Provable Edge-of-Stability for Adam on a One-Dimensional Quadratic

Yiman Fong* Heng Yang†

August 24, 2026

Abstract

The edge-of-stability (EoS) phenomenon of Adam has been widely observed, while its underlying dynamical mechanism is not yet fully understood. We study uncorrected Adam on a one-dimensional quadratic, a clean setting where constant curvature isolates the optimizer-induced dynamics behind the EoS. We characterize the resulting dynamics across the parameter space. In broad regimes, we prove that Adam exhibits a restoring tendency toward its frozen stability threshold $2(1+\beta_1)/[\eta(1-\beta_1)]$. We also identify settings in which this edge-seeking mechanism breaks down, including strictly subcritical periodic orbits and specially tuned trajectories that converge to the optimum while remaining uniformly supercritical. These results give a concrete dynamical explanation for Adam's EoS in a setting free of evolving loss geometry, while also exposing its limitations.

1 Introduction

Adam [Kingma and Ba, 2015] is among the most widely used optimizers in modern deep learning. When neural networks are trained with first-order methods such as gradient descent or Adam, a striking empirical regularity emerges—the *edge of stability* (EoS): the sharpness (the largest eigenvalue of the Hessian of the loss function) rises toward the stability threshold and then oscillates around it (Figure 1, left). For gradient descent, this behavior has been observed extensively in neural-network training, where the sharpness approaches the classical threshold $2/\eta$ for learning rate η [Cohen et al., 2021]. For adaptive methods, Cohen et al. [2022] observed a similar phenomenon for Adam: the *preconditioned sharpness* S_t—the largest eigenvalue of the Hessian after Adam's coordinatewise rescaling—rises toward and then stays near the stability threshold of the corresponding frozen dynamics. These results establish the EoS as a robust but so far largely empirical phenomenon and a strong departure from classical optimization theory. While EoS attracts extensive theoretical investigation on gradient descent, the mechanism of EoS for Adam remains mysterious. This leaves a basic question unresolved for Adam: *why* does this happen?

In this work, we provide theoretical evidence that, in contrast to gradient descent, Adam's adaptive nature by itself induces EoS. More specifically, we work in the simplest possible setting, the one-dimensional quadratic

$$f(x) = \tfrac{1}{2}x^2, \tag{1}$$

and prove that edge-of-stability behavior actually occurs, identifying the mechanism that produces it, illustrated by the negative-feedback loop in Figure 1 (right). On this objective, uncorrected Adam iterates

$$m_{t+1} = \beta_1 m_t + (1-\beta_1)x_t, \qquad v_{t+1} = \beta_2 v_t + (1-\beta_2)x_t^2, \qquad x_{t+1} = x_t - \frac{\eta}{\sqrt{v_{t+1}}+\varepsilon}m_{t+1}, \tag{2}$$

*School of Engineering and Applied Sciences, Harvard University. Email: `yiman_fong@seas.harvard.edu`, `fangyimin05@gmail.com`

†School of Engineering and Applied Sciences, Harvard University. Email: `hankyang@seas.harvard.edu`

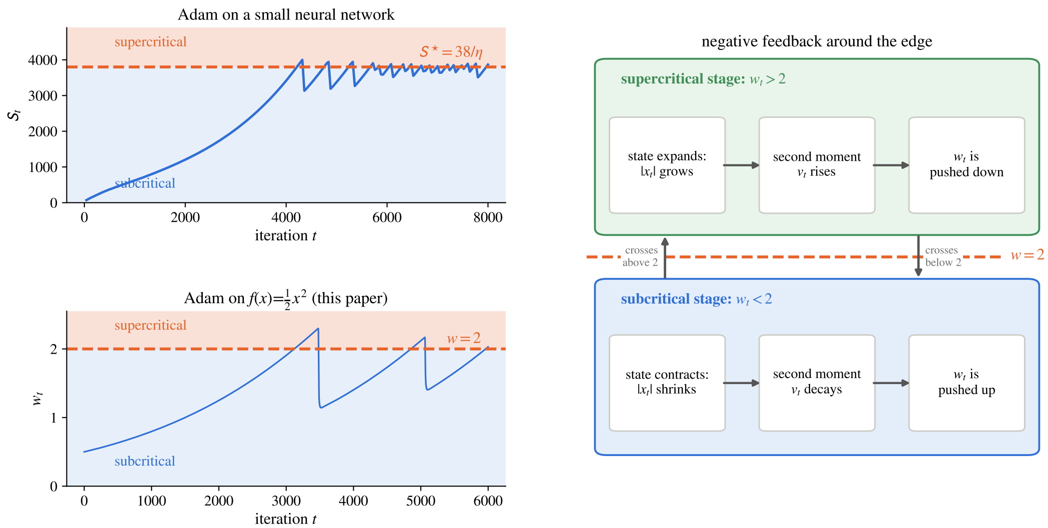

Figure 1: Edge-of-stability behavior and its negative-feedback mechanism. Left: Adam's sharpness oscillates around its frozen stability threshold in both neural-network training and the one-dimensional quadratic studied here. Right: above the edge, expansion raises v_t and pushes w_t down; below the edge, contraction lowers v_t and pushes w_t up. Experimental details are given in Appendix H.

with momentum and second-moment parameters $0 \le \beta_1 < 1$ and $0 \le \beta_2 < 1$, learning rate $\eta > 0$, stabilizer $\varepsilon > 0$, and initial state (x_0, m_0, v_0). We use the normalized sharpness $w_t = c\eta\,\mathsf{S}_t$ with $c = (1-\beta_1)/(1+\beta_1)$, for which the frozen stability threshold becomes the parameter-free boundary $w = 2$. The key mechanism is a negative-feedback loop driven by Adam's second-moment adaptation: above the edge the iterates expand, v_t inflates, the effective step size shrinks, and w_t is pushed back down; below the edge the iterates contract, v_t decays, and w_t is pushed back up. In broad regimes we prove that this feedback forces w_t to keep fluctuating around 2: it cannot persistently stay well above or well below the edge. This built-in self-stabilization of the effective step size may also be part of why Adam performs so well in practice.

Turning this mechanism into proofs requires arguments on the two sides of the edge; Figure 2 summarizes the resulting proof structure (Section 3). In the subcritical regime (left), an adaptive Lyapunov function certifies contraction up to an explicit cutoff $\overline{W} < 2$, which weakens the second-moment forcing and pushes w_t upward toward the edge. This mechanism is not universal: in the complementary parameter regime, under certain conditions, strictly subcritical periodic orbits can arise when $\varepsilon = 0$. In the supercritical regime (right), the sign geometry separates two distinct behaviors. Aligned trajectories expand and are forced out of every fixed supercritical band in finite time, whereas exceptional persistently misaligned trajectories can contract to the origin while remaining uniformly supercritical.

Taken together, these results prove that Adam's adaptive state can generate the negative feedback toward the frozen stability boundary illustrated in Figure 1, even when the loss curvature is completely fixed. At the same time, positive momentum introduces precise ways in which this edge-seeking mechanism can fail, illustrating the additional complexity of the general setting. Thus, our analysis identifies both an optimizer-induced mechanism for the EoS and its limitations already in the simplest quadratic setting. Extending this exact discrete analysis to higher dimensions is an important direction for future work.

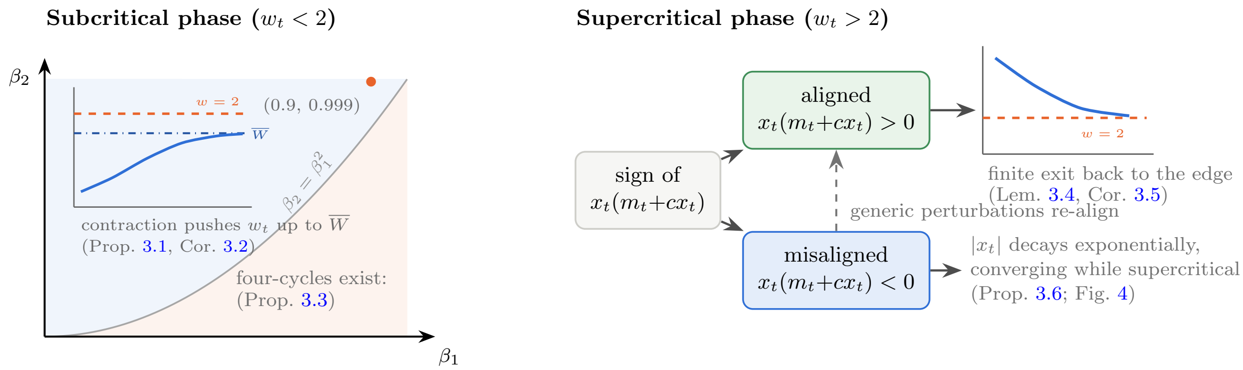

Figure 2: Proof roadmap for $\beta_1 > 0$. Left: below the edge, an adaptive Lyapunov function certifies contraction up to a cutoff $\overline{W} < 2$ when $\beta_2 > \beta_1^2$ (Proposition 3.1, Corollary 3.2), while strictly subcritical four-cycles can arise for certain complementary parameter choices when $\varepsilon = 0$ (Proposition 3.3). Right: above the edge, the sign of $x_t h_t$ separates aligned trajectories, which exit every fixed supercritical band in finite time, from exceptional persistently misaligned trajectories that converge while remaining uniformly supercritical (Lemma 3.4, Corollary 3.5, Proposition 3.6).

2 Mathematical formulation and the case of $\beta_1 = 0$

2.1 The stability threshold

To see where the threshold comes from, freeze the second moment in (2) at a constant value $v \geq 0$. Define

$$c := \frac{1-\beta_1}{1+\beta_1}, \qquad w_t := \frac{c\eta}{\sqrt{v_t}+\varepsilon}, \qquad w_{\max} := \frac{c\eta}{\varepsilon}. \tag{3}$$

Then $0 < w_t \leq w_{\max}$, and stability of the frozen update is decided by this scalar alone:

Lemma 2.1 (Frozen stability in one dimension)**.** *Fix $v \geq 0$ and let $\mathcal{A}_v$ denote the map $(x_t, m_t) \mapsto (x_{t+1}, m_{t+1})$ obtained from (2) by freezing v_{t+1} at v. Then $\mathcal{A}_v$ is linear, and its spectral radius (the largest modulus of its eigenvalues) is smaller than 1 if and only if*

$$w := \frac{c\eta}{\sqrt{v}+\varepsilon} < 2.$$

The lemma is the one-dimensional case of Lemma B.2 in Appendix B, which treats general objectives in any dimension. Thus $w = 2$ is the frozen stability boundary, free of all parameters. In practice ε is tiny (e.g. 10^{-8}), so $w_{\max} > 2$ in essentially every realistic configuration; this is the regime we study, and the globally subcritical case $w_{\max} < 2$ is treated in Appendix G.

