+ \rho}{(1-\rho)^2} + \eta\left\{\frac{\ell_3(R)F}{(1-\beta)(1-\rho)} + \frac{\ell_4(R)Bd_+}{1-\rho}\right\}\right]. \tag{E.9}$$

Also set

$$\lambda_0 = \max\{\rho, \bar{c}, \beta\} < 1, \quad \mu = \eta j_- \kappa a_-^2, \quad X = A\eta d_+ a_+,$$
$$Y = AR_\Delta + \eta\ell_4(R)Bd_+, \quad Z = \eta\ell_3(R)F.$$

and

$$\mathcal{H}_N^{\text{hist}} = (X\lambda_0 + Z)\binom{N}{2}\lambda_0^{N-2} + YN\lambda_0^{N-1}.$$

For $N > 1 + \kappa$, the explicit tail bound is

$$\mathscr{L}_N = \mu - \frac{R_\Delta^2}{4\eta j_-(N - 1 - \kappa)} - a_+\mathcal{H}_N^{\text{hist}} - L_a(R)B\rho^N(\eta d_+ a_+ N + R_\Delta). \tag{E.10}$$

Lemma E.3 (Uniform coefficient lower bound in the tail)**.** *For every initialization in the radius-e ball, $\mathcal{C}_N(\theta_0) \ge \mathscr{L}_N$. Let*

$$N_{\text{dec}} = \max\left\{3, \left\lceil\frac{1 + \lambda_0}{1 - \lambda_0}\right\rceil\right\}.$$

Any $L \ge N_{\text{dec}}$ with $\mathscr{L}_L > 0$ certifies $\mathcal{C}_N(\theta_0) \ge \mathscr{L}_L > 0$ for every $N \ge L$.

Proof. Use a fitted reference response starting at θ_∞ with zero covariance and zero response at time zero; superscript f denotes this reference. This is an auxiliary comparison, not a reset of the original dynamics. Write ν_∞ for its prediction direction and $a_\infty = \nabla\lambda(\theta_\infty)$. Its covariance is $D_{j,t}^f = d_{j,t}^f \nu_\infty \nu_\infty^\top$, with total scalar coefficient at most d_+, and $\|r_t^f\| \le \eta d_+ a_+ t$. The deterministic attraction bounds give $\|A_t - A_\infty\| \le A\rho^t$ and $\|v_t - v_\infty\| \le QB\rho^t$. Subtracting Equation (E.2) from its fitted counterpart in nuclear norm yields

$$\delta D_{t+1} \le \beta\,\delta D_t + F\rho^t, \qquad \delta D_t := \sum_{j>1}\|D_{j,t} - D_{j,t}^f\|_*.$$

In bounding the fitted covariance term, A_∞ acts only on ν_∞, giving $2\bar{c}A + A^2$, not the full-space bound $2A + A^2$. The forcing difference uses $\|v_t v_t^\top - v_\infty v_\infty^\top\|_* \le 2VQB\rho^t$. Consequently

$$\delta D_t \le F \sum_{u<t} \beta^{t-1-u}\rho^u, \qquad \sum_{t\ge 0} \delta D_t \le \frac{F}{(1-\beta)(1-\rho)}.$$

For $\delta r_t = r_t - r_t^f$, subtracting Equation (E.3) gives

$$\delta r_{t+1} = A_t \delta r_t + (A_t - A_\infty) r_t^f - \eta\{T_t : (D_t - D_t^f) + (T_t - T_\infty) : D_t^f\}.$$

Use the propagator bound P, $\|T_t\| \le \ell_3(R)$, $\|T_t - T_\infty\| \le \ell_4(R)B\rho^t$, and $\sum t\rho^t = \rho/(1-\rho)^2$ to obtain $\|\delta r_t\| \le R_\Delta$.

The normal component has a sharper bound. Rewrite its recursion using A_∞ as the propagator, so the remaining term $(A_t - A_\infty)\delta r_t$ is bounded by $AR_\Delta\rho^t$. The forcing is bounded by $Xt\rho^t + Y\rho^t + Z\sum_{u<t}\beta^{t-1-u}\rho^u$. Convolution with $|1-\eta h_\infty| \le \bar{c}$ yields $|\nu_\infty^\top \delta r_N| \le \mathcal{H}_N^{\mathrm{hist}}$.

The positive fitted sum in Equation (D.17) gives

$$\mathcal{C}_N^f \ge \eta j_- \left\{\kappa a_{\infty,n}^2 + (N-1)\|a_{\infty,\perp}\|^2\right\}.$$

The tangent history error costs at most $R_\Delta\|a_{\infty,\perp}\|$. Since $\|a_\infty\| \ge a_-$, completing the square in $u = \|a_{\infty,\perp}\|$ gives

$$\eta j_-\{\kappa a_{\infty,n}^2 + (N-1)u^2\} - R_\Delta u \ge \mu - \frac{R_\Delta^2}{4\eta j_-(N-1-\kappa)}.$$

Finally, $\|\nabla\lambda(b_N) - a_\infty\| \le L_a(R)B\rho^N$ and $\|r_N\| \le \eta d_+ a_+ N + R_\Delta$ give Equation (E.10). For $N \ge N_{\mathrm{dec}}$, each loss term in that bound is nonincreasing: this follows by taking successive ratios for $N\lambda_0^{N-1}$ and $\binom{N}{2}\lambda_0^{N-2}$, and using $\rho \le \lambda_0$. Thus $\mathscr{L}_N$ is nondecreasing and tends to $\mu > 0$. □

A finite search suffices to find L. For an explicit search limit put $z = \sqrt{\lambda_0}$ and $\log_+(x) = \max\{0, \log x\}$. Define

$$C_{\mathrm{geo}} = a_+\left\{\frac{X\lambda_0 + Z}{z^2(1-z)^3} + \frac{Y}{z(1-z)^2}\right\} + L_a(R)B\left\{\frac{\eta d_+ a_+ z}{(1-z)^2} + R_\Delta\right\}.$$

The identities $\sum_{n\ge 0} nz^n = z/(1-z)^2$ and $\sum_{n\ge 0}\binom{n}{2}z^n = z^2/(1-z)^3$ bound the exponential losses by $C_{\mathrm{geo}}z^N$. Hence

$$L_{\mathrm{safe}} = \max\left\{N_{\mathrm{dec}}, \left\lceil 1 + \kappa + \frac{R_\Delta^2}{\eta j_- \mu}\right\rceil, \left\lceil \frac{\log_+(4C_{\mathrm{geo}}/\mu)}{-\log z}\right\rceil\right\} \tag{E.11}$$

gives $\mathscr{L}_{L_{\mathrm{safe}}} \ge \mu/2$. One can instead use an earlier L with a positive margin. This L is a proof interface; the final theorem will include every $N \ge 2$.

E.4 The finite prefix and the explicit initialization radius

Keep e, R, L fixed once the tail has been certified. At the original fit, let $c = 1 - \eta h_*$, $a_n = \nu^\top a$, and $a_\perp = (I - \nu\nu^\top)a$. The known fitted response is

$$d_{j,t}^* = \frac{\eta^2 h_*}{mK} q_j \sum_{\ell=0}^{t-1} (q_j c^2)^\ell, \qquad d_t^* = \sum_{j>1} d_{j,t}^*,$$
$$r_t^* = -\eta \sum_{u<t} d_u^*\{c^{t-1-u} a_n \nu + a_\perp\}, \qquad R_t^* = \|r_t^*\|, \qquad C_t^* = -a^\top r_t^* > 0 \quad (t \ge 2). \tag{E.12}$$

Define $\alpha_t = 1 + s\rho^t$, $\delta_t = \eta\mathcal{H}(z_t)$, and $v_t^e = \mathcal{V}(z_t)$. The following finite sums involve only derivatives evaluated at the fitted reference point and the fixed large-ball constants:

$$f_{j,t} = d_{j,t}^*(2|c|\delta_t + \delta_t^2) + \frac{\eta^2}{mK} v_t^e (2\sqrt{h_*} + v_t^e),$$

$$E^D_{j,t} = \sum_{u<t} q_j^{t-u} f_{j,u} \prod_{k=u+1}^{t-1} \alpha_k^2,$$
$$E^r_t = \sum_{u<t} \left[\delta_u R^*_u + \eta \left\{ \ell_3(z_u) \sum_{j>1} E^D_{j,u} + (\ell_3(z_u) - L_3) d^*_u \right\} \right] \prod_{k=u+1}^{t-1} \alpha_k,$$
$$\Lambda_t(e) = L_a(z_t) z_t R^*_t + (\|a\| + L_a(z_t) z_t) E^r_t. \tag{E.13}$$

Empty products equal one and empty sums equal zero. To verify the bounds, subtract the original-fit covariance recurrence from Equation (E.2). The covariance error is at most $q_j \alpha_t^2$ times its preceding value plus $q_j f_{j,t}$. Subtracting the mean-response recurrences similarly gives the forcing in E^r_t. Finally, split the difference in sharpness-reduction coefficients into changes in the sharpness gradient and changes in the mean response to obtain $|\mathcal{C}_t(\theta_0) - C^*_t| \le \Lambda_t(e)$.

For a smaller radius $\varepsilon \le e$, the deterministic distance bound improves to $(\varepsilon/e) z_t$ while keeping ρ, s, P unchanged. Every source error in these comparisons has a nonnegative power series in this distance with no constant term. The same is true of the spectral derivative majorant inside the positive-gap ball. Therefore

$$|\mathcal{C}_t(\theta_0) - C^*_t| \le (\varepsilon/e) \Lambda_t(e) \quad \text{for all } \|\theta_0 - \theta_*\| \le \varepsilon. \tag{E.14}$$

Choose any safety fraction $0 < \vartheta < 1$, and set

$$\varepsilon_* = \min \left\{ e,\ (1-\vartheta) e \min_{2 \le t < L} \frac{C^*_t}{\Lambda_t(e)} \right\}. \tag{E.15}$$

A ratio with zero denominator is interpreted as $+\infty$. This radius is positive and independent of N. The prefix obeys $\mathcal{C}_t \ge \vartheta C^*_t > 0$. The tail bound already proved on the larger ball remains valid on the smaller one, without recomputing L. Thus

$$c_* = \min \left\{ \mathscr{L}_L,\ \min_{2 \le t < L} \left[C^*_t - \frac{\varepsilon_*}{e} \Lambda_t(e) \right] \right\} > 0 \tag{E.16}$$

satisfies $\mathcal{C}_N(\theta_0) \ge c_*$ for every $N \ge 2$ and every initialization in the ball. A smaller reported constant is also valid. This finite-prefix construction avoids taking an infinite minimum of radii that might approach zero.

A positive starting radius always exists. These checks impose no additional structural assumption on the fitted reference θ_*. For a direct construction, use the compatible unit-ball majorants $V_0 = G_1(1)$, $Q_0 = G_2(1)$, and $M_3 = \ell_3(1)$ from Equation (E.5), and let $\delta = 1 - |1 - \eta h_*| > 0$. Take

$$R_0 = \min \left\{ \frac{1}{2}, \frac{h_*}{8M_3}, \frac{\sqrt{h_*}}{2Q_0}, \frac{\delta}{4\eta Q_0 (2\sqrt{h_*} + Q_0)} \right\},$$
$$e_0 = \min \left\{ \frac{R_0}{4}, \frac{\delta}{2\eta^2 Q_0 V_0^3}, \frac{\delta}{4\eta Q_0 V_0}, \frac{q_{\max}^{-1/2} - 1}{2\eta Q_0 V_0} \right\}. \tag{E.17}$$

Using $E \le V_0 e_0$, $\mathcal{V}(R_0) \le Q_0 R_0$, and $\mathcal{H}(R_0) \le M_3 R_0$ verifies Equation (E.7). Specifically, variation of $\|\nabla g\|^2$ is at most $Q_0(2\sqrt{h_*} + Q_0) R_0 \le \delta/(4\eta)$, the nonlinear residual correction is at most $\delta/4$, and hence $\rho \le 1 - \delta/2$. Also $s \le \delta/4$, so $r_f \le e_0 e^{1/4} < R_0$, and $\beta \le (1 + \sqrt{q_{\max}})^2/4 < 1$. The eigenvalue gap is positive by $\mathcal{H}(R_0) \le h_*/8$. Equations (E.11) and (E.15) then finish a finite construction for every permitted reference, graph, and step size. Tighter checked local bounds can give a larger radius.

E.5 From the coefficient to the actual noisy algorithms

For each fixed N, sampled updates are finite polynomial compositions in bounded samples and initial parameters. Compactness of the initialization ball and the positive deterministic eigenvalue gap give a common noise neighborhood where the final sharpness is smooth. Equation (E.4) is therefore uniform over the ball. We next make its noise threshold computable, without requiring a stochastic invariant region for infinite time.

Let $|\zeta| \le Z_0$. Using the deterministic envelope z_t, define

$$\phi_t(z) = \sum_{k=2}^{2D-1} \ell_{k+1}(z_t) z^k/k!, \qquad \psi_t(z) = \sum_{k=0}^{D-1} G_{k+1}(z_t) z^k/k!.$$

Let $\chi_{\mathcal{W}} = \max_i \sum_j |\mathcal{W}_{ij} - 1/m|$. The nonnegative-coefficient polynomials below start from $\mathsf{R}_0 = \mathsf{S}_0 = \mathsf{E}_0 = 0$:

$$\begin{aligned}
\mathsf{R}_{t+1} &= \alpha_t \mathsf{R}_t + \eta\phi_t(\mathsf{R}_t) + \eta Z_0 z\psi_t(\mathsf{R}_t),\\
\mathsf{S}_{t+1} &= \chi_{\mathcal{W}}\{(\alpha_t + \eta\phi_t'(\mathsf{R}_t))\mathsf{S}_t + \eta Z_0 z\psi_t(\mathsf{R}_t)\},\\
\mathsf{E}_{t+1} &= (\alpha_t + \eta\phi_t'(\mathsf{R}_t))\mathsf{E}_t + \eta\{\phi_t(\mathsf{R}_t + \mathsf{S}_t) - \phi_t(\mathsf{R}_t) - \phi_t'(\mathsf{R}_t)\mathsf{S}_t\}\\
&\quad + \eta Z_0 z\{\psi_t'(\mathsf{R}_t)\mathsf{E}_t + \psi_t(\mathsf{R}_t + \mathsf{S}_t) - \psi_t(\mathsf{R}_t)\}. \qquad \text{(E.18)}
\end{aligned}$$

For $z = |\sigma|$, these bound, respectively, the maximum noise-induced displacement from b_t, the maximum worker-to-mean parameter deviation, and the paired mean–central parameter difference. The S recurrence follows by applying $\mathcal{W} - J$ after the local update. For E, terms linear in the worker-to-mean parameter deviations cancel in the mean population gradient, leaving the quadratic bracket; the last bracket preserves the own-state difference of stochastic gradients. Taylor's formula and induction prove these bounds, including coefficientwise bounds for the analytic series. All bracket differences are evaluated through their nonnegative binomial expansions. In particular, E_t starts at order two.

