Momentum Acceleration of Normalized Steepest Descent at the Edge of Stability

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Optimizers based on normalized steepest descent (NSD) with momentum have seen growing success in training large-scale language models. Despite their widespread empirical adoption, the role of momentum in the full-batch regime remains unclear. In this paper, we identify a novel mechanism by which momentum accelerates NSD in the oscillatory training regime where the loss value does not decrease monotonically. Specifically, momentum suppresses the oscillatory component in the momentum buffer, so the stable progress direction becomes dominant in the unit-norm update. Theoretically, we provide a rigorous justification for this mechanism using a two-dimensional quadratic objective that captures the essential features of the oscillatory dynamics. Our analysis also extends to the sign-based variant of NSD, where momentum is provably essential for making progress in the oscillatory regime. Empirically, we validate the theory with full-batch training of an MLP network, where momentum significantly improves the final loss and delays the onset of the Edge of Stability.

1. Introduction

Normalized steepest descent with momentum (NSD-M) is a family of first-order methods that covers several modern optimizers, including sign-based optimizers, spectral optimizers, and their variants [4, 8, 21, 27, 32]. Despite the empirical success, existing theories for NSD are primarily based on smoothness assumptions and one-step descent lemmas [9, 18, 25, 40]. Within this framework, momentum is typically regarded as a variance-reduction device in stochastic settings [3, 12, 13, 35]. However, in deterministic settings, the same analysis implies that momentum introduces an unnecessary lag behind the current gradient, suggesting that momentum may slow convergence.

This conclusion relies on a one-step view of optimization, where progress is certified by comparing the update direction with the current gradient. Practical finite-step training can violate this picture even in deterministic full-batch dynamics. In particular, neural network training often enters the Edge of Stability (EoS), where sharp directions of the Hessian oscillate near a stability threshold while the loss decreases non-monotonically on a slower time scale [10, 11, 19]. Classical smoothness-based analysis does not describe this regime, leaving the role of momentum within it unexplained. This motivates the following question:

Can momentum help normalized optimizers by averaging deterministic oscillations?

In this work, we approach this through a simple high-level model that aligns with the observed oscillatory dynamics. We claim that momentum acceleration is a consequence of the normalization-induced trade-off between the oscillatory and stable components in the gradient. Indeed, once

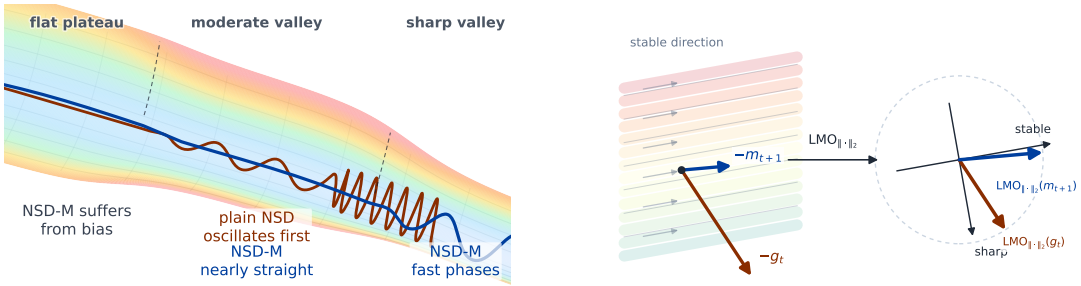

Figure 1: Visualization of stable and oscillatory decompositions and the LMO trade-off.

oscillation sets in, the normalization step in NSD significantly reduces the effective progress along the stable direction. However, momentum helps by damping the alternating sharp component, and the resulting linear minimization oracle can choose a more stable update direction. As detailed in Section 2, this mechanism delays the onset of oscillation and accelerates average progress thereafter.

Contributions. In Section 2, we formulate a stable-oscillatory decomposition of the gradient and identify two complementary mechanisms behind momentum acceleration in NSD: delaying the onset of oscillation, and accelerating average progress once oscillation sets in. We rigorously justify both mechanisms on a two-dimensional quadratic-linear model in Theorem 1, and further extend the analysis to the sign-based variant in Section D, where momentum is provably essential for making progress. In Section 3, we empirically validate these mechanisms on full-batch `NormGD` training of an MLP, where momentum significantly improves the final loss and delays the onset of the EoS.

2. A general mechanism for momentum acceleration in NSD

Throughout, we work with a differentiable loss f and write $g_t = \nabla f(\theta_t)$ for its gradient at t-th iterate. For any norm $\|\cdot\|$ on the parameter space $\mathbb{R}^d$, the associated linear minimization oracle (LMO), as in Frank–Wolfe and norm-constrained optimizer geometry [16, 20, 32], returns a unit-norm direction most aligned with the negative of its input: $\mathsf{LMO}_{\|\cdot\|}(v) \in \arg\min_{\|u\|\le 1}\langle v, u\rangle$. The stated NSD-M maintains an exponential moving average momentum of past gradients and updates using momentum LMO. Suppose that $m_0 = 0$ and $\beta \in [0, 1)$. Then the update is

$$m_{t+1} = \beta m_t + (1-\beta) g_t, \qquad \theta_{t+1} = \theta_t + \eta\, \mathsf{LMO}_{\|\cdot\|}(m_{t+1}). \tag{1}$$

Setting $\beta = 0$ recovers vanilla NSD. Different norms instantiate familiar algorithms: the Euclidean norm gives `NormGD`, the ℓ_∞ norm gives `SignGD`, and the spectral norm gives `SpecGD`-type updates [6], including the recent `Muon` optimizer [21]. Due to space constraints, our analysis below focuses on `NormGD` and its momentum variant and defers the analysis of the sign-based variant to Section D.

Our starting point is a stable-oscillatory decomposition of the gradient, $g_t = h_t + \xi_t$, where h_t varies slowly and ξ_t oscillates. Empirically, ξ_t dominates in the dual norm once the training dynamic enters the oscillatory regime, namely $\|\xi_t\|_* \gg \|h_t\|_*$. Within this model, momentum plays two complementary roles. First, it delays the onset of oscillation by enlarging the local-stability window. Second, once oscillation sets in, it accelerates the average progress by suppressing ξ_t in the momentum buffer. Indeed, both effects hinge on the trade-off between stable and oscillatory components in the normalization step of NSD, which we isolate next.

Normalization: trade-off between stable and oscillatory directions. The key feature for NSD is that the normalization step, or LMO in Eq. (1), retains unit-norm updates throughout the entire

trajectory. Therefore, a dominating oscillatory component in the LMO input significantly suppresses the progress along the stable direction. Assume for simplicity that h_t and ξ_t are nearly orthogonal. Then the stable and oscillatory fractions of the ℓ^2-normalized direction scale as

$$\alpha_t^{(2)} \approx \frac{\|h_t\|_2}{\left(\|h_t\|_2^2+\|\xi_t\|_2^2\right)^{1/2}}, \qquad \zeta_t^{(2)} \approx \frac{\|\xi_t\|_2}{\left(\|h_t\|_2^2+\|\xi_t\|_2^2\right)^{1/2}}.$$

Clearly, a large sharp component reduces down the stable component after ℓ_2 normalization. The right panel of Figure 1 depicts this trade-off in $\mathsf{LMO}_{\|\cdot\|_2}$. This phenomenon is the core reason for the momentum acceleration effect, as we will elaborate below.

Mechanism 1: local stability with momentum. Momentum first helps by delaying the onset of oscillation in the sharp direction. Near a valley floor, the onset of oscillation is governed by the linearized dynamics along the sharp direction. Plain NSD is effectively a one-step feedback system: the iterates respond instantaneously to the sharp-direction gradient, and oscillate once that gradient becomes too strong. With momentum, the LMO is driven by the momentum buffer rather than the instantaneous gradient, so the sharp-direction feedback enters the update in a non-instantaneous, averaged manner. This two-state inertial dynamics has a larger linear stability threshold. In the quadratic model of Section 2.1, we show that the stability window expands from $\rho < 2$ to $\rho < 2(1+\beta)/(1-\beta)$, where $\rho := \eta L/b$ is the rescaled sharpness. This explains why `NormGD-M` remains stable in the moderately sharp area where plain `NormGD` already oscillates.

Mechanism 2: acceleration in oscillatory regime. The main acceleration effect in the oscillatory regime takes place through *cancellation of the sharp component in the momentum buffer*. We illustrate this mechanism with a period-two oscillation model. Suppose that for some large t, it holds that $\xi_t \approx (-1)^t r_t$. Then the momentum buffer is approximately $m_{t+1} \approx h_t + \frac{1-\beta}{1+\beta}(-1)^t r_t$. Based on this, we can deduce that momentum significantly reduces the sharp component in the buffer when β is close to one. As a result, the normalized update aligns more with the stable direction and makes NSD-M faster.

Interestingly, the practical EoS regime witnesses period-two oscillation in the sharp direction as shown in our experiments in Section 3. However, more complex oscillation patterns exist in practice, especially for sign-based NSDs like `SignGD`. Under those circumstances, momentum still helps by suppressing the oscillatory components.

2.1. Case study: quadratic loss with a linear component

We rigorously establish both mechanisms on a two-dimensional loss $f(x, y) = Lx^2/2 - by$ with $L, b > 0$, which separates a sharp direction x from a flat progress direction y in their cleanest form. The flat coordinate y drives long-run descent, while the sharp coordinate x oscillates once the rescaled sharpness $\rho := \eta L/b$ exceeds the stability threshold. We further rescale $z_t := Lx_t/b$ for simplicity. We summarize the main results here; full proofs are in Section C, and the parallel analysis for the sign-oracle variant is in Section D.

`NormGD` baseline. Theorem 6 establishes a phase transition at $\rho = 2$. When $\rho \le 2$, the sharp coordinate satisfies $x_t \to 0$, and the flat speed converges to the full normalized stepsize η. When $\rho > 2$, the sharp coordinate converges to an explicit period-two orbit, and the flat speed saturates at the geometry-limited value $2b/L$. This period-two orbit is precisely the alternating sharp signal that Mechanism 2 takes as input.

Local stability of `NormGD-M`. The fixed point of the exact EMA-momentum map is $E = (0, 0, 1)$. The strict local-stability condition is $\rho < \rho_M^\star(\beta) := 2(1+\beta)/(1-\beta)$, or equivalently $\beta > \beta_c(\rho) := (\rho-2)/(\rho+2)$ for fixed $\rho > 2$. Theorem 9 also gives the consequence for sufficiently small initial sharp coordinate $|x_0|$: under this strict condition, $x_t \to 0$ and the iterate moves at the full flat speed η. This realizes Mechanism 1 in the quadratic model: in the band $\rho \in (2, \rho_M^\star(\beta))$, `NormGD-M` remains locally stable while plain `NormGD` has already entered oscillation.

Provable acceleration in the oscillatory regime. Beyond the local-stability region, `NormGD-M` still strictly exceeds the Cesàro flat-progress rate of plain `NormGD`, even when both methods remain oscillatory. Let $\mathcal{E}_\rho$ denote the countable, measure-zero set of initial states that lead plain `NormGD` to land exactly on the unstable fixed point; generic z_0 avoid it. Let $\kappa_{\rho,\beta} := (1+\beta)/(1+\beta+\beta\rho m_{\rho,\beta})$, where $m_{\rho,\beta} := 1/\{4[1+4\rho^2/(1-\beta)^2]^{3/2}\}$, so that $\kappa_{\rho,\beta} \in (0,1)$ quantifies the EMA-induced shrinkage of the sharp amplitude entering the LMO.

Theorem 1 (Asymptotic acceleration of `NormGD-M`) *Fix $\rho > 2$ and $\beta \in (0,1)$. Let $(\theta_t^{\mathrm{M}})_{t\ge 0}$ and $(\theta_t^{\mathrm{G}})_{t\ge 0}$ denote the iterates of `NormGD-M` and vanilla `NormGD`, respectively. Assume that both methods start from the common state $\theta_0 = (x_0, y_0)$, that the momentum method uses $m_0 = 0$, and that $z_0 = Lx_0/b \notin \mathcal{E}_\rho$. Then*

$$\liminf_{N\to\infty} \frac{y_N^{\mathrm{M}} - y_N^{\mathrm{G}}}{N} \ge \eta \min\left\{1, \frac{2}{\rho\,\kappa_{\rho,\beta}}\right\} - \frac{2b}{L} > 0.$$

Theorem 1 is the comparison form of Theorem 15. Intuitively, EMA averaging shrinks the sharp amplitude entering the LMO, so the effective dimensionless sharpness is $\rho\kappa_{\rho,\beta} < \rho$ rather than ρ. In particular, whenever $\rho\kappa_{\rho,\beta} \le 2$, `NormGD-M` recovers the full normalized stepsize η on the flat coordinate, strictly exceeding the geometric ceiling $2b/L$ of plain `NormGD`. The bound holds without classifying the global attractor of the coupled (z, p, q) map, isolating the cancellation effect as the source of acceleration.

