ADAM AT THE EDGE OF STABILITY: ADAPTIVE FEEDBACK, PROVABLE OSCILLATION, AND GRADIENT REVERSAL

Anonymous authors
Paper under double-blind review

ABSTRACT

The edge-of-stability (EoS) phenomenon of full-batch Adam has been widely observed, yet its underlying dynamical mechanism remains poorly understood. In this paper, we identify Adam's second-moment adaptation as a negative-feedback mechanism that drives the dynamics toward the stability boundary. We characterize this mechanism through the *active curvature*, namely, the preconditioned curvature along the preconditioned gradient direction, and establish rigorous characterizations in progressively richer settings: rank-one quadratics with momentum, diagonal quadratics, on which the active curvature separates from the sharpness, and general objectives. Importantly, the mechanism predicts *gradient reversal* of full-batch Adam near the edge: consecutive gradients repeatedly point in nearly opposite directions, as we observe across fully connected networks, ResNets, ViTs, LSTMs, GPT-2 medium, and Adam-family optimizers. Consistent with this picture, averaging iterates suppresses these fast oscillations and produces smoother and lower loss curves. Together, these results provide an important first step towards fully understanding the dynamical behavior of Adam's EoS through active curvature and gradient reversal.

1 INTRODUCTION

Adam (Kingma & Ba, 2015) is among the most widely used optimizers in modern deep learning. Despite its ubiquity, however, the behavior of Adam on neural networks and the mechanisms underlying that behavior remain poorly understood. One particularly striking phenomenon is the *edge of stability* (EoS), whereby first-order optimization methods operate near the boundary of dynamical stability. For gradient descent (GD), Cohen et al. (2021) observed that, during neural-network

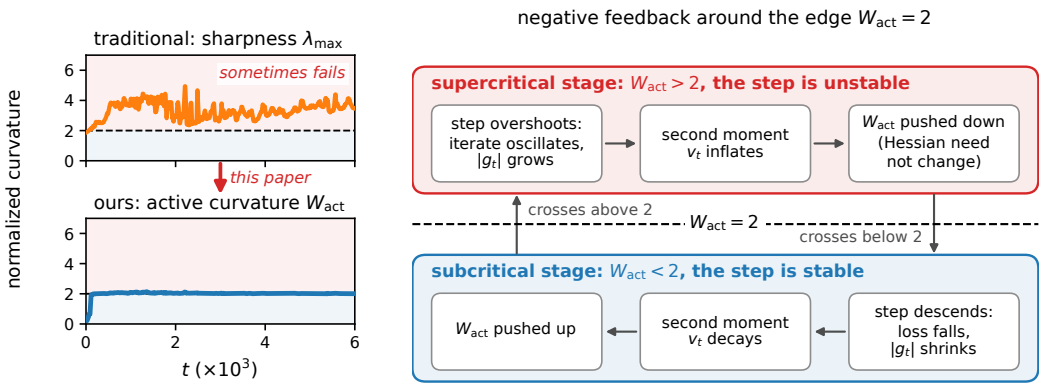

Figure 1: Left: the normalized sharpness (Eq. (4), orange, top) and the active curvature (Eq. (5), blue, bottom) along an Adam training trajectory; the stability boundary of the frozen step is normalized to 2 (dashed). Right: the negative-feedback loop of the second moment that holds the active curvature around the threshold (Section 3). Experimental details are in Appendix B.2.

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.

training, the sharpness $\lambda_{\max}(H)$—the largest eigenvalue of the Hessian $H = \nabla^2 L(\theta)$—increases toward the classical stability threshold $2/\eta$, where η is the learning rate.

For full-batch Adam, Cohen et al. (2022) observed a similar phenomenon: the *preconditioned sharpness* S_t—the largest eigenvalue of the Hessian after Adam's coordinatewise rescaling—increases toward and subsequently remains near the stability threshold of the corresponding linearized dynamics. These observations suggest that EoS is a robust feature of neural-network optimization, yet its underlying mechanism remains largely unexplained. While the EoS phenomenon for gradient descent has received substantial theoretical attention, considerably less is understood for Adam. This leaves a basic question unresolved: *why does full-batch Adam operate near the edge of stability?*

A natural intuition is that Adam's second-moment normalization is a *negative-feedback mechanism* that regulates the optimization dynamics around the stability boundary. In this paper, we formalize this intuition through a theoretical characterization of Adam's dynamics, as illustrated in Figure 1 (right). Specifically, our theory naturally leads to the notion of *active curvature*, $W_{\mathrm{act},t}$, which is closely related to the preconditioned sharpness but measures curvature along the gradient direction. The provable EoS mechanism can be sketched as follows. When $W_{\mathrm{act},t}$ exceeds the stability threshold, the dynamics become unstable: the iterates overshoot, the gradients grow, and Adam's second-moment estimate v_t increases; the resulting stronger preconditioning pushes $W_{\mathrm{act},t}$ downward. When $W_{\mathrm{act},t}$ lies below the stability threshold, the dynamics are stable and contracting: the gradients shrink, v_t decays, and $W_{\mathrm{act},t}$ is pushed upward. Thus, Adam's second-moment adaptation corrects deviations from the stability boundary in both directions, naturally driving the dynamics toward the edge.

Specifically, we establish this feedback mechanism rigorously in progressively richer settings: first on quadratic objectives, which provide the simplest setting in which Adam's momentum and nonlinear preconditioning already lead to highly nontrivial dynamical behaviors, and then, without momentum, on general objectives. Empirically, the mechanism extends further, in line with the broader perspective that quadratic models can capture important aspects of neural-network optimization (Meterez et al., 2026). We test this prediction across a broad range of neural-network training settings, varying architecture, model size, learning rate, momentum, and optimizer variants. Across these settings, whenever Adam operates at the edge, W_{act} remains close to 2, even when the preconditioned sharpness lies far above its nominal stability threshold (Section 3; Appendix B). These results corroborate our explanation that the quantity regulated by Adam is not necessarily the largest curvature of the preconditioned Hessian, but rather the curvature encountered along the active optimization direction.

A further analysis of the dynamics underlying W_{act} suggests another characteristic behavior: *gradient reversal*. Near $W_{\mathrm{act}} = 2$, the dynamics along the active direction have a multiplier close to -1, suggesting that consecutive gradients should repeatedly reverse direction rather than decay smoothly. Guided by these theoretical insights, we examine gradient dynamics in neural-network training and find the predicted reversal behavior across a broad range of neural-network architectures and variants of the Adam family (Section 4.1; Appendix C). Consistent with this oscillatory picture, we further find that averaging the iterates, either by exponential moving average or simple iterate averaging, substantially smooths the loss trajectory and suppresses the oscillations induced by gradient reversal.

In summary, our contributions are twofold.

- **Theory: a dynamical analysis of Adam.** We begin with quadratic objectives and characterize Adam's behavior in both the subcritical ($W_{\mathrm{act},t} < 2$) and supercritical ($W_{\mathrm{act},t} > 2$) regimes. We overcome the challenges posed by momentum and nonlinear preconditioning by viewing Adam as a time-varying linear dynamical system and carefully tracking the evolution of its state-dependent dynamics. This analysis yields a rigorous characterization of the restoring behavior around the stability boundary. To the best of our knowledge, no prior work provides such a characterization of Adam's EoS under standard parameter choices, even on quadratic objectives.
- **Empirics: active curvature and gradient reversal in deep learning.** Across fully connected networks, ResNets, ViTs, LSTMs, GPT-2 medium, and several Adam-family optimizers, we find that W_{act} stays near the stability boundary whenever training operates at the edge, even when the preconditioned sharpness is substantially larger. The theory further predicts *gradient reversal*, which we observe broadly across models and optimizer variants. Consistent with this

oscillatory dynamics, midpoint and exponential moving averages substantially smooth the loss trajectory and often achieve lower loss than the raw iterates.

2 SETUP AND STABILITY QUANTITIES

Adam. Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable, and write $g_t := \nabla f(x_t)$ and $H_t := \nabla^2 f(x_t)$ along a trajectory. We consider the uncorrected (full-batch) Adam recursion

$$\begin{aligned} m_{t+1} &= \beta_1 m_t + (1-\beta_1)\, g_t, \\ v_{t+1} &= \beta_2 v_t + (1-\beta_2)\, g_t \odot g_t, \\ x_{t+1} &= x_t - \eta\, D_t^{-1} m_{t+1}, \qquad D_t := \operatorname{diag}\big(\sqrt{v_{t+1}} + \varepsilon\big), \end{aligned} \tag{1}$$

with $0 \le \beta_1 < 1$, $0 \le \beta_2 < 1$, $\eta > 0$, $\varepsilon > 0$, and initial state $(x_0, m_0, v_0) \in \mathbb{R}^d \times \mathbb{R}^d \times [0,\infty)^d$. Here $\odot$ and the square root act coordinatewise. The recursion preserves $v_t \ge 0$, so $D_t \succeq \varepsilon I$ and every step is well defined.

Frozen stability and normalization. To assess local stability, freeze the second moment D_t. The resulting map on (x, m) is governed by the preconditioned Hessian $D_t^{-1} H_t$, and

$$\mathsf{S}_t := \lambda_{\max}(D_t^{-1} H_t) \tag{2}$$

is the *preconditioned sharpness*. When $H_t \succ 0$, $D_t^{-1} H_t$ is similar to the symmetric matrix $D_t^{-1/2} H_t D_t^{-1/2}$, and the linearized Adam map is Schur stable iff

$$\mathsf{S}_t < \mathsf{S}^\star := \tfrac{2(1+\beta_1)}{\eta(1-\beta_1)} \tag{3}$$

(Appendix D; Lemma D.1). This is the adaptive stability threshold of Cohen et al. (2022).

For convenience, let $c := (1-\beta_1)/(1+\beta_1)$ and normalize the preconditioned Hessian as

$$P_t := c\eta\, D_t^{-1/2} H_t D_t^{-1/2}, \qquad \lambda_{\max}(P_t) = c\eta\, \mathsf{S}_t. \tag{4}$$

The frozen stability boundary is then simply $\lambda_{\max}(P_t) = 2$.

Active curvature. Let $\hat{z}_t := D_t^{-1/2} g_t / \|D_t^{-1/2} g_t\|$ be the preconditioned gradient direction. Then

$$W_{\mathrm{act},t} := \hat{z}_t^\top P_t \hat{z}_t = c\eta\, \frac{g_t^\top D_t^{-1} H_t D_t^{-1} g_t}{g_t^\top D_t^{-1} g_t} \tag{5}$$

is the *active curvature*: the curvature of P_t along the preconditioned gradient. By expanding the unit direction $\hat{z}_t$ in the eigenbasis of P_t, we can regard $W_{\mathrm{act},t}$ as the mean of the eigenvalues under the gradient's own spectral measure, and therefore $\lambda_{\min}(P_t) \le W_{\mathrm{act},t} \le \lambda_{\max}(P_t)$, with equality on the right exactly when the gradient lies in the top eigenspace.

3 ADAM OPERATES NEAR AN ACTIVE STABILITY BOUNDARY

Cohen et al. (2022) observed that along full-batch Adam trajectories the preconditioned sharpness S_t holds roughly at the frozen threshold $\mathsf{S}^\star$, that is, $\lambda_{\max}(P_t) \approx 2$ for large t. While such behavior has long been equated to the EoS, we observe that the sharpness itself may be only *part* of the picture, empirically or theoretically. To capture the full picture of Adam EoS, we provide an in-depth investigation of the preconditioned sharpness and the active curvature in this section.

Motivating experiments. As an illustration, we compare the behavior of preconditioned sharpness S_t (or equivalently, the normalized sharpness $\lambda_{\max}(P_t) = c\eta\mathsf{S}_t$) and the active curvature $W_{\mathrm{act},t}$ on various settings (Figure 2; implementation details in Appendix B.3, more experiments in Appendices B.2–B.4). The sharpness can stay well above the threshold predicted by the linearized stability analysis, particularly when $\beta_1 = 0$. In contrast, the active curvature W_{act} is much more stable and remains close to the threshold for most of training. A natural explanation is that preconditioned sharpness provides a worst-case spectral view and can therefore be dominated by high-curvature directions that carry little gradient mass. Active curvature instead measures the curvature along the direction actually explored by the optimizer, complementing this worst-case perspective (Section 3.2). In Appendix B, we provide additional experiments showing that active curvature more closely tracks the stability of the Adam trajectory and study how these observations are affected by mini-batch noise.

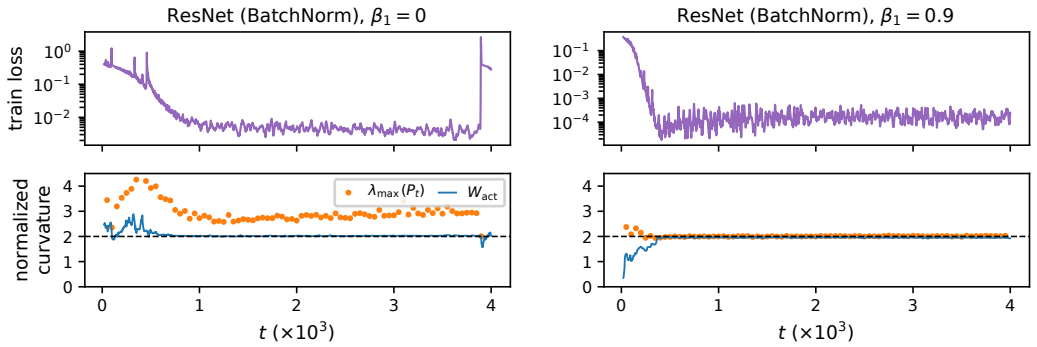

Figure 2: Sharpness and active curvature along the same run. BatchNorm ResNet, full-batch MSE on 1000 CIFAR-10 images, $\eta = 3 \cdot 10^{-4}$, $\beta_2 = 0.999$, 4000 steps, at $\beta_1 = 0$ (left) and $\beta_1 = 0.9$ (right). Training details are in Appendix B.3.

Our theory. The rest of this section explains the mechanism behind Adam's EoS. The roadmap is as follows:

- In Section 3.1 and 3.2, we start with *quadratic* objectives. We first provide rigorous characterization of Adam's dynamics on rank-one quadratics (Section 3.1), thereby establishing the *negative-feedback* loop between the iterate and the curvature. To the best of our knowledge, this is the first rigorous explanation of Adam's EoS with momentum ($\beta_1 > 0$)[1].
- In Section 3.2, we study 2-dimensional diagonal quadratics, where the normalized sharpness can hold well above the threshold while the active curvature provably cannot stay below an explicit cutoff $\overline{W}_{\text{diag}}$ (≥ 1.9 at Adam's default parameters), nor, at the default parameters, above the threshold unless the iterates converge.
- In Section 3.3, we generalize our analysis to general objectives with $\beta_1 = 0$ and argue that the active curvature is indeed a natural measure of stability.
- Finally, in Section 4, we relate the active curvature to the phenomenon of *gradient reversal*, identifying it as a common signature of Adam's EoS.

3.1 Rank one: the feedback loop that pins the active curvature

We begin with the rank-one quadratic, the simplest objective on which the direction that carries the gradient need not be a coordinate axis. On this objective, the edge-of-stability behavior of Adam admits a precise characterization. Gradient descent, by contrast, shows no such behavior, because the curvature of a quadratic is fixed.

Reduction to scalar dynamics. Let $f(x) = \frac{1}{2}\lambda(u^\top x)^2$ with $\lambda > 0$ and $\|u\| = 1$, and write $x_t^\parallel := u^\top x_t$ and $m_t^\parallel := u^\top m_t$. Since $g_t = \lambda x_t^\parallel u$, the standard initialization $m_0 = v_0 = 0$ keeps the momentum in $\operatorname{span}\{u\}$ and the second moment proportional to $u \odot u$, $v_t = \lambda^2 v_t^\parallel \, u \odot u$ for a scalar $v_t^\parallel \geq 0$ (Appendix D). Adam therefore reduces to the scalar system

$$\begin{aligned} m_{t+1}^\parallel &= \beta_1 m_t^\parallel + (1-\beta_1)\lambda x_t^\parallel, \\ v_{t+1}^\parallel &= \beta_2 v_t^\parallel + (1-\beta_2)(x_t^\parallel)^2, \qquad m_0^\parallel = v_0^\parallel = 0, \\ x_{t+1}^\parallel &= x_t^\parallel - \eta\big(u^\top D_t^{-1} u\big) m_{t+1}^\parallel, \end{aligned} \tag{6}$$

while the component of x_t orthogonal to u does not feed back. On this objective the normalized sharpness and the active curvature coincide at every step (Appendix D):

$$W_{\text{act},t} = \lambda_{\max}(P_t) = c\eta\lambda \, u^\top D_t^{-1} u = w_{t+1} := c\eta\lambda \sum_{i=1}^{d} \frac{u_i^2}{\sqrt{v_{i,t+1}} + \varepsilon}. \tag{7}$$

[1]Existing work (Cohen et al., 2025; Bai et al., 2026b) either analyzes Adam at degenerate hyperparameter $\beta_1 = 0$ or adopts heuristics (e.g., central flow).

In particular $0 < w_t \le w_{\max} := c\eta\lambda/\varepsilon$, the value at $v_t^{\parallel} = 0$. Therefore, to understand the EoS behavior of Adam, it is sufficient to study the dynamics of $(x_t^{\parallel}, m_t^{\parallel}, w_t)$. As our starting point, we first point out that the threshold $w = 2$ is not merely a level the trajectory crosses: it is realized exactly, by a 2-cycle of this dynamics on which the active curvature equals 2 at every step.

Proposition 3.1 (2-cycle at the edge)**.** *Assume $w_{\max} > 2$, and let $x_\star > 0$ be the unique value for which the second-moment level $v_t^{\parallel} = x_\star^2$ gives $w_t = 2$. Then*

$$(x_\star,\, -c\lambda x_\star,\, x_\star^2) \;\longleftrightarrow\; (-x_\star,\, c\lambda x_\star,\, x_\star^2)$$

is a 2-cycle of the reduced Adam dynamics of $(x_t^{\parallel}, m_t^{\parallel}, v_t^{\parallel})$, and along it $W_{\mathrm{act},t} = w_{t+1} \equiv 2$ and the gradient reverses at every step.

Illustrative example: RMSProp ($\beta_1 = 0$). As an illustrative example, we start with the case $\beta_1 = 0$, in which Adam reduces to RMSProp (Tieleman & Hinton, 2012). This is the simpler case because $c = 1$ and $m_{t+1} = g_t$, so the recursion reduces to

$$x_{t+1}^{\parallel} = (1 - w_{t+1})\, x_t^{\parallel}. \tag{8}$$

In this setting, we can describe the negative-feedback loop of EoS as follows:

- **Subcritical** regime: For time steps t such that $w_{t+1} \in [\delta, 2-\delta]$, we have $|x_{t+1}^{\parallel}| \le (1-\delta)|x_t^{\parallel}|$, and hence the sequence $(x_t^{\parallel})$ contracts at such steps. Then the second moments (v_t) also shrink, which pushes (w_t) upward.
- **Supercritical** regime: For time steps t such that $w_{t+1} > 2 + \delta$, we have $|x_{t+1}^{\parallel}| \ge (1+\delta)|x_t^{\parallel}|$, and hence the sequence $(x_t^{\parallel})$ blows up exponentially at such steps. Then the second moments (v_t) also blow up, which pushes (w_t) downward.

In other words, for every $\delta > 0$, a trajectory cannot remain indefinitely below $2 - \delta$ or above $2 + \delta$, and the number of steps it can spend on either side is bounded explicitly. Since δ may be chosen arbitrarily close to 0, the result formalizes the restoring behavior toward the edge. The following theorem makes this quantitative, with explicit bounds on the number of steps spent on either side of the threshold; the proof is in Appendix E.[2]

Theorem 3.2 (RMSProp in rank one)**.** *Consider the rank-one model with $w_{\max} > 2$, $\beta_1 = 0$, and $x_t^{\parallel} \neq 0$ for all t. If $w_T < \widehat{w} < 2$ then $w_{T+n} \ge \widehat{w}$ for some $1 \le n \le N_{\mathrm{sub}}$, and if $w_T \ge 2 + \delta$ with $\delta > 0$ then $w_{T+n} < 2 + \delta$ for some $1 \le n \le N_{\mathrm{sup}}$.*

Consequently, for RMSProp, we always have $\liminf_t w_t \le 2 \le \limsup_t w_t$. Our analysis is *non-asymptotic*: N_{sub} and N_{sup} are defined in (35) and (39), and explicit closed-form upper bounds on them are given in (36) and (39).

General case. With $\beta_1 > 0$ the momentum no longer drops out, and we have to consider the two-dimensional dynamics

$$x_{t+1}^{\parallel} = x_t^{\parallel} - (c\lambda)^{-1} w_{t+1}\, m_{t+1}^{\parallel}, \qquad m_{t+1}^{\parallel} = \beta_1 m_t^{\parallel} + (1-\beta_1)\lambda x_t^{\parallel}. \tag{9}$$

The frozen threshold is still $w = 2$, but the presence of momentum introduces complications. In our analysis, in the subcritical regime we need to control the joint evolution of position and momentum; in the supercritical regime the key distinction is whether the two are aligned or misaligned. To this end, we introduce the normalized momentum coordinate $h_t := (c\lambda)^{-1} m_t^{\parallel} + x_t^{\parallel}$. We generalize Theorem 3.2 to this setting as follows.

Theorem 3.3 (Adam in rank one; Proposition F.2, Corollaries F.4 and F.7, Lemma F.6)**.** *Consider the rank-one model with $w_{\max} > 2$, $0 < \beta_1 < 1$ and $\beta_2 > \beta_1^2$.*

- ***Subcritical regime.*** *Write*

$$\overline{W} := \frac{2(\sqrt{\beta_2} - \beta_1)}{\sqrt{\beta_2}(1-\beta_1) - 2\beta_1(1-\sqrt{\beta_2})/w_{\max}}. \tag{10}$$

Then $0 < \overline{W} < 2$, and for any $b < \overline{W}$, if $w_T \le b$ then $w_{T+n} > b$ for some finite n.

[2] In particular, this recovers the asymptotic result of Bai et al. (2026b) on *one-dimensional* quadratics.

- ***Supercritical regime, aligned.*** *Suppose that $x_T^{\parallel} h_T > 0$ and $w_T \geq 2 + \delta$ for some $\delta > 0$. Then $w_{T+n} < 2 + \delta$ for some $1 \leq n \leq \overline{N}_{\sup}$, with $\overline{N}_{\sup}$ explicit in* (89).
- ***Supercritical regime, misaligned.*** *Suppose that $w_t > 2$ and $x_t^{\parallel} h_t < 0$ for all $t \geq T$. Then $\left|x_{t+1}^{\parallel}\right| < \left|x_t^{\parallel}\right| / (w_{t+1} - 1)$ for $t \geq T$, so $\left|x_t^{\parallel}\right|$ decreases at every step although the frozen dynamics remains unstable. In Proposition F.8, we show that such trajectories do exist for suitable initial states (Assumption D.2).*

We briefly discuss each point of Theorem 3.3 as follows. In the subcritical regime, Theorem 3.3 introduces the cutoff $\overline{W} < 2$, below which the contraction is certified. For the standard choice $(\beta_1, \beta_2) = (0.9, 0.999)$ and large $w_{\max}$ it is $\overline{W} \approx 1.991$. Such a cutoff arises because in the presence of momentum, the Adam dynamics is not contracting under a single norm in the regime $w_t < 2$, and we have to introduce an adaptive Lyapunov function (Appendix F) to analyze the contraction behavior. In the supercritical regime, the sign of $x_t^{\parallel} h_t$ separates two sharply different behaviors. An aligned step with $x_t^{\parallel} h_t > 0$ preserves alignment and expands $\left|x_t^{\parallel}\right|$ by a factor at least $w_{t+1} - 1 > 1$. Consecutive misaligned steps instead imply contraction of $\left|x_t^{\parallel}\right|$ even though the dynamics is unstable, and this is a genuine and important phenomenon for the *positive* momentum.

The full details of Theorem 3.3 are presented in Appendix F, where we also show that the condition $\beta_2 > \beta_1^2$ cannot be dropped: for $\varepsilon = 0$, $\beta_1 \geq \sqrt{2} - 1$ and $\beta_2 \leq \beta_1^2$ there are orbits of exact period four along which $w_t < 2$ throughout (Proposition F.5).

3.2 Separating sharpness from the active curvature

On the rank-one quadratic the preconditioned sharpness and the active curvature coincide. However, as we observed in our experiments, their behaviors can diverge beyond such a simple setting. In the following, we demonstrate such separation in a simple yet representative setting: the diagonal quadratic, with $H = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. On this objective, the Adam dynamics is n decoupled 1-dimensional dynamics, each with its own normalized curvature $w_{i,t} := c\eta\lambda_i / (\sqrt{v_{i,t}} + \varepsilon)$, and

$$P_t = \mathrm{diag}(w_{1,t+1}, \ldots, w_{n,t+1}), \qquad \lambda_{\max}(P_t) = \max_i w_{i,t+1}, \qquad W_{\mathrm{act},t} = \textstyle\sum_{i=1}^n \hat{z}_{i,t}^2 \, w_{i,t+1}. \tag{11}$$

For each coordinate i, the behavior of $w_{i,t}$ is characterized in Section 3.1. The normalized sharpness $\lambda_{\max}(P_t) = \max_i w_{i,t+1}$ is therefore the maximum of n coordinates that each fluctuate around the threshold 2, so one may expect the maximum to stay above 2, as we indeed observe in Figure 5 (Appendix B.1). The following example makes this intuition concrete.

Sharpness can be conservative. Consider the following illustrative example: Let $H = \mathrm{diag}(1, (c\eta)^{-1} A\varepsilon)$ with $A > 2$ and let Adam start at $x_0 = (1, 0)$ with $m_0 = 0, v_0 = 1$. The dynamics of the second coordinate is trivial as $x_{2,t} = g_{2,t} \equiv 0$, and hence $w_{2,t} = \frac{A}{\beta_2^{t/2} \varepsilon^{-1} + 1} \to A$ converges exponentially as $t \to \infty$. In particular, the normalized sharpness $\lambda_{\max}(P_t) \geq w_{2,t+1}$, and hence it stays well above the threshold 2 for large t. The example is extreme, but it illustrates how directions that carry little gradient can dominate the sharpness. In our experiments (Figure 7 of Appendix B.2 and Table 3 of Appendix B.5.1) this effect is most pronounced when β_1 is small (e.g., RMSProp with $\beta_1 = 0$).

Active curvature towards the edge. As the simple example above illustrates, the "edge of stability" behavior in fact has to do with the coordinates that carry gradient. This intuition leads to the following theorem, which extends the two regimes of Theorem 3.3 to the active curvature.

Theorem 3.4 (Active curvature on diagonal quadratics)**.** *Let $H = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) \succ 0$, $\varepsilon > 0$, let $w_{i,\max} := c\eta\lambda_i / \varepsilon > 2$ for every i, and let $x_t \neq 0$ for all t. Let $W_{\mathrm{act},t}$ be built from D_t as in* (11)*.*

1. ***Supercritical.*** *Let $0 < \beta_1 < 1$ and $\beta_1^2 < \beta_2 < 1$, and assume that one-dimensional Adam with these parameters admits a potential (Definition H.1). By Lemma H.5, this holds at the default parameters $(\beta_1, \beta_2) = (0.9, 0.999)$. If $W_{\mathrm{act},t} \geq 2 + \delta$ for some $\delta > 0$ and all sufficiently large t, then $\sum_t \|x_t\|^2 < \infty$; in particular $x_t \to 0$ and $w_{i,t} \to w_{i,\max}$. Consequently, every trajectory with $\limsup_t \|x_t\| > 0$ has $\liminf_t W_{\mathrm{act},t} \leq 2$.*

2. ***Subcritical, dimension 2.*** *Let $n = 2$, $0 < \beta_1 < 1$ and $\beta_1^2 < \beta_2 < 1$. There is an explicit cut-off $\overline{W}_{\mathrm{diag}}(\beta_1, \beta_2) \in (0, 2)$, defined in Appendix H, such that for every $q < \overline{W}_{\mathrm{diag}}(\beta_1, \beta_2)$ the bound $W_{\mathrm{act},t} \le q$ cannot hold for all sufficiently large t; hence $\limsup_t W_{\mathrm{act},t} \ge \overline{W}_{\mathrm{diag}}(\beta_1, \beta_2)$. For the standard choice $(\beta_1, \beta_2) = (0.9, 0.999)$, $\overline{W}_{\mathrm{diag}} \ge 1.9$.*

Both regimes can be interpreted similarly to Theorem 3.3, though their proofs are much more involved because the active curvature aggregates the coordinate-wise curvature in a highly non-linear way. We refer to Appendix H for a detailed discussion on our techniques. We also remark that for RMSProp ($\beta_1 = 0$) stronger results are established in the subsequent section.

3.3 General objectives: the active curvature is the one-step loss criterion

In the following, we describe how to generalize our theory of the active curvature to general objectives, focusing on the case $\beta_1 = 0$ (RMSProp, $c = 1$). The update then reduces to $x_{t+1} = x_t - \eta z_t$ with $z_t := D_t^{-1} g_t$, and it is straightforward to see

$$f(x_{t+1}) = f(x_t) - \eta g_t^\top z_t + \frac{\eta^2}{2} z_t^\top \overline{H}_t z_t, \qquad \overline{H}_t := 2\int_0^1 (1-s)\,\nabla^2 f(x_t - s\eta z_t)\,ds. \tag{12}$$

Therefore, instead of investigating $W_{\mathrm{act},t}$ directly, it is more straightforward to study the *zeroth-order active curvature* defined as

$$W'_{\mathrm{act},t} := \frac{2\big(f(x_{t+1}) - f(x_t)\big)}{\eta g_t^\top D_t^{-1} g_t} + 2 = \eta\,\frac{g_t^\top D_t^{-1} \overline{H}_t D_t^{-1} g_t}{g_t^\top D_t^{-1} g_t}. \tag{13}$$

To interpret this, we first note that $W'_{\mathrm{act},t} = W_{\mathrm{act},t}$ when f is a quadratic function, and for general f, the two active curvatures differ only in whether the Hessian is read at x_t or averaged along the realized step. Further,

$$f(x_{t+1}) - f(x_t) = \frac{\eta}{2}\, g_t^\top D_t^{-1} g_t \big(W'_{\mathrm{act},t} - 2\big). \tag{14}$$

The threshold 2 is therefore exact for every twice differentiable f: a step raises the loss if and only if $W'_{\mathrm{act},t} > 2$.

Theorem 3.5. *Let $\beta_1 = 0$ and $\varepsilon > 0$. Suppose that (a) $f \in C^2$ is bounded below with $\sup_x \big\|\nabla^2 f(x)\big\|_{\mathrm{op}} \le \Lambda < \infty$, (b) $\|\nabla f(x)\| \to \infty$ as $\|x\| \to \infty$, and (c) every stationary point of f is nondegenerate. Then, along every Adam trajectory that does not reach a stationary point in finite time, we have*

$$\liminf_{t\to\infty} W'_{\mathrm{act},t} \;\le\; 2 + C_f \eta^{-1} \varepsilon, \qquad 2 - C_f \eta^{-1} \varepsilon \;\le\; \limsup_{t\to\infty} W'_{\mathrm{act},t}, \tag{15}$$

where C_f is a constant only depending on f (the proof is in Appendix I.1).

While the above result is stated for $W'_{\mathrm{act},t}$, the two active curvatures differ only by a small relative error. If $\nabla^2 f$ is Lipschitz, then $\big|W'_{\mathrm{act},t} - W_{\mathrm{act},t}\big| \le \frac{C'_f \eta}{\varkappa_t} W_{\mathrm{act},t}$, where C'_f depends only on f and β_2, and $\varkappa_t$ is the Rayleigh quotient of the raw Hessian $\nabla^2 f(x_t)$, with no preconditioner, along the update direction ((159) in Appendix I.2). The error is therefore $O(\eta)$ whenever $\varkappa_t$ stays bounded away from zero (Proposition I.2 in Appendix I.2).

4 Gradient reversal at EoS

Oscillatory or period-two behavior near a stability boundary has appeared in several prior analyses (Chen & Bruna, 2023; Song & Yun, 2023; Kalra et al., 2025; Mulayoff & Stich, 2026). In this section, we further identify the phenomenon of *gradient reversal*, which is predicted by our theoretical analysis in Section 3 and confirmed by our experiments (Section 4.1). In the following, we first illustrate how our theory predicts such phenomena on quadratic objectives and how it is closely connected with the active curvature.

Gradient reversal on rank-one quadratics. On the rank-one quadratic, we can show that in the *supercritical* regime (i.e., above the edge), the gradient at every step must reverse sign as long as $x_t^{\parallel} h_t > 0$ (with h_t as in Section 3.1). We formalize this observation in the following lemma.

Lemma 4.1 (Gradient and momentum reversal in the supercritical regime)**.** *Let $0 < \beta_1 < 1$ and consider the rank-one quadratic as in Section 3.1. Suppose that for step T, it holds that $x_T^{\parallel} h_T > 0$ and $W_{\mathrm{act},t} > 2$ for every $t \in [T, T+n]$, then*

$$\cos(g_t, g_{t+1}) = -1,\ t \in [T, T+n], \qquad \cos(m_t, m_{t+1}) = -1,\ t \in [T+1, T+n]. \tag{16}$$

The proof is in Appendix F.2.1. While rank-one quadratics are very special in that the gradients always belong to the 1-dimensional subspace $\mathrm{span}(u)$, the behavior indicated by Lemma 4.1 is in fact very close to what we observed in training neural networks with full-batch gradients (Section 4.1).

Active curvature as a lens for gradient reversal. The active curvature of Section 3 also explains gradient reversal, to some extent, beyond rank one. More specifically, on quadratic objectives with $\beta_1 = 0$ (RMSProp), we can express

$$\frac{\langle g_t, g_{t+1} \rangle_{D_t^{-1}}}{\|g_t\|^2_{D_t^{-1}}} = 1 - W_{\mathrm{act},t}. \tag{17}$$

We note that the LHS resembles the cosine between the preconditioned directions $D_t^{-1/2} g_t$ and $D_t^{-1/2} g_{t+1}$. In particular, $W_{\mathrm{act},t} > 1$ implies that $\langle g_t, g_{t+1} \rangle_{D_t^{-1}} < 0$ ("weak reversal"). Furthermore, we can in fact express

$$\cos(D_t^{-1/2} g_t, D_t^{-1/2} g_{t+1}) = \tfrac{1 - W_{\mathrm{act},t}}{\sqrt{(1-W_{\mathrm{act},t})^2 + V_t}}, \tag{18}$$

where $V_t = \|(P_t - W_{\mathrm{act},t} I)\hat{z}_t\|^2$ can be regarded as the variance of the preconditioned curvature under the distribution induced by the normalized vector $\hat{z}_t$. Therefore, gradient reversal is implied when $W_{\mathrm{act},t} - 1 \gg \sqrt{V_t}$. On a general objective f with Lipschitz Hessian, the analogue of Eq. (17) holds up to an $O(\eta)$ error: for $\beta_1 = 0$, reading the Hessian along the realized step instead of at x_t gives $\langle g_t, g_{t+1} \rangle_{D_t^{-1}} = (1 - W_{\mathrm{act},t} - \hat{e}_t)\ \|g_t\|^2_{D_t^{-1}}$ with $|\hat{e}_t| \le \frac{C\eta}{\varkappa_t} W_{\mathrm{act},t}$, where C depends only on f and β_2 and $\varkappa_t$ is the curvature of f along the step, as in Section 3.3. The error is thus $O(\eta)$ whenever $\varkappa_t$ stays bounded away from zero, and Eq. (18) holds verbatim with the step-averaged preconditioned Hessian in place of P_t. This is the same Hessian-averaging argument that relates $W'_{\mathrm{act},t}$ to $W_{\mathrm{act},t}$ (Proposition I.3 in Appendix I.2).