Choose a reference amplitude $\bar\sigma_N > 0$ for which

$$g_N^{\mathrm{tube}} := h_* - 2\mathcal{H}(z_N + \mathsf{R}_N(\bar\sigma_N)) > 0.$$

Write $R_N^{\mathrm{noise}} = \mathsf{R}_N(\bar\sigma_N)$, $E_N^{\mathrm{noise}} = \mathsf{E}_N(\bar\sigma_N)$, $r_1 = [z]\mathsf{R}_N$, and $e_j = [z^j]\mathsf{E}_N$. At the noise-free iterate b_N, set $A_1 = \|a\| + L_a(z_N)z_N$ and $A_2 = L_a(z_N)$. A third-derivative bound for the sharpness over the full noise tube, of radius $r_o = z_N + R_N^{\mathrm{noise}}$ about θ_*, is

$$A_3 = \ell_5(r_o) + \frac{6\ell_3(r_o)\ell_4(r_o)}{g_N^{\mathrm{tube}}} + \frac{12\ell_3(r_o)^3}{(g_N^{\mathrm{tube}})^2}.$$

Eigenvalue differentiation with the reduced resolvent gives this bound. The paired fourth-order remainder constant is

$$M_{4,N} = A_1\frac{E_N^{\mathrm{noise}} - e_2\bar\sigma_N^2 - e_3\bar\sigma_N^3}{\bar\sigma_N^4} + A_2\frac{R_N^{\mathrm{noise}}E_N^{\mathrm{noise}} - r_1e_2\bar\sigma_N^3}{\bar\sigma_N^4} + \frac{A_3(R_N^{\mathrm{noise}})^2E_N^{\mathrm{noise}}}{2\bar\sigma_N^4}. \qquad \text{(E.19)}$$

To check it, for the two noisy parameter displacements x, y, expand the corresponding difference in sharpness directly:

$$\begin{aligned}
\lambda(b+x) - \lambda(b+y) &= \nabla\lambda(b)^\top(x-y) + D^2\lambda(b)[(x+y)/2, x-y] + \mathcal{Q},\\
|\mathcal{Q}| &\le A_3(R_N^{\mathrm{noise}})^2E_N^{\mathrm{noise}}/2.
\end{aligned}$$

The first term starts at order two and the second at order three. Symmetry of the effective label noise cancels odd expected coefficients. Nonnegative polynomial tails evaluated at $\bar\sigma_N$ bound the remaining coefficients at every smaller amplitude. Thus

$$\left|\Delta_N + \frac{\sigma^2}{2}\mathcal{C}_N\right| \le M_{4,N}\sigma^4 \qquad (|\sigma| \le \bar\sigma_N).$$

Let $\underline{C}_N = C_N^* - (\varepsilon_*/e)\Lambda_N(e)$ for $2 \le N < L$ and $\underline{C}_N = \mathscr{L}_N$ for $N \ge L$. Then one may take

$$\sigma_N = \min\left\{\bar\sigma_N,\ \sqrt{\frac{\underline{C}_N}{4M_{4,N}}}\right\}, \qquad \text{(E.20)}$$

with the usual infinite-ratio convention if $M_{4,N} = 0$. This yields $\Delta_N \le -\underline{C}_N\sigma^2/4 \le -c_*\sigma^2/4$ and completes the proof of Theorem E.1.

A positive reference amplitude also has a direct construction. Choose $R_o > r_f$ with $h_* - 2\mathcal{H}(R_o) > 0$, and set

$$d = (R_o - r_f)/3, \quad b_0 = \eta Z_0 G_1(R_o), \quad K_0 = \eta\ell_3(R_o)/2,$$

$$A_0 = 2Pb_0, \quad \chi = \min\{1, d/A_0, 3/(4P^2K_0b_0)\}.$$

For $\bar{\sigma}_N = \chi/N^2$, unfolding the R recurrence gives $\mathsf{R}_t \le A_0\bar{\sigma}_N t \le d$ for $t \le N$: the linear forcing is at most $Pb_0\bar{\sigma}_N t$ and $\sum_{u<t} u^2 \le t^3/3$ bounds the nonlinear part by the remaining half. The tube check follows. This construction supplies a positive sufficient noise threshold, not a universal scaling law for the largest admissible noise.

E.6 An explicit width-two, depth-three instance

Consider

$$W_{1,*} = (1,1)^\top, \qquad W_{2,*} = \begin{pmatrix} 4/5 & 1/5 \\ 1/10 & 9/10 \end{pmatrix}, \qquad W_{3,*} = (3/5, 2/5).$$

Here $g(\theta_*) = 1$, $h_* = 2213/625$, $\eta = 1875/4426$, $m = 2$, $K = 1$, and $\mathcal{W} = \left(\begin{smallmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{smallmatrix}\right)$. Thus $\eta h_* = 3/2$ and $q_{\max} = 1/4$. Convenient derivative bounds at the fitted reference point are $G_1 = 941/500$, $G_2 = 17/8$, and $G_3 = 7/6$. The first follows directly from h_*. For G_2, bound the off-diagonal layer-block coefficients by the matrix operator norm of the remaining layer and take the operator norm of that three-by-three coefficient matrix (approximately $2.12344 < 17/8$); for G_3, the six permutations of one derivative per layer give $\|D^3 g\| \le 2/\sqrt{3} < 7/6$. These are bounds in the full eight-dimensional parameter space.

With $e = 0.01$, $R = 0.01025$, the explicit expressions give

$$\rho \simeq 0.548662, \quad r_f \simeq 0.010191625 < R, \quad \beta \simeq 0.258640, \quad h_* - 2\mathcal{H}(R) > 3.2924.$$

The history bound is $R_\Delta \simeq 0.471057$, $\mu \simeq 1.628099$, $N_{\mathrm{dec}} = 4$, and $\mathscr{L}_8 \simeq 0.0870644 > 0$. Thus $L = 8$ works (the more conservative closed limit is $L_{\mathrm{safe}} = 21$). The prefix lower bounds at $N = 2, \dots, 7$ are approximately $3.77544, 1.69381, 2.17659, 1.25646, 0.95944, 0.26875$. For $\vartheta = 1/20$, the prefix radius cap exceeds 0.0101397, so

$$\varepsilon_* = 0.01, \qquad c_* = 0.08$$

are conservative choices. The radius is about 0.499% of $\|\theta_*\|$, and it allows $|g(\theta_0) - 1| < 0.018927$. It is a sufficient full-ball radius, not the largest possible one.

For unit-bounded effective noise, evaluation of the paired remainder gives the following noise thresholds valid for every observation time $2 \le N \le T$. Values shown for $M_{4,T}$ are rounded upward; the final column is rounded down.

T	$\sigma_{\le T}$	$M_{4,T}$	$\min_{2\le N\le T}(\underline{C}_N - 4M_{4,N}\sigma^2_{\le T})$
2	0.005	3519.564	3.4234
4	0.0015	116198.485	1.1308
16	$5 \cdot 10^{-5}$	$7.749135 \cdot 10^7$	0.05775
32	10^{-5}	$2.134853 \cdot 10^9$	0.08589
64	$2 \cdot 10^{-6}$	$6.279982 \cdot 10^{10}$	0.08701

The top Hessian eigengap exceeds 3.09 throughout each noisy-iterate neighborhood. These are numerical evaluations of the explicit sufficient inequalities, independently reconstructed with high-precision arithmetic; the symbolic proof supplies the all-time coefficient claim. The relatively small sufficient noise thresholds for long windows should not be confused with the larger noise levels used to test the mechanism in Figure 4. In particular, dividing the gap by σ^2 displays its leading coefficient, not its absolute magnitude.

E.7 Why common worker initialization is substantive

The theorem does not require equal layers or entries, but it does require a shared starting network. To see why arbitrary worker-specific starts cannot be admitted, consider two uvx workers initialized at $(1+\tau, 1-\tau)$ and $(1-\tau, 1+\tau)$. Initialize the centralized comparator at their mean $(1,1)$, take $\mathcal{W} = J + \rho_{\mathcal{W}}(I - J)$ with $0 < \rho_{\mathcal{W}} < 1$, and $0 < \eta < 1/2$. At zero training noise, worker states remain $(b_t + d_t, b_t - d_t)$ and its swap, with

$$b_{t+1} = b_t - \eta(b_t^2 - d_t^2 - 1)b_t, \qquad d_{t+1} = \rho_{\mathcal{W}} d_t[1 + \eta(b_t^2 - d_t^2 - 1)].$$

For each fixed N, induction at small τ gives

$$d_t = \rho_{\mathcal{W}}^t \tau + O(\tau^3), \qquad b_N = 1 + \eta\tau^2 \sum_{t=0}^{N-1} (1-2\eta)^{N-1-t} \rho_{\mathcal{W}}^{2t} + O(\tau^4).$$

The centralized point remains fitted. Since the largest eigenvalue of the full Hessian at (b, b) is $3b^2 - 1$ for b near one,

$$\Delta_N(\tau, 0) = 6\eta\tau^2 \sum_{t=0}^{N-1} (1-2\eta)^{N-1-t} \rho_{\mathcal{W}}^{2t} + O(\tau^4) > 0$$

for every fixed $N \geq 1$ and sufficiently small nonzero τ. Continuity preserves this adverse sign at sufficiently small training noise. Thus shrinking the noise alone cannot repair unrestricted differences between worker initializations.

F Discrete Phase and Graph-Dependent Slow Dynamics

This appendix identifies the long-time continuation of the tangential response in Subsection D.7.1. The result concerns the original gradient-then-mix iterates and centralized SGD, whose gradients are evaluated at its own parameters. The phase maps introduced below assign each admitted state its zero-noise limiting point. They are used only in the analysis and are never applied to either algorithm.

Theorem 1 gives a comparison of full-Hessian sharpness at the actual iterates for every fixed horizon, with a horizon-dependent noise threshold. Here one small-noise limit controls a local interval of the rescaled time $s = N\sigma^2$. The uniform sharpness-tracking estimate also yields a comparison at some intermediate horizons (Corollary F.3). These results do not supply one noise threshold covering every $2 \le N \le S_0/\sigma^2$ or a uniform second-order composite expansion.

F.1 Local phase and its normal Hessian

Write

$$L_0(\theta) = \tfrac{1}{2}(g(\theta) - p)^2, \qquad \mathcal{M} = \{\theta : g(\theta) = p\}, \qquad h(\theta) = \|\nabla g(\theta)\|^2.$$

Since $p \neq 0$ and g is homogeneous, ∇g does not vanish on $\mathcal{M}$. On this manifold,

$$\nabla^2 L_0 = \nabla g \nabla g^\top, \qquad \lambda_{\max}(\nabla^2 L_0) = h, \qquad \operatorname{grad}_{\mathcal{M}} h = P \nabla h, \tag{F.1}$$

where $P = I - \nu\nu^\top$ and $\nu = \nabla g/\sqrt{h}$.

For the centralized deterministic map, set $F_C(\theta) = \theta - \eta \nabla L_0(\theta)$. For the decentralized map, F_D applies this update at every worker and then multiplies the worker index by $\mathcal{W}$. It has the fixed-point manifold $\mathcal{M}_D = \{\theta \mathbf{1}^\top : \theta \in \mathcal{M}\}$, the consensus copy of $\mathcal{M}$, which is locally normally attracting under the conditions below: directions transverse to this fixed-point manifold contract. We use the worker-normalized inner product $\langle X, Z\rangle_m = m^{-1} \sum_i \langle X_i, Z_i\rangle$ only in this proof.

Lemma F.1 (Local discrete phase)**.** *Let $\mathcal{K}$ be a compact subset of $\mathcal{M}$ on which*

$$0 < h_- \le h \le h_+, \qquad \eta h_+ < 2, \qquad \|\mathcal{W} - J\|_2 < 1. \tag{F.2}$$

After enlarging $\mathcal{K}$ slightly inside $\mathcal{M}$, there are tubular neighborhoods of $\mathcal{K}$ and its consensus copy and C^3 maps Π_C, Π_D into $\mathcal{M}$ such that

$$\Pi_A \circ F_A = \Pi_A, \qquad \Pi_A(\iota_A(y)) = y, \qquad A \in \{C, D\}.$$

For every admitted point, $\Pi_A(X)$ is the endpoint of the zero-noise orbit starting at X. In particular, a common initialization has the same phase for the two algorithms.

Let $c = 1 - \eta h(y)$, let e_j be a unit eigenvector of $\mathcal{W}$ with eigenvalue λ_j, and put $q_j = \lambda_j^2$. Number the consensus mode by $j = 1$, so $q_1 = 1$. For

$$E_j = \sqrt{m}\, \nu e_j^\top$$

and for the centralized unit normal ν,

$$D^2\Pi_C[\nu, \nu] = -\frac{\eta}{1 - c^2} \operatorname{grad}_{\mathcal{M}} h, \tag{F.3}$$

$$D^2\Pi_D[E_j, E_j] = -\frac{\eta}{1 - q_j c^2} \operatorname{grad}_{\mathcal{M}} h. \tag{F.4}$$

Mixed graph modes have zero contraction against the worker-independent noise covariance used below.

Proof. At a fitted point, DF_C is the identity on $T_y\mathcal{M}$ and multiplies ν by c. The derivative of F_D multiplies a tangent graph mode by λ_j and a normal graph mode by $\lambda_j c$. Hence every direction transverse to the fixed manifold has modulus at most

$$r_0 = \max\{\sup_{\mathcal{K}} |1 - \eta h|, \|\mathcal{W} - J\|_2\} < 1.$$

Choose local coordinates (y, z) adapted to the tangent and stable directions, with y along the fitted manifold and z transverse to it. Let $r \in (r_0, 1)$. Shrinking the tube gives

$$y^+ = y + p(y, z), \qquad z^+ = A(y)z + r_2(y, z), \qquad \|A(y)\| \le r_0, \tag{F.5}$$

with $\|p(y,z)\|+\|r_2(y,z)\| \le L\|z\|^2$. Taking the tube radius at most $(r-r_0)/L$ yields $\|z_t\| \le r^t\|z_0\|$. Moreover,

$$\sum_{t\ge 0}\|y_{t+1}-y_t\| \le \frac{L\|z_0\|^2}{1-r^2}.$$

Thus y_t has a limit, which defines Π_A and proves phase invariance.

