MOMENTUM ACCELERATION OF NORMALIZED STEEPEST DESCENT AT THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimizers based on normalized steepest descent with momentum are widely used to train large language models, yet existing analyses attribute the benefit of momentum to stochastic noise reduction and favor no momentum in deterministic training. We show that full-batch training nevertheless benefits from momentum and explain why. Once training reaches the Edge of Stability, normalization pins the effective learning rate at twice the inverse curvature transverse to the gradient, so progress is limited by the loss geometry rather than by the learning rate. Momentum attenuates the alternating gradient component before normalization, which raises the threshold at which oscillations start and multiplies the progress after it by $(1+\beta)/(1-\beta)$. A switched flow built from this mechanism predicts the onset of oscillations, the speed afterward, and, integrated from initialization, the training loss in full-batch experiments. We prove the mechanism on high-dimensional quadratic valleys, including a strict speedup from almost every common initialization without assuming periodic behavior, and show that on general smooth losses normalized descent follows the flow locally once it oscillates.

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 (Bernstein et al., 2018; Chen et al., 2023a; Jordan et al., 2024; Liu et al., 2025; Pethick et al., 2025). Despite the empirical success, existing theories for normalized steepest descent (NSD) are primarily based on smoothness assumptions and one-step descent lemmas (Hazan et al., 2015; Levy, 2016; Chen et al., 2023b; Pethick et al., 2025; Xie et al., 2025). Within this framework, momentum is typically regarded as a variance-reduction device in stochastic settings (Cutkosky & Orabona, 2019; Cutkosky & Mehta, 2020; Arnold et al., 2019; Qian et al., 2025). However, in deterministic settings, the same analysis implies that momentum introduces an unnecessary lag behind the current gradient, suggesting that momentum may slow convergence.

Outside this framework, the classical deterministic benefit of momentum does not carry over either. On quadratics, heavy-ball momentum accelerates gradient descent by improving its linear convergence rate (Polyak, 1964), but with a fixed learning rate η, normalized descent has no such convergence rate to improve: its updates have a fixed size, so near a minimizer the iterates keep oscillating at the scale of η. Both views suggest that momentum is useless for deterministic normalized descent, yet Figure 1 shows the opposite. In full-batch training of a transformer, momentum lowers the training loss of `NormGD-M`, `Signum`, and `Muon` at the same learning rate. The benefit of momentum must therefore come from an effect that neither view captures, which raises the question:

Why does momentum help normalized optimizers in deterministic training?

Recall that a one-step analysis certifies progress by comparing the update direction with the current gradient, so any difference between the momentum buffer and the gradient reduces the certified progress. Practical finite-step training can undermine the effectiveness of this analysis even in deterministic full-batch dynamics. In particular, neural network training often enters the Edge of Stability (EoS), where the iterates oscillate along sharp directions of the Hessian, whose curvature stays near a stability threshold, while the loss decreases non-monotonically on a slower time scale (Cohen et al., 2021; 2022; Lu et al., 2024; Islamov et al., 2026). Normalized descent enters this regime as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

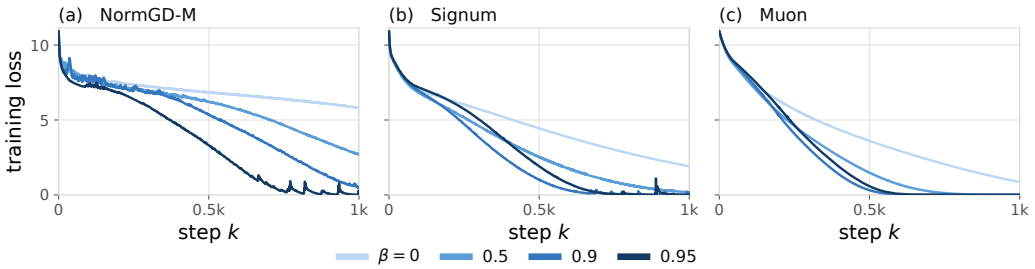


Figure 1: Momentum improves full-batch language-model training. A 14.6M-parameter transformer is trained on 1,024 fixed C4 sequences of length 256, from a common initialization. Learning rates are constant: 0.48, 1.25×10^{-4}, and 3.125×10^{-4} in (a)–(c), respectively; `Signum` uses $\beta_1 = \beta_2 = \beta$. Only (a) is EMA-smoothed (decay 0.3); (b,c) show raw training loss. Details are in Section B.4.

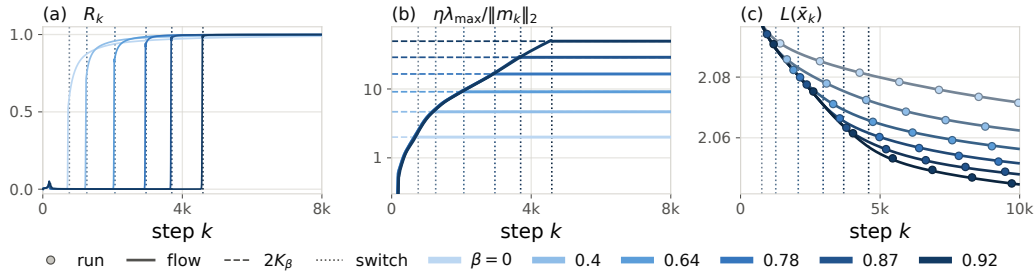


Figure 2: Oscillations start at a *switch* time, after which a flow predicts the training loss. Full-batch `NormGD-M` on the teacher–student task of Section 3, $\eta = 8 \times 10^{-3}$; colors encode β, and the dotted lines mark the switch. (a) Flip ratio $R_k = \|g_k - g_{k-1}\|_2 / (\|g_k\|_2 + \|g_{k-1}\|_2)$. (b) $\eta\lambda_{\max}/\|m_k\|_2$, which locks at $2K_\beta$ (dashed) after the switch, $K_\beta = (1+\beta)/(1-\beta)$. (c) Loss at the de-oscillated center, the mean of the last eight iterates (markers), and the switched flow of Equation (7) integrated from initialization (lines); no parameter is fitted.

well. When the iterates oscillate across a sharp direction, consecutive gradients point to opposite sides of it, and their components along this direction cancel. We measure this cancellation by the flip ratio, the fraction of the gradient that alternates between consecutive updates. In the full-batch runs of Figure 2(a), the flip ratio stays near zero at first, then rises quickly to nearly one and remains there. From then on, most of the difference between the momentum buffer and the current gradient is the alternating part that the buffer averages out, so the one-step analysis no longer describes the progress of the method.

To see when the oscillation sets in, we track the sharpness $\lambda_{\max}$ (top Hessian eigenvalue) times the effective learning rate $\eta/\|m_k\|_2$, where m_k is the momentum buffer, since a gradient step is stable only if this product is below 2. In Figure 2(b), it first grows steadily and then flattens, around the same time as the flip ratio rises. We call this transition the *switch*. The ratio stays flat because the oscillation regulates itself: a growing oscillation enlarges the buffer norm and thus lowers the effective learning rate until the sharp direction is only marginally stable. The plateau is higher for larger β, so a run with more momentum switches later and, after the switch, keeps a larger effective learning rate.

We further distill these observations into a continuous switched flow that describes the trajectory of `NormGD-M`. Before the switch, the flow is the normalized gradient flow, along which `NormGD-M` advances by η per update. After it, the effective learning rate stays at the plateau divided by the sharpness $\lambda_\perp$ in directions orthogonal to the gradient. The plateau of Figure 2(b) is $2K_\beta$, where $K_\beta := (1+\beta)/(1-\beta)$: the switch occurs when $\eta\lambda_\perp/\|\nabla L\|_2$ reaches $2K_\beta$ instead of 2, which prolongs the phase of full speed η, and the speed after it is K_β times that without momentum; note that this speed does not depend on η. Figure 2(c) shows that the flow, integrated from initialization,

accurately predicts the training loss. This answers the question above: momentum helps normalized optimizers in deterministic training through two mechanisms, an extended stability window, which delays the switch, and accelerated progress while the oscillation persists after the switch.

Our main contributions are summarized as follows.

- **Two mechanisms of momentum.** We identify how momentum accelerates deterministic normalized descent: it extends the stability window from 2 to $2K_\beta$, which delays the onset of oscillation and keeps the full speed η for longer, and it multiplies the progress per update by K_β once oscillations persist, a gain that tuning η cannot reproduce (Section 2).
- **A switched flow that predicts training.** We derive a continuous flow with one phase on each side of the switch that captures both mechanisms. It predicts when oscillations start, how fast the iterates move afterward, and the training loss of full-batch runs (Section 3).
- **Provable flow tracking and acceleration.** We prove that on general smooth losses, `NormGD-M` follows the switched flow once it oscillates. On a family of quadratic losses that admits an explicit analysis, we further prove both mechanisms and show that momentum is strictly faster than normalized descent without momentum whenever the latter oscillates (Section 4).

Related work is discussed in Section A, and a sign-based case study is in Section C.

2 A SWITCHED-FLOW MODEL OF `NormGD-M`

This section proposes a switched flow that approximates the trajectory of NSD-M both before and after oscillations set in. We derive it heuristically on a two-dimensional valley (Section 2.2) and then extend it to general losses (Section 2.3).

2.1 NORMALIZED STEEPEST DESCENT WITH MOMENTUM

Let $L\colon \mathbb{R}^d \to \mathbb{R}$ be differentiable and write $g_k = \nabla L(x_k)$. For a norm $\|\cdot\|$ on $\mathbb{R}^d$, the linear minimization oracle (LMO) returns a unit-norm direction most aligned with the negative of its input, $\mathsf{LMO}_{\|\cdot\|}(v) \in \arg\min_{\|u\|\le 1}\langle v, u\rangle$, as in Frank–Wolfe methods and norm-constrained optimizer geometry (Frank & Wolfe, 1956; Jaggi, 2013; Pethick et al., 2025). NSD-M applies the oracle to an exponential moving average of past gradients: with $m_0 = 0$, $\beta \in [0, 1)$, and learning rate $\eta > 0$,

$$m_{k+1} = \beta m_k + (1-\beta) g_k, \qquad x_{k+1} = x_k + \eta\, \mathsf{LMO}_{\|\cdot\|}(m_{k+1}). \tag{1}$$

Setting $\beta = 0$ recovers NSD. The Euclidean norm gives `NormGD` and `NormGD-M`, the ℓ_∞ norm gives `SignGD` and `Signum`, and the spectral norm gives `SpecGD`-type updates (Carlson et al., 2015), including `Muon` (Jordan et al., 2024). Our theory and the experiments of Section 3 use the Euclidean norm, for which $x_{k+1} = x_k - \eta\, m_{k+1}/\|m_{k+1}\|_2$; the sign-based case is studied in Section C.

2.2 AN ILLUSTRATIVE EXAMPLE: A TWO-DIMENSIONAL VALLEY

The original EoS heuristic starts from a one-dimensional positive definite quadratic $L(x) = \frac{1}{2}\lambda x^2$, on which gradient descent is stable exactly when $\eta\lambda < 2$ (Cohen et al., 2021). Normalized descent methods have no such threshold there: their update size is fixed, so near the minimizer they oscillate on the order of η. To help characterize progress while retaining this oscillatory behavior, we extend the quadratic loss by adding a linear term along a flat direction of the quadratic. As an illustrative case, we consider a minimalist example of a two-dimensional *open quadratic*:

$$L(u, v) = \frac{1}{2}\lambda u^2 + bv, \qquad \lambda, b > 0. \tag{2}$$

Intuitively, this depicts a valley that descends with slope b along v and has curvature λ across it. The negative gradient $-\nabla L(u, v) = (-\lambda u, -b)$ pulls the iterate back toward the valley bottom $u = 0$ in the sharp direction (i.e., the u direction) and is constant in the flat direction (i.e., the v direction), and the loss has no minimizer.

Normalization stabilizes the effective learning rate. Under (2), a NormGD update is a gradient step with the *effective learning rate* $\eta_k := \eta/\|\nabla L(u_k, v_k)\|_2$. The two coordinates evolve as

$$u_{k+1} = (1 - \eta_k\lambda)u_k, \qquad v_k - v_{k+1} = \eta_k b. \tag{3}$$

Critically, the effective learning rate and the oscillation in u form a negative feedback loop, since $\|\nabla L(u_k, v_k)\|_2^2 = b^2 + \lambda^2 u_k^2$: if $\eta_k\lambda > 2$, then $|u_{k+1}| > |u_k|$, so the gradient norm increases and η_k decreases; if $\eta_k\lambda < 2$, the opposite happens until η_k reaches its ceiling η/b. Since $\eta_k\lambda$ never exceeds $\eta\lambda/b$, the loop can settle at $\eta_k\lambda = 2$ only when $\eta\lambda/b > 2$, in which case $u_{k+1} = -u_k$ and the iterates oscillate with period two. When $\eta\lambda/b < 2$, the ceiling is stable, the sharp coordinate decays, and the iterates move straight down the valley. In both cases,

$$\eta_k \longrightarrow \min\Big\{\frac{\eta}{b}, \frac{2}{\lambda}\Big\}, \qquad v_k - v_{k+1} \longrightarrow \min\Big\{\eta, \frac{2b}{\lambda}\Big\}. \tag{4}$$

The flat component is essential: it bounds the gradient norm from below and thereby creates the ceiling that separates the two regimes. When $\eta\lambda/b$ exceeds the threshold 2, the progress per update equals $2b/\lambda$ and *no longer depends on* η; a larger learning rate only enlarges the oscillation. With several sharp directions, the largest curvature pins the effective learning rate and the other sharp directions decay; Theorem 4.2 proves this for almost every initialization.

Momentum tames the alternating gradient. Momentum changes this picture through its buffer, which enters the normalization in place of the gradient. For NormGD-M, the effective learning rate is therefore $\eta/\|m_{k+1}\|_2$. Write $m_{k+1} = (p_{k+1}, q_{k+1})$. From $m_0 = 0$, the flat component of the buffer equals $q_{k+1} = (1 - \beta^{k+1})b$ and converges to b. Suppose that, as for NormGD above, the iterates oscillate with period two, $u_{k+1} = -u_k$. The sharp component of the buffer, $p_{k+1} = \beta p_k + (1 - \beta)\lambda u_k$, averages this alternating input, so after a transient of $O(1/(1 - \beta))$ updates it alternates as well, i.e., $p_k = -p_{k+1}$. This implies $(1 + \beta)p_{k+1} = (1 - \beta)\lambda u_k$, and thus

$$m_{k+1} = \Big(\frac{\lambda u_k}{K_\beta}, b\Big), \qquad \text{where } K_\beta := \frac{1 + \beta}{1 - \beta}. \tag{5}$$

In words, the momentum buffer preserves the flat component and divides the alternating one by K_β. With this buffer, the updates take the form of Equation (3) with $\eta_k := \eta/\|m_{k+1}\|_2$ and with λ replaced by λ/K_β. Since straight motion also loses stability through an alternating perturbation, the same replacement in Equation (4) gives

$$\eta_k \longrightarrow \min\left\{\frac{\eta}{b}, \frac{2K_\beta}{\lambda}\right\}, \qquad v_k - v_{k+1} \longrightarrow \min\left\{\eta, \frac{2K_\beta b}{\lambda}\right\}. \tag{6}$$

Comparing (6) with (4), momentum increases the progress per update over NormGD in two ways:

- **Mechanism I: an extended stability window.** Straight motion remains stable up to $\eta\lambda/b = 2K_\beta$ instead of 2, so for $2 < \eta\lambda/b < 2K_\beta$ momentum still advances η per update.
- **Mechanism II: accelerated progress while oscillations persist.** Beyond $2K_\beta$, both NormGD and NormGD-M oscillate, but the latter advances K_β times farther per update because of momentum.

2.3 From open quadratics to general losses

In the previous two-dimensional case, the curvature stays constant throughout the optimization process, so only one of the two mechanisms can be active in a run. For a general loss, the roles of b and λ are played by the gradient norm and the sharpness in directions orthogonal to the gradient,

$$\lambda_\perp(x) := \max\big\{u^\top \nabla^2 L(x) u : \|u\|_2 = 1,\ \langle u, \nabla L(x)\rangle = 0\big\},$$

which may change along the optimization trajectory. If $\|\nabla L\|_2$ and $\lambda_\perp$ change slowly from one update to the next, the iterates behave at each moment as in a valley with slope $\|\nabla L\|_2$ and curvature $\lambda_\perp$. The center of the oscillation then moves along $-\nabla L$ by $\min\{\eta, 2K_\beta\|\nabla L\|_2/\lambda_\perp\}$ per update, as in (4) and (6). Treating the iteration count as a continuous time t gives the *switched flow* $X_\beta(t)$,

$$\frac{\mathrm{d}X_\beta(t)}{\mathrm{d}t} = -\min\left\{\frac{\eta}{\|\nabla L(X_\beta(t))\|_2}, \frac{2K_\beta}{\lambda_\perp(X_\beta(t))}\right\}\nabla L(X_\beta(t)). \tag{7}$$

Where $\nabla L \neq 0$ and $\lambda_\perp > 0$, its phase is selected by the *normalized sharpness*

$$\theta(x) := \frac{\eta\, \lambda_\perp(x)}{\|\nabla L(x)\|_2}. \tag{8}$$

Where $\theta < 2K_\beta$, the flow is in its *stable phase*, the normalized gradient flow at speed η; where $\theta > 2K_\beta$, it is in its *oscillatory phase*, and its speed $2K_\beta\|\nabla L\|_2/\lambda_\perp$ no longer depends on η. Unlike in the valley, θ varies along the path, and the flow switches phase whenever θ crosses $2K_\beta$, in either direction. In particular, for $\beta = 0$, the oscillatory phase follows the curvature-adaptive flow

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = -\frac{2}{\lambda_\perp(X(t))}\nabla L(X(t)). \tag{9}$$

Note that all these flows are time changes of the gradient flow, so they trace the same path and differ only in their speed; in the oscillatory phase, $X_\beta(t) = X(K_\beta t)$ for a common starting point.

Remark 2.1 (Relation to the central flow). The switched flow is closely related to the central flow of Cohen et al. (2025), which describes the time-averaged trajectory of an optimizer at the edge of stability. For $\beta = 0$, both flows follow a stable flow below the threshold and pin the effective learning rate at $2/\lambda$ above it; indeed, Equation (9) is the central flow for scalar RMSProp in Cohen et al. (2025, Section 4.3, Figure 16) with $\lambda_{\max}$ replaced by $\lambda_\perp$ and its sharpness-reduction term removed. Our flow incorporates momentum, which enters only through K_β and changes both the threshold and the speed of the oscillatory phase. It also omits the sharpness-reduction term, which steers the trajectory toward flatter regions; Remark 2.2 adds its counterpart.

Remark 2.2 (Finite learning rates). At finite learning rates, the oscillation is also affected by how the curvature varies across the valley, and the switched flow acquires a sharpness term (Section B.2):

$$\frac{\mathrm{d}X_\beta}{\mathrm{d}t} = -\min\left\{\frac{\eta}{\|\nabla L\|_2}, \frac{2K_\beta}{\lambda}\right\}\nabla L - \frac{K_\beta a^2}{\lambda}\nabla\lambda, \tag{10}$$

$$a^2 = \left[\frac{\eta^2}{4} - \frac{K_\beta^2}{\lambda^2}\left\|\nabla L + \frac{a^2}{2}\nabla\lambda\right\|_2^2\right]_+,$$

where λ is the top eigenvalue of the Hessian at $X_\beta(t)$ and a is the amplitude of the oscillation, which vanishes in the stable phase; while $a > 0$, the minimum is replaced by $2K_\beta/\lambda$ (Section B.2). The sharpness term pushes the trajectory toward flatter regions, as the sharpness-reduction term of Cohen et al. (2025) does; it is of higher order in η and vanishes on open quadratics, so the effect of K_β and the results of Section 4 do not rely on it. With it, the flows for different β and η no longer share one path, so statements that rest on a common path, such as that tuning η cannot substitute for momentum, hold where the term is negligible.

3 The switched flow in deterministic training of NNs

We empirically validate the switched flow of Equation (7) in neural network training. Specifically, the theory developed in the previous section yields three predictions:

Prediction I. *Switch.* When the normalized sharpness θ_k is below $2K_\beta$, the iterates do not oscillate and follow the normalized gradient flow, advancing by η per update. In this stage, iterations with different β are predicted to move along the same path. Along this path, it is expected that the normalized sharpness θ_k grows gradually, and therefore momentum postpones the switch.

Prediction II. *Acceleration in oscillation.* After the switch, the effective learning rate locks at $\eta/\|m_k\|_2 = 2K_\beta/\lambda_{\max}$ due to oscillation. Accordingly, the center advances by $\eta\min\{1, 2K_\beta/\theta_k\}$ per update. From the same checkpoint, `NormGD-M` is therefore K_β times faster than `NormGD` whenever $\theta_k > 2K_\beta$.

Prediction III. *Loss prediction.* Integrated from the same initialization, the switched flow with the same β predicts the loss at the de-oscillated center, including where the switch occurs.

We test these predictions in Sections 3.1 to 3.3. We train a teacher–student MLP on a softmax cross-entropy loss by full-batch `NormGD-M`, over a grid of momentum parameters and learning rates. See Section B.1 for details of the setting. We evaluate the normalized sharpness of Equation (8) at the de-oscillated center $\bar{x}_k$, the mean of the last eight iterates, $\theta_k := \theta(\bar{x}_k)$. We use $\lambda_{\max}$ in place of $\lambda_\perp$ (Section B.1).

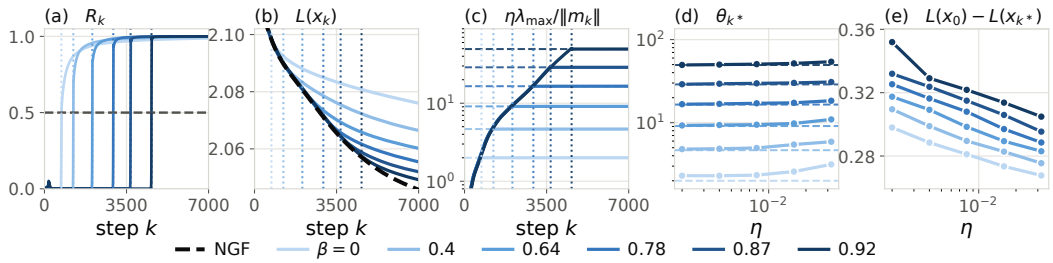


Figure 3: Momentum postpones the switch. (a)–(c) At $\eta = 8 \times 10^{-3}$, with the switch dotted: (a) flip ratio R_k; (b) training loss of `NormGD-M`, with the normalized gradient flow (NGF) dashed. (c) $\eta\lambda_{\max}/\|m_k\|_2$, with $2K_\beta$ dashed. (d) θ_{k^*} and (e) loss decrease before the switch against η.

3.1 Pre-switch: extended stable phase from momentum

In the stable phase, the switched flow (7) reduces to the normalized gradient flow. In this way, the effect of momentum does not lie in changing the dynamics in this phase, but in changing the time at which a run exits this phase. To look into this phase transition, we define the flip ratio

$$R_k := \frac{\|g_k - g_{k-1}\|_2}{\|g_k\|_2 + \|g_{k-1}\|_2} \in [0, 1]. \tag{11}$$

It is clear that when the gradient oscillates along the sharp direction, as $u_{k+1} = -u_k$ in Equation (3), it holds that $R_k \approx 1$. In the stable phase, where consecutive gradients align, R_k stays near zero. We define the switch $k^* = \min\{k : R_k > 1/2\}$. The switch marked in this way is exactly where the run departs from the normalized gradient flow: it follows the flow in loss until k^* and falls behind it afterward (Figure 3(b)).