4.1 Gradient reversal across scales

We therefore measure the cosine between consecutive full-batch gradients across a broad range of neural-network training settings. Our experiments include fully connected networks of different sizes, ResNets, ViTs, LSTMs, GPT-2 medium, and several variants of the Adam family, while varying learning rates, β_1, β_2, the stabilizer, activation, dataset, training-set size, and batch size (Figures 3 and 4; the full set of settings is in Appendix C). Across these substantially different settings, gradient reversal appears repeatedly and often persists for long stretches of training. Its occurrence across architectures, scales, optimizers, and hyperparameters suggests that the phenomenon is not specific to a particular model or parameter choice. Figure 3 places the reversal next to the active curvature in six of these settings: in four of them the gradient reverses on exactly the stretches on which W_{act} sits at the threshold (the other two are discussed in Appendix C).

A striking feature of these runs is that strong gradient reversal does not prevent optimization progress. Even when consecutive gradients are nearly antiparallel for many steps, the training loss can continue to decrease smoothly (Figure 4, where the stretches of reversal are shaded under the loss curve; Figure 18 of Appendix C.1, bottom row against top row). This suggests that much of the motion is a fast, approximately period-two oscillation, superimposed on a smaller non-reversing component that continues to move the model toward lower loss.

Several observations support this interpretation. More directly, if consecutive iterates lie on opposite sides of a slowly moving center, their midpoint should cancel much of the alternating motion. Indeed, evaluating the loss at the midpoint of consecutive iterates produces a smoother and typically

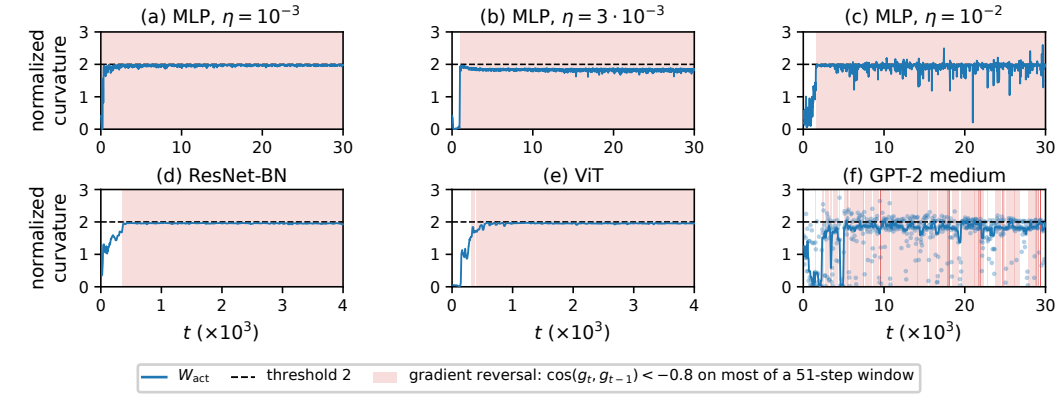

Figure 3: Active curvature and gradient reversal in six full-batch Adam runs with $\beta_1 = 0.9$, $\beta_2 = 0.999$. (a–c) The width-200 fc-tanh MLP of Appendix C.1. (d, e) The ResNet with BatchNorm and the ViT of Appendix B.3. (f) GPT-2 medium of Appendix B.4. Blue: W_{act}; dashed line at 2. Red shading: the stretches on which consecutive full-batch gradients reverse.

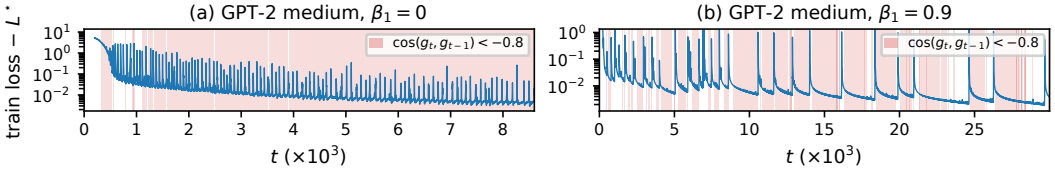

Figure 4: Period-two orbits at the edge in GPT-2 medium, trained full-batch on 64 fixed sequences with label-smoothed cross-entropy, $\eta = 3 \cdot 10^{-4}$: (a) $\beta_1 = 0$, $\beta_2 = 0.95$; (b) $\beta_1 = 0.9$, $\beta_2 = 0.999$. Blue: excess training loss $L - L^\star$ over the entropy of the smoothed target. Red shading: the stretches on which the full-batch gradient reverses. Training details are in Appendix B.4.

lower loss trajectory than evaluating it at the raw iterates (Figure 19; Figure 18 of Appendix C.1, top row, dashed). Exponential moving averages show the same behavior: they suppress the rapid oscillation and expose a smoother trajectory underneath (Figure 19, dotted; the three learning rates and the numbers are in Appendix C.1).

Taken together, these experiments suggest a simple picture of Adam's late-stage dynamics. A large component of the motion can be spent repeatedly moving back and forth in an approximately period-two fashion, while a much slower component carries the net optimization progress. The smooth decrease of the averaged loss is consistent with the model continuing to descend through this slowly moving center even while the raw gradients reverse from step to step (Figure 20: the height of the orbit stays constant while the loss at its center falls).

5 Conclusion

We studied why Adam operates near the edge of stability and identified its second-moment adaptation as a negative-feedback mechanism that continually pushes the dynamics back toward the critical regime. This perspective leads to the notion of active curvature, which captures the curvature encountered along the optimization trajectory and remains close to the stability boundary across a wide range of settings. Our analysis further predicts gradient reversal as a characteristic signature of this regime, a phenomenon that we observe broadly across architectures and Adam-family optimizers. Together, these results provide a unified dynamical picture of Adam near the edge of stability and suggest that its late-stage behavior is governed by persistent oscillation around a slowly evolving optimization trajectory. Extending the momentum analysis beyond quadratics, and the theory to stochastic gradients, remains open.

AI DISCLOSURE

In this work, we used AI tools to polish the paper and write code for generating the figures based on experimental data and verifying the numerical claims in our proof. We have reviewed all AI-assisted work carefully. We take responsibility for the final content of this work, including text, claims and 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, pp. 169–195, 2023.

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent, 2022. URL https://arxiv.org/abs/2204.01050.

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via the "edge of stability", 2023. URL https://arxiv.org/abs/2212.07469.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd, 2025. 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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 948–1024. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

Zhiwei Bai, Jiajie Zhao, Zhangchen Zhou, Zhi-Qin John Xu, and Yaoyu Zhang. Towards understanding adam convergence on highly degenerate polynomials, 2026a. URL https://arxiv.org/abs/2603.09581.

Zhiwei Bai, Zhangchen Zhou, Jiajie Zhao, Xiaolong Li, Zhiyu Li, Feiyu Xiong, Hongkang Yang, Yaoyu Zhang, and Zhi-Qin John Xu. Adaptive preconditioners trigger loss spikes in adam, 2026b. URL https://arxiv.org/abs/2506.04805.

Anas Barakat and Pascal Bianchi. Convergence and dynamical behavior of the adam algorithm for non-convex stochastic optimization, 2020. URL https://arxiv.org/abs/1810.02263.

Sebastian Bock and Martin Weiß. *Non-convergence and Limit Cycles in the Adam Optimizer*, pp. 232–243. Springer International Publishing, 2019. ISBN 9783030304843. doi: 10.1007/978-3-030-30484-3_20. URL http://dx.doi.org/10.1007/978-3-030-30484-3_20.

Sebastian Bock and Martin Georg Weiß. Local convergence of adaptive gradient descent optimizers, 2021. URL https://arxiv.org/abs/2102.09804.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4330–4391. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23b.html.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1x-x309tm.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sIE2rI3ZPs.

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, 2022. URL https://arxiv.org/abs/2207.14484.

Andre Belotto da Silva and Maxime Gazeau. A general system of differential equations to model first-order adaptive algorithms. *Journal of Machine Learning Research*, 21(129):1–42, 2020. URL http://jmlr.org/papers/v21/18-808.html.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nhKHA59gXz.

Steffen Dereich, Robin Graeber, Arnulf Jentzen, and Adrian Riekert. Asymptotic stability properties and a priori bounds for adam and other gradient descent optimization methods, 2025. URL https://arxiv.org/abs/2509.10476.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

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, 2019. URL https://arxiv.org/abs/1807.05031.

Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos, 2025. URL https://arxiv.org/abs/2311.02076.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Sungyoon Lee and Cheongjae Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=bH-kCY6LdKg.

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.

Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold, 2025. URL https://arxiv.org/abs/2511.02773.

Zhouzi Li, Zixuan Wang, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability, 2022. URL https://arxiv.org/abs/2207.12678.

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

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information Processing Systems*, 37:14810–14848, 2024.

Jaehong Moon. State-dependent lyapunov analysis of rank-1 matrix factorization. *arXiv preprint arXiv:2604.26993*, 2026.

Rotem Mulayoff and Sebastian U. Stich. On the stability of nonlinear dynamics in gd and sgd: Beyond quadratic potentials. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 5210–5243. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/mulayoff26a.html.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1904.09237.

Eric Regis and Sinho Chewi. A rod flow model for adam at the edge of stability, 2026. URL https://arxiv.org/abs/2605.06821.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory, 2023. URL https://arxiv.org/abs/2307.04204.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—RMSProp: Divide the gradient by a running average of its recent magnitude. Coursera course lecture, Neural Networks for Machine Learning, 2012. URL https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Lei Wu, Chao Ma, et al. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules, 2023. URL https://arxiv.org/abs/2208.09632.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p7EagBsMAEO.

CONTENTS

1 Introduction **1**

2 Setup and stability quantities **3**

3 Adam operates near an active stability boundary **3**

3.1 Rank one: the feedback loop that pins the active curvature 4

3.2 Separating sharpness from the active curvature 6

3.3 General objectives: the active curvature is the one-step loss criterion 7

4 Gradient reversal at EoS **7**

4.1 Gradient reversal across scales 8

5 Conclusion **9**

A Related work **14**

B Experimental evidence across settings **15**

B.1 Quadratics 15

B.2 Fully connected networks: learning rate, momentum and loss 16

B.3 Residual, attention and recurrent networks 17

B.4 GPT-2 medium 17

B.5 Ablation study 18

B.5.1 The threshold across the Adam family 18

B.5.2 Learning-rate decay 20

B.5.3 Batch size 20

C Gradient reversal across settings **21**

C.1 Fully connected networks: the runs of Section 4.1 22

C.2 Fully connected network: sixteen one-factor changes 23

C.3 Residual, attention and recurrent networks 23

C.4 GPT-2 medium 24

C.5 The Adam family 24

C.6 Batch size 25

D Additional preliminaries **34**

D.1 Proof of Proposition 3.1 35

E Proofs for RMSProp ($\beta_1 = 0$) **36**

E.1 Proof of Theorem 3.2 36

E.2 Eventual reversal at every step 38

F Proofs for general Adam (Theorem 3.3) **39**

F.1 Proof of the subcritical regime of Theorem 3.3 40

F.1.1 Existence of strictly subcritical four-cycles 43

F.2 Proof of the supercritical regime of Theorem 3.3 44

F.2.1 Proof of Lemma 4.1 45

F.2.2 Existence of misaligned supercritical trajectories 45

G Numerical illustration of Section 3.1 **47**

H Proof of Theorem 3.4 **47**

H.1 Proof of the supercritical regime 49

H.1.1 Main argument 49

H.1.2 Lemmas for Proposition H.2 50

H.2 Proof of the subcritical regime 55

H.2.1 Main argument 55

H.2.2 Lemmas for Proposition H.6 57

H.2.3 The constants and the explicit cutoff 59

I Proofs from Section 3.3 **60**

I.1 Proof of Theorem 3.5 61

I.2 Analyzing the active curvature 62

A Related work

Edge of stability. The edge-of-stability (EoS) phenomenon was systematically documented by Cohen et al. (2021), with earlier observations of sharpness dynamics along SGD trajectories by Jastrzębski et al. (2019) and precursors in the catapult phase (Lewkowycz et al., 2020); the role of dynamical stability in selecting minima was highlighted by Wu et al. (2018). Subsequent work has studied its underlying mechanisms, including implicit regularization (Arora et al., 2022), self-stabilization through higher-order geometry (Damian et al., 2023), progressive sharpening (Li et al., 2022), and EoS behavior in simplified models (Agarwala et al., 2023; Ahn et al., 2023; Zhu et al., 2023; Chen & Bruna, 2023); for rank-one matrix factorization, Moon (2026) analyzes the edge with a state-dependent Lyapunov function, the kind of argument our subcritical analysis also uses (Appendix F). Other works connect EoS to bifurcation and nonlinear oscillatory dynamics (Song & Yun, 2023; Mulayoff & Stich, 2026; Kalra et al., 2025). Related work has also studied directional, gradient-aware, or path-wise notions of curvature and smoothness that characterize optimization behavior along the trajectory rather than solely through the largest Hessian eigenvalue (Ahn et al., 2022; Mishkin et al., 2024; Lee & Jang, 2023). In stochastic optimization, Andreyev & Beneventano (2025) showed that minibatch SGD operates near an edge characterized by *batch sharpness*, which measures curvature along stochastic gradient directions rather than through the full-batch Hessian alone. These results primarily concern gradient descent or SGD. For adaptive methods, Cohen et al. (2022) identified an analogous adaptive EoS characterized by the preconditioned Hessian. Our work instead studies the exact finite-step Adam dynamics and identifies the active curvature along the optimization direction as the quantity regulated by adaptive preconditioning, first rigorously on quadratic objectives and then beyond quadratics theoretically and empirically.

Adam. AdaGrad (Duchi et al., 2011) introduced coordinatewise adaptive learning rates based on accumulated gradients, and Kingma & Ba (2015) later introduced Adam by combining adaptive second-moment scaling with momentum. Despite its empirical success, Adam can fail to converge

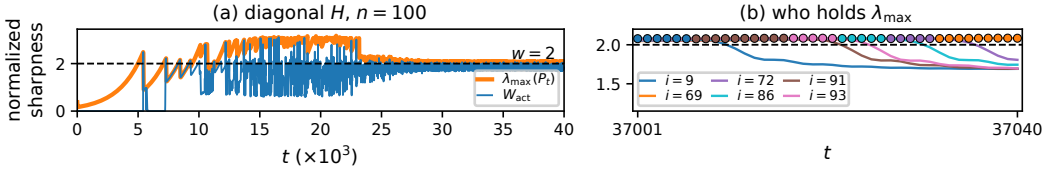

Figure 5: The sharpness and the active curvature on a diagonal quadratic. Diagonal quadratic in $n = 100$ dimensions, Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\eta = 10^{-3}$. (a) The sharpness $\lambda_{\max}(P_t) = \max_i w_{i,t+1}$ (orange) and the active curvature W_{act} (blue). (b) Forty consecutive plateau steps, one color per coordinate i: the line is its normalized curvature $w_{i,t+1}$ and the dots mark the steps on which it attains the sharpness. Spectrum, initialization and the readings are described below.

(Reddi et al., 2018), motivating convergence analyses under additional assumptions (Chen et al., 2019; Zhang et al., 2023; Dereich et al., 2025). Adam has also been studied through continuous-time dynamical models (da Silva & Gazeau, 2020; Barakat & Bianchi, 2020). More recent work has investigated Adam's behavior on degenerate objectives and its implicit effect on sharpness (Bai et al., 2026a; Li et al., 2025). Little work characterizes how Adam's adaptive preconditioner drives its sharpness relative to a finite-step stability threshold.

Adam at the edge of stability. The works most closely related to ours study Adam directly through stability and dynamical perspectives. Bai et al. (2026b) explain loss spikes through the evolution of the adaptive preconditioner, with their theoretical analysis focusing on a one-dimensional quadratic setting with $\beta_1 = 0$. Cohen et al. (2025) derive a central-flow description of the time-averaged oscillatory dynamics near the edge, while Regis & Chewi (2026) develop a continuous-time model for Adam in the EoS regime. From a discrete dynamical perspective, Bock & Weiß (2019) showed that Adam can admit non-convergent limit cycles, including quadratic examples, while Bock & Weiß (2021) analyzed local convergence through linear stability near fixed points. Recently, Dereich et al. (2025) established a priori bounds and asymptotic stability properties for Adam on strongly convex quadratic objectives. In contrast, we study the exact finite-step dynamics across both subcritical and supercritical regimes, identify the active curvature as the relevant stability quantity, and connect its regulation near the edge to gradient reversal, with the predicted dynamics further examined beyond quadratics and across neural-network training settings.

B Experimental evidence across settings

This appendix reports $\lambda_{\max}(P_t)$ and W_{act} on quadratics (B.1), fully connected networks (B.2), residual, attention and recurrent networks (B.3) and GPT-2 medium (B.4), and then varies the optimizer (B.5.1), the learning-rate schedule (B.5.2) and the batch size (B.5.3).

B.1 Quadratics

The objective is $f(x) = \frac{1}{2} x^\top H x$ with $H = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $n = 100$, and λ_i log-uniformly spaced on $[10^{-3}, 1]$. The initial point is $x_0 \sim \mathcal{N}(0, I_n)$ with $m_0 = v_0 = 0$. We run Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\eta = 10^{-3}$, $\varepsilon = 0$ for 40,000 steps in double precision. As we discuss in Section 3, because H is diagonal the recursion decouples: coordinate i runs the one-dimensional problem of curvature λ_i, and $P_t = c\eta D_t^{-1/2} H D_t^{-1/2}$ is diagonal with entries $w_{i,t+1} = c\eta\lambda_i / \sqrt{v_{i,t+1}}$.

Panel (a) of Figure 5 draws every tenth step of the two readings over the whole run. Panel (b) is the window of steps 37,001–37,040: each coordinate that holds the maximum at some step of the window receives a color, its $w_{i,t}$ is drawn as a line, and the steps on which it is the holder carry a dot of that color. This figure explains why the sharpness can remain above the threshold while W_{act} does not: the top eigendirection can be continually handed off between coordinates whose gradients have become small, whereas W_{act} weights each direction by the gradient mass it currently carries.

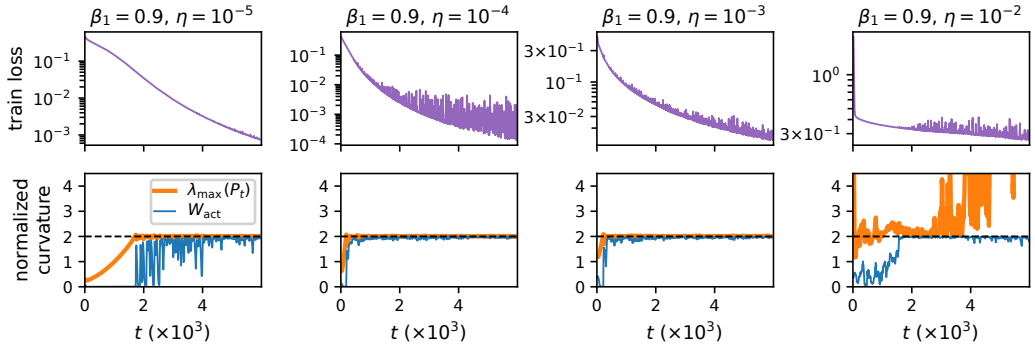

Figure 6: Width-200 network ($3072-200-200-10$, $6.6{\cdot}10^5$ parameters), full-batch MSE, $\beta_1 = 0.9$, 6000 steps, at four learning rates: training loss (purple) above $\lambda_{\max}(P_t)$ (orange) and W_{act} (blue), 31-step rolling medians, threshold 2 dashed. The curvature axis is cut at 4.5; at $\eta = 10^{-2}$ the sharpness rises above it (Table 1).

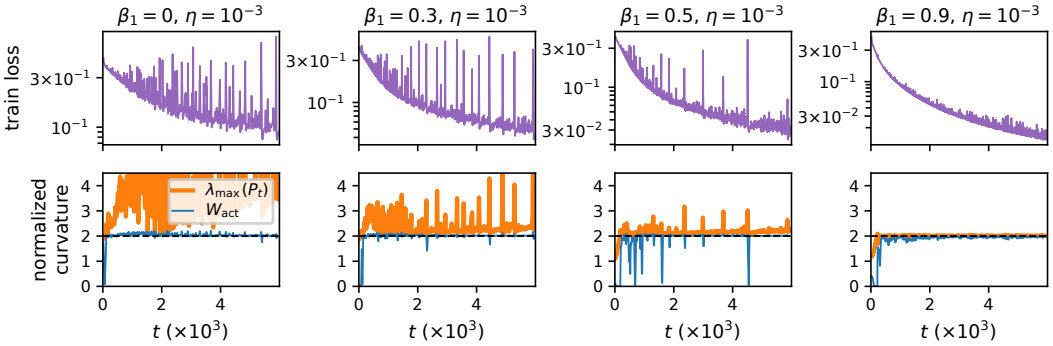

Figure 7: The network and readings of Figure 6, MSE, $\eta = 10^{-3}$, at $\beta_1 \in \{0, 0.3, 0.5, 0.9\}$. The $\beta_1 = 0.9$ panel is the $\eta = 10^{-3}$ run of Figure 6.

B.2 Fully connected networks: learning rate, momentum and loss

Implementation. Unless stated otherwise, the network runs of this appendix use the following protocol. The network is an fc-tanh MLP on the first 5000 CIFAR-10 images (standardized), trained with MSE against one-hot targets and full-batch gradients by uncorrected Adam with $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$, in float32 with seed 0. At every step $\lambda_{\max}(P_t)$ is a 25-step warm-started Lanczos iteration on $P_t = c\eta D_t^{-1/2} H_t D_t^{-1/2}$, and W_{act} is exact, one Hessian-vector product along $\hat{z}_t$; both are drawn as 31-step rolling medians without the first 20 steps, where $v \approx 0$ makes P_0 huge. Figures 6–8 and Table 1 use the width-200 network $3072-200-200-10$ ($6.6 \cdot 10^5$ parameters) for 6000 steps and change one factor at a time from $\beta_1 = 0.9$, $\eta = 10^{-3}$, MSE: $\eta \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$, $\beta_1 \in \{0, 0.3, 0.5, 0.9\}$, and cross-entropy with label smoothing 0.1 and 0.2, reported as the excess $L - L^\star$ over the entropy of the smoothed target ($L^\star = 0.500$ and 0.867). The left column of Figure 1 is the $\beta_1 = 0$, $\eta = 10^{-3}$ run, drawn as centered 100-step rolling medians because it jitters at every step. Table entries are medians over steps 3000–6000; the training accuracy is the median over steps 5800–5999.

Summary. All seven MSE runs reach the edge: W_{act} settles at 2 (at $\eta = 10^{-5}$ only after step 2100, with median 1.89), while the sharpness ranges from 2 to well above it (Table 1). The sharpness sits at 2 only when the gradient occupies the top eigenspace ($\beta_1 = 0.9$, $\eta \le 10^{-3}$); at $\beta_1 = 0$ or $\eta = 10^{-2}$ it drifts upward while W_{act} stays at 2 (Figures 6 and 7). Label-smoothed cross-entropy gives the same picture.

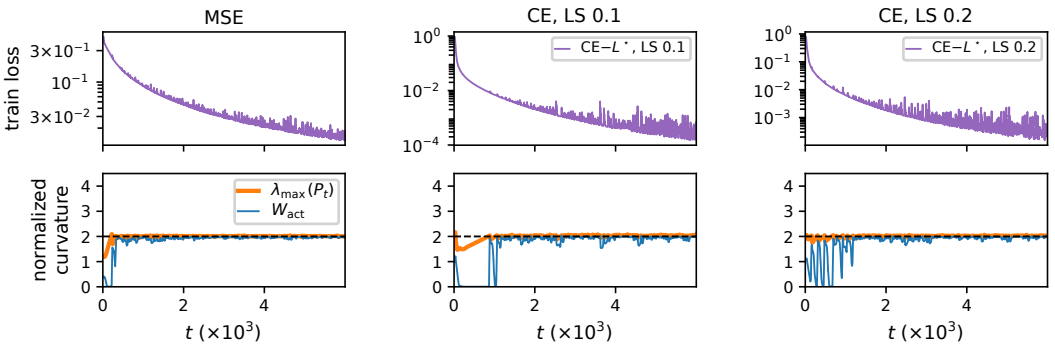

Figure 8: The network and readings of Figure 6, $\beta_1 = 0.9$, $\eta = 10^{-3}$, under MSE and under cross-entropy with label smoothing (LS) 0.1 and 0.2; the label-smoothed panels show the excess loss $L - L^\star$ over the entropy of the smoothed target. The MSE panel is the $\eta = 10^{-3}$ run of Figures 6 and 7.

Table 1: The runs of Figures 6–8: width-200 MLP on 5000 CIFAR-10 images, full batch, 6000 steps; medians of $\lambda_{\max}(P_t)$ and W_{act} over steps 3000–6000 and the training accuracy (median over steps 5800–5999). "CE, LS 0.1" and "CE, LS 0.2" are cross-entropy with label smoothing 0.1 and 0.2.

loss	β_1	η	$\lambda_{\max}(P_t)$	W_{act}	train acc
MSE	0	10^{-3}	3.35	2.018	0.973
MSE	0.3	10^{-3}	2.25	2.013	0.992
MSE	0.5	10^{-3}	2.17	2.011	0.996
MSE	0.9	10^{-3}	2.01	1.965	0.996
MSE	0.9	10^{-5}	2.00	1.885	1.000
MSE	0.9	10^{-4}	2.01	1.970	1.000
MSE	0.9	10^{-2}	5.07	1.967	0.638
CE, LS 0.1	0.9	10^{-3}	2.03	1.957	1.000
CE, LS 0.2	0.9	10^{-3}	2.02	1.962	1.000

B.3 Residual, attention and recurrent networks

The vision networks are a ResNet (128-channel stem, two basic blocks of 128 and 256 channels, BatchNorm, 1.22M parameters) and a ViT (patch 4, width 128, four heads, two pre-LN blocks, $4.1 \cdot 10^5$ parameters) on the first 1000 CIFAR-10 images with one-hot MSE. The LSTM has two layers of width 128 and a tied embedding (6.70M parameters, 6.43M in the embedding) and is trained on 256 FineWeb sequences of 64 GPT-2 tokens with label-smoothed cross-entropy (smoothing 0.1). All runs are full batch, uncorrected Adam with $\eta = 3 \cdot 10^{-4}$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, a 100-step warm-up and 4000 steps, at $\beta_1 = 0$ and 0.9. Attention, LayerNorm and the LSTM cell are written in elementary operations, since the fused kernels have no second derivative. The full batch is processed in two fixed chunks of 500 images, and BatchNorm couples the examples within a chunk, so H_t is the Hessian of that chunked objective.

B.4 GPT-2 medium

GPT-2 medium (355M parameters, 24 layers), the transformer of Section 4.1, is trained full batch on 64 FineWeb sequences of 1024 tokens with label-smoothed cross-entropy (smoothing 0.1; the vocabulary is padded to 50,304 tokens, so the entropy floor is 1.4076 rather than the LSTM's 1.4075). The optimizer is PyTorch AdamW without weight decay or clipping, $\eta = 3 \cdot 10^{-4}$ after a 200-step warm-up, at $(\beta_1, \beta_2) = (0, 0.95)$ for 8600 steps and at $(0.9, 0.999)$ and $(0.9, 0.95)$ for 30,000 steps (Figure 11). W_{act} of (5) is exact and taken every 10 steps at $\beta_1 = 0$ and every 50 at $\beta_1 = 0.9$; $\lambda_{\max}(P_t)$ is taken every 100 or 250 steps by ten warm-started power iterations; at $\beta_1 = 0.9$ these converge except on some spike steps, at $\beta_1 = 0$ about a quarter of the readings have not converged. The runs re-train the two runs of Figure 4 and a $\beta_2 = 0.95$ twin; their reversal is in Appendix C.4.

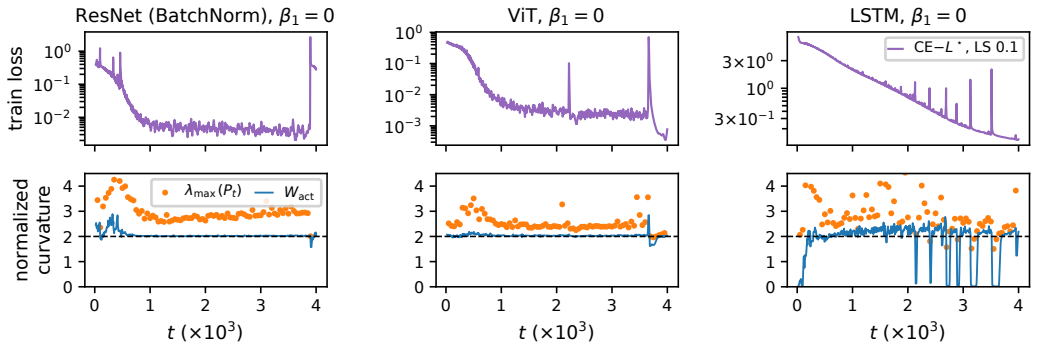

Figure 9: BatchNorm ResNet and ViT (full-batch MSE on 1000 CIFAR-10 images) and LSTM at $\beta_1 = 0$, $\eta = 3 \cdot 10^{-4}$, $\beta_2 = 0.999$, 4000 steps. Top: training loss (purple). Bottom: $\lambda_{\max}(P_t)$ (orange, every 50 steps) and the 31-step rolling median of W_{act} (blue), dashed line at 2.

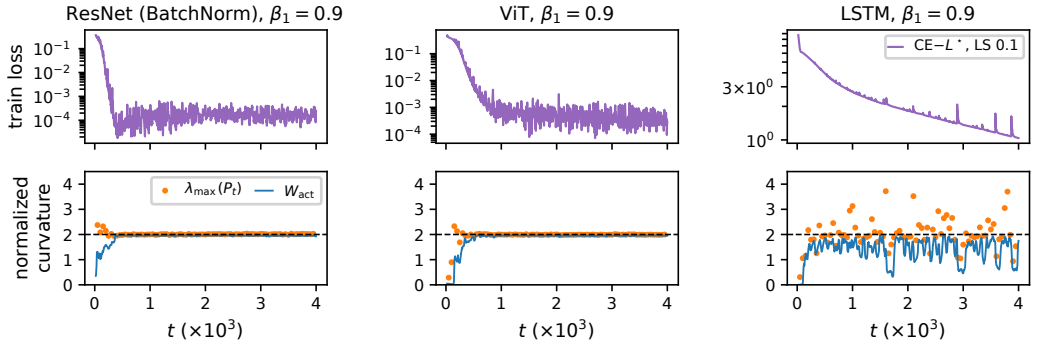

Figure 10: As Figure 9 at $\beta_1 = 0.9$.

B.5 Ablation study

The runs so far use Adam at a constant learning rate and full batch. We now vary the optimizer (Appendix B.5.1), the learning-rate schedule (Appendix B.5.2) and the batch size (Appendix B.5.3). The first two leave the regulation intact: eight of the ten variants hold W_{act} at 2, and a run at the edge stays there through a decay. The effect of the batch size depends on β_1.

B.5.1 The threshold across the Adam family

The variants have different stability thresholds even with a frozen preconditioner, so we normalize each by its own constant, which puts its edge at 2.

The ten updates. Every variant runs the same loop: from the full-batch gradient $g_t = \nabla f(x_t)$ it advances a second-moment state, which gives the diagonal preconditioner D_t, then a momentum state, then the iterate. Uncorrected Adam is the recursion (1),

$$m_{t+1} = \beta_1 m_t + (1-\beta_1) g_t, \qquad v_{t+1} = \beta_2 v_t + (1-\beta_2)\, g_t \odot g_t,$$
$$D_t = \operatorname{diag}\big(\sqrt{v_{t+1}}\big) + \varepsilon I, \qquad x_{t+1} = x_t - \eta D_t^{-1} m_{t+1},$$

and each of the nine others replaces only the lines written below; every line not written is the corresponding line above. Throughout $m_0 = v_0 = 0$, $\beta_1 = 0.9$ (momentum $\mu = 0.9$ for the standard-parameterization variant), $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, weight decay $\lambda = 0.1$ (in this subsection only, λ is the weight-decay coefficient), and all operations on vectors are coordinatewise.

Table 2: Second-half medians (steps 2000–3999) of the two curvature readings for the six runs of Figures 9 and 10; loss is the training loss over the last 400 steps (MSE for the vision networks, excess cross-entropy over the floor 1.4075 for the LSTM).

architecture	params	β_1	$\lambda_{\max}(P_t)$	W_{act}	loss
ResNet-BN	1.22M	0	2.89	2.023	0.00406
ResNet-BN	1.22M	0.9	2.02	1.957	0.000164
ViT	413k	0	2.39	2.021	0.000777
ViT	413k	0.9	1.99	1.955	0.000377
LSTM	6.70M	0	2.46	2.102	0.139
LSTM	6.70M	0.9	1.95	1.466	1.11

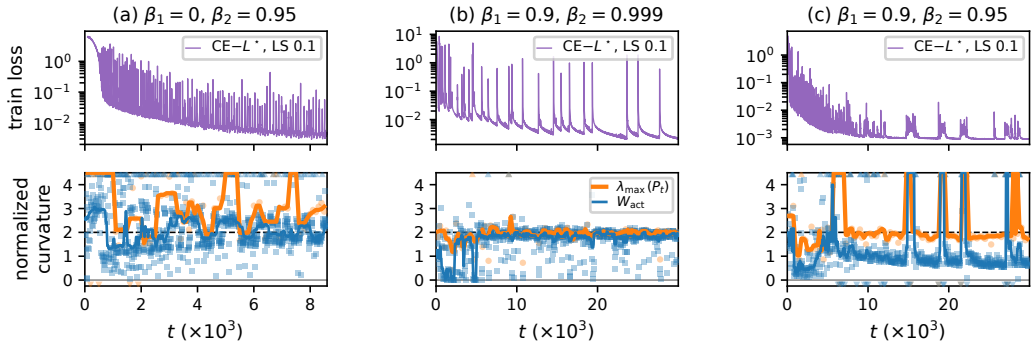

Figure 11: GPT-2 medium, full batch, over the whole of each run: (a) $\beta_1 = 0$, $\beta_2 = 0.95$, 8600 steps; (b) $\beta_1 = 0.9$, $\beta_2 = 0.999$, and (c) $\beta_1 = 0.9$, $\beta_2 = 0.95$, 30,000 steps. Top: training loss, excess over the label-smoothing floor 1.4076 (purple). Bottom: the exact full-batch sharpness $\lambda_{\max}(P_t)$ (orange; one reading every 100 steps in (a), every 250 in (b, c)) and active curvature W_{act} (blue; every 10 steps in (a), every 50 in (b, c)); faint markers are the individual readings, lines their rolling medians over about 500 steps (750 for the sharpness in (b, c)); dashed line at 2; readings outside $[-0.25, 4.5]$ are drawn as triangles at the edge.