For completeness, differentiability follows from the same contraction rather than from invertibility of F_A. Choose $r_0 < \rho < r < 1$ and differentiate (F.5). On a sufficiently small closed tube, the first variational recurrences satisfy

$$\|Dz_{t+1}\| \le \rho\|Dz_t\| + C\|z_t\|\|Dy_t\|, \qquad \|Dy_{t+1}-Dy_t\| \le C\{\|z_t\|\|Dz_t\| + \|z_t\|^2\|Dy_t\|\}.$$

The higher derivatives retain the same center–stable coupling. More explicitly, put $a_{k,t} = \|D^k y_t\|$, $b_{k,t} = \|D^k z_t\|$, and $d_t = \|z_t\| \le \delta r^t$. Differentiating the map equations, for $k \le 3$, gives

$$\begin{aligned} a_{k,t+1} &\le (1+Cd_t^2)a_{k,t} + Cd_t b_{k,t} + F_{k,t},\\ b_{k,t+1} &\le \rho b_{k,t} + Cd_t a_{k,t} + G_{k,t}. \end{aligned}$$

At $k=1$ the two forcing terms vanish. At higher orders they contain only derivatives of y_t and z_t with respect to the initial state, of orders below k. Terms with a differentiated stable input (the z coordinate) contain a decaying lower-order derivative of z_t; derivatives with respect to y alone of p and $A(y)z + r_2$ vanish at $z=0$. Induction therefore bounds $F_{k,t}+G_{k,t}$ by $C_k(1+t)^k r^t$. For $M_{k,T} = \max_{t\le T} a_{k,t}$, geometric convolution in the stable inequality gives

$$b_{k,t} \le C_k(1+t)^k r^t + C\delta M_{k,T} r^t.$$

Summing the center inequality and using the geometric sums yields $M_{k,T} \le C'_k + C'\delta^2 M_{k,T}$. Shrink the tube so that $C'\delta^2 \le 1/2$. Induction on k then bounds the derivatives $D^k y_t$ independently of T and proves, for $k \le 3$,

$$\|D^k z_t\| \le C_k(1+t)^k r^t, \qquad \|D^k(y_{t+1}-y_t)\| \le C_k(1+t)^k r^t.$$

The constants are obtained recursively from the chart derivative bounds and $\sum_{t\ge 0}(1+t)^k r^t$. The derivative series converge uniformly, proving that Π_A is C^3 with uniform bounds on the closed tube.

Differentiate $\Pi_A \circ F_A = \Pi_A$ twice at a fixed point. For stable vectors U, V,

$$D^2\Pi_A[U,V] - D^2\Pi_A[DF_A U, DF_A V] = D\Pi_A D^2 F_A[U,V]. \tag{F.6}$$

The phase derivative is P on a single-model perturbation and, for a worker state, is P applied to its worker mean. At a fit, $D^2F_C[\nu,\nu] = -\eta D^3 L_0[\nu,\nu]$. The full vector $D^3L_0[\nu,\nu]$ contains an additional normal term, but

$$PD^3L_0[\nu,\nu] = P\nabla h = \operatorname{grad}_{\mathcal{M}} h.$$

Since $DF_C\nu = c\nu$, (F.6) gives (F.3).

For E_j, the worker average of the second derivative of the unmixed update is $-\eta D^3 L_0[\nu,\nu]$: the ith squared amplitude is me_{ij}^2 and $m^{-1}\sum_i me_{ij}^2 = 1$. Mixing preserves this average, and the phase derivative projects it to $-\eta \operatorname{grad}_{\mathcal{M}} h$, while $DF_D E_j = \lambda_j c E_j$. Equation (F.6) now gives (F.4). For two distinct graph modes the corresponding average contains $\sum_i e_{ij}e_{ik} = 0$. □

F.2 The slow law for the original stochastic iterates

Write the stochastic update as $X_A^+ = F_A(X_A) + \sigma G_A(X_A,\omega)$. Here $G_C = \eta\bar{\omega}\nabla g(\theta^C)$ and the ith decentralized component is $\eta\sum_\ell \mathcal{W}_{i\ell}\omega_\ell \nabla g(\theta_\ell^D)$, where $\omega_i = K^{-1}\sum_{k=1}^K x_{ik}\zeta_{ik}$ and $\bar{\omega} = m^{-1}\sum_i \omega_i$. Each ω_i is centered, bounded, and has variance $1/K$. At a fit, projecting the postmix injection onto E_j gives variance

$$\frac{\eta^2 h q_j}{mK}\sigma^2. \tag{F.7}$$

For the central comparator only the consensus variance $\eta^2 h\sigma^2/(mK)$ remains. Combining Eqs. (F.3), (F.4) and (F.7) defines

$$b_C(h) = \frac{\eta^3 h}{2mK(1-c^2)} = \frac{\eta^2}{2mK(2-\eta h)},$$

$$b_D(h) = b_C(h) + \frac{\eta^3 h}{2mK} \sum_{j=2}^{m} \frac{q_j}{1 - q_j c^2}, \qquad c = 1 - \eta h. \tag{F.8}$$

Remark 1 (Learning rate and mixing)**.** For our update, which takes a gradient step before mixing, each nonconsensus mode along the prediction normal and a worker-space eigenvector of $\mathcal{W}$ propagates with multiplier $\lambda_j(\mathcal{W})(1 - \eta h)$, and its fresh noise is also multiplied by $\lambda_j(\mathcal{W})$. These effects give the spectral factors in Equation (4). In the update analyzed by Zhu et al. (2023), current weights are averaged but local gradients are not mixed, giving multiplier $\lambda_j(\mathcal{W}) - \eta h$ at the same fitted linearization. At a common fitted position, b_D/b_C is largest at $\eta h = 1$; this does not locate the maximum equal-time sharpness gap or the strongest perturbation-sensitivity reduction.

Theorem F.2 (Local slow limit of the actual updates)**.** *Assume the model, common initialization, matched sampling, and centralized SGD comparator of Theorem E.1. Assume also that the effective label noise is bounded. Let a compact family of common initializations lie in the phase tubes of Lemma F.1; suppose their common zero-noise orbits remain in the tubes and converge to phases $y^\dagger$ a fixed positive distance from the boundary of $\mathcal{K}$. All constants below may be chosen uniformly over this family.*

There is $S_0 > 0$ such that the solutions

$$\dot{y}_A(s) = -b_A(h(y_A(s)))\,\mathrm{grad}_{\mathcal{M}}\, h(y_A(s)), \qquad y_A(0) = y^\dagger, \qquad A \in \{C, D\}, \tag{F.9}$$

remain in $\mathcal{K}$ for $0 \le s \le S_0$. As $\sigma \to 0$, the unmodified stochastic iterates satisfy, for every $0 < s_- < S \le S_0$,

$$\sup_{s_- \le s \le S} \left| \mathbb{E}\lambda_{\max}\Big(\nabla^2 L_\sigma(\bar{\theta}^D_{\lfloor s/\sigma^2 \rfloor})\Big) - h(y_D(s)) \right| \longrightarrow 0, \tag{F.10}$$

and the analogous statement holds for θ^C and y_C. If every initialization is fitted, one may take $s_- = 0$.

Consequently, for each fixed admitted initialization with $\mathrm{grad}_{\mathcal{M}}\, h(y^\dagger) \neq 0$ and every fixed $0 < s_- < S \le S_0$, there is a noise threshold such that

$$\mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\bar{\theta}^D_N)) < \mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(\theta^C_N)) \tag{F.11}$$

for every integer N with $s_- \le N\sigma^2 \le S$.

Proof. We give the estimates because the horizon grows as σ^{-2}. Let $X_{A,t}$ denote the full state and $Y_{A,t} = \Pi_A(X_{A,t})$. Here d_t is the stable-coordinate norm, uniformly comparable on the closed tube to $\|X_{A,t} - \iota_A(\Pi_A(X_{A,t}))\|$. Reserve a slightly larger chart collar containing the segments $F_A(X) + u\sigma G_A(X, \omega)$, $0 \le u \le 1$, used in Taylor expansion; bounded noise and a smaller fixed noise threshold ensure this inclusion. On a smaller preliminary tube, transverse contraction and bounded noise give, pathwise until exit,

$$d_{t+1} \le r d_t + C_G|\sigma|, \qquad d_t \le r^t d_0 + \frac{C_G|\sigma|}{1 - r}. \tag{F.12}$$

Phase invariance and $D\Pi_A G_A = 0$ on the fitted consensus manifold give

$$\|Y_{t+1} - Y_t\| \le C_1|\sigma| d_t + C_2\sigma^2. \tag{F.13}$$

Let the initialization family lie in an inner transverse tube of radius δ_0, while the preliminary tube has radius $\delta > \delta_0$. First choose $|\sigma|$ so that $C_G|\sigma|/(1 - r) < \delta - \delta_0$. If a is the phase distance to the boundary, choose

$$S_0 < \frac{a}{4[C_1 C_G/(1 - r) + C_2]}.$$

Then further reduce $|\sigma|$ so that the summed initial-layer contribution $C_1|\sigma|d_0/(1 - r)$ is below one quarter of the phase margin. Equations (F.12)–(F.13) then close an induction showing that every sample path of the original algorithms stays in the tube for $N\sigma^2 \le S_0$. No stopped process or tail estimate is used.

Taylor expansion of the exact phase at $F_A(X_t)$ gives

$$Y_{t+1} - Y_t = \sigma^2 f_A(Y_t) + \Delta M_{t+1} + R_{t+1}, \tag{F.14}$$

where $\Delta M_{t+1} = \Delta Y_{t+1} - \mathbb{E}[\Delta Y_{t+1} \mid \mathcal{F}_t]$ and $R_{t+1} = \mathbb{E}[\Delta Y_{t+1} \mid \mathcal{F}_t] - \sigma^2 f_A(Y_t)$. Thus $\mathbb{E}[\Delta M_{t+1} \mid \mathcal{F}_t] = 0$ and, with uniform tube constants,

$$\|\mathbb{E}[Y_{t+1} - Y_t \mid \mathcal{F}_t] - \sigma^2 f_A(Y_t)\| \le C(\sigma^2 d_t + |\sigma|^3),$$
$$\mathbb{E}[\|\Delta M_{t+1}\|^2 \mid \mathcal{F}_t] \le C(\sigma^2 d_t^2 + \sigma^4). \tag{F.15}$$

Indeed, the first-order Taylor term is conditionally centered at every state. Its coefficient $D\Pi_A(F_A(X_t))G_A(X_t, \omega)$ vanishes on the fixed manifold and is therefore $O(d_t)$. The conditional mean of the second-order term differs from its value at $\iota_A(Y_t)$ by $O(d_t)$, and the third-order Taylor remainder is $O(|\sigma|^3)$ because the noise and $D^3\Pi_A$ are bounded. At $\iota_A(y)$, Lemma F.1 and (F.7) give

$$f_A(y) = -b_A(h(y)) \operatorname{grad}_{\mathcal{M}} h(y).$$

For $N\sigma^2 \le S_0$, (F.12) implies

$$\sum_{t<N} (\sigma^2 d_t + |\sigma|^3) \le C_{S_0}|\sigma|,$$
$$\sum_{t<N} (\sigma^2 d_t^2 + \sigma^4) \le C_{S_0}\sigma^2.$$

Doob's L^2 inequality therefore bounds the expected maximal martingale by $C_{S_0}|\sigma|$. The vector fields f_A are Lipschitz on $\mathcal{K}$. Comparing (F.14) with the Euler discretization of (F.9), whose accumulated local error is $O(N\sigma^4)$, and applying discrete Gronwall yields

$$\mathbb{E} \max_{t \le S_0/\sigma^2} \|Y_{A,t} - y_A(t\sigma^2)\| \le C_{S_0}|\sigma|. \tag{F.16}$$

Define $X_{C,t}^{\mathrm{read}} = \theta_t^C$ and $X_{D,t}^{\mathrm{read}} = \bar{\theta}_t^D$ as the parameters at which sharpness is evaluated. On the common tube, full-Hessian sharpness at these parameters is Lipschitz, and $L_\sigma = L_0 + \sigma^2/2$ has the same Hessian as L_0. Hence

$$\left|\mathbb{E}\lambda_{\max}(\nabla^2 L_\sigma(X_{A,t}^{\mathrm{read}})) - h(y_A(t\sigma^2))\right| \le C\big(|\sigma| + r^t d_0\big). \tag{F.17}$$

For $t = \lfloor s/\sigma^2 \rfloor$ and $s \ge s_- > 0$, the second term converges to zero uniformly; it is absent when $d_0 = 0$. This proves Equation (F.10) and its centralized counterpart.

It remains to justify strictness of the limiting gap. Both fields in (F.9) are positive scalar multiples of $-\operatorname{grad}_{\mathcal{M}} h$. Starting from a nonstationary point, uniqueness shows that neither flow reaches a stationary point in finite time. The decentralized scalar multiple is strictly larger because $\mathcal{W} \ne J$, so it is strictly farther along the same integral curve. Since h strictly decreases on that curve, $h(y_D(s)) < h(y_C(s))$ for every $s > 0$. Its continuous gap has a positive minimum on $[s_-, S]$; (F.17) then proves (F.11). □

Corollary F.3 (Sharpness ordering at intermediate horizons)**.** *Under the conditions of Theorem F.2, fix an admitted initialization with* $\operatorname{grad}_{\mathcal{M}} h(y^\dagger) \ne 0$ *and set*

$$G(s) = h(y_C(s)) - h(y_D(s)), \qquad \gamma_0 = G'(0) = \big(b_D(h(y^\dagger)) - b_C(h(y^\dagger))\big)\|\operatorname{grad}_{\mathcal{M}} h(y^\dagger)\|^2 > 0.$$

Let $r \in (0,1)$ *and* d_0 *be the contraction factor and initial stable-coordinate distance in Equation (F.12). Summing the two bounds in Equation (F.17) gives a constant* $C_\Delta > 0$ *such that*

$$|\Delta_N(\theta_0, \sigma) + G(N\sigma^2)| \le C_\Delta(\sigma + r^N d_0), \qquad N\sigma^2 \le S_0, \tag{F.18}$$

for sufficiently small $\sigma > 0$*. There is* $s_1 \in (0, S_0]$ *such that every integer* $N \ge 2$ *satisfying*

$$r^N d_0 \le \sigma, \qquad N \ge \frac{8C_\Delta}{\gamma_0 \sigma}, \qquad N\sigma^2 \le s_1$$

obeys

$$\Delta_N(\theta_0, \sigma) \le -\frac{\gamma_0}{4} N\sigma^2 < 0. \tag{F.19}$$

The constants apply to this fixed initialization; they are not uniform over initializations whose limiting fits have tangential sharpness gradients approaching zero.

Proof. The flow equation gives the displayed formula for $G'(0)$. Since $G(0) = 0$ and $\gamma_0 > 0$, choose s_1 such that $G(s) \geq \gamma_0 s/2$ on $[0, s_1]$. The first two conditions on N and Equation (F.18) then imply

$$\Delta_N \leq -\frac{\gamma_0}{2} N\sigma^2 + 2C_\Delta \sigma \leq -\frac{\gamma_0}{4} N\sigma^2.$$

□

Remark 2 (Relation to the finite-step comparison)**.** The bound in Equation (F.18) has error $C_\Delta(\sigma + r^N d_0)$. Once the initial transient satisfies $r^N d_0 \leq \sigma$, Corollary F.3 certifies growing horizons such as $N = \lceil \sigma^{-3/2} \rceil$, because the accumulated gap exceeds the $O(\sigma)$ error. This conclusion follows from the slow estimate: the horizon-dependent thresholds in Theorem 1 do not, by themselves, guarantee $\sigma \leq \sigma_{N(\sigma)}$ along that sequence.