Effective learning rate. We start from the flip ratio, which is measured without the Hessian. After the switch, the gradient alternates, and consecutive iterates differ by about η along the top eigenvector, so $\|g_k - g_{k-1}\|_2 \approx \eta\lambda_{\max}$; the buffer divides the alternating gradient by K_β (Equation (5)), so $\|g_k\|_2 \approx K_\beta\|m_k\|_2$. Accordingly, $R_k \approx \eta\lambda_{\max}/(2K_\beta\|m_k\|_2)$, and $R_k \approx 1$ in Figure 3(a) implies that the effective learning rate $\eta/\|m_k\|_2$ locks at $2K_\beta/\lambda_{\max}$, as predicted by Prediction II. It is clear from Figure 3(c) that $\eta\lambda_{\max}/\|m_k\|_2$ indeed stays at $2K_\beta$ after the switch. Before the switch, $m_k \approx \nabla L(x_k)$, so the same ratio is approximately θ_k, which grows mainly because the gradient norm decreases, not because of progressive sharpening (Cohen et al., 2021); the switch occurs once it reaches $2K_\beta$, as predicted by Prediction I. In Figure 3(d), θ_{k^*} stays close to $2K_\beta$ at small η and exceeds it at larger η, because the oscillation is detected only after it has grown. We remark that R_k reaches $1/2$ slightly after the threshold, at $\theta_k = 2(K_\beta^2 + 1/3)^{1/2}$, which explains the gap for small β at the smallest η (Section B.1).

Loss progress before the switch. As shown by the switched flow and the experiments above, momentum postpones the switch, at which the oscillation starts, and therefore provides an extended window for stable and fast progress along the normalized gradient flow. We visualize this progress for different β and η (Figure 3(e)). At every η, the loss decrease in the stable phase grows with β, as Mechanism I predicts. At every β, it decreases with η, because θ_k reaches $2K_\beta$ earlier along the path when η is larger.

3.2 Post-switch acceleration from momentum and learning-rate cap

In the oscillatory phase, the effective learning rate locks at $\eta/\|m_k\|_2 = 2K_\beta/\lambda_{\max}$ (Figure 3(c)). Training thus selects its own oscillation amplitude, and the curvature caps the speed of the center at $2K_\beta\|\nabla L\|_2/\lambda_{\max}$. By Prediction II, this speed is K_β times that of `NormGD` and independent of the learning rate. Starting from the same initialization, the runs switch at different times and places and then progress from different states, so comparing them at a common iteration would confound speed with position. To test whether momentum still accelerates once the runs oscillate, we therefore take a checkpoint of the run with the largest β once its θ_k reaches 2.5 times its threshold, and continue all `NormGD-M` runs from this state with a zero buffer.

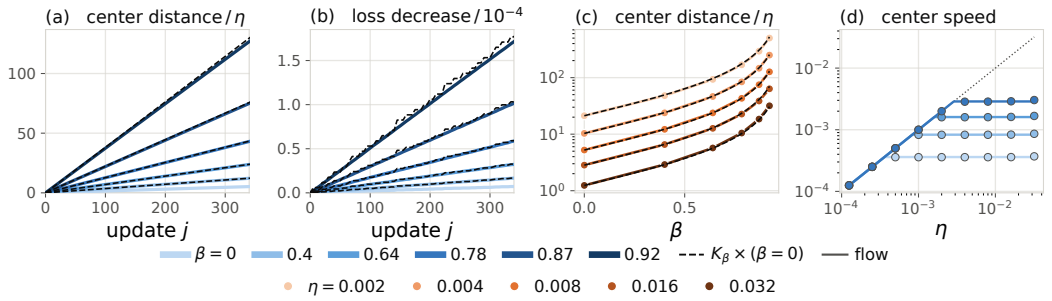


Figure 4: After the switch, momentum multiplies the speed of the center by K_β, and the learning rate does not change it. (a) Distance traveled by the de-oscillated center, in units of η, and (b) loss decrease, at $\eta = 8 \times 10^{-3}$ after a common restart; dashed: K_β times the curve of $\beta = 0$. (c) The distance of (a) at the end of the window against β, for every η; dashed: K_β times the value of $\beta = 0$. (d) Speed of the center against η for $\beta \leq 0.784$, all runs restarted from one state; lines: $\min\{\eta, 2K_\beta\|\nabla L\|_2/\lambda_{\max}\}$; dotted: η.

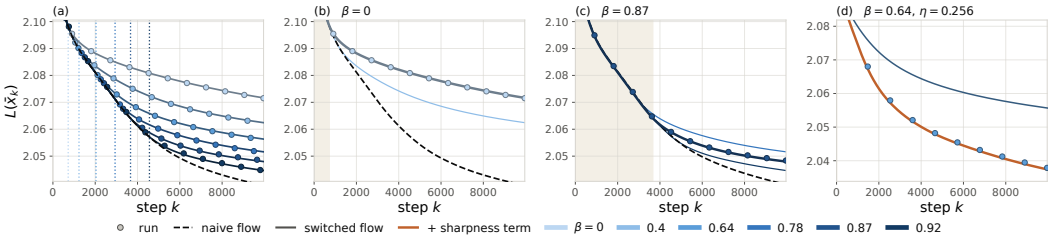


Figure 5: Loss prediction with the switched flow. Markers: loss at the de-oscillated center; lines: flows integrated from initialization, with no fitted parameter; shading: before the switch. (a)–(c) At $\eta = 8 \times 10^{-3}$: (a) every run against the normalized gradient flow (dashed) and the switched flow with its own β; dotted lines mark the switch. (b) $\beta = 0$ against the normalized gradient flow (dashed) and the switched flow with $\beta' = 0$ (thick) and $\beta' = 0.4$ (thin). (c) $\beta = 0.8704$ against the momentum flow (dashed) and the switched flow with $\beta' = 0.8704$ (thick) and $\beta' \in \{0.784, 0.9222\}$ (thin). (d) $\beta = 0.64$ at $\eta = 0.256$ against the switched flow (thin) and the flow (10) with the sharpness term (orange).

Momentum acceleration. On a period-two orbit, the speed of `NormGD` predicts that of `NormGD-M`, which is K_β times larger; this is the change of clock $X_\beta(t) = X(K_\beta t)$ in the switched flow (Section 2.3). Accordingly, the distance and the loss decrease of each β follow K_β times those of $\beta = 0$ (dashed in Figure 4(a,b)), and so does the distance over the whole grid of learning rates (Figure 4(c)).

Learning-rate cap. To vary η at a fixed state, we restart runs with a range of learning rates from one oscillating state (Section B.1). The speed follows the minimum in Equation (7) (Figure 4(d)): below the knee $2K_\beta\|\nabla L\|_2/\lambda_{\max}$, the center advances by η per update, and above it, the speed stays at the knee as η grows. A larger learning rate thus enlarges the oscillation rather than the displacement of its center, whereas a larger β raises the knee itself, by K_β.

3.3 Loss prediction with switched flow

The flow in Equation (7) provides a unified view of the pre- and post-switch behavior of `NormGD-M` with different hyperparameters. The predictions for the discrete iterates in the two phases are validated separately in prior experiments (Sections 3.1 and 3.2). Now, our target here is to provide a comprehensive validity check on the loss tracking property of this flow. We integrate the stated flow from the same initialization as the discrete runs, and compare their loss curves; the errors are reported in Section B.1.

Loss and switch prediction. We highlight the overall capability that the stated switched flow has to track the discrete iterations through both phases: with its own β, it tracks the loss of every run before and after the switch, and its switch nearly coincides with that of the run (Figure 5(a)).

For $\beta = 0$, the switched flow with $\beta' = 0$ detaches from the normalized gradient flow where the run does, while the flow with a larger β' detaches from it later (Figure 5(b)): the flow predicts the onset of the oscillation. After the switch, the proposed flow still tracks the discrete iterates in loss, whereas the normalized gradient flow departs from them.

For a large β, only the switched flow with the matching β' tracks the run, and those with the two neighboring values of β' significantly deviate from the target curve (Figure 5(c)). We also remark that the momentum flow $\dot{m} = -\log\beta\,(\nabla L - m)$, $\dot{x} = -\eta m/\|m\|_2$, the continuous-time limit of `NormGD-M`, is not a good proxy: it cannot oscillate with period two, so it keeps advancing by η per update after the switch, and its loss departs from that of the run. The effect of momentum at the edge of stability is thus captured not by a continuous buffer but by the factor K_β in the switched flow.

Large learning-rate calibration. At learning rates far above the ones shown in Figure 5(a)–(c), the discrete iterates advance faster than the switched flow. We remark that this is because of the higher-order sharpness regularization effect that can be captured with Equation (10). Figure 5(d) shows that adding this regularization term significantly improves loss tracking in this large learning-rate regime.

4 Provable flow tracking and acceleration

This section proves the two ingredients of the model. On a general smooth loss, `NormGD-M` follows the switched flow once it oscillates (Section 4.1). On high-dimensional open quadratics, the two mechanisms of Section 2.2 can be analyzed exactly (Section 4.2).

4.1 General losses: flow tracking

In Section 3.3, the flow integrated from initialization followed the training loss after the switch. The following result is the theoretical counterpart: once `NormGD-M` oscillates, it follows the switched flow started from its current state.

Theorem 4.1 (Informal version of Theorem F.8 and Proposition F.6)**.** *Under proper conditions on the loss geometry, suppose that* `NormGD-M` *oscillates with period two from the start $k = 0$, and let X_β solve Equation* (7) *started from the projection of x_0 onto the path. If $\theta(X_\beta(t)) > 2K_\beta$ for all $t \le T$, then for all $k \le T$,*

$$\|x_k - X_\beta(k)\|_2 \le C_\beta\,\eta\,(1 + \delta k),$$

where $\delta > 0$ measures the per-update change of the loss geometry and C_β does not depend on η, k.

The iterates thus stay within $O(\eta)$ of the flow for $O(\delta^{-1})$ updates. The $O(\eta)$ term is the amplitude of the oscillation around the flow, and the δk term accumulates the effect of the changing geometry; the precise conditions and the definition of δ are in Assumption F.1 and Equation (51). These conditions are sufficient, with unoptimized constants; rather than checking them in training, Section 3 tests the conclusion directly.

The theorem assumes that the oscillation is already established, so it does not cover the passage through the switch. For `NormGD`, the assumption of an established oscillation can be removed where $\theta > 2$: started near the path with a nonzero component in the sharpest direction orthogonal to the gradient, it enters the period-two oscillation within a bounded number of updates (Theorem F.7), after which Theorem 4.1 applies. For `NormGD-M`, the buffer must also reach its periodic state, and we have not proved that this happens from $m_0 = 0$. Even so, `NormGD-M` provably needs fewer updates than `NormGD` to traverse the same sufficiently long part of the path from a common prepared position, by a factor smaller than K_β that does not require the buffer to reach its periodic state (Remark F.11).

4.2 Open quadratics: exact analysis

Section 2.2 derived the two mechanisms on a two-dimensional valley by assuming that the iterates either move straight down the valley or oscillate with period two. On the high-dimensional version

of this valley,

$$L(x) = \frac{1}{2} x^\top H x + b^\top x, \tag{12}$$

the dynamics of NormGD and NormGD-M can be analyzed exactly. Here $H \succeq 0$ has a nontrivial kernel and $b \in \mathrm{Ker}(H)$ is nonzero, so that $\mathrm{Range}(H)$ and $\mathrm{Ker}(H)$ are the sharp and flat directions and $\mathrm{Ker}(H)$ is the floor of the valley; the valley of (2) is the case of rank one. Since this open quadratic loss has no minimizer, we measure progress by the length-averaged loss decrease:

$$V_T := \frac{L(x_0) - L(x_T)}{\eta T},$$

and write V_T^0 and V_T^β for NormGD and NormGD-M with the same η, where β is the momentum factor.

Theorem 4.2 (Informal version of Theorems E.1, E.3 and E.4)**.** *For the open quadratic loss defined in* (12)*, let λ_1 be the largest eigenvalue of H and $\rho_1 := \eta\lambda_1/\|b\|_2$. Then the following hold.*

- Two regimes of NormGD. *For almost every initialization,* NormGD *moves straight down the valley if $\rho_1 < 2$ and settles into a period-two oscillation in the top eigenspace of H if $\rho_1 > 2$, so that $V_T^0 \to \|b\|_2 \min\{1, 2/\rho_1\}$.*
- Momentum extends the stability window. *For $\rho_1 < 2K_\beta$,* NormGD-M *started close to the valley floor converges to it and moves straight down the valley, so that $V_T^\beta \to \|b\|_2$; for $2 < \rho_1 < 2K_\beta$, this is $\rho_1/2$ times the speed of* NormGD.
- Momentum accelerates progress during oscillation. *For $\rho_1 > 2K_\beta$ and a simple top eigenvalue of H,* NormGD-M *has a stable period-two oscillation, to which it returns after small perturbations of its position and buffer; on it, $V_T^\beta \to 2K_\beta\|b\|_2/\rho_1$, which is K_β times the speed of* NormGD.

Here, a period-two oscillation *means that the iterates alternate between the two sides of the valley floor, $Hx_{k+1} = -Hx_k$, while advancing along the valley.*

These limits are the speeds predicted by Equations (4) and (6), extended to high dimensions. For NormGD, the theorem shows that the period-two oscillation assumed in Section 2.2 is actually reached from almost every initialization, and that only the sharpest direction keeps oscillating while the others decay. For NormGD-M, the results are local: they concern initializations close to the valley floor or to the period-two oscillation. The next theorem addresses this restriction and shows that from every initialization the speed of NormGD-M exceeds $2\|b\|_2/\rho_1$, the speed of NormGD from almost every initialization, by a guaranteed factor smaller than K_β.

Theorem 4.3 (Acceleration without periodicity; informal version of Theorem E.2)**.** *Under the setting of Theorem 4.2, if $\rho_1 > 2$ and $0 < \beta < 1$, then for every x_0,* NormGD-M *started with $m_0 = 0$ satisfies*

$$\liminf_{T\to\infty} V_T^\beta \geq \frac{2\|b\|_2}{\rho_1} \min\left\{\frac{\rho_1}{2}, 1 + \frac{2\beta}{\rho_1 + 1 - \beta}\right\} > \frac{2\|b\|_2}{\rho_1}.$$

Since $2\|b\|_2/\rho_1$ is the speed that NormGD *attains for almost every initialization,* NormGD-M *is strictly faster than* NormGD *from the same initialization, whether or not it reaches a period-two oscillation.*

5 Conclusion

We have explained why deterministic normalized descent benefits from momentum. Once training reaches the Edge of Stability, normalization pins the effective learning rate at $2/\lambda_\perp$, so the speed is set by the loss geometry rather than by the learning rate. Momentum attenuates the alternating gradient in its buffer before normalization, which raises the switching threshold and the speed after the switch by the same factor K_β. A switched flow built from this mechanism predicts the switch, the speed after it, and, integrated from initialization, the training loss in full-batch experiments. On open quadratics, these statements are theorems, including a strict speedup from a zero buffer that assumes no periodic behavior; on general smooth losses, the flow is followed locally from a prepared oscillation. The passage through the switch, including the acquisition of the momentum period-two orbit from a zero buffer, remains open. Extending the analysis to stochastic gradients and to non-Euclidean oracles beyond the sign-based case study is left for future work.

Reproducibility statement

Complete assumptions and proofs are provided in the main text and appendices. We report the experimental setup, implementation details, and relevant hyperparameters needed to reproduce the empirical results. Code and configurations for the main experiments will be released upon publication.

AI use statement

The core research ideas, experimental questions, and overall narrative of this work were developed by the authors. Large language models were used to assist with the implementation of some experimental code, to help derive and check parts of the mathematical proofs, and, to a limited extent, to improve the writing and organization of the manuscript. All AI-assisted code, proofs, and written content were reviewed, validated, and revised by the authors as necessary. The authors take full responsibility for the correctness and content of the paper.

References

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

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.

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, pp. 5391–5402, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1dba5eed8838571e1c80af145184e515-Abstract.html.

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

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.

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

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.

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, 2023a. URL https://arxiv.org/abs/2302.06675.

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*, pp. 5396–5427. PMLR, 2023b. URL https://proceedings.mlr.press/v202/chen23ar.html.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

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

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.24206.

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*, pp. 2260–2268. PMLR, 2020.

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.

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.

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.

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.

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.

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

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

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*, pp. 427–435. PMLR, 2013. URL https://proceedings.mlr.press/v28/jaggi13.html.

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.

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.

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.

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.

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

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.

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.

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

Miao Lu, Beining Wu, Xiaodong Yang, and Difan Zou. Benign oscillation of stochastic gradient descent with large learning rates. In *International Conference on Learning Representations*, 2024.

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.

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.

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.

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.

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

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.

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.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon & Scion great again! (Bridging theory and practice of LMO-based optimizers for LLMs), 2025. URL https://arxiv.org/abs/2505.13416.

B. P. Sommeijer, L. F. Shampine, and J. G. Verwer. RKC: An explicit solver for parabolic PDEs. *Journal of Computational and Applied Mathematics*, 88(2):315–326, 1998.

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*, pp. 1139–1147. PMLR, 2013. URL https://proceedings.mlr.press/v28/sutskever13.html.

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*, pp. 10816–10827. PMLR, 2021. URL https://proceedings.mlr.press/v139/wang21n.html.

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

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

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.

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.

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.

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

1 Introduction **1**

2 A switched-flow model of `NormGD-M` **3**

2.1 Normalized steepest descent with momentum 3

2.2 An illustrative example: a two-dimensional valley 3

2.3 From open quadratics to general losses 4

3 The switched flow in deterministic training of NNs **5**

3.1 Pre-switch: extended stable phase from momentum 6

3.2 Post-switch acceleration from momentum and learning-rate cap 6

3.3 Loss prediction with switched flow 7

4 Provable flow tracking and acceleration **8**

4.1 General losses: flow tracking 8

4.2 Open quadratics: exact analysis 8

5 Conclusion **9**

A Related work **16**

B Additional details on experiments **17**

B.1 Setup and measurements of the experiments in Section 3 17

B.2 The sharpness term at large learning rates 18

B.3 Ablations to other NSD variants: `Signum` and `Muon` 19

B.4 Details of the language-model experiment 20

C Analysis of `SignGD` on the rotated quadratic loss **20**

C.1 Rotated coordinates and sign structure 20

C.2 Plain `SignGD`: exact dynamics in the sharp direction 20

C.3 `Signum`: `SignGD` with EMA momentum 21

D Key lemma on NGM speed **22**

E Proof for quadratic loss **29**

E.1 Proof for NGD dynamics 30

E.2 Proof for NGM dynamics 32

F Proof for general loss **38**

F.1 Preliminary analysis 38

F.2 Proof for NGD dynamics 55

F.3 Proof for branch-free NGM 61

F.4 Post-entry comparisons 68

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 (Frank & Wolfe, 1956; Jaggi, 2013; Mokhtari et al., 2020; Braun et al., 2022). In deep learning, the same geometric template appears in coordinate-wise sign methods, Euclidean normalized descent, and spectral updates such as `Muon` (Bernstein et al., 2018; Chen et al., 2023a; Jordan et al., 2024; Liu et al., 2025; Pethick et al., 2025). 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 (Xie & Li, 2024; Xie et al., 2025; Fan et al., 2025; Riabinin et al., 2025). 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 (Polyak, 1964; Nesterov, 1983). For plain gradient descent, the standard acceleration story is parameter-coupled. On quadratics and local quadratic models, momentum turns each eigenvector 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 (Qian, 1999; Lessard et al., 2016; Muehlebach & Jordan, 2021). 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 (Kovachki & Stuart, 2021). In stochastic or online constant-learning-rate settings, momentum can even act mainly as an effective learning-rate rescaling (Yuan et al., 2016). Recent neural-network theory proves analogous Polyak-momentum acceleration when the learning rate and momentum are tuned to the problem geometry (Wang et al., 2021). Our comparisons instead fix the learning rate and vary only β.

In deep learning, momentum is empirically important, but its benefit is known to be regime-dependent (Sutskever et al., 2013; Fu et al., 2023). For stochastic normalized methods, Cutkosky & Mehta (2020) 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 (Yan et al., 2018; Liu et al., 2020; Qian et al., 2025). Other works emphasize limitations, schedules, or optimizer-specific variants of momentum, including stochastic counterexamples, momentum decay, `Lion`, `Signum`, and cautious optimizers (Kidambi et al., 2018; Chen et al., 2021; Zhao et al., 2025; Yu et al., 2026; Liang et al., 2026). 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 (Cohen et al., 2021; Lu et al., 2024). Follow-up work extends this viewpoint to adaptive optimizers, self-stabilizing gradient descent, non-Euclidean descent geometries, and mini-batch stochastic stability (Cohen et al., 2022; Damian et al., 2023; Islamov et al., 2026; Andreyev & Beneventano, 2025). The closest momentum-side work studies how momentum changes sharpness plateaus at the Edge of Stochastic Stability (Andreyev et al., 2026). 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.

B ADDITIONAL DETAILS ON EXPERIMENTS

This appendix gives the setup and measurements of Section 3, the sharpness term at large learning rates, the `Signum` and `Muon` experiments, and the language-model experiment of Figure 1.

B.1 SETUP AND MEASUREMENTS OF THE EXPERIMENTS IN SECTION 3

Task. The inputs are $450{,}000$ standard Gaussian vectors in $\mathbb{R}^{16}$. A teacher MLP with widths $[16, 256, 256, 10]$, $\tanh$ activations, weights drawn from $\mathcal{N}(0, 1/\text{fan-in})$, and zero biases assigns each input a softmax distribution, with logits scaled by 1.5. One label per input is sampled from this distribution once and then held fixed. The student has widths $[16, 128, 128, 10]$ and $19{,}978$ parameters, with the same activation and initialization, and is trained on the full-batch softmax cross-entropy loss. With sampled labels, the gradient norm does not collapse together with the loss, as it does for a realizable regression target, so the runs descend a long valley. We use 22.5 samples per parameter; with 9, the student memorizes the labels and the runs with large β leave the edge of stability. Computations use single precision with full-precision matrix products, which we validated against double precision.

Runs and measurements. All runs start from one initialization with $m_0 = 0$. The momentum parameters are $\beta = 1 - 0.6^i$ for $i = 0, \dots, 5$, that is, $\beta \in \{0, 0.4, 0.64, 0.784, 0.8704, 0.9222\}$, evenly spaced in $\log(1-\beta)$, and the learning rates are $\eta \in \{2, 4, 8, 16, 32\} \times 10^{-3}$; unless stated otherwise, figures use $\eta = 8 \times 10^{-3}$. The de-oscillated center $\bar{x}_k$ is the mean of the last eight iterates, and $\lambda_{\max}$ is computed by power iteration with Hessian–vector products, warm-started from the previous estimate. If the gradient alternates around an orthogonal constant component, $g_k = \bar{g} + (-1)^k \delta$ with $\bar{g} \perp \delta$, then the flip ratio of Equation (11) equals $\|\delta\|_2 / (\|\bar{g}\|_2^2 + \|\delta\|_2^2)^{1/2}$, the fraction of the gradient norm carried by the alternating part. On the period-two orbit, the buffer is $m_k = \bar{g} + (-1)^k \delta / K_\beta$, and the lock $\eta\lambda_{\max}/\|m_k\|_2 = 2K_\beta$ gives $\|\delta\|_2^2 / \|\bar{g}\|_2^2 = \theta^2/4 - K_\beta^2$ with $\theta = \eta\lambda_{\max}/\|\bar{g}\|_2$, so that

$$R_k^2 = \frac{\theta^2 - 4K_\beta^2}{\theta^2 - 4K_\beta^2 + 4}.$$

The flip ratio thus rises from zero at the threshold $\theta = 2K_\beta$ and reaches $1/2$ at $\theta = 2(K_\beta^2 + 1/3)^{1/2}$, which exceeds the threshold by a factor $(1 + 1/(3K_\beta^2))^{1/2}$, that is, 1.15 for $\beta = 0$, 1.03 for $\beta = 0.4$, and less than 1.01 for $\beta \geq 0.64$. We use $\lambda_{\max}$ in place of $\lambda_\perp$; in the slowly varying regime of Section 4.1, the two differ by a second-order geometric error (Equation (57)). The switch k^* is the first iteration at which R_k exceeds $1/2$; with a fifty-iteration mean of R_k in place of R_k, it moves later by 8 to 46 iterations for $\eta \in [2 \times 10^{-3}, 3.2 \times 10^{-2}]$.

