FROM THE EDGE OF STABILITY TO CONVERGENCE THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient descent is usually analyzed under smoothness conditions that guarantee a decrease in the objective at every step. In modern neural-network training, however, gradient descent often operates at the Edge of Stability (EoS), where the largest Hessian eigenvalue oscillates around or above the classical stability threshold and the loss is not monotone. We study what controls convergence in this regime. Our analysis uses an averaged directional sharpness that takes into account the Hessian spectrum along the gradient-descent update. We show that a squared-gradient-weighted stability margin built from this quantity exactly determines the cumulative loss decrease along the trajectory. Lower bounds on this margin directly give gradient-stationarity rates, as well as convergence rates under PL condition. To control the margin when the loss oscillates, we introduce a sufficient condition based on comparing iterates k steps apart. The condition is derived using classical smoothness, but can still hold when individual gradient-descent steps increase the loss. Importantly, we show that it is satisfied by various theoretical neural-network models. Experimentally, we verify lower margin bounds in MLPs and Vision Transformers. Across full-batch experiments on CIFAR–10 and Tiny ImageNet, the cumulative margins is shown to control the dynamics, and we further characterize how mini-batching affects averaged sharpness.

1 INTRODUCTION

Gradient descent (GD) with step size η is typically analyzed under smoothness conditions that ensure every sufficiently small step decreases the objective. Yet, modern neural networks mostly operate in a different regime, where the largest eigenvalue of the Hessian crosses and oscillates slightly above the usual stability threshold $2/\eta$, a phenomenon known as the *Edge of Stability* (EoS) (Jastrzębski et al., 2019; Cohen et al., 2021). In this regime, the loss may oscillate and even increase over individual iterations while still decreasing in the long term. These observations challenge the classical optimisation picture of gradient descent, in which convergence is explained through a uniform upper bound on curvature that guarantees monotone descent. This raises a basic question: *what property of the optimization trajectory controls convergence when GD operates at or beyond the classical stability threshold?*

To answer this question, we move beyond worst-case curvature and instead consider sharpness averaged according to the full Hessian spectrum and the update direction. For a single GD step, this Hessian spectrum is naturally averaged along the segment between two consecutive iterates, introducing the *averaged directional sharpness* $\tilde{\lambda}^t$. It is closely related to directional smoothness (Mishkin et al., 2024), used to estimate local gradient smoothness in adaptive methods. Figure 1 illustrates its distinction from traditional sharpness: later in training, the top eigenvalue stays above $2/\eta$, while $\tilde{\lambda}^t$ oscillates around it. Our main premise is that convergence at the EoS is governed by the directional curvature encountered along the GD trajectory, rather than by the largest Hessian eigenvalue.

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.

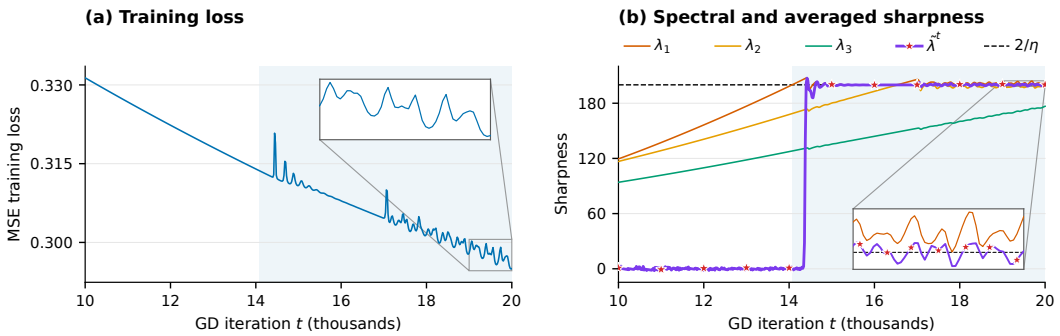


Figure 1: Training loss and sharpness diagnostics for a two-layer tanh MLP of width 200, trained on CIFAR-10 with the MSE objective by full-batch GD; see Appendix G.1 for details. The insets magnify the last 1000 optimizer updates, and the right inset compares the largest Hessian eigenvalue with the averaged directional sharpness $\tilde{\lambda}^t$. The dashed line marks $2/\eta$, and the blue zone indicates EoS.

More precisely, we define the margin $m_t := 1 - \frac{\eta\tilde{\lambda}^t}{2}$, and let $M^g_{s:T}$ denote its average weighted by the squared gradient norms. A key observation is that this quantity exactly characterizes the loss decrease: $f(x_s) - f(x_T) = \eta M^g_{s:T} \sum_{t=s}^{T-1} \|g_t\|_2^2$.

This identity directly connects the averaged directional sharpness to convergence. In particular, any lower bound on $M^g_{s:T}$ yields a gradient stationarity rate. The rate at which $M^g_{s:T}$ can approach zero determines how quickly the average squared gradient norm vanishes. If the trajectory additionally satisfies a local μ-PL inequality, the same argument yields a convergence rate for the loss. We summarize these in Theorem 1. To our knowledge, such convergence rates are new.

The remaining question therefore becomes how to control $M^g_{s:T}$. To account for the fact that the loss does not decrease monotinically, we compare iterates that are k steps apart. We then derive a sufficient condition using classical L-smoothness applied to a Taylor expansion between these iterates (Definition 1). Unlike the standard one-step descent condition, this condition is compatible with oscillatory GD trajectories and can hold even when individual steps increase the loss and provably controls $M^g_{s:T}$ (Theorem 2). This condition relates to a few already existent ones, see (Chen & Bruna, 2023; Song & Yun, 2023; Ahn et al., 2023). The main differences are: i) Our condition is of non-asymptotic nature and ii) Our condition covers the case of k-step progress, while the cited ones cover only the case of two-cycles and they cannot be generalized in a straightforward way. We discuss these relationships in Appendix E.

Importantly, we show that the aforementioned condition is satisfied by a range of neural-network architectures. In Proposition 4, we derive three scalar constructions exhibiting the full EoS phenomenology, with persistent loss and sharpness oscillations, while our condition still guarantees convergence every two iterations. In Proposition 5, we construct a family of functions whose first loss improvement happens after exactly k steps and for which our condition holds strictly at size k. This example illustrates our main distinction particularly clearly: sharpness remains above the $2/\eta$ threshold and slowly approaches it, whereas averaged sharpness eventually stays strictly below the threshold, yielding monotonic convergence. Existing constructions typically exhibit either long-term progress or exact cyclic behavior; see, e.g., Chen & Bruna (2023); Meng et al. (2024); Liang & Montúfar (2026); Mulayoff & Stich (2026). The novelty of our constructions is that a general condition for long-term progress coexists with explicit loss and sharpness oscillations, or with sharpness persistently above the threshold. Establishing this oscillatory behavior and GD invariance, beyond simply verifying our condition, is an important technical challenge.

We finally validate these predictions in neural networks trained with full-batch gradient descent, thereby isolating the deterministic EoS dynamics studied by our theory. More precisely, we show empirical lower bounds for the cumulative margin and local PL inequalities. Moreover, we test mini-batch gradient descent for large-scale training. We show that a mini-batch generalization of a relevant quantity in (Lee & Lee, 2026) showcases some main points of our deterministic theory in stochastic optimization. This notion is related but not the same with ones proposed in (Ahn et al., 2022; Lee & Jang, 2023; Andreyev & Beneventano, 2024).

2 FROM AVERAGED SHARPNESS TO CONVERGENCE

2.1 THEORETICAL SETTING AND LOSS DYNAMICS

Throughout this work, we consider full-batch gradient descent (GD) applied to a C^2 function $f : \mathbb{R}^d \to \mathbb{R}$ and denote its updates by

$$x_{t+1} = x_t - \eta g_t, \qquad g_t := \nabla f(x_t), \qquad \forall t \in \mathbb{N}.$$

While we focus on full-batch GD to keep the analysis simple, we discuss the extension to stochastic (mini-batch) gradient descent in Appendix H.

Applying Taylor's theorem with integral remainder to f gives

$$f(x_{t+1}) - f(x_t) = -\eta \|g_t\|^2 + \eta^2 \int_0^1 (1-\tau) g_t^\top \nabla^2 f(x_t - \tau\eta g_t) g_t \, \mathrm{d}\tau.$$

Equivalentlty, we can normalize both sides by the squared gradient norm and change the sign to track loss decrease. This gives

$$\frac{f(x_t) - f(x_{t+1})}{\|g_t\|^2} = \eta \left(1 - \eta \frac{\int_0^1 (1-\tau) g_t^\top \nabla^2 f(x_t - \tau\eta g_t) g_t \, \mathrm{d}\tau}{\|g_t\|^2} \right)$$

One can see, that the dynamics of the loss decrease are captured by the integral fraction normalized by the gradient on the right side. We call it the *averaged directional sharpness* and denote it by $\tilde{\lambda}^t$:

$$\tilde{\lambda}^t := \frac{g_t^\top \left(2 \int_0^1 (1-\tau) \nabla^2 f(x_t - \tau\eta g_t) \, \mathrm{d}\tau \right) g_t}{\|g_t\|^2}.$$

So far, we have shown that $f(x_t) - f(x_{t+1}) = \eta \|g_t\|^2 \left(1 - \frac{\eta}{2} \tilde{\lambda}_t \right)$.

Summing between iterations s and T yields

$$f(x_s) - f(x_T) = \eta \sum_{t=s}^{T-1} \|g_t\|^2 \left(1 - \frac{\eta}{2} \tilde{\lambda}^t \right).$$

This last identity motivates defining the threshold gap as $m_t := 1 - \frac{\eta}{2} \tilde{\lambda}^t$ and the *cumulative margin* $M^g_{s:T}$ by

$$M^g_{s:T} := \frac{\sum_{t=s}^{T-1} \|g_t\|^2 m_t}{\sum_{t=s}^{T-1} \|g_t\|^2}.$$

When $g_t = 0$ for all t, we consider $M^g_{s:T} = 0$.

The loss identity therefore takes the form

$$f(x_s) - f(x_T) = \eta M^g_{s:T} \sum_{t=s}^{T-1} \|g_t\|^2 . \tag{1}$$

which gives us a precise rule for loss decrease: for $s < T$, $f(x_T) < f(x_s) \iff M^g_{s:T} > 0$. $M^g_{s:t}$ is the central quantity in this work and is used to formulate progress for GD at the EoS.

Figure 2(a,b) shows that stepwise margins change sign, whereas the cumulative margins in panels (c,d) remain positive on the displayed horizons.

2.2 CONVERGENCE FROM A CUMULATIVE MARGIN LOWER BOUND

We now derive a general convergence result from the exact identity (1). The result does not require any smoothness-type assumption and therefore applies directly to the EoS regime. We consider both the non-convex and PL settings; analogous guarantees under convexity or an Aiming condition can also be obtained. In the PL case, it is sufficient for the PL inequality to hold only along the GD trajectory; the empirical diagnostic is described in Appendix G.3.

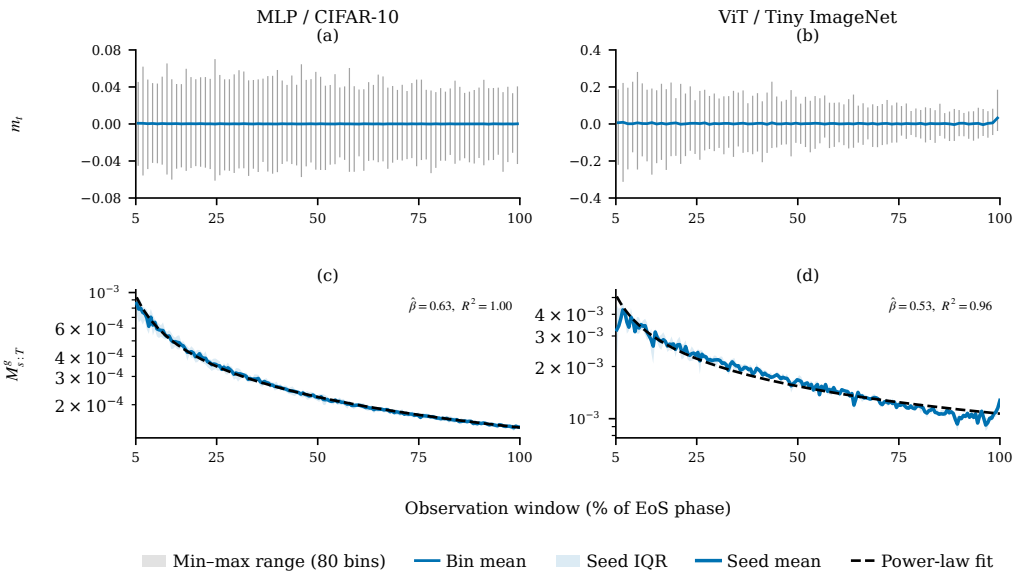


Figure 2: (a,b) Stepwise margins $m_t = 1 - \eta\tilde{\lambda}^t/2$ for $\eta = 0.018$ (left) and $\eta = 1.6$ (right). (c,d) Corresponding cumulative margins M^g with power-law fits. Despite sign-changing stepwise margins, the cumulative margins have positive empirical lower envelopes γ_T.

Theorem 1. *Consider gradient descent on the objective function f, with step-size $\eta > 0$, starting from some $x_s \in \mathbb{R}^d$. Denote $f_\star := \inf_{x\in\mathbb{R}^d} f(x) \in \mathbb{R}$ and $\Delta_t := f(x_t) - f_\star \in [0, \infty)$, $t \in \mathbb{N}$, and write $N := T - s$ for the length of the considered trajectory. Assume $M^g_{s:T} \geq \gamma_T > 0$. Then:*

(i)

$$\min_{s\leq t<T} \|g_t\|^2 \leq \frac{1}{N}\sum_{t=s}^{T-1} \|g_t\|^2 \leq \frac{\Delta_s}{\eta N \gamma_T}.$$

(ii) If in addition f satisfies the Polyak–Lojasiewicz (PL) condition over the GD trajectory, i.e. $\|g_t\|^2 \geq 2\mu\Delta_t$ for some $\mu > 0$, then

$$\min_{s\leq t<T} \Delta_t \leq \frac{1}{N}\sum_{t=s}^{T-1} \Delta_t \leq \frac{\Delta_s}{2\eta\mu N \gamma_T}.$$

This result reveals that convergence behavior is ultimately controlled by a lower bound on the cumullative margin. The proof is a direct consequence of (1) and is presented in Appendix C.

Remark 1 (Classical smoothness)**.** *For $L > 0$, an L-smooth objective with $\eta \in (0, 2/L)$ and $g_0 \neq 0$ satisfies $M^g_{0:T} \geq 1 - \eta L/2 > 0$, for all T. Thus, $\gamma_T > 0$ and Theorem 1 provides the usual rates known in convex optimization. If $\eta > 2/L$, then $1 - \eta L/2 < 0$, and relaxing $M^g_{s:T}$ to $1 - \eta L/2$ makes Theorem 1 inapplicable. In this regime, we instead need a more direct way to control $M^g_{s:T}$.*

For a fixed start s, suppose the cumulative margin is lower bounded as $M^g_{s:s+N} \geq cN^{-\beta}$ for all sufficiently large N, with $c > 0$ and $0 \leq \beta < 1$. By applying Theorem 1, one derives $O(N^{-(1-\beta)})$ gradient stationarity rates. In particular, a constant margin of $\beta = 0$ recovers the classic $O(1/N)$. Under trajectory-local PL with a constant $\mu > 0$ valid over these horizons, the same orders hold for the average loss gap. The constants and exponents may depend on η, s and the trajectory, but not on the endpoint N. Empirically, the lower panels of Figure 2(c,d) compare seed-mean cumulative margins with descriptive power-law fits across various NN architectures and datasets. For a fixed fitted exponent, lowering the prefactor to the minimum of the positive measured values multiplied by the corresponding horizon powers gives a lower envelope on those fitted points. Here these values are seed means; such an envelope is not a bound on each individual trajectory. Experiment-specific stopping rules and exclusions are given in Appendix G.1.

A more interesting empirical finding is that a positive trajectory-local PL constant persists throughout the observed EoS phase, even though a global PL condition is generally not assumed for practical neural-network objectives. Since the losses considered here are nonnegative, $f(x_t) - f_\star \le f(x_t)$, and we can obtain a conservative empirical trajectory-level lower bound on the PL constant by defining

$$\widehat{\mu}_{\mathrm{PL}} := \min_{s \le t < T} \frac{\|g_t\|^2}{2f(x_t)}.$$

Indeed, for every recorded iterate, $\frac{\|g_t\|^2}{2\big(f(x_t)-f_\star\big)} \ge \frac{\|g_t\|^2}{2f(x_t)} \ge \widehat{\mu}_{\mathrm{PL}}$, so that $\|g_t\|^2 \ge 2\widehat{\mu}_{\mathrm{PL}}\big(f(x_t) - f_\star\big)$ throughout the observed EoS window. We find $\widehat{\mu}_{\mathrm{PL}} > 0$ in every EoS phase, see Table 1 and Appendix G.3. Inserting the measured $\widehat{\mu}_{\mathrm{PL}}$ and fitted margin $\widehat{c}N^{-\widehat{\beta}}$ into Theorem 1 suggests the empirical scaling $N^{-(1-\widehat{\beta})}/(\widehat{\mu}_{\mathrm{PL}}\widehat{c})$ for the average loss gap.

Table 1: Observed margin scaling and realized trajectory-local PL constants.

Dataset / model	Loss	$\widehat{\beta}_\eta$: median [range]	$\widehat{\mu}_r$: [min, max]
CIFAR–10k / ViT	CE	$0.556\,[0.542, 0.574]$	$[0.0654, 4.42]$
CIFAR–10k / MLP	MSE	$0.666\,[0.632, 0.841]$	$[0.00200, 2.19]$
Tiny ImageNet / ViT	CE	$0.534\,[0.516, 0.655]$	$[1.10, 24.3] \times 10^{-3}$
Tiny ImageNet / CNN	MSE	$0.570\,[0.541, 0.597]$	$[8.92, 10.0] \times 10^{-6}$

3 Route to margin lower bounds

In the previous section, we showed that lower bounds on $M^g_{s:T}$ translate directly into convergence rates at the Edge of Stability. This leaves a natural question: *under what conditions can such lower bounds be guaranteed?* We now introduce a general condition that addresses this question.

For a C^2 objective, fix an integer $k \ge 2$ and let

$$G(x) = x - \eta\nabla f(x), \quad g(x) = \nabla f(x), \quad S_k(x) := \sum_{r=0}^{k-1} g(G^r(x)), \quad B_k(x) := \sum_{r=0}^{k-1} \|g(G^r(x))\|^2,$$

so that $G^k(x) = x - \eta S_k(x)$. Here, G^k denotes a k-fold composition while G^0 is the identity. A uniform bound $\|\nabla^2 f\|_2 \le L$ on a neighborhood containing $[x, G^k(x)]$ gives, by Taylor's theorem,

$$f(x) - f(G^k(x)) \ge \eta\left(\langle g(x), S_k(x)\rangle - \frac{\eta L}{2}\|S_k(x)\|^2\right). \tag{2}$$

Definition 1. *Consider an arbitrary set Ω, fix $\eta, L, c > 0$ and an integer $m \ge 1$. For a fixed integer k, we say that f satisfies condition (C^k_m) on Ω if all $G^r(x)$, $0 \le r \le k$, belong to Ω, it holds $\|\nabla^2 f\|_2 \le L$ on an open neighborhood in that domain containing $\bigcup_{r=0}^{k-1} G^r(\Omega)$ and every segment $[x, G^k(x)]$ and*

$$\mathcal{C}^{(k)}_L(x) := \langle g(x), S_k(x)\rangle - \frac{\eta L}{2}\|S_k(x)\|^2 \ge c\, B_k(x)^{m+1}, \qquad x \in \Omega. \tag{3}$$

As discussed in the introduction, condition (C^k_m) is closely related to several conditions proposed in recent work. We provide a detailed comparison and clarify some of these connections in Appendix E.

Condition (C^k_m) guarantees a lower bound on $M^g_{s:T}$. This provides a direct connection with the convergence theory of Section 2.2.

Theorem 2. *Assume that condition (C^k_m) holds along the GD trajectory for all $t \ge s$, with $g_s \ne 0$. Then, for every $T > s$ such that $T - s$ is a multiple of k, we have*

$$M^g_{s:T} = \Omega\Big((T-s)^{-m/(m+1)}\Big). \tag{4}$$

For $k = 2$ and $\eta L \ge 1$, we have the stronger bounds

$$M^g_{s:T} = \begin{cases} \Omega\big((\log(2+T-s))^{-1}\big), & m = 1, \\ \Omega\big((T-s)^{-(m-1)/m}\big), & m > 1. \end{cases} \tag{5}$$

The constants may depend on the fixed starting index s, but not on T.

The requirement that $T - s$ be a multiple of k avoids incomplete k-step segments, over which the loss is not guaranteed to decrease. Notice that the case $k = 2$ yields a stronger lower bound for $M^g_{s:T}$. The reason is an artifact of the proof technique: when $k = 2$, condition (C^2_m) provides direct control of the gradient $g(x_{t+1})$ in terms of $g(x_t)$. This yields a tight bound using integration. In the case $k > 2$, such control is absent and we resort to using Hölder's inequality. The proof is in Appendix D.

Connection to semi-algebraicity. On a compact semi-algebraic set Ω and a semi-algebraic function f, one can guarantee a bound of the form (3), assuming only simple positivity of $\mathcal{C}^{(k)}_L(x)$ via a Łojasiewicz-type inequality, using (Basu & Mohammad-Nezhad, 2024, Theorem 1.1). A semi-algebraic function is, in high-level, a function whose graph is created by finitely many polynomial equations and inequalities. We defer exact definitions in Appendix D. Feedforward neural networks with polynomial or ReLU activations and linear attention with MSE loss are indeed semi-algebraic.

Proposition 3. *Let f be C^2 and semi-algebraic on an open semi-algebraic domain, and let Ω be a nonempty compact semi-algebraic set. Fix $\eta, L > 0$ and an integer $k \geq 2$, and assume the domain and uniform Hessian-bound requirements of Definition 1 on Ω. If*

$$\mathcal{C}^{(k)}_L(x) > 0 \qquad \text{for every } x \in \Omega \text{ with } g(x) \neq 0,$$

then there exist $c > 0$ and an integer $p \geq 2$ such that

$$\mathcal{C}^{(k)}_L(x) \geq c\, B_k(x)^p, \qquad x \in \Omega. \tag{6}$$

Thus (C^k_m) holds on Ω with $m = p - 1$.

If $G(\Omega) \subseteq \Omega$, Theorem 2 applies to trajectories starting in Ω and gives lower bounds on $M^g_{s:T}$ in terms of p. This result simply says that if $\mathcal{C}^{(k)}_L(x) > 0 \quad$ for every $x \in \Omega$, then it cannot be smaller than any power of the gradient norm. We defer the proof to Appendix D.

4 Neural-network examples that satisfy condition (C^k_m)

We now verify condition (C^k_m) for several simple neural-network objectives whose gradient-descent dynamics can be analyzed explicitly. In each case, Theorems 1 and 2 then yield an explicit convergence rate. These examples show that (C^k_m) can hold along GD trajectories even when sharpness oscillates around $2/\eta$. The condition therefore captures optimization progress across k-step oscillations in the EoS regime, and more generally in regimes where the sharpness may exceed $2/\eta$.

Our first result is for $k = 2$ and illustrates a construction strategy leading to different values of m. For transparency, we express the objectives in terms of a scalar residual, but each corresponds exactly to the training objective of a small neural network with a standard loss; explicit realizations are given in Remark 2. While the first construction has appeared in various forms, the second and third seem to be novel. More precisely, they are not covered by Lemma 1 of Chen & Bruna (2023), since the higher-order behavior is not created by derivatives vanishing exactly.

Proposition 4. *We write points in $\mathbb{R}^2$ as $w = (w_1, w_2)$ and denote $r(w) := w_1 w_2 - 1$. Consider gradient descent initialized at w^0 with $\eta = 1$ to minimize the function $F_m(w) = \phi_m(r(w))$. We consider the following three models:*

1. $\phi_1(r) = \frac{1}{2}r^2, \quad w_0 = (1, 1-h), \quad 0 < h < \frac{1}{2}.$
2. $\phi_2(r) = \log\cosh\left(r + \frac{r^3}{3}\right), \quad w_0 = \sqrt{1-h}\,(1,1), \quad 0 < h \leq \frac{1}{100}.$
3. $\phi_3(r) = \frac{1}{2}r^2 + \frac{1}{4}r^4, \quad w_0 = \sqrt{1-h}\,(1,1), \quad 0 < h \leq \frac{1}{100}.$

For each $m \in \{1, 2, 3\}$, there exists a compact forward-invariant set $\Omega_m \ni w_0$ on which F_m is L-smooth for $L = 4$ and

- *the following uniform bound holds for $(c_1, c_2, c_3) = (\frac{24}{1681}, \frac{1}{160}, \frac{5}{128})$:*

$$\mathcal{C}^{(2)}_L(w) \geq c_m B_2(w)^{m+1}, \qquad w \in \Omega$$

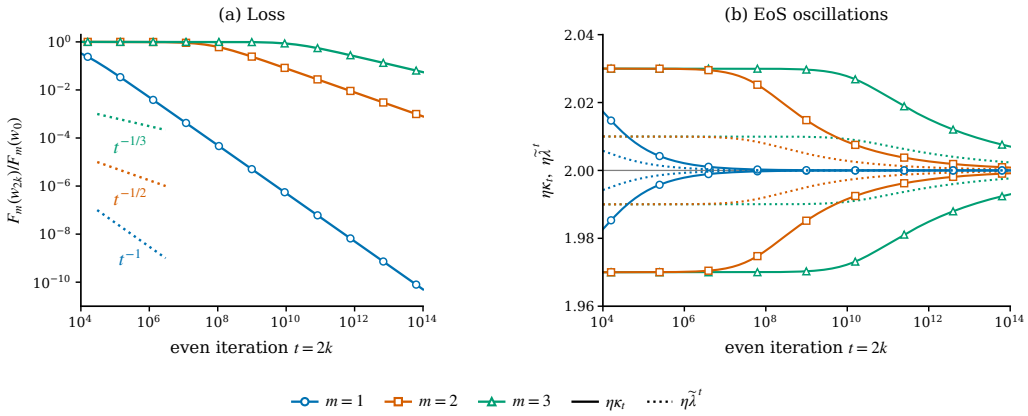


Figure 3: Loss and sharpness behavior for the $m = 1, 2, 3$ models of Proposition 4, with $\eta = 1$ and $h_0 = 10^{-2}$. (a) The normalized loss envelopes exhibit the respective $t^{-1/m}$ scales. (b) The even and odd branches of sharpness approach the EoS threshold 2 from below and above, respectively.