Even for a fitted initialization, the $O(\sigma)$ tracking bound does not resolve the fixed-step $O(\sigma^2)$ gap. At fixed N, the error bound for a nonfitted initialization additionally contains $r^N d_0$, which need not vanish as $\sigma \to 0$. The finite-step theorem also covers tangentially stationary fits: the balanced two-factor example has a motionless limiting flow but a strictly negative fixed-step gap (Subsection D.10).

Relation to noise-induced curvature dynamics. Label-noise-induced motion along a manifold of minimizers has been studied by Li et al. (2022) and Damian et al. (2021). Li et al. derive a tangential drift governed by the Hessian trace in a vanishing-learning-rate limit, while Damian et al. identify a learning-rate-dependent curvature regularizer. On our fitted manifold, the loss Hessian has rank one, so its trace and largest eigenvalue both equal h. Damian et al.'s regularizer therefore reduces to

$$R_\eta(h) = -\frac{1}{2\eta}\log(1 - \eta h/2), \qquad b_C(h) = \frac{\eta^2}{mK} R'_\eta(h).$$

This identity connects the centralized coefficient in (F.8) to their curvature regularizer. Our analysis additionally determines the graph-dependent contribution $b_D(h) - b_C(h)$ and transfers the slow-flow comparison to expected full-Hessian sharpness at the actual iterates.

F.3 Same-orbit and integrated-gap consequences

The mobility ratio at a common fitted position is

$$\mathcal{R}_\mathcal{W}(h) = \frac{b_D(h)}{b_C(h)} = 1 + (1 - c^2)\sum_{j=2}^{m} \frac{q_j}{1 - q_j c^2}, \qquad c = 1 - \eta h. \tag{F.20}$$

For $\mathcal{W} \neq J$,

$$1 < \mathcal{R}_\mathcal{W}(h) < m. \tag{F.21}$$

Every summand is strictly increasing in q_j. Writing $x = c^2$, its derivative with respect to x is $-q_j(1 - q_j)/(1 - q_j x)^2$. Thus the relative mobility is maximal at the same position when $\eta h = 1$. This is a relative-speed fact, not a claim about the learning rate that maximizes an equal-time curvature gap.

Corollary F.4 (Integrated curvature gap)**.** *Let $y_0 \in \mathcal{M}$ be nonstationary with $0 < \eta h(y_0) < 2$. The two ODEs in Equation (F.9), continued while $0 < \eta h < 2$, traverse the same geometric curve. If*

$$G(s) = h(y_C(s)) - h(y_D(s)),$$

then $G(s) > 0$ for every finite $s > 0$. For a deep linear network this continuation exists for every finite slow time, and $G(s) \to 0$ as $s \to \infty$. Consequently G attains a positive maximum at an interior time; it need not grow monotonically.

More quantitatively, if $R_- \leq \mathcal{R}_\mathcal{W}(h) \leq R_+$ on the traversed portion of the curve, then

$$h_C(s) - h_C(R_- s) \leq G(s) \leq h_C(s) - h_C(R_+ s). \tag{F.22}$$

Proof. Let $z'(\tau) = -\operatorname{grad}_\mathcal{M} h(z(\tau))$ and $z(0) = y_0$. There are increasing clocks τ_A such that $y_A(s) = z(\tau_A(s))$ and $\tau'_A = b_A(h(z(\tau_A)))$. Scalar comparison gives $\tau_D(s) > \tau_C(s)$ before a stationary point, proving $G(s) > 0$.

For a depth-D matrix network, let $b_i = W_i \cdots W_1$ and $a_i = W_D \cdots W_{i+1}$. The layer-i gradient has norm $\|a_i\|\|b_{i-1}\|$. Multiplying $|p| \le \|a_i\|\|b_i\|$ over the $D-1$ internal cuts and applying AM–GM gives

$$h = \sum_{i=1}^{D} \|\nabla_{W_i} g\|_F^2 \ge D|p|^{2-2/D}. \tag{F.23}$$

Along z, $dh/d\tau = -\|\operatorname{grad}_{\mathcal{M}} h\|^2$. On every finite τ interval, Cauchy–Schwarz bounds the path length by $\sqrt{\tau[h(0) - \inf h]}$, excluding finite-time escape. Thus the base curve exists globally. Since h decreases from $h_0 < 2/\eta$ and Equation (F.21) holds,

$$\frac{\eta^2}{4mK} \le b_C(h) < b_D(h) < m b_C(h) \le \frac{\eta^2}{2K(2-\eta h_0)}.$$

The upper bound prevents either clock from reaching infinite base time at a finite s, while the lower bound makes both clocks tend to infinity as $s \to \infty$. The decreasing bounded function $h(z(\tau))$ has one limit, so $G(s) \to 0$.

Finally define $u(s)$ by $y_D(s) = y_C(u(s))$. Differentiating the two clocks gives $u'(s) = \mathcal{R}_{\mathcal{W}}(h_D(s))$ and hence $R_- s \le u(s) \le R_+ s$. Since h_C decreases, this is exactly (F.22). □

F.4 Exact two-factor clock

For $g(u,v) = uv$, $p = 1$, and the positive fitted branch, set

$$B = u^2 - v^2, \qquad H(B) = \sqrt{B^2 + 4} = h(u,v).$$

The zero-noise transient from a common nonfitted initialization generally changes B. For residual $e_t = u_t v_t - 1$,

$$B^\dagger = B_0 \prod_{t \ge 0} (1 - \eta^2 e_t^2). \tag{F.24}$$

Thus $B^\dagger$, rather than the gradient-flow projection B_0, is the common initial value of the slow equations.

Corollary F.5 (Two-factor clocks and hitting times)**.** *Starting from $B^\dagger \ne 0$, the slow equations are*

$$\dot{B}_A = -4 b_A(H(B_A)) B_A, \tag{F.25}$$

or, equivalently,

$$\begin{aligned} \dot{h}_C &= -V_C(h_C), & V_C(h) &= \frac{2\eta^2}{mK} \frac{h^2 - 4}{h(2 - \eta h)}, \\ \dot{h}_D &= -V_D(h_D), & V_D(h) &= \mathcal{R}_{\mathcal{W}}(h) V_C(h). \end{aligned} \tag{F.26}$$

For $2 < h \le h_0 := H(B^\dagger)$, define

$$\begin{aligned} T_C(h) &= \frac{mK}{2\eta^2} \{J_\eta(h_0) - J_\eta(h)\}, \\ J_\eta(h) &= \log(h^2 - 4) - \eta h - \eta \log \frac{h-2}{h+2}, \\ T_D(h) &= \frac{mK}{2\eta^2} \int_h^{h_0} \frac{x(2 - \eta x)}{(x^2 - 4)\mathcal{R}_{\mathcal{W}}(x)}\, dx. \end{aligned} \tag{F.27}$$

Then $h_A(s) = T_A^{-1}(s)$. In particular, the slow hitting times of any target $2 < h_{\mathrm{tar}} < h_0$ obey

$$\frac{T_C(h_{\mathrm{tar}})}{R_+} \le T_D(h_{\mathrm{tar}}) \le \frac{T_C(h_{\mathrm{tar}})}{R_-} < T_C(h_{\mathrm{tar}}). \tag{F.28}$$

Here $1 < R_- \le \mathcal{R}_{\mathcal{W}}(h) \le R_+$ on $[h_{\mathrm{tar}}, h_0]$; such bounds exist by continuity and $\mathcal{W} \ne J$. Every interior maximum of $h_C(s) - h_D(s)$ satisfies

$$V_C(h_C(s)) = V_D(h_D(s)). \tag{F.29}$$

Proof. The exact sample update with residual r gives $B^+ = B(1 - \eta^2 r^2)$, proving (F.24) at zero noise. On $uv = 1$, direct differentiation along the unit tangent gives $\dot{B}_A = -4b_A(H(B_A))B_A$. Since $dH/dB = B/H$ and $B^2 = h^2 - 4$, this yields (F.26). The derivative of J_η is

$$J'_\eta(h) = \frac{h(2 - \eta h)}{h^2 - 4},$$

which proves (F.27). The hitting-time bounds follow by bounding $\mathcal{R}_\mathcal{W}$ inside the integral. Finally, $d(h_C - h_D)/ds = -V_C(h_C) + V_D(h_D)$, giving (F.29). □

The slow law describes accumulated tangential motion. At a fitted point with $\operatorname{grad}_\mathcal{M} h = 0$, it is stationary and supplies no strict ordering. The normal raw-Hessian response proved in Theorems 1 and D.1 can remain nonzero there. This separation is why the finite-horizon theorem is retained rather than presented as a corollary of the slow dynamics.

G Perturbation Sensitivity: Proof and Directional Boundaries

G.1 Proof of sensitivity reduction along the initial prediction gradient

Geometry and history condition. At a fitted reference, write $q_* = \nabla g(\theta_*)/\sqrt{h_*}$ and $Q_* = D^2 g(\theta_*)$. The injected direction q_0 is evaluated at the initialization θ_0, whereas q_* is evaluated at the fixed fitted reference θ_*; they coincide when $\theta_0 = \theta_*$. By construction, q_0 is normal to the initial prediction level set $\{\theta : g(\theta) = g(\theta_0)\}$. At an exact fitted initialization, $\nabla^2 L_\sigma(\theta_*) = h_* q_0 q_0^\top$, so q_0 spans the Hessian's nonzero eigenspace. The quantities

$$\chi = q_*^\top Q_* q_*, \qquad b_\perp = Q_* q_* - \chi q_*, \qquad \zeta_* = \sqrt{h_*}\, D^3 g(\theta_*)[q_*, q_*, q_*]$$

measure prediction curvature along the normal, its coupling to the perpendicular space, and the third directional derivative. For $1 < \alpha := \eta h_* < 2$, let $r = \alpha - 1$, and write the nonconsensus eigenvalues of $\mathcal{W}$ as $\lambda_j(\mathcal{W})$, $j = 2, \dots, m$. Set

$$s_j = \lambda_j(\mathcal{W})^2, \quad a_j = s_j r^2, \quad d_{j,1} = \frac{\eta^2 h_* s_j}{mK}, \quad d_{j,t} = d_{j,1}\frac{1 - a_j^t}{1 - a_j}, \quad d_t = \sum_{j=2}^{m} d_{j,t}.$$

For an injection time $n \geq 2$, let n_o be the smallest odd integer no smaller than n, and define the exact-history quantities

$$\delta_n = \frac{\sum_{j=2}^{m} d_{j,1}(a_j^{n_\mathrm{o}} + r^{n_\mathrm{o}})/(r + a_j)}{d_{n_\mathrm{o}}}, \qquad J_n = \frac{\sum_{t=0}^{n-1} d_t}{d_n}.$$

Here $d_n > 0$ because $\mathcal{W} \neq J$. The fitted-geometry and history margin is

$$\widetilde{\mathfrak{m}}_n = 3\left(1 - \tfrac{3}{2}\delta_n\right)\chi^2 + 2\alpha J_n \|b_\perp\|^2 - 2\zeta_*. \tag{G.1}$$

Theorem G.1 (Sensitivity reduction along the initial prediction gradient in deep linear networks)**.** *Under the model, SGD comparison, and label and mixing conditions of Theorem E.1, fix a fitted reference θ_*, any integer injection time $n \geq 2$, and a finite window $T \geq 1$. Use the response quantities in Equation (5) and the geometry/history margin in Equation (G.1). Suppose $1 < \eta h_* < 2$ and $\widetilde{\mathfrak{m}}_n > 0$. There exist $\varepsilon_{\mathrm{resp}} > 0$ and $\sigma_{\mathrm{resp}} > 0$ such that every common initialization $\|\theta_0 - \theta_*\| \leq \varepsilon_{\mathrm{resp}}$ and every $0 < \sigma \leq \sigma_{\mathrm{resp}}$ satisfy*

$$\mathcal{G}_{n,k}^{q_0}(\sigma) < 0 \quad (1 \leq k \leq T), \qquad \mathcal{E}_{n,T}^{q_0}(\sigma) < 0.$$

The neighborhood is in the full space of layer matrices; its radius and noise threshold may depend on n, T and the reference geometry.

Response expansion. At the fit, $H_* := \nabla^2 L_\sigma(\theta_*) = h_* q_* q_*^\top$, so the prediction normal spans the top-Hessian eigenspace. Near a fit, the residual Hessian term makes this identification only approximate; the initial direction is fixed rather than recomputed during training. All derivatives in the fitted calculation are evaluated at θ_*. We use Taylor coefficients without factorials: a state is $\theta_* + \sigma z + \sigma^2 B + \cdots$. Put $v = \nabla g$, $h = h_*$, $q = q_*$, $Q = Q_*$, $c = 1 - \alpha = -r$, and $A = I - \eta v v^\top$. The half-squared population loss has

$$H = vv^\top, \quad a = D^3 L[q, q] = \sqrt{h}(3\chi q + 2b_\perp), \quad D^4 L[q, q, q, q] = 3\chi^2 + 4\zeta_*.$$

The effective minibatch label innovations are independent across workers and times, centered, symmetric, bounded, and have variance $1/K$: multiplication by the independent Rademacher input makes them symmetric under the stated label law.

State history and probe recursions. At zero noise every state stays at the fit. Write the first-order parameter response of worker i as $q x_{i,t}$. Choose an orthonormal graph eigenbasis with consensus eigenvector $\mathbf{1}/\sqrt{m}$. For each nonconsensus mode $j > 1$, let $d_{j,t}$ be the scalar variance of its first-order response coefficient, divided by m, and set $d_t = \sum_{j>1} d_{j,t}$. We retain $s_j = \lambda_j(\mathcal{W})^2$ as defined above. If β_t is the expected second-order mean-state difference, then

$$d_{j,t+1} = s_j\left(c^2 d_{j,t} + \frac{\eta^2 h}{mK}\right), \qquad d_{j,0} = 0, \tag{G.2}$$

$$\beta_{t+1} = A\beta_t - \frac{\eta}{2} d_t a, \qquad \beta_0 = 0. \tag{G.3}$$

These follow by expanding the sampled gradient; the fresh label is independent of the pre-update state. In particular $d_1 > 0$ because $\mathcal{W} \neq J$. No history is discarded at injection.