3. Training neural networks using NSD with momentum

We test the mechanism from Section 2 on a deterministic teacher–student MLP with full-batch mean-squared-error loss. We run `NormGD-M` from Eq. (1) over a momentum sweep at two learning-rate anchors, denoted by η_{s} and $\eta_{\mathrm{l}} = 8\eta_{\mathrm{s}}$. The numerical configuration is deferred to Section B.2.

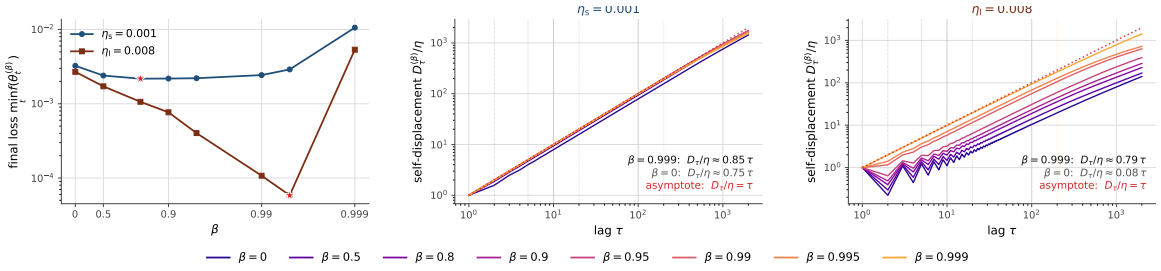

Figure 2: The left panel shows final loss against β. The middle and right panels show $D_\tau^{(\beta)}/\eta$ against lag τ at the two learning-rate anchors.

Average progress. The left panel of Figure 2 shows that the loss-optimal momentum at η_{l} lies above that at η_{s}, matching the prediction from Section 2 that momentum is more useful at a larger learning rate because the dynamics is more oscillatory. Besides loss, we report the normalized lag-τ self-displacement $D_\tau^{(\beta)}/\eta$ (defined in Section B.2), which should be suppressed at short lags and grow linearly for sufficiently large τ in an oscillatory regime. The middle and right panels of Figure 2 highlight this contrast: at η_{s} all β curves grow nearly linearly in τ, while at η_{l} the small-β curves develop strong short-lag self-cancellation and recover only a much smaller long-lag slope. Remarkably, $D_\tau^{(\beta)}/\eta$ at η_{l} dips at $\tau = 2, 4, 6, \ldots$ for small β, certifying a period-two oscillation along a sharp direction.

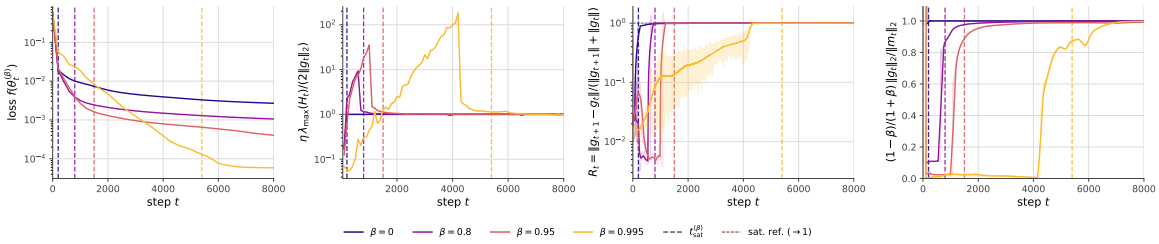

Figure 3: EoS diagnostics. Left to right: loss, the generalized sharpness, normalized gradient-diff norm and gradient-to-momentum ratio. Dashed vertical lines mark the EoS onset.

Delayed EoS and oscillation. Figure 3 verifies the two hypotheses from Section 2. First, the generalized sharpness [19] shows that larger β enters the EoS regime later, matching the enlarged stability window $\rho < 2(1+\beta)/(1-\beta)$ from Theorem 9. The loss column confirms the same delay: smaller β wins before its own EoS onset and loses after, certifying the two-fold effect of momentum — delaying EoS entry while accelerating progress in the oscillatory regime. Second, R_t saturates to 1 after EoS, which implies that the dynamics converges into a period-two oscillation and that the oscillatory component dominates the stable component in the dual norm. The rescaled gradient-to-momentum ratio $\frac{1-\beta}{1+\beta} \cdot \frac{\|g_t\|_2}{\|m_t\|_2}$ converges to 1 after EoS across the β sweep, directly verifying the momentum suppression factor $\frac{1-\beta}{1+\beta}$ predicted by $m_{t+1} \approx h_t + \frac{1-\beta}{1+\beta}(-1)^t r_t$ from Section 2, being an evidence for the period-two oscillation and oscillatory direction dominance.

Additional experiments. The full EoS diagnostics at both learning rates in Section B.2 confirms the same picture at η_{s} once the EMA window settles, and Section B.3 provides a continuous-time counterfactual where the momentum gain disappears. Section B.4 extends the oscillation analysis to `Signum` and `Muon`. In this part, we observe a more complex behavior with multiple oscillation periods. Section B.5 extends the `NormGD-M` experiment to a 273k-parameter ResNet-20.

4. Conclusion

This paper identifies a deterministic mechanism for momentum's acceleration of NSD near the Edge of Stability: the EMA suppresses the sharp alternating component before the LMO is applied, so more of the unit-norm update is spent on stable progress. Both the enlarged local stability window in the linear-flat model and the average-progress advantage observed in the full-batch teacher–student experiments at the Edge-of-Stability anchor support this mechanism.

References

[1] Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for SGD, 2025. URL https://arxiv.org/abs/2412.20553.

[2] Arseniy Andreyev, Advikar Ananthkumar, Marc Walden, Tomaso Poggio, and Pierfrancesco Beneventano. Momentum further constrains sharpness at the edge of stochastic stability, 2026. URL https://arxiv.org/abs/2604.14108.

[3] Sébastien M. R. Arnold, Pierre-Antoine Manzagol, Reza Babanezhad, Ioannis Mitliagkas, and Nicolas Le Roux. Reducing the variance in online optimization by transporting past gradients. In *Advances in Neural Information Processing Systems*, volume 32, pages 5392–5403, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1dba5eed8838571e1c80af145184e515-Abstract.html.

[4] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 560–569. PMLR, 2018.

[5] Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods. *arXiv preprint arXiv:2211.14103*, 2022.

[6] David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted boltzmann machines. In *Artificial intelligence and statistics*, pages 111–119. PMLR, 2015.

[7] John Chen, Cameron Wolfe, Zhao Li, and Anastasios Kyrillidis. Demon: Improved neural network training with momentum decay, 2021. URL https://arxiv.org/abs/1910.04952.

[8] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms, 2023. URL https://arxiv.org/abs/2302.06675.

[9] Ziyi Chen, Yi Zhou, Yingbin Liang, and Zhaosong Lu. Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5396–5427. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23ar.html.

[10] Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability, 2021.

[11] 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. In *Advances in Neural Information Processing Systems 35*, 2022.

[12] Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 2020.

[13] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b8002139cdde66b87638f7f91d169d96-Paper.pdf.

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

[15] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data, 2025. URL https://arxiv.org/abs/2502.04664.

[16] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1–2):95–110, 1956. doi: 10.1002/nav.3800030109.

[17] Jingwen Fu, Bohan Wang, Huishuai Zhang, Zhizheng Zhang, Wei Chen, and Nanning Zheng. When and why momentum accelerates SGD: An empirical study, 2023. URL https://arxiv.org/abs/2306.09000.

[18] Elad Hazan, Kfir Y. Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems 28*, 2015.

[19] Rustem Islamov, Michael Crawshaw, Jeremy Cohen, and Robert Gower. Non-Euclidean gradient descent operates at the edge of stability, 2026.

[20] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435. PMLR, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

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

[22] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham M. Kakade. On the insufficiency of existing momentum schemes for stochastic optimization, 2018. URL https://arxiv.org/abs/1803.05591.

[23] Nikola B. Kovachki and Andrew M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17):1–40, 2021. URL https://jmlr.org/papers/v22/19-466.html.

[24] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. doi: 10.1137/15M1009597.

[25] Kfir Y. Levy. The power of normalization: Faster evasion of saddle points, 2016.

[26] Kaizhao Liang, Lizhang Chen, Bo Liu, and Qiang Liu. Cautious optimizers: Improving training with one line of code, 2026. URL https://arxiv.org/abs/2411.16085.

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

[28] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum, 2020. URL https://arxiv.org/abs/2007.07989.

[29] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of Machine Learning Research*, 21(105):1–49, 2020.

[30] Michael Muehlebach and Michael I. Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22 (73):1–50, 2021. URL https://jmlr.org/papers/v22/20-207.html.

[31] Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[32] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs, 2025. URL https://arxiv.org/abs/2502.07529.

[33] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[34] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. doi: 10.1016/S0893-6080(98)00116-6.

[35] Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtárik. Muon is provably faster with momentum variance reduction, 2025. URL https://arxiv.org/abs/2512.16598.

[36] Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon and Scion great again!, 2025. URL https://arxiv.org/abs/2505.13416.

[37] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147. PMLR, 2013. URL https://proceedings.mlr.press/v28/sutskever13.html.

[38] Jun-Kun Wang, Chi-Heng Lin, and Jacob D. Abernethy. A modular analysis of provable acceleration via Polyak's momentum: Training a wide ReLU network and a deep linear network. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10816–10827. PMLR, 2021. URL https://proceedings.mlr.press/v139/wang21n.html.

[39] Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞ norm constrained optimization, 2024. URL https://arxiv.org/abs/2404.04454.

[40] Shuo Xie, Tianhao Wang, Beining Wu, and Zhiyuan Li. A tale of two geometries: Adaptive optimizers and non-Euclidean descent, 2025.

[41] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning, 2018. URL https://arxiv.org/abs/1808.10396.

[42] Dingzhi Yu, Hongyi Tao, Yuanyu Wan, Luo Luo, and Lijun Zhang. Sign-based optimizers are effective under heavy-tailed noise, 2026. URL https://arxiv.org/abs/2602.07425.

[43] Kun Yuan, Bicheng Ying, and Ali H. Sayed. On the influence of momentum acceleration on online learning. *Journal of Machine Learning Research*, 17(192):1–66, 2016. URL https://jmlr.org/papers/v17/16-157.html.

[44] Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for language models, 2025. URL https://arxiv.org/abs/2407.07972.

Contents

Appendix A. Related work

Normalized steepest descent and modern optimizer geometry. The linear-minimization-oracle view underlying NSD is classical in projection-free optimization. Frank–Wolfe and conditional-gradient methods choose descent directions by solving linear problems over constraint sets [5, 16, 20, 29]. In deep learning, the same geometric template appears in coordinate-wise sign methods, Euclidean normalized descent, and spectral updates such as `Muon` [4, 8, 21, 27, 32]. Several recent papers analyze variants of this template through adaptive or non-Euclidean geometry, including ℓ_∞-style interpretations of `AdamW`, norm-constrained linear-minimization oracles, and spectral-descent implicit bias [15, 36, 39, 40]. These works motivate NSD-M as a unifying abstraction for modern optimizers. Our focus is different: we study how momentum changes the deterministic oscillatory dynamics of the normalized oracle rather than proposing a new optimizer or proving a standard smoothness-based convergence rate.