Note that Schur stability of the frozen iteration is only a pointwise criterion: v_t, and hence w_t, evolves along the trajectory, so it does not by itself determine the stability of the adaptive dynamics—yet our results show that this local boundary is precisely the level around which w_t oscillates. In the general setting of neural-network training, the frozen operator, the preconditioned sharpness S_t, and its threshold $\mathsf{S}^\star$ are defined analogously in Appendix B.

2.2 Illustrative example: $\beta_1 = 0$ case

As an illustrative example, in this section we focus on the case $\beta_1 = 0$, in which Adam reduces to RMSProp [Tieleman and Hinton, 2012]. This is the simpler case because the momentum variable disappears from the position recursion and $c = 1$. Adam then reduces to

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)x_t^2, \qquad x_{t+1} = (1-w_{t+1})x_t, \tag{4}$$

where we recall $w_t = \frac{\eta}{\sqrt{v_t+\varepsilon}}$. Bai et al. [2026b] studied a related threshold-crossing mechanism for a momentum-free one-dimensional quadratic under additional conditions. Appendix D strengthens this intuition through a two-sided finite-passage result: for every $\underline{w} < 2 < \overline{w}$, a trajectory cannot remain indefinitely below $\underline{w}$ or above $\overline{w}$. Since these levels may be chosen arbitrarily close to 2, the result formalizes the restoring behavior toward the edge. If w_t remains uniformly below 2, the position contracts, reducing the forcing of v_t and pushing w_t upward; if it remains uniformly above 2, the reverse mechanism pushes w_t downward. This is the feedback illustrated in Figure 1. The positive-momentum analysis in Section 3 follows the same high-level strategy, but requires an adaptive Lyapunov argument below the edge and sign geometry above it.

3 Analysis of Adam with general parameters

The main focus of this paper is the positive-momentum regime, corresponding to the practical choice $(\beta_1, \beta_2) = (0.9, 0.999)$ of Adam. In this regime, we identify subtler EoS behavior of Adam, which requires more careful analysis. The frozen threshold is still $w = 2$, but momentum changes what happens near it. 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. All proofs for this section are deferred to Appendix E (subcritical results) and Appendix F (supercritical results).

Normalized dynamics and phase decomposition Introduce the normalized momentum coordinate

$$h_t := c^{-1}m_t + x_t, \qquad z_t := \begin{pmatrix} x_t \\ h_t \end{pmatrix}. \tag{5}$$

Then, the update rule on (z_t) can be written as

$$z_{t+1} = A(w_{t+1})z_t, \qquad A(w) := \begin{pmatrix} 1-w & -\beta_1 w \\ 2-w & \beta_1(1-w) \end{pmatrix}. \tag{6}$$

Since $A(w)$ is similar to the frozen map of Lemma 2.1, its spectral radius satisfies $\rho(A(w)) < 1$ exactly when $0 < w < 2$. The matrices encountered along an Adam trajectory, however, vary with w_t and need not commute. Stability of every individual matrix is therefore not sufficient to control their product. Our subcritical argument uses the exact restriction imposed on successive values of w_t by the second-moment recursion, whereas the supercritical argument uses the sign geometry of (x_t, h_t).

3.1 Subcritical analysis

In the subcritical regime, decay of (x_t, m_t) reduces the forcing in the second-moment recursion. The resulting decrease of v_t raises w_t and therefore moves the trajectory toward the frozen boundary. The principal difficulty is that the family $\{A(w) : 0 < w < 2\}$ does not admit a common Euclidean contraction estimate. We instead use the Lyapunov functional

$$\Psi_t := x_t^2 + \frac{\beta_1 w_t}{2-w_t} \cdot h_t^2. \tag{7}$$

The coefficient of h_t^2 is chosen as a function of the current w_t. The exact one-step quadratic inequality, together with the admissible change in w_t, gives the following cutoff:

$$\overline{W} := \frac{2(\sqrt{\beta_2} - \beta_1)}{\sqrt{\beta_2}(1-\beta_1) - 2\beta_1(1-\sqrt{\beta_2})/w_{\max}}. \tag{8}$$

Proposition 3.1. *Assume $w_{\max} > 2$ and $\beta_2 > \beta_1^2$. Then $0 < \overline{W} < 2$; if $\Psi_t > 0$ and $w_t, w_{t+1} < \overline{W}$, then $\Psi_{t+1} < \Psi_t$. More quantitatively, for every $0 < a \le w < \overline{W}$, if $a \le w_t, w_{t+1} \le w$ then $\Psi_{t+1} \le \big(1 - \delta(a, w)\big)\Psi_t$, where the explicit quantity $\delta(a, w) \in (0, 1)$ is defined in Appendix E.*

When $1-\beta_2 \ll 1-\beta_1$ and $\varepsilon/(c\eta) = 1/w_{\max}$ is negligible, the certified contraction region reaches close to the frozen boundary: for $\beta_1 = 0.9$ and $\beta_2 = 0.999$ the cutoff is $\overline{W} \approx 1.991$. An expansion of $\overline{W}$ as $\beta_2 \to 1$ is given in (56) of Appendix E.

Because Ψ_t keeps decreasing, the trajectory cannot stay below any fixed level $w < \overline{W}$ forever.

Corollary 3.2 (Finite passage toward the subcritical cutoff)**.** *Assume the hypotheses of Proposition 3.1. Then, for any $w < \overline{W}$ and $T \in \mathbb{N}_0$ such that $w_T \leq w$, there exists $\tau < \infty$ such that $w_{T+\tau} > w$.*

A quantitative version of Corollary 3.2 is stated in Corollary E.2.

3.1.1 Existence of subcritical cycles

Proposition 3.1 gives a sufficient contraction region for $\beta_2 > \beta_1^2$ which is indeed the regime of the standard parameter choice $(\beta_1, \beta_2) = (0.9, 0.999)$. In the complementary parameter range, strictly subcritical cycling can occur.

Proposition 3.3 (Existence of strictly subcritical four-cycles)**.** *In the scale-free case $\varepsilon = 0$, suppose that*

$$\sqrt{2} - 1 \leq \beta_1 < 1, \qquad 0 \leq \beta_2 \leq \beta_1^2. \tag{9}$$

Then, for every $\eta > 0$, there exists an initial state (x_0, m_0, v_0) whose Adam orbit has prime period four and satisfies $w_t < 2$ for every $t \geq 0$.

Figure 3 in Appendix E displays the construction at a concrete point of the parameter region (9).

3.2 Supercritical analysis

In the supercritical region $w > 2$, the sign of $x_t h_t$ separates two sharply different behaviors. When x_t and h_t are aligned, the next step preserves alignment and expands the position. This expansion feeds the second-moment recursion, increases v_t, and therefore pushes w_t back toward the edge. When they are misaligned, continued misalignment instead forces the position to contract even though the frozen matrix is unstable. The following lemma records the one-step geometry.

Lemma 3.4 (Supercritical sign geometry)**.** *Suppose that $w_{t+1} > 2$.*

(i) *If $x_t h_t > 0$, then $x_{t+1} h_{t+1} > 0$ and*

$$|x_{t+1}| > (w_{t+1} - 1)\,|x_t|\,. \tag{10}$$

(ii) *If $x_t h_t < 0$ and $x_{t+1} h_{t+1} < 0$, then*

$$|x_{t+1}| < \frac{|x_t|}{w_{t+1} - 1}. \tag{11}$$

In particular, an aligned trajectory that remains above a fixed margin $2+\delta$ expands at least at rate $1+\delta$. Such expansion cannot persist indefinitely: it drives the second moment so high that w_t can no longer stay above $2+\delta$. This yields a quantitative exit from every fixed band above the edge.

Corollary 3.5 (Finite exit from a uniformly supercritical band)**.** *Suppose that $x_T h_T > 0$ and $w_T \geq 2+\delta$ for some $\delta \in (0,1]$, and define*

$$\tau := \min\{n \in \mathbb{N}_{\geq 1} : w_{T+n} < 2+\delta\}. \tag{12}$$

Then

$$\tau \leq 1 + \delta^{-1} \log\left(1 + \frac{\delta (c\eta)^2}{(1-\beta_2) x_T^2}\right). \tag{13}$$

Consequently, if an aligned trajectory remains supercritical for all future times, then $\liminf_{t\to\infty} w_t = 2$.

The unstable convergence stage Part (ii) of Lemma 3.4 raises the question of whether misalignment must eventually end. The answer is no: for every initial v_0 with $w_0 > 2$, one can choose the initial position and momentum so that the trajectory remains misaligned forever. It then converges to the origin along an exceptional contracting trajectory even though its frozen dynamics remain unstable.

Proposition 3.6 (Persistent supercritical misalignment)**.** *Suppose that $0 < \beta_1 < 1$ and $w_{\max} > 2$. For every $v_0 > 0$ satisfying $w_0 = c\eta/(\sqrt{v_0} + \varepsilon) > 2$, there exist $x_0 \neq 0$ and $m_0 \in \mathbb{R}$ such that*

$$x_t(m_t + cx_t) < 0 \qquad \text{for every } t \geq 0. \tag{14}$$

Along this trajectory,

$$w_t \geq w_0, \qquad |x_t| \leq (w_0 - 1)^{-t} |x_0|, \qquad t \geq 0. \tag{15}$$

In particular, $(x_t, m_t, v_t) \to (0, 0, 0)$ and $w_t \to w_{\max}$, although the trajectory remains uniformly supercritical for all time.

4 Conclusion

We studied the exact discrete dynamics of uncorrected Adam on a one-dimensional quadratic. Our analysis identifies a restoring mechanism toward the frozen stability boundary. With momentum, we establish this behavior up to an explicit subcritical cutoff close to the threshold under standard parameter regimes. We also show its limitations through strictly subcritical periodic orbits and persistently supercritical yet convergent trajectories. These results demonstrate that Adam can generate edge-of-stability behavior intrinsically, while momentum introduces dynamical mechanisms that can prevent universal convergence to the edge. Extending this analysis to higher-dimensional settings is an important direction for future work.

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, pages 169–195, 2023.

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.

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, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 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*, page 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, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 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 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.

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.

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.

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, editors, *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 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.

Zhonglin Xie, Yiman Fong, Haoran Yuan, and Zaiwen Wen. Accelerating optimization via differentiable stopping time. *Advances in Neural Information Processing Systems*, 38:72185–72212, 2026.

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.