- **Adam, bias-corrected.** Both corrections are folded into the preconditioner, as in Cohen et al. (2022):
$$D_t = \left(1 - \beta_1^{t+1}\right) \left[\operatorname{diag}\left(\sqrt{\frac{v_{t+1}}{1 - \beta_2^{t+1}}}\right) + \varepsilon I\right].$$
- **AdamW.** Decoupled weight decay:
$$x_{t+1} = (1 - \eta\lambda)\, x_t - \eta D_t^{-1} m_{t+1}.$$
- **Adam with L_2 regularization.** Coupled weight decay: the gradient
$$\widetilde{g}_t = \nabla f(x_t) + \lambda x_t = \nabla \left(f(x_t) + \frac{\lambda}{2} \|x_t\|^2\right)$$
replaces g_t in both moments, so the objective is $f + \frac{\lambda}{2}\|\cdot\|^2$ and the curvature read is $\widetilde{H}_t = H_t + \lambda I$.
- **Padam.** A weaker preconditioner:
$$D_t = \operatorname{diag}\left(v_{t+1}^{p}\right) + \varepsilon I, \qquad p = \tfrac{1}{4}.$$
- **Nadam.** m_t is updated as usual, but the parameter update uses $\widetilde{m}_{t+1}$:
$$\widetilde{m}_{t+1} = \beta_1 m_{t+1} + (1 - \beta_1) g_t, \qquad x_{t+1} = x_t - \eta D_t^{-1} \widetilde{m}_{t+1}.$$
- **RMSProp.** No momentum:
$$x_{t+1} = x_t - \eta D_t^{-1} g_t.$$

Table 3: Ten variants, with $\beta_1 = 0.9$ or momentum $\mu = 0.9$ where the variant has a momentum term: the normalization constant c and the raw threshold of each, and their readings on the width-200 network (fc-tanh $3072{-}200{-}200{-}10$, 5000 CIFAR-10 images, MSE, 3000 steps, last quarter): means of the normalized $\lambda_{\max}$ and W_{act}, and the fraction of plateau steps with $W_{\text{act}} \in (1.8, 2.1)$ ("locked").

variant	c	threshold	$\lambda_{\max}$	W_{act}	locked
Adam (uncorrected)	$(1-\beta_1)/(1+\beta_1)$	$38/\eta$	2.01	1.95	100%
Adam, bias-corrected	c_{Adam}	$38/\eta$	2.02	1.94	99%
AdamW	$c_{\text{Adam}}/(1-\eta\lambda/2)$	$38(1-\eta\lambda/2)/\eta$	2.00	1.92	94%
Adam + L_2	c_{Adam}	$38/\eta$	1.87	1.76	43%
Padam ($p = 1/4$)	c_{Adam}	$38/\eta$	2.01	1.97	100%
Nadam	$c_{\text{Adam}}(1+2\beta_1)$	$13.57/\eta$	2.09	2.00	100%
RMSProp	1	$2/\eta$	3.89	2.07	77%
RMSProp + momentum	$1/(1+\mu)$	$3.8/\eta$	3.80	2.04	37%
AMSGrad	c_{Adam}	$38/\eta$	2.00	1.98	100%
Adagrad	1	$2/\eta$	2.08	2.03	100%

- **RMSProp with momentum.** The momentum sits after the preconditioner and in the standard parameterization:
$$b_{t+1} = \mu b_t + D_t^{-1} g_t, \qquad x_{t+1} = x_t - \eta\, b_{t+1}.$$
- **AMSGrad.** The second moment ratchets:
$$\bar{v}_{t+1} = \max\{\bar{v}_t, v_{t+1}\}, \qquad D_t = \operatorname{diag}\big(\sqrt{\bar{v}_{t+1}}\big) + \varepsilon I,$$
the maximum taken coordinatewise, so that $D_{t+1}^{-1} \preceq D_t^{-1}$.
- **Adagrad.** The second moment accumulates and there is no momentum:
$$v_{t+1} = v_t + g_t \odot g_t, \qquad x_{t+1} = x_t - \eta D_t^{-1} g_t.$$

The learning rate is $\eta = 10^{-3}$, except for Padam and Adagrad, whose weaker preconditioners need $\eta = 10^{-2}$ to train. The two readings are then taken from (5) and $\lambda_{\max}(P_t)$ with that variant's own D_t.

With the preconditioner D frozen, each variant is preconditioned gradient descent with its own momentum, and its threshold for $\lambda_{\max}(D^{-1/2} H D^{-1/2})$ is that of the unpreconditioned method. We set $c := 2/(\eta \cdot \text{threshold})$, so that the normalized threshold is 2; here c is the variant's own constant, and for Adam it is the c of Section 2, $c_{\text{Adam}} = (1-\beta_1)/(1+\beta_1)$. Table 3 lists c, the threshold and the readings of each variant.

B.5.2 Learning-rate decay

Figure 13 decays η exponentially over steps 2000–5000 on width-512 MLPs at $\beta_1 = 0$ and 0.9, with both readings normalized by the instantaneous $c\eta_t$. Runs that are at the edge before the decay are back at $W_{\text{act}} = 2$ after it; at $\beta_1 = 0$ they hold 2 throughout the decay, while at $\beta_1 = 0.9$ with $\eta : 10^{-3} \to 2 \cdot 10^{-4}$ the reading dips briefly inside the window. The exception is $\beta_1 = 0.9$ with $\eta : 10^{-2} \to 10^{-3}$, which had not reached the edge: W_{act} approaches 2 and then drops well below it, climbing back only slowly after the decay (1.9 at the end), while the constant-rate twin (not shown) settles at the edge. Decay thus changes when the edge is reached, not whether the regulation holds once it is reached; at constant rate every $\eta \in [10^{-5}, 10^{-2}]$ reaches the edge on the width-200 network of Table 1.

B.5.3 Batch size

Figures 14–17 subsample a fixed training set to a fraction B/N at each step: a width-200 MLP ($N = 5000$, MSE, 30,000 steps) and GPT-2 medium ($N = 64$ sequences, label-smoothed cross-entropy, 8000 steps). Table 4 gives medians over steps 20,000–30,000 (MLP) and 4000–8000 (GPT-2). Each run has two readings. The full-set reading uses g_t and H_t as in Section 2 and is comparable with the full-batch runs. The minibatch reading uses the step's own gradient $g_{B,t}$ and Hessian $H_{B,t}$ in the same formulas,
$$c\eta\, \frac{g_{B,t}^\top D_t^{-1} H_{B,t} D_t^{-1} g_{B,t}}{g_{B,t}^\top D_t^{-1} g_{B,t}} \qquad \text{and} \qquad \lambda_{\max}\big(c\eta\, D_t^{-1/2} H_{B,t} D_t^{-1/2}\big).$$

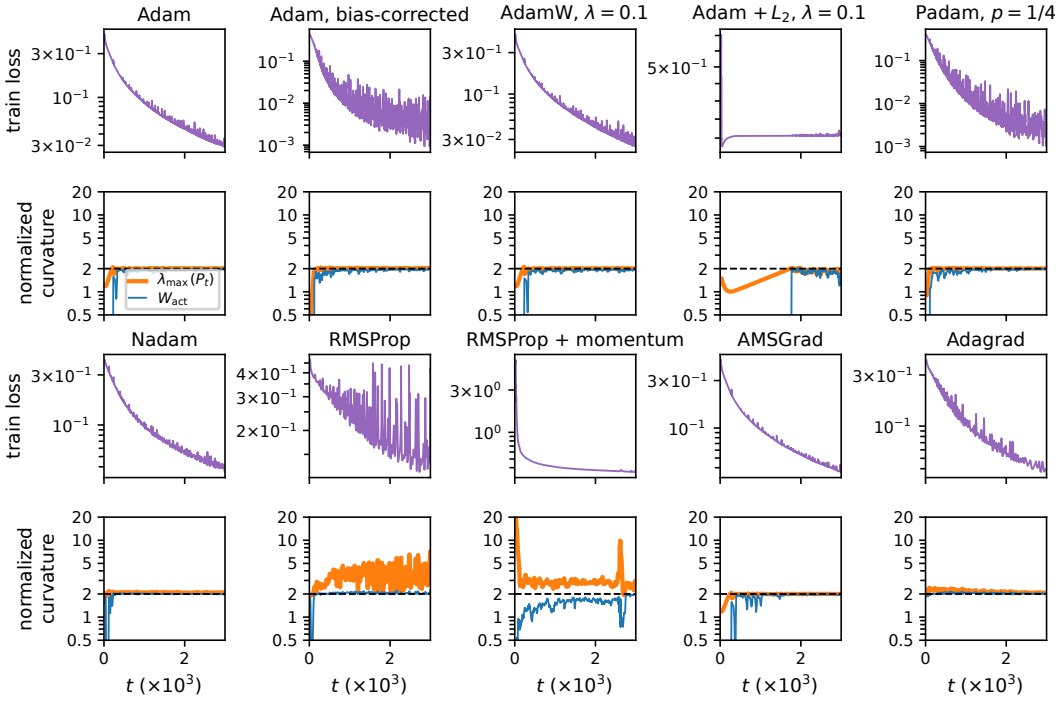

Figure 12: The ten variants of Table 3 on the width-200 MLP (5000 CIFAR-10 images, MSE, full batch, 3000 steps), one column per optimizer: training loss (purple) above $\lambda_{\max}(P_t)$ (orange) and W_{act} (blue), 31-step rolling medians, each normalized by the variant's own constant c so that the threshold (dashed) is 2 for all. AdamW and L_2 use decay 0.1; Padam and Adagrad run at $\eta = 10^{-2}$, the others at 10^{-3}.

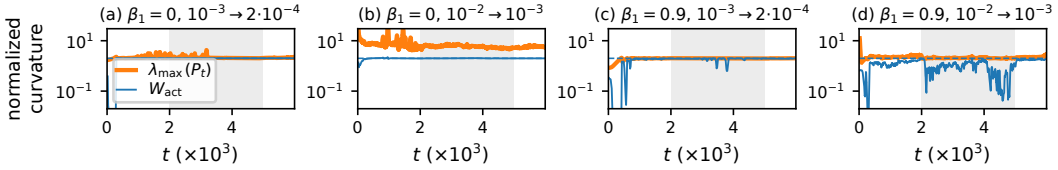

Figure 13: Width-512 MLPs with an exponential learning-rate decay over the shaded window (steps 2000–5000), 31-step rolling medians of $\lambda_{\max}(P_t)$ (orange) and W_{act} (blue) normalized by the instantaneous $c\eta_t$.

On the MLP the full-set sharpness is a 15-step Lanczos iteration every 20 steps; on GPT-2 it is recomputed only at saved checkpoints.

C GRADIENT REVERSAL ACROSS SETTINGS

Section 4 predicts that at the edge the full-batch gradient reverses at every step, $\cos(g_t, g_{t-1}) \to -1$, and Section 4.1 shows this on the width-200 MLP and GPT-2 medium (Figures 3, 4 and 18). This appendix gives the protocol of those runs (C.1) and the reversal on the small network under sixteen changes (C.2), the other architectures (C.3), GPT-2 (C.4), the Adam family (C.5) and the batch-size scan (C.6) of Appendix B. Cosines are between consecutive full-batch quantities, curves are 101-step rolling medians, and "reversal steps" is the fraction of steps with $\cos(g_t, g_{t-1}) < -0.5$. In Figure 3, W_{act} is a 31-step rolling median in (a–e) and, in (f), the exact reading every 50 steps with its rolling median over about 550 steps; (f) uses the bias-corrected AdamW of Appendix B.4, the other five runs uncorrected Adam.

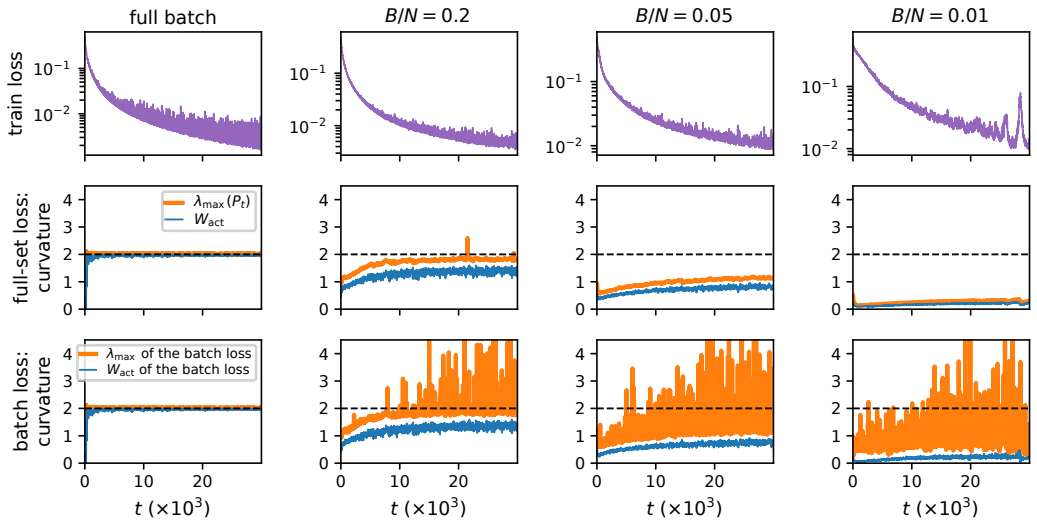

Figure 14: Batch-size scan on the width-200 MLP, $\beta_1 = 0.9$, $\eta = 10^{-3}$, MSE, 30,000 steps: minibatch training, in which Adam sees at each step the gradient of a random subset of the fixed training set ($N = 5000$ images) of size B, for $B/N = 1$ (full batch), 0.2, 0.05 and 0.01 (four of the six fractions of Table 4). Top: the training loss on all N images (purple). Middle: the two readings on the *full-set* loss of that minibatch trajectory, the sharpness $\lambda_{\max}(P_t)$ of the full-set Hessian (orange, Lanczos every 20 steps) and the active curvature W_{act} from the full-set gradient and Hessian (blue). Bottom: the same two readings on the loss of the step's own *minibatch*, its Hessian and its gradient ($\lambda_{\max}$ orange; W_{act} blue, every step).

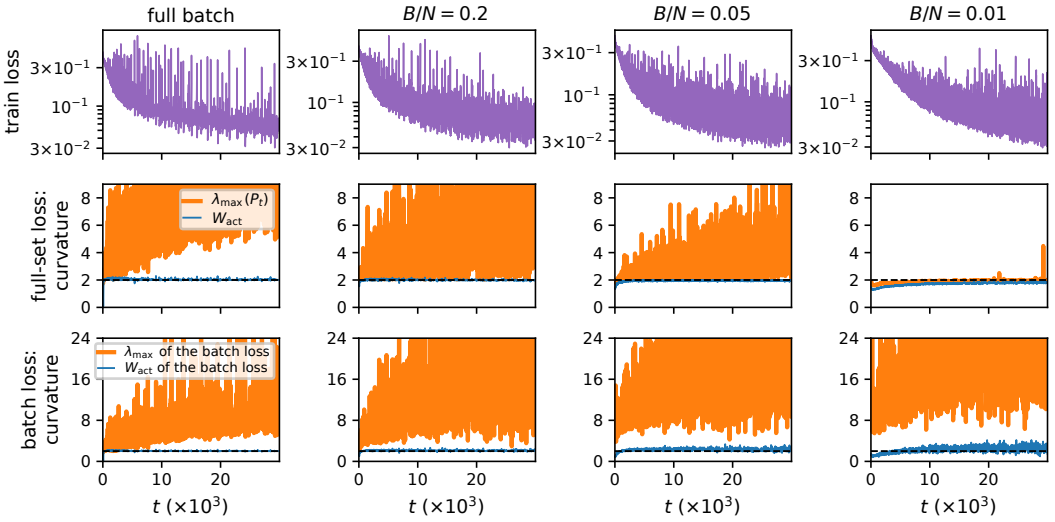

Figure 15: Batch-size scan on the width-200 MLP at $\beta_1 = 0$ (otherwise as Figure 14: minibatch training on random subsets of the fixed $N = 5000$ images, $B/N = 1$, 0.2, 0.05 and 0.01, $\eta = 10^{-3}$, MSE, 30,000 steps). Top: training loss on all N images (purple). Middle: sharpness and active curvature (blue) of the full-set loss along the minibatch trajectory. Bottom: the same two readings on the loss of the step's own minibatch.

C.1 Fully connected networks: the runs of Section 4.1

Figures 18 and 19 use the width-200 network of Appendix B.2 (5000 CIFAR-10 images, MSE, full batch) with uncorrected Adam, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, seed 0, for 30,000 steps at a constant learning

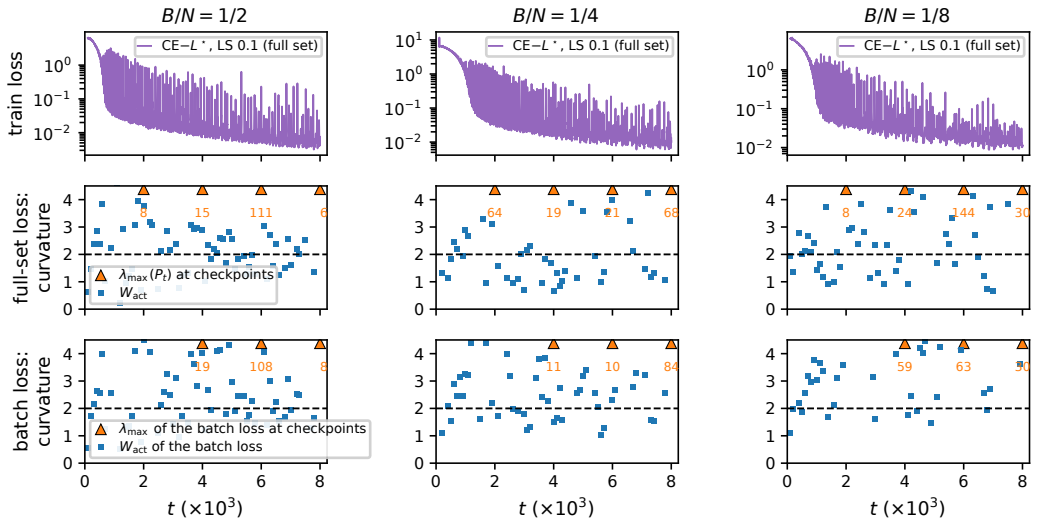

Figure 16: Batch-size scan on GPT-2 medium, $\beta_1 = 0$, $\beta_2 = 0.95$, 8000 steps: minibatch training in which Adam sees at each step the gradient of $B = 32$, 16 or 8 of the fixed 64 training sequences ($B/N = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}$; the full-batch run, $B/N = 1$, is Figure 11(a)). Top: the training loss on all 64 sequences, as excess over the label-smoothing floor 1.4076 (purple). Middle: the readings on the *full-set* loss of the minibatch trajectory, the exact active curvature W_{act} from the full-set gradient and Hessian-vector product (blue squares, one reading every 100 steps) and the sharpness $\lambda_{\max}(P_t)$ of the full-set Hessian, which was not recorded along the run and is recomputed at the saved checkpoints x_{2000}, x_{4000}, x_{6000} and x_{8000} (orange circles where it fits the axis, orange triangles at the top edge with the value written below them where it does not). Bottom: the same two readings on the loss of the step's own *minibatch*: W_{act} of the batch loss (blue squares, one sample every 100 steps) and its sharpness recomputed at x_{4000}, x_{6000} and x_{8000} on two random batches of the run's size each (orange, median of the two). Dashed lines at 2.

rate. Along each run we record the loss at the iterate, at the midpoint $\frac{1}{2}(x_t + x_{t-1})$ and at the EMA iterate (decay 0.99), and the cosines of consecutive gradients, momenta and displacements.

Figure 19: the fixed-height orbit. Panels (a) and (b) are the $\eta = 10^{-3}$ runs at $\beta_1 = 0$ and 0.9 (EMA drawn from step 500; it starts at the first iterate and lags early on). Over the last 10,000 steps the midpoint and the EMA iterate reach about half the loss at the iterate at $\beta_1 = 0$ and about two thirds at $\beta_1 = 0.9$. The orbit height $E_{\text{osc}} = L(x_t) - L(\frac{1}{2}(x_t + x_{t-1}))$ (Figure 20) is flat from step 1000 on at $\beta_1 = 0$ and from about step 3000 on at $\beta_1 = 0.9$, as is the gradient norm, while the loss keeps falling.

C.2 Fully connected network: sixteen one-factor changes

Figures 21–24 and Table 5 use the small network $3072{-}64{-}64{-}10$ (tanh, 5000 CIFAR-10 images, MSE, full batch, 4000 steps) at $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\eta = 10^{-3}$, $\varepsilon = 10^{-8}$, seed 0, and change one factor at a time: β_1 (Figure 21), β_2 and ε (Figure 22), η (Figure 23), and the activation, dataset, training-set size and depth (Figure 24). In all sixteen settings the median gradient cosine over steps 2000–4000 is strongly negative, close to -1.

C.3 Residual, attention and recurrent networks

At $\beta_1 = 0$ all three networks reverse at every step once at the edge (Figures 25 and 26, Table 6): $\cos(g_t, g_{t-1})$ is close to -1 and $\cos(g_t, g_{t-2})$ close to $+1$; on the LSTM, loss spikes interrupt the orbit. At $\beta_1 = 0.9$ the vision networks reverse and their momentum inherits it, while the LSTM, still off the edge, reverses on only part of its steps and its momentum not at all.

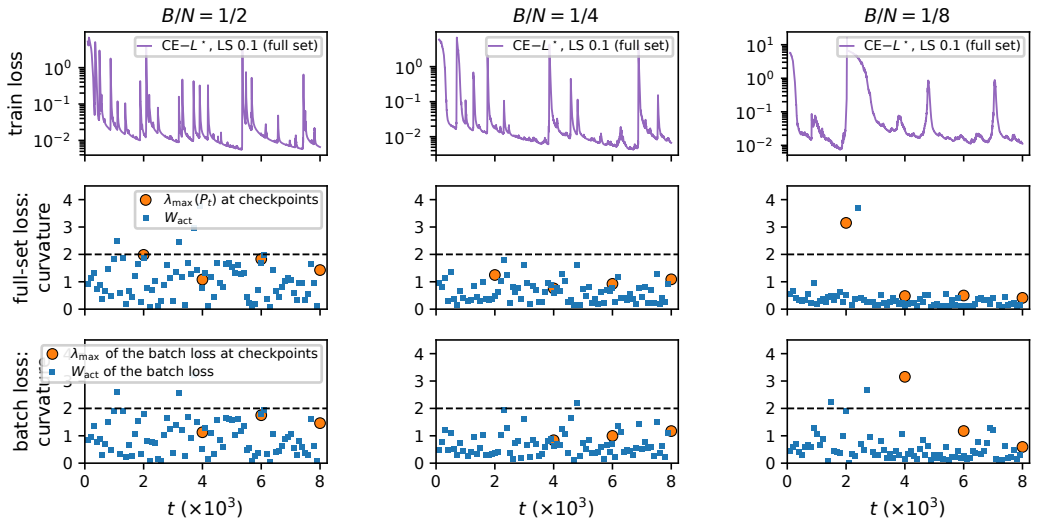

Figure 17: Batch-size scan on GPT-2 medium at $\beta_1 = 0.9$, $\beta_2 = 0.999$ (otherwise as Figure 16: minibatch training on 32, 16 or 8 of the fixed 64 sequences per step, 8000 steps; the full-batch run is Figure 11(b)). Top: training loss on all 64 sequences, excess over the floor 1.4076 (purple). Middle: full-set readings, exact W_{act} (blue squares, every 100 steps) and the full-set sharpness recomputed at the checkpoints x_{2000}–x_{8000} (orange circles). Bottom: the readings on the step's own minibatch loss, W_{act} of the batch loss (blue squares, every 100 steps) and its sharpness at the checkpoints x_{4000}–x_{8000} (orange circles). Dashed lines at 2.

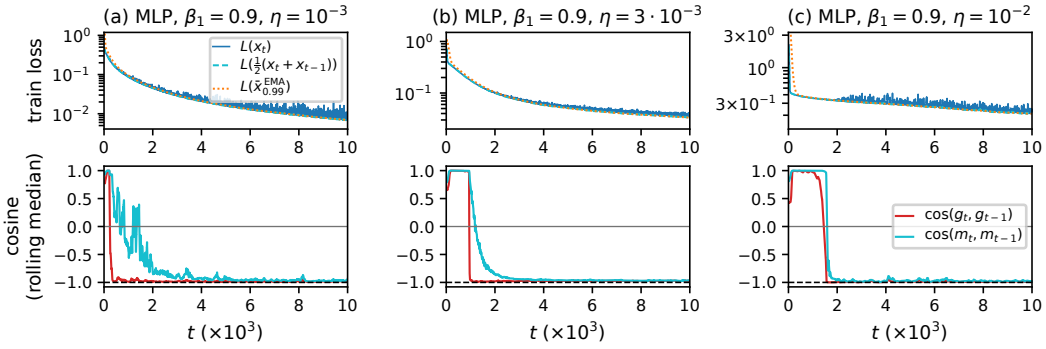

Figure 18: Width-200 fc-tanh MLP on CIFAR-10, full-batch MSE, $\beta_1 = 0.9$, $\beta_2 = 0.999$, at $\eta = 10^{-3}$, $3 \cdot 10^{-3}$ and 10^{-2} (first 10,000 steps). Top: training loss at the iterate $L(x_t)$ (blue, solid), at the midpoint of consecutive iterates (cyan, dashed) and at the EMA iterate (orange, dotted). Bottom: $\cos(g_t, g_{t-1})$ (red) and $\cos(m_t, m_{t-1})$ (cyan). The run protocol and the numbers are in Appendix C.1.

C.4 GPT-2 MEDIUM

Table 7 and Figure 27 give the reversal on the three runs of Appendix B.4; Figure 28 gives the traces of the two runs of Figure 4. At $\beta_1 = 0$ the reversal is clean: the gradient cosine stays close to -1 and most steps reverse. At $\beta_1 = 0.9$, $\beta_2 = 0.999$ it is intermittent, and the momentum inherits it only partly.

C.5 THE ADAM FAMILY

Figures 29 and 30 and Table 8 give the reversal of the ten variants of Appendix B.5.1, and of six of them rerun at $\beta_1 = 0$. The gradient reverses in every configuration that reaches the edge

Table 4: Batch-fraction scan: late-window medians of the active curvature and of the sharpness. MLP: width 200, $N = 5000$, steps 20,000–30,000. GPT-2 medium: fixed 64 sequences, exact full-set W_{act}, steps 4000–8000. The full-batch GPT-2 rows are the runs of Appendix B.4 over the same window; on the subsampled runs the sharpness was not recorded during training and is recomputed at their saved checkpoints x_{4000}, x_{6000} and x_{8000}. Minibatch-loss columns: the same two readings taken on the loss of a minibatch of the run's size, with its own gradient and Hessian. On the MLP both are logged along the run (W_{act} every step, $\lambda_{\max}$ by Lanczos every 20 steps, on the step's own batch).

			full-set loss		minibatch loss	
model	β_1	B/N	W_{act}	$\lambda_{\max}(P_t)$	W_{act}	$\lambda_{\max}$
GPT-2	0	full	2.14	2.76	2.14	2.76
GPT-2	0	1/2	2.11	15.09	2.59	19.18
GPT-2	0	1/4	3.58	21.12	3.36	10.80
GPT-2	0	1/8	4.22	29.69	6.36	58.58
GPT-2	0.9	full	1.82	1.91	1.82	1.91
GPT-2	0.9	1/2	0.94	1.43	0.63	1.46
GPT-2	0.9	1/4	0.46	0.92	0.57	1.00
GPT-2	0.9	1/8	0.18	0.48	0.29	1.17
MLP	0	full	2.02	8.16	2.02	8.16
MLP	0	0.2	2.00	5.87	2.05	15.11
MLP	0	0.05	1.96	3.62	2.22	19.92
MLP	0	0.01	1.80	1.95	2.53	20.60
MLP	0.9	full	1.98	2.01	1.98	2.01
MLP	0.9	0.7	1.71	1.98	1.72	2.04
MLP	0.9	0.4	1.55	1.93	1.54	1.99
MLP	0.9	0.2	1.39	1.82	1.36	1.88
MLP	0.9	0.05	0.81	1.12	0.74	1.36
MLP	0.9	0.01	0.21	0.31	0.22	1.08

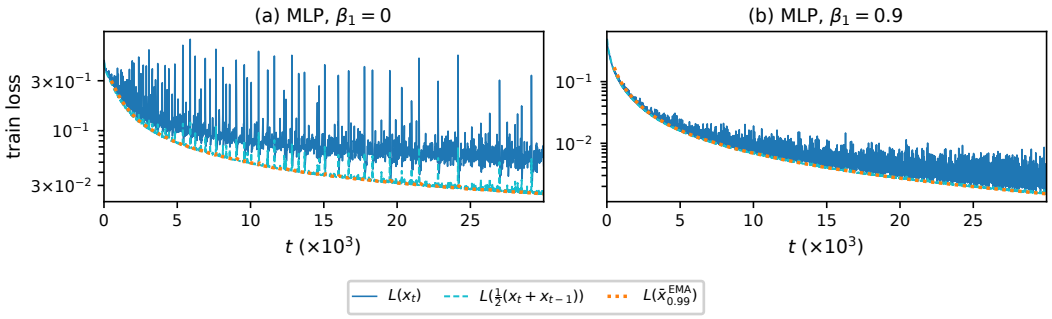

Figure 19: Width-200 fc-tanh MLP on CIFAR-10, full-batch MSE, $\eta = 10^{-3}$, at $\beta_1 = 0$ (a) and $\beta_1 = 0.9$ (b): loss at the iterate x_t (blue, solid), at the midpoint of consecutive iterates (cyan, dashed) and at the EMA of the iterate with rate 0.99 (orange, dotted). The orbit height $E_{\text{osc}} = L(x_t) - L(\text{mid})$ settles to a constant, at once at $\beta_1 = 0$ and after about 3000 steps at $\beta_1 = 0.9$, while the loss keeps falling (Figure 20); details are in Appendix C.1.

and stays there, including Adam, AdamW, AMSGrad, Nadam, Padam, Adagrad and RMSProp. The exception is Adam with coupled L_2 decay, which stays below the threshold. RMSProp with momentum reaches the edge only late and starts to reverse near the end of the run.

C.6 Batch size

Figures 32–35, Table 9 and Figure 31 give the reversal on the subsampled runs of Appendix B.5.3, over the same windows.At $\beta_1 = 0$ the reversal survives every batch size on the MLP and. At $\beta_1 = 0.9$ it weakens as the batch shrinks, together with W_{act}, and disappears at milder subsampling on GPT-2 than on the MLP. Noise alone weakens the reversal only slowly; noise with momentum removes it.

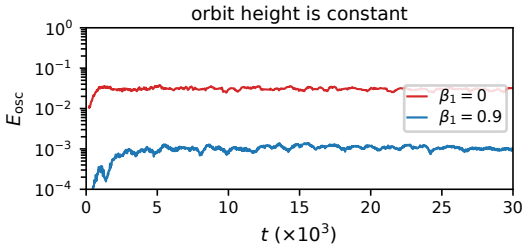

Figure 20: The orbit height $E_{\text{osc}} = L(x_t) - L(\frac{1}{2}(x_t + x_{t-1}))$ of the two runs of Figure 19, as a 501-step rolling median: $\beta_1 = 0$ (red) and $\beta_1 = 0.9$ (blue). It is flat from step 1000 ($\beta_1 = 0$) and from about step 3000 ($\beta_1 = 0.9$) on, while the loss at the iterate keeps falling, by a factor of 6 or more.

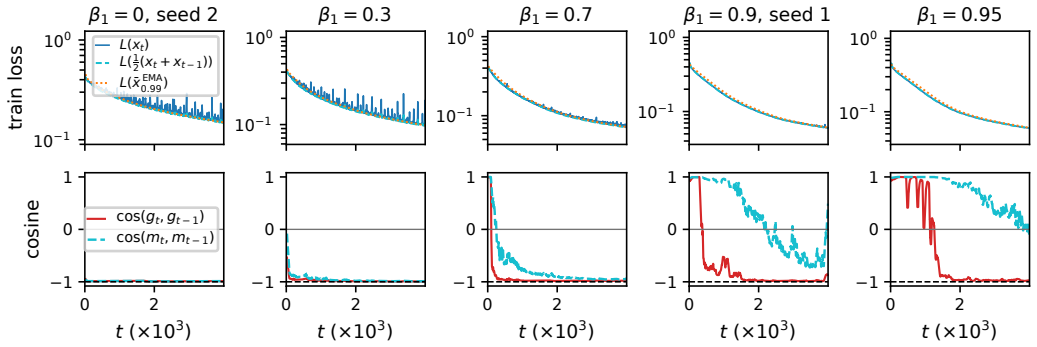

Figure 21: Momentum. The small network at the reference setting of Appendix C.2 with $\beta_1 \in \{0, 0.3, 0.7, 0.9, 0.95\}$; the $\beta_1 = 0$ and $\beta_1 = 0.9$ runs use seeds 2 and 1, the others seed 0. Top: training loss at the iterate (blue, solid), at the midpoint of consecutive iterates $\frac{1}{2}(x_t + x_{t-1})$ (cyan, dashed) and at the EMA of the iterate with rate 0.99 (orange, dotted). Bottom: 101-step rolling medians of $\cos(g_t, g_{t-1})$ (red) and $\cos(m_t, m_{t-1})$ (cyan, dashed); the dashed line is -1.

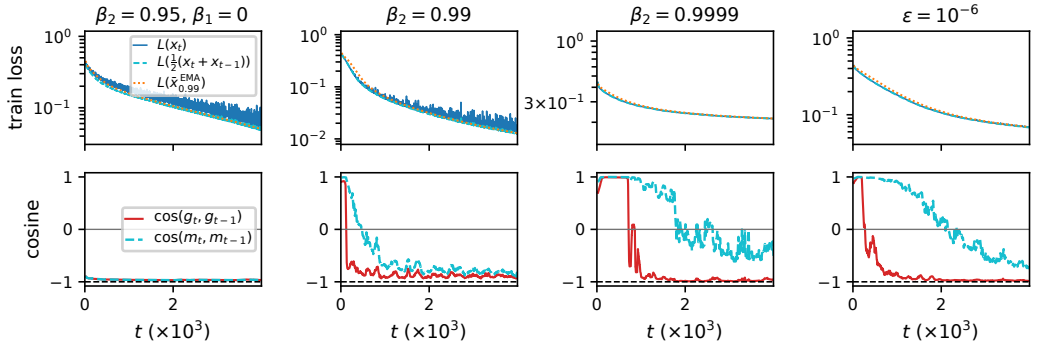

Figure 22: Second-moment constants. The network and readings of Figure 21 with $\beta_2 = 0.95$ (at $\beta_1 = 0$), 0.99 and 0.9999, and with $\varepsilon = 10^{-6}$.

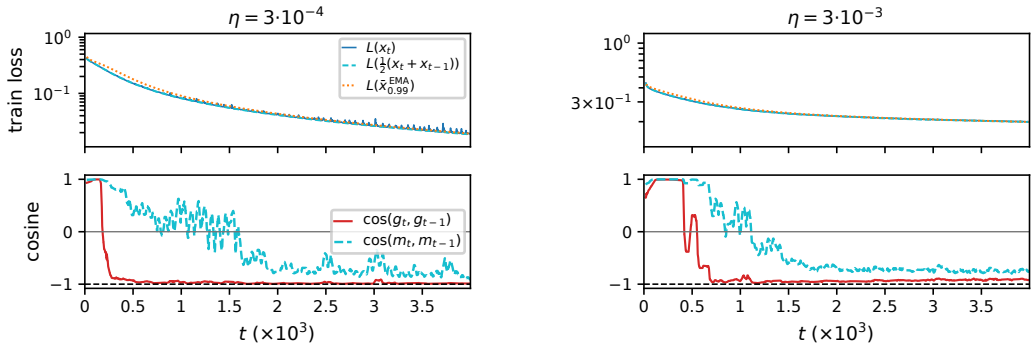

Figure 23: Learning rate. The network and readings of Figure 21 at $\eta = 3 \cdot 10^{-4}$ and $3 \cdot 10^{-3}$.

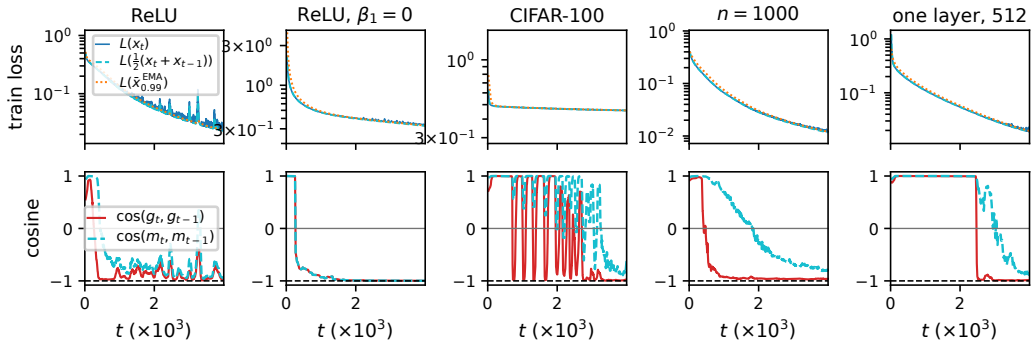

Figure 24: Network and data. The readings of Figure 21 with ReLU (at $\beta_1 = 0.9$ and 0), on CIFAR-100, with $n = 1000$ training images, and with a single hidden layer of 512.