- *for every integer $j \geq 0$, the trajectory satisfies*

$$F(w_{2j+1}) < F(w_{2j+2}) < F(w_{2j}), \quad \eta\kappa_{2j} < 2 < \eta\kappa_{2j+1}, \quad \eta\widetilde{\lambda}^{2j} < 2 < \eta\widetilde{\lambda}^{2j+1},$$

where $\kappa_t = \|\nabla^2 F(w_t)\|_2$ denotes the sharpness at iteration t.

Thus (C_m^2) holds in each case, with the displayed initialization ranges and constants. Along these trajectories, the one-step loss increments alternate in sign, the loss decreases strictly over every two iterations, and the sharpness and averaged sharpness alternate across the threshold $2/\eta$.

Proof sketch: Let $r = r(w^t)$ be residual at point t and $r^+ = r(w^{t+1})$. The main ingredient in all three constructions is a two-step residual update rule of the form $r^{++} = r - cr^{2m+1} + o(r^{2m+2})$, $c > 0$. The r^{2m+1} term has the opposite sign to r and this is exactly the dissipation force that self-stabilizes the dynamics. The power $2m + 1$ determines exactly the power of B_2 in (3). This generates loss oscillation and consequently averaged sharpness oscillation. The exact sharpness excess $\kappa_t - 2/\eta$ can be written as a positive multiple of the residual, thus it also follows the loss oscillation. $1/100$ is a suboptimal constant, the main point is that the result holds non-asymptotically. The full proof is in Appendix F.1.

The examples of Proposition 4 are on scalar parameters and concern only progress in cycles of size two. However, they showcase to some extent the richness of our condition (C_m^k). Below, we demonstrate a general construction, which satisfies our condition for any even $k \geq 4$ and any $m \geq 1$ and seems absent from the literature. Moreover, the sharpness of the GD trajectory stays constantly above the $2/\eta$ threshold and the loss makes its first improvement exactly k steps after initialization. Thus, looking at the whole trajectory, the smallest possible block length such that GD improves the loss over is k.

Proposition 5. *Fix integers $d \geq 2$ and $m \geq 1$, and an even integer $k \geq 4$. Let $\beta, \chi, r_0, \epsilon, c > 0$ be the constants defined in* (43), (63), *and* (65). *For $U, V \in \mathbb{R}^{d\times d}$, consider*

$$F_m(U, V) = \frac{1}{2}\|\sigma_m(U)\|_F^2 + \frac{1}{8}\|V - U^2\|_F^2, \quad \text{with } \sigma_m(s) = s - \frac{\beta d^{2m}}{2m+2}s^{2m+1}$$

applied entry-wise.

Then, the set

$$\Omega := \left\{ (U, V) : \begin{matrix} \|U\|_F \leq r_0, & \|V - U^2\|_F \leq \chi\|U\|_F^{2m}, \\ \|U\|_F^2 - 4\langle U^2, V - U^2\rangle_F \leq \epsilon^2 & \end{matrix} \right\}. \tag{7}$$

is compact and forward invariant under GD and $\|\nabla^2 f(U, V)\|_2 \leq 4/3$ in a convex open neighborhood of it. Moreover, condition (C_m^k) holds:

$$\mathcal{C}_{4/3}^{(k)}(U, V) \geq cB_k(U, V)^{m+1}, \qquad (U, V) \in \Omega. \tag{8}$$

Furthermore, if GD is initialized as

$$U_0 = hP, \quad V_0 = (h^2 - \chi h^{2m})P, \quad \text{with } P = ee^\top,\ e = d^{-1/2}(1,\dots,1)^\top,\ h \in (0, \epsilon/2], \tag{9}$$

then the loss F_j and sharpness κ_j at iterations j satisfy

$$F_j > F_0 > F_k \text{ for } 1 \le j < k,\ \eta\kappa_t > 2 \text{ for } t \ge 0,\ \text{and } \eta\kappa_t \longrightarrow 2. \tag{10}$$

In particular, the first loss value below the initial one occurs exactly at step k, and no shorter block admits a positive sufficient-decrease certificate for this particular initialization. Importantly, the sharpness stays above the $2/\eta$ threshold over the full trajectory. In contrast, the averaged sharpness is below the $2/\eta$ threshold after the first few iterations, creating monotone decrease. This perfectly justifies its use for studying convergence in EoS-type situations. Again, a separate major technical challenge is to identify a non-asymptotic region where the aforementioned happens that contains the dynamics.

Proof sketch: The damping is created as a competition between two terms of order $2m+1$, one from the expansion of U and one from $V - U^2$. In our scalar initialization, it holds roughly $u^+ = -u + 2\beta u^{2m+1} + u(v - u^2) +$ higher order terms. $-u$ term reflection, but since F_m is even in u, it does not change the loss. Loss decrease is driven by the $2\beta u^{2m+1}$ term, while loss increase by $u(v-u^2)$. At the beginning, $u(v-u^2) \approx -\chi u^{2m+1}$, while at iteration j the sum of the two terms is roughly $(j\beta - (1-2^{-j})\chi)u^{2m+1}$. Our parameters are particularly tuned such that $-(j\beta - (1-2^{-j})\chi)$ is negative for exactly the first k iterations. The full proof is in Appendix F.2.

Remark 2. *The first case of Proposition 4 is a simple linear two-layer network with MSE loss. The second and the third case feature neural networks of the form $N_{w_1,w_2}(x) = \tau_m(w_1 w_2 x - 1)$ with $\tau_2(z) = 2(z + z^3/3)$ and $\tau_3(z) = z\sqrt{1 + z^2/2}$, with cross-entropy (labels $(1,0)$ and $(1,1)$) and MSE loss (with single observation $(1,0)$) respectively. The objective of Proposition 5 is the squared training loss of a two-head, shared weight network $N_{U,V}(x) = \begin{pmatrix} \sigma_m(Ux) \\ \frac{1}{2}(Vx - U(Ux)) \end{pmatrix}$ with observations $(e_i, 0)$, where e_i form the standard basis.*

5 Mini-batch averaged sharpness

We finish our exposition with a short discussion of stochastic gradients and arbitrary first-order optimizers. Beyond full-batch GD, the quantity $\eta\tilde{\lambda}^t$ is no longer directly defined: the update $d_t := x_t - x_{t+1}$ need not be proportional to the stochastic gradient $g_t := \nabla f(x_t, \xi_t)$. We therefore retain the normalized finite-step quantity that controls the loss change along the executed update. Let ξ_t be the mini-batch used at update t and set $f_t(x) := f(x, \xi_t)$. When $a_t := \langle g_t, d_t\rangle \neq 0$, we define

$$\zeta_t := 2\left(1 + \frac{f_t(x_{t+1}) - f_t(x_t)}{a_t}\right). \tag{11}$$

If f_t is C^2 along the update segment, Taylor's formula gives $\zeta_t = \frac{2}{a_t}\int_0^1 (1-\tau) d_t^\top \nabla^2 f_t(x_t - \tau d_t) d_t \, d\tau$. Thus ζ_t is the optimizer-normalized analogue of averaged directional sharpness. In particular, for full-batch GD, $d_t = \eta g_t$ and $\zeta_t = \eta\tilde{\lambda}^t$. To connect ζ_t with the stochastic loss dynamics, define the batch displacement term $\varepsilon_{t+1} := f_{t+1}(x_{t+1}) - f_t(x_{t+1})$:

$$f_{t+1}(x_{t+1}) - f_t(x_t) = a_t\left(\frac{\zeta_t}{2} - 1\right) + \varepsilon_{t+1}. \tag{12}$$

Hence $\zeta_t = 2$ is always the loss-change boundary on the mini-batch that produced the update, whereas batch replacement can shift the boundary for consecutive losses (for more details see Appendix H).

This viewpoint is complementary to existing extensions of EoS to stochastic and adaptive optimization, as discussed in the introduction, see also Appendix A for more details.

We first illustrate this distinction on SST-2 from the GLUE benchmark (Wang et al., 2019), using a two-block Transformer with width 64 and eight attention heads, following the small-Transformer setup of Song et al. (2025). We train it with SGD and Adam under both CE and MSE losses. In the four representative trajectories, the two SGD runs remain closer to the same-mini-batch boundary $\zeta = 2$ than the Adam runs (see Appendix H.2 for more details).

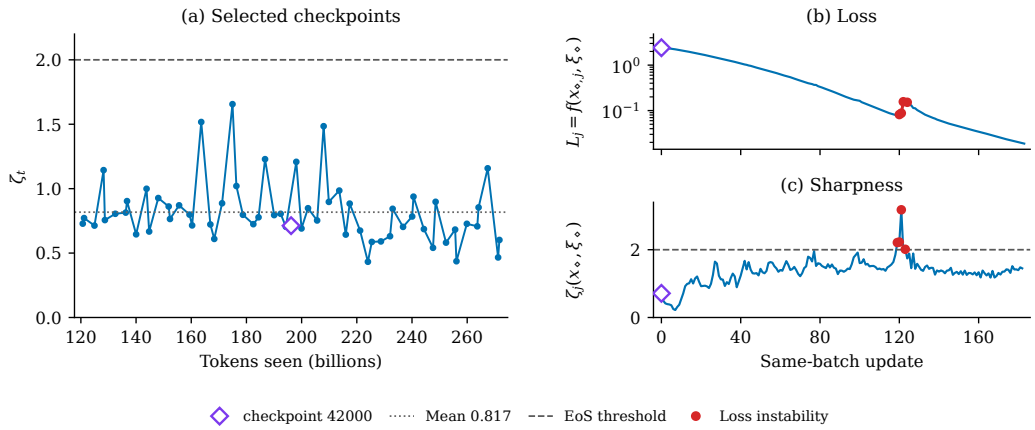


Figure 4: Selected OPEN-1B checkpoints during the stable phase of pre-training. (a) All 60 checkpoint ζ_t residuals lie below 2 (dotted mean 0.817); the left diamond shows ζ_0 at step 42,000 (196.180 billion tokens). (b,c) Continuing AdamW on one fixed mini-batch eliminates batch replacement, i.e. $\varepsilon_{t+1} = 0$; diamonds mark L_0 and ζ_0. The 2 line is the EoS boundary: same-batch updates 119–121 and 123 cross it when loss increases.

Finally, we evaluate our statistics of interest on the published OPEN-1B pre-training trajectory (Donaghy et al., 2026). Starting from the released checkpoints, we replay the corresponding pre-training updates using the original training recipe and the same stochastic mini-batches as in the official run. We verify each replayed update by matching the resulting state hash to the published OPEN-1B hash, ensuring that the statistics are computed on exactly reproduced pre-training states.

For the fixed-checkpoint continuation, let $x_\diamond := x_{42,000}$ and let $\xi_\diamond$ denote the fixed mini-batch. Write $x_{\diamond,j}$ for the state after j continuation updates, with $x_{\diamond,0} = x_\diamond$, and put $L_j := f(x_{\diamond,j}, \xi_\diamond)$. We denote by $\zeta_j(x_\diamond, \xi_\diamond)$ the quantity defined in (11), evaluated at the j-th continuation update on the fixed mini-batch $\xi_\diamond$.

Across 60 uniformly selected checkpoints during the stable phase of pre-training, every realized ζ_t lies below 2. To test the loss-change boundary in Eq. (12), we then restart from $x_\diamond$ and repeatedly optimize the mini-batch $\xi_\diamond$. This intervention removes batch replacement, $\varepsilon_{t+1} = 0$, so that $\zeta = 2$ becomes exactly the observed loss-change boundary Fig. 4.

6 Discussion and Future Work

Our work is the first to establish convergence rates for gradient descent in terms of the averaged directional sharpness $\tilde{\lambda}_t$ that apply to the EoS regime. To quantify progress in blocks of size k, we formulate a sufficient condition that is realistic for many neural networks and can co-exist with loss and sharpness oscillations or sharpness persistently above the threshold. We finally verified the main notions from our theory in practical large-scale training and discussed the effect of mini-batches. There are several promising directions for extending our work. One natural direction is to investigate how the theory could be adapted to stochastic optimization algorithms, where additional sources of randomness make the analysis more complicated. Another direction is to develop continuous-time descriptions of EoS based on averaged sharpness together with convergence guarantees. While several continuous-time models of EoS have been proposed, to the best of our knowledge they rely on worst-case sharpness and do not provide such guarantees. Finally, further work is needed to construct neural-network examples that exhibit persistent oscillations over genuine k-cycles while satisfying our condition (C_m^k). Proposition 5 provides a first step in this direction, but its construction exhibits only a single k-cycle rather than persistent k-cyclic oscillations.

AI USE STATEMENT

In this work, we used generative AI tools, namely the latest models of chatGPT, for literature review, brainstorming new ideas, facilitate mathematical proofs, writing code and polishing the writeup. We have reviewed all AI-assisted work carefully, by verifying that the math is correct, the code does not have bugs, and carefully reviewing the literature for potential plagiarism. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 169–195. PMLR, 2023. URL https://proceedings.mlr.press/v202/agarwala23b.html.

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 247–257. PMLR, 2022. URL https://proceedings.mlr.press/v162/ahn22a.html.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19540–19569, 2023. doi: 10.52202/075280-0858. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3e592c571de69a43d7a870ea89c7e33a-Abstract-Conference.html.

Foivos Alimisis, Rustem Islamov, and Aurélien Lucchi. Why do we need warm-up? a theoretical perspective. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63104.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for SGD. *arXiv preprint arXiv:2412.20553*, 2024. doi: 10.48550/arXiv.2412.20553. URL https://arxiv.org/abs/2412.20553.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024. PMLR, 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

Saugata Basu and Ali Mohammad-Nezhad. Improved effective łojasiewicz inequality and applications. *Forum of Mathematics, Sigma*, 12:e115, 2024. doi: 10.1017/fms.2024.66. URL https://www.cambridge.org/core/journals/forum-of-mathematics-sigma/article/improved-effective-lojasiewicz-inequality-and-applications/022BF859F5714FDA8050F6DC1992E48B.

Yuhang Cai, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L. Bartlett. Large stepsize gradient descent for non-homogeneous two-layer networks: Margin improvement and fast optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2278. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/835a0185f61867a1ea0f86155489839a-Abstract-Conference.html.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23b.html.

Xuxing Chen, Krishnakumar Balasubramanian, Promit Ghosal, and Bhavya Agrawalla. From stability to chaos: Analyzing gradient descent dynamics in quadratic regression. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=Wiklo5VpG7.

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://iclr.cc/virtual/2021/poster/2577.

Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022. doi: 10.48550/arXiv.2207.14484. URL https://arxiv.org/abs/2207.14484.

Michael Crawshaw and Mingrui Liu. Tight bounds for logistic regression with large stepsize gradient descent in low dimension. In *Proceedings of the 39th Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1575–1610. PMLR, 2026. URL https://proceedings.mlr.press/v336/crawshaw26a.html.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11425.

Shenyang Deng, Boyao Liao, Zhuoli Ouyang, Tianyu Pang, Minhak Song, and Yaoqing Yang. Suspicious alignment of SGD: A fine-grained step size condition analysis. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–66. PMLR, 2026. URL https://proceedings.mlr.press/v313/deng26a.html.

John Donaghy, Brian Wilcox, Oğuzhan Ersoy, Shikhar Rastogi, Adam St Arnaud, Alexey Titov, Jordan Greenberg, Ben Fielding, and Harry Grieve. OPEN-1B: A fully auditable training run. *arXiv preprint arXiv:2609.17380*, 2026. URL https://arxiv.org/abs/2609.17380.

Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert Gower. Non-euclidean gradient descent operates at the edge of stability. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/61486.

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *International Conference on Learning Representations*, 2019. URL https://iclr.cc/virtual/2019/poster/875.

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. doi: 10.48550/arXiv.2601.16979. URL https://arxiv.org/abs/2601.16979.

Jaerin Lee and Kyoung Mu Lee. The road taken: The role of optimizers at the edge of stability. *arXiv preprint arXiv:2608.18415*, 2026. URL https://arxiv.org/abs/2608.18415.

Sungyoon Lee and Cheongjae Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/10856.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. doi: 10.48550/arXiv.2003.02218. URL https://arxiv.org/abs/2003.02218.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1749. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7e8bb8d17bb1cb24dfe972a2f8ff2500-Abstract-Conference.html.

Shuang Liang and Guido Montúfar. Gradient descent with large step sizes: Chaos and fractal convergence region. In *International Conference on Learning Representations*, volume 2026, pp. 123445–123492, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c88ccdf819f9afb2ce799f501326150d-Abstract-Conference.html.

Elon Litman. The origin of edge of stability. *arXiv preprint arXiv:2604.20446*, 2026. doi: 10.48550/arXiv.2604.20446. URL https://arxiv.org/abs/2604.20446v1.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 1(3):247–267, 2022. doi: 10.4208/jml.220404. URL https://global-sci.com/index.php/jml/article/view/13179.

Lachlan Ewen MacDonald, Hancheng Min, Leandro Palma, Salma Tarmoun, Ziqing Xu, and René Vidal. Convergence rates for gradient descent on the edge of stability for overparametrised least squares. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4951. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d9af4d6ac714626b652da5616ca71f99-Abstract-Conference.html.

Si Yi Meng, Antonio Orvieto, Daniel Yiming Cao, and Christopher De Sa. Gradient descent on logistic regression with non-separable data and large step sizes. *arXiv preprint arXiv:2406.05033*, 2024. URL https://arxiv.org/abs/2406.05033.

Hancheng Min, Lachlan Ewen MacDonald, and René Vidal. On the convergence, implicit bias, and edge of stability of gradient descent in deep learning: Reviewing recent progress. *IEEE Signal Processing Magazine*, 43(3):37–51, May 2026. doi: 10.1109/MSP.2026.3665010.

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M. Gower. Directional smoothness and gradient methods: Convergence and adaptivity. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0473. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1ac83203e88eb6cf6b30642f0239b932-Abstract-Conference.html.

Rotem Mulayoff and Sebastian U. Stich. On the stability of nonlinear dynamics in GD and SGD: Beyond quadratic potentials. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5210–5243. PMLR, 2026. URL https://proceedings.mlr.press/v336/mulayoff26a.html.

Vincent Roulet, Atish Agarwala, Jean-Bastien Grill, Grzegorz Swirszcz, Mathieu Blondel, and Fabian Pedregosa. Stepping on the edge: Curvature aware learning rate tuners. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1511. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/555479a201da27c97aaeed842d16ca49-Abstract-Conference.html.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3137. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e2a9256bd816ab9e082dfaa22f1f62a2-Abstract-Conference.html.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1757af1fe1429801bdf3abf5600f8bba-Abstract-Conference.html.

Alexander Tyurin. Toward a unified theory of gradient descent under generalized smoothness. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60493–60514. PMLR, 2025. URL https://proceedings.mlr.press/v267/tyurin25a.html.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019. URL https://mlanthology.org/iclr/2019/wang2019iclr-glue/.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0725. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/40bb79c081828bebdc39d65a82367246-Abstract-Conference.html.

Kaiyue Wen, Zhiyuan Li, Jason Wang, David Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6a1fe80a9e2dcda0b3e5fd0fd87eb097-Abstract-Conference.html.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-3247. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/eb189151ced0ff808abafd16a51fec92-Abstract-Conference.html.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 5019–5073. PMLR, 2024. URL https://proceedings.mlr.press/v247/wu24b.html.

Jingfeng Wu, Pierre Marion, and Peter L. Bartlett. Large stepsizes accelerate gradient descent for regularized logistic regression. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3490. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/96b8167534ef3cc30c230bbeb55a524d-Abstract-Conference.html.

Daniyar Zakarin and Sidak Pal Singh. Accelerating neural network training along sharp and flat directions. *arXiv preprint arXiv:2505.11972*, 2025. doi: 10.48550/arXiv.2505.11972. URL https://arxiv.org/abs/2505.11972.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_BJgnXpVYwS.html.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *International Conference on Learning Representations*, 2023. URL https://iclr.cc/virtual/2023/poster/11908.

Appendix

A Related Work

Edge of Stability Early evidence of the interaction between step-size and sharpness was given by Jastrzębski et al. (2019), while Lewkowycz et al. (2020) studied the catapult mechanism through which large learning rates can lead to flatter minima. Cohen et al. (2021) identified the edge of stability (EoS) in full-batch neural-network training: the largest Hessian eigenvalue remains near $2/\eta$, while the loss oscillates over individual iterations and decreases over longer intervals. Subsequent work extended this description to mini-batch SGD through interaction-aware sharpness (Lee & Jang, 2023) and batch sharpness (Andreyev & Beneventano, 2024), to adaptive methods through the preconditioned Hessian (Cohen et al., 2022), and to non-Euclidean gradient methods through sharpness in the corresponding norm (Islamov et al., 2026). Kalra et al. (2026) introduced a computationally inexpensive critical sharpness measure and observed analogous behavior in large language models. The dependence of sharpness on the optimization procedure also matters for learning-rate selection: Roulet et al. (2024) showed that tuning for immediate loss reduction can disrupt curvature stabilization and studied tuners that account for curvature dynamics.

The dominant–bulk experiment of Song et al. (2025) shows why the leading Hessian eigenspace alone does not determine descent: updates restricted to that subspace can stall while updates in its complement make progress. Deng et al. (2026) analyzed the step-size conditions for this separation in quadratic models, and Zakarin & Singh (2025) studied interpolated dominant–bulk updates. Appendix B takes a first step toward explaining why Dom-GD can stall while Bulk-GD descends, sometimes faster than full GD. Its finite-step loss identities identify the required directional sharpness margins and clarify how this separation can relate to progressive sharpening at EoS.

Bifurcation-type conditions Several landscape conditions studied in the EoS literature have been proposed and are related to our k-step sufficient condition presented in Section 3. Chen & Bruna (2023) identify scalar derivative conditions for period-two dynamics, including higher-order cases, while Mulayoff & Stich (2026) derive a multivariate flip coefficient that accounts for coupling to transverse directions and generalized the Chen-Bruna condition to higher dimensions. Complementary conditions describe the shape of scalar losses: Ahn et al. (2023) impose a quantitative deficit of the normalized gradient from its quadratic approximation; Ma et al. (2022) study curvature decreasing away from a minimum; and Song & Yun (2023) use a decreasing gradient ratio to analyze bifurcation and trajectory alignment. In Appendix E, we derive local instances of our proposed condition (C_m^k) at the critical step-size from these assumptions, for scalar objectives or on a critical invariant curve. We distinguish quantitative power bounds from qualitative positivity, for which semialgebraicity can supply a finite exponent. These implications neither identify strict progress on nonstationary cycles nor automatically transfer scalar reductions to the original network parameters. For the stated nonstationary initializations, the examples in Section 4 verify the domain, Hessian, and inequality hypotheses on invariant sets. Theorem 2 then gives cumulative margin lower bounds on complete k-step horizons; its stronger two-step bound is used only when $k = 2$ and $\eta L \geq 1$.

Toy models at EoS Simplified objectives permit a direct analysis of oscillatory descent, see for instance Wang et al. (2022) who analyzed progressive sharpening in two-layer linear networks under additional assumptions, or Zhu et al. (2023) who constructed a minimal example whose limiting sharpness lies close to the stability threshold. Agarwala et al. (2023) showed that models quadratic in their parameters already exhibit progressive sharpening and EoS behavior. For logistic regression on separable data, Wu et al. (2023) established convergence and a max-margin implicit bias for arbitrary fixed step-sizes. Wu et al. (2024) quantified the optimization benefit of a large step-size and the subsequent transition to monotone descent, while Crawshaw & Liu (2026) obtained sharper transition-time bounds in two dimensions. Large-step acceleration has also been established for regularized logistic regression (Wu et al., 2025) and suitable non-homogeneous two-layer networks (Cai et al., 2024). Under a river-valley landscape assumption, Wen et al. (2025) separated oscillations across a valley from progress along it to explain warmup–stable–decay schedules.

Closer to our constructions in Section 4, Chen & Bruna (2023) study exact two-step periods in scalar linear networks, MacDonald et al. (2025) prove convergence at the critical step-size for a deep scalar network, Ahn et al. (2023) and Song & Yun (2023) prove convergence to a minimizer with sharpness close to $2/\eta$ for simple scalar networks. Chen et al. (2024) give also similar scalar constructions of residual alternation, without detailing how they might produce loss and sharpness oscillations. More recently, Liang & Montúfar (2026) and Meng et al. (2024) construct specific examples with general exact or approximate k-cycles, for any integer k. What seems to be missing from the literature is a general condition that guarantees persistent long-term optimization progress and provably coincides with loss and sharpness oscillations in specific constructions. Our work is the first to provide such condition and examples for general k.

Theory of EoS In contrast to classic optimization theory, which assumes that the step-size sits comfortably inside a stable region, EoS analyses address the interaction between finite steps and evolving curvature. Arora et al. (2022) derived a sharpness-reducing flow on a manifold of minimizers for normalized GD and GD on a transformed loss. Damian et al. (2023) identified a self-stabilization mechanism by which unstable oscillations induce a reduction in sharpness through third-order terms. Min et al. (2026) surveyed the connections between convergence, implicit bias, and EoS. A complementary line of work relaxes uniform smoothness through gradient-dependent curvature bounds (Zhang et al., 2020; Li et al., 2023; Tyurin, 2025) or loss-gap-dependent bounds that justify warm-up (Alimisis et al., 2026). Mishkin et al. (2024) used directional smoothness to obtain trajectory-dependent rates and adaptive step-sizes.

Litman (2026) derived the same triangularly weighted directional curvature used here and the corresponding telescoping loss identity. For a C^2 objective bounded below, his Theorem 4.1 shows that the squared-step-weighted curvature average converges to $2/\eta$ if the loss is finite above along the fixed-step GD trajectory and $\sum_t \|x_{t+1} - x_t\|^2 = \infty$.

Lee & Lee (2026) studied curvature along the executed update across full-batch optimizers, including momentum and adaptive methods. They normalized the curvature contribution by gradient–update alignment, giving a common loss-change boundary when this alignment is positive. Their normalization places the boundary at 1, corresponding to 2 in Appendix H. In their experiments, this diagnostic exhibits smaller optimizer-dependent offsets than maximum-eigenvalue diagnostics. For example, Adam's reported median is below, but close to, the realized boundary.