A Related Work

Edge of stability. The edge-of-stability (EoS) phenomenon was systematically documented by Cohen et al. [2021], with early 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 and Bruna, 2023]. Other works connect EoS to bifurcation and nonlinear oscillatory dynamics [Song and Yun, 2023, Mulayoff and Stich, 2026, Kalra et al., 2025]. These results primarily concern gradient descent. For adaptive methods, Cohen et al. [2022] identified an analogous adaptive EoS characterized by the preconditioned Hessian. Our work studies this phenomenon for the exact discrete Adam dynamics on a quadratic objective, where the underlying curvature is fixed and the effective sharpness evolves solely through adaptive preconditioning.

Adam. AdaGrad [Duchi et al., 2011] introduced coordinatewise adaptive learning rates based on accumulated gradients, and Kingma and Ba [2015] later introduced Adam by combining adaptive second-moment scaling with momentum. Despite its empirical success, Adam can fail to converge [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 and Gazeau, 2020, Barakat and Bianchi, 2020]. A separate line of work studies how to tune Adam's hyperparameters, for instance adapting its learning rate online [Xie et al., 2026]. 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]. However, comparatively 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 $\beta_1 = 0$. Regis and Chewi [2026] instead develop a continuous-time model for Adam in the EoS regime. From a discrete dynamical perspective, Bock and Weiß [2019] showed that Adam can admit non-convergent limit cycles, including quadratic examples, while Bock and 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, including finite threshold passage, periodic orbits, and trajectories whose behavior cannot be inferred from frozen stability alone.

B The general Adam iteration and the frozen stability threshold

For a twice continuously differentiable objective $f : \mathbb{R}^d \to \mathbb{R}$, uncorrected Adam maintains a momentum estimate m_t and a second-moment estimate v_t and iterates

$$\begin{aligned} m_{t+1} &= \beta_1 m_t + (1-\beta_1)\nabla f(x_t), \\ v_{t+1} &= \beta_2 v_t + (1-\beta_2)\big(\nabla f(x_t) \odot \nabla f(x_t)\big), \\ x_{t+1} &= x_t - \eta\big(\mathrm{diag}(\sqrt{v_{t+1}}) + \varepsilon I\big)^{-1} m_{t+1}, \end{aligned} \tag{16}$$

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 of v_t are taken coordinatewise. The one-dimensional iteration (2) of the main text is the special case $f(x) = \frac{1}{2}x^2$.

A useful EoS perspective is to freeze the second-moment variable and study the resulting operator on position and momentum:

$$\mathcal{A}_v : \begin{pmatrix} x \\ m \end{pmatrix} \longmapsto \begin{pmatrix} x - \eta\big(\mathrm{diag}(\sqrt{v}) + \varepsilon I\big)^{-1}\big(\beta_1 m + (1-\beta_1)\nabla f(x)\big) \\ \beta_1 m + (1-\beta_1)\nabla f(x) \end{pmatrix}. \tag{17}$$

The following preconditioned Hessian and its largest eigenvalue identify the stability boundary of the frozen linearization.

Definition B.1 (Preconditioned sharpness)**.** Define

$$\widetilde{H}_t := \big(\mathrm{diag}(\sqrt{v_t}) + \varepsilon I\big)^{-1}\nabla^2 f(x_t), \qquad \mathsf{S}_t := \lambda_{\max}(\widetilde{H}_t). \tag{18}$$

Lemma B.2 (Frozen stability)**.** *Suppose that $\nabla^2 f(x_t) \succ 0$. The linearization of $\mathcal{A}_{v_t}$ is Schur stable if and only if*

$$\mathsf{S}_t < \mathsf{S}^\star, \qquad \mathsf{S}^\star := \frac{2(1+\beta_1)}{\eta(1-\beta_1)}. \tag{19}$$

Proof. Recall that when $\nabla^2 f(x_t) \succ 0$, the matrix $\widetilde{H}_t$ of (18) is similar to a symmetric positive-definite matrix, so its eigenvalues are real and positive. Fix t and write $D = \mathrm{diag}(\sqrt{v_t}) + \varepsilon I$ and $H = \nabla^2 f(x_t)$. In the coordinates $y = D^{1/2}\Delta x$ and $p = D^{-1/2}\Delta m$, the linearization of $\mathcal{A}_{v_t}$ is

$$p^+ = \beta_1 p + (1-\beta_1)D^{-1/2}HD^{-1/2}y, \qquad y^+ = y - \eta p^+. \tag{20}$$

The middle matrix is symmetric positive definite and has the same eigenvalues as $D^{-1}H = \widetilde{H}_t$. Along an eigenvector with eigenvalue $\lambda > 0$, the characteristic polynomial is

$$z^2 - \big(1 + \beta_1 - \eta(1-\beta_1)\lambda\big)z + \beta_1. \tag{21}$$

The quadratic Jury criterion gives Schur stability exactly when $0 < \eta(1-\beta_1)\lambda < 2(1+\beta_1)$. This holds for every mode if and only if $\lambda_{\max}(\widetilde{H}_t) < 2(1+\beta_1)/(\eta(1-\beta_1))$, as claimed. ∎

C Notation retained for Section 2.2

This section only records the quantities used in the statements of Section 2.2; no proof for that section is included. For $n \in \mathbb{N}_{\geq 1}$ and $r, s \geq 0$, let

$$G_n(r,s) := \begin{cases} (r^n - s^n)/(r-s), & r \neq s, \\ nr^{n-1}, & r = s. \end{cases} \tag{22}$$

Also set

$$\begin{aligned} \sigma_{\widehat{w}}(T) &:= \inf\{n \in \mathbb{N}_0 : w_{T+n} \geq \widehat{w}\}, \\ A_\delta &:= (1+\delta)^2, \\ \tau_\delta(T) &:= \inf\{n \in \mathbb{N}_{\geq 1} : w_{T+n} < 2+\delta\}, \\ L_n(T,a) &:= \beta_2^n v_T + (1-\beta_2)a^2 G_n(A_\delta, \beta_2), \end{aligned} \tag{23}$$

with $\inf \varnothing = +\infty$.

D Proofs for the $\beta_1 = 0$ case

Here we prove the results announced in Section 2.2: for the $\beta_1 = 0$ dynamics, we bound the precise number of steps needed to approach the threshold from either side.

Subcritical stage If $w_T \geq \widehat{w}$, then by definition $\sigma_{\widehat{w}}(T) = 0$. We therefore consider the nontrivial case $w_T < \widehat{w}$.

Theorem D.1 (Finite passage to a strict subcritical target)**.** *Assume $\beta_1 = 0$, $w_{\max} > 2$, and*

$$w_T < \widehat{w} < 2.$$

Define

$$M_T := \max\{v_T, x_T^2\}, \qquad \underline{w}_T := \frac{\eta}{\sqrt{M_T} + \varepsilon}, \tag{24}$$

$$\rho_T := \max\{|1 - \underline{w}_T|, |1 - \widehat{w}|\}, \qquad r_T := \rho_T^2, \tag{25}$$

and

$$U_n^{(T)} := \beta_2^n v_T + (1-\beta_2)x_T^2 G_n(r_T, \beta_2), \qquad n \in \mathbb{N}_{\geq 1}. \tag{26}$$

Then

$$N_T^{\text{imp}} := \min\left\{n \in \mathbb{N}_{\geq 1} : U_n^{(T)} \leq \left(\frac{\eta}{\widehat{w}} - \varepsilon\right)^2\right\} < \infty, \tag{27}$$

and

$$1 \leq \sigma_{\widehat{w}}(T) \leq N_T^{\text{imp}}. \tag{28}$$

Moreover, introduce

$$\begin{aligned} C_{\beta,T} &:= v_T + \frac{1-\beta_2}{\beta_2 - r_T} x_T^2 && \text{if } \beta_2 > r_T, \\ C_{r,T} &:= v_T + \frac{1-\beta_2}{r_T - \beta_2} x_T^2 && \text{if } \beta_2 < r_T, \\ C_{=,T} &:= v_T + \frac{1-r_T}{(1-\sqrt{r_T})^2} x_T^2 && \text{if } \beta_2 = r_T. \end{aligned}$$

The implicit bound satisfies

$$N_T^{\mathrm{imp}} \le \overline{N}_T := \begin{cases} \left\lceil \dfrac{\log\left(\dfrac{C_{\beta,T}}{(\eta/\widehat{w}-\varepsilon)^2}\right)}{\log(1/\beta_2)} \right\rceil, & \beta_2 > r_T, \\[2ex] \left\lceil \dfrac{\log\left(\dfrac{C_{r,T}}{(\eta/\widehat{w}-\varepsilon)^2}\right)}{\log(1/r_T)} \right\rceil, & \beta_2 < r_T, \\[2ex] 1 + \left\lceil \dfrac{\log\left(\dfrac{C_{=,T}}{(\eta/\widehat{w}-\varepsilon)^2}\right)}{\log(1/\sqrt{r_T})} \right\rceil, & \beta_2 = r_T. \end{cases} \tag{29}$$

Proof. Because $\beta_1 = 0$, we have

$$w_t = \frac{\eta}{\sqrt{v_t}+\varepsilon}, \qquad w_{\max} = \frac{\eta}{\varepsilon}.$$

Since $\widehat{w} < 2 < w_{\max}$,

$$\frac{\eta}{\widehat{w}} - \varepsilon > 0.$$

The assumption $w_T < \widehat{w}$ therefore gives

$$v_T > \left(\frac{\eta}{\widehat{w}} - \varepsilon\right)^2.$$

In particular, $M_T > 0$, and since $M_T \ge v_T$,

$$0 < \underline{w}_T \le w_T < \widehat{w} < 2.$$

Hence

$$0 < \rho_T < 1, \qquad 0 < r_T < 1.$$

Also, since $w_T < \widehat{w}$, we have $\sigma_{\widehat{w}}(T) \ge 1$.

Fix $n \in \mathbb{N}_{\ge 1}$. If $\sigma_{\widehat{w}}(T) < n$, the target has already been reached before time $T+n$, so no further estimate is needed. Suppose instead that

$$\sigma_{\widehat{w}}(T) \ge n.$$

Then

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

We claim that throughout this stopped trajectory,

$$v_{T+j} \le M_T, \qquad |x_{T+j}| \le \rho_T^j \, |x_T| \,, \qquad j = 0, \dots, n-1. \tag{30}$$

We prove the two estimates simultaneously by induction. They are immediate at $j = 0$. Suppose they hold up to some $j-1$ with $1 \le j \le n-1$. Since $w_{T+j} < \widehat{w} < 2$,

$$|x_{T+j}| = |1 - w_{T+j}| \, |x_{T+j-1}| < |x_{T+j-1}| \,,$$

and hence $|x_{T+j-1}| \le |x_T|$. Using

$$v_{T+j} = \beta_2 v_{T+j-1} + (1-\beta_2) x_{T+j-1}^2,$$

together with $v_{T+j-1} \le M_T$ and $x_{T+j-1}^2 \le x_T^2 \le M_T$, gives

$$v_{T+j} \le M_T.$$

Consequently,

$$\underline{w}_T = \frac{\eta}{\sqrt{M_T} + \varepsilon} \le w_{T+j} < \widehat{w}.$$

Since $w \mapsto |1-w|$ is convex, its maximum on $[\underline{w}_T, \widehat{w}]$ is attained at one of the endpoints. Therefore

$$|1 - w_{T+j}| \le \max\{|1 - \underline{w}_T|, |1 - \widehat{w}|\} = \rho_T,$$

and thus

$$|x_{T+j}| \le \rho_T |x_{T+j-1}| \le \rho_T^j |x_T|.$$

This proves (30).