Table 5: The sixteen settings of Figures 21–24, grouped as the figures: medians over steps 2000–4000 of $\cos(g_t, g_{t-1})$, $\cos(g_t, g_{t-2})$ and $\cos(m_t, m_{t-1})$, the fraction of steps with $\cos(g_t, g_{t-1}) < -0.5$, and the training loss (median over steps 3800–3999).

setting	β_1	β_2	η	$\cos(g_t, g_{t-1})$	$\cos(g_t, g_{t-2})$	$\cos(m_t, m_{t-1})$	reversal steps	loss
seed 2, β_1=0	0	0.999	0.001	-0.993	+0.990	-0.993	100%	0.156
β_1=0.3	0.3	0.999	0.001	-0.992	+0.982	-0.985	100%	0.102
β_1=0.7	0.7	0.999	0.001	-0.982	+0.938	-0.941	100%	0.0742
seed 1	0.9	0.999	0.001	-0.982	+0.943	-0.402	99%	0.0611
β_1=0.95	0.95	0.999	0.001	-0.971	+0.908	+0.480	98%	0.0608
β_2=0.95, β_1=0	0	0.95	0.001	-0.963	+0.904	-0.963	100%	0.062
β_2=0.99	0.9	0.99	0.001	-0.891	+0.553	-0.801	89%	0.0152
β_2=0.9999	0.9	0.9999	0.001	-0.977	+0.938	-0.347	98%	0.216
ε=10^{-6}	0.9	0.999	0.001	-0.972	+0.905	-0.466	100%	0.0696
η=$3{\cdot}10^{-4}$	0.9	0.999	0.0003	-0.985	+0.944	-0.753	97%	0.0196
η=$3{\cdot}10^{-3}$	0.9	0.999	0.003	-0.927	+0.796	-0.734	99%	0.201
ReLU	0.9	0.999	0.001	-0.910	+0.697	-0.838	88%	0.0264
ReLU, β_1=0	0	0.999	0.001	-0.997	+0.999	-0.997	100%	0.34
CIFAR-100	0.9	0.999	0.001	-0.942	+0.933	-0.223	73%	0.458
n=1000	0.9	0.999	0.001	-0.963	+0.866	-0.645	100%	0.0126
one layer, 512	0.9	0.999	0.001	-0.969	+0.989	+0.006	74%	0.0204

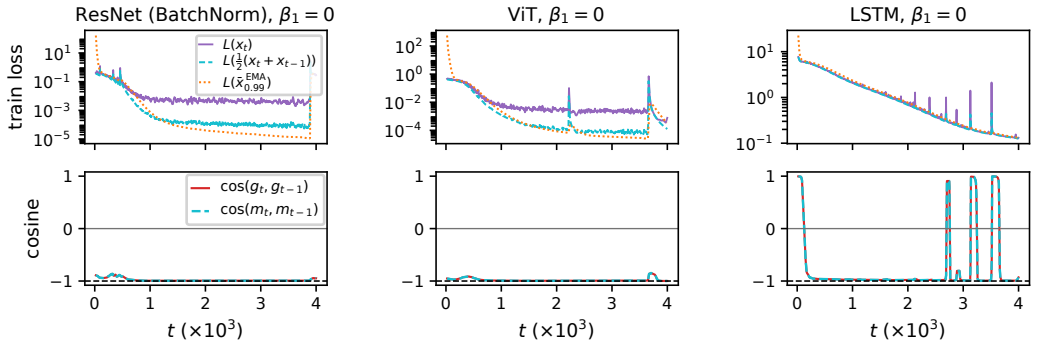

Figure 25: Reversal on the three runs of Figure 9 ($\beta_1 = 0$). Top: training loss at the iterate (purple, solid), at the midpoint of consecutive iterates (cyan, dashed) and at the EMA of the iterate with rate 0.99 (orange, dotted); MSE for the vision networks, excess cross-entropy over the floor 1.4075 for the LSTM. Bottom: 101-step rolling medians of $\cos(g_t, g_{t-1})$ (red, solid) and $\cos(m_t, m_{t-1})$ (cyan, dashed); dashed line at -1. At $\beta_1 = 0$ the momentum is the gradient, $m_{t+1} = g_t$, so the two curves coincide.

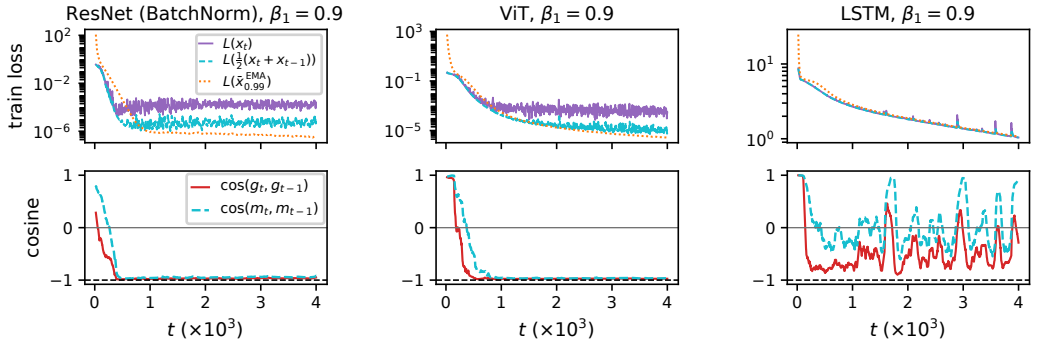

Figure 26: As Figure 25 for the three runs of Figure 10 ($\beta_1 = 0.9$).

Table 6: Second-half medians (steps 2000–3999) of the cosines for the six runs of Figures 9 and 10, and the fraction of steps with $\cos(g_t, g_{t-1}) < -0.5$.

architecture	β_1	$\cos(g_t, g_{t-1})$	$\cos(g_t, g_{t-2})$	$\cos(m_t, m_{t-1})$	reversal steps
ResNet-BN	0	-0.998	+0.994	-0.998	100%
ResNet-BN	0.9	-0.966	+0.872	-0.944	100%
ViT	0	-0.996	+0.990	-0.996	100%
ViT	0.9	-0.975	+0.905	-0.966	100%
LSTM	0	-0.986	+0.991	-0.986	79%
LSTM	0.9	-0.498	-0.286	-0.027	50%

Table 7: Gradient reversal on the three full-batch GPT-2 runs of Appendix B.4: median $\cos(g_t, g_{t-1})$ and $\cos(m_t, m_{t-1})$, and fraction of steps with $\cos(g_t, g_{t-1}) < -0.5$ over every step of the window; at $\beta_1 = 0$ the momentum is the gradient.

β_1	β_2	steps	$\cos(g_t, g_{t-1})$	$\cos(m_t, m_{t-1})$	reversal
0	0.95	2,000–4,000	-0.97	-0.97	89%
		4,000–8,000	-0.98	-0.98	93%
		8,000–8,600	-0.99	-0.99	94%
0.9	0.999	2,000–8,000	-0.90	0.36	61%
		8,000–12,000	-0.97	-0.41	78%
		12,000–20,000	-0.93	-0.30	75%
		20,000–30,000	-0.82	-0.37	73%
0.9	0.95	2,000–8,000	-0.01	0.71	19%
		8,000–12,000	0.01	0.36	5%
		12,000–20,000	0.09	0.49	2%
		20,000–30,000	0.12	0.47	2%

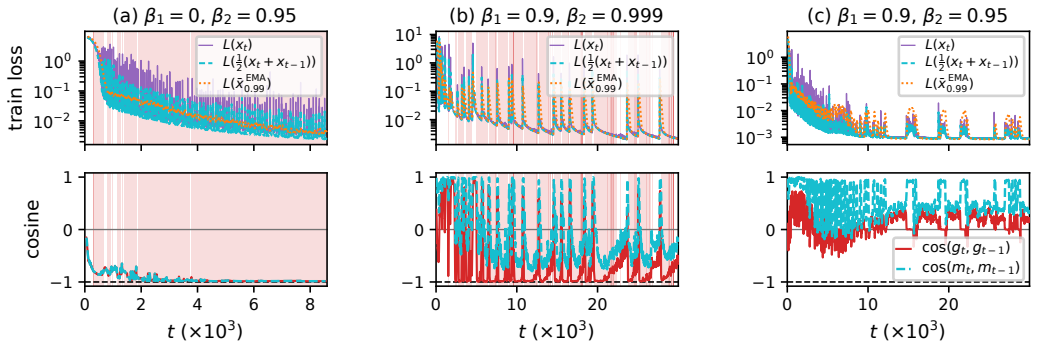

Figure 27: Gradient reversal along the three full-batch GPT-2 runs of Figure 11: (a) $\beta_1 = 0$, $\beta_2 = 0.95$, 8600 steps; (b) $\beta_1 = 0.9$, $\beta_2 = 0.999$, and (c) $\beta_1 = 0.9$, $\beta_2 = 0.95$, 30,000 steps. Top: training loss, excess over the label-smoothing floor 1.4076, at the iterate (purple, solid), at the midpoint of consecutive iterates (cyan, dashed) and at the EMA of the iterate with rate 0.99 (orange, dotted). Bottom: 101-step rolling medians of $\cos(g_t, g_{t-1})$ (red, solid) and $\cos(m_t, m_{t-1})$ (cyan, dashed); the dashed line is -1; at $\beta_1 = 0$ the two curves coincide. Red shading in both rows: stretches on which $\cos(g_t, g_{t-1}) < -0.8$ on the majority of a 51-step window, as in Figure 4; no stretch qualifies in (c).

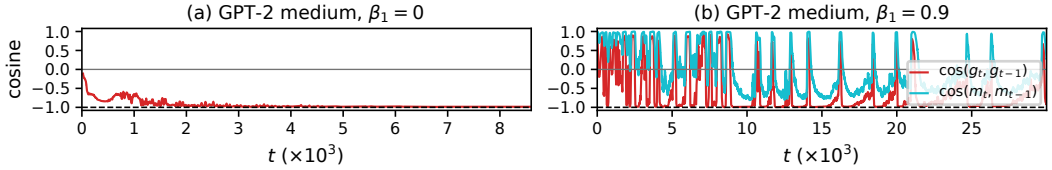

Figure 28: Cosine traces of the two full-batch GPT-2 runs of Figure 4: 101-step rolling medians of $\cos(g_t, g_{t-1})$ (red) and $\cos(m_t, m_{t-1})$ (cyan); the dashed line is -1. (a) $\beta_1 = 0$, $\beta_2 = 0.95$ (b) $\beta_1 = 0.9$, $\beta_2 = 0.999$

Table 8: Reversal of the ten variants of Table 3, last quarter of the run (steps 2250–2999): medians of $\cos(g_t, g_{t-1})$, $\cos(g_t, g_{t-2})$ and $\cos(m_t, m_{t-1})$, the fraction of steps with $\cos(g_t, g_{t-1}) < -0.5$ ("reversal"), and the means of the normalized W_{act} and $\lambda_{\max}$, the statistics of Table 3.

	$\beta_1 = 0.9$ or $\mu = 0.9$						$\beta_1 = 0$				
variant	$\cos(g_t, g_{t-1})$	$\cos(g_t, g_{t-2})$	$\cos(m_t, m_{t-1})$	reversal	W_{act}	$\lambda_{\max}$	$\cos(g_t, g_{t-1})$	$\cos(g_t, g_{t-2})$	reversal	W_{act}	$\lambda_{\max}$
Adam (uncorrected)	-0.987	+0.947	-0.909	99%	1.97	2.01	-0.998	+0.995	100%	2.05	3.70
Adam, bias-corrected	-0.959	+0.832	-0.905	99%	1.94	2.02	-0.983	+0.982	100%	2.28	2.84
AdamW	-0.972	+0.872	-0.753	94%	1.92	2.00	-0.996	+0.994	100%	2.09	3.69
Adam + L_2	+0.677	+0.942	-0.128	1%	1.76	1.87	-0.119	+0.844	3%	1.94	2.90
Padam ($p = 1/4$)	-0.982	+0.930	-0.957	100%	1.97	2.01	-1.000	+0.999	100%	2.00	2.06
Nadam	-0.998	+0.992	-0.989	100%	2.00	2.09	–	–	–	–	–
RMSProp	-0.998	+0.995	–	100%	2.06	3.95	–	–	–	–	–
RMSProp + momentum	-0.231	+0.322	+0.392	37%	2.04	3.80	–	–	–	–	–
AMSGrad	-0.993	+0.981	-0.698	100%	1.98	2.00	-0.998	+0.998	100%	2.06	2.94
Adagrad	-0.998	+0.998	–	100%	2.03	2.08	–	–	–	–	–

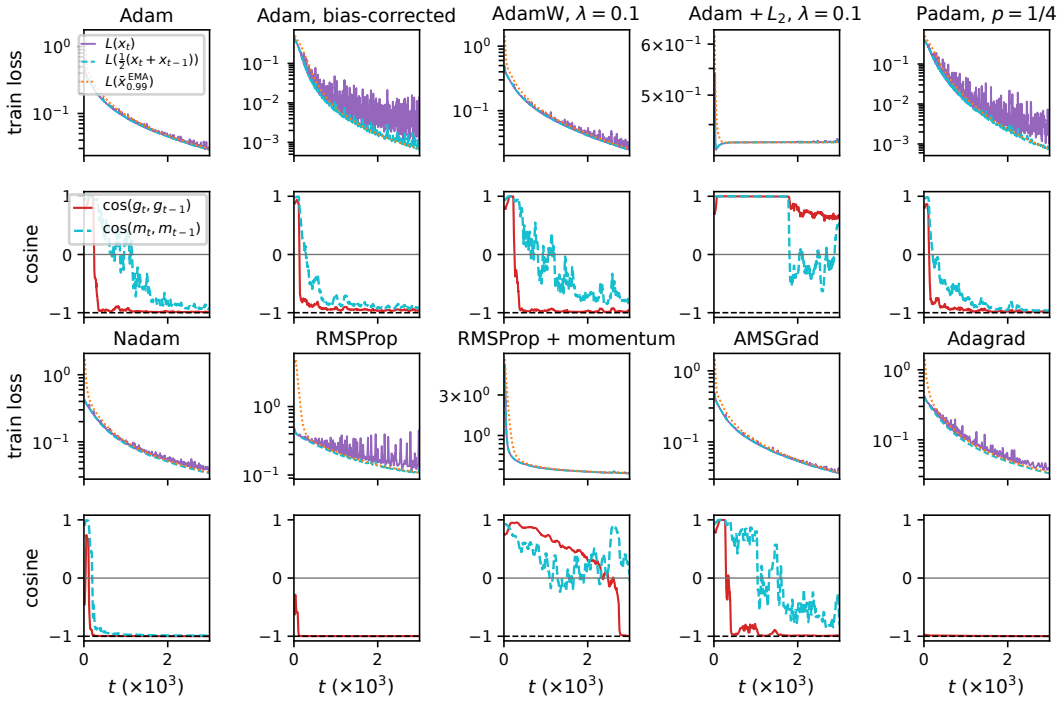

Figure 29: Reversal on the ten configurations of Figure 12 ($\beta_1 = 0.9$ or $\mu = 0.9$), one column per optimizer in the order of Table 3. Top of each block: training loss at the iterate (purple, solid), at the midpoint of consecutive iterates (cyan, dashed) and at the EMA of the iterate with rate 0.99 (orange, dotted). Bottom: 101-step rolling medians of $\cos(g_t, g_{t-1})$ (red, solid) and $\cos(m_t, m_{t-1})$ (cyan, dashed); dashed line at -1. RMSProp and Adagrad have no momentum and show the gradient cosine only.

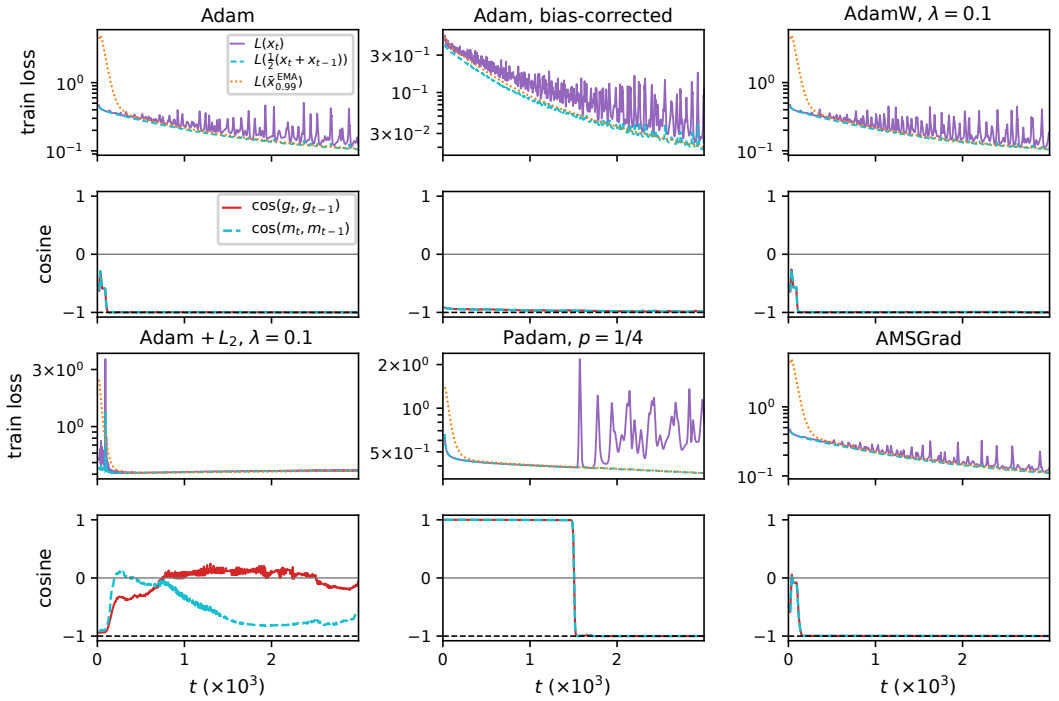

Figure 30: As Figure 29 at $\beta_1 = 0$ for the six variants that are run again there (Table 8); $m_{t+1} = g_t$, so the two cosine curves coincide.

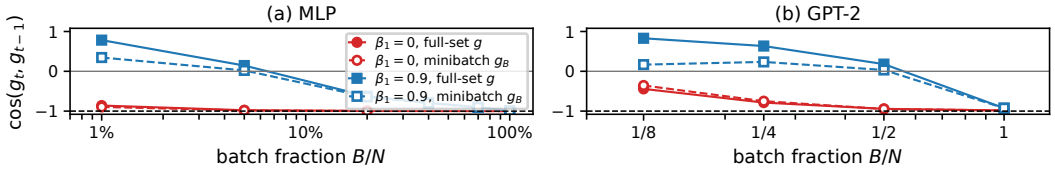

Figure 31: Gradient cosine against the batch fraction B/N, late-window medians, for (a) the width-200 MLP and (b) GPT-2 medium, at $\beta_1 = 0$ (red circles) and $\beta_1 = 0.9$ (blue squares). Filled markers, solid lines: consecutive full-set gradients; hollow markers, dashed lines: consecutive mini-batch gradients along the same runs (at $B/N = 1$ the two coincide). Dashed horizontal line at -1.

Table 9: Batch-fraction scan, the reversal read on the full-set gradient and on the minibatch gradient of the step: late-window medians of $\cos(g_t, g_{t-1})$ and of $\cos(g_{B,t}, g_{B,t-1})$, the fraction of steps on which each is below -0.5 ("reversal"), and the alignment $\cos(g_{B,t}, g_t)$ of the batch gradient with the full gradient at the same iterate. MLP: width 200, $N = 5000$, steps 20,000–30,000, every column logged at every step. GPT-2 medium: fixed 64 sequences, steps 4000–8000.

			full-set gradient		minibatch gradient		
model	β_1	B/N	$\cos(g_t, g_{t-1})$	reversal	$\cos(g_{B,t}, g_{B,t-1})$	reversal	$\cos(g_{B,t}, g_t)$
GPT-2	0	full	-0.98	93%	-0.98	93%	+1.00
GPT-2	0	1/2	-0.94	90%	-0.95	89%	+0.95
GPT-2	0	1/4	-0.79	75%	-0.75	70%	+0.90
GPT-2	0	1/8	-0.44	43%	-0.36	30%	+0.62
GPT-2	0.9	full	-0.93	65%	-0.93	65%	+1.00
GPT-2	0.9	1/2	+0.18	28%	+0.03	26%	+0.88
GPT-2	0.9	1/4	+0.63	4%	+0.24	3%	+0.89
GPT-2	0.9	1/8	+0.83	0%	+0.17	0%	+0.48
MLP	0	full	-1.00	100%	-1.00	100%	+1.00
MLP	0	0.2	-1.00	100%	-1.00	100%	+1.00
MLP	0	0.05	-0.98	100%	-0.98	100%	+0.99
MLP	0	0.01	-0.86	99%	-0.90	100%	+0.95
MLP	0.9	full	-0.99	100%	-0.99	100%	+1.00
MLP	0.9	0.7	-0.90	99%	-0.90	99%	+1.00
MLP	0.9	0.4	-0.79	92%	-0.80	93%	+0.99
MLP	0.9	0.2	-0.63	69%	-0.64	72%	+0.99
MLP	0.9	0.05	+0.14	3%	+0.02	4%	+0.95
MLP	0.9	0.01	+0.78	0%	+0.35	0%	+0.76

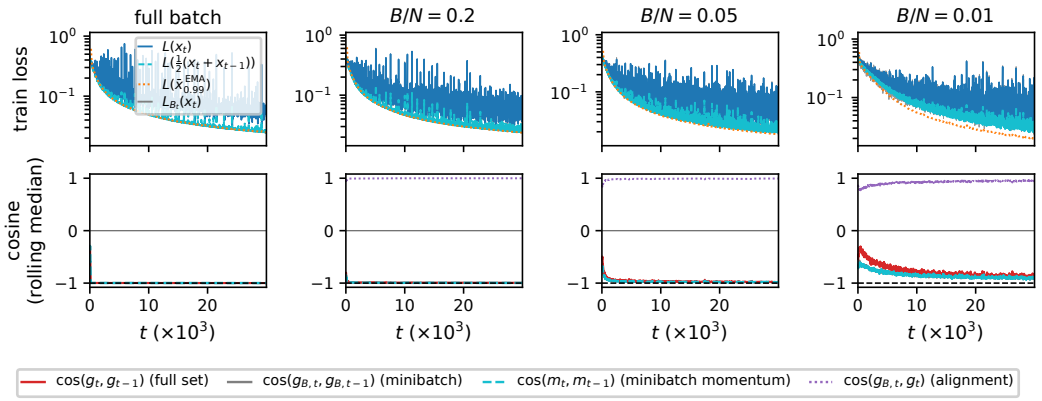

Figure 32: Gradient reversal in the batch-size scan on the width-200 MLP, $\beta_1 = 0$, the runs of Table 9 ($\eta = 10^{-3}$, MSE, 30,000 steps; at each step Adam sees a random subset of size B of the fixed $N = 5000$ images, $B/N = 1$, 0.2, 0.05, 0.01). Top: the training loss on all N images at the iterate (blue, solid), at the midpoint of consecutive iterates (cyan, dashed) and at the EMA of the iterate with rate 0.99 (orange, dotted), and the loss of the step's own minibatch (grey, thin). Bottom: 101-step rolling medians of the cosine of consecutive full-set gradients $\cos(g_t, g_{t-1})$ (red), of consecutive momenta $\cos(m_t, m_{t-1})$ (cyan, dashed; at $\beta_1 = 0$ the momentum is the gradient, $m_{t+1} = g_t$, and the two curves coincide), of consecutive minibatch gradients $\cos(g_{B,t}, g_{B,t-1})$ (grey) and of the alignment $\cos(g_{B,t}, g_t)$ of the batch gradient with the full gradient at the same iterate (purple, dotted). Dashed line at -1. In the full-batch panel the minibatch is the full set and only the full-set curves are drawn.

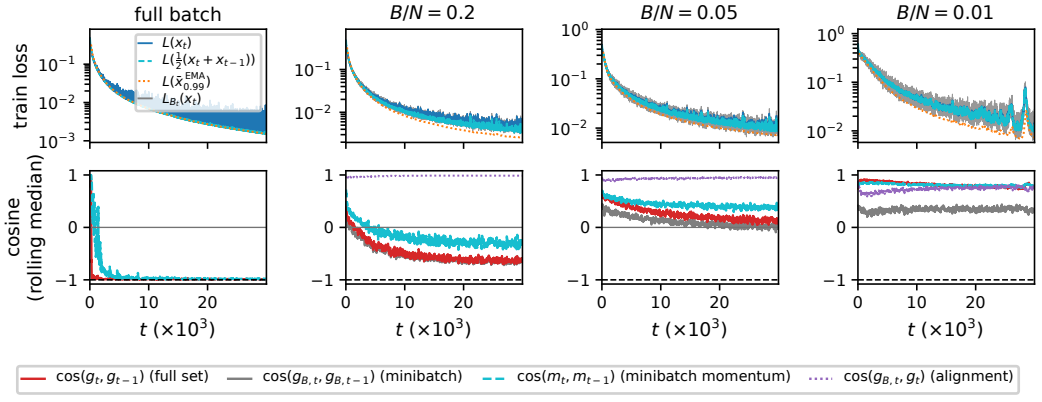

Figure 33: As Figure 32 at $\beta_1 = 0.9$ (the same fractions and readings).

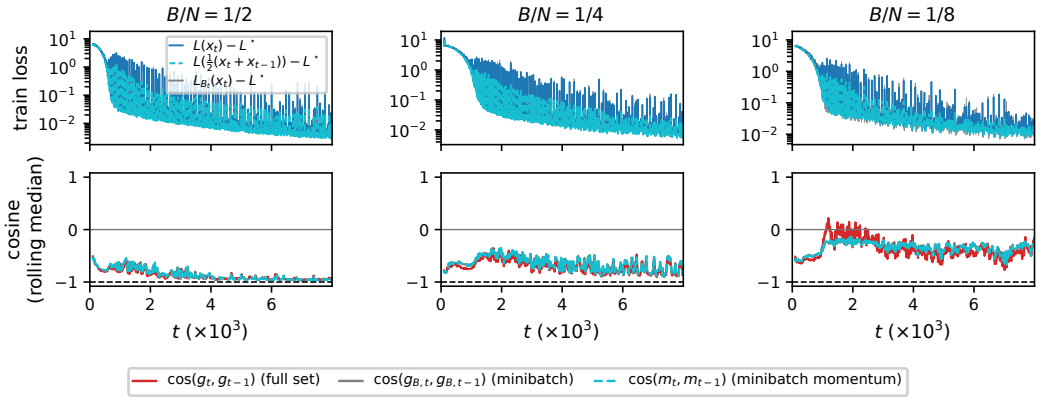

Figure 34: Gradient reversal in the batch-size scan on GPT-2 medium, $\beta_1 = 0$, $\beta_2 = 0.95$, 8000 steps, the runs of Table 9 (at each step Adam sees $B = 32$, 16 or 8 of the fixed 64 training sequences; the full-batch run is Figure 27(a)). Top: the training loss on all 64 sequences as excess over the label-smoothing floor 1.4076, at the iterate (blue, solid) and at the midpoint of consecutive iterates (cyan, dashed), and the loss of the step's own minibatch (grey, thin); 101-step rolling medians of $\cos(g_t, g_{t-1})$ on the full set (red), of $\cos(m_t, m_{t-1})$ (cyan, dashed; at $\beta_1 = 0$ it coincides with the gradient cosine) and of $\cos(g_{B,t}, g_{B,t-1})$ on the minibatches (grey). The alignment $\cos(g_{B,t}, g_t)$ was not logged along these runs; its checkpoint values are in Table 9. Dashed line at -1.

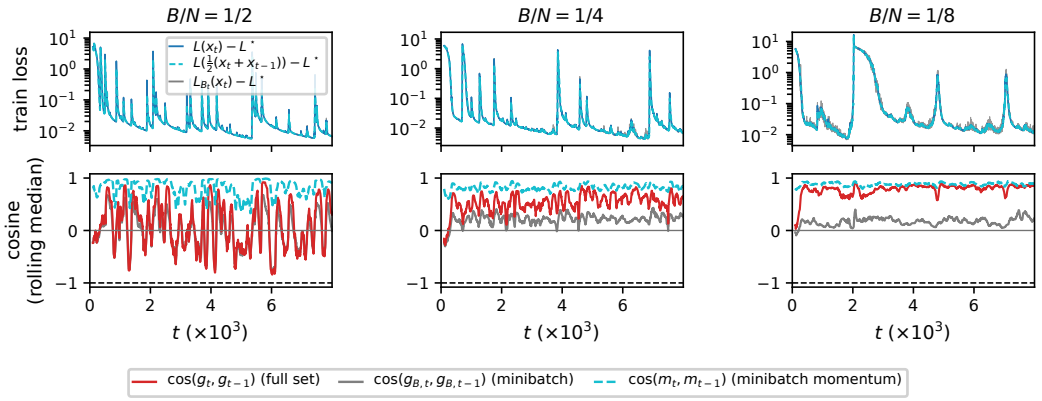

Figure 35: As Figure 34 at $\beta_1 = 0.9$, $\beta_2 = 0.999$

D ADDITIONAL PRELIMINARIES

Freezing the second-moment variable turns one Adam step into a fixed map on position and momentum,

$$\mathcal{A}_v : \begin{pmatrix} x \\ m \end{pmatrix} \longmapsto \begin{pmatrix} x - \eta\big(\mathrm{diag}(\sqrt{v}) + \varepsilon I\big)^{-1}\big(\beta_1 m + (1-\beta_1)\nabla f(x)\big) \\ \beta_1 m + (1-\beta_1)\nabla f(x) \end{pmatrix}, \tag{19}$$

whose linearization is governed by a preconditioned Hessian. The following lemma is stated for a general twice continuously differentiable f, with $\mathsf{S}_t = \lambda_{\max}(D_t^{-1}H_t)$ and $\mathsf{S}^\star$ as in (3).

Lemma D.1 (Frozen stability)**.** *Suppose $H_t \succ 0$. The linearization of $\mathcal{A}_{v_{t+1}}$ at (x_t, m_t) is Schur stable if and only if $\mathsf{S}_t < \mathsf{S}^\star$.*

Proof. Fix t and write $D := D_t$, $H := H_t$ and $\widetilde{H} := D^{-1/2}HD^{-1/2}$, which is symmetric positive definite and has the same eigenvalues as $D^{-1}H$. In the coordinates $x' = D^{1/2}\Delta x$ and $m' = D^{-1/2}\Delta m$, the linearization of (19) at (x_t, m_t) reads

$$m'_+ = \beta_1 m' + (1-\beta_1)\widetilde{H}x', \qquad x'_+ = x' - \eta m'_+. \tag{20}$$

In the eigenbasis of $\widetilde{H}$ this decouples into 2×2 blocks

$$\begin{pmatrix} 1-\eta(1-\beta_1)\sigma & -\eta\beta_1 \\ (1-\beta_1)\sigma & \beta_1 \end{pmatrix}, \tag{21}$$

one per eigenvalue $\sigma > 0$ of $\widetilde{H}$, with characteristic polynomial

$$z^2 - \big(1 + \beta_1 - \eta(1-\beta_1)\sigma\big)z + \beta_1. \tag{22}$$

For a real quadratic $z^2 + a_1 z + a_0$ the Jury criterion places both roots inside the open unit disc exactly when $|a_0| < 1$ and $|a_1| < 1 + a_0$. Here $a_0 = \beta_1 \in [0,1)$, and the second condition reads $|1 + \beta_1 - \eta(1-\beta_1)\sigma| < 1 + \beta_1$, that is

$$0 < \eta(1-\beta_1)\sigma < 2(1+\beta_1). \tag{23}$$

This holds for every mode if and only if it holds for the largest, $\sigma = \mathsf{S}_t$, which is $\mathsf{S}_t < \mathsf{S}^\star$. □

The rank-one model. Throughout Appendices D–G we fix the rank-one objective of Section 3.1,

$$f(x) = \tfrac{1}{2}x^\top H x, \qquad H = \lambda u u^\top, \qquad \lambda > 0, \qquad \|u\|_2 = 1, \tag{24}$$

so that $g_t = Hx_t = \lambda x_t^{\parallel} u$, and run (1).

Assumption D.2 (Standing initialization)**.** $m_0 \in \mathrm{span}\{u\}$ and $v_0 = \lambda^2 v_0^{\parallel}\, u \odot u$ for some $v_0^{\parallel} \geq 0$. The standard initialization $m_0 = 0$, $v_0 = 0$ of Section 3.1 is the case $v_0^{\parallel} = 0$.

Both properties propagate, because g_t is a multiple of u: $m_{t+1} = \beta_1 m_t + (1-\beta_1)\lambda x_t^{\parallel} u \in \mathrm{span}\{u\}$, and $v_{t+1} = \beta_2 v_t + (1-\beta_2)\lambda^2 (x_t^{\parallel})^2\, u \odot u$ is again proportional to $u \odot u$. Hence $m_t = m_t^{\parallel} u$ and

$$v_t = \lambda^2 v_t^{\parallel}\, u \odot u, \qquad v_{t+1}^{\parallel} = \beta_2 v_t^{\parallel} + (1-\beta_2)(x_t^{\parallel})^2 \tag{25}$$

for every t, which is (6) (stated there for the standard initialization $v_0^{\parallel} = 0$). The transverse position $x_t^{\perp}$ never feeds back. Under Assumption D.2 the trajectory is therefore described by the three scalars $(x_t^{\parallel}, m_t^{\parallel}, v_t^{\parallel}) \in \mathbb{R}\times\mathbb{R}\times[0,\infty)$, which evolve by (9) and (25) with

$$w_t = F(v_t^{\parallel}), \qquad F(v^{\parallel}) = c\eta\lambda \sum_{i=1}^{d} \frac{u_i^2}{\lambda\,|u_i|\,\sqrt{v^{\parallel}} + \varepsilon}, \tag{26}$$

as in (7). The loss is $f(x_t) = \lambda (x_t^{\parallel})^2/2$. Every statement about the trajectory below is a statement about this three-dimensional recursion. Unrolling (25) gives, for all T and $n \geq 0$,

$$v_{T+n}^{\parallel} = \beta_2^n v_T^{\parallel} + (1-\beta_2)\sum_{j=0}^{n-1} \beta_2^{n-1-j} (x_{T+j}^{\parallel})^2. \tag{27}$$

When u is a coordinate axis, $F(v^{\parallel}) = c\eta\lambda/(\lambda\sqrt{v^{\parallel}} + \varepsilon)$ and the recursion is the one-dimensional Adam recursion.