MacDonald et al. (2025) proved local convergence rates for codimension-one least-squares objectives, including deep scalar factorization. Their analysis uses maximum-eigenvalue sharpness on the solution manifold and relates its reduction to tangential GD dynamics. Under assumptions including analyticity, an invariant normal line, geodesic strong convexity of sharpness, and a nondegenerate flip bifurcation, they derive eventual linear convergence in a subcritical regime, nonmonotone power-law convergence at the critical step-size, and convergence to a nonstationary two-cycle in a supercritical regime. These results are local and restricted to codimension-one least squares; the paper does not establish the required solution-manifold geometry for practical neural-network training problems.

Our contribution is to turn the exact loss identity into convergence guarantees through a lower envelope γ_T for the gradient-energy-weighted cumulative margin. Theorem 1 gives best-iterate stationarity bounds and, under a trajectory-local PL inequality, best-iterate loss-gap bounds, while allowing individual loss-increasing updates. These best-iterate bounds vanish when $(T-s)\gamma_T \to \infty$. They do not require the solution-manifold geometry above, but are conditional on the margin envelope. Definition 1 and Theorem 2 give a sufficient route to such an envelope on complete k-step horizons.

Minibatch extensions. Previous work studies expected directional progress of a fixed full objective under SGD (Ahn et al., 2022), or extends exact averaged-curvature identities to stochastic updates of that objective (Litman, 2026). Batch sharpness instead averages the current-iterate directional curvature of minibatch losses over sampled batches (Andreyev & Beneventano, 2024). Our approach adapts the executed-update residual of Lee & Lee (2026) by evaluating the same minibatch before and after the actual optimizer update, retaining the exact finite-step loss change rather than a pointwise curvature approximation. We then separate this same-minibatch change from the change caused by replacing the minibatch at the updated parameters. The distinction is therefore between tracking a fixed objective along stochastic updates and accounting for the changing minibatch objectives in the logged loss sequence.

B Why the Largest Hessian Eigenvalue Is Not Enough for Explaining Convergence

EoS is commonly diagnosed through a worst-case spectral sharpness. For full-batch Euclidean GD, this is the largest eigenvalue of the current Hessian approaching the quadratic threshold $2/\eta$ (Cohen et al., 2021). Spectral sharpness and the top Hessian eigenspace record worst-case instantaneous curvature, but not how the gradient aligns with the spectrum. The projected updates of Song et al. (2025) make this distinction transparent: Dom-GD, which retains the top Hessian eigenspace, can stall, whereas Bulk-GD, which removes it, can keep reducing the loss. In the two examples of Figure 5, Bulk-GD also descends faster than full GD. We examine the finite-step curvature conditions behind these observations and their relation to sharpening at EoS.

Fix $d \geq 2$, $x \in \mathbb{R}^d$, and $k \in \{1, \ldots, d-1\}$. Let $u_i(x) \in \mathbb{R}^d$, $i \in \{1, \ldots, d\}$, form an orthonormal basis of Hessian eigenvectors corresponding to $\lambda_1(\nabla^2 f(x)) \geq \cdots \geq \lambda_d(\nabla^2 f(x))$, respectively. Set $g(x) := \nabla f(x) \in \mathbb{R}^d$ and define $P_R(x) \in \mathbb{R}^{d\times d}$ and $x_R^+ \in \mathbb{R}^d$ by

$$P_{\mathrm{dom}}(x) := \sum_{i=1}^{k} u_i(x)u_i(x)^\top, \qquad P_{\mathrm{bulk}}(x) := I - P_{\mathrm{dom}}(x),$$
$$x_R^+ := x - \eta P_R(x)g(x), \qquad R \in \{\mathrm{dom}, \mathrm{bulk}\}.$$

For $R \in \{\mathrm{dom}, \mathrm{bulk}\}$, the relevant projected gradient is $d_R := P_R(x)g(x) \in \mathbb{R}^d$. Then

$$\forall R \in \{\mathrm{dom}, \mathrm{bulk}\} : \qquad d_R \neq 0 \quad \Longrightarrow \quad \langle g(x), d_R\rangle = \|d_R\|^2 > 0.$$

For $d_R \neq 0$, let $\tilde{\lambda}_R(x) \in \mathbb{R}$ denote the averaged sharpness in the direction d_R.

Proposition 6. *For $R \in \{\mathrm{dom}, \mathrm{bulk}\}$ with $d_R \neq 0$, it holds*

$$f(x_R^+) - f(x) = -\eta \|d_R\|^2 \left(1 - \frac{\eta\tilde{\lambda}_R(x)}{2}\right). \tag{13}$$

If $d_{\mathrm{dom}} \neq 0$ and $d_{\mathrm{bulk}} \neq 0$, then

$$\left[f(x_{\mathrm{dom}}^+) \geq f(x) \text{ and } f(x_{\mathrm{bulk}}^+) < f(x)\right] \Longleftrightarrow \tilde{\lambda}_{\mathrm{dom}}(x) \geq \frac{2}{\eta} > \tilde{\lambda}_{\mathrm{bulk}}(x). \tag{14}$$

Proof of Proposition 6. Orthogonality of P_R gives $\langle g, d_R\rangle = \langle g, P_R g\rangle = \|d_R\|^2$. Taylor's theorem with integral remainder therefore yields

$$\begin{aligned} f(x_R^+) - f(x) &= -\eta\langle g, d_R\rangle + \eta^2 \int_0^1 (1-\tau) d_R^\top \nabla^2 f(x - \tau\eta d_R) d_R \, \mathrm{d}\tau \\ &= -\eta \|d_R\|^2 \left(1 - \frac{\eta\tilde{\lambda}_R(x)}{2}\right). \end{aligned}$$

Since $\eta \|d_R\|^2 > 0$ for $d_R \neq 0$, the signs of the two loss changes give (14). □

Thus, the deciding quantity is not the spectral maximum alone, but the step-averaged curvature in the update direction. This is the one-step counterpart of $M_{s:T}^g$: (13) determines the sign of each projected step, while (1) aggregates the same margins over the GD trajectory.

When $d_{\mathrm{bulk}} \neq 0$, compare Bulk-GD with a full GD step from the same point x. Let $\tilde{\lambda}_{\mathrm{full}}(x)$ be the averaged directional sharpness along g. The same loss identity gives

$$f(x_{\mathrm{bulk}}^+) < f(x - \eta g) \quad \Longleftrightarrow \quad \|d_{\mathrm{bulk}}\|^2 \left(1 - \frac{\eta\tilde{\lambda}_{\mathrm{bulk}}(x)}{2}\right) > \|g\|^2 \left(1 - \frac{\eta\tilde{\lambda}_{\mathrm{full}}(x)}{2}\right).$$

Lower bulk curvature alone does not ensure a larger decrease, because the bulk step also removes part of the gradient.

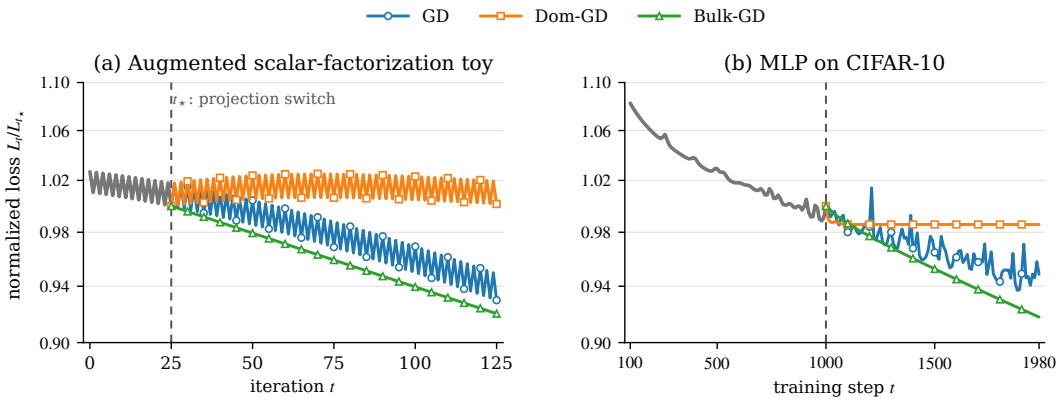


Figure 5: Theory–empirical correspondence for dom–bulk separation. **(a)** The scalar factorization model of of Proposition 4 (i), augmented by one stable bulk coordinate, $\Phi(w_1, w_2, s) = (1 - w_1 w_2)^2 + s^2/2000$, with $\eta = 0.501$ and $(w_{1,0}, w_{2,0}, s_0) = (1, 0.95, 5)$. Raw post-switch losses are shown at every step to expose the EoS two-cycle. **(b)** A representative full-batch GD run for a two-layer width-200 MLP on CIFAR-10 with MSE loss, $\eta = 0.05$, and a rank-10 dominant subspace, switched at $t_\star = 1000$. Each panel uses a shared GD prefix and normalizes loss by its value at the switch. In both panels, Dom-GD stalls while Bulk-GD descends smoothly and faster than the oscillatory full-GD trajectory.

Pointwise curvatures provide a limited connection to sharpening. For $\eta > 0$ and x with $d_{\mathrm{dom}}, d_{\mathrm{bulk}} \neq 0$, set $d_{\mathrm{full}} = g$ and define

$$\chi_R(\tau) := \frac{d_R^\top \nabla^2 f(x - \tau\eta d_R) d_R}{\|d_R\|^2}, \qquad R \in \{\mathrm{dom}, \mathrm{full}, \mathrm{bulk}\}, \quad \tau \in [0, 1].$$

Proposition 7. *Let* $\omega_R = \|d_R\|^2 / \|g\|^2$ *for* $R \in \{\mathrm{dom}, \mathrm{bulk}\}$. *Then*

$$\chi_{\mathrm{full}}(0) = \omega_{\mathrm{dom}} \chi_{\mathrm{dom}}(0) + \omega_{\mathrm{bulk}} \chi_{\mathrm{bulk}}(0).$$

If $\lambda_k(\nabla^2 f(x)) > \lambda_{k+1}(\nabla^2 f(x))$, *then* $\chi_{\mathrm{dom}}(0) > \chi_{\mathrm{full}}(0) > \chi_{\mathrm{bulk}}(0)$.

Proof. The spectral projectors have orthogonal ranges and commute with $\nabla^2 f(x)$. Hence $d_{\mathrm{dom}}^\top \nabla^2 f(x) d_{\mathrm{bulk}} = 0$. Expanding $g = d_{\mathrm{dom}} + d_{\mathrm{bulk}}$ proves the first identity, with positive weights summing to one. The Rayleigh quotient bounds give $\chi_{\mathrm{dom}}(0) \geq \lambda_k(\nabla^2 f(x)) > \lambda_{k+1}(\nabla^2 f(x)) \geq \chi_{\mathrm{bulk}}(0)$. The full directional sharpness, being a strict convex combination, lies strictly between the projected values. □

If training sharpens the leading modes, this pointwise identity shows how their curvature enters the full-GD direction while the bulk direction can remain less curved. The three updates follow different segments, however, so ordering their curvatures at $\tau = 0$ does not establish the finite-step loss comparisons above. Likewise, when $\eta L \geq 2$, an L-smoothness bound gives only $\tilde{\lambda}_R(x) \leq L$ for each direction and does not determine the signs or relative sizes of their loss changes.

C PROOF OF THEOREM 1

We provide the proof of Theorem 1 here. The proof is simple and is based on classic techniques from convex optimization. We just avoid to create bounds based on smoothness-type conditions.

Proof of Theorem 1. If the accumulated gradient energy on $[s, T)$ is zero, then the non-convex rate is obtained immediately. Otherwise, equation (1) and the assumption $M^g_{s:T} \geq \gamma_T$ give

$$\eta\gamma_T \sum_{t=s}^{T-1} \|g_t\|^2 \leq f(x_s) - f(x_T) \leq f(x_s) - f_\star = \Delta_s.$$

Dividing by $\eta N \gamma_T$ proves a squared-gradient average bound and the first claim follows since the minimum squared gradient is smaller than the average.

Under PL inequality, we can use the previous bound and get

$$\frac{1}{N} \sum_{t=s}^{T-1} \Delta_t \leq \frac{1}{2\mu N} \sum_{t=s}^{T-1} \|g_t\|^2 \leq \frac{\Delta_s}{2\eta\mu N \gamma_T}.$$

This includes the case where gradients are all zero, since then the PL inequality implies $\Delta_t = 0$ on the window. $\min_{s \leq t \leq T} \Delta_t$ is upper bounded by the average of Δ_t and the second claim follows.

□

D PROOFS FOR SECTION 3

Proof of Theorem 2. For ease of notation, we denote $A_t := \|g_t\|^2$. Let $T - s = nk$, where $n \geq 1$, and define

$$E_j := B_k(x_{s+jk}) = \sum_{r=0}^{k-1} A_{s+jk+r}, \quad j \geq 0.$$

By (2) and (3), we get an inequality that quantifies function value progress between iterations that are k steps apart:

$$f(x_{s+jk}) - f(x_{s+(j+1)k}) \geq \eta c E_j^{m+1}. \tag{15}$$

We now define $P_n := \frac{f(x_s)-f(x_T)}{\eta}$. and telescoping (15) gives

$$P_n \geq c \sum_{j=0}^{n-1} E_j^{m+1} \geq c E_0^{m+1}. \tag{16}$$

It holds $\sum_{t=s}^{T-1} A_t = \sum_{j=0}^{n-1} E_j$ since the first sum sums all gradients from s to $T-1$, while the second sums first inside each block and then across blocks. Hölder's inequality then yields

$$\sum_{t=s}^{T-1} A_t \leq n^{m/(m+1)} \left(\sum_{j=0}^{n-1} E_j^{m+1} \right)^{1/(m+1)} \leq n^{m/(m+1)} (P_n/c)^{1/(m+1)}. \tag{17}$$

Using the exact definition of $M_{s:T}^g$ and equations (16) and (17), we obtain

$$M_{s:T}^g = \frac{P_n}{\sum_{t=s}^{T-1} A_t} \geq c^{1/(m+1)} P_n^{m/(m+1)} n^{-m/(m+1)} \geq c E_0^m n^{-m/(m+1)}.$$

Thus the explicit finite-window estimate is

$$M_{s:T}^g \geq c\, B_k(x_s)^m \left(\frac{k}{T-s} \right)^{m/(m+1)} \geq c\, \|g_s\|^{2m} \left(\frac{k}{T-s} \right)^{m/(m+1)}.$$

This proves (4).

The stronger bounds for $k = 2$. Suppose now that $k = 2$ and $\eta L \geq 1$, and write $k_t := g_t + g_{t+1} = S_2(x_t)$. Since $g_{t+1} = k_t - g_t$, we have

$$A_t - A_{t+1} = 2\langle g_t, k_t \rangle - \|k_t\|^2 = 2\mathcal{C}_L^{(2)}(x_t) + (\eta L - 1)\|k_t\|^2 \geq 2\mathcal{C}_L^{(2)}(x_t).$$

Since $\mathcal{C}_L^{(2)}(x_t) \geq c B_2(x_t)^{m+1} \geq c A_t^{m+1}$, we have

$$A_{t+1} \leq A_t - 2c A_t^{m+1}.$$

If $A_{t+1} > 0$, then $2cA_t^m < 1$. The elementary inequality $(1-z)^{-m} \geq 1 + mz$ for $0 \leq z < 1$ yields

$$A_{t+1}^{-m} \geq A_t^{-m}(1 - 2cA_t^m)^{-m} \geq A_t^{-m} + 2mc.$$

Summing these increments and taking the negative $1/m$ power gives

$$A_{s+j} \leq \frac{A_s}{(1 + 2mcA_s^m j)^{1/m}}, \qquad j \geq 0.$$

If a gradient becomes zero, all subsequent gradients vanish, thus the same upper bound remains valid.

Let $T - s \geq 2$ be even. The function $u \mapsto A_s/(1 + 2mcA_s^m u)^{1/m}$ is decreasing. By bounding the sum of A_t by an integral, we obtain the bound

$$\sum_{t=s}^{T-1} A_t \leq A_s + \int_0^{T-s-1} \frac{A_s\, du}{(1 + 2mcA_s^m u)^{1/m}}$$

$$= \begin{cases} A_s + \dfrac{1}{2c}\log\big(1 + 2cA_s(T-s-1)\big), & m = 1, \\ A_s + \dfrac{A_s^{1-m}}{2c(m-1)}\left(\big(1 + 2mcA_s^m(T-s-1)\big)^{(m-1)/m} - 1\right), & m > 1. \end{cases}$$
$$:= D_{T-s}.$$

By (2) and (3), we get

$$f(x_s) - f(x_T) \geq \eta c \sum_{j=0}^{(T-s)/2-1} B_2(x_{s+2j})^{m+1} \geq \eta c A_s^{m+1}.$$

Consequently,

$$M_{s:T}^g \geq \frac{cA_s^{m+1}}{D_{T-s}}.$$

For fixed s, the numerator is strictly positive and independent of N. The denominator is $O(\log(2 + T - s))$ when $m = 1$ and $O((T-s)^{(m-1)/m})$ when $m > 1$. This proves (5) and the proof is complete.

□

We now move to the result concerning semi-algebraic functions. We provide here the rigorous definitions of a semi-algebraic set and a semi-algebraic function, which are omitted in the main text.

Definition 2.

- ***Semi-algebraic set:*** *A subset $S \subset \mathbb{R}^n$ is called a* semi-algebraic set *if it can be written as a finite union of basic semi-algebraic sets of the form*

$$\{x \in \mathbb{R}^n : P_1(x) = \cdots = P_a(x) = 0, \quad Q_1(x) > 0, \ldots, Q_b(x) > 0\},$$

where $P_1, \ldots, P_a, Q_1, \ldots, Q_b$ are polynomials in $\mathbb{R}[X_1, \ldots, X_n]$.

- ***Semi-algebraic function:*** *Let $S \subset \mathbb{R}^n$ be a semi-algebraic set. A function $f : S \to \mathbb{R}$ is called a* semi-algebraic function *if its graph*

$$\Gamma(f) := \{(x, y) \in \mathbb{R}^{n+1} : x \in S,\ y = f(x)\}$$

is a semi-algebraic set in $\mathbb{R}^{n+1}$.

We are ready to present the proof of Proposition 3.

Proof of Proposition 3. We use the following comparison theorem for two functions: if Q is compact and semi-algebraic, $u, v : Q \to \mathbb{R}$ are continuous semi-algebraic functions, and $u^{-1}(0) \subseteq v^{-1}(0)$, then there are $C > 0$ and an integer $p_0 \geq 1$ such that

$$|v(x)|^{p_0} \leq C|u(x)|.$$

For a version of this result please look at (Basu & Mohammad-Nezhad, 2024, Theorem 1.1, equation (1.1)).

A C^2 semi-algebraic function has continuous semi-algebraic first derivatives. Composition, sums, and products preserve semi-algebraicity. Thus, we apply this result with

$$u(x) := \mathcal{C}_L^{(k)}(x), \qquad v(x) := B_k(x) = \sum_{r=0}^{k-1} \|g(G^r(x))\|^2,$$

which are continuous semi-algebraic functions on Ω. All compositions are defined by the domain assumptions. If $g(x) = 0$, then $G^r(x) = x$ for every r, so $S_k(x) = 0$, $B_k(x) = 0$, and $\mathcal{C}_L^{(k)}(x) = 0$. If $g(x) \neq 0$, the positivity hypothesis gives $\mathcal{C}_L^{(k)}(x) > 0$, while $B_k(x) \geq \|g(x)\|^2 > 0$. Therefore $\mathcal{C}_L^{(k)}(x)$ vanishes exactly when $g(x)$ vanishes. The comparison theorem consequently gives

$$B_k(x)^{p_0} \leq C\mathcal{C}_L^{(k)}(x), \quad x \in \Omega.$$

Since we are in a compact set, the exponent p can be taken $p \geq 1$. To see this, set

$$p := \max\{2, p_0\}, \qquad R := \max\left\{1, \max_{x\in\Omega} B_k(x)\right\}.$$

Compactness gives $R < \infty$, and

$$B_k(x)^p \leq R^{p-p_0} B_k(x)^{p_0} \leq C R^{p-p_0} \mathcal{C}_L^{(k)}(x).$$

Taking $c := (CR^{p-p_0})^{-1}$ proves (6), hence condition (C_m^k) with $m = p - 1$.

□

E Comparison of (C_m^k) with existent local EoS conditions

We compare condition (3) with landscape assumptions from five relevant papers. The hypotheses cited below belong to those papers, but the implications for our condition are derived by us. We isolate the assumptions needed for these implications, rather than restating the papers' complete network-convergence theorems.

Throughout this section, $k = 2$ and

$$G(x) = x - \eta g(x), \qquad S_2(x) = g(x) + g(G(x)), \qquad B_2(x) = \|g(x)\|^2 + \|g(G(x))\|^2.$$

We use the original Euclidean gradient $g(x) = \nabla f(x)$ and

$$\mathcal{C}_L^{(2)}(x) = \langle g(x), S_2(x)\rangle - \frac{\eta L}{2}\|S_2(x)\|^2.$$

Every step-size below is chosen to be critical at the accumulation point of the trajectory. A positive limit of $\mathcal{C}_L^{(2)}/B_2^{m+1}$ implies a uniform (C_m^2) inequality on an asymptotically small set. The size of this set cannot be specified by the limit alone.

We start with a useful lemma.

Lemma 8. *Let f be even and C^2 near zero, with $h := f''(0) > 0$. Fix $\eta = 2/h$ and $L > h$, and set*

$$\delta(z) := 1 - \frac{f'(z)}{hz}, \text{ if } z \neq 0, \quad \delta(0) := 0.$$

On a sufficiently small interval, it holds

$$\mathcal{C}_L^{(2)}(z) > 0 \quad \Longleftrightarrow \quad \delta(z) > 0. \tag{18}$$

For every fixed $\nu > 0$, we also have

$$\delta(z) \geq a|z|^\nu \quad \Longleftrightarrow \quad \mathcal{C}_L^{(2)}(z) \geq cB_2(z)^{1+\nu/2} \tag{19}$$

locally. If $\delta(z) = az^{2m} + o(|z|^{2m})$, where $a > 0$ and $m \geq 1$, then

$$\frac{\mathcal{C}_L^{(2)}(z)}{B_2(z)^{m+1}} \longrightarrow \frac{a}{2^m h^{2m}}. \tag{20}$$

If $\delta(z) > 0$ for all sufficiently small nonzero z, a sufficiently small closed interval around zero is GD invariant.

Proof. Define $a_z := 1 - 2\delta(z)$, so that $G(z) = -a_z z$ and take $a_z \to 1$ to capture an arbitrarily small movement. Thus, $\delta(z) \to 0$ and $z \to 0$. Since f' is odd, the fundamental theorem of calculus gives

$$S_2(z) = f'(z) - f'(a_z z) = 2hz\delta(z)\mu(z),$$

where

$$\mu(z) := \frac{1}{h}\int_0^1 f''\big((a_z + t(1-a_z))z\big)\, dt.$$

by the mean value theorem for f'. This identity also holds when $\delta(z) = 0$. When $a_z \to 1$, we have $\mu(z) \to 1$. Consequently,

$$\mathcal{C}_L^{(2)}(z) = 2h^2 z^2 \delta(z)\left[(1-\delta(z))\mu(z) - \frac{2L}{h}\delta(z)\mu(z)^2\right].$$

The expression in brackets tends to 1. Also,

$$B_2(z) = 2h^2 z^2 + o(z^2).$$

After shrinking the interval, the bracket is bounded above and below by positive constants, and $B_2(z)$ is comparable to z^2. These estimates prove (18) and both implications in (19). They also give the limit (20). If $\delta(z) > 0$, shrink further so that $0 < \delta(z) < 1/2$. Then $|G(z)| = (1 - 2\delta(z))|z| < |z|$, proving interval invariance. Finally, continuity of f'' supplies the ambient bound $|f''| \leq L$ on a slightly larger open interval containing all required segments. □

E.1 AHN ET AL. (2023)

Assumptions (A1)–(A2) of Ahn et al. (2023) require an even, convex, 1-Lipschitz scalar loss ℓ, with $\ell \in C^2$ near zero and $\ell''(0) = 1$. The finite-exponent part of (A2) implies that there exist $a, \rho > 0$ and $\nu > 1$ such that

$$1 - \frac{\ell'(z)}{z} \geq a|z|^\nu, \qquad 0 < |z| \leq \rho. \tag{21}$$

We have renamed the exponent to avoid confusion with other constants. Only the local properties displayed here are needed below.

Proposition 9. *For the scalar objective $f = \ell$, $\eta = 2$, and any fixed $L > 1$,* (21) *implies locally*

$$\mathcal{C}_L^{(2)}(z) \geq cB_2(z)^{1+\nu/2}.$$

In particular, if $\nu = 2m$, then (C_m^2) holds on a sufficiently small closed invariant interval. For general finite $\nu > 1$, the integer exponent $p := \lceil 1 + \nu/2 \rceil$ gives C_{p-1}.

Proof. Apply Lemma 8 with $h = 1$. If $p > 1 + \nu/2$, shrink the interval so that $B_2 \leq 1$. Then $B_2^{1+\nu/2} \geq B_2^p$. The same lemma also supplies invariance. □

E.2 MA ET AL. (2022)

The scalar subquadratic setting of (Ma et al., 2022, Section 4.1 and Theorem 1) uses an even C^2 loss whose curvature strictly decreases with distance from the minimum. The local hypothesis needed for our comparison is

$$h := f''(0) > 0, \qquad f''(s) < f''(t) \quad \text{whenever } 0 \leq |t| < |s| < \rho. \tag{22}$$

Their assumptions at large distances are not needed for this local implication.

Proposition 10. *Under* (22)*, for $\eta = 2/h$ and any fixed $L > h$, the certificate $\mathcal{C}_L^{(2)}$ is strictly positive at every sufficiently small nonzero point. If f is also semi-algebraic near zero, then (C_m^2) holds for some integer $m \geq 1$ on a sufficiently small closed invariant interval.*

Proof. For $z > 0$ sufficiently small, we can define

$$\delta(z) = \frac{1}{hz} \int_0^z \big(h - f''(u)\big)\, du > 0.$$

Evenness gives the same inequality for $z < 0$. Lemma 8 proves positivity and invariance. On a sufficiently small closed interval, semi-algebraicity and Proposition 3 give $\mathcal{C}_L^{(2)} \geq cB_2^p$ for some integer $p \geq 2$, thus we can take $m = p - 1$. □

Strict curvature decrease alone does not specify a finite power. For example, a quantitative strengthening $h - f''(z) \geq a|z|^{2m}$ gives $\delta(z) \geq a|z|^{2m}/(h(2m+1))$ and hence (C_m^2) directly, without semi-algebraicity.