Unrolling the second-moment recursion over n steps and using (30) gives

$$\begin{aligned}
v_{T+n} &= \beta_2^n v_T + (1-\beta_2) \sum_{j=0}^{n-1} \beta_2^{n-1-j} x_{T+j}^2 \\
&\le \beta_2^n v_T + (1-\beta_2) x_T^2 \sum_{j=0}^{n-1} \beta_2^{n-1-j} r_T^j \\
&= U_n^{(T)}.
\end{aligned}$$

Since $\beta_2 \in [0,1)$ and $r_T \in (0,1)$, $U_n^{(T)} \to 0$. Indeed, when $\beta_2 \neq r_T$, this follows directly from the closed form of $G_n(r_T, \beta_2)$, while for $\beta_2 = r_T$, $G_n(r_T, r_T) = n r_T^{n-1} \to 0$. Because

$$\left(\frac{\eta}{\widehat{w}} - \varepsilon\right)^2 > 0,$$

there exists a finite n for which

$$U_n^{(T)} \le \left(\frac{\eta}{\widehat{w}} - \varepsilon\right)^2.$$

Hence $N_T^{\text{imp}} < \infty$.

Now take $n = N_T^{\text{imp}}$. If $\sigma_{\widehat{w}}(T) < n$, then certainly $\sigma_{\widehat{w}}(T) \le N_T^{\text{imp}}$. Otherwise the stopped estimate applies and yields

$$v_{T+n} \le U_n^{(T)} \le \left(\frac{\eta}{\widehat{w}} - \varepsilon\right)^2.$$

Since $\eta/\widehat{w} - \varepsilon > 0$, this implies

$$\sqrt{v_{T+n}} + \varepsilon \le \frac{\eta}{\widehat{w}},$$

and therefore

$$w_{T+n} = \frac{\eta}{\sqrt{v_{T+n}} + \varepsilon} \ge \widehat{w}.$$

Thus

$$1 \le \sigma_{\widehat{w}}(T) \le N_T^{\text{imp}}.$$

It remains to prove the explicit bounds.

If $\beta_2 > r_T$, then

$$G_n(r_T, \beta_2) = \frac{\beta_2^n - r_T^n}{\beta_2 - r_T} \le \frac{\beta_2^n}{\beta_2 - r_T},$$

so

$$U_n^{(T)} \le C_{\beta,T} \beta_2^n.$$

Hence

$$n \geq \frac{\log\left(\dfrac{C_{\beta,T}}{(\eta/\widehat{w}-\varepsilon)^2}\right)}{\log(1/\beta_2)}$$

is sufficient to ensure

$$U_n^{(T)} \leq \left(\frac{\eta}{\widehat{w}} - \varepsilon\right)^2,$$

which gives the first case of (29).

If $\beta_2 < r_T$, then

$$G_n(r_T, \beta_2) = \frac{r_T^n - \beta_2^n}{r_T - \beta_2} \leq \frac{r_T^n}{r_T - \beta_2}, \qquad \beta_2^n \leq r_T^n.$$

Consequently,

$$U_n^{(T)} \leq C_{r,T} r_T^n.$$

Thus

$$n \geq \frac{\log\left(\dfrac{C_{r,T}}{(\eta/\widehat{w}-\varepsilon)^2}\right)}{\log(1/r_T)}$$

is sufficient, giving the second case.

Finally, suppose $\beta_2 = r_T$. Then

$$G_n(r_T, r_T) = n r_T^{n-1}.$$

For every $n \geq 1$,

$$\begin{aligned} n r_T^{n-1} &= \left(\sqrt{r_T}\right)^{n-1} \left(n\left(\sqrt{r_T}\right)^{n-1}\right) \\ &\leq \frac{\left(\sqrt{r_T}\right)^{n-1}}{(1-\sqrt{r_T})^2}, \end{aligned}$$

because

$$n\left(\sqrt{r_T}\right)^{n-1} \leq \sum_{k=1}^{\infty} k\left(\sqrt{r_T}\right)^{k-1} = \frac{1}{(1-\sqrt{r_T})^2}.$$

Moreover,

$$r_T^n = \left(\sqrt{r_T}\right)^{2n} \leq \left(\sqrt{r_T}\right)^{n-1}.$$

Therefore

$$U_n^{(T)} \leq C_{=,T}\left(\sqrt{r_T}\right)^{n-1}.$$

Hence it is sufficient that

$$n - 1 \geq \frac{\log\left(\dfrac{C_{=,T}}{(\eta/\widehat{w}-\varepsilon)^2}\right)}{\log(1/\sqrt{r_T})},$$

which gives the final case of (29). □

Supercritical stage We next show that a nonzero trajectory cannot remain indefinitely in a uniformly supercritical band. The argument is the reverse of the subcritical mechanism: as long as w_t remains above $2 + \delta$, the position expands geometrically, which forces the second moment upward and eventually makes such a large value of w_t impossible.

Theorem D.2 (Finite supercritical exit for $\beta_1 = 0$)**.** *Let $T \in \mathbb{N}_0$ and let $\delta > 0$ satisfy*

$$2 + \delta \le w_{\max}.$$

Assume

$$\beta_1 = 0, \qquad w_T \ge 2 + \delta, \qquad x_T \ne 0.$$

Define

$$N^{\mathrm{imp}}_{\delta,T} := \min\left\{ n \in \mathbb{N}_{\ge 1} : L_n(T, |x_T|) > \left(\frac{\eta}{2+\delta} - \varepsilon\right)^2 \right\}. \tag{31}$$

Then this minimum is finite and

$$\tau_\delta(T) \le N^{\mathrm{imp}}_{\delta,T} \le N^{\mathrm{exp}}_{\delta,T} < \infty, \tag{32}$$

where

$$N^{\mathrm{exp}}_{\delta,T} := 1 + \left\lceil \frac{\log\left(1 + \frac{(A_\delta - \beta_2)\left(\frac{\eta}{2+\delta} - \varepsilon\right)^2}{(1-\beta_2)x_T^2}\right)}{\log A_\delta} \right\rceil. \tag{33}$$

Proof. Because $\beta_1 = 0$,

$$w_t = \frac{\eta}{\sqrt{v_t} + \varepsilon}.$$

The assumption

$$2 + \delta \le w_{\max} = \frac{\eta}{\varepsilon}$$

implies

$$\frac{\eta}{2+\delta} - \varepsilon \ge 0.$$

Consequently,

$$w_t \ge 2 + \delta \iff v_t \le \left(\frac{\eta}{2+\delta} - \varepsilon\right)^2. \tag{34}$$

Fix $n \in \mathbb{N}_{\ge 1}$. If $\tau_\delta(T) \le n$, then the trajectory has already left the prescribed supercritical band by time $T + n$, so there is nothing to prove. Suppose instead that

$$\tau_\delta(T) > n.$$

By the definition of $\tau_\delta(T)$,

$$w_{T+k} \ge 2 + \delta, \qquad k = 1, \dots, n.$$

Using the zero-momentum position recursion

$$x_{t+1} = (1 - w_{t+1})x_t,$$

we obtain, for every $k = 1, \dots, n$,

$$\begin{aligned} |x_{T+k}| &= |1 - w_{T+k}|\, |x_{T+k-1}| \\ &= (w_{T+k} - 1)\, |x_{T+k-1}| \\ &\ge (1+\delta)\, |x_{T+k-1}|. \end{aligned}$$

Iterating gives

$$|x_{T+j}| \ge (1+\delta)^j\, |x_T|, \qquad j = 0, \dots, n,$$

and hence

$$x_{T+j}^2 \geq A_\delta^j x_T^2, \qquad j = 0, \ldots, n, \tag{35}$$

where $A_\delta = (1+\delta)^2$.

Unrolling the second-moment recursion over n steps yields

$$\begin{aligned}
v_{T+n} &= \beta_2^n v_T + (1-\beta_2) \sum_{j=0}^{n-1} \beta_2^{n-1-j} x_{T+j}^2 \\
&\geq \beta_2^n v_T + (1-\beta_2) x_T^2 \sum_{j=0}^{n-1} \beta_2^{n-1-j} A_\delta^j \\
&= \beta_2^n v_T + (1-\beta_2) x_T^2 G_n(A_\delta, \beta_2) \\
&= L_n(T, |x_T|).
\end{aligned}$$

On the other hand, $\tau_\delta(T) > n$ implies $w_{T+n} \geq 2+\delta$. Therefore, by (34),

$$v_{T+n} \leq \left(\frac{\eta}{2+\delta} - \varepsilon \right)^2.$$

Hence survival in the supercritical band through time $T+n$ necessarily implies

$$L_n(T, |x_T|) \leq \left(\frac{\eta}{2+\delta} - \varepsilon \right)^2. \tag{36}$$

Thus, whenever

$$L_n(T, |x_T|) > \left(\frac{\eta}{2+\delta} - \varepsilon \right)^2,$$

the trajectory must have exited the band no later than time $T+n$.