Lemma D.3 (Properties of the response function)**.** *Let $\varepsilon > 0$.*

1. ***Monotone bijection.*** *F is continuous and strictly decreasing, with $F(0) = w_{\max} = c\eta\lambda/\varepsilon$ and $F(v^{\parallel}) \to 0$ as $v^{\parallel} \to \infty$. Hence for $0 < \omega \le w_{\max}$ the level*
$$Q_\omega := F^{-1}(\omega) \ \ge 0 \tag{28}$$
is well defined, and $w_t \ge \omega \iff v_t^{\parallel} \le Q_\omega$, $w_t \le \omega \iff v_t^{\parallel} \ge Q_\omega$.
2. ***Scaling.*** *For $\kappa \ge 1$ and $v^{\parallel} \ge 0$, $F(\kappa^2 v^{\parallel}) \ge F(v^{\parallel})/\kappa$, with strict inequality when $\kappa > 1$.*
3. ***One-step restriction.*** *For $v^{\parallel} \ge 0$,*
$$\frac{1}{F(\beta_2 v^{\parallel})} - \frac{1}{w_{\max}} \ \ge \ \sqrt{\beta_2}\left(\frac{1}{F(v^{\parallel})} - \frac{1}{w_{\max}}\right). \tag{29}$$

Proof. (i) is read off from (26). (ii) Termwise, $\lambda\,|u_i|\,\kappa\sqrt{v}^{\parallel} + \varepsilon \le \kappa(\lambda\,|u_i|\,\sqrt{v}^{\parallel} + \varepsilon)$, strictly when $\kappa > 1$ because $\varepsilon > 0$. (iii) Put $y_i := \varepsilon/(\lambda\,|u_i|\,\sqrt{v}^{\parallel} + \varepsilon) \in (0, 1]$ and $\alpha_i := u_i^2$, so that $\sum_i \alpha_i = 1$ and $F(v^{\parallel})/w_{\max} = \sum_i \alpha_i y_i$. Substituting $\lambda\,|u_i|\,\sqrt{v}^{\parallel} = \varepsilon(1 - y_i)/y_i$ gives the exact identity
$$\frac{\varepsilon}{\lambda\,|u_i|\,\sqrt{\beta_2 v^{\parallel}} + \varepsilon} = \Phi(y_i), \qquad \Phi(y) := \frac{y}{\sqrt{\beta_2} + (1 - \sqrt{\beta_2})\,y},$$
and Φ is concave on $[0, 1]$: for $\beta_2 > 0$, $\Phi''(y) = -2\sqrt{\beta_2}(1 - \sqrt{\beta_2})\big(\sqrt{\beta_2} + (1 - \sqrt{\beta_2})y\big)^{-3} < 0$, while for $\beta_2 = 0$, $\Phi \equiv 1$ and (29) is trivial. Jensen's inequality with the weights α_i gives $F(\beta_2 v^{\parallel})/w_{\max} = \sum_i \alpha_i \Phi(y_i) \le \Phi\big(F(v^{\parallel})/w_{\max}\big)$, and taking reciprocals is (29). □

The regime of interest is $w_{\max} > 2$; otherwise no supercritical state is reachable at all.

The only nonzero preconditioned curvature. The symmetric representative of the preconditioned Hessian is
$$D_t^{-1/2} H D_t^{-1/2} = \lambda\big(D_t^{-1/2}u\big)\big(D_t^{-1/2}u\big)^{\top}, \tag{30}$$
which has rank one with unique nonzero eigenvalue $\lambda u^{\top} D_t^{-1} u$ and eigenvector $D_t^{-1/2}u$. This is also the only nonzero eigenvalue of $D_t^{-1}H$, which is similar to (30), so $\lambda_{\max}(P_t) = c\eta\lambda u^{\top} D_t^{-1} u = w_{t+1}$. The preconditioned gradient $D_t^{-1/2} g_t = \lambda x_t^{\parallel} D_t^{-1/2} u$ is a multiple of that eigenvector whenever $x_t^{\parallel} \ne 0$, so the unit vector $\hat{z}_t$ of (5) is $\pm D_t^{-1/2}u / \left\| D_t^{-1/2}u \right\|$ and
$$W_{\mathrm{act},t} = \hat{z}_t^{\top} P_t \hat{z}_t = c\eta\lambda\,\frac{\big(u^{\top} D_t^{-1} u\big)^2}{u^{\top} D_t^{-1} u} = c\eta\lambda\, u^{\top} D_t^{-1} u = w_{t+1}, \tag{31}$$
which is the coincidence (7) of the sharpness and the active curvature used in Section 3.1.

D.1 Proof of Proposition 3.1

Proof. By Lemma D.3(i) and $2 < w_{\max}$ the level $x_\star^2 = F^{-1}(2)$ is well defined and $x_\star > 0$ is unique. Start one step from
$$(x_0^{\parallel}, m_0^{\parallel}, v_0^{\parallel}) = (x_\star,\ -c\lambda x_\star,\ x_\star^2).$$
The second moment does not move, $v_1^{\parallel} = \beta_2 x_\star^2 + (1 - \beta_2)(x_0^{\parallel})^2 = x_\star^2$, so
$$w_1 = F(v_1^{\parallel}) = F(x_\star^2) = 2.$$
The momentum recursion of (9) gives, using $1 - \beta_1 = c(1 + \beta_1)$,
$$m_1^{\parallel} = \beta_1 m_0^{\parallel} + (1 - \beta_1)\lambda x_0^{\parallel} = \lambda x_\star\big(-\beta_1 c + 1 - \beta_1\big) = c\lambda x_\star,$$
and the position update then gives
$$x_1^{\parallel} = x_0^{\parallel} - (c\lambda)^{-1} w_1 m_1^{\parallel} = x_\star - 2x_\star = -x_\star.$$
So the state after one step is exactly the second point $(-x_\star,\ c\lambda x_\star,\ x_\star^2)$ of the asserted cycle. The orbit is therefore 2-periodic, with $v_t^{\parallel} \equiv x_\star^2$ and hence $w_t \equiv 2$, which by (7) is $W_{\mathrm{act},t} = \lambda_{\max}(P_t) = 2$ at every step. Finally $g_t = \lambda x_t^{\parallel} u$ alternates in sign, so $\cos(g_t, g_{t+1}) = -1$. □

E PROOFS FOR RMSPROP ($\beta_1 = 0$)

Here we prove the results announced in Section 3.1 for the zero-momentum dynamics: the finite passage times toward the threshold from either side (Appendix E.1, proving Theorem 3.2), and the fact that the reversing band $w > 1$ is entered in finite time and never left (Appendix E.2). Throughout this section $\beta_1 = 0$, hence $c = 1$, $m^{\parallel}_{t+1} = \lambda x^{\parallel}_t$, and (6) reads $x^{\parallel}_{t+1} = (1 - w_{t+1})x^{\parallel}_t$ with $w_{t+1} = F(v^{\parallel}_{t+1})$.

Notation. For $n \geq 1$ and $a, b \geq 0$, let

$$G_n(a,b) := \sum_{j=0}^{n-1} a^j b^{n-1-j} = \begin{cases} \dfrac{a^n - b^n}{a-b}, & a \neq b, \\ n a^{n-1}, & a = b. \end{cases} \tag{32}$$

For a time T, a subcritical target $\widehat{w} < 2$ and a supercritical margin $\delta > 0$, set

$$\sigma_{\widehat{w}}(T) := \inf\{n \geq 0 : w_{T+n} \geq \widehat{w}\}, \quad \tau_\delta(T) := \inf\{n \geq 1 : w_{T+n} < 2 + \delta\}, \quad A_\delta := (1+\delta)^2, \tag{33}$$

with $\inf \emptyset = +\infty$, and

$$M_T := \max\{v^{\parallel}_T, (x^{\parallel}_T)^2\}, \qquad \rho_T := \max\big\{|1 - F(M_T)|, |1 - \widehat{w}|\big\},$$
$$U_n(T) := \beta_2^n v^{\parallel}_T + (1-\beta_2)(x^{\parallel}_T)^2\, G_n(\rho_T^2, \beta_2), \qquad L_n(T) := \beta_2^n v^{\parallel}_T + (1-\beta_2)(x^{\parallel}_T)^2\, G_n(A_\delta, \beta_2). \tag{34}$$

Here $F(M_T)$ is a lower bound on the curvature while $v^{\parallel}_t \leq M_T$, ρ_T is the resulting contraction factor of $\big|x^{\parallel}_t\big|$ below the target, and $U_n(T)$, $L_n(T)$ are the levels of $v^{\parallel}_{T+n}$ after n contracting steps and after n expanding steps. Recall $Q_{\widehat{w}} = F^{-1}(\widehat{w})$ and $Q_\delta := F^{-1}(2+\delta)$ from (28).

E.1 PROOF OF THEOREM 3.2

Subcritical stage. If $w_T \geq \widehat{w}$ then $\sigma_{\widehat{w}}(T) = 0$; we consider $w_T < \widehat{w}$. While w_t stays below $\widehat{w} < 2$ the position contracts at a rate bounded away from one, the contraction drains $v^{\parallel}_t$, and a drained $v^{\parallel}_t$ cannot support a small w_t.

Theorem E.1 (Finite passage to a strict subcritical target)**.** *Assume $\beta_1 = 0$, $w_{\max} > 2$ and $w_T < \widehat{w} < 2$. Then*

$$N_{\mathrm{sub}} := \min\big\{n \geq 1 :\, U_n(T) \leq Q_{\widehat{w}}\big\} < \infty \qquad \textit{and} \qquad 1 \leq \sigma_{\widehat{w}}(T) \leq N_{\mathrm{sub}}. \tag{35}$$

Moreover, with $\gamma_T := \max\{\rho_T^2, \beta_2\} < 1$,

$$N_{\mathrm{sub}} \leq \overline{N}_{\mathrm{sub}} := \begin{cases} \left\lceil \dfrac{\log\big(C_T/Q_{\widehat{w}}\big)}{\log(1/\gamma_T)} \right\rceil, \quad C_T := v^{\parallel}_T + \dfrac{(1-\beta_2)(x^{\parallel}_T)^2}{|\beta_2 - \rho_T^2|}, & \beta_2 \neq \rho_T^2, \\ 1 + \left\lceil \dfrac{\log\big(C_T/Q_{\widehat{w}}\big)}{\log(1/\sqrt{\gamma_T})} \right\rceil, \quad C_T := v^{\parallel}_T + \dfrac{(1-\beta_2)(x^{\parallel}_T)^2}{(1-\sqrt{\gamma_T})^2}, & \beta_2 = \rho_T^2. \end{cases} \tag{36}$$

Proof. Since $w_T < \widehat{w} < 2$ and $w_T = F(v^{\parallel}_T)$, Lemma D.3(i) gives

$$v^{\parallel}_T > Q_{\widehat{w}} > 0.$$

Recall $M_T = \max\{v^{\parallel}_T, (x^{\parallel}_T)^2\}$. Since F is decreasing,

$$0 < F(M_T) \leq w_T < \widehat{w} < 2.$$

Hence

$$\rho_T := \max_{w \in [F(M_T),\, \widehat{w}]} |1 - w| < 1.$$

Now suppose that the trajectory stays below $\widehat{w}$ for the next n steps:

$$w_{T+j} < \widehat{w}, \qquad j = 0, \ldots, n-1.$$

We show inductively that, throughout this interval,

$$v^{\parallel}_{T+j} \le M_T, \qquad |x^{\parallel}_{T+j}| \le \rho_T^j |x^{\parallel}_T|. \tag{37}$$

Indeed, if these bounds hold at step $j-1$, then $v^{\parallel}_{T+j}$ is a convex combination of $v^{\parallel}_{T+j-1}$ and $(x^{\parallel}_{T+j-1})^2$, both bounded by M_T. Thus $v^{\parallel}_{T+j} \le M_T$, so

$$F(M_T) \le w_{T+j} < \widehat{w}.$$

By the definition of ρ_T, $|1 - w_{T+j}| \le \rho_T$, and the scalar recursion gives

$$|x^{\parallel}_{T+j}| \le \rho_T |x^{\parallel}_{T+j-1}| \le \rho_T^j |x^{\parallel}_T|.$$

Unrolling the second-moment recursion therefore yields

$$v^{\parallel}_{T+n} \le \beta_2^n v^{\parallel}_T + (1-\beta_2)(x^{\parallel}_T)^2 \sum_{j=0}^{n-1} \beta_2^{n-1-j} \rho_T^{2j} =: U_n(T). \tag{38}$$

Because $\beta_2 < 1$ and $\rho_T < 1$, we have $U_n(T) \to 0$. Hence there exists a finite N_{sub} such that

$$U_{N_{\mathrm{sub}}}(T) \le Q_{\widehat{w}}.$$

If the trajectory has not crossed $\widehat{w}$ before that time, then (38) gives

$$v^{\parallel}_{T+N_{\mathrm{sub}}} \le Q_{\widehat{w}},$$

and therefore, by Lemma D.3(i),

$$w_{T+N_{\mathrm{sub}}} \ge \widehat{w}.$$

Thus the trajectory must leave the region $w < \widehat{w}$ within at most N_{sub} steps.

It remains only to bound N_{sub} explicitly. Let

$$\gamma_T := \max\{\beta_2, \rho_T^2\} < 1.$$

If $\beta_2 \ne \rho_T^2$, the geometric sum in (38) satisfies

$$U_n(T) \le C_T \gamma_T^n,$$

so it is sufficient to choose

$$n \ge \frac{\log(C_T/Q_{\widehat{w}})}{\log(1/\gamma_T)}.$$

If $\beta_2 = \rho_T^2 = \gamma_T$, the sum contributes an additional factor n, which can be bounded by a slower geometric rate, giving

$$U_n(T) \le C_T \gamma_T^{(n-1)/2}.$$

Hence it suffices that

$$n - 1 \ge \frac{\log(C_T/Q_{\widehat{w}})}{\log(1/\sqrt{\gamma_T})}.$$

This proves the claimed finite-time bound. □

Supercritical stage. As long as w_t remains above $2+\delta$ the position expands geometrically, which forces $v_t^{\parallel}$ upward and eventually makes such a large value of w_t impossible.

Lemma E.2 (Geometric expansion forces exit)**.** *Fix $\delta > 0$ with $2 + \delta \le w_{\max}$, T with $x_T^{\parallel} \ne 0$, and any $0 \le \beta_1 < 1$. Then*

$$N_{\sup} := \min\{n \ge 1 :\ L_n(T) > Q_\delta\} \ \le\ \overline{N}_{\sup},$$
$$\overline{N}_{\sup} := 1 + \left\lfloor \frac{\log\big(1 + (A_\delta - \beta_2) Q_\delta / ((1-\beta_2)(x_T^{\parallel})^2)\big)}{\log A_\delta} \right\rfloor < \infty. \tag{39}$$

If moreover, for some $n \ge 1$,

$$w_{T+k} \ge 2 + \delta \quad \textit{and} \quad \left|x^{\parallel}_{T+k}\right| \ge (1+\delta)\left|x^{\parallel}_{T+k-1}\right|, \qquad k = 1, \dots, n, \tag{40}$$

then $n < N_{\sup}$.

Proof. Suppose the trajectory remains in the supercritical region

$$w_{T+j} \geq 2 + \delta, \qquad j = 0, \ldots, n.$$

Under (40), the scalar state expands geometrically:

$$(x^{\parallel}_{T+j})^2 \geq A_\delta^j (x^{\parallel}_T)^2, \qquad j = 0, \ldots, n,$$

where $A_\delta > 1$.

Unrolling the second-moment recursion therefore gives, for every $k \leq n$,

$$v^{\parallel}_{T+k} \geq \beta_2^k v^{\parallel}_T + (1-\beta_2)(x^{\parallel}_T)^2 \sum_{j=0}^{k-1} \beta_2^{k-1-j} A_\delta^j = L_k(T). \tag{41}$$

Since $A_\delta > \beta_2$ and $v^{\parallel}_T \geq 0$, summing the geometric series gives

$$L_k(T) \geq (1-\beta_2)(x^{\parallel}_T)^2 \frac{A_\delta^k - \beta_2^k}{A_\delta - \beta_2} \geq (1-\beta_2)(x^{\parallel}_T)^2 \frac{A_\delta^k - 1}{A_\delta - \beta_2}.$$

Because $A_\delta > 1$, the lower bound grows without bound with k.

On the other hand, as long as $w_{T+k} \geq 2 + \delta$, Lemma D.3(i) implies

$$v^{\parallel}_{T+k} \leq Q_\delta.$$

Hence the trajectory can remain in the region $w \geq 2 + \delta$ only while

$$L_k(T) \leq Q_\delta.$$

By the definition of $\overline{N}_{\sup}$, the lower bound above exceeds Q_δ at $k = \overline{N}_{\sup}$. Therefore the trajectory must leave the supercritical region before that time, and thus

$$N_{\sup} \leq \overline{N}_{\sup}.$$

□

Theorem E.3 (Finite supercritical exit)**.** *Assume $\beta_1 = 0$, $w_T \geq 2 + \delta$ with $\delta > 0$, and $x^{\parallel}_T \neq 0$. Then $\tau_\delta(T) \leq N_{\sup} \leq \overline{N}_{\sup} < \infty$.*

Proof. Note $2 + \delta \leq w_T \leq w_{\max}$. If $\tau_\delta(T) > n$, then $w_{T+k} \geq 2 + \delta$ for $k = 1, \ldots, n$, and (6) gives $\left|x^{\parallel}_{T+k}\right| = (w_{T+k} - 1)\left|x^{\parallel}_{T+k-1}\right| \geq (1+\delta)\left|x^{\parallel}_{T+k-1}\right|$, so (40) holds and Lemma E.2 gives $n < N_{\sup}$. Hence $\tau_\delta(T) \leq N_{\sup}$. □

The two-sided statement. Theorems E.1 and E.3 are the two halves of Theorem 3.2. Because the two levels may be chosen arbitrarily close to 2, they combine into the restoring statement made after it.

E.2 Eventual reversal at every step

The growth threshold is 2; the sign threshold is 1. The band between them is never re-crossed downward, so from a finite time on the component along u, and with it the gradient, reverses at every step.

Lemma E.4 (Eventual reversal at every step)**.** *Assume $\beta_1 = 0$, $w_{\max} > 1$ and $x^{\parallel}_t \neq 0$ for all t.*

1. ***Forward invariance.*** *If $w_t > 1$ for some $t \geq 1$, then $w_{t+1} > 1$.*
2. ***Finite entry.*** *There is a finite $T \geq 1$ with $w_T > 1$. If $w_{\max} > 2$ and $w_1 < 1$, then $T \leq 1 + \overline{N}_{\mathrm{sub}}$, where $\overline{N}_{\mathrm{sub}}$ is the bound* (36) *of Theorem E.1 at time* 1 *and level $\widehat{w} = 1$.*
3. ***Reversal.*** *For all $t \geq T$: $w_t > 1$, $x^{\parallel}_{t-1} x^{\parallel}_t < 0$ and $\cos(g_{t-1}, g_t) = -1$.*

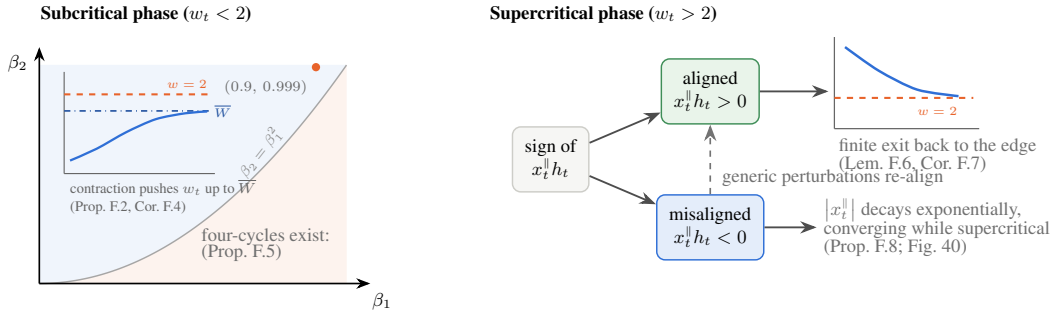

Figure 36: Proof roadmap for $\beta_1 > 0$ (Theorem 3.3). Left: below the edge, an adaptive Lyapunov function certifies contraction up to a cutoff $\overline{W} < 2$ when $\beta_2 > \beta_1^2$ (Proposition F.2, Corollary F.4), while strictly subcritical four-cycles exist for complementary parameter choices when $\varepsilon = 0$ (Proposition F.5). Right: above the edge, the sign of $x_t^{\parallel} h_t$ separates aligned trajectories, which exit every fixed supercritical band in finite time (Lemma F.6, Corollary F.7), from exceptional persistently misaligned trajectories that converge while remaining uniformly supercritical (Proposition F.8).

Proof. (i) Fix $t \geq 1$ with $w_t > 1$. By (25) at time $t-1$, $(1-\beta_2)(x_{t-1}^{\parallel})^2 \leq v_t^{\parallel}$, and by (6), $(x_t^{\parallel})^2 = (w_t - 1)^2 (x_{t-1}^{\parallel})^2$. Hence

$$v_{t+1}^{\parallel} = \beta_2 v_t^{\parallel} + (1-\beta_2)(x_t^{\parallel})^2 \leq \left[\beta_2 + (w_t - 1)^2\right] v_t^{\parallel} \leq w_t^2 \, v_t^{\parallel}, \tag{42}$$

the last step because $w_t^2 - \beta_2 - (w_t - 1)^2 = 2w_t - 1 - \beta_2 > 0$. By Lemma D.3(i) and then (ii) with $\kappa = w_t > 1$,

$$w_{t+1} = F(v_{t+1}^{\parallel}) \;\geq\; F(w_t^2 v_t^{\parallel}) \;>\; \frac{F(v_t^{\parallel})}{w_t} = 1. \tag{43}$$

(ii) By (6), $w_t = 1$ for some $t \geq 1$ would force $x_t^{\parallel} = 0$; so $w_t \neq 1$ for all $t \geq 1$. If $w_1 > 1$ take $T = 1$. Otherwise suppose $w_t < 1$ for all $t \geq 1$. Then $|x_{t+1}^{\parallel}| = (1 - w_{t+1}) |x_t^{\parallel}| \leq |x_t^{\parallel}|$, so as in (37) $v_t^{\parallel} \leq M_1$ and $w_t \geq F(M_1) > 0$ for all $t \geq 1$, whence $|x_t^{\parallel}| \leq (1 - F(M_1))^{t-1} |x_1^{\parallel}| \to 0$; by (27), $v_t^{\parallel} \to 0$ and $w_t \to F(0) = w_{\max} > 1$, contradicting $w_t < 1$. When $w_{\max} > 2$, Theorem E.1 at time 1 with $\widehat{w} = 1$ gives $1 \leq n \leq \overline{N}_{\mathrm{sub}}$ with $w_{1+n} \geq 1$, hence $w_{1+n} > 1$.

(iii) Parts (i) and (ii) give $w_t > 1$ for all $t \geq T$, so (6) reads $x_t^{\parallel} = -(w_t - 1) x_{t-1}^{\parallel}$ with $w_t - 1 > 0$; hence $x_{t-1}^{\parallel} x_t^{\parallel} < 0$, and since $g_t = \lambda x_t^{\parallel} u$, $\cos(g_{t-1}, g_t) = \mathrm{sign}(x_{t-1}^{\parallel} x_t^{\parallel}) = -1$. □

F Proofs for general Adam (Theorem 3.3)

Assume $0 < \beta_1 < 1$ throughout this section.

Figure 36 summarizes the structure of the proofs in this section.

The normalized map. With the normalized momentum coordinate $h_t = (c\lambda)^{-1} m_t^{\parallel} + x_t^{\parallel}$ of Section 3.1, write

$$z_t := (x_t^{\parallel}, h_t)^{\top}. \tag{44}$$

The recursion (9) is then exactly linear in z_t once w_{t+1} is known:

$$z_{t+1} = \mathcal{A}(w_{t+1})\, z_t, \qquad \mathcal{A}(w) := \begin{pmatrix} 1 - w & -\beta_1 w \\ 2 - w & \beta_1 (1 - w) \end{pmatrix}. \tag{45}$$

Indeed, since $1 - \beta_1 = c(1 + \beta_1)$,

$$m_{t+1}^{\parallel} = \beta_1 m_t^{\parallel} + (1-\beta_1)\lambda x_t^{\parallel} = \beta_1 c\lambda\,(h_t - x_t^{\parallel}) + (1-\beta_1)\lambda x_t^{\parallel} = c\lambda\,(x_t^{\parallel} + \beta_1 h_t), \tag{46}$$

so that

$$\begin{aligned} x_{t+1}^{\parallel} &= x_t^{\parallel} - \frac{w_{t+1}}{c\lambda}\, m_{t+1}^{\parallel} = (1 - w_{t+1}) x_t^{\parallel} - \beta_1 w_{t+1} h_t, \\ h_{t+1} &= \frac{m_{t+1}^{\parallel}}{c\lambda} + x_{t+1}^{\parallel} = (2 - w_{t+1}) x_t^{\parallel} + \beta_1 (1 - w_{t+1}) h_t, \end{aligned} \tag{47}$$

which is (45). The matrix $\mathcal{A}(w)$ is similar to the frozen block (21) with $\sigma = w/(c\eta)$, so its spectral radius is below one exactly when $0 < w < 2$. The matrices encountered along a trajectory, however, vary with w_t and need not commute, so stability of every individual matrix does not control their product.

Roadmap. Appendix F.1 proves the subcritical item of Theorem 3.3: the restriction that the second-moment recursion places on successive values of w_t (Lemma F.1) makes an adaptive Lyapunov functional contract below the cutoff $\overline{W}$ (Proposition F.2), which forces finite passage (Corollary F.4); Appendix F.1.1 shows that the parameter condition $\beta_2 > \beta_1^2$ cannot be dropped. Appendix F.2 proves the two supercritical items through the sign geometry of $(x_t^{\parallel}, h_t)$ (Lemma F.6, Corollary F.7), proves the momentum-reversal Lemma 4.1 of the body (Appendix F.2.1), and exhibits the persistently misaligned trajectories (Appendix F.2.2).

F.1 Proof of the subcritical regime of Theorem 3.3

In one dimension the subcritical argument opens with an exact identity relating w_{t+1} to w_t through the scalar second moment. Here F is a sum over coordinates and one side of it survives, as the one-step restriction of Lemma D.3(iii).

Lemma F.1 (Successive-sharpness restriction)**.** *For every t,*

$$w_{t+1} \;\le\; \frac{w_t}{\sqrt{\beta_2} + (1-\sqrt{\beta_2})\, w_t/w_{\max}}, \tag{48}$$

equivalently

$$\frac{1}{w_{t+1}} - \frac{1}{w_{\max}} \;\ge\; \sqrt{\beta_2}\left(\frac{1}{w_t} - \frac{1}{w_{\max}}\right). \tag{49}$$

Proof. By (25), $v_{t+1}^{\parallel} \ge \beta_2 v_t^{\parallel}$, so $w_{t+1} = F(v_{t+1}^{\parallel}) \le F(\beta_2 v_t^{\parallel})$ by Lemma D.3(i), and (29) at $v^{\parallel} = v_t^{\parallel}$ gives (49); (48) is the same inequality solved for w_{t+1}. □

The adaptive Lyapunov functional. The family $\{\mathcal{A}(w) : 0 < w < 2\}$ admits no common Euclidean contraction estimate. We instead use a quadratic form whose weight on h_t^2 is chosen as a function of the current w_t:

$$\begin{gathered}\theta(w) := \frac{w}{2-w} \quad (0<w<2), \qquad P(w) := \operatorname{diag}\big(1, \beta_1\theta(w)\big),\\ \Psi_t := z_t^{\top} P(w_t) z_t = (x_t^{\parallel})^2 + \beta_1\theta(w_t) h_t^2.\end{gathered} \tag{50}$$

Recall the cutoff $\overline{W}$ of Theorem 3.3, defined in (10). The contraction rate on a band $0 < a \le w \le b < \overline{W}$ is expressed through

$$K(b) := \frac{2 - b\big(\sqrt{\beta_2} + 2(1-\sqrt{\beta_2})/w_{\max}\big)}{\sqrt{\beta_2}\,(2-b)}, \tag{51}$$

$$\Delta(a,b) := \frac{(1-\beta_1)\min\{a(2-a),\, b(2-b)\} \cdot \dfrac{\beta_1 a}{2-a}\big(1-\beta_1 K(b)\big)}{2\left(1+\dfrac{\beta_1 b}{2-b}\right)\max\left\{1, \dfrac{\beta_1 b}{2-b}\right\}}. \tag{52}$$

The proof below shows that $K(b)$ bounds the ratio $\theta(w_{t+1})/\theta(w_t)$ on the band, that $\beta_1 K(b) < 1$ is exactly $b < \overline{W}$, and that $\Delta(a,b)$ is the resulting lower bound on $1 - \Psi_{t+1}/\Psi_t$.

Proposition F.2 (Uniform contraction below an explicit cutoff)**.** *Assume* $0 < \beta_1 < 1$, $w_{\max} > 2$ *and* $\beta_2 > \beta_1^2$. *Then* $0 < \overline{W} < 2$. *Fix* $0 < a \le b < \overline{W}$ *and suppose* $a \le w_t, w_{t+1} \le b$. *Then* $0 < \Delta(a,b) < 1$ *and*

$$\Psi_{t+1} \;\le\; \big(1 - \Delta(a,b)\big)\Psi_t. \tag{53}$$

In particular, if $\Psi_t > 0$ *and* $w_t, w_{t+1} < \overline{W}$, *then* $\Psi_{t+1} < \Psi_t$.

Proof. We first record the restriction that the second-moment recursion places on the pair (w_t, w_{t+1}). Rearranging Lemma F.1,

$$w_t \;\geq\; \frac{\sqrt{\beta_2}\, w_{t+1}}{1-(1-\sqrt{\beta_2})w_{t+1}/w_{\max}}. \tag{54}$$

We next solve the one-step quadratic inequality. For $y, Y > 0$ set $P_y := \mathrm{diag}(1, y)$ and

$$\mathcal{E}(y, Y; w) := P_y - \mathcal{A}(w)^\top P_Y \mathcal{A}(w). \tag{55}$$

Direct multiplication gives the exact identities

$$\mathcal{E}_{11}(y, Y; w) = (2-w)\big(w - (2-w)Y\big), \tag{56}$$

$$\det \mathcal{E}(y, Y; w) = \big(w - (2-w)Y\big)\big((2-w)y - \beta_1^2 w\big). \tag{57}$$

Because $0 < w < 2$, Sylvester's criterion therefore shows that

$$\mathcal{A}(w)^\top P_Y \mathcal{A}(w) \prec P_y \iff Y < \theta(w) \;\text{ and }\; y > \beta_1^2 \theta(w). \tag{58}$$

For the Lyapunov functional (50) the source and target weights are $y = \beta_1\theta(w_t)$ and $Y = \beta_1\theta(w_{t+1})$. The first condition in (58) is automatic because $\beta_1 < 1$, while the second becomes

$$\theta(w_t) > \beta_1 \theta(w_{t+1}). \tag{59}$$

Since θ is strictly increasing on $(0, 2)$, (54) implies

$$\theta(w_t) \;\geq\; \theta\!\left(\frac{\sqrt{\beta_2}\, w_{t+1}}{1-(1-\sqrt{\beta_2})w_{t+1}/w_{\max}}\right),$$

and therefore, after direct simplification,

$$\frac{\theta(w_{t+1})}{\theta(w_t)} \;\leq\; \frac{2 - w_{t+1}\big(\sqrt{\beta_2} + 2(1-\sqrt{\beta_2})/w_{\max}\big)}{\sqrt{\beta_2}\,(2-w_{t+1})}. \tag{60}$$

Write $A := \sqrt{\beta_2} + 2(1-\sqrt{\beta_2})/w_{\max}$, so that the right-hand side of (60) is $(2 - Aw_{t+1})/\big(\sqrt{\beta_2}(2 - w_{t+1})\big)$, whose derivative in w_{t+1} has the sign of $2 - 2A$; and $A < 1$ exactly when $w_{\max} > 2$. The right-hand side is thus increasing in w_{t+1}, so if $w_t, w_{t+1} \leq b < 2$ then

$$\frac{\theta(w_{t+1})}{\theta(w_t)} \;\leq\; K(b). \tag{61}$$

The product $\beta_1 K(b)$ is strictly below one exactly when $b\big(\sqrt{\beta_2} - \beta_1 A\big) < 2(\sqrt{\beta_2} - \beta_1)$, that is exactly when $b < \overline{W}$. Since $\sqrt{\beta_2} > \beta_1$ and $w_{\max} > 2$,

$$\sqrt{\beta_2}(1-\beta_1) - \frac{2\beta_1(1-\sqrt{\beta_2})}{w_{\max}} - (\sqrt{\beta_2} - \beta_1) = \beta_1(1-\sqrt{\beta_2})\left(1 - \frac{2}{w_{\max}}\right) > 0, \tag{62}$$

so the denominator of $\overline{W}$ in (10) is positive and strictly larger than $\sqrt{\beta_2} - \beta_1$, whence $0 < \overline{W} < 2$.

It remains to make the contraction factor explicit. Set

$$\mathcal{E} := P(w_t) - \mathcal{A}(w_{t+1})^\top P(w_{t+1}) \mathcal{A}(w_{t+1}). \tag{63}$$

Specializing (56)–(57), or equivalently taking a Schur complement, gives

$$\mathcal{E}_{11} = (1-\beta_1) w_{t+1}(2 - w_{t+1}), \qquad \mathcal{E}_{22} - \frac{\mathcal{E}_{12}^2}{\mathcal{E}_{11}} = \beta_1\big(\theta(w_t) - \beta_1\theta(w_{t+1})\big). \tag{64}$$

If $a \leq w_t, w_{t+1} \leq b < \overline{W}$, then

$$\mathcal{E}_{11} \geq (1-\beta_1)\min\{a(2-a), b(2-b)\}, \qquad \mathcal{E}_{22} - \frac{\mathcal{E}_{12}^2}{\mathcal{E}_{11}} \;\geq\; \frac{\beta_1 a}{2-a}\big(1 - \beta_1 K(b)\big) \;>\; 0, \tag{65}$$

the second by (61) and $\theta(w_t) \geq \theta(a)$, and the positivity because $\beta_1 K(b) < 1$. Hence $\mathcal{E} \succ 0$, and all factors of $\Delta(a, b)$ in (52) are strictly positive. Moreover

$$P(w_t) - \mathcal{E} = \mathcal{A}(w_{t+1})^\top P(w_{t+1}) \mathcal{A}(w_{t+1}) \succ 0,$$

because $P(w_{t+1}) \succ 0$ and $\det \mathcal{A}(w_{t+1}) = \beta_1 > 0$. Thus $0 \prec \mathcal{E} \prec P(w_t)$, and since $w_t \le b$ and θ is increasing,

$$\operatorname{tr} \mathcal{E} < \operatorname{tr} P(w_t) \le 1 + \frac{\beta_1 b}{2-b}. \tag{66}$$

Also, by (65),

$$\det \mathcal{E} = \mathcal{E}_{11}\left(\mathcal{E}_{22} - \frac{\mathcal{E}_{12}^2}{\mathcal{E}_{11}}\right) \;\ge\; (1-\beta_1)\min\{a(2-a), b(2-b)\}\,\frac{\beta_1 a}{2-a}\big(1-\beta_1 K(b)\big). \tag{67}$$

A positive definite 2×2 matrix satisfies $\lambda_{\min} \ge \det / \operatorname{tr}$, so (52), (66) and (67) give

$$\lambda_{\min}(\mathcal{E}) \;\ge\; 2\Delta(a,b)\max\left\{1, \frac{\beta_1 b}{2-b}\right\}.$$

On the other hand $P(w_t) \preceq \max\{1, \beta_1 b/(2-b)\} I$, and therefore

$$\mathcal{E} \;\succeq\; 2\Delta(a,b)\max\left\{1, \frac{\beta_1 b}{2-b}\right\} I \;\succeq\; 2\Delta(a,b)\,P(w_t) \;\succeq\; \Delta(a,b)\,P(w_t). \tag{68}$$

Since $\mathcal{E} \prec P(w_t)$, the first inequality in (68) forces $2\Delta(a,b) < 1$, so in particular $0 < \Delta(a,b) < 1$. Finally,

$$\Psi_{t+1} = z_t^\top \mathcal{A}(w_{t+1})^\top P(w_{t+1}) \mathcal{A}(w_{t+1}) z_t = \Psi_t - z_t^\top \mathcal{E} z_t \le \big(1 - \Delta(a,b)\big)\Psi_t,$$

which is (53). The last assertion follows by taking $a := \min\{w_t, w_{t+1}\}$ and $b := \max\{w_t, w_{t+1}\}$. □

Remark F.3 (Expansion of the cutoff). Substituting $\sqrt{\beta_2} = 1 - \frac{1}{2}(1-\beta_2) + O\big((1-\beta_2)^2\big)$ into (10) gives, for fixed β_1 and $w_{\max} > 2$,

$$\overline{W} = 2 - \frac{\beta_1\big(1 - 2/w_{\max}\big)}{1-\beta_1}(1-\beta_2) + O\big((1-\beta_2)^2\big) \qquad (\beta_2 \to 1). \tag{69}$$

For $(\beta_1, \beta_2) = (0.9, 0.999)$ and large $w_{\max}$ this gives $\overline{W} \approx 1.991$: a restoring tendency close to, but not all the way up to, 2.