Classical interpretation of momentum. Classical momentum theory begins with heavy-ball and accelerated methods in deterministic optimization [31, 33]. For plain gradient descent, the standard acceleration story is parameter-coupled. On quadratics and local quadratic models, momentum turns each eigendirection into a damped second-order recurrence; acceleration comes from choosing the learning rate and momentum so that high-curvature modes are damped while low-curvature modes move faster [24, 30, 34]. Continuous-time and modified-equation analyses make the same point: as the learning rate vanishes, fixed momentum approaches a time-rescaled gradient flow, and the distinctive benefits of momentum appear only in finite-learning-rate corrections [23]. In stochastic or online constant-learning-rate settings, momentum can even act mainly as an effective learning-rate rescaling [43]. Recent neural-network theory proves analogous Polyak-momentum acceleration when the learning rate and momentum are tuned to the problem geometry [38]. These mechanisms are not the comparison made here: our experiments keep the learning rate fixed and ask whether changing β changes progress by altering the normalized oracle input, rather than by allowing a larger or retuned gradient-descent step.

In deep learning, momentum is empirically important, but its benefit is known to be regime-dependent [17, 37]. For stochastic normalized methods, Cutkosky and Mehta [12] show that momentum improves normalized stochastic gradient descent by reducing gradient noise, and related analyses of stochastic momentum or momentum variance reduction develop comparable estimator-based explanations [28, 35, 41]. Other works emphasize limitations, schedules, or optimizer-specific variants of momentum, including stochastic counterexamples, momentum decay, `Lion`, `Signum`, and cautious optimizers [7, 22, 26, 42, 44]. These papers mainly explain momentum through stochastic averaging, stability of standard gradient methods, or empirical optimizer design. Our mechanism is instead full-batch and geometric: momentum cancels an alternating sharp component before the NSD oracle normalizes the update.

Edge of Stability and oscillatory dynamics. The Edge-of-Stability literature shows that large-step training can settle near a sharpness threshold and decrease loss through non-monotone oscillatory dynamics [10]. Follow-up work extends this viewpoint to adaptive optimizers, self-stabilizing gradient descent, non-Euclidean descent geometries, and mini-batch stochastic stability [1, 11, 14, 19]. The closest momentum-side work studies how momentum changes sharpness plateaus at the Edge of Stochastic Stability [2]. Our setting is complementary: we isolate a deterministic full-batch mechanism for NSD-M, where oscillation is not merely a stability artifact but the signal that momentum averages before normalization. This distinction is important because the usual descent-lemma analysis treats the same averaging as lag, while the EoS viewpoint reveals why it can increase stable progress.

Appendix B. Additional details on experiments

This appendix contains the extended versions of the figures from Section 3 and the continuous-time control experiment.

B.1. Diagnostic metrics

We collect in one place every metric used in the EoS diagnostics across Section 3, Section B.4 and Section B.5. All metrics are reported in the optimizer's primal/dual norm pair $(\|\cdot\|, \|\cdot\|_*)$:

- `NormGD-M` (ℓ_2 primal): $\|\cdot\|_* = \ell_2$.

- `Signum` (ℓ_∞ primal): $\|\cdot\|_* = \ell_1$.
- `Muon` (per-block max operator norm): $\|\cdot\|_*$ = sum-of-block-nuclear (sum of singular values per matrix block; ℓ_1 per bias block, since `Muon` applies sign to biases).

Generalized sharpness ratio (Cohen-style). The generalized sharpness in primal norm $\|\cdot\|$ is $S_{\|\cdot\|}(w) \coloneqq \sup_{\|d\|\le 1} d^\top H(w)\, d$ [19], and the dimensionless EoS ratio is

$$S_t \;\coloneqq\; \frac{\eta\, S_{\|\cdot\|}(w_t)}{2\,\|g_t\|_*}, \tag{2}$$

which approaches 1 at saturated EoS in any NSD. For `NormGD-M` we have $S_{\|\cdot\|} = \lambda_{\max}(H_t)$ and compute it by 25-step Hessian-vector power iteration every $T/200$ steps. For `Signum` and `Muon`, $S_{\|\cdot\|}$ is a constrained maximum over a non-quadratic ball, so we approximate it by Frank-Wolfe with the per-optimizer linear oracle (sign for `Signum`, per-block polar for `Muon`): five restarts and 50 iterations for `Signum`, three restarts and 30 iterations for `Muon`.

Lag-k gradient-flip ratio. Our own Hessian-free diagnostic is

$$R_{t,k} \;\coloneqq\; \frac{\|g_{t+k} - g_t\|_*}{\|g_{t+k}\|_* + \|g_t\|_*} \;\in\; [0,1], \tag{3}$$

which approaches 1 at a saturated period-$2k$ EoS in the optimizer's dual norm. The default lag $k = 1$ reduces to $R_t = R_{t,1}$ and detects period-two oscillation. Because momentum can stabilize period-doubled cycles (especially for `Signum` and `Muon` at large β), we report the half-period lag

$$k^*(\beta) \;\coloneqq\; \arg\min_k \left|\, \overline{R_{t,k}} - 1 \,\right| \quad \text{(smallest-}k\text{ tie-break)}, \tag{4}$$

with the average taken over the late half of training. Both R_t and the lag selector are unique to this paper.

Buffer-equilibrium ratio and period-$2k$ shrinkage. The EMA buffer $m_t = \beta m_{t-1} + (1-\beta) g_t$ contracts the gradient along any direction in which the gradient alternates with period $2k$. At steady state in such a direction, summing the geometric series gives

$$\frac{\|m_t\|}{\|g_t\|} \;\to\; \frac{1-\beta^k}{1+\beta^k},$$

which reduces to the familiar $(1-\beta)/(1+\beta)$ when $k = 1$ (period-2 alternation, the canonical EoS for NGD). Inverting and rescaling we obtain a dimensionless buffer-equilibrium ratio

$$B_t \;\coloneqq\; \frac{1-\beta^{k^*}}{1+\beta^{k^*}} \cdot \frac{\|g_t\|_*}{\|m_t\|_*} \;\to\; 1$$

once the buffer settles, where k^* is the half-period selected by Eq. (4). For NGD ($k^* = 1$ throughout) this simplifies to $(1-\beta)/(1+\beta) \cdot \|g\|/\|m\|$; for SIGNUM and Muon the half-period grows with β so the corresponding β^{k^*} factor stays in a more uniform range. The warm-up phase is the only β-dependent transient: B_t approaches 1 over $\mathcal{O}(1/(1-\beta))$ steps.

Saturation onset. We use S_t to mark the first time the dynamics enters the saturated EoS regime,

$$t_{\text{sat}}^{(\beta)} \;:=\; \min\{\, t \geq 50 \;:\; S_t \in [0.85, 1.15] \,\}.$$

The dashed verticals in Figures 3, 4 and 8 are these onsets, repeated across columns to compare the saturation moment against the loss-descent acceleration and the buffer-equilibrium transient.

Self-displacement (`NormGD-M` only). For Figure 2 we compute the lag-τ self-displacement

$$D_\tau^{(\beta)} \;:=\; \frac{1}{N-\tau} \sum_{t=0}^{N-\tau-1} \left\| \theta_{t+\tau}^{(\beta)} - \theta_t^{(\beta)} \right\|_2,$$

directly from raw iterates with no smoothing. At saturated EoS, $D_\tau^{(\beta)}/\eta$ grows linearly in τ once τ exceeds the EMA window $1/(1-\beta)$, with a β-dependent slope prefactor.

B.2. Full details for the MLP experiment

We train a deterministic teacher–student MLP with architecture $[16, 32, 32, 1]$ under the full-batch mean-squared-error loss, and sweep `NormGD-M` (Eq. (1)) over $\beta \in \{0, 0.5, 0.8, 0.9, 0.95, 0.99, 0.995, 0.999\}$ at two learning-rate anchors $\eta_{\text{s}} = 10^{-3}$ and $\eta_{\text{l}} = 8 \times 10^{-3}$, with horizons $T = 40{,}000$ and $T = 20{,}000$ respectively. All reported metrics ($D_\tau^{(\beta)}$, S_t, R_t, B_t, $t_{\text{sat}}^{(\beta)}$) follow the definitions in Section B.1.

Figure 4 reports the full 2×4 version of Figure 3, with the top row at the smaller anchor η_{s} and the bottom row at the larger anchor η_{l} zoomed to the first $5{,}000$ steps. At η_{s}, every plotted β enters saturation within the horizon, and the onset shifts later as β grows, in agreement with the EMA warm-up window $1/(1-\beta)$. At η_{l}, saturation arrives within the first few thousand steps for every β in the subsample, including the high-β runs whose loss curves have already separated by that point. Across both anchors and the entire β-sweep, the Hessian-free R_t tracks S_t within visual tolerance, which confirms that the cheap gradient-flip diagnostic captures the same EoS transition as the Cohen-style sharpness.

B.3. A continuous flow experiment with varying momentum

The continuous-time limit of `NormGD-M` is the autonomous flow

$$\tau\, \dot{m} = \nabla f(\theta) - m, \qquad \dot{\theta} = -m/\|m\|_2, \tag{5}$$

obtained from the discrete update by setting $\eta = \mathrm{d}t \to 0$ and $\beta = \exp(-\mathrm{d}t/\tau) \approx 1 - \mathrm{d}t/\tau$, so $\tau = \eta/(1-\beta)$ is the only surviving parameter. Forward-Euler discretization of Eq. (5) at step h is mathematically identical to `NormGD-M`$(\eta = h, \beta = 1 - h/\tau)$, so any "Euler-flow" claim is really a small-η `NormGD-M` claim. We therefore integrate Eq. (5) on the same MLP task with an adaptive 5th-order Runge–Kutta–Tsitouras solver (`diffrax Tsit5`, $\mathrm{rtol} = 10^{-6}$, $\mathrm{atol} = 10^{-8}$).

We match the discrete β-sweep at η_{l} by setting $\tau = \eta_{\text{l}}/(1-\beta)$ for $\beta \in \{0, 0.5, 0.8, 0.9, 0.95\}$. The resulting flow trajectories at $T = 20$ collapse into a $\sim 1.4\times$ band, as shown in Figure 5. The omitted entries $\tau \in \{0.8, 1.6\}$, corresponding to $\beta \sim \{0.99, 0.995\}$, sit an order of magnitude above this band because their EMA equilibration time exceeds T. Compared at the same $T = 20$, the discrete `NormGD-M`(η_{l}, β) trajectories are all *above* the flow band: $\beta = 0$ reaches loss 5.18×10^{-3} ($\sim$

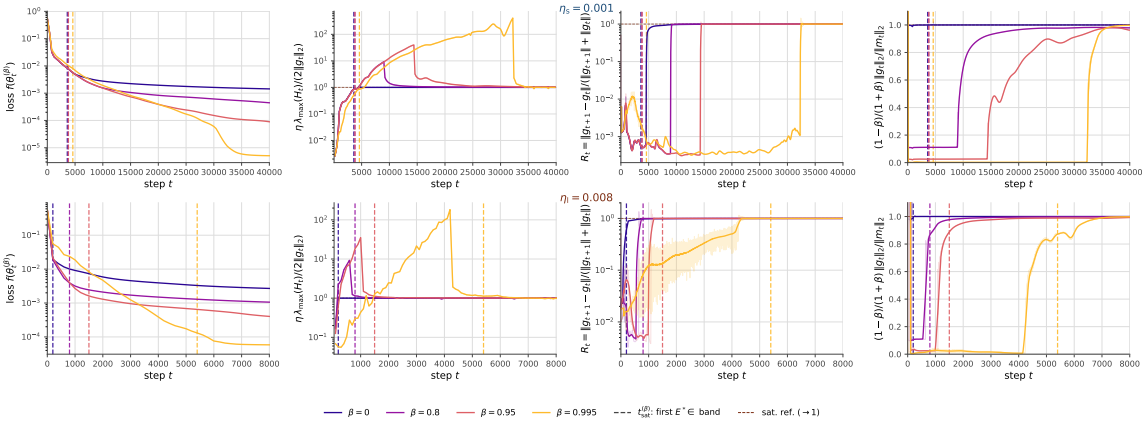

Figure 4: Companion to Figure 3: full 2×4 EoS diagnostic. Rows show η_s and η_l; columns show loss, the generalized sharpness ratio S_t, the gradient-flip ratio R_t, and $\|g_t\|_2$. Per-β vertical lines are the saturation onsets $t_{\text{sat}}^{(\beta)}$, repeated across columns.

$16\times$ above the flow), and the loss-optimal discrete $\beta = 0.95$ still sits $\sim 3\times$ above. Two conclusions: (i) in the continuous limit, momentum confers no acceleration; (ii) the dramatic β-acceleration of Section 3 is therefore a strictly discrete, finite-η EoS phenomenon, not a continuous-flow effect.