Before the switch (Figure 3). At $\eta = 8 \times 10^{-3}$, R_k rises from 0.1 to 0.5 within 9 iterations for every β, and k^* grows from 722 for $\beta = 0$ to 4566 for $\beta = 0.9222$. Over the 500 updates after the switch, the number of updates by which a run trails the normalized gradient flow at equal loss grows by 82 to 186. After the switch, $\eta\lambda_{\max}/\|m_k\|_2$ settles at $2K_\beta$ to within 0.04% for every β. At $\eta = 8 \times 10^{-3}$, from initialization to the last switch, $\|\nabla L(\bar{x}_k)\|_2$ falls by a factor of 500 and $\lambda_{\max}$ grows by a factor of 7; before the first switch, $\lambda_{\max}$ grows by 5%, while the gradient norm falls by a factor of 160. At the same learning rate, the loss decrease before the switch grows from 0.281 for $\beta = 0$ to 0.322 for $\beta = 0.9222$. At $\eta = 2 \times 10^{-3}$, $\theta_{k^*}/(2K_\beta)$ is 1.151 for $\beta = 0$ and 1.030 for $\beta = 0.4$, against 1.155 and 1.030 from the value $2(K_\beta^2 + 1/3)^{1/2}$ above, and it is below 1.017 for $\beta \geq 0.64$; at larger η, it exceeds $2K_\beta$, by up to a factor of 1.6 for $\beta = 0$ at $\eta = 3.2 \times 10^{-2}$. The oscillation is detected only after it has grown, and meanwhile θ_k keeps increasing, by a relative amount per update that is proportional to η and smaller later along the path, where the runs with larger β switch.

Speed after the switch (Figure 4). At each η, we run $\beta = 0.9222$ until θ_k reaches 2.5 times its threshold $2K_{0.9222}$ and restart all six values of β from that state with a zero buffer, so that every run starts at least 2.5 times beyond its own threshold. After a burn-in of 400 updates, which refills the buffer, j counts updates. The window is chosen per η so that the fastest run travels the same

arclength; it is 1265, 635, 340, 190, and 110 updates for $\eta = 2, 4, 8, 16$, and 32×10^{-3}. Per η, the median relative error of the K_β law over β is between 0.6% and 3.9% for the arclength and between 1.2% and 7.0% for the loss decrease. For panel (d), the common state is the de-oscillated center of the state of $\eta = 3.2 \times 10^{-2}$ above, after 300 further updates of $\beta = 0.9222$. Every pair (β, η) with $\eta = 1.25 \times 10^{-4} \cdot 2^i$, $i = 0, \dots, 8$, starts there with $m_0 = 0$ and a displacement of $\eta_0/100$ along the top eigenvector, where $\eta_0 = \min\{\eta, 1.15\,\eta_c\}$ and η_c is the knee $2K_\beta\|\nabla L\|_2/\lambda_{\max}$ at the start of the window. It holds the learning rate η_0 for 250 updates, which lets it enter the period-two orbit as it does from initialization, raises it geometrically to η by update 350, and is measured over updates 400 to 740. The knee is evaluated at the center at the start of the window, for the smallest learning rate above it, where the oscillation is smallest. Below the knee, the center advances by η per update to within 0.03%; above it, the speed stays within 6% of the knee while η grows by up to a factor of 64, and doubling η changes it by at most 5%.

Loss prediction (Figure 5). Every switched flow moves along $-\nabla L/\|\nabla L\|_2$, so all of them trace the path of the normalized gradient flow and differ only in their speed along it, $\mathrm{d}s/\mathrm{d}k = \min\{\eta, 2K_\beta\|\nabla L\|_2/\lambda_{\max}\}$, where s is the arclength. We compute the path once with a second-order Runge–Kutta–Chebyshev method (Sommeijer et al., 1998), which remains stable where θ and hence the stiffness of the flow are large, and obtain each flow by quadrature of its speed. The momentum flow is integrated by the classical fourth-order Runge–Kutta method with step 0.2; with step 0.4, its loss changes by less than 5×10^{-7}. The runs last 9,960 updates. The error of a flow is the root-mean-square difference between its loss and the loss at the de-oscillated center over the updates after the switch of the run, divided by the decrease of the latter over the same updates. At $\eta = 8 \times 10^{-3}$, the switched flow with the run's own β has an error between 0.3% and 2.3%, and its loss departs from that of the normalized gradient flow by 10^{-5} between 12 and 66 updates before the switch of the run. For $\beta = 0$, its error is 0.3%, against 92% for the normalized gradient flow; for $\beta = 0.8704$, it is 1.4%, against 18% to 21% for $\beta' \in \{0.784, 0.9222\}$ and 35% for the momentum flow.

B.2 The sharpness term at large learning rates

Derivation. The switched flow treats the loss at each point as the valley of Section 2.2, whose curvature does not vary across it; on a general loss, the oscillating iterates, which sit on the walls of the valley, feel the variation of the curvature. Let $x_k = \bar{x} + (-1)^k a\, v$ be a period-two orbit whose center $\bar{x}$ moves slowly, where v is the top eigenvector of $\nabla^2 L(\bar{x})$ and λ its eigenvalue. The average of the loss over the two iterates is $L(\bar{x}) + \frac{1}{2}a^2\lambda(\bar{x}) + O(a^4) = \bar{L}(\bar{x}) + O(a^4)$, the floor of the valley plus the energy of the oscillation, since the odd orders cancel, and the gradient is

$$g_k = \nabla L(\bar{x}) + (-1)^k a\lambda v + \tfrac{1}{2}a^2\nabla^3 L(\bar{x})[v, v, \cdot\,] + O(a^3) = \nabla\bar{L}(\bar{x}) + (-1)^k a\lambda v + O(a^3),$$

where $\nabla^3 L(\bar{x})[v, v, \cdot\,] = \nabla\lambda(\bar{x})$ by the Hellmann–Feynman identity and $\nabla\bar{L}$ is taken at fixed a. The buffer passes the average $\nabla\bar{L}$ and, as in Equation (5), divides the alternating part by K_β, so $m_k \approx \nabla\bar{L} + (-1)^k(a\lambda/K_\beta)v$ with $\nabla\bar{L} \perp v$, and the center descends $\bar{L}$ instead of L. Each update $-\eta m_k/\|m_k\|_2$ does two things. It flips the sharp coordinate from a to $-a$, which requires its component along v to have length $2a$, that is, $\|m_k\|_2 = \eta\lambda/(2K_\beta)$, the lock of Equation (6). It also moves the center, by a component along $\nabla\bar{L}$ of length $2K_\beta\|\nabla\bar{L}\|_2/\lambda$ at this locked effective learning rate, which gives the drift in Equation (10). The two legs are orthogonal, so $\|m_k\|_2^2 = \|\nabla\bar{L}\|_2^2 + a^2\lambda^2/K_\beta^2$ gives $(2a)^2 + (2K_\beta\|\nabla\bar{L}\|_2/\lambda)^2 = \eta^2$, the second relation in Equation (10). Since $\nabla\bar{L}$ depends on a^2, the triangle is a quadratic equation for a^2, which we solve exactly. At the switch, $\theta = 2K_\beta$, so the second leg alone has length η and $a = 0$: the flow starts as the switched flow, and beyond the switch the oscillation takes a growing share of each update. Since $a \le \eta/2$, the added drift is of higher order in η, and it vanishes on open quadratics, where λ is constant.

Integration. We integrate Equation (10) from initialization in steps of $H = \min\{\max\{10, 0.64/\eta\}, 150\}$ updates. Within a step, λ and v (from ten warm-started power iterations), $\nabla\lambda$ (by automatic differentiation of $v^\top\nabla^2 L(x)v$ with v held fixed), and a^2 are frozen, and one step of the second-order Runge–Kutta–Chebyshev method is taken. In the stable phase, the flow is the normalized gradient flow, integrated as in Section B.1. The flow enters the oscillatory phase where θ reaches $2K_\beta$ and stays in it as long as the equation for a^2 has a positive root; with

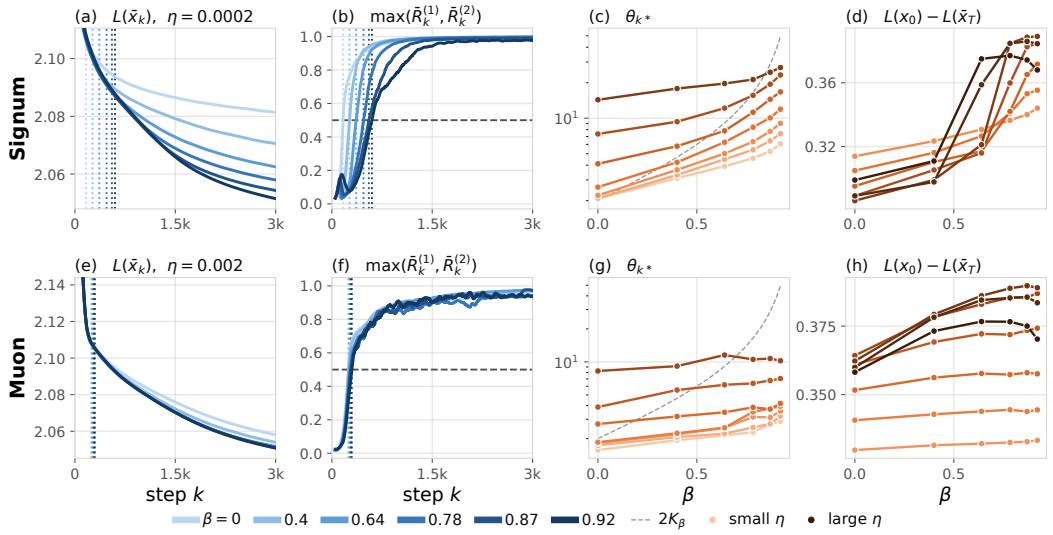


Figure 6: `Signum` (top) and `Muon` (bottom) on the task of Section 3, measured in each optimizer's own norm. (a), (e) Loss at the de-oscillated center at $\eta = 2 \times 10^{-4}$ and $\eta = 2 \times 10^{-3}$; dotted lines mark the switch. (b), (f) The larger of the flip ratios at lags one and two, averaged over the last fifty updates. (c), (g) Normalized sharpness at the switch against β, one curve per η (light to dark: small to large); dashed: $2K_\beta$ for reference. (d), (h) Loss decrease over 14,000 updates against β, one curve per η.

the term, $\|\nabla \bar{L}\|_2$ can be smaller than $\|\nabla L\|_2$, so a run can oscillate at θ slightly below $2K_\beta$. The switched flow is integrated by the same code with the term removed.

B.3 Ablations to other NSD variants: `Signum` and `Muon`

We run `Signum`, which updates by $-\eta\,\mathrm{sign}(m_{k+1})$, and `Muon`, which updates each weight matrix by $-\eta$ times the polar factor of its buffer (five Newton–Schulz iterations) and each bias by $-\eta\,\mathrm{sign}(m_{k+1})$, on the task of Section 3 with the same six values of β, for 14,000 updates at $\eta \in \{1, 2, 4, \ldots, 64\} \times 10^{-4}$ for `Signum` and $\eta \in \{0.5, 1, 2, \ldots, 64\} \times 10^{-3}$ for `Muon`; separate runs to the switch, used in panels (c) and (g), extend the range down to $\eta = 2.5 \times 10^{-5}$ and $\eta = 1.25 \times 10^{-4}$. Every quantity is measured in the optimizer's own norm pair $(\|\cdot\|, \|\cdot\|_*)$. The dual norm is ℓ_1 for `Signum`; for `Muon`, it is the sum of the nuclear norms of the weight matrices and the ℓ_1 norm of the biases. The flip ratio becomes $R_k = \|g_k - g_{k-1}\|_* / (\|g_k\|_* + \|g_{k-1}\|_*)$, and the normalized sharpness becomes $\theta = \eta S / \|\nabla L\|_*$, where $S(x) = \sup_{\|d\| \le 1} d^\top \nabla^2 L(x)\, d$ is the generalized sharpness (Islamov et al., 2026). We estimate S from below by the generalized power iteration $d \leftarrow \mathrm{LMO}(\nabla^2 L\, d)$ from several starts, and locate the switch as in Section 3.1, with one change: at large β, consecutive gradients of `Muon` alternate less than gradients two updates apart, so we use the larger of the flip ratios at lags one and two, which is near one whenever the gradient alternates between consecutive updates or between every other update.

Both methods start to oscillate, and a larger β postpones the switch, as for `NormGD-M` (Figure 6(b,f)). At the learning rates of panels (a) and (e), the switch moves from update 166 to update 599 for `Signum` and from update 247 to update 298 for `Muon` as β grows from 0 to 0.9222. The flip ratio rises more gradually than for `NormGD-M`, since the sign and polar maps normalize each coordinate or singular direction separately. The normalized sharpness at the switch also grows with β (Figure 6(c,g)): at the smallest learning rate, from 2.1 to 6.1 for `Signum` and from 1.6 to 2.9 for `Muon`.

Momentum still lowers the loss (Figure 6(a,d)). For `Signum`, the loss decrease over the run grows with β, most at the larger learning rates: at $\eta = 1.6 \times 10^{-3}$, it grows from 0.286 for $\beta = 0$ to 0.389 for $\beta = 0.9222$. For `Muon`, the gain is smaller but has the same sign: at $\eta = 1.6 \times 10^{-2}$, the loss decrease grows from 0.362 to 0.389.

B.4 Details of the language-model experiment

The model of Figure 1 is a decoder-only transformer with 8 layers, width 256, 4 attention heads, and feed-forward width 1024, with 14,579,200 parameters. It is trained on a fixed set of 1,024 sequences of 256 tokens from C4, and every update uses the gradient of the whole set, accumulated over eight token-weighted micro-batches of 128 sequences. All runs start from one initialization with a zero buffer and use a constant learning rate, with no warmup, decay, clipping, weight decay, or Nesterov momentum. The model computes in bfloat16, and the parameters and buffers are stored in single precision. `NormGD-M` normalizes the buffer by its global ℓ_2 norm and uses $\eta = 0.48$; `Signum` is Lion with $\beta_1 = \beta_2 = \beta$ and $\eta = 1.25 \times 10^{-4}$; `Muon` uses five Newton–Schulz iterations with shape-dependent scaling on every matrix, including the embeddings, sign updates on the other parameters, and $\eta = 3.125 \times 10^{-4}$. Each learning rate is half of the best value of an eight-point, factor-of-two search at $\beta = 0$, selected by the training loss, and it is not retuned for individual values of β. The learning rates are fixed within each optimizer, so the figure compares values of β at a common learning rate, not the optimizers with each other. Panel (a) shows an exponential moving average of the loss with decay 0.3; panels (b) and (c) show the raw loss.

C Analysis of SignGD on the rotated quadratic loss

This section records a sign-oracle analog of the momentum mechanism of Section 2.2. Unlike the Euclidean normalized oracle, 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).

C.1 Rotated coordinates and sign structure

Let (u, v) denote river coordinates for the loss $f(u, v) = \lambda u^2/2 - bv$, where $\lambda > 0$ 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 $f(x, y) = \lambda(x + y)^2/4 - b(y - x)/\sqrt{2}$. We write $z_t \coloneqq \lambda u_t/b$ for the dimensionless sharp coordinate and $\gamma \coloneqq \sqrt{2}\eta\lambda/b$ for the dimensionless stepsize. A direct calculation gives

$$\nabla_{x,y} f(x, y) = \frac{b}{\sqrt{2}} (z + 1, z - 1). \tag{13}$$

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

C.2 Plain SignGD: exact dynamics in the sharp direction

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

$$z_{t+1} = z_t - \gamma \operatorname{sgn}(z_t) \mathbf{1}_{\{|z_t|>1\}}, \tag{14}$$

$$v_{t+1} - v_t = \sqrt{2}\eta \mathbf{1}_{\{|z_t|<1\}}. \tag{15}$$

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

Proof of Proposition C.1. From Equation (13), 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 Equation (14)–Equation (15). This completes the proof. □

Theorem C.2 (A zero-progress sharp-direction period-two orbit 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 period-two orbit along the sharp direction with zero downstream progress.*

Proof of Theorem C.2. Since $\gamma > 2$, one has $A > 1$. If $z_t = A$, then Equation (14) gives the identity $z_{t+1} = A - \gamma = -A$. If $z_t = -A$, then Equation (14) gives the identity $z_{t+1} = -A + \gamma = A$. This proves the claimed period-two orbit. Since $|z_t| = A > 1$ for every t, Equation (15) gives $v_{t+1} - v_t = 0$ for every t, proving the zero-progress claim. This completes the proof. □

C.3 `Signum`: `SignGD` WITH EMA MOMENTUM

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

$$m_{t+1} = \beta m_t + (1-\beta)\nabla f(x_t, y_t), \qquad \beta \in (0,1), \tag{16}$$

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

$$m_t = \frac{b}{\sqrt{2}}(p_t + q_t, p_t - q_t). \tag{17}$$

Proposition C.3 (`Signum` dynamics)**.** *With $m_0 = 0$, the momentum coordinates satisfy*

$$p_{t+1} = \beta p_t + (1-\beta) z_t, \qquad q_{t+1} = \beta q_t + (1-\beta), \tag{18}$$

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

$$z_{t+1} = z_t - \gamma \operatorname{sgn}(p_{t+1})\, \mathbf{1}_{\{|p_{t+1}| > q_{t+1}\}}, \tag{19}$$

$$v_{t+1} - v_t = \sqrt{2}\eta\, \mathbf{1}_{\{|p_{t+1}| < q_{t+1}\}}. \tag{20}$$

Proof of Proposition C.3. Combining Equation (13), Equation (17), and Equation (16) gives the two scalar recursions in Equation (18). 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 along the sharp direction 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 along the sharp direction 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 Equation (19)–Equation (20). This completes the proof. □

The next theorem gives a fully actual-momentum separation result on the same symmetric bad initialization as Theorem C.2. 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 C.4 (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 move along the sharp direction.

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

$$v_N - v_0 \geq \sqrt{2}\eta \left\lfloor \frac{N}{2} \right\rfloor. \tag{21}$$

Combining with Theorem C.2, for the same initialization the downstream-progress gap between the `Signum` *iterate* v_N^{M} *and the plain* `SignGD` *iterate* v_N^{G} *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 C.4. 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 sharp-direction case is $\sigma p_{t+1} > q_{t+1}$. The switching equality is excluded by assumption.

If the downstream case holds, then Equation (19) gives $z_{t+1} = z_t = \sigma A$, and Equation (20) gives one downstream step. Since $|p_{t+1}| < q_{t+1}$, time $t + 1$ is calm again.

If the sharp-direction case holds, then Equation (19) gives $z_{t+1} = \sigma A - \gamma\sigma = -\sigma A$. Thus the step moves along the sharp direction and the sign of z flips. The scalar recursion gives $q_{t+2} = \beta q_{t+1} + s = \beta^2 q_t + s(1 + \beta)$. The next step must be downstream because

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

The equalities use the scalar recursion, and the inequalities use $|p_t| \leq q_t$ and the window assumption. 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 moves along the sharp direction 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 move along the sharp direction. 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 Equation (20). This proves Equation (21). The gap statements follow from the zero-progress claim in Theorem C.2. This completes the proof. □

Remark C.5 (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 sharp-direction period-two orbit $z_t = \pm\gamma/2$, the EMA buffer prevents consecutive moves along the sharp direction under the window in Theorem C.4. Thus actual momentum converts a zero-progress sharp-direction oscillation into a sequence with at least one downstream step every two iterations.

D KEY LEMMA ON NGM SPEED

Notation. The following conventions apply in Sections D to F. We write NGD for normalized gradient descent (`NormGD`) and NGM for its momentum version (`NormGD-M`). Unless another norm is specified, $\langle u, v\rangle = u^\top v$ and $\|v\| = \|v\|_2$ are the Euclidean inner product and norm. We write

$\|\cdot\|_{\mathrm{op}}$ for the induced Euclidean operator norm, including for multilinear maps. For a symmetric matrix $M \succ 0$, the Mahalanobis norm is $\|v\|_M := (v^\top M v)^{1/2}$. We write I or I_m for the identity matrix and $A^\dagger$ for the Moore–Penrose inverse; for symmetric matrices, $A \succeq B$ means that $A - B$ is positive semidefinite. The derivative of a map f is denoted by Df.

For open quadratics, plain `NormGD` admits a global dynamical classification and an exact asymptotic loss speed, as shown in Theorem E.1. To obtain the strict comparison in Section 4.2, it therefore suffices to establish a momentum speed lower bound that exceeds the corresponding plain speed, without determining the eventual momentum orbit. The storage argument below controls average progress through energy identities and buffer boundedness, without selecting a recurrent branch.

The lemma below treats a fixed positive-semidefinite matrix and a perturbed momentum recurrence over a finite window. Theorem E.2 applies it with zero perturbations and lets the window and prefix lengths tend to infinity. The proof of Theorem F.10 extends the elementary storage calculation to a changing normal-sharpness matrix, retaining explicit one-step and endpoint errors. The proof gives two lower bounds: one uses only the maximum growth of the normalization, and the other also uses the distribution of its jumps.

Fix $0 < \beta < 1$ and a symmetric matrix $\Theta \succeq 0$ with positive spectrum in $[\theta_r, \theta_1]$, where $\theta_r > 0$. For an integer $T \geq 1$, consider buffers $p_k \in \mathrm{Range}(\Theta)$ for $0 \leq k \leq T+1$ and tangential coefficients $h_k \in \mathbb{R}$ for $0 \leq k \leq T$. Assume

$$\begin{aligned} \|p_k\| &\leq P, \qquad 0 \leq k \leq T+1, \\ h_{k+1} &\geq h_- > 0, \qquad 0 \leq k < T. \end{aligned}$$

For $0 \leq k < T$, the normalization is $F_k = \sqrt{h_{k+1}^2 + \|p_{k+1}\|^2}$, and the normalized tangential progress is $h_{k+1}/F_k \in (0, 1]$. Suppose the recurrence residuals satisfy, for a common tolerance $\delta \geq 0$,

$$\begin{aligned} \left\| p_{k+2} - (1+\beta)p_{k+1} + \beta p_k + (1-\beta)\Theta \frac{p_{k+1}}{F_k} \right\| &\leq \delta, \\ |h_{k+1} - \beta h_k - (1-\beta)| &\leq \delta, \qquad 0 \leq k < T. \end{aligned}$$

Define the auxiliary states

$$z_k := \frac{p_{k+1} - \beta p_k}{1-\beta}, \qquad 0 \leq k \leq T;$$

in the quadratic recurrence they are exactly the rescaled sharp gradient components, and in the general-loss application they are the frozen sharpness applied to the normal coordinates, up to the buffer residual. Fix an integer k_0 with $0 \leq k_0 < T$, set $T' := T - k_0$, and assume $\|z_k\| \leq Z$ for $k_0 \leq k \leq T$. The prefix of length k_0 absorbs the transients of the buffer velocity and of the tangential coefficient, and the lag parameter

$$a := \theta_1 + 2P\beta^{k_0+1} + \frac{\delta}{1-\beta}$$

bounds the one-step growth of the auxiliary normalization $\sqrt{1 + \|p_{k+1}\|^2}$ on the remaining window.