E.3 SONG & YUN (2023)

We denote by ϱ the ratio called r in (Song & Yun, 2023, Assumption 2.4). Its basic hypotheses are

$$\begin{gathered} \varrho \in C^1(\mathbb{R}), \qquad \varrho \text{ even}, \qquad \varrho(0) = 1, \quad \varrho'(0) = 0, \\ \varrho'(z) < 0 \quad (z > 0), \qquad \lim_{|z| \to \infty} \varrho(z) = 0. \end{gathered} \tag{23}$$

Their Assumptions 4.2 and C.1 use, respectively,

$$\varrho(z) = \frac{\ell'(z)}{z}, \qquad \varrho(z) = \frac{\phi(z)\phi'(z)}{z},$$

corresponding here to the scalar objectives $f = \ell$ and $f = \frac{1}{2}\phi^2$. Their stronger EoS assumptions also include $\varrho \in C^4$ and

$$z \longmapsto \frac{z\varrho(z)}{\varrho'(z)} \quad \text{nonincreasing on } z > 0. \tag{24}$$

Only this part of the stronger assumptions is needed below.

Proposition 11. *For either scalar objective above, assume it is C^2 near zero and $f'(z) = z\varrho(z)$ where ϱ satisfies the assumption of* (23)*. Then, at $\eta = 2$, for every fixed $L > 1$, the certificate is strictly positive near zero away from the minimum. If f is semi-algebraic there, it satisfies (C_m^2) for some integer $m \geq 1$ on a sufficiently small closed invariant interval. If, instead, ϱ is C^2 near zero and satisfies* (24)*, then $\varrho''(0) < 0$ and*

$$\frac{\mathcal{C}_L^{(2)}(z)}{B_2(z)^2} \longrightarrow -\frac{\varrho''(0)}{4} > 0. \tag{25}$$

In this stronger case, C_1 follows without semi-algebraicity.

Proof. The derivative $f'(z) = z\varrho(z)$ is odd, so f is even, and $f''(0) = 1$. Condition (23) gives $0 < \varrho(z) < 1$ for $z \neq 0$. Thus $\delta(z) = 1 - \varrho(z) > 0$, and Lemma 8, followed when applicable by Proposition 3, proves the first claims.

For the stronger claim, fix a sufficiently small $b > 0$. Since $\varrho > 0$ and $\varrho' < 0$ on $(0, b]$, (24) implies

$$-\frac{\varrho'(z)}{z\varrho(z)} \geq -\frac{\varrho'(b)}{b\varrho(b)} > 0, \qquad 0 < z < b.$$

Taking $z \to 0$ gives $-\varrho''(0) \geq -\varrho'(b)/(b\varrho(b)) > 0$. Taylor's theorem now yields

$$\delta(z) = -\frac{1}{2}\varrho''(0)z^2 + o(z^2),$$

and (25) follows from (20). □

E.4 Chen & Bruna (2023)

Let $x_\star$ be a scalar minimum, $h := f''(x_\star) > 0$, and set

$$D := 3f'''(x_\star)^2 - hf''''(x_\star) > 0. \tag{26}$$

This is the derivative sign condition in (Chen & Bruna, 2023, Theorem 1 and Appendix C). Their cycle-existence statement also assumes $f'''(x_\star) \neq 0$; the following critical-step implication does not require that extra assumption.

Proposition 12. *Suppose $f \in C^4$ near $x_\star$ and $f'(x_\star) = 0$. At $\eta = 2/h$, for any fixed $L > h$,*

$$\frac{\mathcal{C}_L^{(2)}(x_\star + z)}{B_2(x_\star + z)^2} \longrightarrow \frac{D}{12h^4}, \quad \textit{as} \quad z \to 0. \tag{27}$$

Consequently, $D > 0$ is equivalent to a local (C_1^2) condition with a positive constant.

For $m \geq 2$, suppose instead that $f \in C^{2m+2}$ near $x_\star$ and

$$f^{(j)}(x_\star) = 0 \quad (3 \leq j \leq 2m+1), \qquad d_m := f^{(2m+2)}(x_\star) < 0. \tag{28}$$

Then, at the same critical step size,

$$\frac{\mathcal{C}_L^{(2)}(x_\star + z)}{B_2(x_\star + z)^{m+1}} \longrightarrow -\frac{d_m}{2^m(2m+1)!h^{2m+1}} > 0. \tag{29}$$

Thus (C_m^2) holds locally in this higher-order case.

Proof. Put $a = f'''(x_\star)$, $b = f''''(x_\star)$, and write $\Phi(z) := G(x_\star + z) - x_\star$. At $\eta = 2/h$, a Taylor expansions gives

$$\Phi(z) = -z - \frac{a}{h}z^2 - \frac{b}{3h}z^3 + o(|z|^3).$$

Composing gives

$$\Phi^2(z) = z - \frac{2D}{3h^2}z^3 + o(|z|^3), \qquad S_2(x_\star + z) = \frac{D}{3h}z^3 + o(|z|^3).$$

Since $g(x_\star + z) = hz + O(z^2)$,

$$\mathcal{C}_L^{(2)}(x_\star + z) = \frac{D}{3}z^4 + o(z^4), \qquad B_2(x_\star + z) = 2h^2 z^2 + o(z^2).$$

This proves (27). A positive limit gives a uniform positive lower bound after shrinking. Conversely, a local inequality $\mathcal{C}_L^{(2)} \geq cB_2^2$, $c > 0$, forces the limit to be at least c, and hence $D > 0$.

Under (28), Taylor expansion instead gives

$$\Phi(z) = -z - \frac{2d_m}{h(2m+1)!}z^{2m+1} + o(|z|^{2m+1}).$$

Because $2m + 1$ is odd,

$$\Phi^2(z) = z + \frac{4d_m}{h(2m+1)!}z^{2m+1} + o(|z|^{2m+1}).$$

Consequently,

$$S_2(x_\star + z) = -\frac{2d_m}{(2m+1)!}z^{2m+1} + o(|z|^{2m+1}),$$
$$\mathcal{C}_L^{(2)}(x_\star + z) = -\frac{2hd_m}{(2m+1)!}z^{2m+2} + o(|z|^{2m+2}).$$

The term involving $L\|S_2\|^2$ has strictly higher order. Dividing by $B_2^{m+1} = (2h^2)^{m+1}z^{2m+2} + o(|z|^{2m+2})$ proves (29). □

Condition (28) is the even-order branch of (Chen & Bruna, 2023, Lemma 1 and Appendix D), not a classification of every case $D = 0$. Their stated cycle results choose a supercritical step size. At an exact nonstationary two-cycle, $G^2(x) = x$ gives $S_2(x) = 0$ and hence $\mathcal{C}_L^{(2)}(x) = 0$; the strict certificate above concerns the critical-step damping instead.

E.5 MULAYOFF & STICH (2026)

Let $f \in C^4$ near a stationary point $x_\star \in \mathbb{R}^d$. Suppose $H := \nabla^2 f(x_\star) \succ 0$ has a simple largest eigenvalue λ, with unit eigenvector $v \in \mathbb{R}^d$. Define $T \in \mathbb{R}^d$ and $Q \in \mathbb{R}$ by

$$\langle T, w\rangle = \nabla^3 f(x_\star)[v, v, w] \quad (w \in \mathbb{R}^d), \qquad Q = \nabla^4 f(x_\star)[v, v, v, v].$$

The coefficient condition of (Mulayoff & Stich, 2026, Theorem 7 and Appendix IV) is

$$\mathfrak{D} := 3T^\top H^{-1}T - Q > 0. \tag{30}$$

Proposition 13. *At $\eta = 2/\lambda$, a local critical invariant curve can be parameterized as $\psi(Z)$, where Z is an interval containing zero and*

$$\psi(z) = x_\star + zv + O(z^2), \qquad G(\psi(z)) = \psi(\varphi(z)).$$

For every fixed $L > \lambda$,

$$\frac{\mathcal{C}_L^{(2)}(\psi(z))}{B_2(\psi(z))^2} \longrightarrow \frac{\mathfrak{D}}{12\lambda^3}, \quad \textit{as} \quad z \to 0. \tag{31}$$

Thus (30) *implies C_1 on a sufficiently small patch of this invariant curve.*

Proof. Translate $x_\star$ to zero and put $P := I - vv^\top$. The Jacobian $I - \eta H$ has eigenvalue -1 along v and all other eigenvalues strictly inside the unit disk. The local center-manifold reduction underlying (Mulayoff & Stich, 2026, Appendix IV) supplies a C^3 invariant curve, parameterized so that $v^\top\psi(z) = z$. Write

$$\psi(z) = zv + z^2 w + O(z^3), \qquad w \perp v, \qquad \varphi(z) = -z + az^2 + bz^3 + o(|z|^3).$$

The transverse quadratic terms in $G(\psi(z)) = \psi(\varphi(z))$ give

$$(I - \eta H)w - \frac{\eta}{2}PT = w, \qquad w = -\frac{1}{2}H^{-1}PT.$$

The terms in the v direction give

$$a = -\frac{\eta}{2}\langle T, v\rangle, \qquad b = -\eta\left(\langle T, w\rangle + \frac{Q}{6}\right).$$

Using $\eta = 2/\lambda$, $Hv = \lambda v$, and $PH = HP$,

$$a^2 + b = \frac{\eta^2}{4}\langle T, v\rangle^2 + \frac{\eta}{2}(PT)^\top H^{-1}(PT) - \frac{\eta Q}{6} = \frac{\eta}{6}(3T^\top H^{-1}T - Q) = \frac{\eta}{6}\mathfrak{D}.$$

Thus

$$\varphi^2(z) = z - 2(a^2 + b)z^3 + o(|z|^3) = z - \frac{\eta\mathfrak{D}}{3}z^3 + o(|z|^3).$$

Returning to the original parameter space,

$$S_2(\psi(z)) = \frac{\psi(z) - \psi(\varphi^2(z))}{\eta} = \frac{\mathfrak{D}}{3}z^3 v + o(|z|^3),$$
$$g(\psi(z)) = \lambda z v + O(z^2), \qquad B_2(\psi(z)) = 2\lambda^2 z^2 + o(z^2).$$

Consequently,

$$\mathcal{C}_L^{(2)}(\psi(z)) = \frac{\lambda\mathfrak{D}}{3}z^4 + o(z^4),$$

which proves the claim. The gradients and displacements in these calculations are Euclidean quantities for f in the original coordinates; we have not replaced them by GD on $f \circ \psi$. □

The curve restriction is essential. Take for example, $\eta = 2$ and

$$f(u, w) = \frac{1}{2}u^2 - \frac{1}{4}u^4 + \frac{1}{6}u^6 + \frac{1}{8}w^2.$$

This function has $H = \mathrm{diag}(1, 1/4)$ and $\mathfrak{D} = 6 > 0$ at zero, but direct calculation gives

$$\mathcal{C}_L^{(2)}(0, w) = \frac{6 - 9L}{64}w^2 < 0 \qquad (w \neq 0,\ L \geq 1).$$

Every local ambient Hessian bound must have $L \geq 1$. Thus a positive multivariate flip coefficient does not imply our condition throughout a whole neighborhood.

F Proofs for Section 4

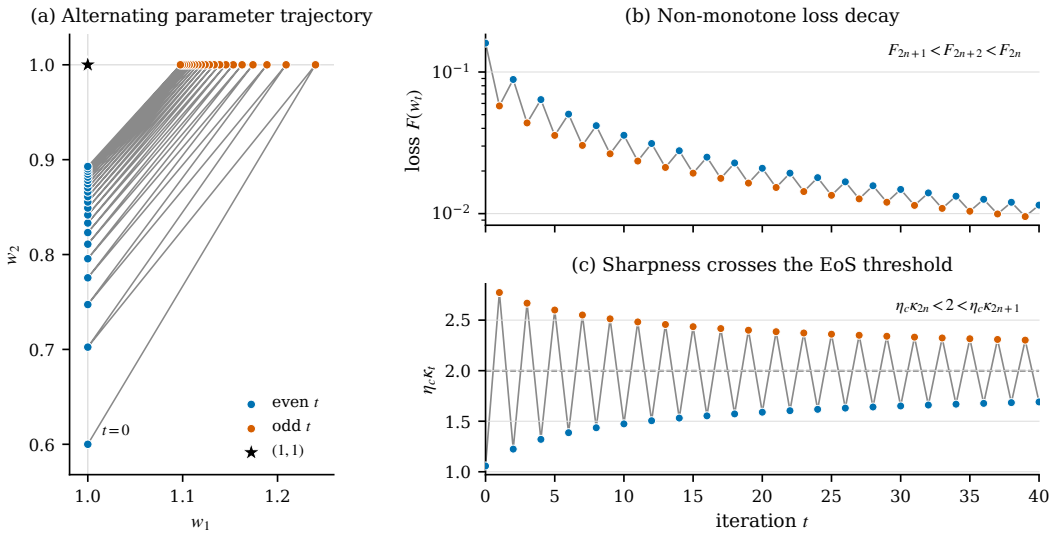


Figure 6: Scalar factorization (case (i) of Proposition 4, with $x = y = 1$, $h = 0.4$, and $\eta_c = 0.5$, shown at every step through $t = 40$. **Left:** the iterates alternate between $w_1 = 1$ and $w_2 = 1$ and approach $(1, 1)$. **Top right:** loss interlacing $F_{2n+1} < F_{2n+2} < F_{2n}$. **Bottom right:** normalized spectral sharpness alternates across the EoS threshold 2.

F.1 Proof of Proposition 4

Proof.

Throughout the proof, we fix

$$\delta := \frac{1}{8}, \qquad h_0 := \frac{1}{100}.$$

All estimates below hold on these fixed intervals.

Normalization and scalar dynamics For $m = 1$, put $z := (u, v) := w/\sqrt{\rho}$ and $f(z) = \frac{1}{2}(uv - 1)^2$. Then

$$F(w) = 2y^2 f(z), \qquad \nabla F(w) = \frac{2y^2}{\sqrt{\rho}} \nabla f(z), \qquad \eta \nabla^2 F(w) = \nabla^2 f(z),$$

and the normalized GD step-size is 1. For $m = 2, 3$, we simply use $z = w$ and $f = F$. We write $r = uv - 1$ for the residual and denote the three functions of the proposition as $f(z) = \phi_m(r)$, where

$$\phi_1(r) = \frac{1}{2} r^2, \qquad \phi_2(r) = \log \cosh(r + r^3/3), \qquad \phi_3(r) = \frac{1}{2} r^2 + \frac{1}{4} r^4.$$

Their derivatives with respect to r are

$$q_1(r) := r, \qquad q_2(r) := (1 + r^2) \tanh(r + r^3/3), \qquad q_3(r) := r + r^3. \tag{32}$$

In all three cases, we have

$$\begin{aligned} \nabla f(u, v) &= q_m(r) \begin{pmatrix} v \\ u \end{pmatrix}, \\ \nabla^2 f(u, v) &= q_m'(r) \begin{pmatrix} v^2 & uv \\ uv & u^2 \end{pmatrix} + q_m(r) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \end{aligned} \tag{33}$$

It holds

$$q_1'(r) = 1, \quad q_2'(r) = 2r \tanh(r + r^3/3) + (1 + r^2)^2 \operatorname{sech}^2(r + r^3/3), \quad q_3'(r) = 1 + 3r^2$$

. Moreover, for $|r| \le \delta$, we can bound q_2' and q_3' as

$$1 \le q_2'(r), q_3'(r) \le 1 + 5r^2.$$

For q_3' the bounds are immediate, while for q_2' the are obtained by elementary bounds on $\tanh$ and the resitriction $|r| \le 1/8$.

As mentioned, in the normalized variables, GD has step-size equal to 1. Thus, writing $z = (u, v)$ and $z^+ = (u^+, v^+)$ for the next iterate, the gradient formula above gives

$$u^+ = u - q_m(r)v, \quad v^+ = v - q_m(r)u.$$

We now restrict these coordinate updates to the sets containing the initializations we use in the proposition statement.

For $m = 1$, we have $q_1(r) = r$, so

$$u^+ = u - rv, \quad v^+ = v - ru.$$

Suppose first that $u = 1$. Since $r = uv - 1$, necessarily $v = 1 + r$. Substituting into the update yields

$$u^+ = 1 - r(1 + r), \quad v^+ = (1 + r) - r = 1.$$

Thus GD maps the line $u = 1$ into the line $v = 1$, and the new residual is

$$r^+ = u^+v^+ - 1 = \big(1 - r(1 + r)\big) \cdot 1 - 1 = -r(1 + r).$$

Conversely, if $v = 1$, then $u = 1 + r$, and

$$u^+ = (1 + r) - r = 1, \quad v^+ = 1 - r(1 + r).$$

Hence GD maps the line $v = 1$ into the line $u = 1$, with exactly the same residual update

$$r^+ = u^+v^+ - 1 = -r(1 + r).$$

The normalized initialization is $(u_0, v_0) = (1, 1 - h)$. Consequently, by induction, every even iterate lies on $u = 1$ and every odd iterate lies on $v = 1$. Thus, the residual satisfies the scalar recurrence

$$r_{t+1} = R_1(r_t), \quad R_1(r) := -r(1 + r), \quad r_0 = -h.$$

Now, applying the formula twice gives the exact two-step residual map

$$T_1(r) := R_1(R_1(r)) = -\big[-r(1 + r)\big]\big[1 - r(1 + r)\big] = r(1 + r)(1 - r - r^2) = r - 2r^3 - r^4.$$

For $m = 2, 3$, the initialization is balanced

$$(u_0, v_0) = \sqrt{1 - h}(1, 1)$$

and at any balanced point of the form $z = s(1, 1)$, we have

$$u = v = s, \quad r = s^2 - 1, \quad \nabla f(s, s) = sq_m(r) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

The two coordinates therefore have identical updates

$$u^+ = v^+ = s - sq_m(r) = s(1 - q_m(r)).$$

Thus, GD iterates maintain the balanced condition $u = v$. Writing the next iterate as $z^+ = s^+(1, 1)$, its common coordinate satisfy $s^+ = s(1 - q_m(r))$ and thus the residual satisfies

$$r^+ = (s^+)^2 - 1 = s^2(1 - q_m(r))^2 - 1 = (1 + r)(1 - q_m(r))^2 - 1.$$

Because the next point is again balanced, the same scalar formula applies at the following iteration. Hence, defining the one-step and two-step residual maps by

$$\begin{aligned} s^+ &= s(1 - q_m(r)), \\ R_m(r) &:= (1 + r)(1 - q_m(r))^2 - 1, \\ T_m(r) &:= R_m(R_m(r)) = (1 + R_m(r))(1 - q_m(R_m(r)))^2 - 1, \end{aligned} \tag{34}$$

we obtain, along each balanced trajectory, the recurrence

$$r_{t+1} = R_m(r_t), \qquad r_{t+2} = T_m(r_t), \qquad r_0 = (\sqrt{1 - h})^2 - 1 = -h.$$

Estimates for the two-step residual maps Recall that, for $m = 2, 3$,

$$R_m(r) = (1 + r)(1 - q_m(r))^2 - 1, \quad T_m = R_m \circ R_m,$$

and that our fixed radii are $\delta = 1/8$ and $h_0 = 1/100$. We now try to estimate the net residual change after two steps, namely $r - T_m(r)$, in all three cases. With $A_2 := \frac{4}{5}$, $A_3 := 20$, our goal is to show

$$\frac{3A_m}{4} \le \frac{r - T_m(r)}{r^{2m+1}} \le \frac{5A_m}{4}, \qquad 0 < |r| \le h_0, \quad m = 2, 3. \tag{35}$$

Besides quantifying the two-step displacement, this will imply that $T_m(r)$ has the sign of r and a strictly smaller absolute value. To obtain this estimate, we first construct polynomial models of R_m, then compose those polynomials, and finally control the error between the polynomial compositions and the actual maps T_m. Only the case $m = 2$ requires an approximation for $\tanh$, since $q_3(r) = r + r^3$ is already a polynomial.

To compute T_2 through degree five, we first need the terms of $q_2(r) = (1+r^2)\tanh(r+r^3/3)$ through degree five, together with a bound on the discarded terms. This motivates using the fifth-degree Taylor approximation of $\tanh$ at 0, which we denote by p:

$$p(z) := z - \frac{1}{3}z^3 + \frac{2}{15}z^5.$$

It still remains to bound the error of this Taylor approximation $E(z) := p(z) - \tanh z$ when $z \in [0, 1/2]$.

$E(z)$ satisfies

$$E'(z) + (p(z) + \tanh z)E(z) = p'(z) + p(z)^2 - 1, \qquad E(0) = 0.$$

and after integration, we get

$$E(z) = \int_0^z \left[p'(t) + p(t)^2 - 1\right] \exp\left(-\int_t^z [p(s) + \tanh s]\, ds\right) dt.$$

Since $p(s) \ge 0$ and $\tanh s \ge 0$ on $[0, 1/2]$, the exponential factor belongs to $(0, 1]$. It also holds

$$p'(z) + p(z)^2 - 1 = \frac{17}{45}z^6 - \frac{4}{45}z^8 + \frac{4}{225}z^{10} \in \left[0, \frac{17}{45}z^6\right].$$

by direct calculation.

Hence

$$0 \le E(z) \le \frac{17}{45}\int_0^z t^6\, dt = \frac{17}{315}z^7.$$

Both p and $\tanh$ are odd, so

$$|\tanh z - p(z)| \le \frac{17}{315}|z|^7, \qquad |z| \le \frac{1}{2}.$$

We now compare q_2 to $(1 + r^2)p(r + r^3/3)$. The terms up to degree five in the latter form the polynomial $Q(r) := r + r^3 - \frac{1}{5}r^5$. There are two errors to control: replacing $\tanh$ by p, and then replacing $(1 + r^2)p(r + r^3/3)$ by Q. For $|r| \le \delta$, the first error satisfies

$$|q_2(r) - (1 + r^2)p(r + r^3/3)| \le (1 + r^2)\frac{17}{315}|r + r^3/3|^7 \le \frac{1}{8}|r|^7.$$

Here $|r + r^3/3| \le \delta(1 + \delta^2/3) < 1/2$ and $(1 + \delta^2)(17/315)(1 + \delta^2/3)^7 < 1/8$. For the second error, exact polynomial expansion gives

$$(1 + r^2)p(r + r^3/3) - Q(r) = \frac{r^7}{3645}\left(-324 + 900r^2 + 675r^4 + 210r^6 + 32r^8 + 2r^{10}\right).$$

Its absolute value is at most $|r|^7/8$ on $|r| \le \delta$, since

$$\frac{324 + 900\delta^2 + 675\delta^4 + 210\delta^6 + 32\delta^8 + 2\delta^{10}}{3645} < \frac{1}{8}.$$

Adding these two errors proves the desired bound

$$|q_2(r) - Q(r)| = \left| q_2(r) - \left(r + r^3 - \frac{1}{5} r^5 \right) \right| \le \frac{1}{4} |r|^7, \qquad |r| \le \delta. \tag{36}$$

We now move to constrcuing polynomial approximations for the one-step maps. Substituting $Q(r)$ into $(1+r)(1-q_2(r))^2 - 1$ and retaining the terms through degree five gives the polynomial

$$P_2(r) := -r - r^2 - r^3 + \frac{12}{5} r^5.$$

For $m = 3$, substituting the exact polynomial $q_3(r) = r + r^3$ gives the polynomial

$$P_3(r) := -r - r^2 - r^3 + 2r^5 + r^6 + r^7 = R_3(r).$$

Thus, P_2 is an approximation of R_2, whereas P_3 is exact. To bound the approximation error between P_2 and R_2, we first observe that

$$(1+r)(1-Q(r))^2 - 1 - P_2(r) = r^6 \left(1 + \frac{3}{5} r - \frac{2}{5} r^2 - \frac{2}{5} r^3 + \frac{1}{25} r^4 + \frac{1}{25} r^5 \right).$$

The factor in parentheses has absolute value at most $9/8$ on $|r| \le \delta$. Thus, $P_2(r) \approx (1+r)(1 - Q(r))^2 - 1$ up to order r^6. Next, using (36) and $|q_2(r)|, |Q(r)| \le 2|r|$, we obtain

$$\begin{aligned} &\left| R_2(r) - \left((1+r)(1-Q(r))^2 - 1 \right) \right| \\ &\quad = (1+r) |q_2(r) - Q(r)| \, |2 - q_2(r) - Q(r)| \\ &\quad \le \frac{1}{4} (1+\delta)(2+4\delta) |r|^7. \end{aligned}$$

Combining the two bounds, and using $9/8 + (1+\delta)(2+4\delta)\delta/4 < 2$, gives

$$|R_2(r) - P_2(r)| \le 2|r|^6, \quad R_3(r) = P_3(r), \quad |r| \le \delta. \tag{37}$$

We also record the one-step bounds needed for the invariant-set argument and for controlling the arguments of the compositions:

$$|R_m(r) + r + r^2| \le \frac{9}{8} |r|^3, \qquad |R_m(r)| \le \frac{6}{5} |r|, \qquad |r| \le \delta, \quad m = 2, 3. \tag{38}$$

For the first inequality, the formulas above bound the coefficient of $|r|^3$ by

$$1 + \frac{12}{5} \delta^2 + 2\delta^3 \le \frac{9}{8} \quad \text{for } m = 2, \quad 1 + 2\delta^2 + \delta^3 + \delta^4 \le \frac{9}{8} \quad \text{for } m = 3.$$

The second inequality follows from $1 + \delta + (9/8)\delta^2 < 6/5$.