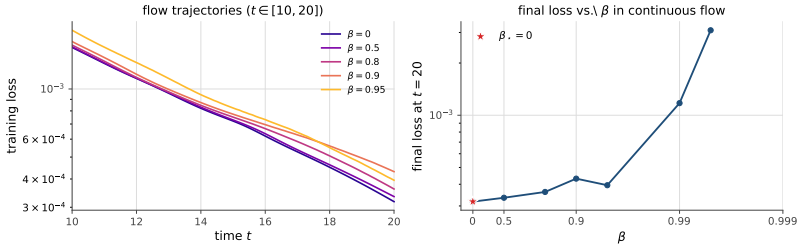

Figure 5: `NormGD-M`-flow (Eq. (5)), τ-sweep solved with `diffrax Tsit5` (rtol $= 10^{-6}$, atol $= 10^{-8}$); $t \in [10, 20]$. $\tau \in \{0.008, 0.016, 0.04, 0.08, 0.16\}$ correspond to $\beta \sim \{0, 0.5, 0.8, 0.9, 0.95\}$ at $\eta_{\text{ref}} = 8 \times 10^{-3}$. All five trajectories collapse into a $\sim 1.4\times$ band: in the continuous limit momentum has no effect.

B.4. Ablations to other NSD variants: `Signum` and `Muon`

The EMA-buffer cancellation picture is not specific to the Euclidean dual norm, and we now check that it survives the transition to two qualitatively different geometries from the normalized-steepest-descent family. We rerun the same $[16, 32, 32, 1]$ teacher–student MLP, full-batch, replacing the ℓ_2 primal of `NormGD-M` with `Signum` (equivalently single-β Lion, ℓ_∞ primal) and with `Muon` (per-block max operator norm). All reported metrics use the optimizer's own dual norm following the conventions of Section B.1.

For `Signum` at $\eta = 10^{-3}$ we observe an order-of-magnitude β-acceleration on the training loss that mirrors the `NormGD-M` phenomenology of Section 3: $\beta = 0$ ends at 9.3×10^{-3} while $\beta = 0.99$ reaches 3.3×10^{-4}, a $\sim 28\times$ gain (Figure 6). The lag-k^* panel confirms that the dynamics period-doubles past $k^* = 1$ at high momentum; for instance, $k^* = 3$ at $\beta = 0.99$ with $\overline{R_{t,k^*}} \approx 0.92$, exactly the longer-period limit cycles predicted by the buffer-equilibrium analysis.

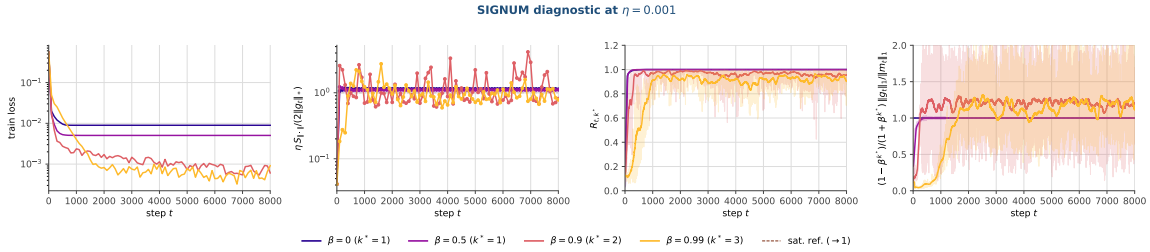

Figure 6: `Signum` (single-β Lion) diagnostic at $\eta = 10^{-3}$ on the $[16, 32, 32, 1]$ MLP, full-batch, $T = 8000$. Panels (left to right): training loss, intermediate sharpness diagnostic, and lag-k^* gradient-flip ratio R_{t,k^*} (Eq. (3)) with per-β half-period k^*.

For `Muon` at $\eta = 10^{-2}$ the β-gain is much milder, only $\sim 1.2\times$ on the final training loss (Figure 7). A plausible interpretation is that the per-block polar factor already performs a spectral averaging analogous to what momentum buys in the scalar case, which leaves less room for further EMA-buffer cancellation. The lag-k^* diagnostic nevertheless still detects period doubling at high momentum ($k^* = 2$ at $\beta = 0.9$ and $k^* = 6$ at $\beta = 0.99$), consistent with momentum stabilizing longer-period limit cycles on top of the polar geometry rather than suppressing oscillation altogether.

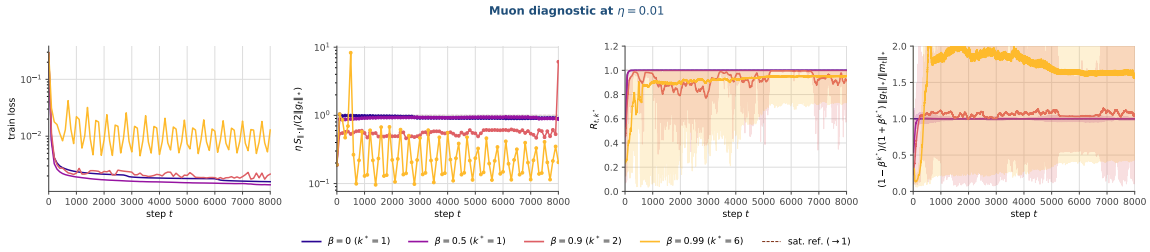

Figure 7: `Muon` diagnostic at $\eta = 10^{-2}$ on the same MLP, full-batch, $T = 8000$. Panels mirror Figure 6.

B.5. Additional experiment: ResNet-20-GN on CIFAR-10

To verify that the picture survives the transition from a toy teacher–student regression to a moderately-sized image classifier, we repeat the diagnostic of Section 3 on a 273k-parameter `ResNet-20` with `GroupNorm`, trained full-batch on a deterministic CIFAR-10 subset of $N = 5{,}000$ samples. The network is a standard 3-stage `ResNet-20` with channel widths $\{16, 32, 64\}$, 3×3

`Conv` + `GroupNorm` (with num_groups $= 4$) + `ReLU` blocks, global average pooling, and a 10-way linear classifier, totalling 273,066 parameters. The training objective is softmax cross-entropy on integer labels, in contrast to the MSE used in Section 3, and the full-batch gradient on $N = 5{,}000$ samples is computed at every step so the trajectory remains deterministic. We sweep the same momentum values $\beta \in \{0, 0.5, 0.8, 0.9, 0.95, 0.99, 0.995\}$ and report the generalized-sharpness ratio and the gradient-flip ratio of Eqs. (2) and (3), together with the buffer-equilibrium ratio $(1-\beta)/(1+\beta) \cdot \|g_t\|_2/\|m_t\|_2$; throughout, $\lambda_{\max}(H_t)$ is estimated by 25-step Hessian-vector power iteration every $T/200$ steps so that the full Hessian is never materialized.

Picking the learning rate requires some care, because the EoS regime is narrow on this architecture. Two preliminary β-sweeps at $\eta \in \{10^{-4}, 10^{-3}\}$ delineate the useful range: at $\eta = 10^{-4}$ the dynamics is strictly sub-EoS for every β ($S_t < 1$ throughout) and the training loss is flat at the random-initialization value, while at $\eta = 10^{-3}$ the Cohen criterion $S_t \to 1$ does fire for every β, but high momentum ($\beta \geq 0.9$) blows up the validation loss. We therefore adopt the intermediate anchor $\eta = 5 \times 10^{-4}$ as the reference, at which every β enters and stays in the saturated EoS regime within a few thousand steps while the training loss remains bounded (Figure 8).

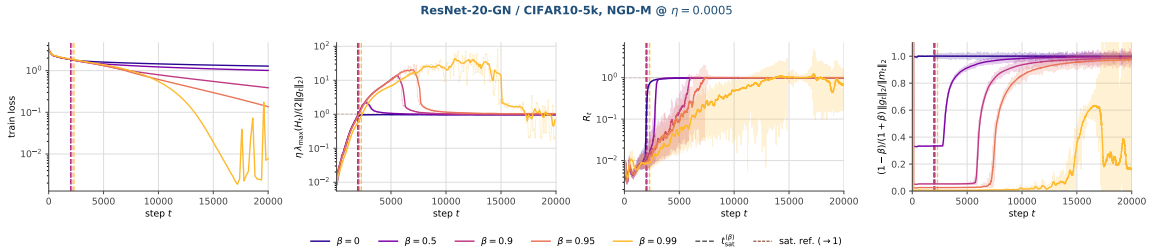

Figure 8: `NormGD-M` on `ResNet-20-GN` (273k parameters) trained full-batch on a deterministic CIFAR-10 subset ($N = 5{,}000$) at $\eta = 5 \times 10^{-4}$, $T = 20{,}000$, with $\beta \in \{0, 0.5, 0.9, 0.95, 0.99\}$. Panels mirror Figure 3: training cross-entropy loss, generalized sharpness $S_t = \eta\, \lambda_{\max}(H_t)/(2\|g_t\|_2)$, the gradient-flip ratio R_t, and the buffer-equilibrium ratio $(1-\beta)/(1+\beta) \cdot \|g_t\|_2/\|m_t\|_2$. The diagnostics behave identically to the toy MLP — every β reaches saturated EoS ($S_t \to 1$, $R_t \to 1$, buffer $\to 1$), with the warm-up window growing as $1/(1-\beta)$ — and the training loss shows a $\sim 700\times$ improvement from $\beta = 0$ to $\beta = 0.99$.

At this anchor the four EoS diagnostics line up cleanly with the toy MLP of Section 3: the generalized sharpness ratio S_t rises monotonically to 1 for every β, the gradient-flip ratio R_t tracks the same value, and the buffer-equilibrium ratio approaches 1 once the EMA has settled, again with a warm-up window growing as $1/(1-\beta)$. The training-loss column reproduces the same β-acceleration pattern as Figure 2: $\beta = 0$ barely moves (final loss 1.28, down from initial ≈ 2.30), $\beta = 0.9$ reaches 0.39, $\beta = 0.95$ reaches 0.14, and $\beta = 0.99$ reaches 1.9×10^{-3} (essentially zero training error), a roughly $700\times$ improvement over $\beta = 0$ on the training objective.

One caveat deserves a remark. On the 5k subset the high-momentum runs strongly overfit, and the validation loss therefore diverges in the opposite direction — it rises from 1.5 to 4.4 at $\beta = 0.99$. This is a generalization artifact of the small training set rather than a contradiction of the EoS-cancellation picture: the optimizer is doing exactly what Section 2 predicts on the training

objective, but the small training set lets the model memorize once the EoS oscillation no longer prevents it from doing so. With a larger training set or explicit regularization we would expect the validation loss to track the training-loss pattern more closely; we report the training-loss view here so the diagnostic comparison with the toy MLP is on equal footing.

Appendix C. Analysis of `NormGD-M` on the quadratic model

C.1. Model and scaled dynamics

We study the exact linear-flat loss $f(x, y) = Lx^2/2 - by$, where $L > 0$ and $b > 0$. Its gradient is $\nabla f(x, y) = (Lx, -b)$. The flat derivative is the constant $-b$, whereas the sharp derivative is the linear term Lx. We scale the sharp coordinate by the flat-gradient magnitude and write $z_t \coloneqq Lx_t/b$. We also write $\rho \coloneqq \eta L/b$ for the dimensionless normalized stepsize.

`NormGD`. `NormGD` takes a unit Euclidean step in the negative gradient direction. After the scaling above, the sharp coordinate follows the autonomous one-dimensional recursion

$$z_{t+1} = F_\rho(z_t) := z_t \left(1 - \frac{\rho}{\sqrt{1 + z_t^2}}\right), \tag{6}$$

and the flat coordinate advances by

$$y_{t+1} - y_t = \frac{\eta}{\sqrt{1 + z_t^2}}. \tag{7}$$

Thus the sharp dynamics decouples from the flat coordinate, while the flat speed is read off from the sharp orbit.