The ideal speed is defined for $\theta > 0$ and $a \geq \theta$ by

$$\begin{aligned} \mathcal{S}(\beta, \theta; a) &:= \max\left\{ \min\left\{1, \frac{2(a+1+\beta)}{\theta(a+1-\beta)}\right\}, \ \frac{2}{\theta}\, \mathcal{G}_\beta\Big(\theta; \frac{a}{\theta}\Big) \right\}, \\ \mathcal{G}_\beta(\theta; \lambda) &:= \inf_{0<\tau\leq 1,\ 0<\vartheta\leq 1} \left\{ \frac{u_{\tau,\vartheta}}{2} + \frac{\beta(1-\beta)\vartheta}{2\lambda^2 \tau}\, b_\tau \Big(b_\tau - \frac{\lambda u_{\tau,\vartheta}\tau}{\beta} \Big)^+ \right\}, \end{aligned} \tag{22}$$

where $b_\tau := 1 - (1+\beta)\tau/(2\beta)$ and $u_{\tau,\vartheta} := 2/[\vartheta(1-\tau) + 2/\theta]$ as in Equation (41). At $\lambda = 1$ the infimum equals $G_\beta(\theta)$ there, since the objective equals $\theta/2$ at $\tau = 1$. The two entries of the maximum are the bounds from the normalization lag and from the distribution of its jumps, respectively. The objective in Equation (22) is nondecreasing in θ and nonincreasing in λ, so $\mathcal{S}$ is nonincreasing in a, and $(\theta/2)\mathcal{S}(\beta, \theta; a) > 1$ whenever $\theta > 2$, because both entries of the minimum then exceed $2/\theta$. For orientation, at $\beta = 0.9$ and $a = \theta$ the rescaled-time speed $(\theta/2)\mathcal{S}$ equals 1.50

at $\theta = 3$, 1.30 at $\theta = 6$, 1.16 at $\theta = 20$, and 1.23 at $\theta = 50$; the plain period-two orbit has speed 1 in these units, and the exact momentum period-two orbit has speed $(1+\beta)/(1-\beta) = 19$ where it exists. All $O(\cdot)$ constants in this section may depend on β, θ_r, θ_1, P, Z, h_-, and $|h_0 - 1|$, but are uniform in T, k_0, and δ.

Lemma D.1 (Perturbed finite-window momentum storage)**.** *Under the preceding setup, there is a nonnegative residual $\mathcal{R}_T$ such that*

$$\frac{1}{T}\sum_{k=0}^{T-1}\frac{h_{k+1}}{F_k} \geq \mathcal{S}(\beta, \theta_1; a) - \mathcal{R}_T. \tag{23}$$

The residual is given explicitly in Equation (35) *and satisfies $\mathcal{R}_T = O\big(\sqrt{(T')^{-1} + \delta} + k_0/T\big)$ for $0 \leq \delta \leq 1$. Moreover, for every $\theta > 0$ the function $\lambda \mapsto \mathcal{G}_\beta(\theta; \lambda)$ is nonincreasing and continuous on $[1, \infty)$. For every $0 < \theta_{\mathrm{lo}} \leq \theta_{\mathrm{hi}}$, there is a constant $C \geq 1$, depending only on $\beta, \theta_{\mathrm{lo}}, \theta_{\mathrm{hi}}$, such that*

$$0 \leq \frac{\theta}{2}\big[\mathcal{S}(\beta, \theta; \theta) - \mathcal{S}(\beta, \theta; \lambda\theta)\big] \leq C(\lambda - 1), \tag{24}$$

uniformly for $\theta \in [\theta_{\mathrm{lo}}, \theta_{\mathrm{hi}}]$ and $1 \leq \lambda \leq 2$.

Proof of Lemma D.1. We first replace the actual normalization by one with tangential component equal to one, then derive two averaged identities and the two speed bounds.

Normalization and perturbations. For $-1 \leq k \leq T$, let $\widehat{F}_k = \sqrt{1 + \|p_{k+1}\|^2}$, and for $0 \leq k \leq T$ let $\phi_k = p_{k+1}/\widehat{F}_k$. Thus $1 \leq \widehat{F}_k \leq 1 + P$ and $\|\phi_k\|^2 = 1 - \widehat{F}_k^{-2}$. The triangle inequality and the tangential recurrence give, for $0 \leq k < T$,

$$\begin{aligned} |F_k - \widehat{F}_k| &\leq |h_{k+1} - 1| \\ &\leq \beta^{k+1}|h_0 - 1| + \delta\sum_{j=0}^{k}\beta^j \leq \beta^{k+1}|h_0 - 1| + \frac{\delta}{1-\beta}. \end{aligned}$$

Averaging this bound over the remaining window gives

$$\frac{1}{T'}\sum_{k=k_0}^{T-1}|h_{k+1} - 1| \leq \overline{\nu} := \frac{\beta^{k_0+1}|h_0 - 1|}{(1-\beta)T'} + \frac{\delta}{1-\beta}. \tag{25}$$

Absorb the recurrence residual and the change of normalization into

$$\begin{aligned} g_k &:= z_{k+1} - z_k + \Theta\phi_k \\ &= \frac{p_{k+2} - (1+\beta)p_{k+1} + \beta p_k}{1-\beta} + \Theta\frac{p_{k+1}}{F_k} + \Theta p_{k+1}\left(\frac{1}{\widehat{F}_k} - \frac{1}{F_k}\right), \\ \|g_k\| &\leq \frac{\delta}{1-\beta} + \frac{\theta_1\|p_{k+1}\||F_k - \widehat{F}_k|}{F_k\widehat{F}_k} \leq \frac{\delta}{1-\beta} + \frac{\theta_1|h_{k+1} - 1|}{h_-}. \end{aligned}$$

By construction, the perturbed recurrence now reads

$$z_{k+1} - z_k = -\Theta\phi_k + g_k, \qquad 0 \leq k < T. \tag{26}$$

We also need a bound on one-step normalization growth. Using the original recurrence, $F_k \geq \|p_{k+1}\|$, and $\|p_1 - p_0\| \leq 2P$, we obtain successively

$$\begin{aligned} \|p_{k+2} - p_{k+1}\| &\leq \beta\|p_{k+1} - p_k\| + (1-\beta)\theta_1 + \delta \\ &\leq 2P\beta^{k+1} + \big((1-\beta)\theta_1 + \delta\big)\sum_{j=0}^{k}\beta^j \\ &\leq 2P\beta^{k+1} + \theta_1 + \frac{\delta}{1-\beta}. \end{aligned}$$

Since $p \mapsto \sqrt{1+\|p\|^2}$ is 1-Lipschitz, this proves

$$\widehat{F}_{k+1} - \widehat{F}_k \le \|p_{k+2} - p_{k+1}\| \le a, \qquad k_0 \le k < T. \tag{27}$$

Energy and normalization identities. Define the Bregman divergence of $p \mapsto \sqrt{1+\|p\|^2}$ between consecutive buffers by

$$B_k := \widehat{F}_{k-1} - \widehat{F}_k - \phi_k^\top (p_k - p_{k+1}) = \widehat{F}_{k-1} - \frac{1 + p_k^\top p_{k+1}}{\widehat{F}_k}, \qquad 0 \le k \le T.$$

Convexity and Cauchy–Schwarz give $0 \le B_k \le 2\widehat{F}_{k-1}$. Writing $\varrho_k = \widehat{F}_{k+1}/\widehat{F}_k$ for $0 \le k < T$, the same definition gives the three inner products

$$\begin{aligned} \phi_k^\top p_{k+1} &= \widehat{F}_k - \widehat{F}_k^{-1}, \\ \phi_k^\top p_k &= \widehat{F}_{k-1} - B_k - \widehat{F}_k^{-1}, \\ \phi_k^\top p_{k+2} &= \widehat{F}_{k+1} - \varrho_k B_{k+1} - \widehat{F}_k^{-1}. \end{aligned} \tag{28}$$

All vectors in the energy calculation lie in $\mathrm{Range}(\Theta)$, on which $\Theta^\dagger\Theta$ is the identity. Using Equations (26) and (28) and $z_k = p_{k+1} + \beta(p_{k+1} - p_k)/(1-\beta)$ gives

$$\begin{aligned} z_{k+1}^\top \Theta^\dagger z_{k+1} - z_k^\top \Theta^\dagger z_k &= (z_{k+1} + z_k)^\top \Theta^\dagger(-\Theta\phi_k + g_k) \\ &= -2z_k^\top \phi_k + \phi_k^\top \Theta \phi_k + (z_{k+1} + z_k)^\top \Theta^\dagger g_k - \phi_k^\top g_k \\ &= -2(\widehat{F}_k - \widehat{F}_k^{-1}) - \frac{2\beta}{1-\beta}(\widehat{F}_k - \widehat{F}_{k-1} + B_k) + \phi_k^\top \Theta \phi_k \\ &\quad + (z_{k+1} + z_k)^\top \Theta^\dagger g_k - \phi_k^\top g_k. \end{aligned}$$

Define the four averages that enter the energy identity:

$$S = \frac{1}{T'}\sum_{k=k_0}^{T-1} \widehat{F}_k^{-1}, \qquad D = \frac{1}{T'}\sum_{k=k_0}^{T-1} \widehat{F}_k,$$
$$X = \frac{1}{T'}\sum_{k=k_0}^{T-1} \phi_k^\top \Theta \phi_k, \qquad J = \frac{1}{T'}\sum_{k=k_0}^{T-1} B_k.$$

Averaging the preceding calculation gives the exact telescoping relation

$$\begin{aligned} D - S - \frac{X}{2} + \frac{\beta}{1-\beta} J &= -\frac{z_T^\top \Theta^\dagger z_T - z_{k_0}^\top \Theta^\dagger z_{k_0}}{2T'} - \frac{\beta(\widehat{F}_{T-1} - \widehat{F}_{k_0-1})}{(1-\beta)T'} \\ &\quad + \frac{1}{2T'}\sum_{k=k_0}^{T-1} \left[(z_{k+1} + z_k)^\top \Theta^\dagger g_k - \phi_k^\top g_k\right]. \end{aligned}$$

Since $0 \le z_k^\top \Theta^\dagger z_k \le Z^2/\theta_r$, $\|\phi_k\| \le 1$, and $|\widehat{F}_{T-1} - \widehat{F}_{k_0-1}| \le P$, the preceding perturbation bound and Equation (25) yield

$$\begin{aligned} \left| D - S - \frac{X}{2} + \frac{\beta}{1-\beta} J \right| &\le \left(\frac{Z}{\theta_r} + \frac{1}{2}\right) \frac{1}{T'} \sum_{k=k_0}^{T-1} \|g_k\| + \frac{Z^2}{2\theta_r T'} + \frac{\beta P}{(1-\beta)T'} \\ &\le \left(\frac{Z}{\theta_r} + \frac{1}{2}\right)\left(\frac{\delta}{1-\beta} + \frac{\theta_1 \overline{\nu}}{h_-}\right) + \frac{Z^2}{2\theta_r T'} + \frac{\beta P}{(1-\beta)T'} =: \mathcal{E}_1. \end{aligned}$$

For the normalization identity, the second-order recurrence gives another direct calculation:

$$\begin{aligned} \widehat{F}_{k+1} - (1+\beta)\widehat{F}_k + \beta \widehat{F}_{k-1} &= \phi_k^\top \big(p_{k+2} - (1+\beta)p_{k+1} + \beta p_k\big) + \varrho_k B_{k+1} + \beta B_k \\ &= -(1-\beta)\phi_k^\top \Theta \phi_k + \varrho_k B_{k+1} + \beta B_k + (1-\beta) g_k^\top \phi_k. \end{aligned}$$

For the normalization jumps, define

$$A = \frac{1}{T'}\sum_{k=k_0}^{T-1} (\varrho_k - 1) B_{k+1}, \qquad M = \frac{1}{T'}\sum_{k=k_0}^{T-1} B_{k+1}.$$

Telescoping the normalization differences yields

$$A + M - (1-\beta)X + \beta J = \frac{\widehat{F}_T - \widehat{F}_{k_0} - \beta(\widehat{F}_{T-1} - \widehat{F}_{k_0-1})}{T'} - \frac{1-\beta}{T'} \sum_{k=k_0}^{T-1} g_k^\top \phi_k,$$

$$|A + M - (1-\beta)X + \beta J| \le \frac{(1+\beta)P}{T'} + \delta + \frac{(1-\beta)\theta_1 \overline{\nu}}{h_-} =: \mathcal{E}_2, \tag{29}$$

$$|M - J| = \frac{|B_T - B_{k_0}|}{T'} \le \mathcal{E}_3 := \frac{2(1+P)}{T'}.$$

Finally, Jensen's inequality and the spectral bounds give

$$D \ge \frac{1}{S}, \qquad 0 < S \le 1,$$

$$\theta_r(1-S) \le \theta_r \frac{1}{T'} \sum_{k=k_0}^{T-1} (1 - \widehat{F}_k^{-2}) \le X \le \theta_1 \frac{1}{T'} \sum_{k=k_0}^{T-1} (1 - \widehat{F}_k^{-2}) \le \theta_1(1-S^2). \tag{30}$$

The bound from the normalization lag. By Equation (27), $\varrho_k - 1 \le a$, so $A \le aM$. Together with the normalization identity, this gives

$$\begin{aligned}(1-\beta)X &\le A + M + \beta J + \mathcal{E}_2 \le (1+a)M + \beta J + \mathcal{E}_2 \\ &\le (1 + a + \beta)J + \mathcal{E}_2 + (1+a)\mathcal{E}_3. \end{aligned} \tag{31}$$

Substitute this lower bound on J into the energy identity:

$$\begin{aligned}\frac{1}{S} - S \le D - S &\le \frac{X}{2} - \frac{\beta}{1-\beta} J + \mathcal{E}_1 \\ &\le \frac{a + 1 - \beta}{2(a+1+\beta)} X + \mathcal{E}_1 + \frac{\beta[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)(1+a+\beta)} \\ &\le \frac{\theta_1(a+1-\beta)}{2(a+1+\beta)}(1-S^2) + \mathcal{E}_1 + \frac{\beta[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)(1+a+\beta)}.\end{aligned}$$

Put $v = \min\{1, 2(a+1+\beta)/[\theta_1(a+1-\beta)]\}$. If $S < v$, then $S^{-1} - v^{-1} = (v-S)/(Sv) \ge v - S$ and $1 - S^2 \ge v - S$, so the preceding inequality implies

$$\begin{aligned}(v-S)^2 &\le (1-S^2)(S^{-1} - v^{-1}) \\ &\le (1-S^2)\left(S^{-1} - \frac{\theta_1(a+1-\beta)}{2(a+1+\beta)}\right) \le \mathcal{E}_1 + \frac{\beta[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)(1+a+\beta)}.\end{aligned}$$

The conclusion is immediate when $S \ge v$, hence

$$S \ge \min\left\{1, \frac{2(a+1+\beta)}{\theta_1(a+1-\beta)}\right\} - \sqrt{\mathcal{E}_1 + \frac{\beta[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)(1+a+\beta)}}. \tag{32}$$

The bound from the distribution of normalization jumps. Set $u = a/D$, let $\mathcal{L} = \{k_0 \le k < T : \varrho_k > 1 + 2u\}$, and write $f = |\mathcal{L}|/T'$. If the complement of $\mathcal{L}$ is nonempty, Cauchy–Schwarz and $\widehat{F}_k \ge 1$ give

$$\begin{aligned}\frac{1}{T'} \sum_{\substack{k_0 \le k < T \\ k \notin \mathcal{L}}} \frac{1}{\widehat{F}_k} &\ge \frac{(1-f)^2}{(T')^{-1} \sum_{\substack{k_0 \le k < T \\ k \notin \mathcal{L}}} \widehat{F}_k} \\ &\ge \frac{(1-f)^2}{D - f} \ge \frac{1 - 2f}{D}.\end{aligned}$$

The last inequality is equivalent to $f[1 + (D-2)f] \ge 0$. Adding the terms in $\mathcal{L}$ yields

$$S \ge \frac{1}{D} + \frac{1}{T'} \sum_{k \in \mathcal{L}} \left(\frac{1}{\widehat{F}_k} - \frac{2}{D}\right).$$

This bound also holds directly when the complement is empty. For $k \in \mathcal{L}$, the lag bound and $B_{k+1} \le 2\widehat{F}_k$ imply

$$\frac{1}{\widehat{F}_k} - \frac{2}{D} \ge \frac{\varrho_k - 1 - 2u}{a} \ge \frac{(\varrho_k - 1)(\varrho_k - 1 - 2u)B_{k+1}}{2a^2}.$$

If $A > 2uM$, then

$$\begin{aligned} A - 2uM &= \frac{1}{T'} \sum_{k=k_0}^{T-1} (\varrho_k - 1 - 2u) B_{k+1} \\ &\le \frac{1}{T'} \sum_{k \in \mathcal{L}} (\varrho_k - 1 - 2u) B_{k+1}. \end{aligned}$$

Consequently $0 < \sum_{k\in\mathcal{L}} B_{k+1} \le T'M$, so the following denominator is positive. Expanding the product and applying Cauchy–Schwarz gives

$$\begin{aligned} &\frac{1}{T'} \sum_{k \in \mathcal{L}} (\varrho_k - 1)(\varrho_k - 1 - 2u) B_{k+1} \\ &\quad = \frac{1}{T'} \sum_{k \in \mathcal{L}} (\varrho_k - 1 - 2u)^2 B_{k+1} + \frac{2u}{T'} \sum_{k \in \mathcal{L}} (\varrho_k - 1 - 2u) B_{k+1} \\ &\quad \ge \frac{\left[\sum_{k \in \mathcal{L}} (\varrho_k - 1 - 2u) B_{k+1}\right]^2}{T' \sum_{k \in \mathcal{L}} B_{k+1}} + \frac{2u}{T'} \sum_{k \in \mathcal{L}} (\varrho_k - 1 - 2u) B_{k+1} \\ &\quad \ge \frac{(A - 2uM)^2}{M} + 2u(A - 2uM) = \frac{A(A - 2uM)}{M}. \end{aligned}$$

If $A \le 2uM$, the same sum is nonnegative. Combining these bounds gives

$$\frac{a}{2} S \ge \frac{u}{2} + \frac{A(A - 2uM)^+}{4aM}, \tag{33}$$

with the last term defined to be zero when $A \le 2uM$, including $M = 0$.

It remains to express this bound in terms of the spectral data; set $\lambda = a/\theta_1$. First suppose $\mathcal{E}_1 < 1$ and $(1 - \beta)X \ge 2[\mathcal{E}_2 + (1 + a)\mathcal{E}_3]$. Since $\mathcal{E}_3 > 0$, we have $X > 0$, and Equation (31) gives

$$\tau := \frac{2\beta J}{(1 - \beta)X} \ge \frac{\beta}{1 + a + \beta} > 0, \qquad \vartheta := \frac{X}{\theta_1} \in (0, 1].$$

Consider first $\tau \le 1$. Since $J = (1 - \beta)X\tau/(2\beta) > 0$, Equation (29), $|M - J| \le \mathcal{E}_3$, and $u \le a$ give

$$\begin{aligned} |A - (1 - \beta)X b_\tau| &\le \mathcal{E}_2 + \mathcal{E}_3, \\ \left| A - 2uM - (1 - \beta)X \left(b_\tau - \frac{u\tau}{\beta} \right) \right| &\le \mathcal{E}_2 + (1 + 2a)\mathcal{E}_3. \end{aligned}$$

The positive part is 1-Lipschitz and $(b_\tau - u\tau/\beta)^+ \le b_\tau^+ \le 1$. Thus the numerator and denominator perturbations can be estimated in one calculation:

$$\begin{aligned} \frac{A(A - 2uM)^+}{M} &\ge \frac{\left((1 - \beta)X b_\tau - \mathcal{E}_2 - \mathcal{E}_3\right)^+}{J + \mathcal{E}_3} \\ &\qquad \cdot \left((1 - \beta)X \left(b_\tau - \frac{u\tau}{\beta} \right) - \mathcal{E}_2 - (1 + 2a)\mathcal{E}_3 \right)^+ \\ &\ge \frac{2\beta(1 - \beta)X}{\tau} b_\tau \left(b_\tau - \frac{u\tau}{\beta} \right)^+ - \frac{2\beta[2\mathcal{E}_2 + (2 + 2a)\mathcal{E}_3]}{\tau} - \frac{4\beta^2 \mathcal{E}_3}{\tau^2}. \end{aligned}$$

The same zero convention applies on the left when $A \le 2uM$. Using $\tau \ge \beta/(1 + a + \beta)$ and $X = \vartheta\theta_1$ in Equation (33) now yields

$$\begin{aligned} \frac{a}{2} S \ge \frac{u}{2} &+ \frac{\beta(1 - \beta)\vartheta\theta_1}{2a\tau} b_\tau \left(b_\tau - \frac{u\tau}{\beta} \right)^+ \\ &- \frac{1 + a + \beta}{a} [\mathcal{E}_2 + (2 + 2a + \beta)\mathcal{E}_3]. \end{aligned}$$

The preceding right side is nondecreasing in $u \geq 0$, with slope at most $1/2$. Where the positive part is active, its slope is $\frac{1}{2}[1-(1-\beta)\vartheta\theta_1 b_\tau/a] \in [\beta/2, 1/2]$; elsewhere it is $1/2$. The energy identity, $S \leq 1$, and $0 < \tau \leq 1$ give

$$D \leq S + \frac{X}{2}(1-\tau) + \mathcal{E}_1 \leq 1 + \frac{\vartheta\theta_1}{2}(1-\tau) + \mathcal{E}_1,$$
$$u = \frac{a}{D} \geq \frac{a}{1+\vartheta\theta_1(1-\tau)/2 + \mathcal{E}_1} \geq \lambda u_{\tau,\vartheta}(1-\mathcal{E}_1).$$

The last inequality uses $\mathcal{E}_1 < 1$ and $\lambda u_{\tau,\vartheta} = a/[1+\vartheta\theta_1(1-\tau)/2]$. Moving the error term to the left, the slope bound permits substitution of $\lambda u_{\tau,\vartheta}$ at a cost of at most $\lambda u_{\tau,\vartheta}\mathcal{E}_1/2$:

$$\begin{aligned}
&\frac{a}{2}S + \frac{1+a+\beta}{a}[\mathcal{E}_2 + (2+2a+\beta)\mathcal{E}_3] \\
&\quad \geq \frac{\lambda u_{\tau,\vartheta}}{2} + \frac{\beta(1-\beta)\vartheta\theta_1}{2a\tau} b_\tau \left(b_\tau - \frac{\lambda u_{\tau,\vartheta}\tau}{\beta}\right)^+ - \frac{\lambda u_{\tau,\vartheta}}{2}\mathcal{E}_1 \\
&\quad = \lambda \left\{ \frac{u_{\tau,\vartheta}}{2} + \frac{\beta(1-\beta)\vartheta}{2\lambda^2\tau} b_\tau \left(b_\tau - \frac{\lambda u_{\tau,\vartheta}\tau}{\beta}\right)^+ \right\} - \frac{\lambda u_{\tau,\vartheta}}{2}\mathcal{E}_1 \\
&\quad \geq \lambda \mathcal{G}_\beta(\theta_1;\lambda) - \frac{a}{2}\mathcal{E}_1.
\end{aligned}$$

The equality uses $a = \lambda\theta_1$; the last inequality uses the defining infimum and $\lambda u_{\tau,\vartheta} \leq \lambda\theta_1 = a$. The omitted cases only require the bound $\mathcal{G}_\beta(\theta_1;\lambda) \leq \theta_1/2$, obtained by taking $\tau = 1$. If $\mathcal{E}_1 \geq 1$, the desired speed bound is trivial; if $(1-\beta)X < 2[\mathcal{E}_2 + (1+a)\mathcal{E}_3]$, then Equation (30) gives

$$1 - S \leq \frac{X}{\theta_r} < \frac{2[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)\theta_r}.$$

In the remaining case $\tau > 1$, the energy identity gives

$$1 - S \leq 1 - S^2 \leq S(D-S) \leq S\left[\frac{X}{2}(1-\tau) + \mathcal{E}_1\right] \leq \mathcal{E}_1.$$

Thus all cases give

$$\begin{aligned}
S \geq \frac{2}{\theta_1}\mathcal{G}_\beta\left(\theta_1; \frac{a}{\theta_1}\right) - \max\Bigg\{ &\mathcal{E}_1 + \frac{2(1+a+\beta)}{a^2}[\mathcal{E}_2 + (2+2a+\beta)\mathcal{E}_3], \\
&\frac{2[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)\theta_r} \Bigg\}.
\end{aligned} \tag{34}$$

Actual progress and the residual. The normalization comparison gives

$$\left|\frac{h_{k+1}}{F_k} - \frac{1}{\widehat{F}_k}\right| = \left|\frac{h_{k+1}-1}{F_k} + \frac{\widehat{F}_k - F_k}{F_k\widehat{F}_k}\right| \leq \frac{2|h_{k+1}-1|}{h_-}.$$

Combining the errors in Equations (32) and (34), define

$$\begin{aligned}
\mathcal{R}_T := \max\Bigg\{ &\sqrt{\mathcal{E}_1 + \frac{\beta[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)(1+a+\beta)}}, \\
&\mathcal{E}_1 + \frac{2(1+a+\beta)}{a^2}[\mathcal{E}_2 + (2+2a+\beta)\mathcal{E}_3], \quad \frac{2[\mathcal{E}_2 + (1+a)\mathcal{E}_3]}{(1-\beta)\theta_r} \Bigg\} + \frac{2\overline{\nu}}{h_-} + \frac{k_0}{T}.
\end{aligned} \tag{35}$$

Since $0 < h_{k+1}/F_k \leq 1$, restoring the prefix and using the two bounds on S yields

$$\begin{aligned}
\frac{1}{T}\sum_{k=0}^{T-1}\frac{h_{k+1}}{F_k} &\geq \frac{1}{T}\sum_{k=k_0}^{T-1}\frac{h_{k+1}}{F_k} \geq \frac{1}{T'}\sum_{k=k_0}^{T-1}\frac{h_{k+1}}{F_k} - \frac{k_0}{T} \\
&\geq S - \frac{2}{h_- T'}\sum_{k=k_0}^{T-1}|h_{k+1}-1| - \frac{k_0}{T} \\
&\geq \mathcal{S}(\beta,\theta_1;a) - \mathcal{R}_T.
\end{aligned}$$

This proves Equation (23). For $0 \le \delta \le 1$, the lag a is uniformly bounded with the declared dependencies, and $\overline{\nu}, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$ are $O((T')^{-1} + \delta)$. Since $a \ge \theta_1 > 0$ and $(T')^{-1} + \delta \le 2$, the explicit residual has the claimed order.