It remains to show that such an n exists. Since

$$A_\delta = (1+\delta)^2 > 1 > \beta_2$$

and $x_T \neq 0$, we have

$$\begin{aligned}
L_n(T, |x_T|) &= \beta_2^n v_T + (1-\beta_2) x_T^2 G_n(A_\delta, \beta_2) \\
&\geq \frac{(1-\beta_2) x_T^2}{A_\delta - \beta_2} \left(A_\delta^n - \beta_2^n \right).
\end{aligned}$$

Because $A_\delta > 1$ and $\beta_2 < 1$, $L_n(T, |x_T|) \to \infty$. Therefore,

$$N_{\delta,T}^{\mathrm{imp}} < \infty.$$

The survival implication above gives

$$\tau_\delta(T) \leq N_{\delta,T}^{\mathrm{imp}}.$$

It remains to derive the explicit bound. Since $0 \leq \beta_2 < 1$, one has $\beta_2^n \leq 1$, and hence

$$\begin{aligned}
G_n(A_\delta, \beta_2) &= \frac{A_\delta^n - \beta_2^n}{A_\delta - \beta_2} \\
&\geq \frac{A_\delta^n - 1}{A_\delta - \beta_2}.
\end{aligned}$$

Thus

$$L_n(T, |x_T|) \geq \frac{(1-\beta_2) x_T^2}{A_\delta - \beta_2} \left(A_\delta^n - 1 \right). \tag{37}$$

Therefore it is sufficient that

$$A_\delta^n > 1 + \frac{(A_\delta - \beta_2)\left(\dfrac{\eta}{2+\delta} - \varepsilon\right)^2}{(1-\beta_2)x_T^2}.$$

By the definition of $N_{\delta,T}^{\exp}$,

$$A_\delta^{N_{\delta,T}^{\exp}} > 1 + \frac{(A_\delta - \beta_2)\left(\dfrac{\eta}{2+\delta} - \varepsilon\right)^2}{(1-\beta_2)x_T^2}.$$

Combining this with (37) yields

$$L_{N_{\delta,T}^{\exp}}(T, |x_T|) > \left(\frac{\eta}{2+\delta} - \varepsilon\right)^2.$$

Hence

$$N_{\delta,T}^{\mathrm{imp}} \le N_{\delta,T}^{\exp},$$

which completes the proof. □

E Proofs for general parameters in the subcritical stage

E.1 Proof of Proposition 3.1

Proof. Let

$$\theta(y) := \frac{y}{2-y}. \tag{38}$$

Recall that in the present section $0 < \beta_1 < 1$.

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

$$\sqrt{v_t} = c\eta\left(\frac{1}{w_t} - \frac{1}{w_{\max}}\right),$$

its exact normalized form is

$$\left(\frac{1}{w_{t+1}} - \frac{1}{w_{\max}}\right)^2 = \beta_2\left(\frac{1}{w_t} - \frac{1}{w_{\max}}\right)^2 + (1-\beta_2)\left(\frac{x_t}{c\eta}\right)^2. \tag{39}$$

Taking nonnegative square roots gives

$$\frac{1}{w_{t+1}} - \frac{1}{w_{\max}} \ge \sqrt{\beta_2}\left(\frac{1}{w_t} - \frac{1}{w_{\max}}\right),$$

and hence

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

or equivalently,

$$w_t \ge \frac{\sqrt{\beta_2}\, w_{t+1}}{1 - (1-\sqrt{\beta_2})w_{t+1}/w_{\max}}. \tag{40}$$

Equality holds exactly when $x_t = 0$.

We next solve the one-step quadratic inequality. For $y, Y > 0$, put

$$P_y := \mathrm{diag}(1, y), \qquad D(y, Y; w_{t+1}) := P_y - A(w_{t+1})^\top P_Y A(w_{t+1}). \tag{41}$$

Direct multiplication gives the exact identities

$$D_{11}(y, Y; w_{t+1}) = (2 - w_{t+1})\big(w_{t+1} - (2 - w_{t+1})Y\big),$$

and

$$\det D(y, Y; w_{t+1}) = \big(w_{t+1} - (2 - w_{t+1})Y\big)\big((2 - w_{t+1})y - \beta_1^2 w_{t+1}\big). \tag{42}$$

Because $0 < w_{t+1} < 2$, Sylvester's criterion shows that

$$A(w_{t+1})^\top P_Y A(w_{t+1}) \prec P_y$$

if and only if

$$Y < \theta(w_{t+1}), \qquad y > \beta_1^2 \theta(w_{t+1}). \tag{43}$$

For the Lyapunov function in the proposition, the source and target weights are

$$y = \beta_1 \theta(w_t), \qquad Y = \beta_1 \theta(w_{t+1}).$$

The first inequality in (43) is automatic because $\beta_1 < 1$, while the second becomes

$$\theta(w_t) > \beta_1 \theta(w_{t+1}). \tag{44}$$

Since θ is strictly increasing on $(0, 2)$, (40) implies

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

Therefore, after direct simplification,

$$\frac{\theta(w_{t+1})}{\theta(w_t)} \leq \frac{2 - w_{t+1}\left(\sqrt{\beta_2} + 2(1 - \sqrt{\beta_2})/w_{\max}\right)}{\sqrt{\beta_2}(2 - w_{t+1})}. \tag{45}$$

The right-hand side is increasing in w_{t+1}, since

$$\sqrt{\beta_2} + \frac{2(1 - \sqrt{\beta_2})}{w_{\max}} < 1$$

when $w_{\max} > 2$. Thus, if

$$w_t, w_{t+1} \leq w < 2,$$

then

$$\frac{\theta(w_{t+1})}{\theta(w_t)} \leq \frac{2 - w\left(\sqrt{\beta_2} + 2(1 - \sqrt{\beta_2})/w_{\max}\right)}{\sqrt{\beta_2}(2 - w)}. \tag{46}$$

The product of the last expression with β_1 is strictly below one exactly when

$$w < \frac{2(\sqrt{\beta_2} - \beta_1)}{\sqrt{\beta_2}(1 - \beta_1) - 2\beta_1(1 - \sqrt{\beta_2})/w_{\max}} =: W. \tag{47}$$

Since $\sqrt{\beta_2} > \beta_1$ and $w_{\max} > 2$,

$$\begin{aligned} &\sqrt{\beta_2}(1 - \beta_1) - \frac{2\beta_1(1 - \sqrt{\beta_2})}{w_{\max}} - (\sqrt{\beta_2} - \beta_1) \\ &= \beta_1(1 - \sqrt{\beta_2})\left(1 - \frac{2}{w_{\max}}\right) > 0. \end{aligned}$$

Hence the denominator in (47) is positive and strictly larger than $\sqrt{\beta_2} - \beta_1$, and therefore

$$0 < W < 2.$$

Equations (43)–(47) prove the strict decrease for every nonzero z_t.

It remains to make the contraction factor explicit. Set

$$P(w_t) := \mathrm{diag}\big(1, \beta_1\theta(w_t)\big),$$

and

$$D := P(w_t) - A(w_{t+1})^\top P(w_{t+1})A(w_{t+1}). \tag{48}$$

Specializing (42), or equivalently taking a Schur complement, gives

$$D_{11} = (1-\beta_1)w_{t+1}(2-w_{t+1}),$$

and

$$D_{22} - \frac{D_{12}^2}{D_{11}} = \beta_1\left(\theta(w_t) - \beta_1\theta(w_{t+1})\right). \tag{49}$$

If

$$a \le w_t, w_{t+1} \le w < W,$$

then

$$D_{11} \ge (1-\beta_1)\min\{a(2-a),\, w(2-w)\},$$

and, using (46),

$$D_{22} - \frac{D_{12}^2}{D_{11}} \ge \frac{\beta_1 a}{2-a}\left[1 - \beta_1\frac{2 - w\left(\sqrt{\beta_2} + 2(1-\sqrt{\beta_2})/w_{\max}\right)}{\sqrt{\beta_2}(2-w)}\right]. \tag{50}$$

The bracket is positive by (47).

For convenience, define

$$K(w) := \frac{2 - w\left(\sqrt{\beta_2} + 2(1-\sqrt{\beta_2})/w_{\max}\right)}{\sqrt{\beta_2}(2-w)},$$

and

$$\delta(a, w) := \frac{(1-\beta_1)\min\{a(2-a),\, w(2-w)\}\dfrac{\beta_1 a}{2-a}\big(1-\beta_1 K(w)\big)}{2\left(1 + \frac{\beta_1 w}{2-w}\right)\max\left\{1, \frac{\beta_1 w}{2-w}\right\}}. \tag{51}$$

All factors in this expression are strictly positive whenever

$$0 < a \le w < W,$$

so

$$\delta(a, w) > 0.$$

Moreover,

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

because $P(w_{t+1}) \succ 0$ and

$$\det A(w_{t+1}) = \beta_1 > 0.$$

Thus

$$0 \prec D \prec P(w_t).$$

Since $w_t \le w$ and θ is increasing,

$$\operatorname{tr} D < \operatorname{tr} P(w_t) \le 1 + \frac{\beta_1 w}{2-w}. \tag{52}$$

Also, by (49)–(50),

$$\begin{aligned}\det D &= D_{11}\left(D_{22} - \frac{D_{12}^2}{D_{11}}\right) \\ &\geq (1-\beta_1)\min\{a(2-a),\, w(2-w)\}\frac{\beta_1 a}{2-a}\big(1-\beta_1 K(w)\big).\end{aligned} \tag{53}$$

For a positive-definite 2×2 matrix,

$$\lambda_{\min}(D) \geq \frac{\det D}{\operatorname{tr} D}.$$

Hence, by (51)–(53),

$$\lambda_{\min}(D) \geq 2\delta(a,w)\max\left\{1, \frac{\beta_1 w}{2-w}\right\}.$$

On the other hand,

$$P(w_t) \preceq \max\left\{1, \frac{\beta_1 w}{2-w}\right\} I.$$

Consequently,

$$D \succeq 2\delta(a,w)\max\left\{1, \frac{\beta_1 w}{2-w}\right\} I \succeq 2\delta(a,w)P(w_t) \succeq \delta(a,w)P(w_t). \tag{54}$$

Since $D \prec P(w_t)$, the inequality above implies

$$2\delta(a,w) < 1,$$

and hence, in particular,

$$0 < \delta(a,w) < 1.$$

Finally,

$$\begin{aligned}\Psi_{t+1} &= z_t^\top A(w_{t+1})^\top P(w_{t+1})A(w_{t+1})z_t \\ &= \Psi_t - z_t^\top D z_t \\ &\leq \big(1-\delta(a,w)\big)\Psi_t.\end{aligned} \tag{55}$$

This proves the quantitative contraction estimate.

Finally,

$$\sqrt{\beta_2} = 1 - \frac{1-\beta_2}{2} + O\big((1-\beta_2)^2\big),$$

and substituting this expansion into the definition of W gives (56). ☐

Remark E.1 (Expansion of the cutoff). For fixed β_1 and $w_{\max}$, the cutoff of Proposition 3.1 has the expansion

$$\overline{W} = 2 - \frac{\beta_1(1-2/w_{\max})}{1-\beta_1}(1-\beta_2) + O\big((1-\beta_2)^2\big) \qquad (\beta_2 \to 1), \tag{56}$$

derived in the final step of the proof above.