After injection let s count subsequent steps. The zero-noise probe is $c^s q$. Let $X_{j,s}$ be the graph-mode cross-covariance between the first-order scalar state response and first-order vector probe response, divided by m, and let u_s be the projection onto q of the expected second-order mean-probe difference. Direct expansion of the sampled state and tangent maps gives

$$X_{j,s+1} = s_j \left(cAX_{j,s} - \eta c^{s+1} d_{j,n+s} a + \frac{\eta^2 \sqrt{h}}{mK} c^s Qq \right), \tag{G.4}$$

$$u_{s+1} = cu_s - \eta \left(c^s Z_{n+s} + a^\top \sum_j X_{j,s} \right), \tag{G.5}$$

$$Z_t = a^\top \beta_t + \left(\tfrac{3}{2}\chi^2 + 2\zeta_* \right) d_t, \qquad X_{j,0} = 0, \quad u_0 = 0. \tag{G.6}$$

For clarity, the local sampled tangent map is $Y^+ = (I - \eta H(\theta) + \eta\sigma\omega D^2 g(\theta))Y$. The first state derivative of $H(\theta)Y$ at the fit is $DH[z]Y$, and $DH[q]q = a$. Projecting the second state derivative onto the zero-order probe contributes $D^4 L[q,q,q,q]/2$, yielding Equation (G.6). The last term of Equation (G.4) is the covariance of the same fresh label in the state and probe innovations; it does not vanish by centering. Fresh labels multiplying only pre-update quantities have zero expectation. Double stochasticity makes both first-order mean variations equal to their centralized counterparts pathwise. Their squared terms cancel in the response difference. Hence the fitted terminal coefficient is

$$\mathcal{G}^q_{n,k}(\sigma) = \sigma^2 \mathcal{B}_{n,k} + O(\sigma^4), \qquad \mathcal{B}_{n,k} = 2c^k u_k. \tag{G.7}$$

A bound on the complete state history. Let $S_t = \sum_{\ell=0}^{t-1} c^{t-1-\ell} d_\ell$. For $a_j = s_j r^2$, solving Equation (G.2) and summing the geometric series gives the exact identity

$$\alpha S_t - d_t = \sum_{j>1} d_{j,1} \frac{a_j^t - c^t}{c - a_j}. \tag{G.8}$$

For even t, every summand on the right is nonnegative. For odd t, define

$$\delta_t^{\mathrm{odd}} = \frac{\sum_{j>1} d_{j,1}(a_j^t + r^t)/(r + a_j)}{d_t};$$

then $\alpha S_t / d_t = 1 - \delta_t^{\mathrm{odd}}$. Its numerator decreases with odd t, whereas d_t increases. Hence the largest odd-time deficit over $t \geq n$ occurs at the first such time n_{o}, and

$$\alpha S_t \geq (1 - \delta_n) d_t \qquad (t \geq n). \tag{G.9}$$

This proves the first exact-history bound without a parity restriction on n.

For the second bound, put $D_t = \sum_{\ell<t} d_\ell$. The sequence d_t is increasing and discretely concave, with $d_0 = 0$. If $\Delta_t = d_{t+1} - d_t$, concavity gives $\Delta_t \leq d_t/t$, while monotonicity gives $D_t \leq (t-1)d_t$. Therefore

$$\frac{D_{t+1}}{d_{t+1}} - \frac{D_t}{d_t} = \frac{d_t^2 - D_t \Delta_t}{d_t d_{t+1}} \geq 0.$$

Thus

$$\sum_{\ell<t} d_\ell \geq J_n d_t \qquad (t \geq n). \tag{G.10}$$

Zero graph modes contribute nothing to either bound.

Solving Equation (G.3) along q and its perpendicular space yields

$$a^\top \beta_t = -\frac{9\alpha}{2} \chi^2 S_t - 2\alpha \|b_\perp\|^2 \sum_{\ell<t} d_\ell.$$

Substituting Eqs. (G.9) and (G.10) into Equation (G.6) now gives

$$Z_t \leq -\widetilde{\mathfrak{m}}_n d_t \qquad (t \geq n). \tag{G.11}$$

This step keeps the true ζ_*, instead of replacing it by a worst-case derivative bound.

The state–probe feedback has the favorable sign. For $s \geq 1$, write

$$X_{j,s} = \sqrt{h}\, c^{s-1} \left(\chi P_{j,s} q + R_{j,s} b_\perp\right).$$

Equation (G.4) is equivalent to

$$P_{j,s+1} = s_j \left(c P_{j,s} + \frac{\eta^2}{mK} - 3\eta c\, d_{j,n+s}\right),$$
$$R_{j,s+1} = s_j \left(R_{j,s} + \frac{\eta^2}{mK} - 2\eta c\, d_{j,n+s}\right), \qquad P_{j,0} = R_{j,0} = 0.$$

Because $c = -r$, both bracketed forcings are positive and nondecreasing in s. Thus $R_{j,s} \geq 0$ by induction. For $P_{j,s}$, set $\tau_j = s_j r \in [0, 1)$ and $p_{j,s} = \eta^2/(mK) + 3\eta r d_{j,n+s}$. Iterating its recursion gives

$$P_{j,s} = s_j \sum_{\ell=0}^{s-1} (-\tau_j)^{s-1-\ell} p_{j,\ell}.$$

If $s = 2L$, group this sum as $s_j \sum_{v=0}^{L-1} \tau_j^{2v}(p_{j,s-1-2v} - \tau_j p_{j,s-2-2v})$. If $s = 2L+1$, the same pairing for $v = 0, \ldots, L-1$ leaves the nonnegative term $s_j \tau_j^{2L} p_{j,0}$. Every pair is nonnegative because $p_{j,s}$ is nondecreasing and $\tau_j < 1$. Hence $P_{j,s} \geq 0$. With $F_s = \sum_j (3\chi^2 P_{j,s} + 2\|b_\perp\|^2 R_{j,s}) \geq 0$ and $F_0 = 0$, solving Equation (G.5) gives

$$\mathcal{B}_{n,k} = -2\eta \left[c^{2k-1} \sum_{s=0}^{k-1} Z_{n+s} + hc^{2k-2} \sum_{s=0}^{k-1} F_s\right] \leq -2\eta k d_1 \widetilde{\mathfrak{m}}_n r^{2k-1} < 0.$$

Here $d_{n+s} \geq d_1$ because the state-variance history is increasing. This proves a strict fitted coefficient margin at every finite terminal step, at any depth satisfying Equation (G.1).

Transfer to the full near-fit neighborhood. The finite-step state and tangent maps are polynomials in initialization and noise before normalizing the initial probe. On a small closed ball around θ_*, ∇g stays nonzero, so $q_0(\theta_0)$ is smooth. The mean response difference is even in σ and vanishes at zero. Uniformly on this ball and on $1 \leq k \leq T$, write $\mathcal{G}_{n,k}^{q_0}(\theta_0, \sigma) = \sigma^2 B_k(\theta_0) + \sigma^4 R_k(\theta_0, \sigma)$, where B_k is continuous and the remainders are bounded for bounded σ. Set $\mu = \min_{1 \leq k \leq T} 2\eta k d_1 \widetilde{\mathfrak{m}}_n r^{2k-1} > 0$. Shrink the initialization ball until every $B_k \leq -\mu/2$. If $|R_k| \leq M$ on its product with $|\sigma| \leq s_0$, choose $\sigma_{\mathrm{resp}} = \min\{s_0, \sqrt{\mu/(4M)}\}$, interpreting the second term as infinite when $M = 0$. Then every terminal gap is at most $-\mu\sigma^2/4$ and every cumulative gap is negative. The initialization radius and noise threshold may depend on the fixed injection time and response window. This completes the proof of Theorem G.1.

G.2 Dependence on the perturbation direction

We examine how the cumulative response gap depends on the injection direction at a fixed fitted reference. For a fixed unit direction u, define $\mathcal{E}_{n,T}^u$ by replacing q_0 with u in Equation (5). Thus the comparator and parameter-response measure are unchanged; only the injected direction differs.

In the depth-four reference of Subsection H.2, set $J_2 = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and define a direction by $\dot{W}_1 = \dot{W}_2 = 0$, $\dot{W}_3 = J_2 W_3$, $\dot{W}_4 = -W_4 J_2$. The product rule gives $Dg[\dot{\theta}] = 0$, while $\|\dot{\theta}\|^2 = 101/50$. The corresponding linear perturbation leaves the initial prediction unchanged to first order, although its sampled-gradient derivative is nonzero. For $\alpha = 3/2, n = 4, T = 16, \rho = 1/2, m = 2, K = 1$, let $u = \dot{\theta}/\sqrt{101/50}$. At $\theta_0 = \theta_*$, the fitted cumulative coefficients for u and the prediction normal q_* have opposite signs:[1]

$$\lim_{\sigma \to 0} \frac{\mathcal{E}_{4,16}^u(\theta_*, \sigma)}{\sigma^2} \approx 1.67787 > 0, \qquad \lim_{\sigma \to 0} \frac{\mathcal{E}_{4,16}^{q_*}(\theta_*, \sigma)}{\sigma^2} \approx -1.77155 < 0.$$

We compute these coefficients using the state and perturbation expansions from Subsection G.1, now for a general fixed injection direction. Double stochasticity again makes the first-order averaged

[1] The exact coefficient for u is $\frac{9166845738195957233136035604187487109375}{5463376576101456582702996962396805791744}$.

perturbation derivative equal its centralized counterpart, so their squared terms cancel in the response difference. Put $v = \nabla g(\theta_*)$, $h = \|v\|^2$, $Q = D^2 g(\theta_*)$, $H = \nabla^2 L_0$, $T_v = DH[v]$, $U_v = D^2 H[v, v]$, $A = I - \eta v v^\top$, and $c = 1 - \eta h$; all derivatives below are evaluated at θ_*. The scalar $v_d = d_t/h$ is the worker-averaged second-moment coefficient of the centered first-order parameter responses in the v normalization, and β is the expected second-order mean-state difference. After injection, Y is the zero-noise perturbation derivative, X is the aggregated graph-mode cross-moment between the centered first-order state response and the first-order perturbation response in the same normalization, and B is the difference of expected second-order mean-perturbation coefficients.

Starting from $v_d = 0$ and $\beta = 0$, run the state recursions for $n = 4$ steps. At injection, set $Y = \dot\theta$ and $X = B = 0$, without resetting v_d or β, then continue for $T = 16$ steps. The superscript $+$ denotes the next step, with every right-hand side evaluated from the current values. Before and after injection, update

$$\beta^+ = A\beta - \tfrac{\eta}{2} v_d T_v v, \qquad v_d^+ = \tfrac{1}{4}(c^2 v_d + \eta^2/2).$$

After injection, also update from the old values

$$B^+ = AB - \eta\big[(DH[\beta] + v_d U_v/2)Y + T_v X\big],$$
$$X^+ = \tfrac{1}{4}(cAX - \eta c v_d T_v Y + \eta^2 QY/2), \qquad Y^+ = AY.$$

Sum $2Y^\top B/\|\dot\theta\|^2$ over the 16 steps after injection, evaluating the term after each update. Here $T_v = hQ + (Qv)v^\top + v(Qv)^\top$, and with $R_v = D^3 g[v, \cdot, \cdot]$, $U_v = 2(Qv)(Qv)^\top + (R_v v)v^\top + v(R_v v)^\top + (v^\top Qv)Q + 2hR_v$. For the normal coefficient, use $Y = v$ at injection and divide the sum of $2Y^\top B$ by h instead of $\|\dot\theta\|^2$. We evaluate the recurrence using exact rational arithmetic.

For either fixed direction $w \in \{q_*, u\}$, finite-step smoothness, symmetric noise, and equality of the zero-noise responses give $\mathcal{E}^w_{4,16}(\theta_*, \sigma) = \sigma^2 C_w + O(\sigma^4)$, where C_w is the corresponding coefficient above. Their opposite strict signs therefore imply opposite expected response gaps for sufficiently small nonzero noise. Thus, changing only the injection direction can reverse the comparison. This argument does not determine the tangent gap at the finite experimental noise amplitude.

G.3 Balanced scalar-layer all-direction complement

This complementary result controls every fixed unit direction in a balanced scalar-layer network. Use the response definitions in Equation (5), replacing q_0 by the fixed direction u.

Theorem G.2 (Multistep sensitivity reduction in balanced deep linear networks)**.** *Under the same SGD comparison, mixing, and sampling conditions, let $f_\theta(x) = x\prod_{\ell=1}^D w_\ell$, with $D \geq 2$ independently trained layer parameters. Use half-squared loss on $x \in \{-1, 1\}$ and $y = x + \sigma\zeta$, where ζ is independent of x, bounded, centered, and has unit variance. Initialize every layer of every model at the fitted value $w_{\ell,*} = 1$. Fix an even injection time $n \geq 2$, a finite response window $T \geq 1$, and a learning rate satisfying*

$$1 < \eta h_* < 1 + \left(\frac{2}{3(D-1)}\right)^{1/n}, \qquad h_* = D. \tag{G.12}$$

There exists $\sigma_0 > 0$, depending on $D, \eta, \mathcal{W}, m, K, n, T$ and the bounded label-noise law, but not on u or k, such that for every $0 < \sigma \leq \sigma_0$, every fixed unit direction u independent of the samples, and every $1 \leq k \leq T$,

$$\mathcal{G}^u_{n,k}(\sigma) < 0, \qquad \mathcal{E}^u_{n,T}(\sigma) < 0.$$

The comparison follows the actual sampled updates after one common injection. Quantitative margins and an explicit finite-horizon noise threshold are given in Subsection G.3.

For the quantitative bound, write $\tau = \eta D$, $r = \tau - 1$, $\kappa_n = 3(D-1)r^n/2$ and $V_1 = \eta^2 (mK)^{-1} \sum_{j=2}^m \lambda_j(\mathcal{W})^2$. Define

$$\gamma_{n,k} = 2k\tau V_1 \min\big\{r^{2k-1}(D-1)(D+1-3\kappa_n),\, 1-\kappa_n\big\} > 0. \tag{G.13}$$

We prove a uniform remainder bound and choose the noise threshold so that $\mathcal{G}^u_{n,k} \leq -\sigma^2\gamma_{n,k}/2$ and $\mathcal{E}^u_{n,T} \leq -\sigma^2 \sum_{k=1}^T \gamma_{n,k}/2$.

Throughout this subsection, the depth is D, the minibatch size is K, and the response window is T. Write

$$\omega_{i,t} = \frac{1}{K}\sum_{b=1}^{K} x_{i,t,b}\zeta_{i,t,b}.$$

Then $\omega_{i,t}$ is centered, has variance $1/K$, is bounded, and is independent across workers and times. The centralized update uses the average of these same variables at its own state.

Actual states and sampled Jacobians. Starting from $w_{\ell,*} = 1$, each unperturbed worker state has the form $w_{i,t}\mathbf{1}_D$. This follows from equality of the coordinate gradients and is preserved by parameter averaging; the parameters are not tied by the algorithm. The scalar state recursion is

$$w_{i,t+1} = \sum_j \mathcal{W}_{ij}\left[w_{j,t} - \eta\{w_{j,t}^{2D-1} - w_{j,t}^{D-1}\} + \eta\sigma\omega_{j,t}w_{j,t}^{D-1}\right]. \tag{G.14}$$

The centralized scalar recursion uses w_t^C and $m^{-1}\sum_i \omega_{i,t}$.