The contraction forces finite passage, which is the subcritical item of Theorem 3.3. Note that our argument can establish non-asymptotic upper bound on the time τ, though we omit the precise upper bound for succinctness.

Corollary F.4 (Finite passage toward the subcritical cutoff)**.** *Under the hypotheses of Proposition F.2, fix $b < \overline{W}$ and $T \ge 0$ with $w_T \le b$. Then $w_{T+\tau} > b$ for some finite $\tau \ge 0$.*

Proof. Suppose, for contradiction, that $w_{T+n} \le b$ for every $n \ge 0$. Put

$$M := \max\{v_T^{\parallel}, \Psi_T\}, \qquad a := F(M), \qquad \varsigma := 1 - \Delta(a,b) \in (0,1), \tag{70}$$

so that $0 < a \le w_T \le b$, because F is decreasing and $v_T^{\parallel} \le M$.

Step 1: induction. We claim that for every $n \ge 0$

$$v_{T+n}^{\parallel} \le M \qquad \text{and} \qquad \Psi_{T+n} \le \varsigma^n \Psi_T. \tag{71}$$

Both hold at $n = 0$. If they hold at n, then $(x_{T+n}^{\parallel})^2 \le \Psi_{T+n} \le M$ by (50), so (25) gives $v_{T+n+1}^{\parallel} \le \beta_2 M + (1-\beta_2)M = M$; hence $w_{T+n}, w_{T+n+1} \in [a,b]$, and Proposition F.2 gives $\Psi_{T+n+1} \le \varsigma \Psi_{T+n} \le \varsigma^{n+1}\Psi_T$.

Step 2: the second moment vanishes. Inserting (71) into (27), with G_n as in (32),

$$v_{T+n}^{\parallel} \;\le\; \beta_2^n v_T^{\parallel} + (1-\beta_2)\,\Psi_T\,G_n(\varsigma, \beta_2) \;\xrightarrow[n\to\infty]{}\; 0, \tag{72}$$

since $\varsigma, \beta_2 < 1$. Hence $w_{T+n} = F(v_{T+n}^{\parallel}) \to F(0) = w_{\max} > 2 > b$, contradicting $w_{T+n} \le b$.

□

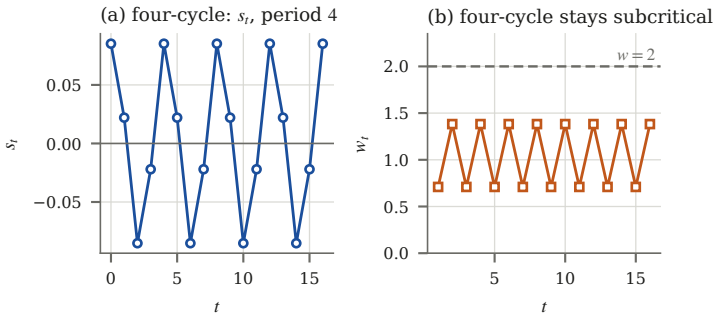

Figure 37: The four-cycle of Proposition F.5 at $\beta_1 = 0.5$, $\beta_2 = 0.2$, $\varepsilon = 0$, $\eta = 0.1$: the constructed $\zeta = 0.25868$ gives an orbit of exact period four (a) whose two sharpness values, 0.7095 and 1.3825, both stay strictly below 2 (b).

F.1.1 Existence of strictly subcritical four-cycles

The parameter condition $\beta_2 > \beta_1^2$ of Proposition F.2 is substantive. This subsection is the one place where the standing assumption $\varepsilon > 0$ is relaxed: we work in the scale-free case $\varepsilon = 0$, where w_t is still defined by (7) but $w_{\max} = +\infty$, and assume $u_i \neq 0$ for every i, so that D_t is invertible and every term of (26) is finite for $v^{\parallel} > 0$; there

$$w_t = F(v_t^{\parallel}) = \frac{c\,\eta_u}{\sqrt{v_t^{\parallel}}}, \qquad \eta_u := \eta\,\|u\|_1\,, \tag{73}$$

so the reduced recursion is exactly one-dimensional Adam with learning rate η_u and $\varepsilon = 0$, and the one-dimensional four-cycle mechanism embeds. We give the construction explicitly.

Proposition F.5 (Strictly subcritical four-cycles)**.** *Assume $\varepsilon = 0$ and*

$$\sqrt{2} - 1 \le \beta_1 < 1, \qquad 0 \le \beta_2 \le \beta_1^2. \tag{74}$$

Then for every $\eta > 0$ there is an initial state satisfying Assumption D.2 whose reduced state $(x_t^{\parallel}, m_t^{\parallel}, v_t^{\parallel})$ has prime period four and satisfies $w_t < 2$ for all t.

Proof. By (73) it suffices to exhibit the one-dimensional four-cycle, which we do explicitly. For $\zeta \in (0, 1)$ define, only within this proof,

$$R(\zeta) := \left[\frac{(1-\beta_1\zeta)(1+\zeta)}{(1-\zeta)(\beta_1+\zeta)}\right]^2, \qquad J(\zeta) := \frac{1 - R(\zeta)\zeta^2}{R(\zeta) - \zeta^2}, \tag{75}$$

and let ζ_0 be the first positive solution of

$$\zeta\,\frac{(1-\beta_1\zeta)(1+\zeta)}{(1-\zeta)(\beta_1+\zeta)} = 1. \tag{76}$$

The unsquared ratio in R exceeds one on $(0, 1)$, because its numerator minus its denominator is

$$(1-\beta_1\zeta)(1+\zeta) - (1-\zeta)(\beta_1+\zeta) = (1-\beta_1)(1+\zeta^2) > 0, \tag{77}$$

and the left-hand side of (76) vanishes at $\zeta = 0$ and tends to $+\infty$ as $\zeta \to 1^-$, so $\zeta_0 \in (0, 1)$ exists. On $(0, \zeta_0]$ we have $R(\zeta) > 1 > \zeta^2$, so the denominator of J is positive and J is continuous there. Equation (76) says $R(\zeta_0)\zeta_0^2 = 1$, hence

$$J(\zeta_0) = 0, \qquad \text{while} \qquad R(0) = \beta_1^{-2} \;\text{ gives }\; J(0) = \beta_1^2.$$

By the intermediate value theorem there is therefore $\zeta \in (0, \zeta_0]$ with $J(\zeta) = \beta_2$ for every β_2 in (74); in the boundary case $\beta_2 = \beta_1^2$ the identity $J'(0) = 2\beta_1(1-\beta_1)^2 > 0$ supplies a second crossing at some $\zeta > 0$.

Fix such a ζ and set, again only within this proof,

$$\nu_+ := \sqrt{\frac{1+\beta_2\zeta^2}{1+\beta_2}}, \quad \nu_- := \sqrt{\frac{\zeta^2+\beta_2}{1+\beta_2}}, \quad \varphi_+ := \frac{1-\beta_1\zeta}{1-\zeta}, \quad \varphi_- := \frac{\beta_1+\zeta}{1+\zeta}, \quad \kappa := \frac{1-\beta_1}{1+\beta_1^2}. \tag{78}$$

The equation $J(\zeta) = \beta_2$ is exactly $\varphi_+/\nu_+ = \varphi_-/\nu_-$, so

$$\varrho := \eta_u \kappa \frac{\varphi_+}{\nu_+} = \eta_u \kappa \frac{\varphi_-}{\nu_-} > 0 \tag{79}$$

is well defined. Initialize

$$x_0^{\parallel} = \varrho, \qquad m_0^{\parallel} = -\lambda\kappa\varrho(\beta_1 + \zeta), \qquad v_0^{\parallel} = \varrho^2\nu_-^2.$$

The identities

$$\beta_2\nu_-^2 + (1-\beta_2) = \nu_+^2, \qquad \beta_2\nu_+^2 + (1-\beta_2)\zeta^2 = \nu_-^2 \tag{80}$$

verify the alternating second moments; the momentum recursion produces the displayed $m_t^{\parallel}$ below; and $\varrho\nu_+ = \eta_u\kappa\varphi_+$, $\varrho\nu_- = \eta_u\kappa\varphi_-$ verify the position updates $\varrho \mapsto \zeta\varrho$ and $\zeta\varrho \mapsto -\varrho$. The remaining two steps follow by half-turn symmetry. Hence

$t \bmod 4$	0	1	2	3
$x_t^{\parallel}$	ϱ	$\zeta\varrho$	$-\varrho$	$-\zeta\varrho$
$m_t^{\parallel}/\lambda$	$-\kappa\varrho(\beta_1+\zeta)$	$\kappa\varrho(1-\beta_1\zeta)$	$\kappa\varrho(\beta_1+\zeta)$	$-\kappa\varrho(1-\beta_1\zeta)$
$v_t^{\parallel}$	$\varrho^2\nu_-^2$	$\varrho^2\nu_+^2$	$\varrho^2\nu_-^2$	$\varrho^2\nu_+^2$

(81)

This also shows why the construction works for every $\eta > 0$: changing η rescales ϱ, $m_t^{\parallel}$ and $\sqrt{v_t^{\parallel}}$ by the same factor.

The two normalized sharpness values along the orbit are

$$\frac{1+\beta_1^2}{1+\beta_1}\cdot\frac{1-\zeta}{1-\beta_1\zeta} \qquad \text{and} \qquad \frac{1+\beta_1^2}{1+\beta_1}\cdot\frac{1+\zeta}{\beta_1+\zeta}. \tag{82}$$

The first is always below 2; the second is below 2 exactly when

$$\zeta > \frac{1-2\beta_1-\beta_1^2}{1+2\beta_1-\beta_1^2}, \tag{83}$$

whose right-hand side is nonpositive precisely when $\beta_1 \geq \sqrt{2}-1$. Hence the orbit is strictly subcritical, and its four values of $x_t^{\parallel}$ are distinct, so the period is prime four. □

F.2 Proof of the supercritical regime of Theorem 3.3

In the supercritical region $w > 2$ the sign of $x_t^{\parallel} h_t$ separates two sharply different behaviors, recorded in the following lemma; the two supercritical items of Theorem 3.3 are its two parts iterated in time.

Lemma F.6 (Supercritical phase geometry)**.** *Suppose* $w_{t+1} > 2$.

1. *If* $x_t^{\parallel} h_t > 0$ *then* $x_t^{\parallel} x_{t+1}^{\parallel} < 0$, $h_t h_{t+1} < 0$, *hence* $x_{t+1}^{\parallel} h_{t+1} > 0$, *and* $\left|x_{t+1}^{\parallel}\right| > (w_{t+1}-1)\left|x_t^{\parallel}\right|$.
2. *If* $x_t^{\parallel} h_t < 0$ *and* $x_{t+1}^{\parallel} h_{t+1} < 0$ *then* $\left|x_{t+1}^{\parallel}\right| < \left|x_t^{\parallel}\right|/(w_{t+1}-1)$.

Proof. Write $w := w_{t+1} > 2$.

(i) Let $\sigma := \operatorname{sign}(x_t^{\parallel}) = \operatorname{sign}(h_t)$, $a := \left|x_t^{\parallel}\right| > 0$ and $b := |h_t| > 0$. By (47),

$$x_{t+1}^{\parallel} = -\sigma\big[(w-1)a + \beta_1 w b\big], \qquad h_{t+1} = -\sigma\big[(w-2)a + \beta_1(w-1)b\big]. \tag{84}$$

Both brackets are strictly positive because $w > 2$, so both components carry the sign $-\sigma$, and

$$\left|x_{t+1}^{\parallel}\right| = (w-1)\left|x_t^{\parallel}\right| + \beta_1 w\,|h_t| > (w-1)\left|x_t^{\parallel}\right|. \tag{85}$$

(ii) If instead $h_t = -\xi x_t^{\parallel}$ with $\xi > 0$, then (47) gives

$$x_{t+1}^{\parallel} = -x_t^{\parallel}\big[(w-1) - \beta_1 w\xi\big], \qquad h_{t+1} = -x_t^{\parallel}\big[(w-2) - \beta_1(w-1)\xi\big], \tag{86}$$

so $x_{t+1}^{\parallel} h_{t+1} < 0$ holds exactly when the two brackets have opposite signs, that is exactly on the strip

$$\ell(w) < \xi < \upsilon(w), \qquad \ell(w) := \frac{w-2}{\beta_1(w-1)}, \qquad \upsilon(w) := \frac{w-1}{\beta_1 w}, \tag{87}$$

where ℓ and υ are used in this appendix only. The strip is nonempty because $w(w-2) < (w-1)^2$. Within it, $\left|x_{t+1}^{\parallel}\right| = \left((w-1) - \beta_1 w\xi\right)\left|x_t^{\parallel}\right|$, and $\xi > \ell(w)$ gives

$$\left|x_{t+1}^{\parallel}\right| < \left(w - 1 - \frac{w(w-2)}{w-1}\right)\left|x_t^{\parallel}\right| = \frac{\left|x_t^{\parallel}\right|}{w-1}, \tag{88}$$

using $(w-1)^2 - w(w-2) = 1$. □

Under $x_t^{\parallel} h_t > 0$ the first component of (84) is already reversed as soon as $w > 1$; the hypothesis $w > 2$ is what keeps the second bracket positive, and hence keeps the cone $x_t^{\parallel} h_t > 0$ invariant so that the lemma can be iterated. Iterating part (ii) gives the misaligned item of Theorem 3.3 directly: if $w_t > 2$ and $x_t^{\parallel} h_t < 0$ for all $t \geq T$, then $\left|x_{t+1}^{\parallel}\right| < \left|x_t^{\parallel}\right| / (w_{t+1} - 1) < \left|x_t^{\parallel}\right|$ for every $t \geq T$. Iterating part (i) gives the aligned item, since the expansion feeds the second moments exactly as in the zero-momentum case.

Corollary F.7 (Aligned finite exit from every supercritical band)**.** *Suppose $x_T^{\parallel} h_T > 0$ and $w_T \geq 2 + \delta$ with $\delta > 0$. Then, with $\tau_\delta(T)$ as in* (33) *and $N_{\sup}$, $\overline{N}_{\sup}$ as in* (39)*,*

$$\tau_\delta(T) \;\leq\; N_{\sup} \;\leq\; \overline{N}_{\sup} = 1 + \left\lfloor \frac{\log\!\big(1 + (A_\delta - \beta_2) Q_\delta / ((1-\beta_2)(x_T^{\parallel})^2)\big)}{\log A_\delta} \right\rfloor \;<\; \infty. \tag{89}$$

Proof. Note first that $x_T^{\parallel} \neq 0$, since $x_T^{\parallel} h_T > 0$, and that $2 + \delta \leq w_T \leq w_{\max}$. Fix $n \geq 1$ and suppose $\tau_\delta(T) > n$, so that $w_{T+k} \geq 2 + \delta > 2$ for $k = 1, \ldots, n$. By Lemma F.6(i), applied successively at $t = T, \ldots, T + n - 1$, alignment is preserved at every such step and

$$\left|x_{T+k}^{\parallel}\right| \;\geq\; (w_{T+k} - 1)\left|x_{T+k-1}^{\parallel}\right| \;\geq\; (1+\delta)\left|x_{T+k-1}^{\parallel}\right|, \qquad k = 1, \ldots, n.$$

Thus (40) holds, and Lemma E.2 gives $n < N_{\sup}$. Hence $\tau_\delta(T) \leq N_{\sup} \leq \overline{N}_{\sup}$, which is (89).

□

F.2.1 Proof of Lemma 4.1

Proof of Lemma 4.1. By (7), $W_{\mathrm{act},t} = w_{t+1}$, so the hypothesis reads $w_{t+1} > 2$ for every $t \in [T, T+n]$.

Step 1: the gradient. Lemma F.6(i) at $t = T$ gives $x_T^{\parallel} x_{T+1}^{\parallel} < 0$ and $x_{T+1}^{\parallel} h_{T+1} > 0$, so its hypothesis holds again at $t = T + 1$. By induction,

$$x_t^{\parallel} h_t > 0 \qquad \text{and} \qquad x_t^{\parallel} x_{t+1}^{\parallel} < 0 \qquad (t \in [T, T+n]). \tag{90}$$

Step 2: the momentum. By (46), $m_{t+1}^{\parallel} = c\lambda\,(x_t^{\parallel} + \beta_1 h_t)$ with $c\lambda > 0$ and $\beta_1 > 0$. Under (90) the two summands have the sign of $x_t^{\parallel}$, so $m_{t+1}^{\parallel}$ has the sign of $x_t^{\parallel}$ for $t \in [T, T+n]$, and therefore

$$m_t^{\parallel} m_{t+1}^{\parallel} < 0 \qquad (t \in [T+1, T+n]). \tag{91}$$

Step 3: from scalars to cosines. Since $g_t = \lambda x_t^{\parallel} u$ and, under Assumption D.2, $m_t = m_t^{\parallel} u$,

$$\cos(g_t, g_{t+1}) = \operatorname{sign}(x_t^{\parallel} x_{t+1}^{\parallel}) = -1, \qquad \cos(m_t, m_{t+1}) = \operatorname{sign}(m_t^{\parallel} m_{t+1}^{\parallel}) = -1 \tag{92}$$

on the ranges of (90) and (91) respectively.

□

F.2.2 Existence of misaligned supercritical trajectories

The contraction of Lemma F.6(ii) is available only while misalignment survives, and the next proposition shows that it can survive forever: there are trajectories that converge to the minimizer manifold while the frozen map is unstable at every step. Together with the four-cycles of Appendix F.1.1, these are the one-dimensional exceptions, and they survive in rank one.

Proposition F.8 (Supercritical convergence by persistent misalignment)**.** *Assume* $0 < \beta_1 < 1$, $w_{\max} > 2$ *and* $\varepsilon > 0$. *Fix a level* $v_0^{\parallel} > 0$ *with* $w_0 = F(v_0^{\parallel}) > 2$. *Then there exist* $x_0^{\parallel} \neq 0$ *and* $m_0^{\parallel}$, *realized by an initialization satisfying Assumption D.2, for which*

$$x_t^{\parallel} h_t < 0, \qquad w_t \geq w_0, \qquad \left|x_t^{\parallel}\right| \leq (w_0 - 1)^{-t} \left|x_0^{\parallel}\right| \qquad (t \geq 0), \tag{93}$$

and along this trajectory $x_t^{\parallel} \to 0$, $m_t^{\parallel} \to 0$, $v_t^{\parallel} \to 0$ *and* $w_t \to w_{\max} > 2$. *The full vector* x_t *converges to a minimizer in* $u^{\perp}$, *not necessarily to the origin.*

Proof. Choose

$$0 < \left|x_0^{\parallel}\right| \leq \sqrt{v_0^{\parallel}}, \tag{94}$$

and, for a trial slope $\zeta_0 > 0$, set

$$h_0 = -\zeta_0 x_0^{\parallel}, \qquad m_0^{\parallel} = c\lambda(h_0 - x_0^{\parallel}) = -c\lambda(1 + \zeta_0) x_0^{\parallel}. \tag{95}$$

These data are realized by $x_0 = x_0^{\parallel} u$, $m_0 = m_0^{\parallel} u$ and $v_0 = \lambda^2 v_0^{\parallel}\, u \odot u$.

The prescribed box is preserved. Whenever misalignment survives one step, Lemma F.6(ii) gives $\left|x_{t+1}^{\parallel}\right| < \left|x_t^{\parallel}\right| / (w_{t+1} - 1) < \left|x_t^{\parallel}\right|$, because $w_{t+1} > 2$. Starting from (94), induction and the convex combination form of (25) therefore give

$$(x_t^{\parallel})^2 \leq v_0^{\parallel}, \qquad v_t^{\parallel} \leq v_0^{\parallel}, \qquad w_t = F(v_t^{\parallel}) \geq w_0 > 2 \tag{96}$$

at every surviving time. Thus the supercritical hypothesis of Lemma F.6(ii) is preserved for as long as the trajectory stays misaligned.

Selection of the slope. Removing the alternating signs by

$$\tilde{x^{\parallel}}_t := \operatorname{sign}(x_0^{\parallel})(-1)^t x_t^{\parallel}, \qquad \tilde{h}_t := \operatorname{sign}(x_0^{\parallel})(-1)^t h_t$$

turns (45) into a nonnegative map, and on a surviving orbit

$$\zeta_t := -\frac{h_t}{x_t^{\parallel}} = -\frac{\tilde{h}_t}{\tilde{x^{\parallel}}_t} > 0.$$

By (87) the next step is misaligned exactly when $\ell(w_{t+1}) < \zeta_t < \upsilon(w_{t+1})$, and then (86) gives

$$\zeta_{t+1} = \Gamma_{w_{t+1}}(\zeta_t), \qquad \Gamma_w(\zeta) := \frac{\beta_1(w-1)\zeta - (w-2)}{(w-1) - \beta_1 w \zeta}, \tag{97}$$

with Γ used in this proof only. For fixed $w > 2$ the map Γ_w is continuous and strictly increasing on the strip, since

$$\Gamma_w'(\zeta) = \frac{\beta_1}{\big((w-1) - \beta_1 w \zeta\big)^2} > 0$$

after the cancellation $(w-1)^2 - w(w-2) = 1$, and its endpoint limits are

$$\lim_{\zeta \downarrow \ell(w)} \Gamma_w(\zeta) = 0, \qquad \lim_{\zeta \uparrow \upsilon(w)} \Gamma_w(\zeta) = +\infty. \tag{98}$$

We now build nested intervals. The value w_1 depends on $x_0^{\parallel}$ and $v_0^{\parallel}$ but not on ζ_0, so

$$\mathcal{J}_1 := \big(\ell(w_1), \upsilon(w_1)\big)$$

is exactly the set of slopes surviving one step; the map $\zeta_0 \mapsto \zeta_1$ is continuous on it and, by (98), maps it onto $(0, \infty)$. Suppose inductively that $\mathcal{J}_N$ is a nonempty open interval such that every $\zeta_0 \in \mathcal{J}_N$ survives through time N, the map $\zeta_0 \mapsto \zeta_N$ is continuous on it, and its limits at the left and right endpoints are 0 and $+\infty$. The quantities w_{N+1}, $\ell(w_{N+1})$ and $\upsilon(w_{N+1})$ depend continuously on ζ_0, and both boundaries increase with w because $\ell'(w) = \beta_1^{-1}(w-1)^{-2} > 0$ and $\upsilon'(w) = \beta_1^{-1} w^{-2} > 0$; hence, by (96),

$$\ell(w_{N+1}) \geq \ell(w_0) > 0, \qquad \upsilon(w_{N+1}) \leq \upsilon(w_{\max}) < \frac{1}{\beta_1}. \tag{99}$$

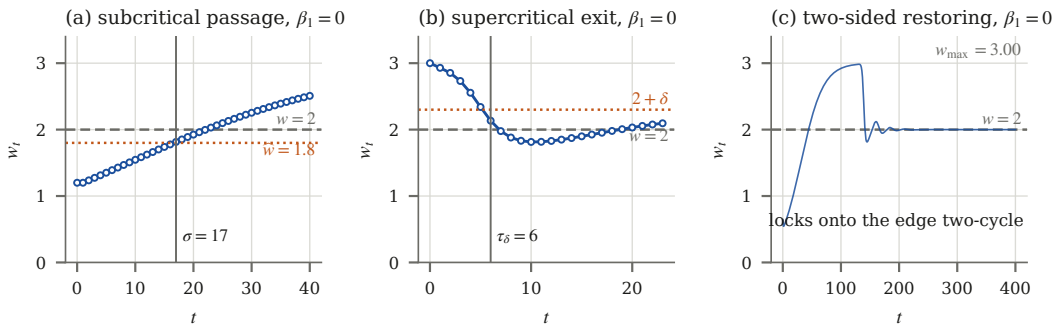

Figure 38: Zero momentum ($\eta = 0.05$, $\beta_2 = 0.9$, $w_{\max} = 3$). (a) From $w_0 = 1.2$ the target $\widehat{w} = 1.8$ is reached at $\sigma = 17$, against the certificate $N_{\text{sub}} = 19$ of Theorem E.1. (b) From $w_0 = w_{\max}$ the band $w \geq 2.3$ is left at $\tau_\delta = 6$, against $N_{\text{sup}} = 10$ of Theorem E.3. (c) A longer run crosses $w = 2$ from both sides and then settles on the exact edge two-cycle of Proposition 3.1.

Thus the graph of ζ_N starts below the lower boundary in (87) and ends above the upper one, so a component $\mathcal{J}_{N+1}$ can be chosen on which $\ell(w_{N+1}) < \zeta_N < \upsilon(w_{N+1})$ and whose left and right endpoints meet the lower and the upper boundary respectively; then $\overline{\mathcal{J}_{N+1}} \subset \mathcal{J}_N$, and (98) restores the endpoint limits 0 and $+\infty$ for ζ_{N+1}. This completes the induction. The nonempty compact sets $\overline{\mathcal{J}_N}$ are nested, so we may choose

$$\zeta_0^\star \in \bigcap_{N \geq 1} \overline{\mathcal{J}_N} \subset \bigcap_{N \geq 1} \mathcal{J}_N,$$

and with this choice misalignment persists at every finite time.

Convergence estimates. Iterating Lemma F.6(ii) and using $w_t \geq w_0$ gives the geometric bound in (93). Since $\zeta_t < \upsilon(w_{t+1}) < 1/\beta_1$ and $m_t^\parallel = -c\lambda(1 + \zeta_t)x_t^\parallel$, also $m_t^\parallel \to 0$ geometrically; (25) then gives $v_t^\parallel \to 0$, and (7) gives $w_t \to w_{\max}$. Finally $m_t = m_t^\parallel u$ and $\left\|D_t^{-1} u\right\|_2 \leq 1/\varepsilon$, so $\sum_t \|x_{t+1} - x_t\|_2 < \infty$ and x_t converges; its limit satisfies $u^\top x_\infty = \lim_t x_t^\parallel = 0$ and is therefore a minimizer. □

G Numerical illustration of Section 3.1

This appendix illustrates the results of Section 3.1 by simulating the rank-one model directly. All runs use $d = 3$, $u = (\frac{2}{3}, \frac{2}{3}, \frac{1}{3})$ and $\lambda = 1.5$, with $\varepsilon = 0.025$ except in the scale-free four-cycle run ($\varepsilon = 0$). Figure 38 shows the $\beta_1 = 0$ statements of Appendix E. In panel (c) the trajectory does not merely oscillate: it locks onto the exact orbit of Proposition 3.1, on which $w_t = 2$ identically. Figure 39 shows the positive-momentum results of Appendices F and F.2, with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $w_{\max} = 2.84$, for which $\overline{W} = 1.9722$. Figure 37 (Appendix F.1.1) shows the four-cycle and Figure 40 the misaligned orbit.

H Proof of Theorem 3.4

Setting. Let $H = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \succ 0$ and $\varepsilon > 0$. Since H and D_t are diagonal, (1) decouples: coordinate i is the rank-one dynamics of Appendix D with $u = e_i$ and $\lambda = \lambda_i$, that is

$$\begin{aligned} m_{i,t+1} &= \beta_1 m_{i,t} + (1 - \beta_1)\lambda_i x_{i,t}, \qquad v_{i,t+1} = \beta_2 v_{i,t} + (1 - \beta_2)\lambda_i^2 x_{i,t}^2, \\ x_{i,t+1} &= x_{i,t} - \frac{w_{i,t+1}}{c\lambda_i}\, m_{i,t+1}, \end{aligned} \tag{100}$$

with $w_{i,t} = c\eta\lambda_i/(\sqrt{v_{i,t}} + \varepsilon)$ as in (11) and $w_{i,\max} = c\eta\lambda_i/\varepsilon$. Every statement of Appendices E–F therefore applies to each coordinate separately, with $x_{i,t}$, $m_{i,t}$, $v_{i,t}/\lambda_i^2$, $w_{i,t}$, $h_{i,t} := (c\lambda_i)^{-1} m_{i,t} + x_{i,t}$ and $w_{i,\max}$ in place of $x_t^\parallel$, $m_t^\parallel$, $v_t^\parallel$, w_t, h_t and $w_{\max}$; the only coupling between coordinates is

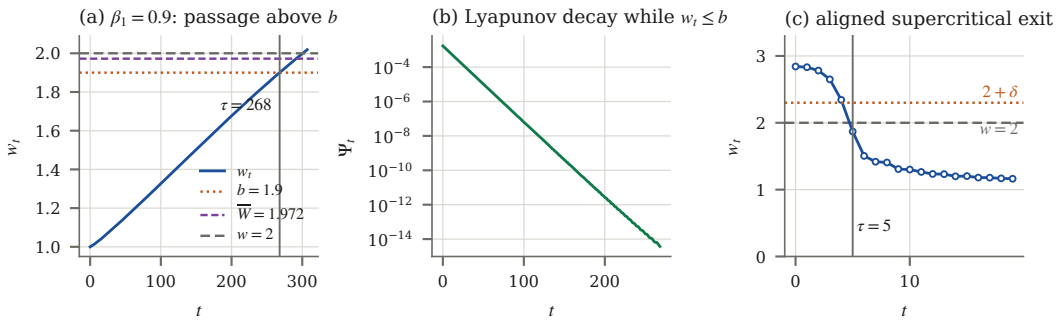

Figure 39: Positive momentum ($\eta = 0.9$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, $w_{\max} = 2.84$, $\overline{W} = 1.9722$). (a) Starting at $w_0 = 1$, the normalized sharpness leaves the band $w \le b = 1.9$ at $t = 268$, as forced by Corollary F.4. (b) The Lyapunov functional Ψ_t falls over that stretch, an observed per-step factor 0.905; the certified factor is $1 - \Delta(a, b)$ with $a = w_0 = 1$, $b = 1.9$ and $\Delta = 1.93 \cdot 10^{-6}$. (c) An aligned supercritical start leaves $w \ge 2.3$ at $\tau = 5$, against the closed-form bound 16 of Corollary F.7.

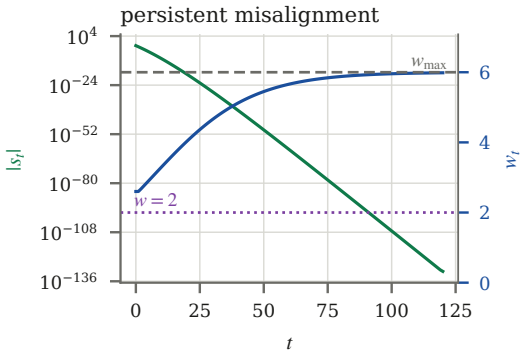

Figure 40: The misaligned orbit of Proposition F.8 at $\beta_1 = 0.5$, $\beta_2 = 0.9$, $\eta = 0.3$, $w_{\max} = 6$: over 120 steps $\left|x_t^{\parallel}\right|$ falls below 10^{-130} while w_t rises from 2.6 to 5.98, so the iterate converges to the minimizer manifold without ever leaving the supercritical region.

the hypothesis on $W_{\mathrm{act},t}$. For $x_t \neq 0$ put

$$a_{i,t} := \lambda_i x_{i,t}^2 \, w_{i,t+1} \ \ge 0, \qquad \sum_{i=1}^{n} a_{i,t} > 0, \tag{101}$$

so that (11) reads $W_{\mathrm{act},t} = \sum_i a_{i,t} w_{i,t+1} / \sum_i a_{i,t}$ and, for every level ω,

$$\sum_{i=1}^{n} a_{i,t} \big(w_{i,t+1} - \omega\big) = \Big(\sum_{i=1}^{n} a_{i,t}\Big)\big(W_{\mathrm{act},t} - \omega\big). \tag{102}$$

Throughout, $f(x) = \frac{1}{2} \sum_i \lambda_i x_i^2$.

Since $v_{i,t}$ is a convex combination of $v_{i,T}$ and of $\lambda_i^2 x_{i,s}^2$, $T \le s < t$,

$$v_{i,t} \ \le \ \max\Big\{ v_{i,T}, \ \lambda_i^2 \max_{T \le s < t} x_{i,s}^2 \Big\} \qquad (t > T). \tag{103}$$

In particular, along a bounded trajectory $v_{i,t} \le \overline{V}_i$ for some constants $\overline{V}_i$, and every coordinate curvature is bounded below: $w_{i,t} \ge \alpha_i := c\eta\lambda_i / (\sqrt{\overline{V}_i} + \varepsilon) > 0$ for all t.

The speed bound. A coordinate curvature cannot increase arbitrarily fast. Applying Lemma F.1 to coordinate i (equivalently, to the rank-one model with $u = e_i$), we obtain

$$w_{i,t+1} \le \frac{w_{i,t}}{\sqrt{\beta_2} + (1-\sqrt{\beta_2})w_{i,t}/w_{i,\max}} \le \frac{w_{i,t}}{\sqrt{\beta_2}}, \qquad \text{and hence} \quad w_{i,t-j} \ge \beta_2^{j/2} w_{i,t}, \quad 0 \le j \le t. \tag{104}$$

Thus, if a coordinate reaches level Q at time t, then j steps earlier it must have been at least $\beta_2^{j/2}Q$. In other words, upward motion is geometrically limited, whereas downward motion need not be: a large $x_{i,t}$ relative to $v_{i,t}$ can cause $w_{i,t}$ to drop sharply.

H.1 Proof of the supercritical regime

H.1.1 Main argument

This subsection proves the supercritical item of Theorem 3.4. The idea is simple. We attach to each coordinate a number $\Phi_{i,t}$, which we call its *potential*. The potential has two properties. It stays bounded. And at every step it grows by at least $x_{i,t}^2\, w_{i,t+1}(w_{i,t+1}-2)/(c\eta)^2$.

This lower bound is positive when $w_{i,t+1} > 2$ and negative when $w_{i,t+1} < 2$. Add the bounds over the coordinates, with the weights λ_i. By (102), the sum is a positive multiple of $W_{\mathrm{act},t} - 2$. So if $W_{\mathrm{act},t} \ge 2 + \delta$ at every step, the total potential grows at every step by at least a fixed multiple of $\|x_t\|^2$, implying $\sum_t \|x_t\|^2 < \infty$.

The rest of the subsection establishes this argument. Proposition H.2 states it with three hypotheses: a floor on the curvatures, a bound on the total potential, and the growth bound. Lemma H.3 gives the floor and shows that trajectories stay bounded. Lemma H.4 is an exact identity that shows where the threshold 2 comes from. Lemma H.5 constructs a potential at the default parameters. All three are proved in Appendix H.1.2. The first two hold whenever $\beta_1^2 < \beta_2$. Only the third uses specific parameters.

Setting and notation. The supercritical item assumes $0 < \beta_1 < 1$, $\beta_1^2 < \beta_2 < 1$, and that the scalar dynamics admits a potential in the sense of Definition H.1 below. Lemma H.5 proves this at $(\beta_1, \beta_2) = (0.9, 0.999)$. Put $\gamma := \beta_1/(1+\beta_1)$.

It is convenient to measure each coordinate in units of the step size. Put $y_{i,t} := x_{i,t}/(c\eta)$, $\mu_{i,t} := m_{i,t}/(c\eta\lambda_i)$, $V_{i,t} := v_{i,t}/(c\eta\lambda_i)^2$ and $e_i := \varepsilon/(c\eta\lambda_i) = 1/w_{i,\max}$. Since $w_{i,\max} > 2$, we have $0 < e_i < \frac{1}{2}$. In these units (100) reads

$$\mu_{i,t+1} = \beta_1\mu_{i,t} + (1-\beta_1)y_{i,t}, \qquad V_{i,t+1} = \beta_2 V_{i,t} + (1-\beta_2)y_{i,t}^2,$$
$$y_{i,t+1} = y_{i,t} - \frac{1+\beta_1}{1-\beta_1}\, w_{i,t+1}\mu_{i,t+1}, \tag{105}$$

with $w_{i,t+1} = (\sqrt{V_{i,t+1}} + e_i)^{-1}$. When we look at one coordinate only, we drop the index i.

Definition H.1 (Potential)**.** Fix β_1, β_2. A *potential* for the scalar dynamics (105) is, for every $e \in (0, \frac{1}{2})$, a real function $\Phi_t = \phi_e(y_{t-1}, y_t, V_t)$ of the state with the following two properties along every trajectory. There is a constant C with $\Phi_t \le C\, y_t^2$ for all $t \ge 1$. And for all sufficiently large t, $\Phi_{t+1} - \Phi_t \ge y_t^2\, w_{t+1}(w_{t+1}-2)$.