E.2 Proof of Corollary 3.2

Proof. Suppose, to the contrary, that $w_{T+n} \leq w$ for every $n \geq 0$, and set

$$M := \max\{v_T, \Psi_T\}, \qquad a := \frac{c\eta}{\sqrt{M}+\varepsilon}. \tag{57}$$

Then $0 < a \leq w_T \leq w$. Proposition 3.1 and the fact $x_t^2 \leq \Psi_t$ show inductively that

$$v_{T+n} \leq M, \qquad \Psi_{T+n} \leq \big(1-\delta(a,w)\big)^n \Psi_T. \tag{58}$$

Unrolling the second-moment recursion and using the second inequality gives

$$v_{T+n} \le \beta_2^n v_T + (1-\beta_2)\Psi_T \sum_{j=0}^{n-1} \beta_2^{n-1-j} \big(1-\delta(a,w)\big)^j. \tag{59}$$

The right-hand side tends to zero. Hence $w_{T+n} \to w_{\max} > 2 > w$, a contradiction. □

The following is the quantitative version referenced after Corollary 3.2.

Corollary E.2 (Quantitative passage toward the subcritical cutoff)**.** *Assume the hypotheses of Proposition 3.1 and $w_{\max} > 4$. Fix $0 < \delta < \overline{W}$ and $T \in \mathbb{N}_0$, and let*

$$\tau := \min\{n \in \mathbb{N}_0 : w_{T+n} > \overline{W} - \delta\}. \tag{60}$$

If $w_T > \overline{W} - \delta$, then $\tau = 0$. Otherwise, set

$$q := 1 - \delta\left(\frac{c\eta}{\sqrt{\max\{v_T, \Psi_T\}} + c\eta/4}, \overline{W} - \delta\right). \tag{61}$$

Then

$$\tau \le 1 + \left\lceil \frac{2}{1-\max\{\beta_2, q\}} \log\left[\frac{16}{c^2\eta^2}\left(v_T + \frac{2(1-\beta_2)}{1-\max\{\beta_2,q\}}\Psi_T\right)\right]\right\rceil. \tag{62}$$

Proof. Only the case $w_T \le \overline{W} - \delta$ needs consideration. Put $\lambda = \max\{\beta_2, q\}$ and $\gamma = (1+\lambda)/2$. Since $w_{\max} > 4$, one has $\varepsilon < c\eta/4$. Until time τ, the argument in (58) therefore applies with the lower endpoint used in (61), and gives

$$v_{T+n} \le \beta_2^n v_T + (1-\beta_2)\Psi_T \sum_{j=0}^{n-1} \beta_2^{n-1-j} q^j. \tag{63}$$

The sum is at most $n\lambda^{n-1}$ and hence at most $2\gamma^n/(1-\lambda)$. Thus the right-hand side is bounded by

$$\gamma^n \left(v_T + \frac{2(1-\beta_2)}{1-\lambda}\Psi_T\right). \tag{64}$$

Moreover, if $v_{T+n} \le (c\eta/4)^2$ then, since $\varepsilon < c\eta/4$ and $\overline{W} - \delta < 2$, one has $w_{T+n} = c\eta/(\sqrt{v_{T+n}} + \varepsilon) > 2 > \overline{W} - \delta$. Finally, $\log(1/\gamma) \ge 1 - \gamma = (1-\lambda)/2$. Taking n equal to the ceiling in (62) makes (64) no larger than $c^2\eta^2/16$, so the target has been crossed by that time. The extra 1 in (62) is harmless. □

E.3 Proof of Proposition 3.3

Proof. For $s \in (0,1)$ define, only within this proof,

$$\begin{aligned} R(s) &:= \left[\frac{(1-\beta_1 s)(1+s)}{(1-s)(\beta_1+s)}\right]^2, \\ G(s) &:= \frac{1 - R(s)s^2}{R(s) - s^2}. \end{aligned} \tag{65}$$

The ratio inside the square is larger than one. Let $s_0 \in (0,1)$ be the first solution of

$$s\frac{(1-\beta_1 s)(1+s)}{(1-s)(\beta_1+s)} = 1. \tag{66}$$

Such a solution exists because the left-hand side is zero at $s = 0$ and tends to $+\infty$ as $s \to 1^-$. On $[0, s_0]$ the denominator in G is positive, and

$$G(0) = \beta_1^2, \qquad G(s_0) = 0. \tag{67}$$

Thus every $\beta_2 \in [0, \beta_1^2)$ equals $G(s)$ for some $s \in (0, s_0]$. When $\beta_2 = \beta_1^2$, direct differentiation gives

$$G'(0) = 2\beta_1(1-\beta_1)^2 > 0;$$

since $G(s_0) = 0$, there is again a solution with $s \in (0, s_0)$. We henceforth fix such a positive s.

Set

$$\begin{aligned} A &:= \sqrt{\frac{1+\beta_2 s^2}{1+\beta_2}}, \quad B := \sqrt{\frac{s^2+\beta_2}{1+\beta_2}}, \\ f &:= \frac{1-\beta_1 s}{1-s}, \qquad g := \frac{\beta_1+s}{1+s}, \qquad k := \frac{1-\beta_1}{1+\beta_1^2}. \end{aligned} \tag{68}$$

The equation $\beta_2 = G(s)$ is equivalent to

$$\frac{f}{A} = \frac{g}{B}. \tag{69}$$

For the prescribed learning rate $\eta > 0$, define

$$a := \eta k \frac{f}{A} = \eta k \frac{g}{B} > 0 \tag{70}$$

and initialize

$$x_0 = a, \qquad m_0 = -ka(\beta_1+s), \qquad v_0 = a^2B^2. \tag{71}$$

The momentum and second-moment recursions give

t mod 4	0	1	2	3
x_t	a	sa	$-a$	$-sa$
m_t	$-ka(\beta_1+s)$	$ka(1-\beta_1 s)$	$ka(\beta_1+s)$	$-ka(1-\beta_1 s)$
v_t	a^2B^2	a^2A^2	a^2B^2	a^2A^2.

(72)

Indeed, $aA = \eta kf$ and $aB = \eta kg$ verify the first two position updates when $\varepsilon = 0$; the remaining two follow by half-turn symmetry. This also shows directly why the construction works for every $\eta > 0$: changing η merely rescales a, m_t, and $\sqrt{v_t}$ by the same factor.

The two values of w_t are

$$\frac{1+\beta_1^2}{1+\beta_1}\frac{1-s}{1-\beta_1 s}, \qquad \frac{1+\beta_1^2}{1+\beta_1}\frac{1+s}{\beta_1+s}. \tag{73}$$

The first is always below 2, while the second is below 2 exactly when

$$s > \frac{1-2\beta_1-\beta_1^2}{1+2\beta_1-\beta_1^2}. \tag{74}$$

Since $\beta_1 \geq \sqrt{2}-1$, the right-hand side is nonpositive. Therefore the orbit is strictly subcritical. The four position values are distinct, so its prime period is four. ☐

F Proofs for the supercritical results

F.1 Proof of Lemma 3.4

Proof. If x_t and h_t have the same sign, the two rows of $A(w_{t+1})$ show that both signs reverse when $w_{t+1} > 2$, and

$$|x_{t+1}| = (w_{t+1}-1)|x_t| + \beta_1 w_{t+1}|h_t| > (w_{t+1}-1)|x_t|. \tag{75}$$

If $x_t h_t < 0$, write $h_t = -rx_t$ with $r > 0$. The next state is also misaligned exactly when

$$\frac{w_{t+1}-2}{\beta_1(w_{t+1}-1)} < r < \frac{w_{t+1}-1}{\beta_1 w_{t+1}}. \tag{76}$$

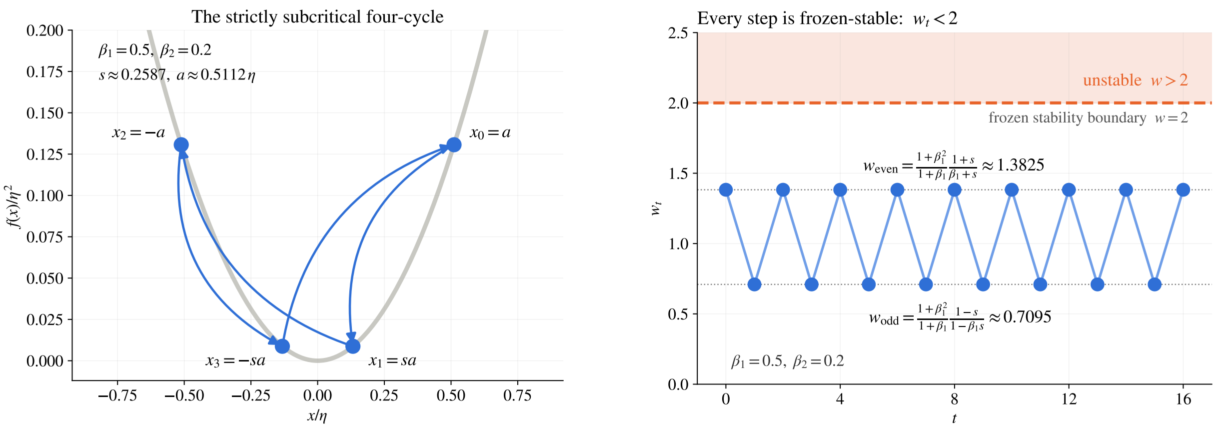

Figure 3: An exact four-cycle of Proposition 3.3 at $(\beta_1, \beta_2) = (\frac{1}{2}, \frac{1}{5})$, $\varepsilon = 0$, initialized at the point (71) with $s \approx 0.2587$ and $a \approx 0.5112\,\eta$. Left: the orbit on $f(x) = \frac{1}{2}x^2$. Right: w_t alternates between ≈ 1.38 and ≈ 0.71, below $w = 2$.

Within this interval,

$$\begin{aligned} |x_{t+1}| &= \big(w_{t+1} - 1 - \beta_1 w_{t+1} r\big)\,|x_t| \\ &< \left(w_{t+1} - 1 - \frac{w_{t+1}(w_{t+1}-2)}{w_{t+1}-1}\right)|x_t| = \frac{|x_t|}{w_{t+1}-1}. \end{aligned} \tag{77}$$

This proves both assertions. □

F.2 Proof of Corollary 3.5

Proof. While $w_{T+j} \geq 2 + \delta$, Lemma 3.4(i) preserves alignment. The following stopped estimate is obtained by substituting $w_{t+1} = c\eta/(\sqrt{v_{t+1}} + \varepsilon)$ in (75) and using $v_{t+1} = \beta_2 v_t + (1-\beta_2)x_t^2$: for every $1 \leq n < \tau$,

$$(1+\delta)^{2n} - 1 < \frac{\delta(c\eta)^2}{(1-\beta_2)x_T^2}. \tag{78}$$

This is a direct induction on n; after clearing the positive denominators, the induction step is precisely the sum of the nonnegative momentum term in (75) and the nonnegative term $\beta_2 v_t$ in the second-moment update.