Normalized momentum. The EMA momentum method first averages gradients with coefficient $\beta \in (0, 1)$ and then steps in the normalized momentum direction. We write the momentum state as $m_t = b(p_t, -q_t)$ so that p_t is the scaled sharp momentum and q_t is the scaled flat momentum. The momentum variables obey

$$p_{t+1} = \beta p_t + (1 - \beta) z_t, \qquad q_{t+1} = \beta q_t + (1 - \beta), \tag{8}$$

and the corresponding state update is

$$z_{t+1} = z_t - \rho \frac{p_{t+1}}{\sqrt{p_{t+1}^2 + q_{t+1}^2}}, \qquad y_{t+1} - y_t = \eta \frac{q_{t+1}}{\sqrt{p_{t+1}^2 + q_{t+1}^2}}. \tag{9}$$

Throughout the momentum comparison we use the standard initialization $m_0 = 0$, equivalently $p_0 = q_0 = 0$.

C.2. Exact analysis of `NormGD`

This section gives a complete classification of `NormGD` on the linear-flat model.

The stable regime $0 < \rho \leq 2$.

Proposition 2 (Global convergence to the sharp optimum when $0 < \rho \leq 2$) *Suppose $0 < \rho \leq 2$. Then for every initialization $z_0 \in \mathbb{R}$,*

$$z_t \to 0, \qquad y_t = y_0 + \eta t + o(t).$$

Equivalently,

$$x_t \to 0, \qquad y_t = y_0 + \eta t + o(t).$$

Proof of Theorem 2. The proof is a magnitude-contraction argument for the sharp coordinate. Fix $z \neq 0$ and write $s = \sqrt{1+z^2}$, so $s > 1$. Since $0 < \rho \leq 2$, the multiplier in Eq. (6) satisfies $0 < \rho/s < 2$, and therefore

$$|F_\rho(z)| < |z| \qquad \text{for every } z \neq 0.$$

It follows that $|z_t|$ is strictly decreasing unless $z_t = 0$, so $|z_t|$ converges to some limit $\ell \geq 0$. Passing to the limit in Eq. (6) yields

$$\ell = \ell\,|1 - \rho/\sqrt{1+\ell^2}|.$$

If $\ell > 0$, then $|1 - \rho/\sqrt{1+\ell^2}| = 1$, which is impossible because $0 < \rho/\sqrt{1+\ell^2} < 2$. Hence $\ell = 0$, so $z_t \to 0$.

The flat-speed statement follows by substituting this convergence into the flat increment.

$$y_{t+1} - y_t = \frac{\eta}{\sqrt{1+z_t^2}} \to \eta.$$

Summing the increments gives $y_t = y_0 + \eta t + o(t)$. This completes the proof. ∎

The oscillatory regime $\rho > 2$. Assume now that $\rho > 2$. Write $r_\rho := \sqrt{\rho^2 - 1}$ for the alternating-strip radius and $a_\rho := \sqrt{\rho^2/4 - 1}$ for the period-two amplitude. The open alternating strip is $\mathcal{I}_\rho := (-r_\rho, r_\rho) \setminus \{0\}$. The exact period-two orbit is

$$F_\rho(a_\rho) = -a_\rho, \qquad F_\rho(-a_\rho) = a_\rho. \tag{10}$$

We first record two elementary geometric facts.

Lemma 3 (Outside the alternating strip, the sharp magnitude decreases) *Assume $\rho > 2$. If $|z| > r_\rho$, then $F_\rho(z)$ has the same sign as z and satisfies*

$$0 < |F_\rho(z)| < |z|.$$

Consequently, every orbit with $z_0 \neq 0$ enters the closed strip $[-r_\rho, r_\rho]$ in finitely many steps.

Proof of Theorem 3. The point is that outside the strip the multiplier in the scalar recursion is positive and smaller than one. If $|z| > r_\rho$, then $\sqrt{1+z^2} > \rho$, so the factor in Eq. (6) satisfies

$$0 < 1 - \frac{\rho}{\sqrt{1+z^2}} < 1.$$

This proves the sign preservation and the strict decrease of the magnitude.

It remains to show finite-time entry into the closed strip. Suppose an orbit with $z_0 \neq 0$ never enters $[-r_\rho, r_\rho]$. By odd symmetry we may assume $z_0 > 0$. Then $z_t > r_\rho$ for all t, and the previous paragraph shows that z_t is strictly decreasing and bounded below by $r_\rho > 0$. Hence $z_t \to \ell$ for some $\ell \geq r_\rho$. Passing to the limit in Eq. (6) gives

$$\ell = \ell \left(1 - \frac{\rho}{\sqrt{1+\ell^2}}\right),$$

which is impossible because $\ell > 0$. Therefore every nonzero orbit enters $[-r_\rho, r_\rho]$ in finite time. This completes the proof. ■

Lemma 4 (The alternating strip is invariant) *Assume $\rho > 2$. If $0 < |z| < r_\rho$, then*

$$0 < |F_\rho(z)| < r_\rho,$$

and $F_\rho(z)$ has the opposite sign from z. Thus $\mathcal{I}_\rho$ is forward invariant.

Proof of Theorem 4. The open strip is the region where the multiplier becomes negative but not large enough to leave the strip. Write $r = |z|$ and $s = \sqrt{1+r^2}$. Since $r < r_\rho$, we have $s < \rho$, so

$$1 - \frac{\rho}{s} < 0.$$

Hence the sign flips. Moreover,

$$|F_\rho(z)| = r\left(\frac{\rho}{s} - 1\right) = (\rho - s)\frac{r}{s}.$$

Because $0 < r/s < 1$, we obtain

$$0 < |F_\rho(z)| < \rho - s < \rho - 1 < \sqrt{\rho^2 - 1} = r_\rho.$$

Thus the new point remains in $\mathcal{I}_\rho$. This completes the proof. ■

The next theorem gives the exact asymptotic dynamics inside the strip.

Theorem 5 (Convergence to the explicit period-two orbit inside $\mathcal{I}_\rho$) *Assume $\rho > 2$ and $0 < z_0 < r_\rho$. Then*

(i) $z_t \in \mathcal{I}_\rho$ for every $t \geq 0$, and the sign alternates at every step;

(ii) the even subsequence satisfies $z_{2t} > 0$ and converges monotonically to a_ρ;

(iii) the odd subsequence satisfies $z_{2t+1} < 0$ and converges monotonically to $-a_\rho$.

By odd symmetry, if $-r_\rho < z_0 < 0$, then $z_{2t} \to -a_\rho$ and $z_{2t+1} \to a_\rho$.

Proof of Theorem 5. By Theorem 4, the entire orbit stays in $\mathcal{I}_\rho$ and the sign alternates. It therefore suffices to study the positive even subsequence through the two-step map

$$G_\rho := F_\rho \circ F_\rho$$

on $(0, r_\rho)$.

Fix $z \in (0, r_\rho)$. Write $s = \sqrt{1+z^2}$, $z_1 = F_\rho(z)$, and $s_1 = \sqrt{1+z_1^2}$. Since $s < \rho$, the first step sends z to a negative value z_1. A direct computation gives

$$\frac{G_\rho(z) - z}{z} = \left(1 - \frac{\rho}{s}\right)\left(1 - \frac{\rho}{s_1}\right) - 1 = \frac{\rho(\rho - s - s_1)}{s s_1}. \tag{11}$$

So the sign of $G_\rho(z) - z$ is the sign of $\rho - s - s_1$.

We now compare s_1 with $\rho - s$. Using $z^2 = s^2 - 1$ and $z_1 = z(1 - \rho/s)$ gives the exact identity

$$s_1^2 - (\rho - s)^2 = \frac{\rho(2s - \rho)}{s^2}. \tag{12}$$

If $s < \rho/2$, equivalently $0 < z < a_\rho$, then the right-hand side of Eq. (12) is negative. Since $\rho - s > 0$, we get $s_1 < \rho - s$, hence $\rho - s - s_1 > 0$. By Eq. (11),

$$0 < z < a_\rho \quad \Longrightarrow \quad G_\rho(z) > z.$$

If $s > \rho/2$, equivalently $a_\rho < z < r_\rho$, then the right-hand side of Eq. (12) is positive. If $\rho - s \geq 0$, this implies $s_1 > \rho - s$. If $\rho - s < 0$, then $\rho - s - s_1 < 0$ is immediate because $s_1 > 0$. Therefore

$$a_\rho < z < r_\rho \quad \Longrightarrow \quad G_\rho(z) < z.$$

Finally, $G_\rho(a_\rho) = a_\rho$ because Eq. (10) holds.

We can now conclude the monotone convergence. If $0 < z_0 < a_\rho$, then $z_{2t} = G_\rho^t(z_0)$ is increasing and bounded above by a_ρ. If $a_\rho < z_0 < r_\rho$, then z_{2t} is decreasing and bounded below by a_ρ. In either case z_{2t} converges to some limit $\ell \in (0, r_\rho)$. Passing to the limit in $z_{2t+2} = G_\rho(z_{2t})$ gives $G_\rho(\ell) = \ell$. The sign characterization above shows that the only fixed point of G_ρ in $(0, r_\rho)$ is a_ρ, so $\ell = a_\rho$. This proves the even-subsequence claim. The odd-subsequence claim follows from continuity:

$$z_{2t+1} = F_\rho(z_{2t}) \to F_\rho(a_\rho) = -a_\rho.$$

The negative-initialization case follows by odd symmetry. This completes the proof. ∎

The only obstruction to convergence to the period-two orbit is the exact-hit set

$$\mathcal{E}_\rho := \{z_0 \in \mathbb{R} : \text{ there exists } n \geq 0 \text{ such that } F_\rho^n(z_0) = 0\}.$$

This set contains 0 and $\pm r_\rho$, and more generally all backward iterates of those points.

Theorem 6 (Global classification of `NormGD`) *Fix $\rho > 2$.*

(i) If $z_0 \in \mathcal{E}_\rho$, then z_t reaches 0 in finitely many steps and remains there afterwards.

(ii) If $z_0 \notin \mathcal{E}_\rho$, then the orbit enters $\mathcal{I}_\rho$ in finitely many steps and converges to the explicit period-two orbit $\{\pm a_\rho\}$.

Consequently, if $z_0 \notin \mathcal{E}_\rho$ *then*

$$y_t = y_0 + \frac{2b}{L}t + o(t),$$

whereas if $z_0 \in \mathcal{E}_\rho$ *then eventually* $z_t = 0$ *and therefore*

$$y_t = y_0 + \eta t + o(t).$$

Proof of Theorem 6. Statement (i) is immediate from the definition of $\mathcal{E}_\rho$.

For (ii), let $z_0 \notin \mathcal{E}_\rho$. Then $z_0 \neq 0$ by definition of $\mathcal{E}_\rho$. By Theorem 3, the orbit enters the closed strip $[-r_\rho, r_\rho]$ after finitely many steps. It cannot land on the boundary $\pm r_\rho$, because either boundary point maps to 0 at the next step. Therefore the orbit enters the open strip $\mathcal{I}_\rho$ in finite time. Theorem 5 then gives convergence to the explicit period-two orbit.

It remains to translate the sharp-coordinate classification into flat speed. In the non-exceptional case, Theorem 5 implies $z_t \to \pm a_\rho$ along the two parities, so Eq. (7) gives

$$y_{t+1} - y_t \to \frac{\eta}{\sqrt{1 + a_\rho^2}}.$$

Using $1 + a_\rho^2 = \rho^2/4$, we obtain

$$\frac{\eta}{\sqrt{1 + a_\rho^2}} = \frac{2\eta}{\rho} = \frac{2b}{L}.$$

Thus $y_t = y_0 + (2b/L)t + o(t)$. If $z_0 \in \mathcal{E}_\rho$, the orbit is eventually at 0, and then Eq. (7) gives $y_{t+1} - y_t = \eta$ from that point onward. This completes the proof. ■

Remark 7 (What the `NormGD` theorem means) *The theorem above is the exact end-to-end law for* `NormGD` *on the linear-flat model. Below the threshold* $\eta = 2b/L$*, every nonzero sharp error decays to zero and the method eventually spends essentially the entire unit step on the flat direction. Above the threshold, every non-exceptional trajectory is forced into a sharp period-two oscillation whose amplitude is completely explicit. In that oscillatory regime the asymptotic flat speed is no longer* η *but the capped value* $2b/L$*.*

C.3. Actual EMA momentum

We now return to the exact EMA momentum system in Eqs. (8) and (9). Unlike `NormGD`, the exact momentum dynamics is genuinely higher-dimensional, so we do not attempt a complete pointwise classification of its global attractors. Instead, the analysis isolates two statements that are enough for the mechanism in the main text. First, the fixed point near the river is locally stable in a strictly larger sharpness window. Second, even without knowing the global attractor, the long-run Cesàro flat progress admits a strict lower bound.