Proposition H.2 (Abstract supercritical criterion)**.** *Suppose that each coordinate i carries a number $\Phi_{i,t}$ at each time t, its potential, and put $\Phi_t^{\mathrm{tot}} := \sum_i \lambda_i \Phi_{i,t}$. Suppose that there are constants $\alpha > 0$ and $\overline{\Phi} < \infty$ such that, for all sufficiently large t and every coordinate i:*

(S1) Curvature floor. $w_{i,t} \ge \alpha$.

(S2) Bounded potential. $\Phi_t^{\mathrm{tot}} \le \overline{\Phi}$.

(S3) Growth bound. $\Phi_{i,t+1} - \Phi_{i,t} \ge y_{i,t}^2\, w_{i,t+1}(w_{i,t+1}-2)$.

If $W_{\mathrm{act},t} \ge 2 + \delta$ for some $\delta > 0$ and all sufficiently large t, then $\sum_t \|x_t\|^2 < \infty$. In particular $x_t \to 0$, $v_{i,t} \to 0$ and $w_{i,t} \to w_{i,\max}$.

Proof. Choose T so large that (S1)–(S3) hold and $W_{\mathrm{act},t} \geq 2+\delta$ for every $t \geq T$. Fix $t \geq T$. By $y_{i,t} = x_{i,t}/(c\eta)$ and (101),

$$\lambda_i y_{i,t}^2 w_{i,t+1} = \frac{a_{i,t}}{(c\eta)^2} \qquad \text{and, by (S1),} \qquad a_{i,t} \;\geq\; (c\eta)^2 \alpha\, \lambda_i y_{i,t}^2. \tag{106}$$

Multiply (S3) by λ_i, sum over i, and use (102) with $\omega = 2$:

$$\begin{aligned}
\Phi_{t+1}^{\mathrm{tot}} - \Phi_t^{\mathrm{tot}} &= \sum_i \lambda_i \big(\Phi_{i,t+1} - \Phi_{i,t}\big) \;\geq\; \sum_i \lambda_i y_{i,t}^2 w_{i,t+1}\big(w_{i,t+1} - 2\big) && \text{(S3)} \\
&= \frac{1}{(c\eta)^2} \sum_i a_{i,t}\big(w_{i,t+1} - 2\big) && (106) \\
&= \frac{1}{(c\eta)^2} \Big(\sum_i a_{i,t}\Big)\big(W_{\mathrm{act},t} - 2\big) && (102) \\
&\geq \frac{\delta}{(c\eta)^2} \sum_i a_{i,t} \;\geq\; \delta\alpha \sum_i \lambda_i y_{i,t}^2 && W_{\mathrm{act},t} \geq 2+\delta,\ (106).
\end{aligned} \tag{107}$$

Sum (107) over $t = T, \ldots, n$; the left-hand side telescopes and (S2) bounds it:

$$\delta\alpha \sum_{t=T}^{n} \sum_i \lambda_i y_{i,t}^2 \;\leq\; \sum_{t=T}^{n} \big(\Phi_{t+1}^{\mathrm{tot}} - \Phi_t^{\mathrm{tot}}\big) \;=\; \Phi_{n+1}^{\mathrm{tot}} - \Phi_T^{\mathrm{tot}} \;\leq\; \overline{\Phi} - \Phi_T^{\mathrm{tot}} \qquad (n \geq T). \tag{108}$$

The right-hand side does not depend on n, so $\sum_t \sum_i \lambda_i y_{i,t}^2 < \infty$, hence

$$\sum_t \|x_t\|^2 \;\leq\; \frac{(c\eta)^2}{\min_i \lambda_i} \sum_t \sum_i \lambda_i y_{i,t}^2 \;<\; \infty \qquad \text{and} \qquad x_t \to 0.$$

Then $v_{i,t+1} = \beta_2 v_{i,t} + (1-\beta_2)\lambda_i^2 x_{i,t}^2 \to 0$ by (100), so $w_{i,t} = c\eta\lambda_i/(\sqrt{v_{i,t}} + \varepsilon) \to w_{i,\max}$. □

H.1.2 Lemmas for Proposition H.2

Lemma H.3 (Moment bound and curvature floor)**.** *Let $0 < \beta_1 < 1$ and $\beta_1^2 < \beta_2 < 1$. In every coordinate the momentum is controlled by the second moment:*

$$|m_{i,t+1}| \;\leq\; \beta_1^{t+1}\,|m_{i,0}| + K\sqrt{v_{i,t+1}}, \qquad K^2 := \frac{(1-\beta_1)^2}{(1-\beta_2)(1-\beta_1^2/\beta_2)}. \tag{109}$$

As a consequence, every trajectory is bounded:

$$\limsup_{t\to\infty} |x_{i,t}| \;\leq\; \frac{\eta K}{1-\beta_1}, \qquad \limsup_{t\to\infty} \sqrt{v_{i,t}} \;\leq\; \frac{\eta K \lambda_i}{1-\beta_1}. \tag{110}$$

Finally, the curvatures have a common floor that does not depend on the coordinate:

$$w_{i,t} \;\geq\; \alpha := \frac{c}{K/(1-\beta_1) + c/2} \;>\; 0 \qquad \textit{for every } i \textit{ and all sufficiently large } t. \tag{111}$$

Proof. Fix i and drop the index. Unrolling (100),

$$m_{t+1} = \beta_1^{t+1} m_0 + (1-\beta_1)\sum_{j=0}^{t} \beta_1^j \lambda x_{t-j}, \qquad v_{t+1} \;\geq\; (1-\beta_2)\sum_{j=0}^{t} \beta_2^j \lambda^2 x_{t-j}^2.$$

Write $\beta_1^j = (\beta_1/\sqrt{\beta_2})^j \beta_2^{j/2}$ and apply Cauchy–Schwarz:

$$\begin{aligned}
\sum_{j=0}^{t} \beta_1^j\,|\lambda x_{t-j}| \;&\leq\; \Big(\sum_{j\geq 0}\Big(\frac{\beta_1^2}{\beta_2}\Big)^j\Big)^{1/2} \Big(\sum_{j=0}^{t} \beta_2^j \lambda^2 x_{t-j}^2\Big)^{1/2} \\
&\leq\; \frac{1}{\sqrt{1-\beta_1^2/\beta_2}} \cdot \frac{\sqrt{v_{t+1}}}{\sqrt{1-\beta_2}} \;=\; \frac{K}{1-\beta_1}\sqrt{v_{t+1}}.
\end{aligned}$$

With the first display this is (109). Hence, by (100) and $w_{t+1} = c\eta\lambda/(\sqrt{v_{t+1}} + \varepsilon)$,

$$|x_{t+1} - x_t| = \frac{\eta\,|m_{t+1}|}{\sqrt{v_{t+1}} + \varepsilon} \;\le\; \eta K + \rho_t, \qquad \rho_t := \frac{\eta\beta_1^{t+1}\,|m_0|}{\varepsilon} \;\to\; 0. \tag{112}$$

Position. Put $\xi_t := m_t/\lambda - x_t$. By (100),

$$\frac{m_{t+1}}{\lambda} = x_t + \beta_1\xi_t, \qquad \xi_{t+1} = \frac{m_{t+1}}{\lambda} - x_{t+1} = \beta_1\xi_t + (x_t - x_{t+1}). \tag{113}$$

Iterating the second identity from a time T and using (112),

$$|\xi_t| \;\le\; \beta_1^{t-T}\,|\xi_T| + \frac{\eta K + \sup_{s\ge T}\rho_s}{1-\beta_1} \quad (t \ge T), \qquad \text{so} \qquad \limsup_{t\to\infty} |\xi_t| \;\le\; \frac{\eta K}{1-\beta_1}. \tag{114}$$

Fix $\zeta > 0$ and put

$$R := \frac{\eta K}{1-\beta_1}, \qquad b := \frac{\beta_1\eta K}{1-\beta_1} + \frac{\zeta}{2}, \qquad \text{so that} \qquad b + \eta K + \frac{\zeta}{2} = R + \zeta.$$

By (114) and (112) there is T with

$$\beta_1\,|\xi_t| \le b, \qquad \rho_t \le \zeta/2, \qquad |x_{t+1} - x_t| \le \eta K + \zeta/2 \qquad (t \ge T). \tag{115}$$

For $t \ge T$ write $x_{t+1} = x_t - d_t$ with $d_t := w_{t+1}m_{t+1}/(c\lambda)$. If $|x_t| > b$, the first identity in (113) and (115) give

$$\operatorname{sign} d_t = \operatorname{sign} x_t, \qquad \frac{|m_{t+1}|}{\lambda} \;\ge\; |x_t| - \beta_1\,|\xi_t| \;\ge\; |x_t| - b, \qquad |x_{t+1}| = \big||x_t| - |d_t|\big|. \tag{116}$$

We first show that the trajectory eventually remains in $\{|x| \le R + \zeta\}$. Suppose $t \ge T$ and $|x_t| \le R + \zeta$. If $|x_t| \le b$, then

$$|x_{t+1}| \le |x_t| + |d_t| \le b + \eta K + \zeta/2 = R + \zeta.$$

If instead $b < |x_t| \le R + \zeta$, (116) gives

$$|x_{t+1}| \le \max\{|x_t|\,, |d_t|\} \le R + \zeta.$$

Thus $\{|x| \le R + \zeta\}$ is forward invariant after time T.

It remains to show that the trajectory reaches this interval in finite time. Suppose that $|x_s| > R + \zeta$ for $T \le s \le t$. By (116) and (115),

$$|d_s| \le \eta K + \zeta/2 < |x_s|\,,$$

so the step points inward and

$$|x_{s+1}| = |x_s| - |d_s| < |x_s|\,.$$

In particular, $|x_s| \le |x_T|$ throughout this period. Hence (103) yields

$$v_{s+1} \le \overline{V} := \max\{v_T, \lambda^2 x_T^2\}.$$

Moreover,

$$|d_s| = \frac{\eta\,|m_{s+1}|}{\sqrt{v_{s+1}} + \varepsilon} \ge \frac{\eta\lambda(|x_s| - b)}{\sqrt{\overline{V}} + \varepsilon} > \frac{\eta\lambda(\eta K + \zeta/2)}{\sqrt{\overline{V}} + \varepsilon} =: d_{\min} > 0.$$

Therefore, as long as the trajectory remains outside $R + \zeta$, its distance from the origin decreases by at least $d_{\min}$ at each step:

$$|x_{t+1}| \le |x_T| - (t - T + 1)d_{\min}.$$

It must consequently enter $\{|x| \le R + \zeta\}$ after finitely many steps, and by the preceding argument it cannot leave afterward. Since $\zeta > 0$ is arbitrary,

$$\limsup_{t\to\infty} |x_t| \le R.$$

Finally, (100) gives

$$v_t \le \beta_2^{t-T}v_T + \lambda^2 \sup_{s\ge T} x_s^2,$$

and therefore

$$\limsup_{t\to\infty} \sqrt{v_t} \le \lambda R = \frac{\eta K \lambda}{1-\beta_1},$$

which proves (110).

Floor. $w_{\max} = c\eta\lambda/\varepsilon > 2$ gives $\zeta' := c\eta\lambda/2 - \varepsilon > 0$. By (110), $\sqrt{v_t} \le \lambda R + \zeta'$ for all sufficiently large t, so

$$w_t = \frac{c\eta\lambda}{\sqrt{v_t}+\varepsilon} \;\ge\; \frac{c\eta\lambda}{\lambda R + \zeta' + \varepsilon} \;=\; \frac{c\eta\lambda}{\eta\lambda K/(1-\beta_1) + c\eta\lambda/2} \;=\; \frac{c}{K/(1-\beta_1)+c/2} = \alpha,$$

which is (111). □

Lemma H.4 (Edge identity)**.** *Let $0 < \beta_1 < 1$. Fix one coordinate and write*

$$p := w_t, \qquad w := w_{t+1}, \qquad u := \frac{w}{p}, \qquad y := y_t, \qquad y_- := y_{t-1}, \qquad \Delta := y - y_-. \tag{117}$$

Thus p and w are the curvatures at two consecutive times, u is their ratio, and Δ is the last step. The positions satisfy the two-step recursion

$$y_{t+1} = \big[1 + \beta_1 u - (1+\beta_1)w\big]\, y - \beta_1 u\, y_-. \tag{118}$$

Moreover, put

$$E_t := \frac{y_t^2 - \beta_1 y_{t-1}^2}{1+\beta_1}, \qquad Q_t := \beta_1(u\Delta - wy)^2 = \frac{[y_{t+1} + (w-1)y]^2}{\beta_1}, \qquad R_t := \gamma(u-1)\big[(u+1)\Delta^2 - 2y\Delta\big], \tag{119}$$

so that $Q_{t-1} = [y + (p-1)y_-]^2/\beta_1$. Then

$$y_t^2\, w_{t+1}(w_{t+1} - 2) \;=\; E_{t+1} - E_t - Q_t + R_t. \tag{120}$$

Proof. The previous update, the last equation of (105) at time $t-1$, gives $\mu_t = -c\Delta/p$. Substituting this into (105) removes the momentum and gives (118). Substituting (118) into $E_{t+1} - E_t$ and collecting terms gives (120). □

The potential is built from the pieces of Lemma H.4 and one cubic term in the second moment. Put $\nu := 1 - \beta_2$ and $\Gamma(V) := \frac{8\gamma}{3}\min\{\sqrt{V},\, 1-e\}^3$. For $t \ge 1$ let

$$\Phi_t := \begin{cases} -\dfrac{c}{2}\,(y_t - y_{t-1})^2, & w_t \le 1, \\[2ex] E_t - \dfrac{\beta_1}{2}\,Q_{t-1} = \dfrac{y_t^2 - \beta_1 y_{t-1}^2}{1+\beta_1} - \dfrac{1}{2}\big[y_t + (w_t - 1)y_{t-1}\big]^2, & w_t > 1, \end{cases} \;-\; \Gamma(V_t). \tag{121}$$

Below curvature one the potential is a single kinetic square. Above curvature one it is E_t corrected by the previous iterate. The cubic term accounts for the change of curvature.

The proof of the growth bound uses only three facts about the dynamics and four inequalities between the parameters. The facts are:

(F1) the second-moment identity and the speed bound, $(1/w_{t+1} - e)^2 = \beta_2(1/w_t - e)^2 + \nu y_t^2$ and $w_{t+1} \le w_t/\sqrt{\beta_2}$;
(F2) $V_t \ge \nu y_{t-1}^2$;
(F3) eventually, $w_t > 1$ implies $w_{t+1} > c$.

(F1) and (F2) hold for all parameters. (F3) is proved in Step 1 below. The parameter inequalities are

$$\beta_1^2 + \beta_1 > 1, \qquad \beta_2 \ge \frac{(1+\beta_1)^2}{4}, \qquad 4\gamma\nu < \frac{c}{2} - \Big(1 - \frac{\beta_1^2}{2}\Big)c^2, \qquad (1+\beta_1)\sqrt{\frac{\beta_2}{\beta_2 - \beta_1^2}} + \frac{1}{2} < \frac{1}{c}. \tag{122}$$

The case in which both curvatures exceed one also uses explicit numerical bounds. All of these hold at Adam's default parameters $(\beta_1, \beta_2) = (0.9, 0.999)$, where $c = 1/19$, $\gamma = 9/19$ and $\nu = 1/1000$. There the four inequalities in (122) read $1.71 > 1$, $0.999 \ge 0.9025$, $0.00189 < 0.02467$ and $4.87 < 19$. We prove the lemma at these parameters.

Lemma H.5 (Potential inequality)**.** *Let* $(\beta_1, \beta_2) = (0.9, 0.999)$*. The potential* (121) *satisfies* $\Phi_t \le 0$ *if* $w_t \le 1$ *and* $\Phi_t \le y_t^2/(1+\beta_1)$ *if* $w_t > 1$*. In particular* $\Phi_t \le y_t^2/(1+\beta_1)$*. Along every trajectory of* (105)*, from any initialization, and for all sufficiently large* t*, it satisfies the growth bound*

$$\Phi_{t+1} - \Phi_t \;\ge\; y_t^2\, w_{t+1}(w_{t+1} - 2). \tag{123}$$

So (121) *is a potential in the sense of Definition H.1, with* $C = 1/(1+\beta_1)$*.*

Proof. The upper bounds hold because every term of (121) except $y_t^2/(1+\beta_1)$ is nonpositive.

For the growth bound, use the shorthand (117) and let $\Psi := \Phi_{t+1} - \Phi_t - y^2 w(w-2)$ be the margin. By (118), $y_{t+1} = y + \beta_1 u\Delta - (1+\beta_1)wy$. We first prove (F3) and a bound for downward steps. Then we show $\Psi \ge 0$ in four cases, according to whether p and w lie above or below 1. Only Step 1 needs t to be large.

Step 1: if $p > 1$*, then eventually* $w > c$*.* Unrolling the momentum in (105) writes y through the earlier positions:

$$y = \big[1 - (1+\beta_1)p\big]y_- - (1+\beta_1)p\sum_{j=1}^{t-1}\beta_1^j y_{t-1-j} - \frac{\beta_1^t p}{c}\,\mu_0.$$

The second moment controls the earlier positions: $V_t \ge \nu\sum_{j=0}^{t-1}\beta_2^j y_{t-1-j}^2$. Cauchy–Schwarz with these weights bounds the linear part of y. The triangle inequality for $\sqrt{V_{t+1}} = \big\|(\sqrt{\beta_2 V_t}, \sqrt{\nu}\, y)\big\|_2$ then gives

$$\sqrt{V_{t+1}} \;\le\; \sqrt{V_t\Big[1 + \beta_2 - 2(1+\beta_1)p + \frac{(1+\beta_1)^2\beta_2}{\beta_2 - \beta_1^2}\,p^2\Big]} + \frac{\beta_1^t p\sqrt{\nu}}{c}\;|\mu_0|\,.$$

Now let $p > 1$. Then $V_t = (1/p - e)^2 \le p^{-2}$, $1 + \beta_2 < 2(1+\beta_1)p$ and $p \le 1/e$. Hence

$$\sqrt{V_{t+1}} + e \;\le\; (1+\beta_1)\sqrt{\frac{\beta_2}{\beta_2 - \beta_1^2}} + e + \frac{\beta_1^t\sqrt{\nu}}{ce}\;|\mu_0|\,.$$

Since $e < \frac{1}{2}$, the last inequality in (122) makes the right-hand side smaller than $1/c$ for all large t. So $w > c$. This holds for every initialization.

Step 2: a bound for downward steps. Let $u \le 1$ and $y \ne 0$. Then $1/w - e \ge u^{-1}(1/p - e)$, so (F1) gives $u^{-2}V_t \le \beta_2 V_t + \nu y^2$. Together with (F2),

$$\Big|\frac{y_-}{y}\Big| \;\le\; \frac{u}{\sqrt{1 - \beta_2 u^2}}. \tag{124}$$

If instead $y = 0$, then $V_{t+1} = \beta_2 V_t \le V_t$, so $w \ge p$. In particular a step from $p > 1$ to $w \le 1$ has $y \ne 0$.

Step 3: both curvatures at most one. Then $\min\{\sqrt{V}, 1-e\} = 1 - e$ at both times, so the cubic term does not change, and

$$\Psi = \frac{c}{2}\Delta^2 - \frac{c}{2}\big[\beta_1 u\Delta - (1+\beta_1)wy\big]^2 + w(2-w)y^2.$$

By Cauchy–Schwarz with the weights $\beta_1^2 u^2$ and $1 - \beta_1^2 u^2$, $\Psi \ge 0$ whenever $\frac{c}{2}(1+\beta_1)^2\, w/(2-w) \le 1 - \beta_1^2 u^2$. This holds because $w \le 1$, $u^2 \le 1/\beta_2$ by (F1), and $\beta_2 \ge (1+\beta_1)^2/4 \ge 2\beta_1^2/(1+\beta_1^2)$. No lower bound on u is needed.

Step 4: upward crossings, $p \le 1 < w$*.* Here $\sqrt{V_t} \ge 1 - e > \sqrt{V_{t+1}}$, so the cubic term $-\Gamma$ increases. The rest of Ψ is the quadratic form $A\Delta^2 + By\Delta + Cy^2$ with

$$A = \frac{c}{2}\big(1 + \beta_1^2 u^2\big), \qquad B = 2\beta_1 u\Big[\frac{1}{1+\beta_1} - \Big(1 - \frac{\beta_1}{2}\Big)w\Big], \qquad C = c + \beta_1\Big(1 - \frac{\beta_1}{2}\Big)w^2.$$

Its determinant factors as

$$AC - \frac{B^2}{4} = \frac{c}{2}\,C + \frac{\beta_1^2 u^2}{1+\beta_1}\Big[\frac{c}{2} - \Big(1 - \frac{\beta_1}{2}\Big)(w-1)^2\Big].$$

Here $1 < w \le \beta_2^{-1/2} \le 2/(1+\beta_1) = 1+c$, so the bracket is at least $\frac{c}{2} - (1-\frac{\beta_1}{2})c^2 = \frac{c}{2(1+\beta_1)}(-1+4\beta_1-\beta_1^2) > 0$ by $\beta_1^2+\beta_1 > 1$. So the form is positive definite and $\Psi \ge 0$. This also covers $y = 0$.

Step 5: downward crossings, $p > 1 \ge w$. Now $c < w \le 1$ by Step 1, and $0 < u = w/p < w$. The cubic term $-\Gamma(V)$ changes by at least $-4\gamma\nu y^2$: as a function of V its derivative lies in $[-4\gamma, 0]$, and $V_{t+1} - V_t \le \nu y^2$. After this bound, $\Psi \ge A\Delta^2 + B\Delta y + Cy^2$ with

$$A = \gamma + \tfrac{1}{2}(p-1)^2 - \tfrac{c}{2}\beta_1^2 u^2, \qquad B = \beta_1(1-\beta_1)uw - 2\gamma - p(p-1),$$
$$C = 2w - \big[1 + \tfrac{1}{2}(1-\beta_1^2)\big]w^2 - c + \tfrac{1}{2}p^2 - 4\gamma\nu.$$

By Step 2, $y \ne 0$. Put $z := y_-/(uy)$, so that $\Delta = (1-uz)y$ and (124) reads $(1-\beta_2 u^2)z^2 \le 1$. We claim

$$\frac{\Psi}{y^2} \;\ge\; \mathcal{H}(w) := w - \Big(1 - \frac{\beta_1^2}{2}\Big)w^2 - \frac{c}{2} - 4\gamma\nu \;>\; 0. \tag{125}$$

For the first inequality, put $\vartheta := \beta_1(1-\beta_1)/2$. Subtract $\mathcal{H}(w) + \frac{1}{2}w(1-w)\big[1-(1-\beta_2u^2)z^2\big]$, which is at least $\mathcal{H}(w)$, from $A(1-uz)^2 + B(1-uz) + C$. What remains is the quadratic az^2+bz+d in z, with

$$a = \frac{w}{2} - wu + \Big[\frac{1}{2} + \gamma - \frac{\beta_2}{2}w(1-w)\Big]u^2 - \gamma\vartheta u^4,$$
$$b = (w-u) - 2\vartheta u^2(w-\gamma u), \qquad d = \frac{w}{2} + \vartheta u(2w - \gamma u).$$

Since $0 < u \le w \le 1$, we have $d \ge w/2 > 0$, and the two terms of b are nonnegative, so $b^2 \le (w-u)^2 + 4\vartheta^2u^4(w-\gamma u)^2$. The exact factorization

$$a - \frac{(w-u)^2}{2w} = \gamma u^2 - \gamma\vartheta u^4 + \frac{u(1-w)}{2}\Big(2 - \frac{u}{w} - \beta_2 wu\Big)$$

has a nonnegative last term. Hence

$$a - \frac{b^2}{4d} \;\ge\; \gamma u^2 - \gamma\vartheta u^4 - 2\vartheta^2u^4w \;\ge\; \gamma u^2(1-2\vartheta) \;>\; 0,$$

using $2\vartheta \le \gamma$ and $u, w \le 1$. So the remaining quadratic is positive, which proves the first inequality in (125). Finally, $\mathcal{H}$ is concave, $\mathcal{H}(c) > 0$ by the third inequality in (122), and $\mathcal{H}(1) - \mathcal{H}(c) = (1-c)(\beta_1^2+\beta_1-1)/(1+\beta_1) > 0$. So $\mathcal{H} > 0$ on $[c, 1]$.

Step 6: both curvatures above one. Put $\chi := \beta_1(1-\beta_1/2) = 99/200$ and $k := \chi - \gamma = 81/3800$. The quadratic part of Ψ is $A\Delta^2 + 2By\Delta + Cy^2$ (here B is half the cross coefficient), with

$$A = ku^2 + \frac{(p-1)^2}{2} + \gamma, \qquad B = -\chi pu^2 - \frac{p(p-1)}{2} + \gamma(u-1), \qquad C = p^2\Big(\chi u^2 + \frac{1}{2}\Big).$$

At both times $\sqrt{V} < 1-e$, so $\Gamma(V) = \frac{8\gamma}{3}V^{3/2}$ there. If $y = 0$, then $w \ge p$ by Step 2, the cubic term does not decrease, and $\Psi \ge A\Delta^2 \ge 0$. So let $y \ne 0$.

The cubic term is worst at $e = 0$. Put $s := \sqrt{V_t} = 1/p - e$ and $d := 1/w - 1/p$, so that $\sqrt{V_{t+1}} = s + d$. Here d is fixed once p and w are. By (F1), $y^2 = [(s+d)^2 - \beta_2 s^2]/\nu$, and the change of the cubic term, divided by y^2, is

$$\rho_e := -\frac{\Gamma(V_{t+1}) - \Gamma(V_t)}{y^2} = -\frac{8\gamma\nu}{3}\,d\,\frac{3s^2+3ds+d^2}{\nu s^2 + 2ds + d^2}.$$

Its derivative in s is $-\frac{8\gamma\nu}{3}\,d^2\big[\beta_2 s^2 + 2(1+\beta_2)s(s+d) + (s+d)^2\big]/\big[(s+d)^2 - \beta_2 s^2\big]^2 \le 0$. The denominator stays positive as s increases to $1/p$: for upward steps this follows from $u < \beta_2^{-1/2}$, and for downward steps it is immediate. So ρ_e is smallest at $s = 1/p$, that is at $e = 0$:

$$\rho_e \;\ge\; \rho_0 := \frac{8\gamma\nu}{3p}\,\frac{u^3-1}{u(1-\beta_2u^2)}. \tag{126}$$

For a downward step,

$$-\rho_0 \;\le\; \frac{8\gamma\nu}{3}\Big(\frac{1}{p} + \frac{1}{w}\Big) \;\le\; \frac{16\gamma\nu}{3} = \frac{6}{2375}. \tag{127}$$

Large downward steps, $u \le \beta_1$. Put $U := 1/p$ and $P := (AC - B^2)/p^2$, as a function of $v := u^2$. Direct differentiation gives

$$\frac{\partial^2 P}{\partial v^2} = -2\chi\gamma + \frac{3\chi\gamma U}{2u} - \frac{\gamma(1-U) + 2\gamma^2 U^2}{4u^3} \;<\; 0 \qquad (U \le u \le 1).$$

At the two ends,

$$P\big|_{u=1} = \frac{\chi}{2}(p-2)^2 \ge 0, \qquad P\big|_{u=U} = \frac{k}{2}(1-2U)^2 + \frac{\gamma}{2}U^2 + \gamma k U^2(1-U)^2 \;\ge\; \frac{81}{8968} > \frac{9}{1000}.$$

By concavity, $P \ge (1-u^2)\cdot\frac{9}{1000} \ge \frac{171}{100000}$. Since $A/p^2 \le \frac{1}{2}$, minimizing over Δ gives $C - B^2/A \ge \frac{171}{50000} > \frac{6}{2375}$. Together with (127), $\Psi > 0$.

All other steps, $\beta_1 \le u < \beta_2^{-1/2} < 1.001$. Put

$$X := p - 2, \qquad z := \frac{\Delta}{y} - 2, \qquad \tau := u - 1, \qquad F_1 := \frac{p-1}{2} - \chi u^2, \qquad F_2 := \chi u^2 + \frac{1}{2}.$$

Expanding about $\Delta = 2y$, the quadratic part divided by y^2 is exactly

$$Az^2 + 2\big[F_1 X - \gamma(2u+1)\tau\big]z + F_2 X^2 - 4\gamma u\tau.$$

The bound (126) satisfies the exact identity

$$\rho_0 - 4\gamma u\tau = -\frac{4\gamma u}{p}X\tau + \frac{8\gamma}{p}T(u)\tau^2, \qquad T(u) := \frac{3\beta_2 u^2(u+1) - \nu(2u+1)}{3u(1-\beta_2 u^2)}.$$

On this interval $T(u) \ge \frac{3}{2}u/(1-\beta_2 u^2)$. So Ψ/y^2 is bounded below by a quadratic form in (X, z, τ). Its (X, z) block is

$$\mathcal{B} = \begin{pmatrix} F_2 & F_1 \\ F_1 & A \end{pmatrix}, \qquad \det\mathcal{B} = \frac{\chi}{2}w^2 + \gamma F_2(1-u^2) \;\ge\; \frac{6}{25}w^2 > 0.$$

(For $u > 1$ the second term is negative, but $u^2 < 1/\beta_2$ and $F_2 < 1$ give $\gamma F_2(1-u^2) > -0.0005$, while $(\chi/2 - 6/25)w^2 > 0.0075$ because $w > 1$.) Also $A/p^2 \le \frac{1}{2}$, $|F_1|/p \le \frac{1}{2}$ and $F_2 < 1$. Minimizing over (X, z) costs at most

$$\begin{aligned}\frac{\gamma^2}{\det\mathcal{B}}\Big[\frac{4u^2 A}{p^2} - \frac{4u(2u+1)F_1}{p} + (2u+1)^2 F_2\Big]\tau^2 \;&\le\; \frac{\gamma^2(10u^2+6u+1)}{(6/25)w^2}\,\tau^2 \\ &\le\; \frac{75}{4}\,\frac{U^2}{u^2}\,\tau^2 \;\le\; \frac{625}{27}\,U^2\tau^2,\end{aligned}$$

since $10u^2 + 6u + 1 < 18$ and $\gamma < \frac{1}{2}$. The available coefficient of τ^2 is larger:

$$\frac{8\gamma}{p}T(u) \;\ge\; \frac{12\gamma U u}{1-\beta_2 u^2} \;\ge\; \frac{243}{10}\,U \;>\; \frac{625}{27}\,U^2,$$

using $u \ge 9/10$, $\gamma \ge 9/20$ and $1 - \beta_2 u^2 < 1/5$. So the quadratic form is nonnegative and $\Psi \ge 0$. This treats upward and downward steps together and finishes the proof. □

H.2 Proof of the subcritical regime

H.2.1 Main argument

This subsection proves the subcritical item of Theorem 3.4. Proposition H.6 reduces it to three hypotheses (H1)–(H3) with constants ρ, CB, R and the condition $RCB < 1$. Lemmas H.7–H.9 (Appendix H.2.2) supply the hypotheses; Proposition H.10 (Appendix H.2.3) chooses the constants and gives the cutoff.

Setting and notation. Let $n = 2$, $0 < \beta_1 < 1$ and $\beta_1^2 < \beta_2 < 1$, and write $\sigma := \sqrt{\beta_2} \in (\beta_1, 1)$. The notation of this subsection is local to it. By (44)–(45), (100) reads

$$z_{i,t+1} = \mathcal{A}(w_{i,t+1})\, z_{i,t}, \qquad \mathcal{A}(w) = \begin{pmatrix} 1-w & -\beta_1 w \\ 2-w & \beta_1(1-w) \end{pmatrix}, \qquad w_{i,t} \geq \sigma w_{i,t+1}. \tag{128}$$

The last inequality follows from $v_{i,t+1} \geq \sigma^2 v_{i,t}$. Put

$$w_\pm := 1 \pm \frac{2\sqrt{\beta_1}}{1+\beta_1}, \qquad \kappa := \frac{1-\beta_1}{c^2} = \frac{(1+\beta_1)^2}{1-\beta_1}, \tag{129}$$

so that $w_- < c^2 < w_+ < 2$ (indeed $w_+ w_- = c^2$ and $2\sqrt{\beta_1} < 1 + \beta_1$).

Fix a level q, and suppose that $W_{\mathrm{act},t} \leq q$ for all sufficiently large t. Write, as in (101),

$$u_{i,t} := \lambda_i x_{i,t}^2\, w_{i,t+1}\, (w_{i,t+1} - q)_+, \qquad d_{i,t} := \lambda_i x_{i,t}^2\, w_{i,t+1}\, (q - w_{i,t+1})_+ \tag{130}$$

for the excess of a coordinate above q and its deficit below q. By (102) with $\omega = q$, the hypothesis $W_{\mathrm{act},t} \leq q$ is exactly

$$u_{i,t} \leq d_{j,t} \quad (j \neq i). \tag{131}$$

Proposition H.6 (Abstract subcritical criterion)**.** *Let $0 < q < 2$, fix a threshold $S \in [c^2, w_+)$, and let*

$$E_{i,t} := \lambda_i\, z_{i,t}^\top \mathcal{P}_S(w_{i,t})\, z_{i,t}, \qquad \mathcal{P}_S(w) := \varphi(\bar{w}) M(\bar{w}), \qquad \bar{w} := \min\{S, \max\{c^2, w\}\},$$
$$\varphi(w) := \frac{w_+ - c^2}{w_+ - w}, \qquad M(w) := \begin{pmatrix} 2-w & \frac{1-\beta_1}{2}(w-1) \\ \frac{1-\beta_1}{2}(w-1) & \beta_1 w \end{pmatrix} \tag{132}$$

be the coordinate energies ($E_{i,t} \geq 0$ since $\mathcal{P}_S(w) \succeq M(c^2) \succ 0$, Lemma H.7). Let $\rho < 1$ and $R, C, B, N > 0$ be constants such that, for all sufficiently large t, each step of each coordinate is of one of two kinds, low *($w_{i,t+1} \leq S$) or* high *($w_{i,t+1} > S$):*

(H1) Low-curvature contraction. *On a low step, $E_{i,t+1} \leq \rho\, E_{i,t}$; a low window of length N contracts by the factor B.*

(H2) Compensation. *If coordinate i is high at time t, then the other coordinate $j \neq i$ is low on the preceding window of length N, and*

$$\max\{u_{i,t}, u_{i,t-1}\} \leq CB \max\{E_{j,t-N}, E_{j,t-N-1}\}. \tag{133}$$

(H3) Excursion reset. *On a high step, $E_{i,t+1} \leq R \max\{u_{i,t}, u_{i,t-1}\}$.*

If $RCB < 1$, then $W_{\mathrm{act},t} \leq q$ cannot hold for all sufficiently large t; that is, $W_{\mathrm{act},t} > q$ for arbitrarily large t.