Since $2\log(1+\delta) \geq \delta$ for $0 < \delta \leq 1$, (78) implies

$$n < \delta^{-1}\log\left(1 + \frac{\delta(c\eta)^2}{(1-\beta_2)x_T^2}\right) \qquad (1 \leq n < \tau). \tag{79}$$

Taking the largest surviving integer gives (13). If an aligned trajectory remained supercritical and had $\liminf_t w_t > 2$, it would eventually remain in $w_t \geq 2 + \delta$ for some $\delta \in (0, 1]$, contradicting the bound just proved. Hence its lower limit is 2. □

F.3 Proof of Proposition 3.6

Proof. Fix the prescribed $v_0 > 0$ and choose any

$$0 < |x_0| \leq \sqrt{v_0}. \tag{80}$$

We will select the momentum by a one-dimensional shooting argument. For a temporary slope $r_0 > 0$, let

$$h_0 = -r_0 x_0, \qquad m_0 = c(h_0 - x_0) = -c(1+r_0)x_0. \tag{81}$$

Let $\sigma = \operatorname{sign}(x_0)$ and remove the alternating sign by setting

$$X_t := \sigma(-1)^t x_t, \qquad H_t := \sigma(-1)^t h_t. \tag{82}$$

Thus $X_0 = |x_0| > 0$ and $H_0 = -r_0 |x_0| < 0$. From (6),

$$\begin{pmatrix} X_{t+1} \\ H_{t+1} \end{pmatrix} = \begin{pmatrix} w_{t+1} - 1 & \beta_1 w_{t+1} \\ w_{t+1} - 2 & \beta_1(w_{t+1} - 1) \end{pmatrix} \begin{pmatrix} X_t \\ H_t \end{pmatrix}. \tag{83}$$

We first verify that every finite misaligned orbit stays in the prescribed supercritical box. If misalignment survives from time t to time $t+1$, Lemma 3.4(ii) gives

$$|x_{t+1}| < \frac{|x_t|}{w_{t+1} - 1} < |x_t|. \tag{84}$$

Starting from (80), induction and the convex-combination identity $v_{t+1} = \beta_2 v_t + (1-\beta_2)x_t^2$ therefore give

$$x_t^2 \le v_0, \qquad v_t \le v_0 \tag{85}$$

at every surviving time. Hence

$$w_t \ge \frac{c\eta}{\sqrt{v_0} + \varepsilon} = w_0 > 2. \tag{86}$$

This closes the small-box argument: the supercritical hypothesis needed in Lemma 3.4(ii) is automatically preserved for as long as the trajectory remains misaligned.

We now construct an orbit for which misalignment never ends. On a finite surviving orbit define

$$r_t := -\frac{H_t}{X_t} = -\frac{h_t}{x_t} > 0. \tag{87}$$

Substitution of $H_t = -r_t X_t$ in (83) shows that the next state is misaligned exactly when

$$\ell(w_{t+1}) < r_t < q(w_{t+1}), \tag{88}$$

where, within this proof,

$$\ell(w) := \frac{w-2}{\beta_1(w-1)}, \qquad q(w) := \frac{w-1}{\beta_1 w}. \tag{89}$$

When (88) holds, the next ratio is

$$r_{t+1} = \Phi_{w_{t+1}}(r_t) := \frac{\beta_1(w_{t+1}-1)r_t - (w_{t+1}-2)}{w_{t+1} - 1 - \beta_1 w_{t+1} r_t}. \tag{90}$$

For each fixed $w > 2$, this map is continuous and strictly increasing on $(\ell(w), q(w))$, and

$$\lim_{r \to \ell(w)^+} \Phi_w(r) = 0, \qquad \lim_{r \to q(w)^-} \Phi_w(r) = +\infty. \tag{91}$$

We next build nested shooting intervals. Since v_1 depends on x_0 and v_0 but not on r_0, the number w_1 is fixed. Define

$$I_1 := (\ell(w_1), q(w_1)). \tag{92}$$

Every $r_0 \in I_1$ survives through time one, the map $r_0 \mapsto r_1$ is continuous, and (91) says that its range is $(0, +\infty)$.

Suppose inductively that $I_N = (a_N, b_N)$ is a nonempty open interval such that every $r_0 \in I_N$ survives through time N, the map $r_0 \mapsto r_N(r_0)$ is continuous, and

$$\lim_{r_0 \to a_N^+} r_N(r_0) = 0, \qquad \lim_{r_0 \to b_N^-} r_N(r_0) = +\infty. \tag{93}$$

The quantities $w_{N+1}(r_0)$, $\ell(w_{N+1}(r_0))$, and $q(w_{N+1}(r_0))$ are continuous on I_N. Furthermore, (86) implies the uniform bounds

$$\ell(w_{N+1}) \geq \frac{w_0 - 2}{\beta_1(w_0 - 1)} > 0, \qquad q(w_{N+1}) < \frac{1}{\beta_1}. \tag{94}$$

Thus the graph of r_N starts below the lower boundary in (88) and ends above its upper boundary. Choose a component $I_{N+1} = (a_{N+1}, b_{N+1})$ on which

$$\ell(w_{N+1}) < r_N < q(w_{N+1}) \tag{95}$$

and whose left and right endpoints meet the lower and upper boundary, respectively. Then

$$\overline{I}_{N+1} \subset I_N. \tag{96}$$

Applying (91) at the two new endpoints shows that $r_0 \mapsto r_{N+1}$ again has the endpoint behavior in (93). This completes the induction.

The nonempty compact intervals $\overline{I}_N$ are nested, so choose

$$r_0^* \in \bigcap_{N \geq 1} \overline{I}_N. \tag{97}$$

Because $\overline{I}_{N+1} \subset I_N$, the selected point actually belongs to every open survival interval I_N. The initialization (81) with $r_0 = r_0^*$ therefore remains strictly misaligned at every finite time. Since $m_t + cx_t = ch_t$, this gives $x_t(m_t + cx_t) = cx_t h_t < 0$ for every $t \geq 0$.

It remains only to record the convergence estimates. Iterating Lemma 3.4(ii) and using $w_t \geq w_0$ gives

$$|x_t| \leq |x_0| \prod_{j=1}^{t} \frac{1}{w_j - 1} \leq (w_0 - 1)^{-t} |x_0|. \tag{98}$$

Also $r_t < q(w_{t+1}) < 1/\beta_1$, and $m_t = c(h_t - x_t) = -c(1 + r_t)x_t$, so

$$|m_t| \leq c\left(1 + \frac{1}{\beta_1}\right)|x_t| \to 0. \tag{99}$$

Finally,

$$v_t = \beta_2^t v_0 + (1 - \beta_2)\sum_{j=0}^{t-1} \beta_2^{t-1-j} x_j^2 \to 0, \tag{100}$$

because x_j^2 decays geometrically. Hence $(x_t, m_t, v_t) \to (0, 0, 0)$ and $w_t \to w_{\max}$. □

F.4 Numerical illustration of Proposition 3.6

Take

$$\beta_1 = \tfrac{1}{2}, \qquad \beta_2 = 0.9, \qquad \eta = 1, \qquad \varepsilon = \tfrac{1}{12},$$

so that $c = \frac{1}{3}$ and $w_{\max} = 4$, and prescribe $v_0 = 0.0025$ and $x_0 = 0.04 \leq \sqrt{v_0}$, hence $w_0 = c\eta/(\sqrt{v_0} + \varepsilon) = \frac{5}{2} > 2$. Following the shooting argument of Appendix F.3, the momentum is initialized as $m_0 = -c(1 + r_0)x_0$ and the persistent slope $r_0^* \in \bigcap_N \overline{I}_N$ is located by bisection on r_0, running the recursion (2). With this initialization the orbit remains misaligned, $x_t(m_t + cx_t) < 0$, over the 12 steps shown in Figure 4 (and beyond): $w_t \geq w_0$ increases monotonically toward $w_{\max} = 4$, the envelope $|x_t| \leq (w_0 - 1)^{-t}|x_0|$ of (15) holds at every step, and already $|x_{12}| < 10^{-9}$: the trajectory converges to the origin without ever leaving the supercritical region.

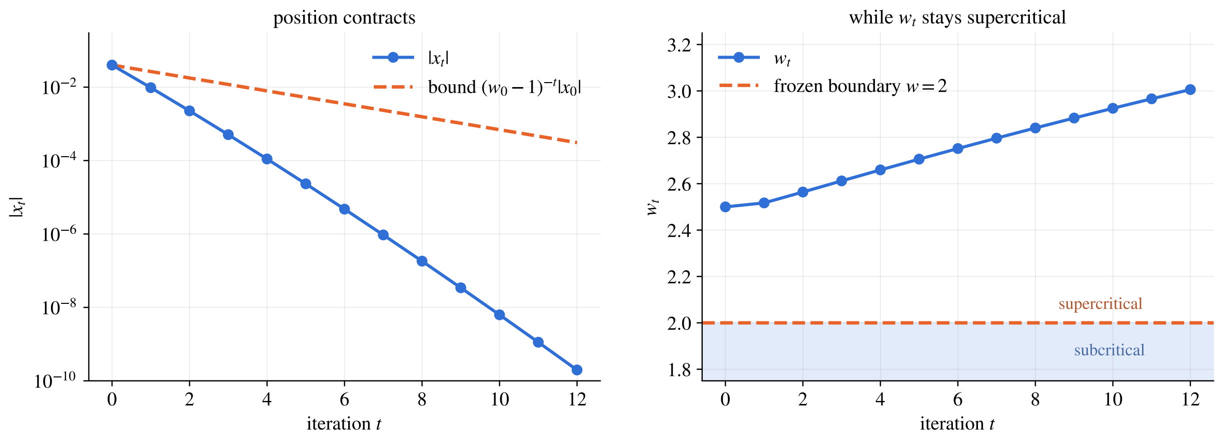

Figure 4: The exceptional orbit of Proposition 3.6 for $\beta_1 = \frac{1}{2}$, $\beta_2 = 0.9$, $\eta = 1$, $\varepsilon = \frac{1}{12}$, $v_0 = 0.0025$, $x_0 = 0.04$, with m_0 determined by the shooting argument of Appendix F.3. Left: the sign of x_t alternates at every step and $|x_t|$ decays well inside the certified envelope $(w_0 - 1)^{-t}\,|x_0|$ of (15) (dashed). Right: along the same steps w_t increases monotonically from $w_0 = \frac{5}{2}$ toward $w_{\max} = 4$ and never leaves the supercritical region.