Two exact identities. The flat-momentum component is explicit under standard initialization.

Proposition 8 (The flat momentum is deterministic) *Assume* $q_0 = 0$*. Then for every* $t \geq 0$*,*

$$q_t = 1 - \beta^t.$$

In particular, $q_t \uparrow 1$ *monotonically.*

Proof of Theorem 8. The recursion for q_t in Eq. (8) is scalar and autonomous:

$$q_{t+1} = \beta q_t + (1-\beta), \qquad q_0 = 0.$$

A direct induction gives $q_t = 1 - \beta^t$. ■

C.4. Local stability of `NormGD-M`

This subsection isolates the first of the two statements promised in Section C.3: the sharp-equilibrium of the exact EMA momentum map is locally asymptotically stable in a strictly larger sharpness window than `NormGD`. The exact momentum dynamics has the equilibrium $E = (z, p, q) = (0, 0, 1)$. Indeed, when $z = 0$ and $p = 0$, the sharp coordinate is already at its optimum, while $q = 1$ means that the flat momentum has matched the constant flat gradient.

Theorem 9 (Strict local sharp-stability threshold for exact EMA momentum) *Define the local momentum threshold by*

$$\rho_M^\star := \frac{2(1+\beta)}{1-\beta}.$$

The equilibrium $E = (0, 0, 1)$ of the exact momentum map is locally asymptotically stable whenever $\rho < \rho_M^\star$. Equivalently, the learning rate satisfies $\eta < (2b/L)(1+\beta)/(1-\beta)$. In this regime, under the standard initialization $m_0 = 0$, there exists $\delta_M > 0$ such that every trajectory with $|z_0| < \delta_M$ satisfies

$$z_t \to 0, \qquad p_t \to 0, \qquad q_t \to 1,$$
$$x_t \to 0, \qquad y_t = y_0 + \eta t + o(t).$$

If $\rho > \rho_M^\star$, then E is unstable. For fixed $\rho > 2$, the strict stability condition is equivalent to $\beta > \beta_c(\rho) := (\rho - 2)/(\rho + 2)$. No conclusion is claimed at the boundary $\rho = \rho_M^\star$.

Proof of Theorem 9. We first analyze the linearization at E, and then translate the local basin statement to the standard initialization $m_0 = 0$. View Eqs. (8) and (9) as a C^1 map on $(z, p, q) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$. Linearizing at E, the normalization denominator equals 1 to first order, so the Jacobian is

$$J = \begin{pmatrix} 1-\rho(1-\beta) & -\rho\beta & 0 \\ 1-\beta & \beta & 0 \\ 0 & 0 & \beta \end{pmatrix}.$$

The third eigenvalue is $\beta \in (0, 1)$. The remaining two are the roots of

$$\lambda^2 - \tau\lambda + \beta = 0, \qquad \tau := 1 + \beta - \rho(1-\beta).$$

For a quadratic polynomial $\lambda^2 + a_1\lambda + a_0$, the Jury criterion states that both roots lie in the open unit disk if and only if

$$1 + a_1 + a_0 > 0, \qquad 1 - a_1 + a_0 > 0, \qquad 1 - a_0 > 0.$$

Here $a_1 = -\tau$ and $a_0 = \beta$, so the three conditions are $\rho(1-\beta) > 0$, $2 + 2\beta - \rho(1-\beta) > 0$, and $1 - \beta > 0$. The first and third always hold because $\rho > 0$ and $\beta \in (0, 1)$. The second is

exactly $\rho < 2(1+\beta)/(1-\beta)$. Thus all eigenvalues of J lie in the open unit disk when $\rho < \rho_M^\star$, while at least one eigenvalue lies outside the closed unit disk when $\rho > \rho_M^\star$. The standard C^1 discrete-time linearization criterion gives local asymptotic stability in the first case and instability in the second. At equality, the quadratic factor has the eigenvalue -1, so this argument gives no nonlinear conclusion.

It remains to translate local state stability into the standard-initialization statement. Assume $\rho < \rho_M^\star$, and let U be an open neighborhood of E contained in its local basin of attraction. By Theorem 8, the trajectory with $z_0 = 0$ and standard initialization is exactly

$$(z_t, p_t, q_t) = (0, 0, 1-\beta^t).$$

Hence for some large N we have $(0, 0, 1-\beta^N) \in U$. Let $\Phi_N(z_0)$ denote the state (z_N, p_N, q_N) obtained after N steps from the initial sharp value z_0 with standard initialization. The map Φ_N is continuous and satisfies

$$\Phi_N(0) = (0, 0, 1-\beta^N) \in U.$$

Therefore there exists $\delta_M > 0$ such that $\Phi_N(z_0) \in U$ whenever $|z_0| < \delta_M$. Every such trajectory then converges to E.

Since $z_t \to 0$ and $z_t = Lx_t/b$, we also have $x_t \to 0$. Finally, $p_t \to 0$ and $q_t \to 1$, so Eq. (9) implies

$$y_{t+1} - y_t = \eta \frac{q_{t+1}}{\sqrt{p_{t+1}^2 + q_{t+1}^2}} \to \eta.$$

Summing the increments gives $y_t = y_0 + \eta t + o(t)$. The equivalence with $\beta > \beta_c(\rho)$ follows by rearranging $\rho < 2(1+\beta)/(1-\beta)$ for fixed $\rho > 2$. This completes the proof. ■

Remark 10 (Why the momentum theorem is local) *The threshold $\eta_M^\star = (2b/L)(1+\beta)/(1-\beta)$ is a theorem about the local basin of the exact momentum fixed point. It is* not *a complete global classification of the momentum map for arbitrary sharp initialization. That distinction matters because the exact momentum dynamics is higher-dimensional, and outside the local basin it can exhibit behavior that does not reduce to a single scalar recursion. The present note therefore focuses on the regime that matches the optimization picture of interest, namely small sharp error and persistent flat slope.*

C.5. Long-run flat progress of `NormGD-M`

This subsection turns to the second statement promised in Section C.3: even without classifying the global attractor, the long-run Cesàro flat progress of exact EMA momentum admits a strict lower bound.

Global boundedness of the exact coupled sharp dynamics. The next theorem is the only global structural result on the exact coupled momentum system Eqs. (8) and (9) that the Cesàro analysis below requires. For every standard initialization, the sharp coordinate z_t remains trapped in an explicit bounded strip. No attractor classification is claimed.

Theorem 11 (Global boundedness of the exact coupled sharp dynamics) *Assume the standard initialization $p_0 = q_0 = 0$. Define*

$$d_t := z_t - p_t, \qquad D_* := \max\left\{|z_0|, \frac{\rho}{1-\beta}\right\}, \qquad M_* := \max\{|z_0|, \beta D_* + \rho\}.$$

Then for every $t \geq 0$,

$$|d_t| \leq D_*, \qquad |z_t| \leq M_*, \qquad |p_t| \leq M_* + D_*.$$

In particular, every exact momentum trajectory is globally bounded in the sharp variables (z_t, p_t).

Proof of Theorem 11. The proof first controls the lag $d_t = z_t - p_t$ and then uses this control to trap z_t. From $p_{t+1} = \beta p_t + (1-\beta) z_t$ we obtain the exact identity

$$p_{t+1} = z_t - \beta d_t. \tag{13}$$

Using Eq. (9),

$$d_{t+1} = z_{t+1} - p_{t+1} = \beta d_t - \rho \frac{p_{t+1}}{\sqrt{p_{t+1}^2 + q_{t+1}^2}}.$$

Therefore

$$|d_{t+1}| \leq \beta |d_t| + \rho.$$

Since $d_0 = z_0$, induction gives $|d_t| \leq D_*$ for all $t \geq 0$.

We next prove the sharp-coordinate bound by induction. Assume first that $|z_t| > \beta D_* + \rho$. Then by Eq. (13),

$$|p_{t+1} - z_t| = \beta |d_t| \leq \beta D_* < |z_t|,$$

so p_{t+1} has the same sign as z_t. Consequently the update in Eq. (9) moves toward the origin in the sharp coordinate, and because its magnitude is at most ρ we get

$$|z_{t+1}| < |z_t|.$$

Assume instead that $|z_t| \leq \beta D_* + \rho$. If $|z_{t+1}| > \beta D_* + \rho$, then the step must have moved away from the origin, so p_{t+1} has the opposite sign from z_t. By Eq. (13), this is possible only when $|z_t| \leq \beta D_*$. The step size in the sharp coordinate is at most ρ, hence $|z_{t+1}| \leq \beta D_* + \rho$ after all. Thus the interval $[-(\beta D_* + \rho), \beta D_* + \rho]$ is forward absorbing, and outside it the sharp magnitude decreases. Since $|z_0| \leq M_*$ by definition, induction yields $|z_t| \leq M_*$ for all $t \geq 0$.

Finally,

$$|p_t| \leq |z_t| + |d_t| \leq M_* + D_*,$$

which proves the last claim. This completes the proof. ∎

A global Cesàro flat-speed theorem without attractor classification. The next theorem controls downstream progress through Cesàro averages without classifying the exact long-run attractor of the momentum map. It proves that, for every initialization, `NormGD-M` is at least as fast as the `NormGD` benchmark in long-run average flat progress, and for every $\beta > 0$ it is strictly faster in the oscillatory regime $\rho > 2$.

We again assume the standard initialization $p_0 = q_0 = 0$. The averaged proof keeps track of the normalized sharp and flat fractions in each momentum step. Write $F_t = \sqrt{p_{t+1}^2 + q_{t+1}^2}$, $\phi_t = p_{t+1}/F_t$, $\psi_t = q_{t+1}/F_t$, and $v_t = p_{t+1} - p_t$. Then $\phi_t^2 + \psi_t^2 = 1$. In this notation, Eq. (9) becomes $z_{t+1} = z_t - \rho\phi_t$ and $y_{t+1} - y_t = \eta\psi_t$. The scalar sharp-momentum recursion follows directly from the first equation in Eq. (8). Indeed, $(1-\beta)z_t = p_{t+1} - \beta p_t$, so substituting the sharp update into the next momentum update gives

$$p_{t+2} = (1+\beta)p_{t+1} - \beta p_t - (1-\beta)\rho\,\phi_t, \tag{14}$$

$$v_{t+1} = \beta v_t - (1-\beta)\rho\,\phi_t. \tag{15}$$

Also, summing the flat increments gives $(y_N - y_0)/(N\eta) = N^{-1}\sum_{t=0}^{N-1}\psi_t$. The averaged argument starts from Eq. (15).

Proposition 12 (Averaged work and dissipation identities) *For $N \geq 1$, let V_N, C_N, Φ_N, and J_N denote the time averages of v_t^2, p_{t+1}^2/F_t, ϕ_t^2, and $\phi_t v_t$, respectively, over $0 \leq t < N$. Then, as $N \to \infty$,*

$$\frac{1+\beta}{2}\,V_N = (1-\beta)\rho\,C_N + O(N^{-1}), \tag{16}$$

and

$$(1-\beta^2)V_N = (1-\beta)^2\rho^2\,\Phi_N - 2\beta(1-\beta)\rho\,J_N + O(N^{-1}). \tag{17}$$

Proof of Theorem 12. Because $v_t = p_{t+1} - p_t$ and (p_t) is globally bounded by Theorem 11, the sequence (v_t) is globally bounded as well. We first derive the work identity. Multiply Eq. (15) by p_{t+1} and use the identities $2v_{t+1}p_{t+1} = p_{t+2}^2 - p_{t+1}^2 - v_{t+1}^2$ and $2v_t p_{t+1} = p_{t+1}^2 - p_t^2 + v_t^2$. This gives

$$2(1-\beta)\rho\frac{p_{t+1}^2}{F_t} = \beta(p_{t+1}^2 - p_t^2 + v_t^2) - (p_{t+2}^2 - p_{t+1}^2 - v_{t+1}^2).$$

Summing from $t = 0$ to $N-1$ gives

$$\begin{aligned} 2(1-\beta)\rho\sum_{t=0}^{N-1}\frac{p_{t+1}^2}{F_t} &= \beta(p_N^2 - p_0^2) - (p_{N+1}^2 - p_1^2) + \beta\sum_{t=0}^{N-1}v_t^2 + \sum_{t=1}^{N}v_t^2 \\ &= (1+\beta)\sum_{t=0}^{N-1}v_t^2 + \beta p_N^2 - p_{N+1}^2 + v_N^2, \end{aligned}$$

because $p_0 = 0$. Dividing by N yields Eq. (16), since the boundary term is $O(N^{-1})$. For the dissipation identity, square Eq. (15), sum from $t = 0$ to $N-1$, rearrange, and divide by N. This yields Eq. (17), again because the boundary term $(v_N^2 - \beta^2 v_0^2)/N$ is $O(N^{-1})$. This completes the proof. ∎

To obtain a strict averaged speed gain in the oscillatory regime $\rho > 2$, we need one more ingredient: an eventual strip bound on the transverse momentum itself.