Proof. Suppose $W_{\mathrm{act},t} \leq q$ for all $t \geq T$, with T so large that (H1)–(H3) hold for $t \geq T$. Put

$$\bar{\rho} := \max\{\rho, RCB\} < 1, \qquad \mathcal{E}_t := \max_{\substack{i \in \{1,2\} \\ t-N-1 \leq s \leq t}} E_{i,s}. \tag{134}$$

For $t \geq T$ and $i = 1, 2$,

$$E_{i,t+1} \;\leq\; \bar{\rho}\, \mathcal{E}_t : \tag{135}$$

on a low step $E_{i,t+1} \leq \rho E_{i,t}$ by (H1); on a high step $E_{i,t+1} \leq R \max\{u_{i,t}, u_{i,t-1}\} \leq RCB \max\{E_{j,t-N}, E_{j,t-N-1}\}$ ($j \neq i$) by (H3), (H2). Hence $\mathcal{E}_{t+1} \leq \mathcal{E}_t$. Every entry of $\mathcal{E}_{t+N+2}$ is some $E_{i,s+1}$ with $t \leq s \leq t + N + 1$, so (135) gives

$$\mathcal{E}_{t+N+2} \;\leq\; \bar{\rho}\, \mathcal{E}_t, \qquad \lambda_i x_{i,T+n}^2 \;\leq\; E_{i,T+n} \;\leq\; \mathcal{E}_T\, \bar{\rho}^{\,\lfloor n/(N+2) \rfloor} \qquad (n \geq 0), \tag{136}$$

the first inequality in the second display by Lemma H.7. By (100) and (136),

$$v_{i,T+n} = \beta_2^n v_{i,T} + (1-\beta_2)\lambda_i^2 \sum_{s=0}^{n-1} \beta_2^{n-1-s} x_{i,T+s}^2 \;\xrightarrow[n\to\infty]{}\; 0,$$

$$w_{i,T+n} = \frac{c\eta\lambda_i}{\sqrt{v_{i,T+n}} + \varepsilon} \;\to\; w_{i,\max} > 2 > q.$$

So $w_{i,t+1} > q$ for $i = 1, 2$ and all large t, and (11) with $x_t \neq 0$ gives $W_{\mathrm{act},t} > q$, a contradiction. □

Proof of the subcritical item. Let $q_0 < \overline{W}_{\mathrm{diag}}(\beta_1, \beta_2)$ as in (143), $N := \lfloor L \rfloor + 1$ and $q := S\sigma^{N+1}$. By Proposition H.10, $q \geq \overline{W}_{\mathrm{diag}} > q_0$, and if $W_{\mathrm{act},t} \leq q$ for all large t then (H1)–(H3) hold with $RCB < 1$, which Proposition H.6 rules out. Hence $W_{\mathrm{act},t} > q > q_0$ at arbitrarily late times. □

H.2.2 Lemmas for Proposition H.6

Lemma H.7 (Coordinate energy). *Let $c^2 \leq S < w_+$, let $E_{i,t}$ be the energy* (132)*, and put*

$$\mu := \frac{\beta_1}{\sigma}\Big(1 + \frac{1-\sigma}{w_+ - 1}\Big), \qquad G(w) := \mu\,\frac{w_+ - \sigma w}{w_+ - w} \tag{137}$$

and, for $\beta_1 < \theta < 1$, the contraction ceiling

$$C(\theta) := \begin{cases} c^2, & \theta \leq G(c^2), \\ w_+ \dfrac{\theta - \mu}{\theta - \sigma\mu}, & \theta > G(c^2), \end{cases} \qquad c^2 \leq C(\theta) < w_+. \tag{138}$$

(G is increasing, and $G(C(\theta)) = \theta$ in the second case.) Then $\lambda_i x_{i,t}^2 \leq E_{i,t}$, and whenever $w = w_{i,t+1} \leq \min\{S, C(\theta)\}$ for some $\theta \in (\beta_1, 1)$,

$$E_{i,t+1} \;\leq\; \big[1 - \min\{\kappa w,\, 1-\theta\}\big]\, E_{i,t}. \tag{139}$$

Proof. Direct computation gives

$$\mathcal{A}(w)^\top M(w)\mathcal{A}(w) = \beta_1 M(w), \qquad 4\det M(w) = (1+\beta_1)^2(w_+ - w)(w - w_-),$$

so $M(w) \succ 0$ for $w_- < w < w_+$ ($M_{11} = 2 - w > 0$) and $M(w_\pm) \succeq 0$. Since M is affine,

$$\frac{d}{dw}\,\frac{M(w)}{w_+ - w} = \frac{M(w_+)}{(w_+ - w)^2} \;\succeq\; 0.$$

Thus $\mathcal{P}_S$ is nondecreasing, $\mathcal{P}_S(w) \succeq M(c^2)$, and $z^\top M(c^2) z = x^2 + \beta_1 c^2\big(h - 2x/(1-\beta_1)\big)^2 \geq x^2$.

For $0 \leq w \leq c^2$, convexity of $X \mapsto X^\top M(c^2) X$ along the affine $w \mapsto \mathcal{A}(w)$ gives

$$\mathcal{A}(w)^\top M(c^2)\mathcal{A}(w) \preceq \Big(1 - \frac{w}{c^2}\Big)\mathcal{A}(0)^\top M(c^2)\mathcal{A}(0) + \frac{w}{c^2}\,\beta_1 M(c^2) \preceq (1 - \kappa w)\, M(c^2),$$

since $\mathcal{A}(0)$ fixes x and multiplies the second square by β_1^2. With $\mathcal{P}_S(w) = M(c^2) \preceq \mathcal{P}_S(w_{i,t})$ this is $E_{i,t+1} \leq (1 - \kappa w) E_{i,t}$.

For $c^2 \leq w \leq S$, put $p := \max\{\sigma w, c^2\} \in [c^2, w]$. Affinity also gives

$$M(p) \;\succeq\; \frac{p - w_-}{w - w_-}\, M(w), \qquad \frac{w - w_-}{p - w_-} \;\leq\; \frac{1}{\sigma}\Big(1 + \frac{1-\sigma}{w_+ - 1}\Big).$$

(the difference of the two sides of the first is $(w-p)M(w_-)/(w - w_-) \succeq 0$; the second uses $w \leq p/\sigma$, $p \geq c^2$ and $w_-/(c^2 - w_-) = 1/(w_+ - 1)$). By (128), $\bar{w}_{i,t} \geq p$, so $\mathcal{P}_S(w_{i,t}) \succeq \mathcal{P}_S(p)$ and

$$E_{i,t+1} = \beta_1 \varphi(w)\, \lambda_i z_{i,t}^\top M(w) z_{i,t} \;\leq\; \beta_1\, \frac{w_+ - p}{w_+ - w}\, \frac{w - w_-}{p - w_-}\, E_{i,t} \;\leq\; G(w)\, E_{i,t},$$

the last step by $p \geq \sigma w$. If $\theta \leq G(c^2)$ then $C(\theta) = c^2$ and only the case $w \leq c^2$ occurs; otherwise $G(w) \leq G(C(\theta)) = \theta$ for $c^2 \leq w \leq C(\theta)$. Both $1 - \kappa w$ and θ are at most $1 - \min\{\kappa w, 1 - \theta\}$. □

Lemma H.8 (Reset). *For $0 < q < \sigma S$ put*

$$\begin{aligned} J(q,S) &:= 1 + 2\Big(\frac{1}{\sigma S} - 1\Big)_+ + \frac{1-\sigma}{\sigma(\sigma S - q)}, \\ R(q,S) &:= (3 - \beta_1)\,\varphi(S)\, J(q,S)^2 \max\Big\{\frac{\beta_1}{S(S-q)},\, (1+\beta_1)^2\Big\}. \end{aligned} \tag{140}$$

and let $u_{i,t} = \lambda_i x_{i,t}^2\, w_{i,t+1}(w_{i,t+1} - q)_+$ be as in (130)*. On a step with $w_{i,t+1} > S$ and $t \geq 1$,*

$$E_{i,t+1} \;\leq\; R(q,S) \max\{u_{i,t}, u_{i,t-1}\}. \tag{141}$$

The bound remains valid with $R(Q,S)$ for $q \leq Q < \sigma S$.

Proof. Write $w = w_{i,t+1}$, $p = w_{i,t} \geq \sigma w > q$, $z = (x, h)^\top = z_{i,t}$, and normalize $\max\{u_{i,t}, u_{i,t-1}\} = \lambda_i = 1$ (the zero case is trivial). Then

$$|x| \leq \frac{1}{\sqrt{w(w-q)}}, \qquad |x_{i,t-1}| \leq \frac{1}{\sqrt{p(p-q)}}, \qquad h = (1-p^{-1})\,x + p^{-1}x_{i,t-1},$$

the last by $x_{i,t} - x_{i,t-1} = -w_{i,t}m_{i,t}/(c\lambda_i)$ (100). With $r := (w-q)/(\sigma w - q) \geq \sqrt{(w-q)/(\sigma(\sigma w - q))}$,

$$|h|\sqrt{w(w-q)} \;\leq\; \bigl|1-p^{-1}\bigr| + \frac{r}{p} \;\leq\; 1 + 2\Bigl(\frac{1}{p}-1\Bigr)_+ + \frac{r-1}{p},$$
$$\frac{r-1}{p} = \frac{(1-\sigma)w}{p(\sigma w - q)} \;\leq\; \frac{1-\sigma}{\sigma(\sigma S - q)}$$

($p \geq \sigma w$, $w > S$, $p \geq \sigma S$), hence $\|z\|_\infty \leq J(q,S)/\sqrt{w(w-q)}$.

Let $\mathbf{1} = (1,1)^\top$, $\mathbf{b} = (1, \beta_1)^\top$ and $\|y\|_{M(S)} := (y^\top M(S) y)^{1/2}$. Then

$$\mathcal{A}(w) = \mathcal{A}(S) - (w-S)\,\mathbf{1}\mathbf{b}^\top, \qquad \|\mathcal{A}(S)z\|_{M(S)} = \sqrt{\beta_1}\;\|z\|_{M(S)}\,,$$
$$\|z\|^2_{M(S)} \leq (3-\beta_1)\,\|z\|^2_\infty\,, \qquad \|\mathbf{1}\|^2_{M(S)} = 1+\beta_1 \leq 3-\beta_1,$$

and the triangle inequality gives

$$\|\mathcal{A}(w)z\|_{M(S)} \;\leq\; \frac{J(q,S)\sqrt{3-\beta_1}}{\sqrt{w(w-q)}}\,\bigl[\sqrt{\beta_1} + (w-S)(1+\beta_1)\bigr].$$

Since $\sqrt{w(w-q)} \geq \sqrt{S(S-q)} + (w-S)$,

$$\frac{\sqrt{\beta_1} + (w-S)(1+\beta_1)}{\sqrt{w(w-q)}} \;\leq\; \max\Bigl\{\frac{\sqrt{\beta_1}}{\sqrt{S(S-q)}},\, 1+\beta_1\Bigr\};$$

squaring and multiplying by $\varphi(S)$ ($\mathcal{P}_S(w) = \varphi(S)M(S)$) gives (141). Replacing q by $Q \geq q$ only enlarges the position bounds. □

Lemma H.9 (Compensation)**.** *Let $\theta \in (\beta_1, 1)$, let $N \geq 1$ be an integer, let $0 < q \leq \min\{S\sigma^{N+1}, C(\theta)\}$, and suppose that $W_{\mathrm{act},s} \leq q$ for all $s \geq T$. If $t \geq T+N+1$ and $w_{i,t+1} > S$, then the other coordinate $j \neq i$ has $w_{j,k+1} \leq q$ for $t-N-1 \leq k \leq t$ (in particular it is low there), and*

$$\max\{u_{i,t}, u_{i,t-1}\} \leq D\max\{E_{j,t-N}, E_{j,t-N-1}\},$$
$$D := \max\Bigl\{\frac{q}{e\kappa N\sigma^N},\; \frac{q^2\theta^N}{4}\Bigr\}. \tag{142}$$

Proof. By (131), $u_{i,s} \leq d_{j,s}$ for $s \geq T$, so at most one curvature exceeds q at any time. By (128), $w_{i,k+1} \geq \sigma^{t-k} w_{i,t+1} > \sigma^{N+1} S \geq q$ for $t-N-1 \leq k \leq t$, hence $w_{j,k+1} \leq q$ there. For $s \in \{t-1, t\}$ put $u := w_{j,s+1}$; the preceding N curvatures of j lie in $[\sigma^N u, q] \subset [0, \min\{S, C(\theta)\}]$, so Lemma H.7 and $\lambda_j x^2_{j,s} \leq E_{j,s}$ give

$$u_{i,s} \;\leq\; d_{j,s} \;\leq\; u(q-u)\bigl[1 - \min\{\kappa\sigma^N u, 1-\theta\}\bigr]^N E_{j,s-N}$$
$$\leq\; \max\Bigl\{qu\,e^{-\kappa\sigma^N N u},\; \frac{q^2}{4}\theta^N\Bigr\}E_{j,s-N} \;\leq\; D\,E_{j,s-N},$$

using $u(q-u) \leq \min\{qu, q^2/4\}$, $1-a \leq e^{-a}$ and $\sup_{u\geq 0} ue^{-Au} = 1/(eA)$. □

H.2.3 THE CONSTANTS AND THE EXPLICIT CUTOFF

The cutoff. The following explicit choices give a cutoff that depends on (β_1, β_2) only:

$$\begin{aligned}\theta &:= \min\{\beta_1^{3/4},\ \beta_1 + \sqrt{(1-\beta_1)(1-\beta_2)}\}, \qquad S := C(\sqrt{\beta_1}), \qquad Q := \min\{\beta_2 S,\ C(\theta)\},\\ R &:= R(Q,S) = (3-\beta_1)\,\frac{w_+ - c^2}{w_+ - S}\Big(1 + 2\Big(\frac{1}{\sigma S} - 1\Big)_+ + \frac{1-\sigma}{\sigma(\sigma S - Q)}\Big)^2\\ &\qquad\qquad \times \max\Big\{\frac{\beta_1}{S(S-Q)},\ (1+\beta_1)^2\Big\},\\ \tau &:= -\log\sigma, \qquad L := \max\Big\{\frac{\log(S/Q)}{\tau} - 1,\ \frac{R\sigma S}{e\kappa},\ \frac{\log(R\beta_2 S^2/4)}{2\tau - \log\theta}\Big\},\\ &\qquad\qquad \boxed{\overline{W}_{\mathrm{diag}}(\beta_1,\beta_2) := S\,\sigma^{L+2}}\,.\end{aligned} \tag{143}$$

Here θ and $\sqrt{\beta_1}$ lie in $(\beta_1, 1)$, C is the contraction ceiling (138), and R is the constant (140) at $q = Q$. All quantities are elementary; $0 < \overline{W}_{\mathrm{diag}} < S < w_+ < 2$.

Proposition H.10 (When $RCB < 1$ holds)**.** *Let θ, S, Q and R be as in* (143)*, and let $N \geq 1$ be an integer with*

$$S\sigma^{N+1} \leq Q, \qquad \frac{R\sigma S}{e\kappa N} < 1, \qquad \frac{R\beta_2 S^2}{4}\,(\beta_2\theta)^N < 1. \tag{144}$$

Put $q := S\sigma^{N+1}$. Then $0 < q \leq Q < 2$, and if $W_{\mathrm{act},t} \leq q$ for all $t \geq T$, then for all sufficiently large t the hypotheses (H1)–(H3) *of Proposition H.6 hold at level q and threshold S, with*

$$\rho = \max\{1 - \kappa\alpha, \sqrt{\beta_1}\} < 1, \qquad CB = D \text{ of } (142), \qquad R = R(Q,S),$$

and $RCB = RD < 1$. The integer $N := \lfloor L \rfloor + 1$ satisfies (144)*, and then $q \geq \overline{W}_{\mathrm{diag}}(\beta_1, \beta_2)$.*

Proof. First, $0 < q \leq Q \leq \beta_2 S < S < w_+ < 2$. We check the three hypotheses.

(H1) A low step has $w_{i,t+1} \leq S = C(\sqrt{\beta_1})$, so Lemma H.7 with $\theta = \sqrt{\beta_1}$ and the curvature floor $w_{i,t+1} \geq \alpha$ of (111) give $E_{i,t+1} \leq \rho E_{i,t}$ for all large t. (The floor is Lemma H.3, which assumes only $\beta_1^2 < \beta_2$.)

(H2) Since $q = S\sigma^{N+1}$ and $q \leq Q \leq C(\theta)$, Lemma H.9 gives (H2) with $CB = D$.

(H3) Since $q \leq Q < \sigma S$, Lemma H.8 gives (H3) with $R = R(Q,S)$.

With $q = S\sigma^{N+1}$, $RD = \max\{Rq/(e\kappa N\sigma^N),\ Rq^2\theta^N/4\} = \max\{R\sigma S/(e\kappa N),\ R\beta_2 S^2(\beta_2\theta)^N/4\} < 1$ by (144). Finally $L < N \leq L + 1$ gives the three inequalities (144), one per entry of the maximum defining L (for the third, $2\tau - \log\theta = -\log(\beta_2\theta)$), and $q = S\sigma^{N+1} \geq S\sigma^{L+2} = \overline{W}_{\mathrm{diag}}$. □

Remark H.11 (The standard parameters)*.* At $(\beta_1, \beta_2) = (9/10, 999/1000)$ one has exactly $c^2 = 1/361$, $\kappa = 361/10$, $3 - \beta_1 = 21/10$ and $(1+\beta_1)^2 = 361/100$. The other quantities in (143) are, first in closed form and then numerically,

$$\begin{aligned}\sigma &= \sqrt{999/1000} && \approx 0.999500,\\ \sqrt{\beta_1} &= 3/\sqrt{10} && \approx 0.948683,\\ w_\pm &= 1 \pm 6\sqrt{10}/19 && \approx 1.998614,\ 0.001386,\\ \theta &= \min\{(9/10)^{3/4},\ 91/100\} = 91/100 && = 0.91,\\ \mu &= \frac{9}{10\sigma}\Big(1 + \frac{19(1-\sigma)}{6\sqrt{10}}\Big) && \approx 0.900901,\\ G(c^2) &= \mu\,\frac{w_+ - \sigma/361}{w_+ - 1/361} && \approx 0.900902,\\ S &= C(\sqrt{\beta_1}) = w_+\,\frac{\sqrt{\beta_1} - \mu}{\sqrt{\beta_1} - \sigma\mu} && \approx 1.979944,\end{aligned}$$

$$
\begin{aligned}
Q = C(\theta) &= w_+ \frac{91/100 - \mu}{91/100 - \sigma\mu} && \approx 1.904313,\\
\varphi(S) &= \frac{w_+ - 1/361}{w_+ - S} && \approx 106.901,\\
J(Q,S) &= 1 + \frac{1-\sigma}{\sigma(\sigma S - Q)} && \approx 1.006704,\\
R &= \frac{21}{10}\,\varphi(S)\,J(Q,S)^2\,\frac{9}{10\,S(S-Q)} && \approx 1367.40,\\
\tau &= \tfrac{1}{2}\log(1000/999) && \approx 5.00250\cdot 10^{-4},\\
L &= \frac{\log(S/Q)}{\tau} - 1 && \approx 76.855.
\end{aligned}
$$

$$
\boxed{1.90336 < \overline{W}_{\mathrm{diag}}(9/10, 999/1000) = S\sigma^{L+2} < 1.90337},
$$

so $\overline{W}_{\mathrm{diag}}(0.9, 0.999) \geq 1.9$ as stated in Theorem 3.4.

I Proofs from Section 3.3

With $\beta_1 = 0$ and $\varepsilon > 0$, (1) reads

$$
\begin{aligned}
&g_t = \nabla f(x_t), \quad v_{t+1} = \beta_2 v_t + (1-\beta_2) g_t^{\odot 2}, \quad D_t = \mathrm{diag}\big(\sqrt{v_{i,t+1}} + \varepsilon\big)_{i=1}^{d},\\
&z_t = D_t^{-1} g_t, \quad x_{t+1} = x_t - \eta z_t,
\end{aligned}
\tag{145}
$$

with $\eta > 0$, $0 \leq \beta_2 < 1$, and $v_0 \geq 0$ coordinatewise. Write the *transition activity*

$$
\Gamma_t := g_t^\top D_t^{-1} g_t = z_t^\top D_t z_t = \sum_{i=1}^{d} \frac{g_{i,t}^2}{\sqrt{v_{i,t+1}} + \varepsilon},
\tag{146}
$$

so that the first-order change of the loss along the step is $g_t^\top (x_{t+1} - x_t) = -\eta \Gamma_t$. If $g_T = 0$ at some step then $z_T = 0$, $x_{T+1} = x_T$, and every later step repeats this: the trajectory is *stationary from T on*. A trajectory that is not stationary in finite time therefore has $g_t \neq 0$ at every step.

Lemma I.1 (Scale-free metric bounds)**.** *Suppose $g_t \neq 0$. Then $z_t \neq 0$, $\|D_t\|_{\mathrm{op}} > 0$, $\Gamma_t > 0$, and*

$$
\|z_t\|_\infty \leq \frac{1}{\sqrt{1-\beta_2}}, \qquad \|z_t\| \leq Z := \sqrt{\frac{d}{1-\beta_2}}, \qquad \frac{\|g_t\|^2}{\|D_t\|_{\mathrm{op}}} \;\leq\; \Gamma_t \;\leq\; \|D_t\|_{\mathrm{op}} \|z_t\|^2 .
\tag{147}
$$

In particular every step has length $\|x_{t+1} - x_t\| \leq \eta Z$.

Proof. We first bound the preconditioned gradient z_t. From the current-gradient term in (145),

$$
v_{i,t+1} \geq (1-\beta_2) g_{i,t}^2.
$$

Hence, on every active coordinate,

$$
|z_{i,t}| = \frac{|g_{i,t}|}{\sqrt{v_{i,t+1}} + \varepsilon} \leq \frac{|g_{i,t}|}{\sqrt{v_{i,t+1}}} \leq (1-\beta_2)^{-1/2}.
$$

On inactive coordinates $z_{i,t} = 0$. Summing over coordinates therefore gives $\|z_t\| \leq Z$.

We next bound Γ_t. Since $g_{i,t} = 0$ outside the active support, (146), together with $\sqrt{v_{i,t+1}} + \varepsilon \leq \|D_t\|_{\mathrm{op}}$, gives the claimed lower bound on Γ_t. For the upper bound, simply use

$$
\Gamma_t = z_t^\top D_t z_t \leq \|D_t\|_{\mathrm{op}} \|z_t\|^2
$$

and the bound on $\|z_t\|$ above.

Finally, if $g_t \neq 0$, then at least one active coordinate satisfies $g_{i,t} \neq 0$, and hence $z_t \neq 0$. Since $D_t \succ 0$, it follows that

$$
\Gamma_t = z_t^\top D_t z_t > 0.
$$

□

I.1 Proof of Theorem 3.5

Throughout, f satisfies hypotheses (a)–(c) of Theorem 3.5.

Under (b) and (c) the stationary set $\mathcal{C}$ of f is finite. For each $\bar{x} \in \mathcal{C}$, we have

$$\|\nabla f(x)\|^2 \;\geq\; \kappa_{\bar{x}}\,|f(x) - f(\bar{x})|\,, \qquad x \in U_{\bar{x}}, \qquad \kappa_{\bar{x}} := \frac{1}{2\Lambda\,\|\nabla^2 f(\bar{x})^{-1}\|_{\mathrm{op}}^2}, \tag{148}$$

on some neighborhood $U_{\bar{x}}$, since $\nabla f(x) = \nabla^2 f(\bar{x})(x - \bar{x}) + o(\|x - \bar{x}\|)$. Put

$$\kappa_f := \min_{\bar{x} \in \mathcal{C}} \kappa_{\bar{x}} > 0, \qquad \delta_\varepsilon := \frac{2\varepsilon}{\eta\,\kappa_f}, \tag{149}$$

with $\kappa_f := \infty$ and $\delta_\varepsilon := 0$ if $\mathcal{C} = \emptyset$. Thus δ_ε is proportional to ε/η and vanishes only in the degenerate case $\mathcal{C} = \emptyset$; it is a constant of (f, η, ε) alone.

We prove Theorem 3.5 with $C_f := 2/\kappa_f$, so that $C_f \eta^{-1} \varepsilon = \delta_\varepsilon$, in the following equivalent form: unless the trajectory is stationary from some finite step on, for every $\delta > \delta_\varepsilon$ neither

$$W'_{\mathrm{act},t} \geq 2 + \delta \quad \text{for all large } t \tag{150}$$

nor

$$W'_{\mathrm{act},t} \leq 2 - \delta \quad \text{for all large } t \tag{151}$$

is possible.

Proof. On a trajectory that is not stationary in finite time, $g_t \neq 0$, $\Gamma_t > 0$ and $\|D_t\|_{\mathrm{op}} > 0$ at every step (Lemma I.1), so $W'_{\mathrm{act},t}$ is defined throughout. Fix $\delta > \delta_\varepsilon$. Two consequences of the recursion are used for both regimes.

(i) The preconditioner is controlled by the current gradient. The Hessian bound makes ∇f Λ-Lipschitz, and each step has length at most ηZ (Lemma I.1), so $\|g_{t-j}\| \leq \|g_t\| + \Lambda\eta Z j$. Unrolling $v_{i,t+1} = \beta_2^{t+1} v_{i,0} + (1-\beta_2)\sum_{j=0}^{t} \beta_2^j g_{i,t-j}^2$ and taking the weighted ℓ^2 norm with weights $(1-\beta_2)\beta_2^j$, whose total mass is at most 1, gives $\sqrt{v_{i,t+1}} + \varepsilon \leq \|g_t\| + C_D$ with $C_D := \varepsilon + \sqrt{\|v_0\|_\infty} + \Lambda\eta Z\sqrt{\beta_2(1+\beta_2)}/(1-\beta_2)$. Hence $\|D_t\|_{\mathrm{op}} \leq C_D + \|g_t\|$ and, by (147),

$$\Gamma_t \;\geq\; \frac{\|g_t\|^2}{C_D + \|g_t\|}, \tag{152}$$

whose right-hand side is increasing and unbounded in $\|g_t\|$: bounded Γ_t forces bounded $\|g_t\|$, and $\Gamma_t \to 0$ forces $g_t \to 0$.

(ii) Vanishing gradients empty the second moment. If $g_t \to 0$ then $v_t \to 0$, being an exponential average of a null sequence, so

$$\|D_t\|_{\mathrm{op}} = \max_i\big(\sqrt{v_{i,t+1}} + \varepsilon\big) \;\longrightarrow\; \varepsilon, \tag{153}$$

and since $\delta > \delta_\varepsilon$, i.e. $\eta\delta\kappa_f > 2\varepsilon$, there is a T_1 with

$$\frac{\eta\delta\kappa_f}{2\,\|D_t\|_{\mathrm{op}}} \;>\; 1 \qquad (t \geq T_1). \tag{154}$$

Supercritical regime. Suppose (150) holds for $t \geq T$. By (13),

$$f(x_{t+1}) - f(x_t) \;\geq\; \frac{\eta\delta}{2}\,\Gamma_t \;\geq\; 0 \qquad (t \geq T), \tag{155}$$

so $f(x_t)$ is nondecreasing. Descent along the step also gives $f(x_{t+1}) - f(x_t) \leq -\eta\Gamma_t + \frac{1}{2}\Lambda\eta^2 Z^2$, so $\Gamma_t \leq \frac{1}{2}\Lambda\eta Z^2$ and, by (152), $\sup_{t \geq T}\|g_t\| < \infty$; coercivity then bounds the tail $\{x_t : t \geq T\}$ and with it f, so $f(x_t) \uparrow f_\infty < \infty$. Summing (155) gives $\sum_{t \geq T}\Gamma_t \leq \frac{2}{\eta\delta}(f_\infty - f(x_T)) < \infty$, hence $\Gamma_t \to 0$, $g_t \to 0$, and (153) and (154) apply.

Every accumulation point of the bounded tail is then a critical point of value f_∞, so $\mathcal{C} \neq \emptyset$ and, for all large t, x_t lies in the union of the neighborhoods $U_{\bar{x}}$ of those accumulation points. Applying (148) there with $\kappa_{\bar{x}} \geq \kappa_f$,

$$\|g_t\|^2 \;\geq\; \kappa_f\, \Delta_t, \qquad \Delta_t := f_\infty - f(x_t) \geq 0, \tag{156}$$

and (155), (147) and (156) combine into

$$\Delta_{t+1} \;\leq\; \Delta_t - \frac{\eta\delta}{2}\,\frac{\|g_t\|^2}{\|D_t\|_{\mathrm{op}}} \;\leq\; \Delta_t\Big(1 - \frac{\eta\delta\kappa_f}{2\,\|D_t\|_{\mathrm{op}}}\Big).$$

The factor is negative for $t \geq T_1$ by (154) while $\Delta_{t+1} \geq 0$, so $\Delta_t = 0$ for all large t; then (155) forces $\Gamma_t = 0$, that is $g_t = 0$, and the trajectory is stationary from a finite step on, a contradiction.

Subcritical regime. Suppose (151) holds for $t \geq T$. By (13),

$$f(x_{t+1}) - f(x_t) \;\leq\; -\frac{\eta\delta}{2}\,\Gamma_t \qquad (t \geq T), \tag{157}$$

and f is bounded below, so $\sum_{t\geq T}\Gamma_t < \infty$, $\Gamma_t \to 0$, $g_t \to 0$ and $f(x_t) \downarrow f_\infty$. The tail is bounded as before and the same local gap gives $\|g_t\|^2 \geq \kappa_f\Delta_t$ with $\Delta_t := f(x_t) - f_\infty$, whence $\Delta_{t+1} \leq \Delta_t\big(1 - \eta\delta\kappa_f/(2\,\|D_t\|_{\mathrm{op}})\big)$ with an eventually negative factor: again $\Delta_t = 0$ for all large t and (157) forces $\Gamma_t = 0$, a contradiction.

Neither (150) nor (151) is therefore possible for any $\delta > \delta_\varepsilon$, which is (15) with $C_f = 2/\kappa_f$. □

I.2 Analyzing the active curvature

Define

$$e_t := \eta\,\frac{z_t^\top(\overline{H}_t - H_t)z_t}{\Gamma_t}, \qquad W'_{\mathrm{act},t} = W_{\mathrm{act},t} + e_t, \qquad f(x_{t+1}) - f(x_t) = \frac{\eta}{2}\Gamma_t\big(W_{\mathrm{act},t} - 2 + e_t\big). \tag{158}$$

Further, introduce the directional curvature

$$\varkappa_t := \frac{z_t^\top H_t z_t}{\|z_t\|^2}. \tag{159}$$

The proposition is straightforward; we omit the proof.

Proposition I.2. *Suppose that*

$$\big\|\nabla^2 f(x) - \nabla^2 f(y)\big\|_{\mathrm{op}} \leq L_H\,\|x - y\|_\infty\,. \tag{160}$$

Then

$$\big\|\overline{H}_t - H_t\big\|_{\mathrm{op}} \leq \frac{L_H\eta}{3}\,\|z_t\|\,, \qquad |e_t| = \big|W'_{\mathrm{act},t} - W_{\mathrm{act},t}\big| \leq \frac{L_H\eta^2}{3}\,\frac{\|z_t\|^3}{\Gamma_t}. \tag{161}$$

If moreover $\varkappa_t > 0$, then

$$\big|W'_{\mathrm{act},t} - W_{\mathrm{act},t}\big| \leq \frac{L_H\eta\,\|z_t\|}{3\varkappa_t}\,W_{\mathrm{act},t} \leq \frac{L_H Z\eta}{3\varkappa_t}\,W_{\mathrm{act},t}, \tag{162}$$

with Z as in Lemma I.1.

The second bound follows from the first: by (5) with $c = 1$ and (159), $W_{\mathrm{act},t} = \eta\, z_t^\top H_t z_t/\Gamma_t = \eta\varkappa_t\,\|z_t\|^2/\Gamma_t$, and $\|z_t\| \leq Z$ by Lemma I.1. Thus, whenever the step direction carries curvature bounded away from zero, the measured $W_{\mathrm{act},t}$ and the loss criterion $W'_{\mathrm{act},t}$ differ by a relative error $O(\eta)$.

Consecutive gradients. The same averaging argument transfers the reversal identity (17) of Section 4 from quadratics to general objectives. Write

$$\widehat{H}_t := \int_0^1 \nabla^2 f(x_t - s\eta z_t)\, ds, \qquad \hat{e}_t := \eta\, \frac{z_t^\top (\widehat{H}_t - H_t) z_t}{\Gamma_t}, \qquad \widehat{P}_t := \eta\, D_t^{-1/2} \widehat{H}_t D_t^{-1/2}, \tag{163}$$

so that $g_{t+1} = g_t - \eta \widehat{H}_t z_t$ by the fundamental theorem of calculus, and $\hat{z}_t^\top \widehat{P}_t \hat{z}_t = W_{\mathrm{act},t} + \hat{e}_t$. The weight in $\widehat{H}_t$ is uniform, whereas $\overline{H}_t$ in (158) carries the weight $2(1-s)$; this is why the constants below are $\frac{1}{2}$ where those of Proposition I.2 are $\frac{1}{3}$.

Proposition I.3 (Gradient reversal on general objectives)**.** *Let $\beta_1 = 0$ and $g_t \neq 0$. Then*

$$\begin{aligned} \frac{\langle g_t, g_{t+1}\rangle_{D_t^{-1}}}{\|g_t\|^2_{D_t^{-1}}} &= 1 - W_{\mathrm{act},t} - \hat{e}_t, \\ \cos(D_t^{-1/2} g_t, D_t^{-1/2} g_{t+1}) &= \frac{1 - W_{\mathrm{act},t} - \hat{e}_t}{\sqrt{(1 - W_{\mathrm{act},t} - \hat{e}_t)^2 + \widehat{V}_t}}, \end{aligned} \tag{164}$$

with $\widehat{V}_t := \left\| (\widehat{P}_t - (W_{\mathrm{act},t} + \hat{e}_t) I) \hat{z}_t \right\|^2$. Under (160)*,*

$$|\hat{e}_t| \le \frac{L_H \eta^2}{2} \frac{\|z_t\|^3}{\Gamma_t}, \qquad \textit{and, if } \varkappa_t > 0, \qquad |\hat{e}_t| \le \frac{L_H \eta \|z_t\|}{2\varkappa_t} W_{\mathrm{act},t} \le \frac{L_H Z \eta}{2\varkappa_t} W_{\mathrm{act},t}. \tag{165}$$

In particular $W_{\mathrm{act},t} > 1 + |\hat{e}_t|$ implies $\langle g_t, g_{t+1}\rangle_{D_t^{-1}} < 0$.

Proof. Since $x_{t+1} - x_t = -\eta z_t$ and $z_t = D_t^{-1} g_t$,

$$\begin{aligned} \langle g_t, g_{t+1}\rangle_{D_t^{-1}} &= g_t^\top D_t^{-1} g_t - \eta\, g_t^\top D_t^{-1} \widehat{H}_t z_t = \Gamma_t - \eta\, z_t^\top H_t z_t - \eta\, z_t^\top (\widehat{H}_t - H_t) z_t \\ &= \Gamma_t \big(1 - W_{\mathrm{act},t} - \hat{e}_t\big) \end{aligned}$$

holds because $\eta\, z_t^\top H_t z_t = \Gamma_t W_{\mathrm{act},t}$ by (5) with $c = 1$. Dividing by $\Gamma_t = \|g_t\|^2_{D_t^{-1}}$ gives the first identity. For the second, $D_t^{-1/2} g_{t+1} = \left\| D_t^{-1/2} g_t \right\| (I - \widehat{P}_t)\hat{z}_t$, so the cosine equals $\hat{z}_t^\top (I - \widehat{P}_t)\hat{z}_t / \left\| (I - \widehat{P}_t)\hat{z}_t \right\|$, and $\left\| (I - \widehat{P}_t)\hat{z}_t \right\|^2 = (1 - \hat{z}_t^\top \widehat{P}_t \hat{z}_t)^2 + \widehat{V}_t$ by the Pythagorean decomposition of $(I - \widehat{P}_t)\hat{z}_t$ along and orthogonal to $\hat{z}_t$. For the bound, (160) and $\|\cdot\|_\infty \le \|\cdot\|$ give $\left\| \nabla^2 f(x_t - s\eta z_t) - H_t \right\|_{\mathrm{op}} \le L_H s \eta \|z_t\|$, hence

$$\left\| \widehat{H}_t - H_t \right\|_{\mathrm{op}} \le L_H \eta \|z_t\| \int_0^1 s\, ds = \frac{L_H \eta}{2} \|z_t\|, \quad |\hat{e}_t| \le \frac{\eta \|z_t\|^2 \left\| \widehat{H}_t - H_t \right\|_{\mathrm{op}}}{\Gamma_t} \le \frac{L_H \eta^2 \|z_t\|^3}{2\Gamma_t}.$$

The relative form follows from $W_{\mathrm{act},t} = \eta \varkappa_t \|z_t\|^2 / \Gamma_t$ and $\|z_t\| \le Z$ (Lemma I.1). □