We now target the two-step map. The advantage of using P_2, P_3 as approximations is that their compositions can be checked by finite polynomial algebra. The cancellations in the two compositions give the exact identities

$$P_2(P_2(r)) = r - \frac{4}{5} r^5 + r^6 H_2(r), \quad P_3(P_3(r)) = r - 20r^7 + r^8 H_3(r),$$

where H_2, H_3 are polynomials. We only need upper bounds on these two remainder polynomials, not their full expansions. Their first coefficients are

$$\begin{aligned} H_2(r) &= -\frac{6}{5} - \frac{162}{5} r - \frac{393}{5} r^2 - \frac{499}{5} r^3 + r^4 J_2(r), \\ H_3(r) &= -50 - 70r - 40r^2 + 60r^3 + 160r^4 + 196r^5 + r^6 J_3(r). \end{aligned}$$

Here J_2, J_3 are polynomials of degrees 15 and 35, respectively. Direct calculation gives bounds on their coefficients in absolute value by 1000 and 25000 respectively. Thus, for $|r| \le \delta$, we have

$$\begin{aligned} |H_2(r)| &\le \frac{6}{5} + \frac{162}{5} \delta + \frac{393}{5} \delta^2 + \frac{499}{5} \delta^3 + \frac{1000 \delta^4}{1-\delta} < 7, \\ |H_3(r)| &\le 50 + 70\delta + 40\delta^2 + 60\delta^3 + 160\delta^4 + 196\delta^5 + \frac{25000 \delta^6}{1-\delta} < 60. \end{aligned}$$

Consequently,

$$|P_2(P_2(r)) - r + \tfrac{4}{5}r^5| \le 7|r|^6 \quad \text{and} \quad |P_3(P_3(r)) - r + 20r^7| \le 60|r|^8, \text{ for } |r| \le \delta. \tag{39}$$

We now transfer these bounds to T_m. For $m = 3$, there is nothing to transfer, since $R_3 = P_3$ and hence $T_3 = P_3 \circ P_3$. For $m = 2$, we must account for the error in both appearances of R_2 in $T_2 = R_2 \circ R_2$. For $|r| \le 1/16$, (38) and the formula for P_2 give

$$|R_2(r)|, |P_2(r)| \le \frac{6}{5}|r| < \delta.$$

Moreover, on $[-\delta, \delta]$, we have

$$|P_2'(r)| \le 1 + 2\delta + 3\delta^2 + 12\delta^4 < \frac{3}{2}.$$

We split the composition error by first changing the outer map and then changing the inner map, i.e.

$$T_2(r) - P_2(P_2(r)) = R_2(R_2(r)) - P_2(R_2(r)) + P_2(R_2(r)) - P_2(P_2(r)).$$

Using (37) for the first difference and the mean value theorem for the second yields

$$|T_2(r) - P_2(P_2(r))| \le 2|R_2(r)|^6 + \frac{3}{2}|R_2(r) - P_2(r)| \le \left(2\left(\frac{6}{5}\right)^6 + 3\right)|r|^6 \le 9|r|^6.$$

Combining this estimate with (39) proves the required bounds in absolute value:

$$|T_2(r) - r + \tfrac{4}{5}r^5| \le 16|r|^6, \quad |r| \le \frac{1}{16}.$$

It also immediately holds

$$|T_3(r) - r + 20r^7| \le 60|r|^8.$$

Finally, for $0 < |r| \le h_0 = 1/100$, these inequalities give

$$\left|\frac{r - T_2(r)}{r^5} - A_2\right| \le 16|r| \le 16h_0 \le \frac{A_2}{4},$$

$$\left|\frac{r - T_3(r)}{r^7} - A_3\right| \le 60|r| \le 60h_0 \le \frac{A_3}{4}.$$

This proves the target bound (35) on the entire stated interval. Since $(5A_m/4)h_0^{2m} \le 1/100$, it also follows that

$$\frac{99}{100} \le \frac{T_m(r)}{r} = 1 - \frac{r - T_m(r)}{r^{2m+1}}r^{2m} < 1, \qquad 0 < |r| \le h_0, \quad m = 2, 3.$$

Thus $T_m(r)$ has the sign of r but it decreases in absolute value $|T_m(r)| < |r|$. This is what guarantees progress.

Invariant sets and loss oscillation For $m = 1$ and $0 < b \le h < 1/2$, consider

$$a := R_1(-b) = b(1 - b), \quad b^+ := -R_1(a) = a(1 + a) = b - b^3(2 - b).$$

Then $b/2 \le a < b^+ < b$. Hence

$$\widehat{\Omega}_1 = \{(1, 1 - b) : 0 \le b \le h\} \cup \{(1 + b(1 - b), 1) : 0 \le b \le h\}$$

is compact and invariant for normalized GD, and $\Omega_h = \sqrt{\rho}\,\widehat{\Omega}_1$ is invariant in the original variables.

For $m = 2, 3$, (38) implies

$$\frac{1}{2}b^2 \le b - R_m(-b) \le \frac{3}{2}b^2, \qquad \frac{1}{2}b^2 \le -R_m(b) - b \le \frac{3}{2}b^2, \qquad 0 \le b \le h_0.$$

Thus, for $a = R_m(-b)$ and $b^+ = -R_m(a)$, we have $b/2 \le a < b$ and $b^+ > a$. Equation (35) also gives

$$\frac{3A_m}{4}b^{2m+1} \le b - b^+ \le \frac{5A_m}{4}b^{2m+1},$$

so $0 < a < b^+ < b$. Since $|q_m| < 1$, all these balanced parameter updates remain positive. Consequently, with G_m denoting the corresponding GD map,

$$\mathcal{A} = \{\sqrt{1-b}(1,1) : 0 \le b \le h_0\}, \qquad \Omega_m = \mathcal{A} \cup G_m(\mathcal{A})$$

are the required compact invariant sets for $m = 2, 3$.

In all three cases, starting from $b_0 = h$, we define

$$a_j = R_m(-b_j), \qquad b_{j+1} = -R_m(a_j).$$

Induction gives

$$r_{2j} = -b_j, \qquad r_{2j+1} = a_j, \qquad \frac{1}{2}b_j \le a_j < b_{j+1} < b_j. \tag{40}$$

For $m = 1$, the normalized iterates are $(1, 1-b_j)$ and $(1+a_j, 1)$.

For $m = 2, 3$, the balanced initialization gives updates

$$w_{2j} = \sqrt{1-b_j}(1,1), \quad w_{2j+1} = \sqrt{1+a_j}(1,1).$$

In all cases we have $F(w_{2j+1}) < F(w_{2j+2}) < F(w_{2j})$.

Hessian bounds For $m = 1$, use the open convex box $U_1 = (0, 9/7)^2$, which contains $\widehat{\Omega}_1$. In this box, the maximum absolute row sum of

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

is at most $194/49 < 4$. Thus $\sqrt{\rho}\,U_1$ gives the required bound $4/\eta$ in the original variables.

For $m = 2, 3$, the Hessian is given by the more general formula

$$\nabla^2 f(u,v) = q_m'(r) \begin{pmatrix} v^2 & uv \\ uv & u^2 \end{pmatrix} + q_m(r) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The first matrix is rank one, with operator norm $u^2 + v^2$, and the second has operator norm 1. Consequently,

$$\|\nabla^2 f(u,v)\|_2 \le |q_m'(r)|(u^2+v^2) + |q_m(r)|.$$

We now apply this bound on the open convex box

$$U = (19/20, 21/20)^2.$$

For $(u,v) \in U$, we have

$$-\frac{39}{400} < r = uv - 1 < \frac{41}{400}, \qquad |r| < \frac{41}{400} < \delta = \frac{1}{8}.$$

The previously established scalar estimates therefore give

$$|q_m'(r)| \le 1 + 5r^2 \le 1 + 5\delta^2 = \frac{69}{64} < \frac{9}{8},$$
$$|q_m(r)| \le |r|(1+2r^2) \le \delta(1+2\delta^2) = \frac{33}{256} < \frac{9}{64}.$$

It follows that

$$\|\nabla^2 f(u,v)\|_2 \le \frac{9}{8}(u^2+v^2) + \frac{9}{64} \le \frac{9}{8}\,2\left(\frac{21}{20}\right)^2 + \frac{9}{64} = \frac{2097}{800} < 3 < 4.$$

Every point of Ω_m has the form $\sqrt{1+r}(1,1)$ with $|r| \le h_0 = 1/100$. Since

$$\left(\frac{19}{20}\right)^2 < 1 - h_0 \le 1 + r \le 1 + h_0 < \left(\frac{21}{20}\right)^2,$$

we have $\Omega_m \subset U$.

Finally, in each case, GD invariance gives $G(\Omega) \subseteq \Omega$ and $G^2(\Omega) \subseteq \Omega$. Thus both endpoints of every segment $[w, G^2(w)]$ belong to the corresponding open convex box. Convexity places the entire segment inside that box, establishing the required ambient Hessian bound.

A lower bound based on B_2 Recall the notation

$$g(w) = \nabla F(w), \qquad k(w) = g(w) + g(G(w)), \qquad B_2(w) = \|g(w)\|^2 + \|g(G(w))\|^2.$$

Since $\eta L = 4$ in all three cases, it holds

$$\mathcal{C}_L^{(2)}(w) = \langle g(w), k(w)\rangle - 2\|k(w)\|^2.$$

Before treating the three cases separately, observe that $g(G(w)) = k(w) - g(w)$ gives the identity

$$\|g(w)\|^2 - \|g(G(w))\|^2 = 2\langle g(w), k(w)\rangle - \|k(w)\|^2 = 2\mathcal{C}_L^{(2)}(w) + 3\|k(w)\|^2. \tag{41}$$

Consequently, once we have proved $\mathcal{C}_L^{(2)}(w) \geq 0$, we may just use $B_2(w) \leq 2\|g(w)\|^2$ to obtain the desired lower bound in terms of B_2.

For $m = 1$, every point of Ω_h, up to exchanging coordinates, has the form

$$w = \sqrt{\rho}(1, 1 + r), \qquad -\frac{1}{2} \leq r \leq \frac{1}{4}.$$

Here $r = w_1 w_2/\rho - 1$, so differentiating $F(w) = (y - x w_1 w_2)^2$ gives

$$g(w) = 2xy\, r \begin{pmatrix} w_2 \\ w_1 \end{pmatrix} = \frac{2y^2}{\sqrt{\rho}} \begin{pmatrix} r(1+r) \\ r \end{pmatrix}.$$

The next iterate and its gradient are

$$G(w) = \sqrt{\rho}(1 - r - r^2, 1), \qquad g(G(w)) = \frac{2y^2}{\sqrt{\rho}} \begin{pmatrix} -r(1+r) \\ -r(1+r)(1-r-r^2) \end{pmatrix}.$$

Adding the two gradients therefore yields

$$k(w) = \frac{2y^2}{\sqrt{\rho}} \begin{pmatrix} 0 \\ r^3(r+2) \end{pmatrix}.$$

Substitution into the certificate, with $L = 4/\eta$, gives

$$\mathcal{C}_{4/\eta}^{(2)}(w) = \langle g(w), k(w)\rangle - 2\|k(w)\|^2 = \frac{4y^4}{\rho}\, r^4(r+2)\big[1 - 2r^2(r+2)\big].$$

On the stated residual interval,

$$r + 2 \geq \frac{3}{2}, \qquad r^2(r+2) \leq \frac{3}{8}, \qquad 1 + (1+r)^2 \leq \frac{41}{16}.$$

Thus

$$\mathcal{C}_{4/\eta}^{(2)}(w) \geq \frac{4y^4}{\rho}\,\frac{3}{8} r^4 = \frac{3y^4}{2\rho} r^4 \geq 0.$$

We can now apply (41) and the computed gradient to obtain

$$B_2(w) \leq 2\|g(w)\|^2 = \frac{8y^4}{\rho} r^2[1 + (1+r)^2] \leq \frac{41y^4}{2\rho} r^2.$$

Combining these two bounds proves

$$\mathcal{C}_{4/\eta}^{(2)}(w) \geq \frac{3y^4}{2\rho} r^4 \geq \frac{6\rho}{1681 y^4} B_2(w)^2.$$

Exchanging coordinates leaves all inner products and norms unchanged, so the bound holds throughout Ω_h.

For $m = 2$, every point of Ω_2 has the form

$$w = \sqrt{1+r}(1, 1), \qquad |r| \leq h_0 = \frac{1}{100}.$$

Using the already defined $q_2(r) = (1 + r^2)\tanh(r + r^3/3)$, its gradient is $g(w) = \sqrt{1+r}\, q_2(r) \begin{pmatrix}1\\1\end{pmatrix}$. The iterates remain positive and balanced, so $G^2(w) = \sqrt{1+T_2(r)}(1,1)$. Since $\eta = 1$, the sum of the two gradients is exactly the two-iterations displacement. We can write

$$k(w) = w - G^2(w) = \left[\sqrt{1+r} - \sqrt{1+T_2(r)}\right] \begin{pmatrix}1\\1\end{pmatrix} = \frac{r - T_2(r)}{\sqrt{1+r} + \sqrt{1+T_2(r)}} \begin{pmatrix}1\\1\end{pmatrix}.$$

Consequently,

$$\mathcal{C}_4^{(2)}(w) = \frac{2\sqrt{1+r}\, q_2(r)[r - T_2(r)]}{\sqrt{1+r} + \sqrt{1+T_2(r)}} - \frac{4[r - T_2(r)]^2}{[\sqrt{1+r} + \sqrt{1+T_2(r)}]^2}.$$

For $0 < |r| \le h_0$, the previously established estimates give

$$1 \le \frac{q_2(r)}{r} \le 1 + 2r^2, \qquad \frac{3}{5} \le \frac{r - T_2(r)}{r^5} \le 1, \qquad |T_2(r)| \le |r| \le h_0.$$

In particular, $q_2(r)$ and $r - T_2(r)$ both have the sign of r, and hence

$$q_2(r)[r - T_2(r)] \ge \frac{3}{5}|r|^6, \qquad |r - T_2(r)| \le |r|^5.$$

The square-root factors satisfy

$$\frac{2\sqrt{1+r}}{\sqrt{1+r} + \sqrt{1+T_2(r)}} \ge \sqrt{\frac{1-h_0}{1+h_0}} \ge \frac{19}{20}, \quad \left(\sqrt{1+r} + \sqrt{1+T_2(r)}\right)^2 \ge 4(1-h_0).$$

Substituting these estimates into the exact certificate yields

$$\mathcal{C}_4^{(2)}(w) \ge \frac{19}{20}\frac{3}{5}|r|^6 - \frac{|r|^{10}}{1-h_0} \ge \left(\frac{57}{100} - \frac{h_0^4}{1-h_0}\right)|r|^6 \ge \frac{14}{25}|r|^6 \ge 0,$$

where the last inequality uses $h_0^4/(1-h_0) \le 1/100$ for $h_0 = 1/100$.

We may therefore apply (41) and get

$$B_2(w) \le 2\|g(w)\|^2 = 4(1+r)q_2(r)^2 \le 4(1+h_0)(1+2h_0^2)^2 r^2 \le \frac{21}{5}r^2.$$

Since $(21/5)^3/160 \le 14/25$, we conclude that

$$\mathcal{C}_4^{(2)}(w) \ge \frac{14}{25}|r|^6 \ge \frac{1}{160}B_2(w)^3.$$

For $m = 3$, every point of Ω_3 likewise has the form

$$w = \sqrt{1+r}(1,1), \qquad |r| \le h_0 = \frac{1}{100}.$$

Here $q_3(r) = r + r^3$, so the gradient is $g(w) = \sqrt{1+r}(r + r^3) \begin{pmatrix}1\\1\end{pmatrix}$. Positivity and balance give $G^2(w) = \sqrt{1+T_3(r)}(1,1)$. Since $\eta = 1$, the gradient sum is

$$k(w) = w - G^2(w) = \left[\sqrt{1+r} - \sqrt{1+T_3(r)}\right] \begin{pmatrix}1\\1\end{pmatrix} = \frac{r - T_3(r)}{\sqrt{1+r} + \sqrt{1+T_3(r)}} \begin{pmatrix}1\\1\end{pmatrix}.$$

Thus the certificate is

$$\mathcal{C}_4^{(2)}(w) = \frac{2\sqrt{1+r}(r+r^3)[r - T_3(r)]}{\sqrt{1+r} + \sqrt{1+T_3(r)}} - \frac{4[r - T_3(r)]^2}{[\sqrt{1+r} + \sqrt{1+T_3(r)}]^2}.$$

For $0 < |r| \le h_0$, the return-map bounds give

$$15 \le \frac{r - T_3(r)}{r^7} \le 25, \qquad |T_3(r)| \le |r| \le h_0.$$

Since $(r+r^3)/r = 1+r^2 \geq 1$, it follows that

$$(r+r^3)[r - T_3(r)] \geq 15|r|^8, \qquad |r - T_3(r)| \leq 25|r|^7.$$

The same bounds on the square-root factors used in the case $m = 2$ apply with T_3 in place of T_2. Consequently,

$$\mathcal{C}_4^{(2)}(w) \geq \frac{19}{20}\,15|r|^8 - \frac{625}{1-h_0}|r|^{14} \geq \left(\frac{57}{4} - \frac{625h_0^6}{1-h_0}\right)|r|^8 \geq 14|r|^8 \geq 0,$$

where $625h_0^6/(1-h_0) \leq 1/4$ for $h_0 = 1/100$. Applying (41) now gives

$$B_2(w) \leq 2\|g(w)\|^2 = 4(1+r)(r+r^3)^2 \leq 4(1+h_0)(1+h_0^2)^2 r^2 \leq \frac{21}{5}r^2.$$

Since $(5/128)(21/5)^4 \leq 14$, we obtain

$$\mathcal{C}_4^{(2)}(w) \geq 14|r|^8 \geq \frac{5}{128}B_2(w)^4.$$

Sharpness and averaged sharpness Recall that $r_{2j} = -b_j < 0 < a_j = r_{2j+1}$ and that $g(w_t) \neq 0$ at every finite iterate. For averaged sharpness, we use the exact one-step identity

$$\eta\widetilde{\lambda}^t - 2 = \frac{2[F(w_{t+1}) - F(w_t)]}{\eta\|g(w_t)\|^2}. \tag{42}$$

For $m = 1$, at $w = \sqrt{\rho}(1, 1+r)$, where $r = w_1w_2/\rho - 1$, the Hessian and its two eigenvalues are

$$\nabla^2 F(w) = \frac{1}{\eta}\begin{pmatrix}(1+r)^2 & 1+2r\\ 1+2r & 1\end{pmatrix},$$

$$\lambda_\pm(r) = \frac{2+2r+r^2 \pm \sqrt{r^2(r+2)^2 + 4(1+2r)^2}}{2\eta}.$$

up to a change of coordinates. Since $\lambda_+ + \lambda_- > 0$, we have $\|\nabla^2 F(w)\|_2 = \lambda_+(r)$. On $-1/2 < r < 1/4$, both $2-2r-r^2$ and $6+5r$ are positive, and

$$r^2(r+2)^2 + 4(1+2r)^2 - (2-2r-r^2)^2 = 4r(6+5r).$$

Comparing the square root in λ_+ with $2-2r-r^2$ therefore shows that $\eta\lambda_+(r) - 2$ has the sign of r/ Thus $\eta\kappa_{2j} < 2 < \eta\kappa_{2j+1}$, since r alternates from positive to negative in each iteration. For averaged sharpness, substituting $F(w_t) = y^2 r_t^2$, $r_{t+1} = -r_t(1+r_t)$, and the computed gradient into (42) gives

$$\eta\widetilde{\lambda}^t - 2 = \frac{r_t(r_t+2)}{1+(1+r_t)^2}, \qquad \eta\widetilde{\lambda}^{2j} < 2 < \eta\widetilde{\lambda}^{2j+1}.$$

For $m = 2$, at $w = \sqrt{1+r}(1,1)$, where $r = w_1w_2 - 1$, recall $q_2(r) = (1+r^2)\tanh(r + r^3/3)$. The Hessian is

$$\nabla^2 F(w) = \begin{pmatrix}(1+r)q_2'(r) & (1+r)q_2'(r) + q_2(r)\\ (1+r)q_2'(r) + q_2(r) & (1+r)q_2'(r)\end{pmatrix}.$$

Its eigenvectors $(1,1)$ and $(1,-1)$ have respective eigenvalues

$$\lambda_+(r) = 2(1+r)q_2'(r) + q_2(r), \quad \lambda_-(r) = -q_2(r).$$

For $|r| \leq \delta = 1/8$, the established bounds on q_2, q_2' give

$$|\lambda_+(r) - 2 - 3r| = |2(1+r)(q_2'(r) - 1) + q_2(r) - r| \leq 12r^2,$$

$$\lambda_+(r) \geq 2 - 3\delta - 12\delta^2 = \frac{23}{16} > \frac{33}{256} \geq |\lambda_-(r)|.$$

Hence $\|\nabla^2 F(w)\|_2 = \lambda_+(r)$. Since $3 - 12\delta > 0$, the first bound shows that $\lambda_+(r) - 2$ has the sign of r, proving $\kappa_{2j} < 2 < \kappa_{2j+1}$. For the averaged sharpness, (42) gives again $\widetilde{\lambda}^{2j} < 2 < \widetilde{\lambda}^{2j+1}$, since the loss oscillates.

For $m = 3$, at $w = \sqrt{1+r}(1,1)$, where $r = w_1 w_2 - 1$, the Hessian is

$$\nabla^2 F(w) = \begin{pmatrix} (1+r)(1+3r^2) & (1+r)(1+3r^2) + r + r^3 \\ (1+r)(1+3r^2) + r + r^3 & (1+r)(1+3r^2) \end{pmatrix}.$$

Its eigenvectors $(1,1)$ and $(1,-1)$ have respective eigenvalues

$$\lambda_+(r) = 2 + 3r + 6r^2 + 7r^3, \qquad \lambda_-(r) = -r - r^3.$$

For $|r| \le \delta = 1/8$,

$$\lambda_+(r) \ge 2 - 3\delta - 7\delta^3 = \frac{825}{512} > \frac{65}{512} = \delta(1+\delta^2) \ge |\lambda_-(r)|,$$

so $\|\nabla^2 F(w)\|_2 = \lambda_+(r)$. Moreover,

$$\lambda_+(r) - 2 = r(3 + 6r + 7r^2), \qquad 3 + 6r + 7r^2 \ge 3 - 6\delta > 0.$$

Thus $\kappa_{2j} < 2 < \kappa_{2j+1}$ since r changes sign in each iteration. Again, the strict loss interlacing in (42) gives $\widetilde{\lambda}^{2j} < 2 < \widetilde{\lambda}^{2j+1}$.

Finally, in each case, the certificate and the two-step Taylor inequality yield, for every $t \ge 0$,

$$F(w_t) - F(w_{t+2}) \ge \eta c_m B_2(w_t)^{m+1} > 0.$$

This is the desired result. □

F.2 Proof of Proposition 5

Proof. We use the Frobenius inner product for matrices and the product norm $\|(X,Y)\|^2 = \|X\|_F^2 + \|Y\|_F^2$ for parameter pairs. All radii and constants below are explicit, i.e. we do not use an asymptotic argument to choose the initialization range.

Parameters determining the block length until the first improvement For $j \ge 1$, put

$$a_j := 1 - 2^{-j}, \qquad b_j := \frac{a_j}{4}\left(\frac{4}{3}a_j - \frac{1}{2}\right).$$

The quantity a_j/j is strictly decreasing. Define

$$\begin{aligned} \delta &:= \frac{a_{k-1}}{k-1} - \frac{a_k}{k} = \frac{1-(k+1)2^{-k}}{k(k-1)} > 0, \qquad \vartheta := \frac{a_k}{k} + \frac{\delta}{2}, \\ \chi &:= \min\left\{1, 4\delta, \frac{k\delta}{4b_k}\right\}, \qquad \beta := \chi\vartheta, \\ \mu &:= k\beta - a_k\chi - b_k\chi^2, \qquad \nu := \chi\delta. \end{aligned} \tag{43}$$

These parameters depend on k, but not on m or d. Since $b_k > 0$, they satisfy

$$\mu \ge \frac{k\nu}{4} > 0, \qquad 2(j\beta - a_j\chi) \le -j\nu \quad (1 \le j < k), \qquad \frac{\chi^2}{8} \le \frac{\nu}{2}.$$

Indeed, $k\beta - a_k\chi = k\nu/2$, and $a_j/j \ge a_{k-1}/(k-1) = \vartheta + \delta/2$ for $j < k$. Also,

$$0 < \chi \le \frac{1}{3}, \qquad 0 < \beta < \frac{\chi}{3}, \tag{44}$$

because $\delta < 1/[k(k-1)] \le 1/12$ and $\vartheta < a_{k-1}/(k-1) < 1/(k-1) \le 1/3$.

Exact gradients and updates For ease of exposition, we use the notation

$$\alpha := \beta d^{2m}, \qquad \gamma := \frac{\alpha}{2m+2}, \qquad \zeta := (2m+1)\gamma^2, \qquad r := \|U\|_F, \quad R := V - U^2.$$

Entrywise powers are denoted by $\odot$. Direct differentiation gives

$$\begin{aligned} g_U &= U - \alpha U^{\odot(2m+1)} + \zeta U^{\odot(4m+1)} - \frac{1}{4}(U^\top R + R U^\top), \\ g_V &= \frac{1}{4}R. \end{aligned}$$

For the simultaneous update with $\eta = 2$, we have

$$U^+ = -U + \Delta, \qquad V^+ = U^2 + \frac{1}{2}R. \tag{45}$$

where

$$\Delta := 2\alpha U^{\odot(2m+1)} + \frac{1}{2}(U^\top R + RU^\top) - 2\zeta U^{\odot(4m+1)}, \tag{46}$$

We also have

$$(U^+)^2 = U^2 - Q, \qquad R^+ = \frac{1}{2}R + Q,$$

where

$$Q := U\Delta + \Delta U - \Delta^2. \tag{47}$$

All matrices in these equations are in $\mathbb{R}^{d\times d}$. In particular, $g_U = U - \Delta/2$.

A bound for the Hessian spectral norm For $X, Y \in \mathbb{R}^{d\times d}$, we have

$$\begin{aligned}\nabla^2 F_m(U,V)[(X,Y),(X,Y)] = \sum_{i,j}\left[1 - (2m+1)\alpha U_{ij}^{2m} + (4m+1)\zeta U_{ij}^{4m}\right] X_{ij}^2 \\ + \frac{1}{4}\|Y - XU - UX\|_F^2 - \frac{1}{2}\langle R, X^2\rangle_F.\end{aligned} \tag{48}$$

Indeed, set $\varphi(z) = \frac{1}{2}\sigma_m(z)^2$ and $R(t) = V + tY - (U + tX)^2$. It holds

$$\begin{aligned}\varphi''(z) &= 1 - (2m+1)\alpha z^{2m} + (4m+1)\zeta z^{4m}, \\ R(t) &= R + t(Y - XU - UX) - t^2X^2, \qquad R''(0) = -2X^2.\end{aligned}$$

The expression for $\phi''(z)$ is obtained as follows. Since $\sigma_m'(z) = 1 - (2m+1)\gamma z^{2m}$, it holds

$$\begin{aligned}\varphi'(z) &= \sigma_m(z)\sigma_m'(z) = (z - \gamma z^{2m+1})\big(1 - (2m+1)\gamma z^{2m}\big) \\ &= z - (2m+2)\gamma z^{2m+1} + (2m+1)\gamma^2 z^{4m+1} = z - \alpha z^{2m+1} + \zeta z^{4m+1}.\end{aligned}$$

Differentiating once more gives the formula.