Proposition 13 (Uniform tail confinement of the transverse momentum) *For every exact momentum trajectory with standard initialization,* $\limsup_{t\to\infty} |p_t| \le 2\rho/(1-\beta)$. *More precisely, for every* $\varepsilon > 0$ *there exists* T_ε *such that* $|p_t| \le 2\rho/(1-\beta) + \varepsilon$ *for all* $t \ge T_\varepsilon$.

Proof of Theorem 13. Recall from Theorem 11 that $d_t := z_t - p_t$ satisfies $d_{t+1} = \beta d_t - \rho\phi_t$. Thus $|d_{t+1}| \le \beta|d_t| + \rho$, and iteration gives $|d_t| \le \beta^t|d_0| + \rho(1-\beta^t)/(1-\beta)$. Hence $\limsup_{t\to\infty} |d_t| \le \rho/(1-\beta)$.

Fix $\varepsilon > 0$. Choose $T \ge 0$ such that $|d_t| \le D_\varepsilon := \rho/(1-\beta) + \varepsilon/(2\beta)$ for all $t \ge T$. Set $A_\varepsilon := \beta D_\varepsilon + \rho$. If $t \ge T$ and $|z_t| > A_\varepsilon$, then using $p_{t+1} = z_t - \beta d_t$ from Eq. (13), we have $|p_{t+1} - z_t| = \beta|d_t| \le \beta D_\varepsilon < |z_t|$. Thus p_{t+1} has the same sign as z_t, and $|p_{t+1}| \ge |z_t| - \beta D_\varepsilon > \rho$. The sharp update then implies

$$|z_{t+1}| = |z_t| - \rho\frac{|p_{t+1}|}{\sqrt{p_{t+1}^2 + q_{t+1}^2}} \le |z_t| - \rho\frac{\rho}{\sqrt{\rho^2+1}}.$$

Thus the sharp magnitude decreases by a fixed positive amount whenever $|z_t| > A_\varepsilon$. Therefore after finitely many steps the orbit enters $[-A_\varepsilon, A_\varepsilon]$.

Once the orbit has entered $[-A_\varepsilon, A_\varepsilon]$, it remains forever in the slightly larger interval $[-A_\varepsilon - \rho, A_\varepsilon + \rho]$. Indeed, if $|z_t| \le A_\varepsilon$, then the update size satisfies $|z_{t+1} - z_t| \le \rho$, so $|z_{t+1}| \le A_\varepsilon + \rho$. If instead $A_\varepsilon < |z_t| \le A_\varepsilon + \rho$, then p_{t+1} still has the same sign as z_t, so $|z_{t+1}| < |z_t| \le A_\varepsilon + \rho$. Hence there exists $T'_\varepsilon \ge T$ such that $|z_t| \le A_\varepsilon + \rho$ for all $t \ge T'_\varepsilon$.

Finally, for $t \ge T'_\varepsilon$ we use Eq. (13) to obtain

$$\begin{aligned} |p_{t+1}| &\le |z_t| + \beta|d_t| \le (A_\varepsilon + \rho) + \beta D_\varepsilon \\ &= 2\beta D_\varepsilon + 2\rho = \frac{2\rho}{1-\beta} + \varepsilon. \end{aligned}$$

Thus the claimed eventual bound on $|p_t|$ holds after an index shift. Letting $\varepsilon \downarrow 0$ proves the limsup bound. ∎

The tail strip from Theorem 13 gives uniform strong convexity for the auxiliary functions used below. This convexity is the only extra input needed for a strict averaged speed gain.

Lemma 14 (Strong-convexity mixed-term bound) *Assume* $M > 2\rho/(1-\beta)$. *Put* $m_M = [4(1+M^2)^{3/2}]^{-1}$. *Then, as* $N \to \infty$,

$$J_N \ge \frac{m_M}{2} V_N + O(N^{-1}). \tag{18}$$

Proof of Theorem 14. The proof turns the eventual momentum strip into a uniform convexity estimate. By Theorem 13, there exists $T_M \ge 0$ such that both $|p_t| \le M$ and $|p_{t+1}| \le M$ for all $t \ge T_M$. By Theorem 8, after increasing T_M if necessary, we may also assume $q_{t+1} \ge 1/2$ for all $t \ge T_M$. For $t \ge T_M$, let $G_t(u) := \sqrt{u^2 + q_{t+1}^2}$. Then $G_t''(u) = q_{t+1}^2/(u^2+q_{t+1}^2)^{3/2} \ge m_M$ for all $|u| \le M$. Hence G_t is m_M-strongly convex on $[-M, M]$, and therefore

$$G_t(p_{t+1}) - G_t(p_t) \le G_t'(p_{t+1})(p_{t+1} - p_t) - \frac{m_M}{2}(p_{t+1} - p_t)^2.$$

Because $G_t'(p_{t+1}) = \phi_t$ and $p_{t+1} - p_t = v_t$, this gives

$$\phi_t v_t \geq G_t(p_{t+1}) - G_t(p_t) + \frac{m_M}{2} v_t^2 \qquad \text{for all } t \geq T_M.$$

Summing from $t = T_M$ to $N-1$ yields

$$\sum_{t=T_M}^{N-1} \phi_t v_t \geq \frac{m_M}{2} \sum_{t=T_M}^{N-1} v_t^2 + \sum_{t=T_M}^{N-1} \big(G_t(p_{t+1}) - G_t(p_t)\big).$$

The final sum is bounded from below by a finite constant independent of N. Indeed, since q_t is increasing by Theorem 8, one has

$$\begin{aligned}\sum_{t=T_M}^{N-1} \big(G_t(p_{t+1}) - G_t(p_t)\big) &= G_{N-1}(p_N) - G_{T_M}(p_{T_M}) + \sum_{t=T_M+1}^{N-1} \big(G_{t-1}(p_t) - G_t(p_t)\big) \\ &\geq -G_{T_M}(p_{T_M}) - \sum_{t=T_M+1}^{\infty} (q_{t+1} - q_t).\end{aligned}$$

Adding the finitely many terms with $t < T_M$ changes the average only by $O(N^{-1})$. Dividing by N gives $J_N \geq m_M V_N/2 + O(N^{-1})$, which is Eq. (18). This completes the proof. ∎

We can now turn the strict lower bound on J_N into a strict lower bound on the averaged flat speed itself.

Theorem 15 (Global strict Cesàro flat-speed gain for exact momentum) *Fix $\rho > 0$, $\beta \in (0,1)$, and let $M > 2\rho/(1-\beta)$. Set $m_M := [4(1+M^2)^{3/2}]^{-1}$ and $\kappa_M := (1+\beta)/(1+\beta+\beta\rho m_M)$. Then $\kappa_M \in (0,1)$ and every exact momentum trajectory with standard initialization satisfies*

$$\liminf_{N\to\infty} \frac{y_N - y_0}{N} \geq \eta \min\left\{1, \frac{2}{\rho\kappa_M}\right\}. \tag{19}$$

The explicit bound is obtained from the limiting tail-strip constant

$$m_{\rho,\beta} := \frac{1}{4\left(1 + \dfrac{4\rho^2}{(1-\beta)^2}\right)^{3/2}}, \qquad \kappa_{\rho,\beta} := \frac{1+\beta}{1+\beta+\beta\rho\, m_{\rho,\beta}}. \tag{20}$$

With this notation, $\kappa_{\rho,\beta} \in (0,1)$ and every exact momentum trajectory with standard initialization also satisfies

$$\liminf_{N\to\infty} \frac{y_N - y_0}{N} \geq \eta \min\left\{1, \frac{2}{\rho\kappa_{\rho,\beta}}\right\}. \tag{21}$$

If $\rho > 2$, then this bound is strictly larger than the non-exceptional `NormGD` *flat speed:*

$$\liminf_{N\to\infty} \frac{y_N - y_0}{N} > \frac{2b}{L}. \tag{22}$$

Moreover, if `NormGD` *and exact momentum start from the same initial state (x_0, y_0), the momentum method uses standard initialization, and $z_0 = Lx_0/b \notin \mathcal{E}_\rho$, then*

$$\liminf_{N\to\infty} \frac{y_N^{\mathrm{M}} - y_N^{\mathrm{G}}}{N} \geq \eta \min\left\{1, \frac{2}{\rho\kappa_{\rho,\beta}}\right\} - \frac{2b}{L} > 0. \tag{23}$$

Proof of Theorem 15. The proof first improves the averaged dissipation inequality, then converts the resulting transverse bound into a flat-speed bound. By Theorem 14, the averaged dissipation identity Eq. (17) gives

$$(1-\beta^2)V_N \le (1-\beta)^2\rho^2\,\Phi_N - \beta(1-\beta)\rho\, m_M V_N + O(N^{-1}).$$

Rearranging this estimate gives a bound on V_N in terms of Φ_N. Substituting that bound into Eq. (16) yields

$$C_N \le \frac{\rho(1+\beta)}{2(1+\beta+\beta\rho m_M)}\,\Phi_N + O(N^{-1}) = \frac{\rho\kappa_M}{2}\,\Phi_N + O(N^{-1}).$$

We now convert C_N into a lower bound on the averaged flat speed. Let $S_N = (y_N - y_0)/(N\eta)$ denote the averaged normalized flat increment. Let $\bar{q}_N$, T_N, R_N, and Q_N denote the time averages of q_{t+1}, ψ_t^2, F_t, and $q_{t+1}\psi_t$, respectively. Since $p_{t+1}^2 + q_{t+1}^2 = F_t^2$, Cauchy–Schwarz gives $C_N = R_N - Q_N$ and $\bar{q}_N^2 \le R_N Q_N$. The same definitions give $0 < Q_N \le S_N$. Therefore $C_N \ge \bar{q}_N^2/S_N - S_N$. Since $\phi_t^2 + \psi_t^2 = 1$ and $u \mapsto u^2$ is convex on $[0,1]$, we also have $\Phi_N = 1 - T_N \le 1 - S_N^2$. Combining these two estimates with the upper bound on C_N gives

$$\frac{\bar{q}_N^2}{S_N} - S_N \le C_N \le \frac{\rho\kappa_M}{2}(1 - S_N^2) + O(N^{-1}).$$

Choose a subsequence $N_k \to \infty$ such that $S_{N_k} \to s_* := \liminf_{N\to\infty} S_N$. Because $\bar{q}_{N_k} \to 1$, the displayed inequality also rules out $s_* = 0$. Passing to the limit therefore gives either $s_* = 1$, or else $1/s_* - s_* \le \rho\kappa_M(1 - s_*^2)/2$. If $s_* < 1$, dividing by $1 - s_*^2 > 0$ yields $s_* \ge 2/(\rho\kappa_M)$. Thus $s_* = \liminf_{N\to\infty}(y_N - y_0)/(N\eta) \ge \min\{1, 2/(\rho\kappa_M)\}$, which is exactly Eq. (19).

The formula Eq. (20) is exactly the limit of m_M as $M \downarrow 2\rho/(1-\beta)$. Hence the corresponding κ_M decreases to $\kappa_{\rho,\beta}$. Since the first bound holds for every admissible M, inequality Eq. (21) follows by taking the supremum of the resulting lower bounds. The strict inequality Eq. (22) is immediate for $\rho > 2$, because $\kappa_{\rho,\beta} < 1$ implies $\min\{1, 2/(\rho\kappa_{\rho,\beta})\} > 2/\rho$. By Theorem 6, the non-exceptional `NormGD` trajectory satisfies $y_N^{\mathrm{G}} = y_0 + (2b/L)N + o(N)$. Subtracting the two statements yields Eq. (23). This completes the proof. ∎

The explicit bound in Theorem 15 is intentionally rough. Its role is not to identify the exact oscillatory speed law of momentum, which still depends on the unresolved global attractor classification, but to prove a *strict* global speed advantage in long-run average flat progress.