Dependence on the lag parameter. Fix $0 < \theta_{\mathrm{lo}} \le \theta_{\mathrm{hi}}$, let $\theta \in [\theta_{\mathrm{lo}}, \theta_{\mathrm{hi}}]$, and denote the objective in Equation (22) by $\Phi_\lambda(\tau, \vartheta)$, suppressing its dependence on θ. For each fixed θ, it is nonincreasing in λ and continuous for fixed (τ, ϑ), so its infimum is nonincreasing and upper semicontinuous; together these imply left-continuity. Fix $\lambda_0 \ge 1$ and restrict to $\lambda \in [\lambda_0, 2\lambda_0]$. The positive part is 1-Lipschitz, and $u_{\tau,\vartheta} \le \theta$, $b_\tau^+ \le 1$ give

$$\begin{aligned} 0 \le \Phi_{\lambda_0}(\tau, \vartheta) - \Phi_\lambda(\tau, \vartheta) &\le \frac{\beta(1-\beta)\vartheta b_\tau^+}{2\tau} \left[\frac{(\lambda - \lambda_0) u_{\tau,\vartheta} \tau}{\beta \lambda_0^2} + \left(\frac{1}{\lambda_0^2} - \frac{1}{\lambda^2} \right) b_\tau^+ \right] \\ &\le (\lambda - \lambda_0) \left[\frac{(1-\beta)\theta}{2} + \frac{3\beta(1-\beta)\vartheta}{2\tau} \right]. \end{aligned}$$

Set $\tau_* := \min\{\beta/[2(1+\beta)], \beta/(8\lambda_0\theta_{\mathrm{hi}})\}$. On the sublevel set $\Phi_\lambda \le \theta/2 + 1$, the ratio ϑ/τ has a bound uniform over the entire curvature interval. Indeed, if $\tau \le \tau_*$, then $b_\tau \ge 3/4$ and $\lambda u_{\tau,\vartheta}\tau/\beta \le 1/4$, whence

$$\frac{\theta_{\mathrm{hi}}}{2} + 1 \ge \Phi_\lambda(\tau, \vartheta) \ge \frac{\beta(1-\beta)\vartheta}{32\lambda_0^2 \tau}.$$

For $\tau > \tau_*$, one has $\vartheta/\tau \le \tau_*^{-1}$. Thus on this sublevel set,

$$\frac{\vartheta}{\tau} \le M := \max\left\{ \frac{32\lambda_0^2(\theta_{\mathrm{hi}}/2 + 1)}{\beta(1-\beta)}, \frac{1}{\tau_*} \right\}.$$

Since $\mathcal{G}_\beta(\theta; \lambda) \le \theta/2$, for each $0 < \epsilon < 1$ choose a point in this sublevel set with $\Phi_\lambda(\tau, \vartheta) \le \mathcal{G}_\beta(\theta; \lambda) + \epsilon$. With $C_{\mathrm{G}} := (1-\beta)\theta_{\mathrm{hi}}/2 + 3\beta(1-\beta)M/2$, the preceding comparison gives

$$\begin{aligned} \mathcal{G}_\beta(\theta; \lambda_0) \le \Phi_{\lambda_0}(\tau, \vartheta) &\le \Phi_\lambda(\tau, \vartheta) + C_{\mathrm{G}}(\lambda - \lambda_0) \\ &\le \mathcal{G}_\beta(\theta; \lambda) + \epsilon + C_{\mathrm{G}}(\lambda - \lambda_0). \end{aligned}$$

Letting $\epsilon \downarrow 0$ gives

$$0 \le \mathcal{G}_\beta(\theta; \lambda_0) - \mathcal{G}_\beta(\theta; \lambda) \le C_{\mathrm{G}}(\lambda - \lambda_0),$$

uniformly in $\theta \in [\theta_{\mathrm{lo}}, \theta_{\mathrm{hi}}]$. In particular, taking $\theta_{\mathrm{lo}} = \theta_{\mathrm{hi}} = \theta$ and then $\lambda \downarrow \lambda_0$ proves right-continuity at every $\lambda_0 \ge 1$ for each $\theta > 0$.

For Equation (24), set $\lambda_0 = 1$ in this estimate. The first entry of the maximum defining $(\theta/2)\mathcal{S}(\beta, \theta; \lambda\theta)$ is $\min\{\theta/2, (\lambda\theta + 1 + \beta)/(\lambda\theta + 1 - \beta)\}$. The minimum with $\theta/2$ is a nondecreasing 1-Lipschitz function, and

$$\begin{aligned} 0 \le \frac{\theta + 1 + \beta}{\theta + 1 - \beta} - \frac{\lambda\theta + 1 + \beta}{\lambda\theta + 1 - \beta} &= \frac{2\beta\theta(\lambda - 1)}{(\theta + 1 - \beta)(\lambda\theta + 1 - \beta)} \\ &\le \frac{2\beta\theta_{\mathrm{hi}}}{(1-\beta)^2}(\lambda - 1). \end{aligned}$$

The second entry is $\mathcal{G}_\beta(\theta; \lambda)$, whose decrement is bounded by $C_{\mathrm{G}}(\lambda - 1)$. Taking the maximum of these two nonincreasing entries bounds its decrement by the larger of their bounds. Thus $C := \max\{1, C_{\mathrm{G}}, 2\beta\theta_{\mathrm{hi}}/(1-\beta)^2\}$ proves Equation (24) with the stated dependencies. □

E Proof for quadratic loss

The quadratic analysis combines the exact plain dynamics with an averaged analysis of the momentum recurrence through the shared finite-window bound in Section D, whose elementary storage calculation is extended to changing geometry in the general-loss proof.

Throughout this section, we use the setup and notation of Section 4.2. Let $r \ge 1$ be the rank of H and $\lambda_1 \ge \cdots \ge \lambda_r > 0$ its positive eigenvalues counted with multiplicity, write $\rho_i := \eta\lambda_i/\|b\|_2$, and let $z_k := Hx_k/\|b\|_2$ denote the sharp-direction state. Let $e := b/\|b\|_2$ and $\Theta := \eta H/\|b\|_2$, whose positive eigenvalues are $\rho_1, \ldots, \rho_r$. Let $\ell_k := -\langle x_k, e \rangle$ denote the downstream coordinate.

The curved component of x_k is $H^\dagger H x_k = \|b\|_2 H^\dagger z_k$, and $b^\top x_k = -\|b\|_2 \ell_k$, giving

$$\begin{aligned} L(x_k) &= \frac{1}{2}(Hx_k)^\top H^\dagger (Hx_k) + b^\top x_k \\ &= \frac{\|b\|_2^2}{2} z_k^\top H^\dagger z_k - \|b\|_2 \ell_k, \\ V_T &= \|b\|_2 \frac{\ell_T - \ell_0}{\eta T} + \frac{\|b\|_2^2}{2\eta T}\big[z_0^\top H^\dagger z_0 - z_T^\top H^\dagger z_T\big]. \end{aligned} \tag{36}$$

The last identity follows by substituting the loss decomposition into $V_T = [L(x_0) - L(x_T)]/(\eta T)$. For bounded z_k, the endpoint term is $O(T^{-1})$, so the asymptotic loss speed is determined by downstream progress.

E.1 Proof for NGD dynamics

Define $q(z) := \sqrt{1 + \|z\|_2^2}$ on $\mathrm{Range}(H)$. Since $z_k \perp e$, we have $\|b + Hx_k\|_2 = \|b\|_2 q(z_k)$, and the Euclidean update is $x_{k+1} = x_k - \eta(e + z_k)/q(z_k)$. Applying $H/\|b\|_2$ and $-\langle \cdot, e\rangle$, with $He = 0$, gives

$$\begin{aligned} z_{k+1} &= \left(I - \frac{\Theta}{q(z_k)}\right) z_k, \\ \ell_{k+1} - \ell_k &= \frac{\eta}{q(z_k)}. \end{aligned} \tag{37}$$

The normalization is defined because $q(z_k) \geq 1$.

For each distinct positive eigenvalue ρ of Θ, let P_ρ denote its orthogonal spectral projector. We say that the eigenmode associated with ρ *survives* if $P_\rho z_k \not\to 0$ as $k \to \infty$.

Theorem E.1 (Global plain dynamics)**.** *Consider the plain recurrence in Equation* (37) *with arbitrary* $z_0 \in \mathrm{Range}(H)$*. If no mode survives, then* $z_k \to 0$ *and* $V_T \to \|b\|_2$*. Otherwise, there is exactly one mode that survives with eigenvalue* $\rho_\sharp > 2$ *being the largest eigenvalue* ρ *for which* $P_\rho z_k \neq 0$ *at every iteration. Let amplitude* $a := \sqrt{\rho_\sharp^2/4 - 1}$ *and the unit vector* $v := P_{\rho_\sharp} z_0 / \|P_{\rho_\sharp} z_0\|_2$*. Then, there exists* $\sigma \in \{-1, 1\}$ *such that*

$$z_{2j} \longrightarrow \sigma a v, \quad z_{2j+1} \longrightarrow -\sigma a v, \quad V_T \longrightarrow \frac{2\|b\|_2}{\rho_\sharp}.$$

Moreover, if $\rho_1 > 2$*,* $\|z_0\|_2 < \sqrt{\rho_1^2 - 1}$*, and* $P_{\rho_1} z_0 \neq 0$*, then the top mode survives and* $\rho_\sharp = \rho_1$*. For* $\rho_1 > 2$*, the set of initial curved states whose top mode does not survive has Lebesgue measure zero in* $\mathrm{Range}(H)$*.*

In general, the surviving curvature $\rho_\sharp$ can be smaller than the original largest curvature ρ_1. The comparison with $2\|b\|_2/\rho_1$ therefore requires the original top mode to survive.

Proof of Theorem E.1. We identify the limit set with a coercive Lyapunov function, determine which mode survives, and then recover the loss speed. Define $f(z) := [I - \Theta/q(z)]z$ on $\mathrm{Range}(H)$. The spectral recurrence in Equation (37) gives $P_\rho f(z) = [1 - \rho/q(z)]P_\rho z$. Thus a nonzero mode is eliminated exactly when $q(z) = \rho$, and an eliminated mode remains zero.

Define the Lyapunov function $\mathcal{J}(z) := z^\top \Theta^\dagger z - q(z)$ on $\mathrm{Range}(H)$. It is coercive because

$$\mathcal{J}(z) \geq \frac{\|z\|_2^2}{\rho_1} - \sqrt{1 + \|z\|_2^2} \longrightarrow \infty \qquad \text{as } \|z\|_2 \to \infty.$$

Write $z^+ := f(z)$. Since $z \in \mathrm{Range}(H)$, we have $\Theta^\dagger \Theta z = z$. Using symmetry of $\Theta^\dagger$ and $z^+ - z = -\Theta z/q(z)$ gives

$$\begin{aligned}\mathcal{J}(z^+) - \mathcal{J}(z) &= (z^+ - z)^\top \Theta^\dagger (z^+ + z) - q(z^+) + q(z) \\ &= -\frac{z^\top (z^+ + z)}{q(z)} - q(z^+) + q(z) \\ &= \frac{1 - z^\top z^+ - q(z)q(z^+)}{q(z)} \\ &= \frac{\langle (1, z), (1, -z^+) \rangle - q(z)q(z^+)}{q(z)} \le 0.\end{aligned}$$

The third equality uses $q(z)^2 = 1 + \|z\|_2^2$. The inequality is Cauchy–Schwarz in $\mathbb{R} \times \mathrm{Range}(H)$. Equality holds exactly when $(1, -z^+) = c(1, z)$ for some $c \ge 0$; the first coordinate forces $c = 1$, so equality is equivalent to $z^+ = -z$.

By monotonicity, $\mathcal{J}(z_k) \le \mathcal{J}(z_0)$ for every k, so the reduced orbit remains in the sublevel set

$$\mathcal{K} := \{z \in \mathrm{Range}(H) : \mathcal{J}(z) \le \mathcal{J}(z_0)\}.$$

Continuity of $\mathcal{J}$ makes $\mathcal{K}$ closed, and coercivity makes it bounded. Since $\mathrm{Range}(H)$ is finite-dimensional, $\mathcal{K}$ is compact. Thus (z_k) is bounded and its orbit is precompact: its closure is a closed subset of the compact set $\mathcal{K}$. Moreover, $\mathcal{J}$ is bounded below on $\mathcal{K}$, so the nonincreasing sequence $\mathcal{J}(z_k)$ converges to a finite value $\mathcal{J}_\infty$.

Let w be any limit point, and choose a subsequence $z_{k_j} \to w$. Since $q(z) \ge 1$, the map f is continuous, and hence $z_{k_j+1} = f(z_{k_j}) \to f(w)$. Both k_j and $k_j + 1$ tend to infinity, so continuity of $\mathcal{J}$ gives

$$\begin{aligned}\mathcal{J}(w) &= \lim_{j \to \infty} \mathcal{J}(z_{k_j}) = \mathcal{J}_\infty, \\ \mathcal{J}(f(w)) &= \lim_{j \to \infty} \mathcal{J}(z_{k_j+1}) = \mathcal{J}_\infty.\end{aligned}$$

Therefore $\mathcal{J}(f(w)) = \mathcal{J}(w)$, and the equality condition above yields $f(w) = -w$.

For a nonzero equality point w, the identity $f(w) = -w$ gives $\Theta w = 2q(w)w$. Thus $w \in \mathrm{Range}(P_\rho)$ for some $\rho > 2$, with $\|w\|_2^2 = \rho^2/4 - 1$, and

$$\mathcal{J}(w) = \frac{\rho^2/4 - 1}{\rho} - \frac{\rho}{2} = -\frac{\rho}{4} - \frac{1}{\rho}.$$

In contrast, $\mathcal{J}(0) = -1$. The displayed Lyapunov value is strictly decreasing for $\rho > 2$. The single limiting value of $\mathcal{J}(z_k)$ therefore confines the limit set to the origin or to one eigensphere.

If the limit set is the origin, then $z_k \to 0$ and no mode survives. Otherwise, let $\rho_\sharp > 2$ denote the eigenvalue of the limiting eigensphere. Every spectral projection remains on its initial line. The multiplier of $P_{\rho_\sharp} z_k$ tends to -1, while $\|P_{\rho_\sharp} z_k\|_2 \to a$ and $P_\rho z_k \to 0$ for $\rho \ne \rho_\sharp$. The multiplier is eventually negative, so the even and odd subsequences converge to the asserted antipodal phases. Thus $\rho_\sharp$ is the unique surviving eigenvalue. Every mode with $\rho > \rho_\sharp$ must be initially absent or eliminated at a finite iteration. Otherwise, its nonzero projection would have multiplier modulus converging to $|1 - 2\rho/\rho_\sharp| > 1$ and would grow geometrically while the full orbit remains bounded. Since an eliminated mode remains zero, $\rho_\sharp$ is the largest eigenvalue whose projection is nonzero at every iteration.

To recover the loss speed, combine Equations (36) and (37). Boundedness makes the curved endpoint contribution vanish, so

$$V_T = \frac{\|b\|_2}{T} \sum_{k=0}^{T-1} \frac{1}{q(z_k)} + o(1).$$

When a mode survives, we have $q(z_k) \to \rho_\sharp/2$, and Cesàro averaging gives $V_T \to 2\|b\|_2/\rho_\sharp$. When $z_k \to 0$, including finite termination at zero, we have $q(z_k) \to 1$ and hence $V_T \to \|b\|_2$.

It remains to prove the two top-mode survival statements for initial states in $\mathrm{Range}(H)$. For $\rho_1 > 2$, a top projection that is nonzero at every iteration cannot coexist with $z_k \to 0$: its multiplier modulus would tend to $|1-\rho_1| > 1$, contradicting boundedness. The classification above therefore makes top-mode survival equivalent to the top projection being nonzero at every iteration. For $0 < \|z\|_2 < \sqrt{\rho_1^2 - 1}$, we have $q(z) < \rho_1$ and $|q(z) - \rho| < \rho_1$ for every positive eigenvalue ρ. Consequently,

$$\begin{aligned}\|f(z)\|_2^2 &= \sum_{\rho>0}\left(1-\frac{\rho}{q(z)}\right)^2 \|P_\rho z\|_2^2 < \frac{\rho_1^2}{q(z)^2}\|z\|_2^2 \\ &= \rho_1^2\big[1 - q(z)^{-2}\big] < \rho_1^2 - 1.\end{aligned}$$

The origin is also fixed, so this open ball is forward invariant. The strict bound $q(z_k) < \rho_1$ prevents elimination of an active top mode. For $\rho_1 > 2$, this proves the claimed top-mode survival.

For generic survival, note that for each $k \geq 0$ the function $z \mapsto 1 + \|f^k(z)\|_2^2 - \rho_1^2$ is real analytic and takes the nonzero value $1 - \rho_1^2$ at the origin. Its zero set is therefore Lebesgue-null. The exceptional set is contained in the proper subspace $\mathrm{Range}(H) \cap \mathrm{Ker}(P_{\rho_1})$ together with the countable union of these zero sets, so it is null. □

E.2 Proof for NGM dynamics

Since $\langle b + Hx_k, e\rangle = \|b\|_2$, projecting the EMA update in Equation (1) onto e from $m_0 = 0$ gives $\langle m_k, e\rangle/\|b\|_2 = (1-\beta)\sum_{j=0}^{k-1}\beta^j = 1 - \beta^k$. We can therefore write $m_k = \|b\|_2[p_k + (1-\beta^k)e]$, where $p_k \in \mathrm{Range}(H)$ and $p_0 = 0$. Set $F_k := \sqrt{(1-\beta^{k+1})^2 + \|p_{k+1}\|_2^2}$ for $k \geq 0$. Then $\|m_{k+1}\|_2 = \|b\|_2 F_k$, so $x_{k+1} = x_k - \eta[p_{k+1} + (1-\beta^{k+1})e]/F_k$. Projecting the buffer and position updates gives

$$\begin{aligned}p_{k+1} &= \beta p_k + (1-\beta)z_k, \\ z_{k+1} &= z_k - \Theta\frac{p_{k+1}}{F_k}, \\ \ell_{k+1} - \ell_k &= \eta\frac{1-\beta^{k+1}}{F_k}.\end{aligned} \tag{38}$$

For $y = (z, p) \in \mathrm{Range}(H) \times \mathrm{Range}(H)$ and $h > 0$, define $\Phi_\beta(y; h) := (z^+, p^+)$ by updating the buffer and then the position:

$$\begin{aligned}p^+ &:= (1-\beta)z + \beta p, \\ z^+ &:= z - \Theta\frac{p^+}{\sqrt{h^2 + \|p^+\|_2^2}}.\end{aligned} \tag{39}$$

Here h is the updated tangential buffer coefficient used in the position normalization, while the subscript β records the momentum parameter. Thus $y_k := (z_k, p_k)$ satisfies $y_{k+1} = \Phi_\beta(y_k; 1-\beta^{k+1})$ and $y_0 = (z_0, 0)$. Since $1 - \beta^{k+1} \to 1$, the map $y \mapsto \Phi_\beta(y; 1)$ is the autonomous limiting map.

Global progress without recurrent-branch selection. The next theorem provides the momentum speed bound that, together with Theorem E.1, yields the strict comparison in Section 4.2.

Theorem E.2 (Global branch-free momentum speed)**.** *Under the setup of Section 4.2, assume* $0 < \beta < 1$ *and* $\rho_1 > 2$ *and consider the momentum recurrence in Equation* (38) *from* $m_0 = 0$*. For every initialization* x_0*, the reduced state* (z_k, p_k) *is bounded and*

$$\liminf_{T\to\infty} V_T \geq \frac{2\|b\|_2}{\rho_1}\max\left\{\min\left\{\frac{\rho_1}{2}, 1 + \frac{2\beta}{\rho_1 + 1 - \beta}\right\},\, G_\beta(\rho_1)\right\} > \frac{2\|b\|_2}{\rho_1}, \tag{40}$$

where, with $b_\tau := 1 - (1+\beta)\tau/(2\beta)$ *and* $u_{\tau,\vartheta} := 2/[\vartheta(1-\tau) + 2/\rho_1]$*,*

$$G_\beta(\rho_1) := \inf_{0<\tau<1,\ 0<\vartheta\leq 1}\left\{\frac{u_{\tau,\vartheta}}{2} + \frac{\beta(1-\beta)\vartheta}{2\tau}\, b_\tau\Big(b_\tau - \frac{u_{\tau,\vartheta}\,\tau}{\beta}\Big)^+\right\}. \tag{41}$$

This result does not assume a momentum period, invariant law, or entry into a selected basin. The gain $G_\beta(\rho_1)$ depends only on β and ρ_1, not on the dimension or on the rest of the spectrum; it is

nondecreasing in ρ_1 and converges to $G_\beta := \lim_{\rho_1\to\infty} G_\beta(\rho_1)$. For $\beta \le 0.76$,

$$G_\beta = \frac{3+\beta+\sqrt{\beta^2+22\beta+25}}{8},$$

while for every β one has $G_\beta \ge (3\beta-1)/2 + \sqrt{(1-\beta)(1+\beta^2)}$. Numerically, $G_\beta \approx 1.190, 1.262, 1.295, 1.284, 1.237$ for $\beta = 0.5, 0.7, 0.8, 0.9, 0.95$. The elementary factor in Equation (40) gives an explicit strict gain for every $\rho_1 > 2$, whereas $G_\beta(\rho_1)$ retains a gain bounded away from one as $\rho_1 \to \infty$ with β fixed. The constant is not sharp. The guarantee concerns asymptotic average loss speed and does not assert finite-horizon trajectory dominance.

Proof of Theorem E.2. We first bound the sharp-direction state and buffer, then apply Lemma D.1 and send the window length and prefix length to infinity, in that order.