By the definition of the Hessian, we have

$$\begin{aligned}\nabla^2 F_m(U,V)[(X,Y),(X,Y)] &= \frac{d^2}{dt^2}F_m(U + tX, V + tY)\Big|_{t=0} \\ &= \sum_{i,j}\varphi''(U_{ij})X_{ij}^2 + \frac{1}{4}\|R'(0)\|_F^2 + \frac{1}{4}\langle R, R''(0)\rangle_F \\ &= \sum_{i,j}\varphi''(U_{ij})X_{ij}^2 + \frac{1}{4}\|Y - XU - UX\|_F^2 - \frac{1}{2}\langle R, X^2\rangle_F\end{aligned}$$

and expression (48) follows after substituting the value of $\phi''(U_{ij})$.

Thus $\nabla^2 F_m(0,0) = \mathrm{diag}(I_{d^2}, I_{d^2}/4)$. Define

$$H_0 := (2m+1)\alpha + (4m+1)\zeta + \frac{3}{2}, \qquad \delta_H := \min\left\{\frac{1}{6}, \frac{1}{\sqrt{6H_0}}\right\}.$$

Now assume that $\sqrt{\|U\|_F^2 + \|V\|_F^2} \le t \le 1$. Then

$$\|Z\|_F \le 2t\|X\|_F, \qquad \|R\|_F \le t + t^2, \qquad \|X^2\|_F \le \|X\|_F^2.$$

Since $m \ge 1$, the activation coefficient satisfies

$$\left|-(2m+1)\alpha U_{ij}^{2m} + (4m+1)\zeta U_{ij}^{4m}\right| \le \left(H_0 - \frac{3}{2}\right)t^2.$$

Expanding $\|Y - XU - UX\|_F^2$ in (48) gives

$$\begin{aligned}
&\left|\nabla^2 F_m(U,V)[(X,Y),(X,Y)] - \|X\|_F^2 - \frac{1}{4}\|Y\|_F^2\right| \\
&\le \left(H_0 - \frac{3}{2}\right) t^2\|X\|_F^2 + \frac{1}{2}\|Y\|_F\|Z\|_F + \frac{1}{4}\|Z\|_F^2 + \frac{1}{2}\|R\|_F\|X^2\|_F \\
&\le \left(\frac{t}{2} + H_0 t^2\right)\|X\|_F^2 + t\|X\|_F\|Y\|_F \\
&\le (t + H_0 t^2)\|X\|_F^2 + \frac{t}{2}\|Y\|_F^2 \\
&\le (t + H_0 t^2)\big(\|X\|_F^2 + \|Y\|_F^2\big),
\end{aligned}$$

where the penultimate inequality uses $2\|X\|_F\|Y\|_F \le \|X\|_F^2 + \|Y\|_F^2$.

Consequently,

$$\|\nabla^2 F_m(U,V)\|_2 \le \frac{4}{3} \qquad \text{whenever } \|(U,V)\| < \delta_H. \tag{49}$$

This is a global bound on the Hessian norm, not only on the curvature in the direction of a GD update.

Construction of an invariant region Define

$$\begin{aligned}
D_0 &:= 2\alpha + \chi + 2\zeta, \qquad C_0 := 4\zeta + 37D_0^2 + (12 + 6\chi)D_0, \\
\rho_{\text{inv}} &:= \min\left\{1, (8mD_0)^{-1/(2m)}, \sqrt{\frac{\chi}{12D_0}}, (8\chi)^{-1/(2m)}, \sqrt{\frac{2\beta}{C_0}}, \frac{\delta_H}{2(2+\chi)}\right\}.
\end{aligned} \tag{50}$$

We first prove invariance for every $0 < r_0 \le \rho_{\text{inv}}$, with $\epsilon := r_0/4$. We will specify the final r_0 in (63).

Suppose $r \le r_0$ and $\|R\|_F \le \chi r^{2m}$ (recall that $r := \|U\|_F$). Since $r \le 1$, it holds $r^{2m+1}, r^{4m+1} \le r^{2m}$ and $r^{4m+2} \le r^{2m+2}$. The exact update from (46) and (47) gives

$$\|\Delta\|_F \le D_0 r^{2m+1}, \qquad \|Q\|_F \le 2r\|\Delta\|_F + \|\Delta\|_F^2 \le 3D_0 r^{2m+2}. \tag{51}$$

Writing $r^+ := \|U^+\|_F$, we get

$$(r^+)^{2m} \ge r^{2m}(1 - D_0 r^{2m})^{2m} \ge r^{2m}(1 - 2mD_0 r^{2m}) \ge \frac{3}{4} r^{2m}.$$

The first inequality uses $U^+ = -U + \Delta$ and $\|\Delta\|_F \le D_0 r^{2m+1}$, the second uses Bernoulli inequality and the third uses that $r \le r_0 \le \rho_{\text{inv}} \le \frac{1}{(8mD_0)^{1/(2m)}}$.

On the other hand, $r_0^2 \le \chi/(12D_0)$ gives

$$\|R^+\|_F \le \frac{\chi}{2} r^{2m} + 3D_0 r^{2m+2} \le \frac{3\chi}{4} r^{2m} \le \chi (r^+)^{2m}.$$

Thus the tube constraint is preserved.

To control its radius, we denote

$$\mathcal{E}(U,V) := r^2 - 4\langle U^2, R\rangle_F.$$

Because $4\chi r_0^{2m} \le 1/2$ from (50), it holds $|\langle U^2, R\rangle_F| \le \|U^2\|_F\|R\|_F \le \frac{r^2}{8}$ and thus

$$\frac{1}{2} r^2 \le \mathcal{E}(U,V) \le \frac{3}{2} r^2 \tag{52}$$

on the tube within radius r_0. The identities (45) give exactly

$$\begin{aligned}
\mathcal{E}(U^+,V^+) - \mathcal{E}(U,V) = &-4\alpha \sum_{i,j} |U_{ij}|^{2m+2} + 4\zeta \sum_{i,j} |U_{ij}|^{4m+2} + \|\Delta\|_F^2 \\
&- 4\langle U^2, Q\rangle_F + 2\langle Q, R\rangle_F + 4\|Q\|_F^2.
\end{aligned} \tag{53}$$

All terms after the first are bounded above in absolute value, in total, by $C_0 r^{2m+4}$: their respective coefficients are $4\zeta, D_0^2, 12D_0, 6\chi D_0, 36D_0^2$, using (51), $r \le 1$, $m \ge 1$ and Cauchy-Schwarz. Moreover, the power-mean inequality over the d^2 entries gives

$$\sum_{i,j} |U_{ij}|^{2m+2} \ge d^{-2m} r^{2m+2}. \tag{54}$$

Since $\alpha = \beta d^{2m}$ by definition and $r_0^2 \le 2\beta/C_0$ by (50), it holds

$$\mathcal{E}(U^+, V^+) \le \mathcal{E}(U, V) - 2\beta r^{2m+2}. \tag{55}$$

If $(U, V) \in \Omega$, then (52) implies $r \le \sqrt{2\mathcal{E}(U,V)} \le \sqrt{2}\epsilon = r_0/(2\sqrt{2})$ since ϵ was taken equal to $r_0/4$. Also,

$$r^+ = \|U^+\|_F \le \|U\|_F + \|\Delta\|_F \le r + D_0 r^{2m} = r(1 + D_0 r^{2m}) \le \frac{9}{8} r \le \frac{9}{8}\frac{1}{2\sqrt{2}} r_0 < r_0.$$

Together with the tube estimate and (55), we have shown that all constraints in (7) are preserved. Thus, Ω is a GD invariant region: $G(\Omega) \subseteq \Omega$. The case $r = 0$ has $R = 0$ and is the fixed point $(0, 0)$. The set is closed and bounded, and points with $R = 0$ and $0 < r < \epsilon$ satisfy all its inequalities strictly. Thus it is compact and has nonempty interior.

Summing (55) gives

$$\sum_{t=0}^{\infty} \|U_t\|_F^{2m+2} \le \frac{\mathcal{E}(U_0, V_0)}{2\beta}.$$

Therefore $U_t \to 0$, and the tube constraint yields $V_t \to 0$. Finally, on Ω, we have

$$\|(U, V)\| \le r + \|V\|_F \le r + \|R\|_F + \|U\|_F^2 \le r + \chi r^{2m} + r^2 \le r + (1+\chi) r^2 \le (2+\chi) r_0 \le \frac{\delta_H}{2}.$$

The last but one inequality follows from $r \le r_0/(2\sqrt{2})$ and the last one by (50).

Invariance and convexity place every required iterate and segment in the open ball of (49).

Uniform finite-block estimates Fix an initial point (U, V) in such an invariant region and put $(U_i, V_i) := G^i(U, V)$, $R_i := V_i - U_i^2$. Let $q := 9/8$ and define

$$\begin{aligned} \mathsf{U} &:= D_0 k q^{(2m+1)k}, \\ \mathsf{R} &:= 6 D_0 q^{(2m+2)k}, \\ \mathsf{D} &:= k\left[(2\alpha(2m+1) + 2\chi) q^{2mk} \mathsf{U} + 2\mathsf{R} + 2\zeta q^{(4m+1)k}\right], \\ \mathsf{V} &:= \frac{(q^k + 1)\mathsf{U} + \mathsf{R}}{2}. \end{aligned} \tag{56}$$

For $0 \le i \le k$, invariance and (51) give

$$\|U_i\|_F \le q^i r, \quad \|U_i - (-1)^i U\|_F \le \mathsf{U} r^{2m+1}, \quad \|R_i - 2^{-i} R\|_F \le \mathsf{R} r^{2m+2}. \tag{57}$$

The first bound follows by applying recursively $r^+ \le \frac{9}{8} r$. For the second bound, we sum the updates $U_{i+1} = -U_i + \Delta_i$ and use $\|\Delta_i\|_F \le D_0 r_i^{2m+1}$. For the third, sum $R_{i+1} = R_i/2 + Q_i$ and use $\|Q_i\|_F \le 3 D_0 r_i^{2m+2}$.

For a positive integer p, the entrywise power map satisfies

$$\|X^{\odot p} - Y^{\odot p}\|_F \le p \max\{\|X\|_F, \|Y\|_F\}^{p-1} \|X - Y\|_F.$$

This is obtained by rewriting the Hadamard products into products of entries and use $x^p - y^p = (x - y)\sum_{k=0}^{p-1} x^k y^{p-1-k}$.

Using this bound in the exact update and applying equation (46), for $0 \le i < k$, we get

$$\left\| \Delta_i - (-1)^i \left[2\alpha U^{\odot(2m+1)} + \frac{2^{-i}}{2}(U^\top R + R U^\top) \right] \right\|_F \le \frac{\mathsf{D}}{k} r^{2m+3}. \tag{58}$$

The activation difference is bounded by $2\alpha(2m+1)q^{2mk}\mathsf{U}r^{4m+1}$, the bilinear difference is bounded by $2\chi q^{2mk}\mathsf{U}r^{4m+1} + 2\mathsf{R}r^{2m+3}$, and the last activation term by $2\zeta q^{(4m+1)k}r^{4m+1}$. Since $r \le 1$ and $4m+1 \ge 2m+3$, their sum yields (58).

Summing the signed updates gives, for every $1 \le j \le k$,

$$(-1)^j U_j = U - 2j\alpha U^{\odot(2m+1)} - a_j(U^\top R + RU^\top) + E_j, \quad \|E_j\|_F \le \mathsf{D}r^{2m+3}. \tag{59}$$

Here the activation accumulates with coefficient j, whereas $\sum_{i=0}^{j-1} 2^{-i} = 2a_j$ controls the accumulated mismatch. Since $S_j = ((U-U_j)/2, (V-V_j)/2)$, for even $j \le k$,

$$\begin{aligned}(S_j)_U &= j\alpha U^{\odot(2m+1)} + \frac{a_j}{2}(U^\top R + RU^\top) + E_{j,U},\\ \|E_{j,U}\|_F &\le \frac{\mathsf{D}}{2}r^{2m+3}, \qquad \|(S_j)_U\|_F \le \frac{\mathsf{U}}{2}r^{2m+1}.\end{aligned} \tag{60}$$

For all $1 \le j \le k$,

$$(S_j)_V = \frac{a_j}{2}R + E_{j,V}, \qquad \|E_{j,V}\|_F \le \mathsf{V}r^{2m+2}. \tag{61}$$

Indeed,

$$\|U_j^2 - U^2\|_F \le (\|U_j\|_F + r)\|U_j - (-1)^j U\|_F \le (q^k+1)\mathsf{U}r^{2m+2},$$

and $V_j = U_j^2 + R_j$.

Remainder constants and final radii Define

$$\begin{aligned}M_C &:= \frac{\mathsf{D}}{2} + \frac{D_0\mathsf{U}}{4} + \frac{\chi\mathsf{V}}{4} + \frac{\mathsf{U}^2}{3} + \frac{4\chi\mathsf{V}}{3} + \frac{4\mathsf{V}^2}{3},\\ M_F &:= \mathsf{D} + (\alpha+\zeta)\mathsf{U} + \frac{2}{3}\mathsf{U}^2 + \frac{\chi\mathsf{R}}{4} + \frac{\mathsf{R}^2}{8},\\ M &:= 1 + M_C + M_F.\end{aligned} \tag{62}$$

The constants M_C and M_F control the certificate remainder and the exact loss remainder proved below respectively. Set

$$r_0 := \min\left\{\rho_{\mathrm{inv}}, \sqrt{\frac{\mu}{2M}}, \sqrt{\frac{\nu}{4M}}\right\}, \qquad \epsilon := \frac{r_0}{4}. \tag{63}$$

All constants are strictly positive explicit functions of (d, m, k). The preceding invariance and Hessian estimates apply to this choice.

The length-k certificate on the whole region. Since $\eta = 2$ and $L = 4/3$, the certificate is exactly

$$\mathcal{C}_{4/3}^{(j)} = \langle U, (S_j)_U\rangle_F + \langle e_U, (S_j)_U\rangle_F + \frac{1}{4}\langle R, (S_j)_V\rangle_F - \frac{4}{3}\|(S_j)_U\|_F^2 - \frac{4}{3}\|(S_j)_V\|_F^2.$$

The previously established estimates (60)–(61) are

$$(S_j)_U = j\alpha U^{\odot(2m+1)} + \frac{a_j}{2}(U^\top R + RU^\top) + E_{j,U}, \quad (S_j)_V = \frac{a_j}{2}R + E_{j,V},$$

$$\|E_{j,U}\|_F \le \frac{\mathsf{D}}{2}r^{2m+3}, \quad \|(S_j)_U\|_F \le \frac{\mathsf{U}}{2}r^{2m+1}, \quad \|E_{j,V}\|_F \le \mathsf{V}r^{2m+2}.$$

The exact gradient can be written as $g_U = U + e_U$ and $g_V = R/4$, where

$$e_U = -\alpha U^{\odot(2m+1)} + \zeta U^{\odot(4m+1)} - \frac{1}{4}(U^\top R + RU^\top).$$

Submultiplicativity, the tube constraint, and $r \le 1$ give

$$\|e_U\|_F \le \alpha r^{2m+1} + \zeta r^{4m+1} + \frac{1}{2}r\|R\|_F \le \left(\alpha + \zeta + \frac{\chi}{2}\right)r^{2m+1} = \frac{D_0}{2}r^{2m+1}.$$

For computing the first inner product $\langle U, (S_j)_U \rangle_F$, we use

$$\langle U, U^{\odot(2m+1)} \rangle_F = \sum_{a,b} |U_{ab}|^{2m+2},$$
$$\langle U, U^\top R + R U^\top \rangle_F = \operatorname{tr}((U^\top)^2 R) + \operatorname{tr}(U^\top R U^\top) = 2\operatorname{tr}((U^2)^\top R) = 2\langle U^2, R \rangle_F.$$

The second identity uses only the cyclic property of the trace. Consequently,

$$\langle U, (S_j)_U \rangle_F = j\alpha \sum_{a,b} |U_{ab}|^{2m+2} + a_j \langle U^2, R \rangle_F + \langle U, E_{j,U} \rangle_F.$$

For the V component, expanding the squared norm gives

$$\frac{1}{4}\langle R, (S_j)_V \rangle_F - \frac{4}{3}\|(S_j)_V\|_F^2 = \left(\frac{a_j}{8} - \frac{a_j^2}{3}\right)\|R\|_F^2 + \frac{1}{4}\langle R, E_{j,V} \rangle_F - \frac{4a_j}{3}\langle R, E_{j,V} \rangle_F - \frac{4}{3}\|E_{j,V}\|_F^2.$$

In particular,

$$\frac{a_j}{8} - \frac{a_j^2}{3} = -\frac{a_j}{4}\left(\frac{4}{3}a_j - \frac{1}{2}\right) = -b_j.$$

Thus, defining the remainder by the exact identity

$$\mathcal{R}_j(U,V) := \langle U, E_{j,U} \rangle_F + \langle e_U, (S_j)_U \rangle_F + \frac{1}{4}\langle R, E_{j,V} \rangle_F - \frac{4}{3}\|(S_j)_U\|_F^2 - \frac{4a_j}{3}\langle R, E_{j,V} \rangle_F - \frac{4}{3}\|E_{j,V}\|_F^2,$$

we obtain

$$\mathcal{C}_{4/3}^{(j)}(U,V) = j\alpha \sum_{a,b} |U_{ab}|^{2m+2} + a_j \langle U^2, R \rangle_F - b_j \|R\|_F^2 + \mathcal{R}_j(U,V), \tag{64}$$

Now, we proceed on bounding $|R_j(U,V)|$.

Cauchy–Schwarz, the bounds recalled above (together with $\|R\|_F \le \chi r^{2m}$, and $0 < a_j < 1$ give

$$|\langle U, E_{j,U} \rangle_F| \le \frac{\mathsf{D}}{2} r^{2m+4}, \quad |\langle e_U, (S_j)_U \rangle_F| \le \frac{D_0 \mathsf{U}}{4} r^{4m+2}, \quad \frac{1}{4}|\langle R, E_{j,V} \rangle_F| \le \frac{\chi \mathsf{V}}{4} r^{4m+2},$$
$$\frac{4}{3}\|(S_j)_U\|_F^2 \le \frac{\mathsf{U}^2}{3} r^{4m+2}, \quad \frac{4a_j}{3}|\langle R, E_{j,V} \rangle_F| \le \frac{4\chi \mathsf{V}}{3} r^{4m+2}, \quad \frac{4}{3}\|E_{j,V}\|_F^2 \le \frac{4\mathsf{V}^2}{3} r^{4m+4}.$$

Since $m \ge 1$ and $0 \le r \le 1$,

$$r^{4m+2} \le r^{2m+4}, \qquad r^{4m+4} \le r^{2m+4}.$$

The triangle inequality therefore yields

$$|\mathcal{R}_j(U,V)| \le \left(\frac{\mathsf{D}}{2} + \frac{D_0 \mathsf{U}}{4} + \frac{\chi \mathsf{V}}{4} + \frac{\mathsf{U}^2}{3} + \frac{4\chi \mathsf{V}}{3} + \frac{4\mathsf{V}^2}{3}\right) r^{2m+4} = M_C r^{2m+4}.$$

This is because the quantity inside the parenthesis was defined to be M_C.

Using (54), $\|U^2\|_F \le r^2$ and the tube constraints, we obtain

$$\mathcal{C}_{4/3}^{(k)}(U,V) \ge (k\beta - a_k\chi) r^{2m+2} - b_k \chi^2 r^{4m} - M_C r^{2m+4} \ge \mu r^{2m+2} - M r^{2m+4} \ge \frac{\mu}{2} r^{2m+2}.$$

The second line uses $r \le 1$ and $4m \ge 2m + 2$. This is a lower bound for our certificate. Now, we also need to bound the gradient norm.

The exact gradient and update estimates give

$$g_U(U_i, V_i) = U_i - \frac{1}{2}\Delta_i, \qquad g_V(U_i, V_i) = \frac{1}{4}R_i, \qquad \|\Delta_i\|_F \le D_0\|U_i\|_F^{2m+1}.$$

Consequently, using the product norm on the parameter space,

$$\begin{aligned}\|g(U_i, V_i)\|^2 &= \left\|U_i - \frac{1}{2}\Delta_i\right\|_F^2 + \frac{1}{16}\|R_i\|_F^2 \\ &\le \left(\|U_i\|_F + \frac{D_0}{2}\|U_i\|_F^{2m+1}\right)^2 + \frac{\chi^2}{16}\|U_i\|_F^{4m} \\ &\le \left(\left(1 + \frac{D_0}{2}\right)^2 + \frac{\chi^2}{16}\right)\|U_i\|_F^2.\end{aligned}$$

The last inequality follows from $m \ge 1$ and $\|U_i\|_F \le 1$, which imply $\|U_i\|_F^{2m+1} \le \|U_i\|_F$ and $\|U_i\|_F^{4m} \le \|U_i\|_F^2$.

Next, we recall

$$\|U_i\|_F \le q^i r, \qquad 0 \le i \le k.$$

and the series sum gives

$$\sum_{i=0}^{k-1}\|U_i\|_F^2 \le r^2\sum_{i=0}^{k-1} q^{2i} = \frac{q^{2k}-1}{q^2-1}\,r^2.$$

Combining this estimate with the gradient bound yields

$$\begin{aligned}B_k(U, V) &= \sum_{i=0}^{k-1}\|g(U_i, V_i)\|^2 \\ &\le \left(\left(1 + \frac{D_0}{2}\right)^2 + \frac{\chi^2}{16}\right)\sum_{i=0}^{k-1}\|U_i\|_F^2 \\ &\le \left(\left(1 + \frac{D_0}{2}\right)^2 + \frac{\chi^2}{16}\right)\frac{q^{2k}-1}{q^2-1}\,r^2.\end{aligned}$$

Thus, with

$$C_B := \left(\left(1 + \frac{D_0}{2}\right)^2 + \frac{\chi^2}{16}\right)\frac{q^{2k}-1}{q^2-1}, \qquad c := \frac{\mu}{2C_B^{m+1}}, \tag{65}$$

we have $B_k(U, V) \le C_B r^2$ and $\mathcal{C}_{4/3}^{(k)}(U, V) \ge cB_k(U, V)^{m+1}$.

This is condition (C_m^k) using a full Hessian norm bound.

Initializations and first loss improvement Fix $0 < h \le \epsilon/2$, let $e = d^{-1/2}(1, \dots, 1)^\top \in \mathbb{R}^d$, and put $P = ee^\top \in \mathbb{R}^{d\times d}$. We first verify that the prescribed initialization belongs to Ω and remains in the stated scalar subspace. We then estimate the two terms of the loss separately for every prefix $1 \le j \le k$.

The initialization belongs to the invariant region. Every entry of P equals $1/d$, and hence

$$P^\top = P, \qquad P^2 = P, \qquad \|P\|_F^2 = \sum_{a,b=1}^{d}\frac{1}{d^2} = 1.$$

For the initialization in (9),

$$U_0 = hP, \qquad V_0 = (h^2 - \chi h^{2m})P, \qquad R_0 = V_0 - U_0^2 = -\chi h^{2m}P.$$

Thus

$$\|U_0\|_F = h \le r_0, \qquad \|R_0\|_F = \chi h^{2m} = \chi\|U_0\|_F^{2m}.$$

By (50) and (63), $4\chi r_0^{2m} \le 1/2$ and $\epsilon = r_0/4$. Since $\langle U_0^2, R_0\rangle_F = -\chi h^{2m+2}$,

$$\mathcal{E}(U_0, V_0) = h^2 + 4\chi h^{2m+2} = h^2(1 + 4\chi h^{2m}) \le \frac{3}{2}h^2 \le \frac{3}{8}\epsilon^2 < \epsilon^2.$$

Therefore all three constraints of (7) hold at $w_0 = (U_0, V_0)$. The forward invariance already proved implies $w_t = (U_t, V_t) \in \Omega$ for every $t \ge 0$. In particular, $\|U_t\|_F \le r_0 \le 1$.

The equal-entry choice of P gives, for every integer $p \ge 1$,

$$(P^{\odot p})_{ab} = d^{-p} = d^{1-p}P_{ab}, \qquad P^{\odot p} = d^{1-p}P.$$

Consequently, at an arbitrary pair $U = uP$, $V = vP$,

$$\begin{aligned} U^{\odot(2m+1)} &= d^{-2m}u^{2m+1}P, \\ U^{\odot(4m+1)} &= d^{-4m}u^{4m+1}P, \\ R &= (v - u^2)P, \\ U^\top R + RU^\top &= 2u(v - u^2)P. \end{aligned}$$

Substituting these expressions into (F.2), and using $\alpha = \beta d^{2m}$ and $\zeta = (2m+1)(\beta d^{2m}/(2m+2))^2$, gives

$$\begin{aligned} g_U(uP, vP) &= \left[u - \beta u^{2m+1} + (2m+1)\left(\frac{\beta}{2m+2}\right)^2 u^{4m+1} - \frac{u}{2}(v - u^2)\right]P, \\ g_V(uP, vP) &= \frac{1}{4}(v - u^2)P. \end{aligned}$$

Both components are scalar multiples of P. Thus the simultaneous update $(U, V) - 2(g_U, g_V)$ preserves the form (uP, vP), without any projection. By induction, $U_t = u_t P$ and $V_t = v_t P$ for all t. Moreover,

$$\sigma_m(uP) = \left(u - \frac{\beta}{2m+2}u^{2m+1}\right)P.$$

Since $\|P\|_F = 1$, the restriction of the objective is exactly

$$f_m(u, v) := F_m(uP, vP) = \frac{1}{2}\left(u - \frac{\beta}{2m+2}u^{2m+1}\right)^2 + \frac{1}{8}(v - u^2)^2. \tag{66}$$

The map $(u, v) \mapsto (uP, vP)$ is isometric because $\|(uP, vP)\|^2 = u^2 + v^2$; therefore this restriction represents the actual full GD trajectory, not a differently scaled scalar update.