Let

$$e = \frac{\mathbf{1}_D}{\sqrt{D}}, \qquad u = u_+ e + u_\perp, \qquad e^\top u_\perp = 0.$$

At a diagonal state, differentiating the full D-parameter sample loss gives two invariant eigenspaces for its update Jacobian:

$$j_+(w,\sigma\omega) = 1 - \eta\left[(2D-1)w^{2D-2} - (D-1)(1+\sigma\omega)w^{D-2}\right], \tag{G.15}$$

$$j_-(w,\sigma\omega) = 1 + \eta\left[w^{2D-2} - (1+\sigma\omega)w^{D-2}\right]. \tag{G.16}$$

The first acts on e; the second acts on every direction perpendicular to e. Consequently, for the once-injected probe,

$$Y_{i,t}^D = u_+ J_{i,t}^+ e + J_{i,t}^- u_\perp, \qquad J_{i,n}^+ = J_{i,n}^- = 1,$$

where

$$J_{i,t+1}^a = \sum_j \mathcal{W}_{ij} j_a(w_{j,t},\sigma\omega_{j,t})J_{j,t}^a, \qquad a \in \{+,-\}.$$

The centralized multipliers obey the corresponding recursion at w_t^C. Thus both channels propagate through all future steps without resetting.

State coefficients. Use Taylor coefficients without factorials:

$$w_{i,t} = 1 + \sigma z_{i,t} + \sigma^2 b_{i,t} + O(|\sigma|^3).$$

Set $c = 1-\tau$, $\tau = \eta D$, and define

$$V_t = \frac{1}{m}\mathbb{E}\sum_i (z_{i,t} - \bar{z}_t)^2, \qquad b_t = \mathbb{E}(\bar{b}_t - b_t^C).$$

Thus V_t is the worker-averaged second moment of the centered first-order parameter responses, while b_t is the second-order Taylor coefficient of the expected mean-parameter difference. The first mean variation satisfies $\bar{z}_t = z_t^C$ pathwise. For an orthonormal graph eigenbasis q_j, $j > 1$, put

$$g_j = \lambda_j(\mathcal{W})^2, \qquad I_j = \frac{\eta^2 g_j}{mK}.$$

The normalized variance of the first-order parameter response in nonconsensus graph mode j is

$$V_{j,t} = I_j \sum_{\ell=0}^{t-1}(g_j c^2)^\ell, \qquad V_t = \sum_{j>1} V_{j,t}. \tag{G.17}$$

In particular, $V_t \geq V_1 > 0$ for $t \geq 1$.

For $F(w) = w^{2D-1} - w^{D-1}$,

$$F'(1) = D, \qquad F''(1) = 3D(D-1).$$

Fresh-sample centering removes the conditional mean of $\omega_{i,t}z_{i,t}$, yielding

$$b_{t+1} = cb_t - \frac{3\tau(D-1)}{2}V_t, \qquad b_0 = 0. \tag{G.18}$$

This is the same state displacement used in the sharpness calculation. Indeed, $d_t = DV_t$, and the raw sharpness gradient at the fit is $3(D-1)\mathbf{1}_D$.

Probe coefficients and the same-sample cross term. Expanding Equation (G.16) at $w = 1$ gives

$$j_a = A_a + \sigma(P_a z + Q_a \omega) + \sigma^2\{P_a b + R_a z^2 + (D-2)Q_a \omega z\} + O(|\sigma|^3),$$

where

a	A_a	P_a	Q_a	R_a
$+$	c	$-3\tau(D-1)$	$\eta(D-1)$	$-\tau(D-1)(7D-11)/2$
$-$	1	τ	$-\eta$	$\tau(3D-5)/2$

and the state dependence of the sampled gradient has been retained.

Write

$$J^a_{i,t} = A_a^{t-n} + \sigma L^a_{i,t} + \sigma^2 B^a_{i,t} + O(|\sigma|^3).$$

The first probe variations satisfy $\bar{L}^a_t = L^{C,a}_t$ pathwise. Define, for $j > 1$,

$$C^a_{j,t} = \frac{1}{m}\mathbb{E}[(q_j^\top z_t)(q_j^\top L^a_t)], \qquad M^a_t = \mathbb{E}(\bar{B}^a_t - B^{C,a}_t).$$

Initially $C^a_{j,n} = M^a_n = 0$. For $t \geq n$, direct multiplication of the first state and probe recursions gives

$$C^a_{j,t+1} = g_j \left[cA_a C^a_{j,t} + A_a^{t-n}\left(cP_a V_{j,t} + \frac{\eta Q_a}{mK} \right) \right]. \tag{G.19}$$

The term $\eta Q_a/(mK)$ is the covariance of the same fresh sample's state and probe injections. Centering does not remove it.

The second mean probe difference is

$$M^a_{t+1} = A_a M^a_t + P_a \sum_{j>1} C^a_{j,t} + A_a^{t-n} Z^a_t, \qquad Z^a_t = P_a b_t + R_a V_t. \tag{G.20}$$

To verify this equation, the second-order products involving $\omega_{i,t} L^a_{i,t}$ and $\omega_{i,t} z_{i,t}$ have zero conditional expectation because the latter variables are measurable before the current samples. The remaining state–probe product is $m^{-1}\sum_i \mathbb{E}[(z_{i,t} - \bar{z}_t)(L^a_{i,t} - \bar{L}^a_t)] = \sum_{j>1} C^a_{j,t}$. The state-square difference is V_t, and the mean second state difference is b_t.

When evaluating the squared response, the first probe means agree pathwise, so their squares cancel. Therefore

$$\mathcal{G}^u_{n,k}(\sigma) = \sigma^2 \left[u_+^2 \mathcal{B}^+_{n,k} + \|u_\perp\|^2 \mathcal{B}^-_{n,k} \right] + O_{n,k}(|\sigma|^3), \qquad \mathcal{B}^a_{n,k} = 2A_a^k M^a_{n+k}. \tag{G.21}$$

The factor two comes from squaring the response; $B^a_{i,t}$ is a Taylor coefficient, not a second derivative.

Control of the initial oscillatory transient. Let

$$B_D = \frac{3(D-1)}{2}, \qquad e_t = b_t + B_D V_t.$$

Equations (G.17) and (G.18) imply

$$e_{t+1} = ce_t + B_D(V_{t+1} - V_t), \qquad e_0 = 0,$$

and hence

$$e_t = B_D \sum_{j>1} I_j \sum_{\ell=0}^{t-1} c^{t-1-\ell}(g_j c^2)^\ell. \tag{G.22}$$

Put $c = -r$, $0 < r < 1$. For each active mode,

$$\sum_{\ell=0}^{t-1} c^{t-1-\ell}(g_j c^2)^\ell = (-r)^{t-1} \sum_{\ell=0}^{t-1} (-g_j r)^\ell.$$

Because $0 \leq g_j r < 1$, the latter geometric sum is positive. It follows that $e_t \leq 0$ for even t, while for odd t,

$$0 \leq e_t \leq B_D r^{t-1} V_1.$$

For even n and every $t \geq n$, this gives

$$e_t \leq \kappa_n V_t, \qquad \kappa_n = B_D r^n < 1.$$

Consequently,

$$Z^-_t = \tau(e_t - V_t) \leq -\tau(1-\kappa_n)V_t < 0, \tag{G.23}$$

$$Z^+_t = \tau(D-1)\{(D+1)V_t - 3e_t\} \tag{G.24}$$

$$\geq \tau(D-1)(D+1-3\kappa_n)V_t > 0. \tag{G.25}$$

The propagated state–probe cross-moments have a favorable sign. We use an elementary recursion fact. If $0 \le g < 1$, $0 < r < 1$, $f_s > 0$ is nondecreasing, and

$$x_{s+1} = g(-rx_s + f_s), \qquad x_0 = 0,$$

then $0 \le x_{s+1} \le gf_s$. Indeed, assuming $0 \le x_s \le gf_{s-1}$ gives

$$x_{s+1} \ge g(f_s - rgf_{s-1}) \ge 0, \qquad x_{s+1} \le gf_s.$$

For the minus channel, apply this fact to $-C^-_{j,n+s}$. Its forcing is $-(cP_-V_{j,n+s} + \eta Q_-/(mK)) > 0$ and is nondecreasing. Thus $C^-_{j,t} \le 0$.

For the plus channel define $H_{j,s} = C^+_{j,n+s}/c^{s-1}$ for $s \ge 1$, with $H_{j,0} = 0$. Equation (G.19) becomes

$$H_{j,s+1} = g_j \left[cH_{j,s} + cP_+V_{j,n+s} + \frac{\eta Q_+}{mK} \right].$$

Here $cP_+ > 0$ and $Q_+ > 0$, so the same recursion fact gives $H_{j,s} \ge 0$.

Solving Equation (G.20) over the actual window yields

$$\mathcal{B}^-_{n,k} = 2 \left[\sum_{t=n}^{n+k-1} Z^-_t + P_- \sum_{s=1}^{k-1} \sum_{j>1} C^-_{j,n+s} \right], \tag{G.26}$$

$$\mathcal{B}^+_{n,k} = 2c^{2k-2} \left[c \sum_{t=n}^{n+k-1} Z^+_t + P_+ \sum_{s=1}^{k-1} \sum_{j>1} H_{j,s} \right]. \tag{G.27}$$

Every displayed feedback contribution is nonpositive. Using $V_t \ge V_1$ gives

$$\mathcal{B}^-_{n,k} \le -2k\tau(1-\kappa_n)V_1, \tag{G.28}$$

$$\mathcal{B}^+_{n,k} \le -2k\tau r^{2k-1}(D-1)(D+1-3\kappa_n)V_1. \tag{G.29}$$

Since $u_+^2 + \|u_\perp\|^2 = 1$, this proves the coefficient margin in Equation (G.13). Summing over terminal times uses the same propagated probe throughout.

Actual finite-horizon remainder and a uniform noise threshold. We give explicit bounds to complete the passage from coefficients to the unmodified SGD processes. Let $N_* = n + T$ and choose a closed Euclidean ball $\mathcal{B}_R$ around $\mathbf{1}_D$ inside the regularity neighborhood. Let

$$L_j = \sup_{\mathcal{B}_R} \|D^j L_0\|, \quad j = 2, \dots, 5, \qquad M_j = \sup_{\mathcal{B}_R, Z} \|D^j \xi(\theta; Z)\|, \quad j = 0, \dots, 3,$$

Here $L_0(\theta) = \frac{1}{2}(g(\theta) - 1)^2$ and $\xi(\theta; Z) = -x\zeta\nabla g(\theta)$, so a sampled gradient is $\nabla L_0 + \sigma\xi$. All derivative norms are induced Euclidean multilinear norms. The latter supremum is an almost-sure deterministic bound; bounded labels and the polynomial network make all these constants finite on the ball. The minibatch fields satisfy the same bounds.

Put $b = 1 + \eta L_2$, $S_q(b) = \sum_{t=0}^{q-1} b^t$, and

$$\delta_* = \min \left\{ 1, \frac{R}{2\eta M_0 S_{N_*}(b)} \right\}.$$

The maximum distance of any worker or central state from the common fitted initialization obeys $E_{t+1} \le bE_t + \eta|\sigma|M_0$. Thus all actual states lie in $\mathcal{B}_{R/2}$ through time N_* when $|\sigma| \le \delta_*$. No projection or truncation is used.

For completeness, define

$$B_s = b + \eta\delta_* M_1, \quad C_s = \eta(L_3 + \delta_* M_2), \quad D_s = \eta(L_4 + \delta_* M_3).$$

Starting from $U_0 = V_0^{\mathrm{cap}} = Q_0^{\mathrm{cap}} = 0$, set

$$\begin{aligned} U_{t+1} &= B_s U_t + \eta M_0, \\ V_{t+1}^{\mathrm{cap}} &= B_s V_t^{\mathrm{cap}} + C_s U_t^2 + 2\eta M_1 U_t, \\ Q_{t+1}^{\mathrm{cap}} &= B_s Q_t^{\mathrm{cap}} + 3C_s U_t V_t^{\mathrm{cap}} + D_s U_t^3 + 3\eta M_1 V_t^{\mathrm{cap}} + 3\eta M_2 U_t^2. \end{aligned}$$

Let U, V, Q be the respective maxima through N_*. They bound the first three amplitude derivatives of each actual state.

Each actual update Jacobian, evaluated along its own state, has an expansion

$$\mathsf{J}_t(\sigma) = A + \sigma \mathsf{J}_t^{[1]} + \sigma^2 \mathsf{J}_t^{[2]} + \mathsf{R}_t(\sigma)$$

where $A = I - \eta \nabla^2 L_0(\mathbf{1}_D)$, with deterministic bounds

$$b_1 = \eta(L_3 U + M_1),$$
$$b_2 = \frac{\eta}{2}(L_3 V + L_4 U^2 + 2M_2 U),$$
$$b_3 = \frac{\eta}{6}(L_3 Q + 3L_4 UV + L_5 U^3) + \frac{\eta}{2}(M_2 V + M_3 U^2),$$

such that $\|\mathsf{J}_t^{[1]}\| \le b_1$, $\|\mathsf{J}_t^{[2]}\| \le b_2$, and $\|\mathsf{R}_t(\sigma)\| \le b_3|\sigma|^3$. For the sampled-field contribution, this uses only a first-order amplitude expansion of $D\xi(\theta_t^\sigma)$ with a second-order remainder, multiplied by σ. Thus $D^3\xi$, not $D^4\xi$, suffices.

Use the maximum worker Euclidean norm for stacked probes. Nonnegative row-stochastic mixing is nonexpansive in this norm, and $\|A\|_2 \le 1$. Define $\mathfrak{b} = 1 + b_1 + b_2 + b_3$ and initialize $p_0 = q_0 = e_0^{\mathrm{cap}} = 0$. For $k \ge 0$, set

$$p_{k+1} = p_k + b_1, \tag{G.30}$$
$$q_{k+1} = q_k + b_1 p_k + b_2, \tag{G.31}$$
$$e_{k+1}^{\mathrm{cap}} = \mathfrak{b}\, e_k^{\mathrm{cap}} + b_1 q_k + b_2 p_k + b_2 q_k + b_3(1 + p_k + q_k). \tag{G.32}$$

These bound the first Taylor coefficient, second Taylor coefficient, and the remainder divided by $|\sigma|^3$ of every propagated unit probe and its mean. The last recursion follows by multiplying $(A + \sigma \mathsf{J}^{[1]} + \sigma^2 \mathsf{J}^{[2]} + \mathsf{R})$ by the probe expansion and bounding the terms of total order at least three; $|\sigma| \le 1$ is used.