Step 1: Boundedness. Set $d_k := z_k - p_k$, so $p_{k+1} = z_k - \beta d_k$ and $d_0 = z_0$. Since $F_k \ge 1-\beta^{k+1} \ge 1-\beta > 0$ and $\|p_{k+1}\|_2 \le F_k$, Equation (38) gives

$$d_{k+1} = z_k - \Theta\frac{p_{k+1}}{F_k} - p_{k+1} = \beta d_k - \Theta\frac{p_{k+1}}{F_k},$$
$$|d_{i,k}| \le \beta^k |z_{i,0}| + \rho_i \sum_{j=0}^{k-1} \beta^j = \beta^k |z_{i,0}| + \rho_i \frac{1-\beta^k}{1-\beta}.$$

Here $d_{i,k}, z_{i,k}, p_{i,k}$ denote the coordinates in an orthonormal eigenbasis of H on $\mathrm{Range}(H)$. For each i, set $R_i := \beta|z_{i,0}| + \rho_i/(1-\beta)$, so $\beta|d_{i,k}| \le R_i - \rho_i$. If $|z_{i,k}| \le R_i - \rho_i$, the step bound $|z_{i,k+1} - z_{i,k}| \le \rho_i$ gives $|z_{i,k+1}| \le R_i$. Otherwise, $p_{i,k+1} = z_{i,k} - \beta d_{i,k}$ has the sign of $z_{i,k}$, so the update moves toward zero and any overshoot has magnitude at most $\rho_i \le R_i$. Thus $|z_{i,k+1}| \le \max\{|z_{i,k}|, R_i\}$, and induction bounds the position. The buffer is an exponentially weighted average of these positions, so with $P := \left(\sum_i \max\{|z_{i,0}|, R_i\}^2\right)^{1/2}$,

$$\|z_k\|_2^2 \le \sum_i \max\{|z_{i,0}|, R_i\}^2 = P^2,$$
$$\|p_{k+1}\|_2 = \|(1-\beta)\sum_{j=0}^{k} \beta^{k-j} z_j\|_2 \le (1-\beta)\sum_{j=0}^{k} \beta^{k-j}\|z_j\|_2 \le P.$$

Together with $p_0 = 0$, these bounds hold at every index and also bound F_k.

Step 2: Applying the storage bound and recovering loss speed. Fix integers $T \ge 2$ and $0 \le k_0 < T$. Eliminating z_k from the buffer update gives the recurrence and sharp-direction states required by Lemma D.1:

$$p_{k+2} - (1+\beta)p_{k+1} + \beta p_k = (1-\beta)(z_{k+1} - z_k) = -(1-\beta)\Theta\frac{p_{k+1}}{F_k},$$
$$\frac{p_{k+1} - \beta p_k}{1-\beta} = z_k.$$

Apply that lemma to this fixed matrix and buffer, with spectral endpoints $(\theta_r, \theta_1) = (\rho_r, \rho_1)$, $h_k = 1-\beta^k$, and $\delta = 0$. The tangential recurrence is exact, $h_{k+1} = \beta h_k + 1 - \beta$, and $h_{k+1} \ge 1-\beta =: h_- > 0$, so its normalization is exactly F_k. Step 1 verifies all bounds on the buffer and sharp-direction state, including the look-ahead index, with $Z = P$. The lag parameter is $\rho_1 + 2P\beta^{k_0+1}$, so Equation (23) yields

$$\frac{1}{T}\sum_{k=0}^{T-1} \frac{1-\beta^{k+1}}{F_k} \ge \mathcal{S}(\beta, \rho_1; \rho_1 + 2P\beta^{k_0+1}) - O\big((T-k_0)^{-1/2} + k_0/T\big),$$

where the implicit constant depends on β, Θ, and the initialization through P, but not on T or k_0. Combining this with Equations (36) and (38) gives

$$\begin{aligned}\frac{V_T}{\|b\|_2} &= \frac{1}{T}\sum_{k=0}^{T-1}\frac{1-\beta^{k+1}}{F_k} + \frac{z_0^\top\Theta^\dagger z_0 - z_T^\top\Theta^\dagger z_T}{2T} \\ &= \frac{1}{T}\sum_{k=0}^{T-1}\frac{1-\beta^{k+1}}{F_k} + O(T^{-1}) \\ &\geq \mathcal{S}(\beta,\rho_1;\rho_1+2P\beta^{k_0+1}) - O\big((T-k_0)^{-1/2}+k_0/T\big).\end{aligned}$$

Here $\Theta^\dagger = \|b\|_2 H^\dagger/\eta$, and Step 1 bounds the endpoint term with the same constant dependencies. First let $T\to\infty$ with k_0 fixed, and then let $k_0\to\infty$. The elementary term of $\mathcal{S}$ is continuous in its lag parameter, and the continuity clause of Lemma D.1 gives the same property for its jump term. Since $\rho_1 + 2P\beta^{k_0+1}\downarrow\rho_1$, these limits give

$$\begin{aligned}\liminf_{T\to\infty} V_T &\geq \lim_{k_0\to\infty}\|b\|_2\mathcal{S}(\beta,\rho_1;\rho_1+2P\beta^{k_0+1}) \\ &= \|b\|_2\,\mathcal{S}(\beta,\rho_1;\rho_1) \\ &= \frac{2\|b\|_2}{\rho_1}\max\Big\{\min\Big\{\frac{\rho_1}{2},\frac{\rho_1+1+\beta}{\rho_1+1-\beta}\Big\},\,G_\beta(\rho_1)\Big\},\end{aligned}$$

where $\mathcal{G}_\beta(\rho_1;1) = G_\beta(\rho_1)$ was noted after Equation (22). Since $(\rho_1+1+\beta)/(\rho_1+1-\beta) = 1+2\beta/(\rho_1+1-\beta)$, this is the bound in Equation (40), and for $\rho_1 > 2$ both entries of the minimum exceed one, which gives the strict inequality. The same argument applies when $\rho_1\leq 2$: the elementary term of $\mathcal{S}(\beta,\rho_1;\rho_1)$ equals one, while each progress fraction $(1-\beta^{k+1})/F_k$ is at most one, so $V_T\to\|b\|_2$.

Step 3: Evaluation of the large-curvature gain. To minimize over ϑ, fix $w := \vartheta(1-\tau)$, so $u_{\tau,\vartheta} = 2/(w+2/\rho_1)$ is fixed. The second term in Equation (41) satisfies

$$\frac{\beta(1-\beta)\vartheta}{2\tau}\,b_\tau\Big(b_\tau - \frac{u_{\tau,\vartheta}\tau}{\beta}\Big)^+ = \frac{\beta(1-\beta)w}{2}\,\frac{b_\tau}{1-\tau}\Big(\frac{b_\tau}{\tau} - \frac{2}{\beta(w+2/\rho_1)}\Big)^+.$$

While this term is positive, both factors after the coefficient decrease with τ, since $b_\tau/(1-\tau) = 1-(1-\beta)\tau/[2\beta(1-\tau)]$ and $b_\tau/\tau = 1/\tau - (1+\beta)/(2\beta)$. It remains zero after the positive part vanishes. Thus increasing τ to $1-w$ cannot increase the objective and gives $\vartheta = 1$. The infimum may therefore be restricted to $\vartheta = 1$ for every ρ_1. The objective in Equation (22) is nondecreasing in its curvature argument, so $G_\beta(\rho_1) = \mathcal{G}_\beta(\rho_1;1)$ is nondecreasing in ρ_1. For fixed β, the minimizing sublevel sets remain in a common compact subinterval of $(0,1)$ as $\rho_1\to\infty$: the penalty diverges as $\tau\downarrow 0$, uniformly for large ρ_1, and the first term is $1/(1-\tau+2/\rho_1)$. Uniform convergence on that compact interval then identifies the limit with the infimum of the limiting objective. The limit is strictly greater than one, since its objective is at least $1/(1-\tau) > 1$ and attains its minimum inside $(0,1)$.

Let τ_β be the smaller root of $(1+\beta)\tau^2 - (5+3\beta)\tau + 2\beta = 0$, where the limiting positive part vanishes. On $0<\tau\leq\tau_\beta$, the limiting objective is

$$h(\tau) := \frac{\beta(1-\beta)}{2\tau} + \frac{1+\beta^2}{2\beta(1-\tau)} + \frac{(1-\beta)(1+\beta)^2}{8\beta}(\tau-4),$$

and for $\tau_\beta\leq\tau<1$ it is $1/(1-\tau)$. Both pieces are strictly convex, and the derivative jumps upward at τ_β, so the full objective is convex. Its second piece is increasing, hence $G_\beta = \min_{0<\tau\leq\tau_\beta} h(\tau)$. At the contact point, differentiate and use $\beta = \tau_\beta(5-\tau_\beta)/[(1-\tau_\beta)(2-\tau_\beta)]$ to obtain

$$\begin{aligned}h'(\tau_\beta) &= -\frac{\beta(1-\beta)}{2\tau_\beta^2} + \frac{1+\beta^2}{2\beta(1-\tau_\beta)^2} + \frac{(1-\beta)(1+\beta)^2}{8\beta} \\ &= \frac{\tau_\beta^4 + 4\tau_\beta^3 - 41\tau_\beta^2 + 54\tau_\beta - 10}{\tau_\beta(5-\tau_\beta)(2-\tau_\beta)(1-\tau_\beta)^3}.\end{aligned}$$

For $\beta \le 0.76$, the contact quadratic is negative at $\tau = 0.221$, hence $\tau_\beta < 0.221$. The numerator in this display is increasing on $[0, 0.221]$, since its derivative is at least $54 - 82(0.221) > 0$, and it is negative at 0.221. Thus $h'(\tau_\beta) < 0$, and convexity gives

$$G_\beta = h(\tau_\beta) = \frac{1}{1-\tau_\beta} = \frac{3+\beta+\sqrt{\beta^2+22\beta+25}}{8}.$$

For every $\beta \in (0,1)$, Cauchy–Schwarz applied to the first two terms of h and nonnegativity of its term proportional to τ give

$$\begin{aligned} G_\beta &\ge \left(\sqrt{\frac{\beta(1-\beta)}{2}} + \sqrt{\frac{1+\beta^2}{2\beta}}\right)^2 - \frac{(1-\beta)(1+\beta)^2}{2\beta} \\ &= \frac{3\beta-1}{2} + \sqrt{(1-\beta)(1+\beta^2)}. \end{aligned}$$

□

Local straightening from a zero momentum buffer. This theorem supplies the momentum conclusion of Theorem 4.2 below the shifted threshold. The shifted-trajectory limit further describes the resulting straight motion.

Theorem E.3 (Subcritical local straightening)**.** *Under the setup of Section 4.2 with $0 < \beta < 1$ and $m_0 = 0$, assume $\rho_1 < 2(1+\beta)/(1-\beta)$. There exists $\delta > 0$, depending only on β and Θ, such that whenever $\|z_0\|_2 < \delta$, there is $x_\infty \in \mathrm{Ker}(H)$ with*

$$\begin{aligned} x_k + \eta k e &\longrightarrow x_\infty, \\ V_T &\longrightarrow \|b\|_2. \end{aligned}$$

The radius is existential; the conclusion does not require distinct positive eigenvalues.

The shifted threshold controls a neighborhood on the actual zero-buffer initialization slice. The proof accounts for the tangential deviation β^{k+1} before applying the limiting contraction. It does not provide a global basin below the shifted threshold or a basin radius uniform as β approaches one.

Proof of Theorem E.3. We prove contraction for the nonautonomous tail and pull the resulting neighborhood back through the finite zero-buffer transient. On $\mathrm{Range}(H) \times \mathrm{Range}(H)$, the linearization of Equation (39) at the limiting straight state is

$$A := D_y \Phi_\beta(0;1) = \begin{bmatrix} I - (1-\beta)\Theta & -\beta\Theta \\ (1-\beta)I & \beta I \end{bmatrix}.$$

Let A_i denote its two-dimensional block associated with the eigenvalue ρ_i. This block has trace $1 + \beta - (1-\beta)\rho_i$ and determinant β. The three strict Jury quantities are

$$\begin{aligned} 1 - \det A_i &= 1 - \beta > 0, \\ 1 - \mathrm{tr}(A_i) + \det A_i &= (1-\beta)\rho_i > 0, \\ 1 + \mathrm{tr}(A_i) + \det A_i &= 2(1+\beta) - (1-\beta)\rho_i > 0. \end{aligned}$$

They are positive for every i by the assumed strict upper bound on ρ_1. Hence all eigenvalues of A lie inside the unit disk. There exist a norm $\|\cdot\|_{\mathrm{c}}$ on the reduced state space and $\alpha \in (0,1)$ such that $\|Ay\|_{\mathrm{c}} \le \alpha \|y\|_{\mathrm{c}}$. Fix $C_* \ge 1$ satisfying $C_*^{-1}\|y\|_2 \le \|y\|_{\mathrm{c}} \le C_* \|y\|_2$.

To control the actual normalization remainder, write $y_{k+1} = Ay_k + \xi_k$ with $\xi_k := \big(\Theta p_{k+1}(1 - F_k^{-1}), 0\big)$. For $1 - \beta^{k+1} \ge 1/2$, rationalization and the buffer update give

$$\begin{aligned} \|\xi_k\|_{\mathrm{c}} &\le C_* \rho_1 \|p_{k+1}\|_2 |1 - F_k^{-1}| \\ &= C_* \rho_1 \|p_{k+1}\|_2 \frac{|F_k^2 - 1|}{F_k(F_k+1)} \\ &\le C_* \rho_1 \|p_{k+1}\|_2 \big(4\beta^{k+1} + 2\|p_{k+1}\|_2^2\big) \\ &\le C_*^2 \rho_1 \big(4\beta^{k+1} + 2C_*^2 \|y_k\|_{\mathrm{c}}^2\big)\|y_k\|_{\mathrm{c}} \\ &\le K\big(\beta^{k+1} + \|y_k\|_{\mathrm{c}}^2\big)\|y_k\|_{\mathrm{c}}, \end{aligned}$$

where $K := 4C_*^4\rho_1$ is finite and positive. The second inequality uses $1-(1-\beta^{k+1})^2 = 2\beta^{k+1} - \beta^{2k+2} \le 2\beta^{k+1}$ and $F_k \ge 1-\beta^{k+1} \ge 1/2$. By Cauchy–Schwarz, the buffer update satisfies $\|p_{k+1}\|_2 \le \sqrt{(1-\beta)^2+\beta^2}\|y_k\|_2 \le \|y_k\|_2$. Together with the norm comparison, this gives the penultimate inequality. Choose $q \in (\alpha, 1)$, then choose an integer N such that $1-\beta^{N+1} \ge 1/2$ and $K\beta^{N+1} \le (q-\alpha)/2$. Choose $\tau > 0$ so that $K\tau^2 \le (q-\alpha)/2$. For $k \ge N$ and $\|y_k\|_{\mathrm{c}} \le \tau$, this gives

$$\begin{aligned}\|y_{k+1}\|_{\mathrm{c}} &\le \|Ay_k\|_{\mathrm{c}} + \|\xi_k\|_{\mathrm{c}} \\ &\le \big[\alpha + K(\beta^{k+1} + \|y_k\|_{\mathrm{c}}^2)\big]\|y_k\|_{\mathrm{c}} \\ &\le q\|y_k\|_{\mathrm{c}}.\end{aligned}$$

Thus the radius-τ ball is invariant on this tail and contracts geometrically.

The finite-prefix map $z_0 \mapsto y_N$, with $y_0 = (z_0, 0)$, is continuous near zero and maps zero to zero. This follows because every denominator has the fixed lower bound $F_k \ge 1-\beta > 0$. Consequently, there exists $\delta > 0$ such that $\|z_0\|_2 < \delta$ implies $\|y_N\|_{\mathrm{c}} \le \tau$. Iteration of the tail contraction gives

$$\|y_k\|_{\mathrm{c}} \le q^{k-N}\|y_N\|_{\mathrm{c}}, \qquad k \ge N.$$

The norm comparison therefore gives $z_k, p_k \to 0$ and summability of both $\|p_k\|_2$ and $\|p_k\|_2^2$. All choices depend only on β and Θ.

To recover the ambient trajectory, the parameter update and rationalization give

$$\begin{aligned}\|x_{k+1} - x_k + \eta e\|_2 &= \eta\Big\|-\frac{p_{k+1}}{F_k} + \left(1 - \frac{1-\beta^{k+1}}{F_k}\right)e\Big\|_2 \\ &\le \frac{\eta\|p_{k+1}\|_2}{F_k} + \eta\left(1 - \frac{1-\beta^{k+1}}{F_k}\right) \\ &= \frac{\eta\|p_{k+1}\|_2}{F_k} + \frac{\eta\|p_{k+1}\|_2^2}{F_k(F_k + 1 - \beta^{k+1})} \\ &\le \frac{\eta}{1-\beta}\|p_{k+1}\|_2 + \frac{\eta}{2(1-\beta)^2}\|p_{k+1}\|_2^2.\end{aligned}$$

Here $F_k \ge 1-\beta^{k+1} \ge 1-\beta$ ensures both that the tangential term is nonnegative and that the denominators have the stated lower bounds. Summability of $\|p_{k+1}\|_2$ and $\|p_{k+1}\|_2^2$ therefore gives $\sum_{k=0}^{\infty}\|x_{k+1} - x_k + \eta e\|_2 < \infty$. Thus $x_k + \eta k e$ converges to some x_∞. Its component in $\mathrm{Range}(H)$ equals $\|b\|_2 H^\dagger z_k \to 0$, so $x_\infty \in \mathrm{Ker}(H)$. Finally, $(1-\beta^{k+1})/F_k \to 1$; the reduced downstream recurrence and the vanishing endpoint term in Equation (36) give $V_T \to \|b\|_2$. □

Existence and local attraction of the period-two orbit. For this local result, allow an initial buffer $m_0 = \|b\|_2(p_0 + h_0 e)$ with $p_0 \in \mathrm{Range}(H)$ and $h_0 \ge 0$. Projecting the EMA update onto e gives $m_k = \|b\|_2\{p_k + [1 + (h_0 - 1)\beta^k]e\}$. On $\mathrm{Range}(H)^2 \times [0, \infty)$, write $Y = (y, h)$ and define the autonomous reduced map

$$\mathcal{F}(Y) := \big(\Phi_\beta(y; \beta h + 1 - \beta),\ \beta h + 1 - \beta\big). \tag{42}$$

The full state $Y_k = (y_k, 1 + (h_0 - 1)\beta^k)$ satisfies $Y_{k+1} = \mathcal{F}(Y_k)$, and $h = 1$ is an invariant slice on which the y update is $y \mapsto \Phi_\beta(y; 1)$. The following theorem supplies the local attraction and momentum speed used in Theorem 4.2 above the shifted threshold. Its invariant-slice calculation also supports the general-loss result in Theorem F.4.

Theorem E.4 (Existence and local attraction of the momentum period-two orbit)**.** *Assume $0 < \beta < 1$, a simple top eigenvalue ρ_1 of Θ, and $\rho_1 > 2(1+\beta)/(1-\beta)$. Choose a unit top eigenvector v_1 and set $a := \sqrt{\rho_1^2/4 - (1+\beta)^2/(1-\beta)^2}$. For $\sigma \in \{-1, 1\}$, define $\bar{Y}_\sigma := (\bar{y}_\sigma, 1)$, where*

$$\bar{y}_\sigma := \left(\sigma a v_1, -\sigma\frac{1-\beta}{1+\beta}a v_1\right).$$

Then the following hold.

(i) Existence. *The states form a period-two orbit of Equation* (42)*: $\mathcal{F}(\bar{Y}_\sigma) = \bar{Y}_{-\sigma}$. The normalized downstream increment at either phase is $2(1+\beta)/[\rho_1(1-\beta)]$.*

(ii) Local attraction. *There exist open neighborhoods $\mathcal{U}_\sigma$ of $\bar{Y}_\sigma$, depending only on β, Θ and the chosen phase labels, such that every $Y_0 \in \mathcal{U}_\sigma$ gives the parity limits as $j \to \infty$ and the speed limit as $T \to \infty$:*

$$\begin{aligned} Y_{2j} &\longrightarrow \bar{Y}_\sigma, \\ Y_{2j+1} &\longrightarrow \bar{Y}_{-\sigma}, \\ V_T &\longrightarrow \frac{2\|b\|_2}{\rho_1}\frac{1+\beta}{1-\beta}. \end{aligned}$$

The neighborhoods are existential local-stability quantities.

For the standard zero-buffer trajectory, the theorem applies after any entry $(y_N, 1-\beta^N) \in \mathcal{U}_\sigma$, with that time taken as the new origin and the buffer retained. After shifting time by N, the tangential coefficient at step k is $1-\beta^{N+k}$, and the position update uses $1-\beta^{N+k+1}$. Since $\mathcal{U}_\sigma$ is open around $(\bar{y}_\sigma, 1)$, entry of y_N into a sufficiently small neighborhood of $\bar{y}_\sigma$ at a sufficiently late time implies this full-state condition. The theorem does not establish entry from the standard initialization slice $(z_0, 0, 0)$. The period is in the reduced state; the position continues to move downstream. At the boundary $\rho_1 = 2(1+\beta)/(1-\beta)$, the amplitude vanishes and the theorem does not apply; Theorem E.2 still gives the unconditional speed guarantee.

Proof of Theorem E.4. For either $\sigma \in \{-1, 1\}$, the buffer update at $\bar{y}_\sigma$ is

$$(1-\beta)\sigma a v_1 - \beta\sigma\frac{1-\beta}{1+\beta}a v_1 = \sigma\frac{1-\beta}{1+\beta}a v_1.$$

The squared normalization denominator is

$$1 + \frac{(1-\beta)^2}{(1+\beta)^2}a^2 = 1 + \frac{(1-\beta)^2}{(1+\beta)^2}\left[\frac{\rho_1^2}{4} - \frac{(1+\beta)^2}{(1-\beta)^2}\right] = \frac{\rho_1^2(1-\beta)^2}{4(1+\beta)^2}.$$

Substitution into the position update gives

$$\sigma a v_1 - \Theta\frac{\sigma(1-\beta)a v_1/(1+\beta)}{\rho_1(1-\beta)/[2(1+\beta)]} = \sigma a v_1 - 2\sigma a v_1 = -\sigma a v_1.$$

The square root of the computed normalization is positive, and the assumed strict threshold makes $a > 0$, so the two phases are distinct. Together with the buffer identity, this gives $\Phi_\beta(\bar{y}_\sigma; 1) = \bar{y}_{-\sigma}$ and hence $\Phi_\beta(\Phi_\beta(\bar{y}_\sigma; 1); 1) = \bar{y}_\sigma$.

The invariant slice $h = 1$ therefore gives $\mathcal{F}(\bar{Y}_\sigma) = \bar{Y}_{-\sigma}$. The reciprocal of the computed positive normalization denominator is the asserted downstream increment.

The derivative of $u \mapsto u/\sqrt{1+\|u\|_2^2}$ equals $(1+\|u\|_2^2)^{-1/2}I - (1+\|u\|_2^2)^{-3/2}uu^\top$. At either phase, its radial coefficient is $c_1 := [2(1+\beta)/(\rho_1(1-\beta))]^3$ and its coefficients in directions orthogonal to v_1 are $c_i := 2(1+\beta)/[\rho_1(1-\beta)]$ for $i \geq 2$. Thus the derivative $D_y\Phi_\beta(\bar{y}_\sigma; 1)$ in each active coordinate pair (z_i, p_i) is

$$J_i := \begin{bmatrix} 1-(1-\beta)\rho_i c_i & -\beta\rho_i c_i \\ 1-\beta & \beta \end{bmatrix}. \tag{43}$$

Write $J := \bigoplus_{i=1}^r J_i$. The determinant is β and the trace is $1+\beta-(1-\beta)\rho_i c_i$, giving the following stability calculation:

$$\begin{aligned} 1 - \det J_i &= 1-\beta, \\ 1 - \operatorname{tr}(J_i) + \det J_i &= (1-\beta)\rho_i c_i, \\ 1 + \operatorname{tr}(J_i) + \det J_i &= 2(1+\beta) - (1-\beta)\rho_i c_i, \\ 0 < \rho_1 c_1 &= \frac{8(1+\beta)^3}{\rho_1^2(1-\beta)^3} < 2\frac{1+\beta}{1-\beta}, \\ 0 < \rho_i c_i &= 2\frac{1+\beta}{1-\beta}\frac{\rho_i}{\rho_1} < 2\frac{1+\beta}{1-\beta}, \qquad i \geq 2. \end{aligned}$$

The radial inequality uses the strict supercritical threshold, and the inequalities for $i \geq 2$ use simplicity of the top eigenvalue. Thus the three Jury quantities are strictly positive, so each J_i has all eigenvalues inside the unit disk. The derivative is identical at the two phases, so the chain rule gives the two-step derivative J^2 on the invariant slice $h = 1$.