G Proofs for the $w_{\max} < 2$ case

G.1 Global convergence in the globally subcritical zero-momentum regime

We record separately the complementary regime $w_{\max} < 2$ for $\beta_1 = 0$. In this case the entire trajectory remains strictly below the frozen stability threshold, and the stopped-trajectory argument used in Theorem D.1 becomes a global contraction argument.

Theorem G.1 (Global exponential convergence for $\beta_1 = 0$)**.** *Assume*

$$\beta_1 = 0, \qquad w_{\max} = \frac{\eta}{\varepsilon} < 2.$$

Then every trajectory of Adam converges to the origin. More precisely, for every initial state $(x_0, m_0, v_0) \in \mathbb{R}^2 \times [0, \infty)$, there exist constants $C > 0$ and $\gamma \in (0, 1)$, depending on the initial state and the parameters, such that

$$|x_t| + |m_t| + v_t \le C\gamma^t, \qquad t \in \mathbb{N}_0.$$

In particular, $(x_t, m_t, v_t) \to (0, 0, 0)$.

Proof. Since $\beta_1 = 0$, one has $c = 1$ and

$$m_{t+1} = x_t, \qquad v_{t+1} = \beta_2 v_t + (1 - \beta_2)x_t^2, \qquad x_{t+1} = (1 - w_{t+1})x_t,$$

where

$$w_t = \frac{\eta}{\sqrt{v_t} + \varepsilon}.$$

Because $v_t \ge 0$,

$$0 < w_t \le w_{\max} < 2 \qquad \text{for every } t \ge 0.$$

Hence

$$|x_{t+1}| = |1 - w_{t+1}|\,|x_t| \le |x_t|,$$

so that $|x_t| \le |x_0|$ for every $t \ge 0$.

Set $M := \max\{v_0, x_0^2\}$. We claim that $v_t \le M$ for every $t \ge 0$. Indeed, this is true at $t = 0$, and if $v_t \le M$, then

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)x_t^2 \le \beta_2 M + (1-\beta_2)M = M.$$

Therefore

$$\underline{w} := \frac{\eta}{\sqrt{M}+\varepsilon} \le w_t \le w_{\max} < 2 \qquad \text{for every } t \ge 0.$$

Define

$$\rho := \max\bigl\{|1-\underline{w}|\,, |1-w_{\max}|\bigr\}.$$

Since $0 < \underline{w} \le w_{\max} < 2$, we have $0 \le \rho < 1$. The convexity of $w \mapsto |1-w|$ therefore gives $|1-w_t| \le \rho$ for every $t \ge 0$. Consequently, $|x_{t+1}| \le \rho\,|x_t|$, and hence

$$|x_t| \le \rho^t\,|x_0|\,. \tag{101}$$

It remains to control the second moment. Unrolling its recursion yields

$$v_t = \beta_2^t v_0 + (1-\beta_2)\sum_{j=0}^{t-1} \beta_2^{t-1-j} x_j^2.$$

Using (101),

$$v_t \le \beta_2^t v_0 + (1-\beta_2)x_0^2 \sum_{j=0}^{t-1} \beta_2^{t-1-j}\rho^{2j} = \beta_2^t v_0 + (1-\beta_2)x_0^2 G_t(\rho^2,\beta_2).$$

Since $\rho^2 < 1$ and $\beta_2 < 1$, the right-hand side converges to zero. Moreover, letting $\lambda := \max\{\rho^2, \beta_2\} < 1$ and choosing any $\gamma \in (\lambda, 1)$, the geometric convolution satisfies $G_t(\rho^2,\beta_2) \le C_1\gamma^t$ for some $C_1 > 0$. Thus $v_t \le C_2\gamma^t$ after enlarging C_2 if necessary.

Finally, $m_{t+1} = x_t$, so $|m_{t+1}| \le \rho^t\,|x_0|$. Combining the preceding estimates, and absorbing the finite initial values into the constant, gives

$$|x_t| + |m_t| + v_t \le C\widetilde{\gamma}^t$$

for some $C > 0$ and $\widetilde{\gamma} \in (\max\{\rho,\gamma\}, 1)$. Therefore $(x_t, m_t, v_t) \to (0,0,0)$ exponentially. $\square$

Remark G.2. The condition $w_{\max} < 2$ is sufficient for global convergence when $\beta_1 = 0$ because the scalar position recursion admits a common contraction factor once v_t is bounded. This conclusion does not extend directly to positive momentum: when $0 < \beta_1 < 1$, pointwise frozen stability $w_t < 2$ does not in general imply contraction of the adaptive two-dimensional (x_t, m_t) dynamics.

G.2 Global convergence in the globally subcritical regime with positive momentum

The next result removes the zero-momentum restriction.

Theorem G.3 (Global convergence in the strictly subcritical regime)**.** *Assume*

$$0 < \beta_1 < 1, \qquad 0 \le \beta_2 < 1, \qquad \sqrt{\beta_2} > \beta_1^2, \qquad w_{\max} < 2.$$

Then every trajectory of Adam on $f(x) = \frac{1}{2}x^2$ converges to the origin. More precisely, for every initial state $(x_0, m_0, v_0) \in \mathbb{R}^2 \times [0,\infty)$, there exist constants $C > 0$ and $\gamma_0 \in (0,1)$, possibly depending on the initial state, such that

$$|x_t| + |m_t| + v_t \le C\gamma_0^t, \qquad t \ge 0.$$

In particular, $(x_t, m_t, v_t) \to (0,0,0)$ and $w_t \to w_{\max}$.

Proof. Recall

$$\xi_t := \frac{m_t + cx_t}{c}, \qquad z_t := \begin{pmatrix} x_t \\ \xi_t \end{pmatrix},$$

so that $z_{t+1} = A(w_{t+1})z_t$. Also recall

$$\theta(w) := \frac{w}{2-w}, \qquad \Phi(w) := \frac{w}{\sqrt{\beta_2} + (1-\sqrt{\beta_2})w/w_{\max}}.$$

Every Adam transition satisfies $0 < w_{t+1} \le \Phi(w_t)$.

We first construct a quadratic weight valid on the entire interval $(0, w_{\max}]$. Define

$$R(w) := \frac{\beta_1^2\theta(\Phi(w))}{\theta(w)} = \beta_1^2 \frac{2-w}{2\sqrt{\beta_2} - \big[1 - 2(1-\sqrt{\beta_2})/w_{\max}\big]w}.$$

A direct differentiation gives

$$R'(w) = \frac{2\beta_1^2(1-\sqrt{\beta_2})(1-2/w_{\max})}{\big\{2\sqrt{\beta_2} - \big[1-2(1-\sqrt{\beta_2})/w_{\max}\big]w\big\}^2}.$$

Since $w_{\max} < 2$, one has $R'(w) < 0$. Moreover,

$$\lim_{w\downarrow 0} R(w) = \frac{\beta_1^2}{\sqrt{\beta_2}} < 1$$

by the assumption $\sqrt{\beta_2} > \beta_1^2$. Hence

$$\beta_1^2\theta(\Phi(w)) < \theta(w), \qquad 0 < w \le w_{\max}.$$

Define

$$y(w) := \beta_1\sqrt{\theta(w)\theta(\Phi(w))}.$$

Then

$$\beta_1^2\theta(\Phi(w)) < y(w) < \theta(w), \qquad 0 < w \le w_{\max}.$$

For any valid transition $w \mapsto u$, monotonicity of θ and $u \le \Phi(w)$ give

$$y(w) > \beta_1^2\theta(u), \qquad y(u) < \theta(u).$$

The exact one-step diagonal quadratic criterion therefore yields

$$A(u)^\top P(y(u))A(u) \prec P(y(w)), \qquad P(y) := \operatorname{diag}(1, y).$$

Set

$$E_t := z_t^\top P(y(w_t))z_t = x_t^2 + \frac{y(w_t)}{c^2}(m_t + cx_t)^2.$$

It follows that $E_{t+1} \le E_t$, with strict inequality whenever $z_t \neq 0$. In particular, $x_t^2 \le E_t \le E_0$.

Let $M := \max\{v_0, E_0\}$. The second-moment recursion then gives inductively $v_t \le M$, and hence

$$0 < \underline{w} := \frac{c\eta}{\sqrt{M} + \varepsilon} \le w_t \le w_{\max} < 2.$$

Thus all transitions of the trajectory lie in the compact set

$$\mathcal{K} := \{(w, u) : \underline{w} \le w, u \le w_{\max},\ u \le \Phi(w)\}.$$

The strict quadratic inequality is continuous on $\mathcal{K}$. Consequently, there exists $q \in (0,1)$ such that $E_{t+1} \le q^2 E_t$, and therefore $E_t \le q^{2t}E_0$. Since y has a positive minimum on $[\underline{w}, w_{\max}]$, both x_t and ξ_t, and hence m_t, decay geometrically.

Finally,

$$v_t = \beta_2^t v_0 + (1-\beta_2)\sum_{j=0}^{t-1}\beta_2^{t-1-j}x_j^2,$$

and the geometric bound on x_j^2 implies geometric decay of v_t. Thus there exist $C > 0$ and $\gamma_0 \in (0,1)$ such that $|x_t| + |m_t| + v_t \le C\gamma_0^t$. Hence $(x_t, m_t, v_t) \to (0,0,0)$ and, consequently, $w_t \to w_{\max}$. $\square$

H Experimental details for Figure 1

Left, top. Standard (bias-corrected) full-batch Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\eta = 10^{-2}$, $\varepsilon = 10^{-8}$ trains a two-layer network $x \mapsto W_2^\top \tanh(W_1 x + b_1) + b_2$ with 16 hidden units on the squared loss for the regression task $y = \sin(2x)$ over 32 equispaced inputs in $[-2, 2]$. Every 40 steps the Hessian is computed by central finite differences and we plot the preconditioned sharpness S_t of (18), with v_t replaced by its bias-corrected estimate; S_t equilibrates near the frozen threshold $\mathsf{S}^\star = 2(1+\beta_1)/\big(\eta(1-\beta_1)\big) = 38/\eta$, the adaptive edge of stability of Cohen et al. [2022].

Left, bottom. Uncorrected Adam (2) on the quadratic (1) with the same (β_1, β_2) and $\eta = 1$, $w_{\max} = 10$ (equivalently $\varepsilon = c\eta/10$), $x_0 = 0.1$, $m_0 = 0$, $v_0 = 0.01$; the normalized sharpness w_t oscillates around the parameter-free boundary $w = 2$.

Right. The negative-feedback loop: supercritical steps expand the state and inflate v_t, pushing w_t down, while subcritical steps contract the state and deflate v_t, pushing w_t up. For $\beta_1 = 0$, Appendix D gives explicit finite-step bounds for both transitions.