Remark 16 (What the new global theorem does and does not use) *This result does* not *require an attractor classification of the exact coupled momentum dynamics. The proof uses only three ingredients: the scalar inertial recurrence Eq. (14), convexity of* $u \mapsto \sqrt{u^2 + q_{t+1}^2}$, *and the eventual strip bound from Theorem 13. The result is therefore orthogonal to the still-open global attractor-classification problem. An exact pointwise asymptotic speed law would require finer late-time information on the transverse geometry.*

Appendix D. Analysis of `SignGD` on the rotated quadratic loss

This section records a sign-oracle analogue of the linear-flat mechanism. Unlike the normalized analysis above, the coordinate-wise sign oracle is not rotation-invariant. A 45° rotation therefore exposes a particularly clean failure mode for plain `SignGD` and a corresponding stabilizing role for `Signum` (`SignGD` with EMA momentum).

D.1. Rotated coordinates and sign structure

Let (u, v) denote river coordinates for the loss $f(u, v) = Lu^2/2 - bv$, where $L > 0$ and $b > 0$. The sharp coordinate is u and the river coordinate is v. Introduce parameter coordinates (x, y) by the $45°$ rotation

$$u = \frac{x+y}{\sqrt{2}}, \qquad v = \frac{y-x}{\sqrt{2}}.$$

In parameter coordinates, the same loss is $f(x, y) = L(x+y)^2/4 - b(y-x)/\sqrt{2}$. We write $z_t := Lu_t/b$ for the dimensionless sharp coordinate and $\gamma := \sqrt{2}\eta L/b$ for the dimensionless stepsize. A direct calculation gives

$$\nabla_{x,y} f(x, y) = \frac{b}{\sqrt{2}}(z+1, z-1). \tag{24}$$

Thus the coordinate-wise signs are determined by $z+1$ and $z-1$. We use the convention $\operatorname{sgn}(0) = 0$.

D.2. Plain `SignGD`: exact transverse dynamics

Plain coordinate-wise `SignGD` is $(x_{t+1}, y_{t+1}) = (x_t, y_t) - \eta \operatorname{sgn}(\nabla f(x_t, y_t))$, where the sign is applied coordinate-wise.

Proposition 17 (Plain `SignGD` dynamics) *Away from the switching points $z_t = \pm 1$, plain `SignGD` satisfies*

$$z_{t+1} = z_t - \gamma \operatorname{sgn}(z_t)\, \mathbf{1}_{\{|z_t|>1\}}, \tag{25}$$

$$v_{t+1} - v_t = \sqrt{2}\eta\, \mathbf{1}_{\{|z_t|<1\}}. \tag{26}$$

Consequently, a plain `SignGD` step is either purely transverse, when $|z_t| > 1$, or purely downstream, when $|z_t| < 1$.

Proof of Theorem 17. From Eq. (24), the coordinate sign vector is $(\operatorname{sgn}(z_t+1), \operatorname{sgn}(z_t-1))$. If $z_t > 1$, this sign vector is $(1, 1)$, hence $\Delta u_t = (\Delta x_t + \Delta y_t)/\sqrt{2} = -\sqrt{2}\eta$ and $\Delta v_t = (\Delta y_t - \Delta x_t)/\sqrt{2} = 0$. Therefore $z_{t+1} = z_t - \gamma$ and $v_{t+1} - v_t = 0$. If $z_t < -1$, the sign vector is $(-1, -1)$, hence $\Delta u_t = \sqrt{2}\eta$ and $\Delta v_t = 0$, giving $z_{t+1} = z_t + \gamma$ and again zero downstream progress. If $-1 < z_t < 1$, the sign vector is $(1, -1)$, hence $\Delta u_t = 0$ and $\Delta v_t = \sqrt{2}\eta$. Combining the three cases gives Eq. (25)–Eq. (26). This completes the proof. ∎

Theorem 18 (A zero-progress transverse two-cycle for plain `SignGD`) *Assume $\gamma > 2$ and initialize $z_0 = A := \gamma/2$. Then plain `SignGD` satisfies $z_t = (-1)^t A$ and $v_t = v_0$ for every $t \geq 0$. Thus this initialization lies on a transverse period-two orbit with zero downstream progress.*

Proof of Theorem 18. Since $\gamma > 2$, one has $A > 1$. If $z_t = A$, then Eq. (25) gives the identity $z_{t+1} = A - \gamma = -A$. If $z_t = -A$, then Eq. (25) gives the identity $z_{t+1} = -A + \gamma = A$. This proves the claimed two-cycle. Since $|z_t| = A > 1$ for every t, Eq. (26) gives $v_{t+1} - v_t = 0$ for every t, proving the zero-progress claim. This completes the proof. ∎

D.3. `Signum`: `SignGD` with EMA momentum

We now analyze the actual momentum method, not an attenuated oracle. Define

$$m_{t+1} = \beta m_t + (1-\beta)\nabla f(x_t, y_t), \qquad \beta \in (0,1), \tag{27}$$

with standard initialization $m_0 = 0$, and update by the coordinate-wise sign of the momentum,

$$(x_{t+1}, y_{t+1}) = (x_t, y_t) - \eta \operatorname{sgn}(m_{t+1}).$$

Write the momentum in the rotated sign coordinates as

$$m_t = \frac{b}{\sqrt{2}}(p_t + q_t, p_t - q_t). \tag{28}$$

Proposition 19 (`Signum` dynamics) *With $m_0 = 0$, the momentum coordinates satisfy*

$$p_{t+1} = \beta p_t + (1-\beta) z_t, \qquad q_{t+1} = \beta q_t + (1-\beta), \tag{29}$$

so that $q_t = 1 - \beta^t$. Moreover, away from the switching planes $|p_{t+1}| = q_{t+1}$,

$$z_{t+1} = z_t - \gamma \operatorname{sgn}(p_{t+1})\, \mathbf{1}_{\{|p_{t+1}| > q_{t+1}\}}, \tag{30}$$

$$v_{t+1} - v_t = \sqrt{2}\eta\, \mathbf{1}_{\{|p_{t+1}| < q_{t+1}\}}. \tag{31}$$

Proof of Theorem 19. Combining Eq. (24), Eq. (28), and Eq. (27) gives the two scalar recursions in Eq. (29). Since $q_0 = 0$, the second recursion gives $q_t = 1 - \beta^t$.

The coordinate-wise momentum signs are $(\operatorname{sgn}(p_{t+1} + q_{t+1}), \operatorname{sgn}(p_{t+1} - q_{t+1}))$. If $p_{t+1} > q_{t+1}$, both signs are positive, so the step is purely transverse with $\Delta z_t = -\gamma$ and $\Delta v_t = 0$. If $p_{t+1} < -q_{t+1}$, both signs are negative, so the step is purely transverse with $\Delta z_t = \gamma$ and $\Delta v_t = 0$. If $|p_{t+1}| < q_{t+1}$, the signs are opposite, so $\Delta z_t = 0$ and $\Delta v_t = \sqrt{2}\eta$. These three cases give Eq. (30)–Eq. (31). This completes the proof. ■

The next theorem gives a fully actual-momentum separation result on the same symmetric bad initialization as Theorem 18. The only genericity condition is that the trajectory avoids the switching planes $|p_{t+1}| = q_{t+1}$, where one coordinate of the sign vector is exactly zero.

Theorem 20 (Momentum recovers downstream steps) *Assume $2 < \gamma < 2(1+\beta)/(1-\beta)$. Set $A := \gamma/2$, and initialize $z_0 = A$ and $p_0 = q_0 = 0$. Assume also that the trajectory never hits a switching plane, namely $|p_{t+1}| \neq q_{t+1}$ for every $t \geq 0$. Then the* `Signum` *trajectory satisfies the following.*

(i) $z_t \in \{A, -A\}$ *for every $t \geq 0$.*

(ii) No two consecutive steps are transverse.

(iii) Consequently, for every $N \geq 1$,

$$v_N - v_0 \geq \sqrt{2}\eta \left\lfloor \frac{N}{2} \right\rfloor. \tag{32}$$

Combining with Theorem 18, for the same initialization the flat-progress gap satisfies

$$v_N^{\mathrm{M}} - v_N^{\mathrm{G}} \geq \sqrt{2}\eta \left\lfloor \frac{N}{2} \right\rfloor .$$

In particular,

$$\liminf_{N\to\infty} \frac{v_N^{\mathrm{M}} - v_N^{\mathrm{G}}}{N} \geq \frac{\sqrt{2}\eta}{2} > 0.$$

Proof of Theorem 20. Let $s := 1 - \beta$. The window assumption is equivalent to $1 < A < (1+\beta)/(1-\beta) = (1+\beta)/s$. We prove the theorem by an induction on the step type.

Call a time t calm if $z_t \in \{A, -A\}$ and $|p_t| \leq q_t$. The initial time is calm because $z_0 = A$ and $p_0 = q_0 = 0$.

Suppose first that time t is calm, and write $z_t = \sigma A$ with $\sigma \in \{\pm 1\}$. Then the scalar recursions give $p_{t+1} = \beta p_t + s\sigma A$ and $q_{t+1} = \beta q_t + s$. Because $|p_t| \leq q_t$,

$$\sigma p_{t+1} = \beta\sigma p_t + sA > -\beta q_t - s = -q_{t+1},$$

where the strict inequality uses $sA > -s$. Hence p_{t+1} cannot have the opposite sign with magnitude above threshold. Therefore one of two alternatives holds. The downstream case is $|p_{t+1}| < q_{t+1}$. The transverse case is $\sigma p_{t+1} > q_{t+1}$. The switching equality is excluded by assumption.

If the downstream case holds, then Eq. (30) gives $z_{t+1} = z_t = \sigma A$, and Eq. (31) gives one downstream step. Since $|p_{t+1}| < q_{t+1}$, time $t+1$ is calm again.

If the transverse case holds, then Eq. (30) gives $z_{t+1} = \sigma A - \gamma\sigma = -\sigma A$. Thus the step is transverse and the sign of z flips. We now show that the next step must be downstream. Using the scalar recursions once more,

$$\begin{aligned} p_{t+2} &= \beta p_{t+1} - s\sigma A \\ &= \beta(\beta p_t + s\sigma A) - s\sigma A = \beta^2 p_t - s^2\sigma A. \end{aligned}$$

Also, $q_{t+2} = \beta q_{t+1} + s = \beta^2 q_t + s(1+\beta)$. Therefore, using $|p_t| \leq q_t$ and the window assumption,

$$\begin{aligned} |p_{t+2}| &\leq \beta^2 q_t + s^2 A \\ &< \beta^2 q_t + s(1+\beta) = q_{t+2}. \end{aligned}$$

Thus the step after a transverse step is necessarily downstream. On that step z remains equal to $-\sigma A$, and time $t+2$ is calm.

We have proved that from every calm time either the next step is downstream and returns immediately to a calm time, or the next step is transverse and the following step is downstream and returns to a calm time. By induction, $z_t \in \{A, -A\}$ for all t, and no two consecutive steps are transverse. Hence among the first N steps at least $\lfloor N/2 \rfloor$ are downstream. Each downstream step contributes exactly $\sqrt{2}\eta$ to v, by Eq. (31). This proves Eq. (32). The gap statements follow from the zero-progress claim in Theorem 18. This completes the proof. ∎

Remark 21 (Interpretation) *Plain* `SignGD` *makes the downstream decision from the instantaneous sharp coordinate z_t: it moves downstream only inside the strip $|z_t| < 1$.* `Signum` *makes*

the downstream decision from the averaged sharp signal p_{t+1}: it moves downstream when $|p_{t+1}| < q_{t+1}$. On the symmetric orbit where plain `SignGD` *is trapped in the transverse two-cycle $z_t = \pm\gamma/2$, the EMA buffer prevents consecutive transverse moves under the window in Theorem 20. Thus actual momentum converts a zero-progress transverse oscillation into a sequence with at least one downstream step every two iterations.*