To include the scalar buffer, differentiate Equation (42). Let $d_\sigma := \partial_h \Phi_\beta(\bar{y}_\sigma; h)|_{h=1}$, a vector in $\mathrm{Range}(H)^2$. Differentiating the denominator and substituting the computed buffer and normalization gives

$$\begin{aligned} d_\sigma &= \left(\frac{\sigma \rho_1 (1-\beta) a/(1+\beta)}{\{\rho_1(1-\beta)/[2(1+\beta)]\}^3} v_1,\, 0 \right) \\ &= \left(\frac{8\sigma(1+\beta)^2 a}{\rho_1^2 (1-\beta)^2} v_1,\, 0 \right), \\ D\mathcal{F}(\bar{Y}_\sigma) &= \begin{bmatrix} J & \beta d_\sigma \\ 0 & \beta \end{bmatrix}. \end{aligned}$$

Since $d_{-\sigma} = -d_\sigma$, the chain rule gives the return derivative

$$\begin{aligned} A_\sigma := D(\mathcal{F}^2)(\bar{Y}_\sigma) &= \begin{bmatrix} J & -\beta d_\sigma \\ 0 & \beta \end{bmatrix} \begin{bmatrix} J & \beta d_\sigma \\ 0 & \beta \end{bmatrix} \\ &= \begin{bmatrix} J^2 & \beta(J - \beta I) d_\sigma \\ 0 & \beta^2 \end{bmatrix}. \end{aligned}$$

Its eigenvalues are those of J^2 together with β^2, so A_σ is Schur stable.

Define $R_\sigma := \sum_{j=0}^{\infty} (A_\sigma^\top)^j A_\sigma^j$ on the full reduced space. The series converges, $R_\sigma \succeq I$, and $A_\sigma^\top R_\sigma A_\sigma = R_\sigma - I$, giving

$$\begin{aligned} \|A_\sigma u\|_{R_\sigma}^2 &= \|u\|_{R_\sigma}^2 - \|u\|_2^2 \\ &\leq \left[1 - \lambda_{\max}(R_\sigma)^{-1}\right] \|u\|_{R_\sigma}^2. \end{aligned}$$

For $|h - 1| < 1/2$, both scalar-buffer updates in $\mathcal{F}^2$ are greater than $1/2$, so both normalization denominators are positive and $\mathcal{F}^2$ is continuously differentiable. Derivative continuity and the mean-value inequality therefore give common constants $\delta \in (0, 1/2)$ and $q \in (0, 1)$ such that

$$\|\mathcal{F}^2(Y) - \bar{Y}_\sigma\|_{R_\sigma} \leq q \|Y - \bar{Y}_\sigma\|_{R_\sigma}$$

whenever $\|Y - \bar{Y}_\sigma\|_{R_\sigma} \leq \delta$. Decrease δ so that the two phase balls are disjoint, and set $\mathcal{U}_\sigma := \{Y : \|Y - \bar{Y}_\sigma\|_{R_\sigma} < \delta\}$. These choices depend only on β, Θ and the chosen phases. The contraction maps each ball into itself, so induction gives

$$\|Y_{2j} - \bar{Y}_\sigma\|_{R_\sigma} \leq q^j \|Y_0 - \bar{Y}_\sigma\|_{R_\sigma} \longrightarrow 0.$$

Continuity of $\mathcal{F}$ gives $Y_{2j+1} \to \bar{Y}_{-\sigma}$.

The normalized downstream increment at $Y = (z, p, h)$ is the continuous function

$$\frac{\beta h + 1 - \beta}{\sqrt{(\beta h + 1 - \beta)^2 + \|\beta p + (1-\beta) z\|_2^2}}.$$

It has the same limit $2(1 + \beta)/[\rho_1(1 - \beta)]$ along both parities. Its Cesàro average gives $(\ell_T - \ell_0)/(\eta T) \to 2(1 + \beta)/[\rho_1(1 - \beta)]$. The curved state is bounded, so the endpoint term in Equation (36) vanishes and yields the asserted limit for V_T. A finite prefix before an entry time has zero contribution to this asymptotic speed. □

F Proof for general loss

F.1 Preliminary analysis

The results below concern NGD and NGM iterates along a prescribed noncritical gradient-flow path segment. We establish the Fermi reduction and geometric error bounds, construct the local period-two orbits, and then prove actual tracking and clock estimates. The dynamics subsections use these

tools for their entry, progress, and iteration bounds. We first specify the loss geometry used in Section 4.1 and the notation needed for the proofs.

Fix a loss $L \in C^3(\mathcal{U};\mathbb{R})$ on an open set $\mathcal{U} \subset \mathbb{R}^d$, with $d \geq 2$, and a learning rate $\eta > 0$. Let $z(t)$ be a noncritical gradient-flow curve, and consider its segment $0 \leq t \leq T_*$, where $T_* > 0$. Choose a compact interval I with $[0, T_*] \subset \operatorname{int} I$, and extend z to an open neighborhood of I, with $B(t) := \|\nabla L(z(t))\| > 0$. Write its positive continuous speed and unit tangent as

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = v(t)e(t), \qquad e(t) := -\frac{\nabla L(z(t))}{B(t)}, \qquad v(t) > 0. \tag{44}$$

Let $N(t) \in \mathbb{R}^{d\times(d-1)}$ be an orthonormal normal frame whose column derivatives are tangential, called a Bishop frame. For a normal coordinate $u \in \mathbb{R}^{d-1}$, write a nearby point as

$$x = \Psi(t, u) := z(t) + \eta N(t)u. \tag{45}$$

Here t locates the point along the curve, while $\|u\|$ measures its normal distance in units of one step. The dimensionless normal sharpness is

$$\Theta(t) := \frac{\eta}{B(t)} N(t)^\top \nabla^2 L(z(t)) N(t). \tag{46}$$

Write $\theta_1(t) \geq \cdots \geq \theta_{d-1}(t)$ for its eigenvalues. For a physical radius $\rho > 0$, define the tube

$$\mathcal{X} := \big\{\Psi(t, u) : t \in \operatorname{int} I,\ \eta\|u\| < \rho\big\}.$$

Assumption F.1 (Loss geometry)**.** Let $L \in C^3(\mathcal{U};\mathbb{R})$ on an open $\mathcal{U} \subset \mathbb{R}^d$, $d \geq 2$, and fix $\rho, B_- > 0$ and $\Gamma_3 \geq 0$. Assume z is a noncritical gradient-flow curve as in Equation (44) on a neighborhood of a compact interval I with $[0, T_*] \subset \operatorname{int} I$, and that (i) $\Theta(t)$ has eigenvalues $\theta_1(t) \geq \cdots \geq \theta_{d-1}(t) > 0$ on I; (ii) Ψ is injective for $t \in \operatorname{int} I$, $\eta\|u\| < \rho$, with image $\mathcal{X}$ whose closure is compact in $\mathcal{U}$; (iii) $B(t) \geq B_-$ on I and $\|\nabla^3 L(y)\|_{\mathrm{op}} \leq \Gamma_3$ on $\overline{\mathcal{X}}$.

Under Assumption F.1(i), use the curvature-adaptive flow in Equation (9), so that

$$v(t) = \frac{2B(t)}{\lambda_\perp(z(t))} = \frac{2\eta}{\theta_1(t)}.$$

Thus t measures iteration time. For two times $a, b \in I$, write $\ell(a, b) := \int_a^b v(q)\,\mathrm{d}q$ for their signed path length, and put $S := \ell(0, T_*)$. For a nonempty compact time interval $J \subset I$ and a physical length $r \geq 0$, write

$$J^{+r} := \big\{t \in I : \inf_{q\in J} |\ell(q, t)| \leq r\big\}.$$

The condition $J^{+r} \subset \operatorname{int} I$ supplies a path-length margin r at both ends of J.

For $\beta \in [0, 1)$, both algorithms use the original loss through

$$\begin{aligned} m_{k+1} &= \beta m_k + (1-\beta)\nabla L(x_k), \\ x_{k+1} &= x_k - \eta \frac{m_{k+1}}{\|m_{k+1}\|}. \end{aligned} \tag{47}$$

At $\beta = 0$, the update is plain `NormGD` and does not depend on the previous buffer. For $\beta > 0$, resolve the current buffer in the orthonormal frame $[-e(t), N(t)]$:

$$\frac{m}{B(t)} = -he(t) + N(t)p. \tag{48}$$

The coefficients are uniquely determined by $h = -e(t)^\top m/B(t)$ and $p = N(t)^\top m/B(t)$. Thus $h \in \mathbb{R}$ measures the buffer along the gradient direction and $p \in \mathbb{R}^{d-1}$ is its normal component. Use $Z = u$ when $\beta = 0$ and $Z = (u, h, p)$ when $\beta > 0$, with the same convention for Z_k.

Under Assumption F.1(i), write $\theta_+ := \sup_I \theta_1$, $\theta_- := \inf_I \theta_{d-1} > 0$, $\theta_{1,-} := \inf_I \theta_1$, $\kappa_\perp := \theta_+/\theta_-$, and $\alpha := 1 - 4/\theta_{1,-}^2$. Denote the unique coordinates of $x \in \mathcal{X}$ by $(\pi(x), u(x))$, and set $t_k := \pi(x_k)$ and $u_k := u(x_k)$ for an iterate. Thus $\pi(x)$ is the time coordinate of the normal

projection onto z, and is not the iteration index. Smoothness, compactness, and Assumption F.1(iii) give finite bounds $B_+ \geq B_-$ and $M_2 \geq 0$ such that, for every $y \in \overline{\mathcal{X}}$,

$$\begin{aligned} \|\nabla L(y)\| &\leq B_+, \\ \|\nabla^2 L(y)\|_{\mathrm{op}} &\leq M_2, \\ \|\nabla^3 L(y)\|_{\mathrm{op}} &\leq \Gamma_3. \end{aligned} \tag{49}$$

Here $\nabla^3 L(y)[v, w]$ is the vector satisfying $r^\top \nabla^3 L(y)[v, w] = D^3 L(y)[r, v, w]$ for every $r \in \mathbb{R}^d$. Write $\kappa(t) := v(t)^{-1} N(t)^\top \mathrm{d}e(t)/\mathrm{d}t$ and $K_\kappa := \sup_I \|\kappa(t)\|$ for path curvature. The variation constants are measured per unit path length:

$$\begin{aligned} L_B &:= \sup_{t \in I} \frac{1}{v(t)} \left| \frac{\mathrm{d} \log B(t)}{\mathrm{d}t} \right|, & L_\kappa &:= \sup_{t \in I} \frac{1}{v(t)} \left\| \frac{\mathrm{d}\kappa(t)}{\mathrm{d}t} \right\|, \\ L_\Theta &:= \sup_{t \in I} \frac{1}{v(t)} \left\| \frac{\mathrm{d}\Theta(t)}{\mathrm{d}t} \right\|_{\mathrm{op}}, & L_{\theta_1} &:= \sup_{a \neq b \in I} \frac{|\theta_1(b) - \theta_1(a)|}{|\ell(a, b)|}. \end{aligned}$$

Remark F.2 (Normal positivity). Assumption F.1(i) requires positive curvature normal to the reference flow, while allowing weak or negative curvature in the flow direction. For example, fix $t \in I$ and write $H := \nabla^2 L(z(t))$ and $n := -e(t)$. Suppose that H has a unit eigenvector v with eigenvalue in $[-\nu, \nu]$, while its other eigenvalues are at least $\mu > \nu \geq 0$. If $\|(I_d - vv^\top)n\| \leq 1/2$, then every unit $w \perp n$ satisfies $|w^\top v|^2 \leq 1/4$ and $w^\top H w \geq \mu - (\mu + \nu)/4 \geq \mu/2$. Thus alignment of the gradient with the weak eigendirection gives $\Theta(t) \succeq \eta\mu I_{d-1}/(2B_+)$. The tube condition and the simple-top and sharpness-threshold conditions used below remain separate.

One-step updates in local coordinates. The NGD and NGM recursions in Equation (47) evaluate the original loss. We express their updates in Fermi coordinates, using the buffer components in Equation (48) and suppressing the superscript β on state coordinates when the method is clear.

For the local model, fix the reference time t and keep its local coefficients constant during the state update. For plain descent, the local arclength increment is $\eta a_0(t, u)$ and the next normal coordinate is $F_{0,t}(u)$, where

$$\begin{aligned} a_0(t, u) &:= \frac{1}{\sqrt{1 + \|\Theta(t)u\|^2}}, \\ F_{0,t}(u) &:= u - a_0(t, u)\Theta(t)u. \end{aligned}$$

For momentum with $\beta > 0$, first form the local buffer update, whose tangential component is $\widehat{h} := \beta h + 1 - \beta$ and whose normal component and norm are

$$\begin{aligned} \widehat{p} &:= \beta p + (1 - \beta)\Theta(t)u, \\ \widehat{r} &:= \sqrt{\widehat{h}^2 + \|\widehat{p}\|^2}. \end{aligned} \tag{50}$$

On $\widehat{r} > 0$, define the local arclength speed $a_\beta(t, Z) := \widehat{h}/\widehat{r}$ and the update of $Z = (u, h, p)$ by

$$F_{\beta,t}(Z) := \left(u - \frac{\widehat{p}}{\widehat{r}}, \widehat{h}, \widehat{p} \right).$$

Lemma F.3 relates these local maps to one step of the algorithms, with errors from nonlinear gradient terms and changes in the Fermi coordinates. To quantify these errors and the variation of the local coefficients, define the dimensionless geometric variation

$$\delta := \max\left\{ \eta K_\kappa,\ \eta L_B,\ \frac{\eta^2 \Gamma_3}{B_-} \right\}. \tag{51}$$

The three terms measure path bending, relative changes in gradient norm, and nonquadratic error over one step.

Lemma F.3 (NGD and NGM in local coordinates)**.** *Under Assumption F.1, fix $R \geq 0$ and consider states $x = \Psi(t, u)$ with $t \in \operatorname{int} I$ and $\|u\| \leq R$. For momentum, also fix $M \geq 0$ and $r_* > 0$, and require $|h| + \|p\| \leq M$ and $\widehat{r} \geq r_*$. Under the smallness and radius conditions below, if the closed ambient ball of radius η centered at x is contained in $\mathcal{X}$, one update of the original-loss algorithm is well defined, remains in $\mathcal{X}$, and satisfies*

$$\ell(t, t^+) = \eta\big[a_\beta(t, Z) + \zeta_\beta\big],$$
$$Z^+ = F_{\beta,t}(Z) + \xi_\beta. \tag{52}$$

The physical increment obeys $|\ell(t, t^+)| \leq 2\eta$. There exist $c_{\mathrm{geo}}, C_{\mathrm{geo}} > 0$, depending only on R and θ_+, with β, M, r_*^{-1}, and B_+/B_- added for momentum, such that whenever $\delta \leq 1$,

$$\max\{\eta L_\Theta, \eta^2 L_\kappa\} \leq C_{\mathrm{geo}}\delta.$$

If $\delta \leq c_{\mathrm{geo}}$ and $\eta(R+1) < \rho$, the update errors satisfy

$$|\zeta_\beta| + \|\xi_\beta\|_2 \leq C_{\mathrm{geo}}\delta.$$

Let $J = [a, b]$ satisfy $J^{+2\eta} \subset \operatorname{int} I$. Under these same smallness and radius conditions, the ball containment holds for every $t \in J$ and $\|u\| \leq R$. These constants are independent of the learning rate, path length, and chosen state when the listed dimensionless data are fixed.

Proof of Lemma F.3. We first bound curvature using the tube width and the loss derivatives, then derive exact update errors under general chart and buffer margins. We then derive the geometric error bounds and verify the ball containment from the geometric margins. Use the curvature and variation bounds $K_\kappa, L_B, L_\kappa, L_\Theta, L_{\theta_1}$ introduced above.

The normalized gradient field is C^2, so $e(t)$ is continuously differentiable along the C^1 curve z. Starting from any orthonormal normal basis, solve $\mathrm{d}N/\mathrm{d}t = -e(\mathrm{d}e/\mathrm{d}t)^\top N$. This ODE preserves $e^\top N = 0$ and $N^\top N = I_{d-1}$, since

$$\frac{\mathrm{d}}{\mathrm{d}t}(e^\top N) = \left(\frac{\mathrm{d}e}{\mathrm{d}t}\right)^\top N - \left(\frac{\mathrm{d}e}{\mathrm{d}t}\right)^\top N = 0,$$
$$\frac{\mathrm{d}}{\mathrm{d}t}(N^\top N) = 0.$$

With $\kappa = v^{-1}N^\top \mathrm{d}e/\mathrm{d}t$, the resulting C^1 Bishop frame satisfies

$$\frac{\mathrm{d}e}{\mathrm{d}t} = vN\kappa,$$
$$\frac{\mathrm{d}N}{\mathrm{d}t} = -ve\kappa^\top. \tag{53}$$

Changing the initial normal basis only applies a constant orthogonal change of coordinates to u and Θ, leaving distances and eigenvalues unchanged. Differentiating Equation (45) gives

$$D\Psi(t, u)[a, b] = v(t)\big(1 - \eta\kappa(t)^\top u\big)e(t)a + \eta N(t)b. \tag{54}$$

The factor $1 - \eta\kappa^\top u$ accounts for the change in arclength away from the center curve.

The Bishop identity gives $K_\kappa = \sup_I \|\kappa\|$. Uniqueness of normal coordinates throughout the tube implies

$$\rho K_\kappa \leq 1. \tag{55}$$

To prove this, fix $t \in \operatorname{int} I$ with $\kappa(t) \neq 0$ and let $r \neq t$ approach t. Write $b_r := e(r) - \langle e(r), e(t)\rangle e(t)$ and $a_r := \langle z(r) - z(t), e(r)\rangle$. For r sufficiently close to t, the vector $w_r := a_r b_r/\|b_r\|^2$ is well defined and satisfies

$$\langle w_r, e(t)\rangle = 0,$$
$$\langle w_r - [z(r) - z(t)], e(r)\rangle = 0.$$

Thus $z(t) + w_r$ has normal representations at both t and r. Since $b_r/\ell(t, r) \to N(t)\kappa(t)$ and $a_r/\ell(t, r) \to 1$, both normal displacements have norm tending to $1/\|\kappa(t)\|$. If this limit were smaller

than ρ, both representations would lie in the operational tube, contradicting Assumption F.1(ii). The bound is automatic when $\kappa(t) = 0$, and continuity includes the endpoints of I.

We now derive the geometric bounds from Assumption F.1(iii) and Equation (49). Differentiating Equation (44) gives

$$\begin{aligned}\frac{1}{v(t)}\frac{\mathrm{d}B(t)}{\mathrm{d}t} &= -e(t)^\top \nabla^2 L(z(t))e(t),\\ \frac{1}{v(t)}\frac{\mathrm{d}e(t)}{\mathrm{d}t} &= -\frac{[I_d - e(t)e(t)^\top]\nabla^2 L(z(t))e(t)}{B(t)},\\ \kappa(t) &= -\frac{N(t)^\top \nabla^2 L(z(t))e(t)}{B(t)},\\ e(t)^\top \nabla^2 L(z(t))N(t) &= -B(t)\kappa(t)^\top. \end{aligned} \tag{56}$$

These identities also compare the normal sharpness with the largest eigenvalue of the full Hessian under Assumption F.1(i). For this comparison, fix $t \in I$ and write $H := \nabla^2 L(z(t))$, $a := -\eta v(t)^{-1}\mathrm{d}\log B(t)/\mathrm{d}t$, and $\widehat{\theta}_1 := \eta\lambda_{\max}(H)/B(t)$. In the orthonormal frame $[e, N]$, suppressing the argument t, the normalized Hessian is

$$\frac{\eta}{B}[e, N]^\top H[e, N] = \begin{pmatrix} a & -\eta\kappa^\top \\ -\eta\kappa & \Theta \end{pmatrix}.$$

If $\delta < \theta_{1,-}$, then $|a| \le \delta$, $\eta\|\kappa\| \le \delta$, and $\theta_1 - a \ge \theta_{1,-} - \delta > 0$. The Rayleigh principle gives the lower bound below, while $\Theta \preceq \theta_1 I_{d-1}$ bounds the full Rayleigh quotient by that of the corresponding two-dimensional matrix:

$$\begin{aligned} 0 \le \widehat{\theta}_1 - \theta_1 &\le \frac{\sqrt{(\theta_1 - a)^2 + 4\eta^2\|\kappa\|^2} - (\theta_1 - a)}{2}\\ &\le \frac{\eta^2\|\kappa\|^2}{\theta_1 - a} \le \frac{\delta^2}{\theta_{1,-} - \delta}. \end{aligned} \tag{57}$$

When $\kappa = 0$ and $a \le \theta_1$, the two sharpnesses coincide exactly.

The curvature identity shows that κ is C^1 without requiring differentiability of v. Differentiating it and estimating its terms gives, with all quantities evaluated at t,

$$\begin{aligned}\frac{1}{v}\frac{\mathrm{d}\kappa}{\mathrm{d}t} &= -\frac{N^\top \nabla^3 L(z)[e, e]}{B} + \frac{2e^\top \nabla^2 L(z)e}{B}\kappa\\ &\quad - \frac{N^\top \nabla^2 L(z)N}{B}\kappa,\\ L_\kappa &\le \frac{\Gamma_3}{B_-} + \left(2L_B + \frac{M_2}{B_-}\right)K_\kappa \le \frac{\Gamma_3}{B_-} + \frac{3M_2^2}{B_-^2}. \end{aligned} \tag{58}$$

Here $K_\kappa, L_B \le M_2/B_-$ by Equation (56) and the Hessian bound.

For the normal sharpness, write $Q_\Theta := \sup_I \|\Theta\|_{\mathrm{op}}$. Differentiating Equation (46) and using $v^{-1}\mathrm{d}N/\mathrm{d}t = -e\kappa^\top$ and Equation (56) gives

$$L_\Theta \le \frac{\eta\Gamma_3}{B_-} + 2\eta K_\kappa^2 + Q_\Theta L_B \le \eta\left(\frac{\Gamma_3}{B_-} + \frac{3M_2^2}{B_-^2}\right). \tag{59}$$

Weyl's inequality gives $L_{\theta_1} \le L_\Theta$, and the definition of Θ gives $Q_\Theta \le \eta M_2/B_-$. Under Assumption F.1(i), $Q_\Theta = \theta_+$; positivity is not needed for these general geometric bounds.

For the general quantitative version, retain the state and complete-step ball conditions, fix $q_0 \in (0, 1]$, and require $\eta(R + 1) < \rho$. For momentum, assume $|h| + \|p\| \le M$ and $\widehat{r} \ge r_*$ for fixed $M \ge 0$ and $r_* > 0$. Define

$$\delta_{\mathrm{geo},\eta} := \sup_{t\in I}\left[\eta K_\kappa R + \frac{\Gamma_3\eta^2 R^2}{2B(t)}\right]. \tag{60}$$

The quantitative Fermi conditions comprise these state and domain requirements and

$$\begin{aligned} \eta K_\kappa(R+1) &\le 1-q_0, \\ \delta_{\mathrm{geo},\eta} &\le 1/2, \\ (1-\beta)\delta_{\mathrm{geo},\eta} &\le r_*/2 \qquad \text{for momentum only.} \end{aligned} \tag{61}$$

These conditions retain the dependence on the actual curvature and any fixed chart margin q_0. For comparison, Equation (55) gives the sufficient conditions

$$\begin{aligned} \eta(R+1) &\le \rho/2, \\ \frac{\eta R}{\rho} + \frac{\Gamma_3 \eta^2 R^2}{2B_-} &\le 1/2. \end{aligned}$$

For momentum, additionally require $(1-\beta)[\eta R/\rho + \Gamma_3\eta^2R^2/(2B_-)] \le r_*/2$. These imply Equation (61) with $q_0 = 1/2$. We prove the normal form and exact defect bounds under the general conditions, without requiring this sufficient specialization.

Throughout the open tube, Equation (55) gives $1-\eta\kappa(t)^\top u \ge 1-\eta\|u\|/\rho > 0$. Thus Equation (54) is invertible everywhere on the coordinate domain. The inverse function theorem and injectivity of Ψ make $\mathcal{X}$ open and its coordinate inverse continuously differentiable. The unique representation in Assumption F.1(ii) gives $\langle x - z(\pi(x)), e(\pi(x))\rangle = 0$ and

$$u(x) = \eta^{-1} N(\pi(x))^\top [x - z(\pi(x))].$$

Differentiating the orthogonality identity and the normal-coordinate formula in an ambient direction w gives

$$\begin{aligned} 0 &= \langle e(t), w\rangle - v(t)\big(1-\eta\kappa(t)^\top u\big) D\pi(x)[w], \\ Du(x)[w] &= \eta^{-1} N(t)^\top w, \end{aligned}$$

where the second identity uses $N(t)^\top e(t) = 0$ and $[\mathrm{d}N(t)/\mathrm{d}t]^\top N(t)u = 0$.