Now fix $j \in \{1, \dots, k\}$ and define

$$\Phi(W) := \frac{1}{2}\|\sigma_m(W)\|_F^2, \qquad \overline{U}_j := (-1)^j U_j, \qquad Z_j := \overline{U}_j - U_0 \in \mathbb{R}^{d\times d}.$$

Because σ_m is odd, $\Phi(-W) = \Phi(W)$, so that $\Phi(U_j) = \Phi(\overline{U}_j)$ for both parities of j. The purpose of $\overline{U}_j$ is to replace the large sign reversal by the small displacement Z_j in the Taylor calculation. It need not itself be a GD iterate.

At the initial point,

$$\alpha U_0^{\odot(2m+1)} = \beta h^{2m+1}P, \qquad U_0^\top R_0 + R_0 U_0^\top = -2\chi h^{2m+1}P.$$

Thus (59) yields

$$\begin{aligned} Z_j &= -2j\alpha U_0^{\odot(2m+1)} - a_j(U_0^\top R_0 + R_0 U_0^\top) + E_j \\ &= -2(j\beta - a_j\chi)h^{2m+1}P + E_j, \qquad \|E_j\|_F \le \mathsf{D}h^{2m+3}. \end{aligned}$$

Independently, (57) gives

$$\|Z_j\|_F = \|U_j - (-1)^j U_0\|_F \le \mathsf{U}h^{2m+1}.$$

These are the two displacement estimates used below: the first identifies the leading term, while the second bounds the Taylor error. Taking the inner product with $U_0 = hP$, we obtain

$$-\langle U_0, Z_j\rangle_F = 2(j\beta - a_j\chi)h^{2m+2} - \langle U_0, E_j\rangle_F. \tag{67}$$

The activation-loss difference and its Taylor error. Differentiating Φ entrywise gives

$$\nabla\Phi(W) = W - \alpha W^{\odot(2m+1)} + \zeta W^{\odot(4m+1)},$$

$$\nabla^2\Phi(W)[X,X] = \sum_{a,b}\left[1 - (2m+1)\alpha W_{ab}^{2m} + (4m+1)\zeta W_{ab}^{4m}\right] X_{ab}^2.$$

The Hessian is diagonal in the matrix-entry coordinates. Hence, for $\|W\|_F \le r_0 \le 1$,

$$\|\nabla^2\Phi(W)\|_2 \le 1 + (2m+1)\alpha r_0^{2m} + (4m+1)\zeta r_0^{4m} \le 1 + H_0 r_0^2 \le \frac{4}{3}.$$

Here $2m \ge 2$ and $4m \ge 2$, H_0 is defined in (F.2), and $r_0 \le \rho_{\mathrm{inv}} \le \delta_H$ implies $H_0 r_0^2 \le 1/6$ by (50) and (63). For $0 \le \theta \le 1$, forward invariance gives

$$\|U_0 + \theta Z_j\|_F = \|(1-\theta)U_0 + \theta\overline{U}_j\|_F \le (1-\theta)h + \theta\|U_j\|_F \le r_0.$$

Thus this Hessian bound holds on the entire Taylor segment, even though we do not require that segment to be contained in Ω.

Taylor's theorem with integral remainder now gives the exact identity

$$\Phi(U_j) = \Phi(\overline{U}_j) = \Phi(U_0) + \langle\nabla\Phi(U_0), Z_j\rangle_F + T_j,$$
$$T_j := \int_0^1 (1-\theta)\nabla^2\Phi(U_0 + \theta Z_j)[Z_j, Z_j]\, d\theta.$$

Therefore

$$|T_j| \le \frac{4}{3}\|Z_j\|_F^2 \int_0^1 (1-\theta)\, d\theta = \frac{2}{3}\|Z_j\|_F^2 \le \frac{2}{3}\mathsf{U}^2 h^{4m+2}.$$

Furthermore, for integers $p \ge 1$, $\|W^{\odot p}\|_F \le \|W\|_F^p$. Applying this to the formula for $\nabla\Phi(U_0)$ and using $h \le 1$ yields

$$\|\nabla\Phi(U_0) - U_0\|_F \le \alpha h^{2m+1} + \zeta h^{4m+1} \le (\alpha+\zeta)h^{2m+1}.$$

Splitting the gradient into $U_0 + (\nabla\Phi(U_0) - U_0)$ and using (67), we consequently have

$$\Phi(U_0) - \Phi(U_j) - 2(j\beta - a_j\chi)h^{2m+2} = -\langle U_0, E_j\rangle_F - \langle\nabla\Phi(U_0) - U_0, Z_j\rangle_F - T_j.$$

Cauchy–Schwarz bounds the three terms on the right by

$$\mathsf{D}h^{2m+4}, \qquad (\alpha+\zeta)\mathsf{U}h^{4m+2}, \qquad \frac{2}{3}\mathsf{U}^2 h^{4m+2},$$

respectively. Since $h^{4m+2} \le h^{2m+4}$ for $m \ge 1$ and $h \le 1$, this proves

$$\left|\Phi(U_0) - \Phi(U_j) - 2(j\beta - a_j\chi)h^{2m+2}\right| \le \left[\mathsf{D} + (\alpha+\zeta)\mathsf{U} + \frac{2}{3}\mathsf{U}^2\right] h^{2m+4}. \tag{68}$$

Define $\widehat{R}_j := R_j - 2^{-j}R_0 \in \mathbb{R}^{d\times d}$. Equation (57) gives

$$R_j = 2^{-j}R_0 + \widehat{R}_j, \qquad \|\widehat{R}_j\|_F \le \mathsf{R}h^{2m+2}.$$

Expanding the square exactly,

$$\|R_j\|_F^2 = 4^{-j}\|R_0\|_F^2 + 2^{1-j}\langle R_0, \widehat{R}_j\rangle_F + \|\widehat{R}_j\|_F^2.$$

As $\|R_0\|_F^2 = \chi^2 h^{4m}$, it follows that

$$\frac{\|R_0\|_F^2 - \|R_j\|_F^2}{8} - \frac{1-4^{-j}}{8}\chi^2 h^{4m} = -\frac{2^{-j}}{4}\langle R_0, \widehat{R}_j\rangle_F - \frac{1}{8}\|\widehat{R}_j\|_F^2.$$

Thus Cauchy–Schwarz and $2^{-j} \le 1$ give

$$\left| \frac{\|R_0\|_F^2 - \|R_j\|_F^2}{8} - \frac{1-4^{-j}}{8}\chi^2 h^{4m} \right| \le \frac{\chi \mathsf{R}}{4} h^{4m+2} + \frac{\mathsf{R}^2}{8} h^{4m+4} \le \left(\frac{\chi \mathsf{R}}{4} + \frac{\mathsf{R}^2}{8} \right) h^{2m+4}. \tag{69}$$

The last inequality again uses $m \ge 1$ and $h \le 1$.

The full loss difference. Since $F_j = \Phi(U_j) + \|R_j\|_F^2/8$, the preceding exact identities add to

$$\begin{aligned} F_0 - F_j &= 2(j\beta - a_j\chi)h^{2m+2} + \frac{1-4^{-j}}{8}\chi^2 h^{4m} + \mathcal{L}_j, \\ |\mathcal{L}_j| &\le M_F h^{2m+4}, \qquad 1 \le j \le k. \end{aligned} \tag{70}$$

Explicitly, the scalar remainder is

$$\mathcal{L}_j = -\langle U_0, E_j \rangle_F - \langle \nabla\Phi(U_0) - U_0, Z_j \rangle_F - T_j - \frac{2^{-j}}{4}\langle R_0, \widehat{R}_j \rangle_F - \frac{1}{8}\|\widehat{R}_j\|_F^2.$$

Adding the bounds (68) and (69) gives exactly

$$M_F = \mathsf{D} + (\alpha + \zeta)\mathsf{U} + \frac{2}{3}\mathsf{U}^2 + \frac{\chi \mathsf{R}}{4} + \frac{\mathsf{R}^2}{8},$$

which is the constant already fixed in (62). The same bound holds for all $1 \le j \le k$, including odd j.

Every shorter prefix has loss above F_0. For $1 \le j < k$, (F.2) gives

$$2(j\beta - a_j\chi) \le -j\nu, \qquad \frac{\chi^2}{8} \le \frac{\nu}{2}.$$

Also, $0 < 1 - 4^{-j} < 1$, and $h^{4m} \le h^{2m+2}$ because $m \ge 1$ and $h \le 1$. Therefore

$$\begin{aligned} F_0 - F_j &\le -j\nu h^{2m+2} + \frac{\chi^2}{8} h^{4m} + M_F h^{2m+4} \\ &\le \left(-j\nu + \frac{\nu}{2} + M_F h^2 \right) h^{2m+2}. \end{aligned}$$

By (62), $M_F \le M$. The choice of radius in (63) gives $Mh^2 \le Mr_0^2 \le \nu/4$. Hence

$$F_0 - F_j \le \left(-j\nu + \frac{\nu}{2} + \frac{\nu}{4} \right) h^{2m+2} \le -\frac{\nu}{4} h^{2m+2} < 0, \qquad 1 \le j < k.$$

We use the length-k certificate already established on Ω. To make its numerical lower bound explicit at w_0, note that

$$\alpha \sum_{a,b} |(U_0)_{ab}|^{2m+2} = \beta h^{2m+2}, \quad \langle U_0^2, R_0 \rangle_F = -\chi h^{2m+2}, \quad \|R_0\|_F^2 = \chi^2 h^{4m}.$$

Thus (64), which applies because k is even, gives

$$\begin{aligned} \mathcal{C}_{4/3}^{(k)}(w_0) &\ge (k\beta - a_k\chi)h^{2m+2} - b_k\chi^2 h^{4m} - M_C h^{2m+4} \\ &\ge \left(k\beta - a_k\chi - b_k\chi^2 - M_C h^2 \right) h^{2m+2} \\ &= \left(\mu - M_C h^2 \right) h^{2m+2} \ge \frac{\mu}{2} h^{2m+2} > 0. \end{aligned}$$

In the second line we used $b_k > 0$ and $h^{4m} \le h^{2m+2}$; the last line uses the definition of μ in (43), $M_C \le M$, and $Mh^2 \le Mr_0^2 \le \mu/2$ from (63). The points w_0, w_k lie in Ω, and the Hessian-bounded ball containing Ω is convex. Hence the bound in (49) holds on $[w_0, w_k]$. Applying (2) with $\eta = 2$ yields

$$F_0 - F_k \ge 2\mathcal{C}_{4/3}^{(k)}(w_0) \ge \mu h^{2m+2} > 0.$$

We have therefore proved the nonasymptotic bounds

$$F_j - F_0 \ge \frac{\nu}{4} h^{2m+2} \quad (1 \le j < k), \qquad F_0 - F_k \ge \mu h^{2m+2}. \tag{71}$$

In particular, the first loss value strictly below F_0 occurs exactly at step k. This compares each prefix endpoint with F_0; it does not assert that the loss increases at every individual step before k.

Finally, for $1 \le j < k$, let $L' > 0$ be any valid Hessian bound on a neighborhood of $[w_0, w_j]$. The same Taylor inequality gives

$$\mathcal{C}_{L'}^{(j)}(w_0) \le \frac{F_0 - F_j}{\eta} \le -\frac{\nu}{8} h^{2m+2} < 0.$$

Thus every shorter valid Taylor certificate is negative at the specified initializations. This obstruction comes from the actual loss increase relative to F_0, not from the particular choice $L = 4/3$.

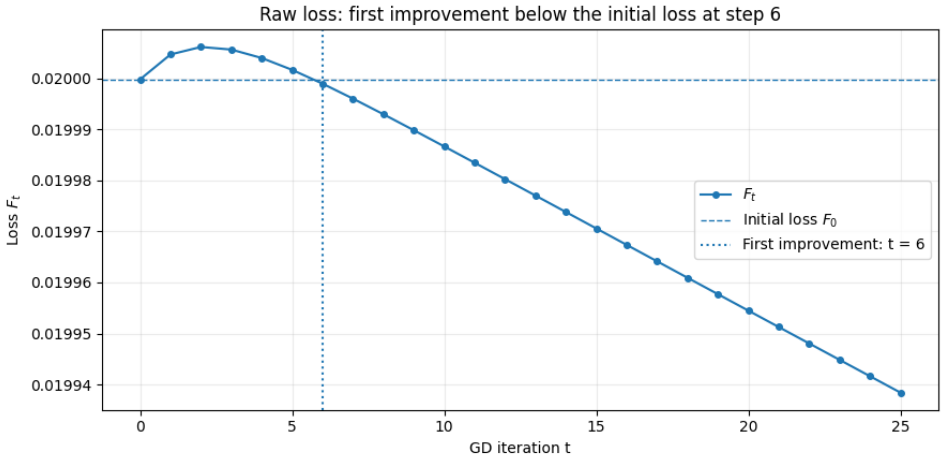


Figure 7: Loss value increases for a few steps at the beginning of training and then starts decreasing monotonically

Spectral sharpness stays strictly above the threshold. Along the invariant scalar subspace, write $U_t = u_t P$, $V_t = v_t P$, and $s_t := v_t - u_t^2$. Since $\|P\|_F = 1$, forward invariance of Ω gives

$$|u_t| = \|U_t\|_F \le r_0 \le 1, \qquad |s_t| \le \chi |u_t|^{2m}.$$

Moreover, the reverse triangle inequality in $U_{t+1} = -U_t + \Delta_t$ and (51) imply

$$\begin{aligned} |u_{t+1}| &\ge |u_t| - \|\Delta_t\|_F \ge |u_t|\big(1 - D_0 |u_t|^{2m}\big) \\ &\ge \left(1 - \frac{1}{8m}\right) |u_t|. \end{aligned}$$

Here $D_0 r_0^{2m} \le 1/(8m)$ follows from (50) and (63). Since $u_0 = h > 0$, induction gives $u_t \neq 0$ at every finite iteration.

The direction $(P, 0)$ has unit norm in the full parameter space. The Rayleigh quotient and differentiation of (66) therefore give

$$\begin{aligned} \kappa_t &\ge \nabla^2 F_m(U_t, V_t)[(P,0),(P,0)] = \partial_{uu}^2 f_m(u_t, v_t) \\ &= 1 + u_t^2 - (2m+1)\beta u_t^{2m} - \frac{s_t}{2} + (4m+1)(2m+1)\left(\frac{\beta}{2m+2}\right)^2 u_t^{4m} \\ &\ge 1 + u_t^2 - \left((2m+1)\beta + \frac{\chi}{2}\right) |u_t|^{2m} \\ &= 1 + \left[1 - \left((2m+1)\beta + \frac{\chi}{2}\right) |u_t|^{2m-2}\right] u_t^2. \end{aligned} \tag{72}$$

In the inequality we used $-s_t/2 \ge -|s_t|/2$ and discarded only the nonnegative term containing u_t^{4m}. We bound the expression in brackets separately for $m = 1$ and for every integer $m \ge 2$.

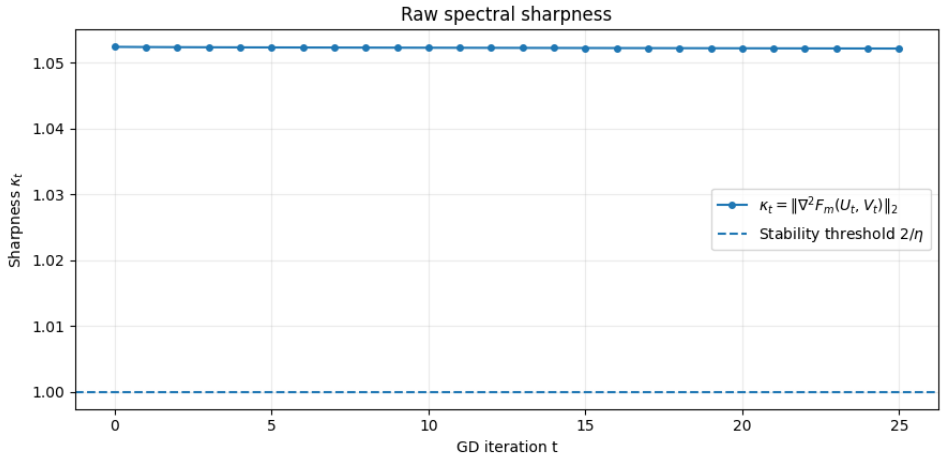


Figure 8: Sharpness is constantly above the threshold and converges slowly to it.

Case $m = 1$. The negative correction is of the same order as u_t^2, so we control its coefficient directly. By (44), $\beta < \chi/3$ and $\chi \le 1/3$, whence

$$3\beta + \frac{\chi}{2} < \chi + \frac{\chi}{2} = \frac{3\chi}{2} \le \frac{1}{2}.$$

Thus (72) gives $\kappa_t \ge 1 + \frac{1}{2}u_t^2$.

Case $m \ge 2$. Now the negative correction has higher order than u_t^2. The definition in (F.2) gives

$$H_0 = (2m+1)\alpha + (4m+1)\zeta + \frac{3}{2} \ge (2m+1)\beta + \frac{\chi}{2},$$

since $\alpha = \beta d^{2m} \ge \beta$, $\zeta \ge 0$, and $\chi \le 1/3$. Furthermore, $m \ge 2$ and $0 < r_0 \le 1$ imply

$$r_0^{2m-2} \le r_0^2.$$

This is equality when $m = 2$, and the same inequality holds for all $m > 2$ because the exponent is larger. Using the radius bounds $r_0 \le \delta_H/[2(2+\chi)]$ and $H_0\delta_H^2 \le 1/6$, from (F.2), (50), and (63), we obtain

$$\begin{aligned}\left((2m+1)\beta + \frac{\chi}{2}\right)|u_t|^{2m-2} &\le \left((2m+1)\beta + \frac{\chi}{2}\right) r_0^{2m-2} \\ &\le H_0 r_0^{2m-2} \le H_0 r_0^2 \\ &\le \frac{H_0\delta_H^2}{4(2+\chi)^2} \le \frac{1}{24(2+\chi)^2} < \frac{1}{2}.\end{aligned}$$

Therefore $\kappa_t \ge 1 + \frac{1}{2}u_t^2$ also holds for every integer $m \ge 2$. The constants H_0, δ_H, and r_0 depend on the chosen m; no radius uniform over all m is assumed.

Combining the two cases and using $u_t \ne 0$ gives

$$\kappa_t \ge 1 + \frac{1}{2}u_t^2 > 1 = \frac{2}{\eta}, \qquad t \ge 0. \tag{73}$$

Finally, the convergence $w_t \to (0,0)$ was proved using (55), and (48) gives

$$\nabla^2 F_m(0,0) = \mathrm{diag}(I_{d^2}, I_{d^2}/4).$$

Continuity of the Hessian and its operator norm therefore imply $\kappa_t \to 1 = 2/\eta$.

The two parameter choices have distinct roles: β/χ compares the linearly accumulating damping $j\beta$ with the bounded mismatch contribution $a_j\chi$, while $\chi \le 4\delta$ also controls the loss released by the mismatch term at shorter prefixes. The exponent m sets the common scale h^{2m+2} by pairing the activation correction with the tube width $\|R\|_F \le \chi\|U\|_F^{2m}$. □

G FULL-BATCH EXPERIMENTS

G.1 EXPERIMENTAL SETUP

We measure cumulative margins and trajectory-local PL quotients on recorded full-batch GD trajectories. Table 2 distinguishes these sweeps from the minibatch experiments in Appendix H. All conclusions concern the stated objectives, selection rules, and observed windows.

Table 2: Experiment inventory. Full-batch SST-2 and minibatch SST-2 are separate campaigns. Checkpoint measurements and the fixed-batch OPEN-1B intervention also have distinct verification scopes.

Experiment	Data and optimizer	Declared coverage and termination
CIFAR–10k, ViT/MLP	Balanced 10,000 examples; full-batch GD, CE/MSE	442 histories; 436 completed, 403 eligible windows; 80,000-update cap and rolling accuracy rule
Tiny ImageNet, CNN/ViT	10,000/100,000 examples; full-batch GD, CE/MSE	120 runs; 116 finite; first 95% accuracy or 6,000/4,000 updates
SST-2 full-batch	First 1,000 training examples; GD, CE/MSE	64 runs; all attain 95% before 20,000 updates
SST-2 minibatch	All 67,349 training examples; SGD/Adam, CE/MSE	128 original runs, 117 complete at 10,000 updates; separate 32-run SGD/MSE study at 20,000
OPEN-1B checkpoints	Published pre-training batches and AdamW states	60 single-update replays; all target hashes match
OPEN-1B continuation	One fixed batch from checkpoint 42,000; AdamW	183 consecutive updates; final 33-update segment is a certified stopped prefix

Image models and objectives. The CIFAR sweep uses a four-block, width-64, eight-head ViT with 208,074 parameters and a capacity-matched tanh MLP with two width-66 hidden layers and 207,910 parameters. Both use the same balanced CIFAR–10 subset, constant learning rates, and seeds 1–11. There are ten rates per loss, from 0.004 to 0.026 for CE and 0.008 to 0.048 for MSE. The Tiny ImageNet CNN has 1,012,296 parameters and uses a balanced 10,000-example subset; its ViT has 5,427,080 parameters and uses all 100,000 training examples, with 200 classes and 64×64 inputs. Its six rates per loss are

$$\begin{aligned} \text{CNN–CE: } & 0.48, 0.52, 0.56, 0.60, 0.65, 0.70, \\ \text{CNN–MSE: } & 1.20, 1.24, 1.28, 1.32, 1.36, 1.40, \\ \text{ViT–CE: } & 1.28, 1.44, 1.60, 1.76, 1.92, 2.08, \\ \text{ViT–MSE: } & 0.015, 0.0158, 0.016, 0.0165, 0.017, 0.018. \end{aligned}$$

Seeds 1–5 are shared across rates within each Tiny ImageNet architecture. Different training-set sizes make this an observational comparison, rather than an architecture-only ablation. Image MSE is half the sum of squared logit errors against one-hot targets, averaged over examples; SST-2 categorical MSE omits the factor $1/2$. CE is averaged over examples. Loss normalizations and frozen configurations are retained in the reproduction materials.

The illustrative trajectory in Figure 1 is separate: a width-200, two-layer tanh MLP trained on all 50,000 CIFAR–10 examples with MSE, $\eta = 0.01$, and seed 1. Its measurements have a 20-update cadence; the displayed interval is $[10\,000, 20\,000]$, with inset $[19\,000, 20\,000]$. Shading begins at $t = 14\,080$, the first recorded *spectral* crossing $\lambda_1 \geq 2/\eta$.

Full-batch SST-2. The first 1,000 examples in canonical GLUE SST-2 training order are encoded once with the frozen bert-base-uncased tokenizer at length 64. The 2,008,066-parameter Song-style Transformer (Song et al., 2025) has two width-64 post-residual-LayerNorm blocks, eight heads, width-64 Swish feed-forward layers, unmasked mean pooling, and a zero-initialized two-class head. Four learning rates and eight shared initialization seeds per loss give 64 runs. GD has no momentum,

clipping, weight decay, or stochastic layers. All runs attain 95% training accuracy before the 20,000-update cap: CE stops at 2,469–6,904 updates and MSE at 3,158–11,527. Validation affects neither selection nor stopping. These variable-horizon runs supply the full-batch SST-2 results below; the separate fixed-horizon campaign is not pooled with them.

Stopping, exclusions, and coverage. For CIFAR, the prescribed criterion is a mean training accuracy of 95% over five readings taken every 50 updates, with an 80,000-update cap. The inventory contains 436 completed histories (216 ViT, 220 MLP) and six crashed ViT histories. All 436 completed runs satisfy the rolling accuracy criterion before the cap. All 216 completed ViT runs and all 110 MLP–MSE runs enter the averaged-directional EoS region. For MLP–CE, none of the 22 runs at $\eta = 0.004, 0.005$ crosses, and only 6/11 at $\eta = 0.0065$ cross; the entire latter condition is excluded. The seven larger rates retain 77 runs, giving 403 eligible CIFAR windows. For Tiny ImageNet, all 60 MSE runs reach the update cap, whereas all 56 finite CE runs reach 95% training accuracy. Four original CE runs terminate after nonfinite proposed updates: CNN at $\eta = 0.65$, seed 4, and ViT at $\eta = 1.76, 1.92, 2.08$, seed 5. They remain in the 120-run declared denominator and are excluded from the 116 eligible windows. Subsequent reruns are separate executions and do not alter these counts, seed means, fits, or confidence summaries.

Measurements and observation windows. The dense image records contain per-update loss L_t, full-gradient norm, and averaged directional sharpness $\tilde{\lambda}^t$; gradient energy is the square of the recorded norm. The 436 completed CIFAR files have consecutive integer steps and 6,106,036 rows. At 122,548 shared steps, loss, norm, and directional sharpness agree with the separate 50-update logs. Every retained Tiny ImageNet window also has complete per-update coverage. Full-batch SST-2 records same-objective loss pairs and full gradient norms, without a Hessian-eigenvalue diagnostic. These are recorded floating-point quantities, not exact arithmetic evaluations.

Let T_r be the final logged finite state (for SST-2, the loss after the last completed update supplies that state), and set

$$t^{(r)}_{\mathrm{EoS}} := \inf\{t : m^{(r)}_t \le 0\} = \inf\{t : \tilde{\lambda}^t_{(r)} \ge 2/\eta_r\}, \qquad s_r := t^{(r)}_{\mathrm{EoS}} + \left\lceil 0.05(T_r - t^{(r)}_{\mathrm{EoS}}) \right\rceil . \quad (74)$$

The primary window is $[s_r, T_r)$; the prespecified 5% burn-in depends on the realized terminal horizon. The separate, post hoc restriction $\rho = (t - s_r)/(T_r - s_r) \in [0.05, 1]$ applies to display and power-law fitting. Early oscillations produce nonpositive cumulative prefixes: in the two conditions in Figure 2, the last such seed-level prefixes occur at approximately 4.16% (MLP) and 2.70% (ViT) of the primary window. Restricting fitted endpoints does not restart the cumulative sum at $\rho = 0.05$. Confidence summaries treat seeds, not updates, as independent units. One-sided 95% seed-bootstrap bounds on condition means are distinct from the pointwise lower curves used in the window-sensitivity fits.

Computation and reproduction. CIFAR uses CUDA, TF32, and `torch.compile` in `reduce-overhead` mode; the archived configurations do not identify its GPU model. Tiny ImageNet uses bfloat16 parameters with PyTorch 2.9.1 and CUDA 12.8 on RTX 4090 and L40S GPUs. The full-batch SST-2 campaign uses A100 GPUs. Configurations, available per-run summaries, plotting inputs, and commands are supplied in `reproduction/experimental_appendix/README.md`; the historical block and full-batch SST-2 summaries are distinguished there from analyses reproducible from the recovered dense records.