The difference in squared parameter responses therefore has remainder at most $\mathfrak{M}_k|\sigma|^3$, uniformly over unit input directions, where

$$\mathfrak{M}_k = 2\left[2p_k q_k + q_k^2 + 2(1 + p_k + q_k)e_k^{\mathrm{cap}} + (e_k^{\mathrm{cap}})^2\right]. \tag{G.33}$$

The factor two covers the decentralized and centralized squared responses. These deterministic bounds justify the required expectation limits.

The probe derivative itself is that of the original process: for fixed σ, a sufficiently small injection $|\epsilon| \le R/(2B_s^T)$ keeps the perturbed trajectories in $\mathcal{B}_R$, by a finite-step Lipschitz induction. The same bound applies to every unit direction.

Finally, take

$$\sigma_0 = \min\left\{\delta_*, \min_{1 \le k \le T} \frac{\gamma_{n,k}}{2\mathfrak{M}_k}\right\}, \tag{G.34}$$

where a ratio with zero denominator is interpreted as infinity. Then each terminal difference is at most $-\sigma^2 \gamma_{n,k}/2$. Summing these inequalities proves the cumulative comparison and completes the proof.

G.3.1 Extensions and boundaries

Other injection times and an explicit margin. For an arbitrary integer $n \ge 1$, let n_{odd} be the smallest odd integer not less than n. The same proof applies under

$$\frac{3(D-1)}{2}(\eta h_* - 1)^{n_{\mathrm{odd}}-1} < 1, \qquad 1 < \eta h_* < 2.$$

The even-n statement in Theorem G.2 is its direct learning-rate form, not a stronger necessity claim.

For a prescribed margin $\vartheta \in (0,1)$, it is sufficient to impose

$$1 < \eta h_* \le 1 + \left(\frac{2(1-\vartheta)}{3(D-1)}\right)^{1/n}$$

at even n. Then $1 - \kappa_n \ge \vartheta$ and $D + 1 - 3\kappa_n \ge D - 2 + 3\vartheta$. A learning-rate-uniform noise threshold additionally requires the lower endpoint to be bounded away from $\eta h_* = 1$. All such statements retain a fixed finite total horizon.

Nonzero targets and balanced signs. Let $p \neq 0$, set $a = |p|^{1/D}$, and initialize $w_{\ell,*} = as_\ell$, where $s_\ell \in \{-1, 1\}$ and $\prod_\ell s_\ell = \operatorname{sign}(p)$. Writing $w = aSz$, with $S = \operatorname{diag}(s_\ell)$, transforms the actual dynamics into the unit-target dynamics with

$$\eta_{\text{eff}} = \eta|p|^{2-2/D}, \qquad \sigma_{\text{eff}} = \sigma/|p|, \qquad h_* = D|p|^{2-2/D}.$$

The possible sign of p is absorbed into the centered effective sample noise. In particular, $\eta_{\text{eff}} D = \eta h_*$.

For an original-coordinate unit probe u, the transformed initial probe is Su/a, while the output transformation multiplies by aS. The common scale cancels in the input–output derivative and S is orthogonal. Thus the gain comparison remains a comparison of the original Euclidean parameter responses. The coefficient is expressed in $(\sigma/|p|)^2$, and the physical noise threshold is $|p|$ times the unit-target threshold.

Corollary G.3 (Layer-scaling probes in wider rank-one networks)**.** *Let the input and output be scalar, let hidden widths be arbitrary, and choose unit vectors v_ℓ with $v_0 = v_D = 1$. Initialize all layer matrices at*

$$W_{\ell,*} = v_\ell v_{\ell-1}^\top.$$

All matrix entries are trained. For probes in

$$\mathcal{U} = \left\{ (\alpha_\ell v_\ell v_{\ell-1}^\top)_{\ell=1}^D : \alpha_\ell \in \mathbb{R} \right\},$$

the conclusion of Theorem G.2 holds with the full parameter Euclidean norm, uniformly over fixed unit $u \in \mathcal{U}$.

Proof. The subspace of factorizations $W_\ell = w_\ell v_\ell v_{\ell-1}^\top$ is preserved by every sampled gradient update and by mixing. Restricted to it, the actual loss and updates are exactly those of $g(w) = \prod_\ell w_\ell$. Each layer basis matrix has Frobenius norm one, so the map from scalar layer coefficients to full parameters is an isometry. An injected probe in $\mathcal{U}$ remains in this invariant subspace for the perturbed process as well. The scalar-chain theorem therefore applies without changing the Euclidean parameter-response norm. □

This corollary does not claim strict improvement in all directions of a wider network. For depth three and width two, take $W_1 = e_1$, $W_2 = e_1 e_1^\top$, $W_3 = e_1^\top$. A probe changing only $(W_2)_{22}$ has no effect on any prediction or sampled gradient along the resulting paths. It remains unchanged by both algorithms, giving exactly $\mathcal{G}^u_{n,k} = \mathcal{E}^u_{n,T} = 0$. Thus a full-direction strict extension over arbitrary widths is false.

A single injection can be initially adverse but cumulatively favorable. For the two-factor process, take $m = 2$, $K = 1$, $\eta = 3/4$, $\rho = 1/2$, and inject at $n = 1$. For the fixed directions $e_\pm = (1, \pm 1)/\sqrt{2}$, the coefficients of σ^2 are

	$\mathcal{G}_{1,1}$	$\mathcal{G}_{1,2}$	$\mathcal{E}_{1,2}$
e_+	$81/512$	$-11259/32768$	$-6075/32768$
e_-	$27/256$	$-2079/4096$	$-1647/4096$

as obtained from Eqs. (G.19) and (G.20). Both channels have a worse first response, but better second terminal and two-step cumulative responses. Because the inequalities are strict, this behavior persists in an open neighborhood of these parameters. It is a once-injected, continuously propagated response, not a sum of separately injected one-step probes.

The oscillatory lower bound is substantive. For the same two-factor model, take $n = 2$, $\eta = 1/4$, $\rho = \sqrt{3}/2$, $m = 2$, and $K = 1$. The normal direction satisfies

$$\mathcal{G}^{e_+}_{2,2}(\sigma) = \frac{675}{524288}\sigma^2 + O(|\sigma|^3), \qquad \mathcal{E}^{e_+}_{2,2}(\sigma) = \frac{3555}{524288}\sigma^2 + O(|\sigma|^3).$$

Both are positive for sufficiently small nonzero noise, although the sharpness theorem still gives a negative expected sharpness gap. Lower sharpness alone therefore does not imply the multistep response comparison.

What the comparison does and does not order. The theorem is uniform over fixed unit input directions and finitely many terminal times. It does not compare sample-dependent worst-case directions, pathwise norms, or expectations of operator suprema. It also does not assert that the decentralized gain is below one. At zero noise, layer-imbalance directions have unit gain at every time, so infinite cumulative squared responses need not be finite. No infinite-horizon energy or invariant-law conclusion is used here.

H Theory Experiments and Reproducibility

H.1 Weight trajectories and mixing in deep linear networks

Figures 5 and 10 use scalar, width-one networks $g(w) = \prod_{\ell=1}^{D} w_\ell$, with target $p = 1$ and prediction loss $L_0(w) = \frac{1}{2}(g(w) - 1)^2$. Inputs and label noise are independent Rademacher variables, as in the matrix-network experiments of Subsection H.3. The population loss is $L_\sigma = L_0 + \sigma^2/2$, so both losses have the same Hessian. Four workers use one fresh sample each per step, noise amplitude $\sigma = .3$, and $\mathcal{W} = \rho I + (1 - \rho)J$, with $\rho \in \{.5, .95\}$. Larger ρ means weaker averaging, as in Figures 4 and 6. The centralized comparator uses the same four samples at its own weights, and decentralized measurements use the mean of the worker weights.

Initialization and exact loss slice. For each $D \in \{3, 6, 9, 12\}$, all methods start at

$$w_{1,0} = 0.22/3.2, \qquad w_{2,0} = \cdots = w_{D,0} = 3.2^{1/(D-1)}, \qquad \eta = \frac{0.3}{\|\nabla g(w_0)\|^2}.$$

Every depth starts with prediction $g(w_0) = .22$ and loss .3042. The learning rate is constant during each run; its values for the four depths are approximately .029211, .028868, .028491, and .028114. All D weights are trainable without parameter tying. The product gradient and mixing preserve equality of layers $2, \ldots, D$, so each worker, the worker mean, and the centralized weights remain on the plane $w_2 = \cdots = w_D$. Coordinates $a = w_1$ and $b = \sqrt{D-1}\, w_2$ preserve Euclidean distances on this plane. The plotted surface is therefore the exact restriction

$$L_0(a, b) = \frac{1}{2}\left[a\left(\frac{b}{\sqrt{D-1}}\right)^{D-1} - 1\right]^2,$$

not a projection of weights from outside the plane. Depth changes both the parameterization and the absolute learning rate; the panels are not an isolated depth ablation.

The factor scale 3.2 was chosen for a clear curvature contrast after exploring scales 2.2, 3.2, and 4.4 at depth three, keeping the initial prediction and normalized step size fixed. This choice is used at every displayed depth. It changes both layer imbalance and the corresponding absolute learning rate, rather than isolating either effect.

Sampling and curvature measurements. Training runs for 5,000 steps with noise throughout. We use 128 independent paired noise streams with seed 2026092412, reusing the samples across depths and mixing values. The displayed stream is index 0, fixed before the exploration; no paths fail or are excluded. We compute the largest eigenvalue of the full $D \times D$ Hessian $H(w) = \nabla g(w)\nabla g(w)^\top + (g(w) - 1)\nabla^2 g(w)$ at the centralized weights or the worker mean, not the average of worker-local sharpnesses. Ensemble curves average these quantities over paths. Each local section in Figure 5 plots $L_0(w + rq)$ for $-.22 \le r \le .22$, where q is a unit top eigenvector of the full Hessian at the starred point. These are exact losses, with curvature $\lambda_{\max}(H(w))$ at $r = 0$. No trajectory smoothing, parameter projection, or vertical shifting is used. The three-dimensional background is omitted above loss .46 for visibility; trajectories are retained.

Table 2: Three-layer step-5,000 measurements over 128 paired realizations. Sharpness and prediction loss are sample means; the gap is the paired sample mean of gossip-minus-centralized sharpness, with one standard error.

Method	Full-Hessian sharpness	Gap ± SE	Prediction loss
Centralized	7.0219	—	.001438
Gossip, $\rho = .5$	6.2104	$-.8115 \pm .0038$	.001245
Gossip, $\rho = .95$	4.2366	$-2.7853 \pm .0072$	.000798

The sharpness values printed in Figure 5 are single-path measurements from stream 0; Table 2 reports the corresponding ensemble estimates. The initialization with residual $-.78$ and noise amplitude .3 are not certified by the local small-noise conditions. These finite-time experiments illustrate the geometry; they do not test the slow-limit speed formula or establish distinct limiting solutions.

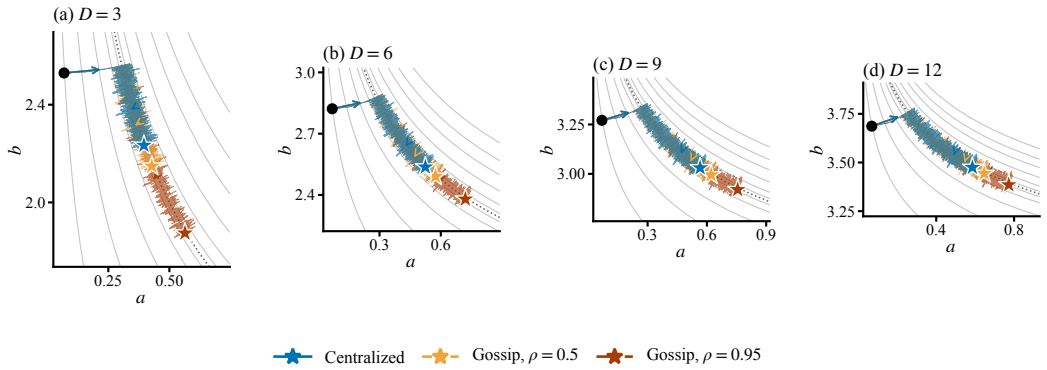


Figure 10: Weight trajectories at depths (a) 3, (b) 6, (c) 9, and (d) 12. Gray contours show prediction loss on the exact plane with coordinates $a = w_1$, $b = \sqrt{D-1}\, w_2$; the dotted curve is the fitted set $g = 1$. Colors and line styles match Figure 5. Black dots mark initialization and colored stars mark step 5,000. Panel (a) reuses the three-layer trajectories and endpoints in Figure 5.

H.2 Sensitivity experiments along the initial prediction gradient

Networks and matched training. The fitted reference has $W_1 = (1,1)^\top$, $W_D = (.6,.4)$, and every intermediate matrix equal to $\left(\begin{smallmatrix} .8 & .2 \\ .1 & .9 \end{smallmatrix}\right)$. Thus $g(\theta_*) = 1$, every hidden width is two, and the $4D - 4$ entries are independently trained. Figure 6 uses depths $D = 2, 4, 8$, whose h_* values are $2.52, 4.54419, 8.70754$. At $\alpha = 1.5, n = 4, \rho = .5$, their margins in Equation (G.1) are $4.9991, 18.3544, 63.3536$. The same reference family also has positive margins $11.2851, 26.7873$ at depths three and five. Here $\delta_4 = 331/6355$ and $J_4 = 12816/4369$ retain the exact graph-filtered history.

Each group starts at $\theta_0 = \theta_* + \varepsilon u_j$, where u_j is a normalized Gaussian direction in the full parameter space. We use 12 independent groups, each with 2,048 paired training-noise paths. The two workers use $K = 1$, effective Rademacher label noise of amplitude $\sigma = .02$, and $\mathcal{W} = \rho I + (1-\rho)J$. The centralized model uses the same two samples at its own state. At $n = 4$, the initial unit normal $q_0 = \nabla g(\theta_0)/\|\nabla g(\theta_0)\|$ is injected once into each tangent process. Exact Hessian–vector products propagate this infinitesimal probe alongside the continuing noisy training for 16 steps. We measure the squared norm of the averaged worker tangent, not the average of the worker tangents' squared norms.

Comparisons and uncertainty. All three panels report $\mathcal{E}^{q_0}_{4,16}(\sigma)/\sigma^2$ on the grid $\alpha \in \{.1, .2, .4, .6, .8, 1, 1.2, 1.4, 1.5, 1.6, 1.8, 1.9\}$. Panel (a) varies depth at $\varepsilon = .1, \rho = .5$. Panel (b) varies $\varepsilon \in \{.01, .05, .1, .3\}$ at depth four and $\rho = .5$, sharing initialization directions and noise across radii. Panel (c) varies $\rho \in \{.2, .5, .8\}$ at depth four and $\varepsilon = .1$; larger ρ means less averaging per step. Means average paired gaps within each group and then across groups. We resample whole groups 5,000 times and use the 95th percentile of the maximum standardized centered bootstrap deviation over 168 estimates: the 120 shown in the figure and 48 from an additional experiment varying the noise amplitude. The resulting approximate simultaneous bands have critical multiplier 2.8816. The replication unit is the initialization/noise group, not an individual noise path. All 1,584 group configurations, including those from the additional experiment, completed without failed or excluded paths. All grid points are retained without smoothing; signed logarithmic axes have a linear region $[-.001, .001]$.

At depth four and $\alpha = 1.5, \rho = .5$, the normalized gaps for the four radii are $-1.7832, -1.8240, -1.9119, -2.8937$. Their simultaneous intervals are $[-1.8461, -1.7203]$, $[-1.9742, -1.6738]$, $[-2.2035, -1.6203]$, and $[-4.2616, -1.5259]$. At radius .3, initial predictions range from .6871 to 1.2751, and variation across groups is larger. At $\varepsilon = .1$, changing ρ from .2 through .5 to .8 gives gaps $-.2523, -1.9119, -6.4628$. Thus the tested initialization displacements

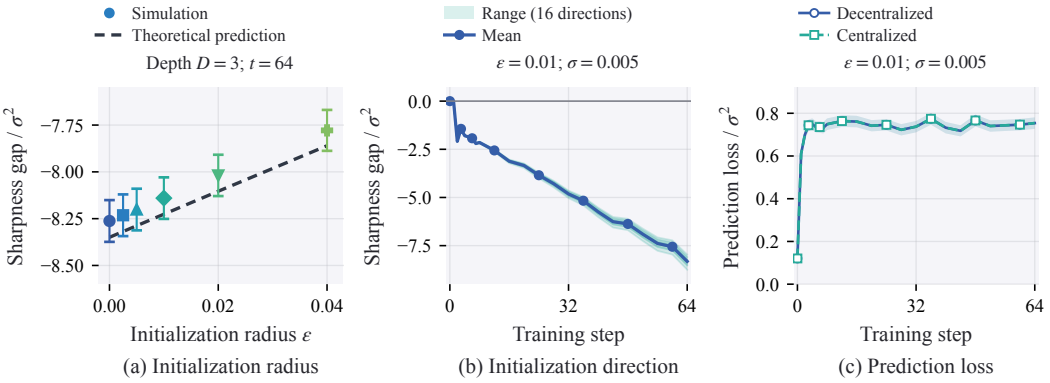


Figure 11: Initialization robustness and prediction loss at $D = 3$, $\sigma = 0.005$, and $\mathcal{W} = J + 0.5(I - J)$. (a) Step-64 sharpness gap along one fixed full-parameter direction, at six initialization radii; the vertical range is enlarged to resolve differences. The dashed line connects moving-baseline theoretical predictions. Error bars are pointwise ± 1.96 paired standard errors from 4,096 noise trajectories at each radius. Radii 0.02 and 0.04 exceed the sufficient radius 0.01 evaluated analytically. (b) Sixteen additional initialization directions at radius 0.01, each using 1,024 paired noise trajectories. The line averages their estimated gaps; shading shows the minimum–maximum range across direction-specific estimates, not a confidence interval. (c) Half-squared prediction residual for the original direction, with 4,096 trajectories and pointwise ± 1.96 standard-error bands. All panels divide their respective observables by σ^2; sharpness gaps are decentralized minus centralized. No trajectories or directions are excluded. The sampled finite-noise diagnostics are not a uniform-ball or infinite-time guarantee.

preserve the high-rate reduction, while mixing changes its magnitude. The actual gaps equal these normalized values times $\sigma^2 = 4 \times 10^{-4}$. These finite-noise observations do not certify that the experimental radii and noise satisfy the initialization and noise thresholds whose existence the theorem guarantees.

Reproducibility. The runner, per-group records, validation checks, plot generator, and statistics are in `illustration/normal_response/preview_sweep/`. Initialization directions use seed $2026092400 + 100D$, noise group j uses seed $2026092500 + j$, and the bootstrap seed is 2026092116. Gradient, Hessian–vector, and finite-difference tangent checks accompany zero-noise and full-averaging equality controls. An additional experiment varies the noise amplitude over $\sigma \in \{.005, .01, .02, .05\}$ at depth four, $\varepsilon = .1$, $\rho = .5$; these data are not an additional manuscript panel.

H.3 Near-fit numerical protocol and initialization robustness

This section gives the protocol for Figure 4 and additional initialization checks.

For every tested depth $D \geq 2$, use the fitted reference

$$W_{1,*} = (1, 1)^\top, \qquad W_{k,*} = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix} \ (2 \leq k < D), \qquad W_{D,*} = (0.6, 0.4).$$

The intermediate rows sum to one, so $g(\theta_*) = 1$. Flatten the entries in layer order, row-major within a layer, and set $\theta_0 = \theta_* + \varepsilon u_{D,j}$, where $u_{D,j}$ is a standard Gaussian vector normalized to unit length. NumPy's default generator uses seed $20260918 + D + 1000j$ for this direction. The main panels use $j = 0$, $\varepsilon = 0.01$. The two-layer baseline has initial prediction 1.0083189, rather than one; the depth-three robustness experiments start at prediction 1.0024537. Across the six main depths, the initial prediction ranges from 0.9956264 to 1.0083189. Directions and grids are fixed before measuring outcomes. All entries are trainable without parameter tying; no balancing, projection, restart, or frozen parameter is introduced.

Each worker performs one minibatch-SGD update before mixing, as in Algorithm 1. There are two workers and one sample per worker per step. All workers and the centralized model share the

same initialization and iid samples from $(x, y) = (\pm 1, x \pm \sigma)$. The centralized model evaluates those same two samples at its own state. Independent effective Rademacher signs are drawn with seed 20260918. Each main configuration uses 4,096 independent paired noise paths; the same sign arrays are reused across configurations to reduce comparison noise. Configuration-level estimates are therefore not treated as independent replicates. The normalization uses σ^2 only, and the step size is fixed by the reference curvature, $\eta h_* = 1.5$. Matching ηh_* does not hold the absolute step size or higher-order geometry fixed across depths. Training runs for 64 steps; measurements occur at $0, 1, \ldots, 8, 12, 16, \ldots, 64$ without smoothing.

The noise grid is $\{0.005, 0.03, 0.05, 0.07, 0.09, 0.12\}$ at depth two. The depth grid is $\{2, 3, 4, 5, 8, 12\}$ at $\sigma = 0.005$. Both use $\mathcal{W} = J + 0.5(I - J)$. At depth two and $\sigma = 0.005$, we compare $\rho \in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 0.9\}$ using measurements at step 64. The theoretical mixing curve evaluates the recurrence on 181 equally spaced values of ρ; it is not a fitted regression.

Full Hessian and moving theoretical prediction. Exact forward differentiation of matrix products computes g, $v = \nabla g$, $Q = D^2 g$, and $D^3 g$. The evaluated Hessian and third loss derivative are

$$H = vv^\top + (g - 1)Q, \qquad D^3 L = (g - 1)D^3 g + \mathrm{sym}(v \otimes Q),$$

where sym sums the three distinct placements of v. At each sampled state we compute the largest eigenvalue of the full Hessian, not a Gauss–Newton, layerwise, or directional proxy. For each paired path, the gap is evaluated at the averaged decentralized parameters and the centralized parameters; worker-local sharpnesses are not averaged. Finite-difference checks verify the third derivative, loss tensor, and top-eigenvalue derivative. The moving response Eqs. (E.2) and (E.3), initialized at zero, yields $\nabla\lambda(b_t)^\top r_t/2$ for the plotted theoretical gap divided by σ^2. At an exact fit it reproduces the earlier fitted recursion. Full averaging ($\rho = 0$) reproduces the centralized process to floating-point precision.

Initialization robustness and uncertainty. Figure 11 reports negative estimated step-64 gaps across all six tested radii and all sixteen additional directions. These finite collections do not verify the entire parameter ball. The radius-0.01 analytical certificate applies to the stated depth-three reference and original graph; the other depths, graphs, and larger radii are additional empirical design points. We vary the initialization distance over $\varepsilon \in \{0, 0.0025, 0.005, 0.01, 0.02, 0.04\}$ along the same $u_{3,0}$, with 4,096 paired paths at each distance. To compare initialization directions, we use $u_{3,j}$ for all $j = 1, \ldots, 16$ at radius 0.01, with 1,024 paths per direction. Their initial predictions range from 0.9912267 to 1.0072168, spanning both under- and overprediction. All 49 evaluated configurations complete with zero failed or excluded paths. The per-configuration standard error is the sample standard deviation of paired gaps divided by the square root of the number of noise paths. The band in Figure 11(b) instead shows variation across direction-specific estimated means, which also contain Monte Carlo error. It is not a noise-path confidence interval or a bound over all directions. The normalized step-64 gaps across these directions range from -8.75 to -7.96. At $\sigma = 0.005$, the depth-three robustness baseline's unnormalized sharpness gap is approximately $-2.04 \cdot 10^{-4}$.

The prediction loss is $\frac{1}{2}(g(\theta) - 1)^2$, excluding the additive population label-noise floor $\sigma^2/2$. At step 64, both methods' prediction losses are about $1.9 \cdot 10^{-5}$. The close loss curves do not assert equality of loss or isolate sharpness at matched loss. These finite-noise measurements complement the analytical certificate rather than verify its conservative sufficient noise thresholds: for the radius-0.01 depth-three ball, the explicit sufficient noise threshold through step 64 is $2 \cdot 10^{-6}$, whereas the main diagnostic uses 0.005 and above.

Reproducibility. The offline experiment runner, complete per-configuration records, manifest, validation checks, and figure generators are in `illustration/nearfit_v8/`; the additional depth-two experiments and revised main figure are in `illustration/depth2_main/`. The manifest records all settings and a source hash; results retain raw sharpnesses, paired gaps and standard errors, prediction losses, initial parameters, and moving-baseline predictions. No external training service is needed for these synthetic experiments. The shared plotting style preserves the main empirical figures' serif typography and light panels, with marker shapes accompanying the blue–teal–green palette.

H.4 Beyond the certified initialization neighborhood

The step-64 sharpness advantage also occurs at the larger initialization displacements $\varepsilon = 0.1$ and 1 shown in Figure 12. We use the same full-parameter directions $u_{D,0}$ and fitted references as in Subsection H.3, with $\theta_0 = \theta_* + \varepsilon u_{D,0}$. At depth three, the initial predictions are 1.0257 and 1.3673, respectively, rather than the target 1. When varying noise amplitude or depth, we fix $\rho = 0.8$; the mixing comparison varies ρ as before. All other sampling and training settings are unchanged, including $\eta h_* = 1.5$ with h_* evaluated at the original fitted reference.

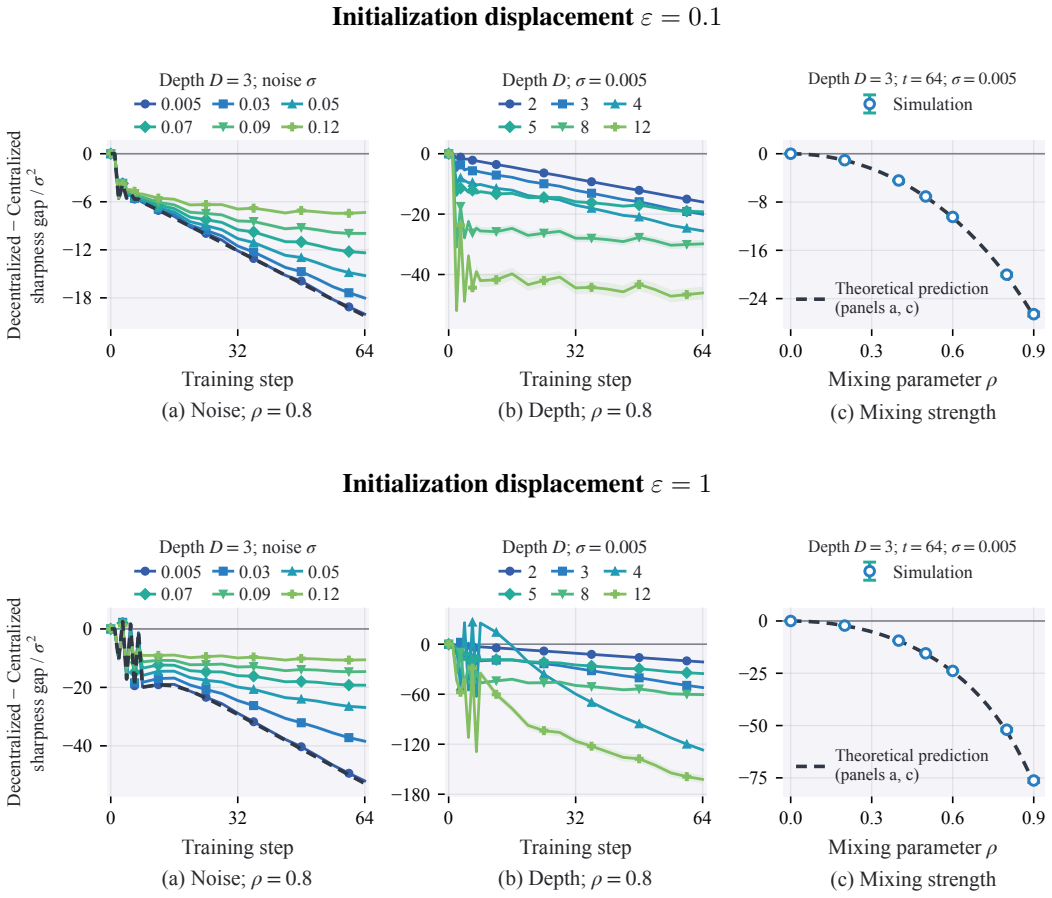


Figure 12: Sharpness gaps across noise amplitudes, depths, and mixing parameters at two larger initialization displacements. Each row follows the three-panel organization of Figure 4: (a) noise levels at $D = 3$, (b) depths at $\sigma = 0.005$, and (c) mixing at $D = 3$, $\sigma = 0.005$, and step 64. Negative normalized sharpness gaps favor decentralized training. Dashed curves are moving-baseline second-order predictions; bands and error bars are pointwise ± 1.96 paired standard errors from 4,096 independent noise paths per configuration. All 34 configurations displayed here completed without path exclusions. Axes are scaled independently for readability; lines connect recorded checkpoints without smoothing.

At both displayed radii, every tested nonzero-mixing configuration has a negative estimated gap at step 64; full averaging agrees with centralized SGD to numerical precision. This endpoint advantage need not hold throughout training: at $\varepsilon = 1$, the depth-three and depth-four curves have positive early gaps before becoming negative. These selected finite-horizon examples demonstrate sharpness reduction beyond the concrete radius-0.01 sufficient certificate, but do not enlarge its certified ball or establish an all-time ordering. The complete paired measurements and plotting scripts accompany these supplementary experiments.