G.2 Cumulative margins in full-batch GD

For each retained run and $s_r < t \le T_r$, the cumulative margin is

$$M^{g,(r)}_{s_r:t} := \frac{\sum_{u=s_r}^{t-1} \|g^{(r)}_u\|^2 m^{(r)}_u}{\sum_{u=s_r}^{t-1} \|g^{(r)}_u\|^2}. \quad (75)$$

The GD identity (1) relates its sign to the net loss change. All 403 eligible CIFAR windows and all 116 eligible original Tiny ImageNet windows have positive terminal margins. The minima supported by the per-run analysis are, for CE and MSE respectively, 4.85×10^{-4} and 2.95×10^{-4} (CIFAR ViT), 1.10×10^{-3} and 8.86×10^{-5} (CIFAR MLP), 1.15×10^{-3} and 3.16×10^{-4} (Tiny ImageNet

ViT), and 9.99×10^{-4} and 3.10×10^{-3} (Tiny ImageNet CNN). All 64 full-batch SST-2 runs also enter the operational EoS window and have positive terminal margins; their minima are 1.16×10^{-3} for CE and 7.23×10^{-4} for MSE. These assertions exclude the failed or ineligible records specified above.

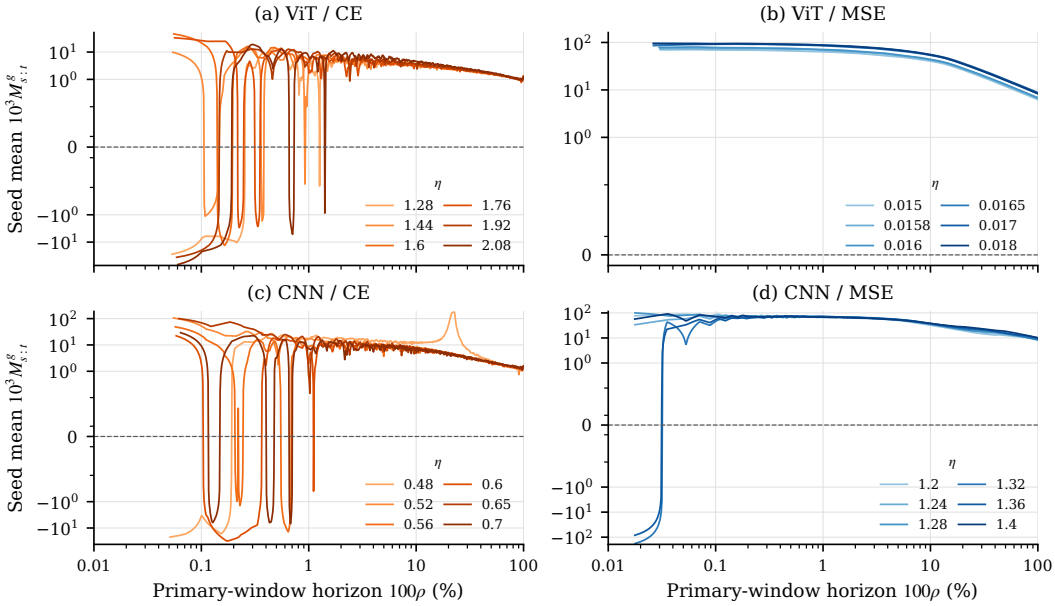


Figure 9: Tiny ImageNet cumulative margins for both architectures and losses. Each curve is a seed mean at one of the six retained learning rates; four or five eligible original seeds contribute to each mean. The signed ordinate retains the early nonpositive behavior, while every terminal mean is positive. The horizon axis is logarithmic and the ordinate is symmetric logarithmic with a linear region near zero. All prefixes accumulate from the same run-specific primary-window start.

To describe the cumulative profiles, eligible seed curves at each rate are interpolated onto 200 normalized horizons in $[0.002, 1]$. Ordinary least squares in log coordinates fits the positive seed means on $\rho \in [0.05, 1]$ as $\overline{M}_\eta(\rho) \approx a_\eta \rho^{-\widehat{\beta}_\eta}$. All 61 image fits are available and use all 190 fitting points; no seed-mean point on this restricted support is nonpositive. The fitted coefficient is the predicted value at $\rho = 1$, rather than the measured terminal margin. Table 3 consolidates the results; Figure 2 supplies the two main-text examples without repeating them here.

For a fixed fitted exponent, replacing a_η by

$$c_\eta := \min\left\{a_\eta, \min_{i:\, \rho_i \in [0.05,1]} \overline{M}_\eta(\rho_i)\rho_i^{\widehat{\beta}_\eta}\right\} > 0$$

preserves that exponent and gives a lower envelope of the positive *seed means at the fitted grid points*. It neither bounds every individual trajectory nor supplies a uniform continuation to unobserved horizons. An analogous envelope for an individual trajectory must use its own positive measured prefixes. Convergence rates beyond the recorded interval require the corresponding margin and PL hypotheses to persist. In particular, the Tiny ImageNet CNN–CE exponent range includes 1.347; a fit with $\widehat{\beta} \geq 1$ does not give a decaying $N^{-(1-\widehat{\beta})}$ bound. The fitted exponents are descriptive, not universal asymptotic laws.

Table 3: All eight image conditions on their primary windows. Counts n refer to eligible original trajectories. Fits count learning-rate-level seed-mean fits; exponent ranges and median log–log R^2 use these fits. Each $\widehat{\mu}_r$ is first minimized over every recorded update of one run, then the range is taken across runs (Appendix G.3).

Dataset / model	Loss	n	fits	$\widehat{\beta}$ range	median R^2
CIFAR–10k / ViT	CE	108	10/10	$[0.542, 0.574]$	0.976
CIFAR–10k / ViT	MSE	108	10/10	$[0.561, 0.721]$	0.984
CIFAR–10k / MLP	CE	77	7/7	$[0.234, 0.584]$	0.978
CIFAR–10k / MLP	MSE	110	10/10	$[0.632, 0.841]$	0.995
Tiny ImageNet / ViT	CE	27	6/6	$[0.516, 0.655]$	0.961
Tiny ImageNet / ViT	MSE	30	6/6	$[0.800, 0.813]$	0.986
Tiny ImageNet / CNN	CE	29	6/6	$[0.736, 1.347]$	0.975
Tiny ImageNet / CNN	MSE	30	6/6	$[0.541, 0.597]$	0.981

Dataset / model	Loss	$\widehat{\mu}_r$: [min, max]
CIFAR–10k / ViT	CE	$[0.0654, 4.42]$
CIFAR–10k / ViT	MSE	$[0.0884, 0.821]$
CIFAR–10k / MLP	CE	$[0.0148, 2.47]$
CIFAR–10k / MLP	MSE	$[0.00200, 2.19]$
Tiny ImageNet / ViT	CE	$[1.10, 24.3]\,10^{-3}$
Tiny ImageNet / ViT	MSE	$[1.56, 2.95]\,10^{-2}$
Tiny ImageNet / CNN	CE	$[4.43\,10^{-6}, 1.14\,10^{-2}]$
Tiny ImageNet / CNN	MSE	$[8.92, 10.0]\,10^{-6}$

G.3 Trajectory-local PL measurements

With the zero-loss reference $L_{\star,r} := 0$ and $L_t^{(r)} > 0$, define

$$\mu_{\mathrm{PL},t}^{(0,r)} := \frac{\|g_t^{(r)}\|^2}{2L_t^{(r)}}, \qquad \widehat{\mu}_r := \min_{s_r \le t < T_r} \mu_{\mathrm{PL},t}^{(0,r)}. \tag{76}$$

Every integer update in the 403 CIFAR and 116 Tiny ImageNet primary windows is recorded; the terminal state T_r is excluded from the minimum. No sampled checkpoints are substituted for intermediate updates. The ranges in Table 3 are ranges of run-specific minima, not minima of seed means. They are all positive, despite substantial variation along each trajectory and between learning rates. Loss normalization changes their numerical scale, so the table is not a scale-invariant comparison of architectures.

Since the objectives are nonnegative, $f(x_t) - f_\star \le L_t$; hence the measured inequality $\|g_t\|^2 \ge 2\widehat{\mu}_r L_t$ is conservative for the loss gap at the recorded iterates. Combining it with the positive measured terminal margin gives the finite-window comparison

$$\frac{1}{T_r - s_r} \sum_{t=s_r}^{T_r-1} (f(x_t^{(r)}) - f_\star) \le \frac{L_{s_r}^{(r)}}{2\eta_r \widehat{\mu}_r (T_r - s_r) M_{s_r:T_r}^{g}}. \tag{77}$$

This is an interpretation of the recorded GD quantities through Theorem 1, not an arithmetic certification of the training computation. The finite minimum gives no common positive constant on later iterates, a neighborhood of the trajectory, or the whole parameter space. It also does not establish the structural block condition.

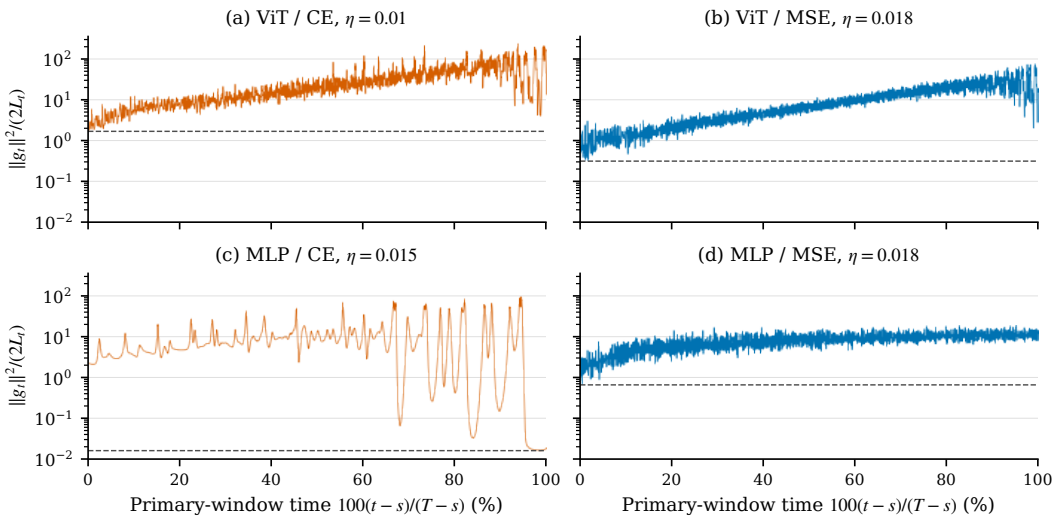


Figure 10: Unsmoothed recorded PL quotients for CIFAR–10k, seed 1: ViT–CE ($\eta = 0.01$), ViT–MSE (0.018), MLP–CE (0.015), and MLP–MSE (0.018). Each panel shows one central retained rate on its normalized primary window; the ordinate is logarithmic. Dashed lines mark the run-specific minima. The zero-loss reference is common to all four panels.

G.4 Sensitivity to the observation window

We retain the prespecified grid with starts at 5%, 10%, and 20% of the post-crossing span and endpoints at 50%, 80%, 90%, 95%, and 100%. For each rate–window cell, seeds supply a one-sided 95% bootstrap lower bound on the mean terminal margin. A separate descriptive fit uses the positive values of $\overline{M}(\rho) - 1.645\,\mathrm{SE}(M(\rho))$ on the 200-point horizon grid, restricted to $\rho \geq 0.05$. This pointwise lower curve is not a simultaneous confidence envelope. A weak window has median log–log $R^2 < 0.8$ across learning rates; weak fits and nonpositive terminal bounds are distinct outcomes.

Table 4: Window sensitivity without selecting a favorable window. The minimum and nonpositive count range over learning-rate–window cells; weak-window counts aggregate fit quality across rates. Dashes denote a fit-quality summary not reported here.

Dataset / model	Loss	smallest lower bound	nonpositive cells	weak windows
CIFAR–10k / ViT	CE	3.52×10^{-4}	0/150	6/15
	MSE	3.00×10^{-4}	0/150	1/15
CIFAR–10k / MLP	CE	8.57×10^{-4}	0/105	8/15
	MSE	7.90×10^{-5}	0/150	2/15
Tiny ImageNet / ViT	CE	6.93×10^{-4}	0/90	11/15
	MSE	8.81×10^{-5}	0/90	0/15
Tiny ImageNet / CNN	CE	3.96×10^{-4}	0/90	10/15
	MSE	2.98×10^{-3}	0/90	5/15
SST-2 full-batch	CE	positive	0/60	—
	MSE	-1.12×10^{-3}	6/60	—

All image terminal bounds remain positive, including windows with weak power-law fits. The six SST-2 MSE failures occur at $\eta = 0.02$, at the 50% and 80% endpoints for each of the three starts. At the 100% endpoint, all SST-2 MSE bounds are positive, with minimum 8.87×10^{-4}. Thus positivity of the primary terminal certificate need not persist on every truncated window. We retain the primary window rather than choose a different start or endpoint from this grid.

G.5 FIXED-LENGTH LOSS CHANGES

This retrospective calculation uses a separate cohort: 51 historical MLP pilot trajectories (21 CE, 30 MSE; seeds 1–3) and the 64 full-batch SST-2 trajectories, hence 115 runs. It does not replace the 403-run CIFAR margin/PL analysis. For every integer start $t \geq s_r$ with $t + B \leq T_r$, we use the recorded endpoint decrease $D_t(B) := L_t - L_{t+B}$ and all B intervening gradient energies. A run is eligible only if $T_r - s_r \geq 20B$. The fixed update lengths are

$$B \in \{1, 2, 3, 4, 5, 8, 9, 16, 17, 32, 33, 64, 65, 128, 129, 256, 257, 512, 513, 1024\}.$$

No missing updates are interpolated. With $\mathcal{G}_t(B) := B^{-1} \sum_{j=0}^{B-1} \|g_{t+j}\|^2 > 0$, the block margin is $D_t(B)/(\eta B \mathcal{G}_t(B))$. A nonpositive $D_t(B)$ precludes a positive lower bound proportional to any nonnegative power of $\mathcal{G}_t(B)$ on all tested starts.

Table 5: Selected block lengths in the separate 115-run cohort. An increase means that at least one eligible block increases the loss. The reproduction materials retain the historical aggregate table; the original per-run block inputs are not included.

Cohort	$B = 1$		$B = 128$	
	eligible	with increase	eligible	with increase
MLP pilot / CE	21	21	20	20
MLP pilot / MSE	30	30	30	30
SST-2 full-batch / CE	32	32	5	5
SST-2 full-batch / MSE	32	32	25	25

Of 1,789 eligible run–length pairs, 1,779 contain a loss increase larger than the numerical sensitivity threshold $10^{-12} + 10^{-7} \max(|L_t|, |L_{t+B}|)$; this threshold leaves the strict sign tests unchanged and is not a training-error bound. The ten remaining pairs all use MLP–CE, $\eta = 0.008$, seed 1, and $B \in \{2, 4, 8, 9, 16, 17, 32, 33, 64, 65\}$. Its window has 1,427 updates, so it becomes ineligible at $B = 128$. All assessed gradient-to-loss block quotients have positive minima, but no pair preserves both first-half extremal constants on the second half. At $B = 32$, changing the burn-in to 20% or 50% leaves all 115 runs eligible and the same sole trajectory positive at every start.

An additional endpoint-only test uses the 50-update CIFAR loss logs, with starts rounded upward after burn-in and exact endpoint spacings. At $B = 50$, 402/403 eligible runs have a material increase; at $B = 100$, all 400 eligible runs do, covering every retained learning rate. Under the domain and Hessian hypotheses of Definition 1, (C_m^B) implies strict loss decrease on each positive-energy block. Thus these increases contradict its necessary descent consequence at the tested starts. Positive full-window margins do not imply this fixed-length condition; passing the endpoint test would not verify its other hypotheses or its validity beyond the measured record.

H MINIBATCH EXPERIMENTS

H.1 SAME-MINIBATCH AND CHANGING-MINIBATCH LOSSES

Using the notation of Section 5, write $L_t = f(x_t, \xi_t)$, $L_t^+ = f(x_{t+1}, \xi_t)$ and $a_t = \langle g_t, x_t - x_{t+1} \rangle$. On positive-work updates, (11) gives

$$L_t^+ - L_t = a_t(\zeta_t/2 - 1), \qquad L_{t+1} > L_t \iff \zeta_t > 2(1 - \varepsilon_{t+1}/a_t). \tag{78}$$

The batch-replacement term need not have mean zero because x_{t+1} depends on ξ_t. For $I_{s:t} = \{u : s \le u < t,\ a_u > 0\}$ and $A_{s:t} = \sum_{u \in I_{s:t}} a_u > 0$, define

$$M_{s:t}^{\mathrm{mb}} := \frac{\sum_{u \in I_{s:t}} (L_u - L_u^+)}{A_{s:t}} = 1 - \frac{\sum_{u \in I_{s:t}} a_u \zeta_u}{2A_{s:t}}. \tag{79}$$

Summing the loss changes yields

$$L_t - L_s = -A_{s:t} M_{s:t}^{\mathrm{mb}} + \sum_{\substack{s \le u < t \\ a_u \le 0}} (L_u^+ - L_u) + \sum_{u=s}^{t-1} \varepsilon_{u+1}. \tag{80}$$

Thus the margin omits nonpositive-work updates and does not alone certify descent of changing-minibatch losses. With one fixed batch and positive work throughout, both additional sums vanish. The Hessian interpretation in Section 5 further requires an actual gradient and a C^2 objective along the update segment.

H.2 SST-2

We use all $67{,}349$ training examples, tokenized to length 64, and the two-block, width-64, eight-head Transformer described in Appendix G.1. Minibatches contain $4{,}096$ examples sampled with replacement. CE is averaged over examples; MSE averages the sum of squared logit errors without a factor $1/2$. SGD has no momentum; Adam uses $(\beta_1, \beta_2) = (0.6, 0.996)$ and $\epsilon = 5 \times 10^{-9}$. Neither uses weight decay or clipping. An 83-update linear warmup precedes constant learning rates. Same-minibatch quantities are recorded at every update; full-training loss and accuracy are evaluated every 500 updates. These L40S runs use float32 training with TF32 enabled and no autocast.

The original 128 runs comprise four rates and eight seeds per optimizer–loss setting: 117 reached the prescribed $10{,}000$ updates, while 11 SGD/MSE runs terminated with nonfinite updates. A separate, post hoc SGD/MSE study completed all 32 runs at four refined rates for $20{,}000$ updates. The 95% training-accuracy target was a logged milestone, not a stopping rule. Of the completed runs, $96/117$ in the original campaign and $27/32$ in the follow-up reached that milestone at a recorded evaluation.

Figure 11 retains seed 0 at the second-lowest rate of each four-rate grid; SGD/MSE uses the separate $20{,}000$-update study. Table 6 reports margins accumulated from update 0, including warmup, with nonpositive-work updates excluded. Their work-weighted terminal ζ averages are closer to 2 for the two displayed SGD runs than for Adam. This comparison concerns these four trajectories. In particular, the displayed SGD/MSE run has no recorded evaluation at 95% training accuracy (maximum 94.39%).

Table 6: Displayed SST-2 trajectories (seed 0). The terminal margin uses all positive-work updates; $N_{\le 0}$ counts omitted updates.

Setting	Target rate	Updates	$M_{0:T}^{\mathrm{mb}}$	$N_{\le 0}$
SGD/CE	0.04	10,000	0.01996	0
SGD/MSE	0.015874	20,000	0.02497	0
Adam/CE	0.001414	10,000	0.13455	16
Adam/MSE	0.001414	10,000	0.14627	7

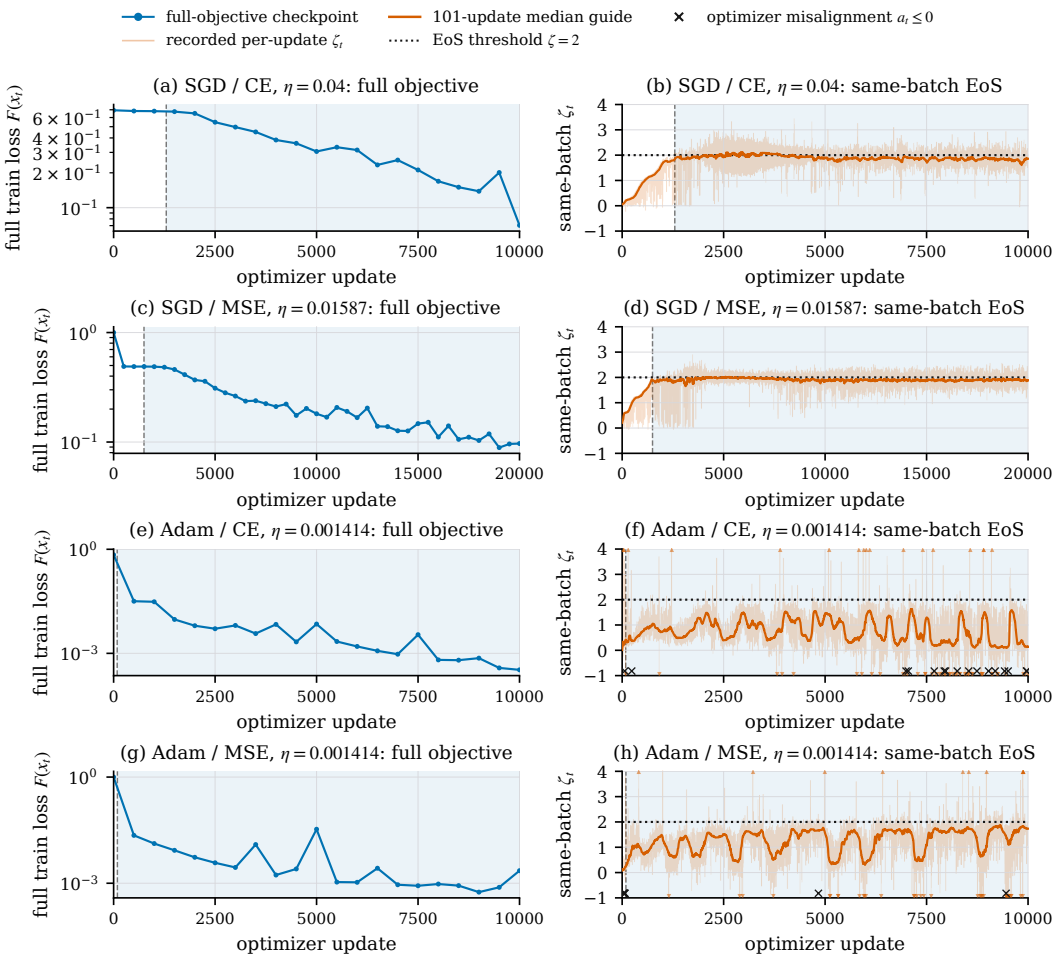


Figure 11: Four SST-2 optimizer–loss settings. Left: full-training loss at 500-update checkpoints. Right: recorded same-minibatch ζ_t and a 101-update median guide. The dotted line marks the same-minibatch loss-change boundary 2 for positive work; crosses mark nonpositive-work updates. Values outside $[-1, 4]$ are marked by edge triangles. Shading begins at the first post-warmup crossing of 2. SGD/MSE uses the separate 20,000-update study.

H.3 OPEN-1B

Measurements and interpretation. We measure the training objective (cross-entropy plus z-loss) and actual AdamW displacement in the published OPEN-1B recipe. Its int8 W8A8 learned step-size quantization uses straight-through (LSQ/STE) backward signals. With the pre-clipping signal g_t^{STE}, the measured work is $a_t = \langle g_t^{\mathrm{STE}}, x_t - x_{t+1} \rangle$ and the reported quantity is

$$\zeta_t^{\mathrm{res}} = 2\left(1 + \frac{L_t^+ - L_t}{a_t}\right), \qquad a_t > 0. \tag{81}$$

The displacement includes decoupled weight decay, which is not added as an L2 loss term. This is an LSQ/STE loss–work residual: surrogate differentiation of the quantized objective does not establish the Hessian-integral interpretation. For positive measured work, $\zeta_t^{\mathrm{res}} > 2$ is algebraically equivalent to $L_t^+ > L_t$; matching loss increases are therefore not independent evidence of curvature or dynamical instability.

Checkpoint replays. We replay one update at each of 60 distinct checkpoints between steps 26,000 and 58,000, selected from published metadata before inspecting the measured residuals. Each replay uses one H200 GPU. These are separated measurements on one published training trajectory. Each

uses its original batch of $4{,}718{,}592$ tokens and the scheduled AdamW rate, which decreases from approximately 3.68×10^{-4} to 1.40×10^{-4} across the interval. All 60 replays completed, had positive measured work, and matched their published target-state hashes; no selected position is omitted. The replay budget is one update per position, with no accuracy stopping criterion.

All 60 residuals lie below 2: their unweighted mean is 0.817, median 0.775, and range $[0.433, 1.656]$. None lies within 0.1 of 2. These dependent, separated observations do not provide seed confidence intervals or a full-trajectory average. They appear in the left panel of Figure 4.

Fixed-batch continuation and verification. Starting at step $42{,}000$, we freeze the canonical batch for the next published update (288 microbatches, $4{,}718{,}592$ tokens) and reuse it for 183 sequential AdamW updates. Four H200 workers accumulate in the canonical virtual-rank order; the scheduled learning rate decreases from 2.553×10^{-4} to 2.539×10^{-4}. There is no accuracy stopping rule: the reported trajectory ends with a certified 33-update prefix of the final planned 50-update segment. Its 184 loss states decrease overall from 2.413 to 0.0188. All measured works are positive. The residual exceeds 2 at zero-based update indices 119, 120, 121, and 123, then remains below 2 through the observed endpoint; see Figure 4. Batch replacement is identically zero, so the four corresponding loss increases follow from (78).

The first update matches the published target hash for step $42{,}001$. Subsequent fixed-batch updates intentionally follow a different trajectory: verification then checks consecutive update indices, chained internal state hashes, and agreement between each post-update and next pre-update loss. These checks establish recorded continuity, not independent published verification of every continuation step. The study measures fitting of one reused training batch and contains no held-out evaluation. The reproduction materials retain the configurations, per-run SST-2 summaries, plotting inputs, and OPEN-1B hash records; see `reproduction/experimental_appendix/minibatch.md`.