Let ω be any unit vector and consider $x_r = x + r\eta\omega$, $0 \le r \le 1$. The complete-step ball condition places the entire segment in $\mathcal{X}$. Write $x_r = \Psi(t_r, u_r)$ and assume $\|u\| \le R$. Applying the inverse derivatives to $\mathrm{d}x_r/\mathrm{d}r = \eta\omega$ gives

$$\begin{aligned} \frac{\mathrm{d}t_r}{\mathrm{d}r} &= \eta \frac{\langle e(t_r), \omega\rangle}{v(t_r)[1-\eta\kappa(t_r)^\top u_r]}, \\ \frac{\mathrm{d}u_r}{\mathrm{d}r} &= N(t_r)^\top \omega. \end{aligned}$$

Since $\|N(t_r)^\top\omega\| \le 1$, integration and the quantitative Fermi conditions give

$$\begin{aligned} \|u_r\| &\le \|u\| + r \le R + r, \\ 1-\eta\kappa(t_r)^\top u_r &\ge 1 - \eta K_\kappa(R+1) \ge q_0, \\ |\ell(t, t_r)| &\le \frac{\eta r}{q_0}. \end{aligned}$$

In particular, every unit update satisfying these geometric conditions obeys

$$|\ell(t, t^+)| \le \eta/q_0, \qquad \|u^+\| \le R+1. \tag{62}$$

Write $\omega = ae(t) - N(t)b + r_{\mathrm{dir}}$, with scalar a, normal vector b, and ambient vector r_{dir}. Set $t^+ = t_1$ and $u^+ = u_1$. Write $n_r := \langle e(t_r), \omega\rangle$ and $q(r) := 1 - \eta\kappa(t_r)^\top u_r$. The path equations give a direct arclength estimate without differentiating κ:

$$\begin{aligned} |n_r - a| &\le \|r_{\mathrm{dir}}\| + \frac{\eta K_\kappa r}{q_0}, \qquad\qquad |1-q(r)| \le \eta K_\kappa(R+r), \\ \left|\frac{\ell(t,t^+)}{\eta} - a\right| &\le \int_0^1 \frac{|n_r - a| + |a|\,|1-q(r)|}{q(r)}\, dr \\ &\le \frac{\|r_{\mathrm{dir}}\|}{q_0} + \eta K_\kappa \left(\frac{R+1/2}{q_0} + \frac{1}{2q_0^2}\right). \end{aligned}$$

Here $|a| \le 1$ for both algorithmic directions below.

For the normal coordinate, define $R_{\mathrm{n}} := \int_0^1 [N(t_r) - N(t)]^\top \omega \, dr$. The arclength bound and $v(q)^{-1}\|\mathrm{d}N(q)/\mathrm{d}q\|_{\mathrm{op}} = \|\kappa(q)\|$ give

$$\begin{aligned}\|R_{\mathrm{n}}\| &\le \int_0^1 \|N(t_r) - N(t)\|_{\mathrm{op}} \, dr \\ &\le K_\kappa \int_0^1 |\ell(t, t_r)| \, dr \le \frac{K_\kappa \eta}{2q_0}.\end{aligned}$$

Integrating the two path equations therefore gives the exact endpoint identities

$$\begin{aligned}\ell(t, t^+) &= \eta \int_0^1 \frac{n_r}{q(r)} \, dr, \\ u^+ - u &= -b + N(t)^\top r_{\mathrm{dir}} + R_{\mathrm{n}}.\end{aligned}$$

We now compare each algorithm's update direction with the approximation from the local model. Write $g = \nabla L$. The radial segment $\Psi(t, ru)$, $0 \le r \le 1$, lies in $\mathcal{X}$ because $r\eta\|u\| < \rho$, so Equation (49) applies throughout it. For $x = z(t) + \eta N(t)u$, write $R_g(t, u)$ for the integral Taylor remainder and expand

$$\begin{aligned}\frac{g(x)}{B(t)} &= -e(t) + \frac{\eta}{B(t)} \nabla^2 L(z(t)) N(t) u + R_g(t, u) \\ &= -\big(1 + \eta \kappa(t)^\top u\big) e(t) + N(t)\Theta(t)u + R_g(t, u), \qquad (63) \\ R_g(t, u) &:= \frac{1}{B(t)} \int_0^1 (1 - r) \nabla^3 L\big(z(t) + r\eta N(t)u\big)[\eta N(t)u, \eta N(t)u] \, dr, \\ \|R_g(t, u)\| &\le \frac{\Gamma_3 \eta^2 \|u\|^2}{2B(t)}.\end{aligned}$$

The second equality uses $I_d = e(t)e(t)^\top + N(t)N(t)^\top$ and the NGF–Bishop identity in Equation (56). Set $y := \Theta(t)u$, $g_{\mathrm{fr}} := -e(t) + N(t)y$, and $e_g := -\alpha e(t) + R_g(t, u)$, where $\alpha = \eta\kappa(t)^\top u$. Equation (63) becomes $g(x)/B(t) = g_{\mathrm{fr}} + e_g$, and its defect satisfies

$$\|e_g\| \le \eta K_\kappa R + \frac{\Gamma_3 \eta^2 R^2}{2B(t)} \le \delta_{\mathrm{geo},\eta}. \qquad (64)$$

Moreover, orthogonality of the frame gives $\|g_{\mathrm{fr}}\| = \sqrt{1 + \|y\|^2} \ge 1$.

For both methods, we use the following normalization estimate: if $v, w \ne 0$ and $\|v - w\| \le \|w\|/2$, then

$$\begin{aligned}\Big\|\frac{v}{\|v\|} - \frac{w}{\|w\|}\Big\| &= \Big\|\frac{v - w}{\|w\|} + \Big(\frac{1}{\|v\|} - \frac{1}{\|w\|}\Big) v\Big\| \\ &\le \frac{\|v - w\|}{\|w\|} + \frac{\big|\|v\| - \|w\|\big|}{\|w\|} \le \frac{2\|v - w\|}{\|w\|}. \qquad (65)\end{aligned}$$

The last inequality is the reverse triangle inequality.

For plain descent, Equations (61) and (64) give $\|e_g\| \le 1/2$. Applying Equation (65) with $v = g_{\mathrm{fr}} + e_g$ and $w = g_{\mathrm{fr}}$ defines a direction defect e_0 through

$$\begin{aligned}-\frac{g(x)}{\|g(x)\|} &= a_0(t, u) e(t) - N(t) \frac{y}{\sqrt{1 + \|y\|^2}} + e_0, \\ \|e_0\| &\le 2\delta_{\mathrm{geo},\eta}. \qquad (66)\end{aligned}$$

In particular, $\|g(x)\|/B(t) \ge \|g_{\mathrm{fr}}\| - \|e_g\| \ge 1/2$, so the NGD direction is well defined.

For momentum, set $e_m := (1-\beta)e_g$. The ideal vector $-\widehat{h}e(t) + N(t)\widehat{p}$ has norm $\widehat{r} \geq r_*$. The NGM buffer and its normalized direction satisfy

$$\frac{m^+}{B(t)} = -\widehat{h}e(t) + N(t)\widehat{p} + e_m, \tag{67}$$

$$\frac{\|m^+\|}{B(t)} \geq \widehat{r} - \|e_m\| \geq \frac{r_*}{2},$$

$$-\frac{m^+}{\|m^+\|} = \frac{\widehat{h}}{\widehat{r}}e(t) - N(t)\frac{\widehat{p}}{\widehat{r}} + e_\beta,$$

$$\|e_\beta\| \leq \frac{2(1-\beta)}{r_*}\delta_{\mathrm{geo},\eta}. \tag{68}$$

The first identity follows from Equation (48) and the EMA recursion. The noncollapse bound uses Equations (61) and (64); applying Equation (65) then defines e_β and gives its bound.

It remains to convert each exact Euclidean step into Fermi coordinates. For the momentum estimates, set $C_{\mathrm{dir}} := 2(1-\beta)/r_*$; for plain descent alone, use $C_{\mathrm{dir}} = 2$. Write $a_\beta = a_\beta(t, Z)$ and set $b_0 := y/\sqrt{1+\|y\|^2}$ for plain descent and $b_\beta := \widehat{p}/\widehat{r}$ for $\beta > 0$. The direction defects in Equations (66) and (68) satisfy $\|e_\beta\| \leq C_{\mathrm{dir}}\delta_{\mathrm{geo},\eta}$. Apply the unit-step calculation with $\omega = a_\beta e(t) - N(t)b_\beta + e_\beta$, which is the algorithm's unit direction by the preceding noncollapse estimates. Let $R_{u,\beta}$ be its normal remainder. Define the scalar arclength defect and the vector normal defect by

$$\zeta_\beta := \int_0^1 \frac{n_r}{q(r)}\,dr - a_\beta,$$

$$\xi^u_\beta := N(t)^\top e_\beta + R_{u,\beta}.$$

The endpoint identities become $\ell(t, t^+) = \eta(a_\beta + \zeta_\beta)$ and $u^+ = u - b_\beta + \xi^u_\beta$, with the normal update of the corresponding local map in each case. The two coordinate defects therefore satisfy

$$|\zeta_\beta| \leq \frac{C_{\mathrm{dir}}}{q_0}\delta_{\mathrm{geo},\eta} + \eta K_\kappa\left(\frac{R+1/2}{q_0} + \frac{1}{2q_0^2}\right),$$

$$\|\xi^u_\beta\| \leq \|N(t)^\top e_\beta\| + \|R_{u,\beta}\| \leq C_{\mathrm{dir}}\delta_{\mathrm{geo},\eta} + \frac{K_\kappa\eta}{2q_0}.$$

These bounds use the preceding integrated path equations. For the momentum bound, let $\widehat{M}$ bound the predicted normalized buffer norm; $\widehat{M} = \beta M + (1-\beta)(1 + Q_\Theta R)$ is always sufficient, as verified below. Define the plain path defect by

$$\varepsilon_{0,\eta} := 2(1+q_0^{-1})\delta_{\mathrm{geo},\eta} + \eta K_\kappa\left(\frac{R+1}{q_0} + \frac{1}{2q_0^2}\right). \tag{69}$$

Define the joint path and buffer defects by

$$\varepsilon_{\mathrm{path},\eta} := (1+q_0^{-1})C_{\mathrm{dir}}\delta_{\mathrm{geo},\eta} + \eta K_\kappa\left(\frac{R+1}{q_0} + \frac{1}{2q_0^2}\right),$$

$$\varepsilon_{\mathrm{buf},\eta} := e^{\eta L_B/q_0}\left[(1-\beta)\delta_{\mathrm{geo},\eta} + \frac{\eta(K_\kappa + L_B)\widehat{M}}{q_0}\right].$$

For momentum, define the full-state defect budget by

$$\varepsilon_\eta := \varepsilon_{\mathrm{path},\eta} + \varepsilon_{\mathrm{buf},\eta}. \tag{70}$$

Adding the two coordinate estimates gives $|\zeta_\beta| + \|\xi^u_\beta\| \leq \varepsilon_{\mathrm{path},\eta}$. For plain descent, Equation (66) permits $C_{\mathrm{dir}} = 2$, giving $|\zeta_0| + \|\xi^u_0\| \leq \varepsilon_{0,\eta}$.

For momentum, set $\mathsf{Q}(t) := [-e(t), N(t)] \in O(d)$ and $\widehat{b} := (\widehat{h}, \widehat{p}^\top)^\top \in \mathbb{R}^d$. The Bishop equations in Equation (53) give

$$\frac{1}{v(t)}\frac{\mathrm{d}\mathsf{Q}(t)}{\mathrm{d}t} = [-N(t)\kappa(t), -e(t)\kappa(t)^\top],$$

$$\frac{1}{v(t)}\mathsf{Q}(t)^\top\frac{\mathrm{d}\mathsf{Q}(t)}{\mathrm{d}t} = \begin{pmatrix} 0 & \kappa(t)^\top \\ -\kappa(t) & 0 \end{pmatrix}.$$

Since $\mathsf{Q}(t)$ is orthogonal, the last identity implies $v(t)^{-1}\|\mathrm{d}\mathsf{Q}(t)/\mathrm{d}t\|_{\mathrm{op}} = \|\kappa(t)\|$.

Define the new buffer-coordinate vector $b^+ = (h^+, (p^+)^\top)^\top \in \mathbb{R}^d$ by the first identity below, and set $\rho_B = B(t)/B(t^+)$ and $\Delta\ell = \ell(t, t^+)$. The change of frame gives

$$\begin{aligned}
\frac{m^+}{B(t^+)} &= \mathsf{Q}(t^+)b^+, \\
b^+ &= \frac{B(t)}{B(t^+)}\mathsf{Q}(t^+)^\top\big[\mathsf{Q}(t)\widehat{b} + e_m\big], \\
b^+ - \widehat{b} &= \rho_B\big[\mathsf{Q}(t^+)^\top\mathsf{Q}(t) - I_d\big]\widehat{b} \\
&\qquad + (\rho_B - 1)\widehat{b} + \rho_B\mathsf{Q}(t^+)^\top e_m.
\end{aligned}$$

The second identity uses Equation (67); subtracting $\widehat{b}$ separates the frame drift, the change in gradient magnitude, and the old-frame buffer error.

The frame-drift term satisfies

$$\begin{aligned}
\mathsf{Q}(t^+)^\top\mathsf{Q}(t) - I_d &= -\mathsf{Q}(t^+)^\top\int_t^{t^+}\frac{\mathrm{d}\mathsf{Q}(r)}{\mathrm{d}r}\,\mathrm{d}r, \\
\|\mathsf{Q}(t^+)^\top\mathsf{Q}(t) - I_d\|_{\mathrm{op}} &\le K_\kappa|\Delta\ell|.
\end{aligned}$$

The ratio of gradient magnitudes satisfies

$$\begin{aligned}
|\log\rho_B| &= \left|\int_t^{t^+}\frac{\mathrm{d}\log B(r)}{\mathrm{d}r}\,\mathrm{d}r\right| \le L_B|\Delta\ell|, \\
0 &< \rho_B \le e^{\eta L_B/q_0}, \\
|\rho_B - 1| &\le e^{\eta L_B/q_0}L_B|\Delta\ell|.
\end{aligned}$$

The last line applies the scalar mean-value theorem over the one-step arclength interval.

The sufficient choice $\widehat{M} = \beta M + (1-\beta)(1 + Q_\Theta R)$ follows from Equation (50):

$$\begin{aligned}
\|\widehat{b}\| &\le |\widehat{h}| + \|\widehat{p}\| \\
&\le \beta\big(|h| + \|p\|\big) + (1-\beta)\big(1 + Q_\Theta\|u\|\big) \\
&\le \beta M + (1-\beta)(1 + Q_\Theta R) = \widehat{M}.
\end{aligned}$$

Define the buffer-coordinate defect by $\xi_\beta^{\mathrm{buf}} := b^+ - \widehat{b}$. Using $\|e_m\| \le (1-\beta)\delta_{\mathrm{geo},\eta}$ and $|\Delta\ell| \le \eta/q_0$, the preceding exact separation gives

$$\begin{aligned}
\|\xi_\beta^{\mathrm{buf}}\| &\le e^{\eta L_B/q_0}\big[\|e_m\| + (K_\kappa + L_B)\widehat{M}|\Delta\ell|\big] \\
&\le e^{\eta L_B/q_0}\left[(1-\beta)\delta_{\mathrm{geo},\eta} + \frac{\eta(K_\kappa + L_B)\widehat{M}}{q_0}\right] \\
&= \varepsilon_{\mathrm{buf},\eta}.
\end{aligned}$$

Define $\xi_0 := \xi_0^u$; for $\beta > 0$, write $\xi_\beta^{\mathrm{buf}} := (\xi_\beta^h, \xi_\beta^p)^\top$ and set $\xi_\beta := (\xi_\beta^u, \xi_\beta^h, \xi_\beta^p)$. The path and buffer identities now give Equation (52).

The plain and momentum defects, respectively, satisfy

$$\begin{aligned}
|\zeta_0| + \|\xi_0\|_2 &\le \varepsilon_{0,\eta}, \\
|\zeta_\beta| + \|\xi_\beta\|_2 &\le |\zeta_\beta| + \|\xi_\beta^u\|_2 + \|\xi_\beta^{\mathrm{buf}}\|_2 \\
&\le \varepsilon_{\mathrm{path},\eta} + \varepsilon_{\mathrm{buf},\eta} = \varepsilon_\eta.
\end{aligned} \tag{71}$$

The Bishop equations and the definition of δ bound ηK_κ, ηL_B, and $\eta^2\Gamma_3/B_-$ by δ. The curvature identity used in Equation (58), together with $v^{-1}\mathrm{d}B/\mathrm{d}t = -e^\top\nabla^2 L(z)e$, gives

$$\begin{aligned}
\frac{1}{v}\frac{\mathrm{d}\kappa}{\mathrm{d}t} &= -\frac{N^\top\nabla^3 L(z)[e,e]}{B} - \frac{2}{vB}\frac{\mathrm{d}B}{\mathrm{d}t}\kappa - \frac{\Theta}{\eta}\kappa, \\
\eta^2 L_\kappa &\le \frac{\eta^2\Gamma_3}{B_-} + (2\eta L_B + \theta_+)\eta K_\kappa \le \delta + 2\delta^2 + \theta_+\delta.
\end{aligned}$$

Similarly, the first inequality in Equation (59) yields

$$\eta L_\Theta \le \frac{\eta^2\Gamma_3}{B_-} + 2(\eta K_\kappa)^2 + \theta_+\eta L_B \le \delta + 2\delta^2 + \theta_+\delta.$$

Thus, for $\delta \le 1$, the exact derivative bound is

$$\max\{\eta L_\Theta, \eta^2 L_\kappa\} \le (\theta_+ + 3)\delta.$$

This proves the stated derivative estimate with a coefficient depending only on θ_+.

The pointwise budget in Equation (60) satisfies

$$\delta_{\mathrm{geo},\eta} \le \eta K_\kappa R + \frac{\eta^2\Gamma_3 R^2}{2B_-} \le \big(R + R^2/2\big)\delta.$$

Take

$$c_{\mathrm{geo}} := \min\left\{\frac{1}{2}, \frac{1}{2(R+1)}, \frac{1}{2\max\{1, R + R^2/2\}}\right\}.$$

For momentum, replace this value by its minimum with $r_*/[2(1-\beta)\max\{1, R + R^2/2\}]$. These explicit choices imply Equation (61) with $q_0 = 1/2$ whenever $\delta \le c_{\mathrm{geo}}$. The radius, complete-ball, and state conditions are the remaining hypotheses of that calculation, so its update and defect identities apply.

Substitution of $q_0 = 1/2$ gives separate coefficients for position, phase, and buffer errors:

$$\begin{aligned} c_u &:= 1 + \frac{2(1-\beta)}{r_*}(R + R^2/2), \\ c_\ell &:= 3 + 2R + \frac{4(1-\beta)}{r_*}(R + R^2/2), \\ c_b &:= e\big[(1-\beta)(R + R^2/2) + 4\widehat{M}\big]. \end{aligned} \tag{72}$$

For plain descent, use $\beta = 0$, $r_* = 1$, and omit c_b. For momentum, any proved upper bound on $\widehat{r}$ may be used for $\widehat{M}$. In particular, $|h-1| \le H < 1$ and $\|\widehat{p}\| \le P$ permit $r_* = 1 - \beta H$ and $\widehat{M} = \sqrt{(1+\beta H)^2 + P^2}$. The exact estimates above and $\eta L_B/q_0 \le 2\delta \le 1$ yield

$$\begin{aligned} \|\xi_\beta^u\| &\le c_u\delta, & |\zeta_\beta| &\le c_\ell\delta, \\ \sqrt{|\xi_\beta^h|^2 + \|\xi_\beta^p\|^2} &\le c_b\delta, & \eta L_\Theta &\le (\theta_+ + 2)\delta. \end{aligned} \tag{73}$$

The buffer inequality is needed only for momentum and uses its Euclidean norm. The derivative inequality uses $\delta + 2\delta^2 + \theta_+\delta \le (\theta_+ + 2)\delta$. Thus $\varepsilon_{0,\eta} \le (c_u + c_\ell)\delta$ for plain descent and $\varepsilon_\eta \le (c_u + c_\ell + c_b)\delta$ for momentum. Taking C_{geo} to be the maximum of $\theta_+ + 3$ and the applicable sum proves the stated joint bound. The estimates hold uniformly over the prescribed state collection because all quantities on their right-hand sides are uniform bounds.

It remains to verify the complete-ball assertion from the normal-radius and arclength margins. The proof uses only $\eta(R+1) < \rho$, $J^{+2\eta} \subset \operatorname{int} I$, and $\delta(R+1) \le 1/2$; it does not need the other smallness conditions imposed on c_{geo}. Fix $x = \Psi(t,u)$ with $t \in J$ and $\|u\| \le R$, and fix $w \in \mathbb{R}^d$ with $\|w\| \le 1$. We lift the segment $x(r) := x + r\eta w$, $0 \le r \le 1$, into Fermi coordinates starting from (t,u). The Fermi differential is invertible at the starting point because $1 - \eta\kappa(t)^\top u \ge 1 - \delta R \ge 1/2$. On every interval where the lift $(t(r), u(r))$ is defined, differentiating $x(r) = \Psi(t(r), u(r))$ gives

$$\begin{aligned} \frac{\mathrm{d}u(r)}{\mathrm{d}r} &= N(t(r))^\top w, \\ \frac{\mathrm{d}t(r)}{\mathrm{d}r} &= \eta\frac{\langle e(t(r)), w\rangle}{v(t(r))[1 - \eta\kappa(t(r))^\top u(r)]}. \end{aligned}$$

Consequently, throughout its maximal lifted interval in $[0,1]$,

$$\begin{aligned} \|u(r)\| &\le \|u\| + r \le R + 1, \\ 1 - \eta\kappa(t(r))^\top u(r) &\ge 1 - \delta(R+1) \ge 1/2, \\ |\ell(t, t(r))| &\le 2\eta r. \end{aligned}$$

The lifted coordinates therefore remain in the compact set

$$J^{+2\eta} \times \{v \in \mathbb{R}^{d-1} : \|v\| \le R+1\},$$

which is contained in the coordinate domain by the arclength margin and $\eta(R+1) < \rho$. The preceding derivative bounds and $\inf_{J^{+2\eta}} v > 0$ make the lifted coordinates converge at any finite endpoint of their maximal interval. At the limit, the Fermi differential remains invertible, and its local inverse extends the lift. Thus the lift exists through $r = 1$, and $x+\eta w \in \mathcal{X}$. Since this argument holds for every $\|w\| \le 1$, the entire closed ball is contained in $\mathcal{X}$. Our choice $c_{\mathrm{geo}}(R+1) \le 1/2$ proves the statement. The more general condition $\delta(R+1) \le 1/2$ also gives a useful numerical specialization: $\rho > 2\eta$, $[0, T_*]^{+4\eta} \subset \operatorname{int} I$, and $\delta \le 1/4$ suffice for the region $t \in [0, T_*]^{+2\eta}$, $\|u\| \le 1$. □

Local period-two orbits and contraction. On any interval where the top eigenvalue is simple, let $v_1(t)$ be a continuously differentiable unit top eigenvector. Such a choice exists on an interval, and changing its sign exchanges the two cycle points used below.

For $\beta \in [0, 1)$, wherever the top eigenvalue is simple and $\theta_1(t) > 2(1+\beta)/(1-\beta)$, define $r_\beta(t) := \sqrt{1/4 - (1+\beta)^2/[(1-\beta)^2\theta_1(t)^2]}$ and the reference state sets

$$\mathcal{C}_0(t) := \{\chi r_0(t) v_1(t) : \chi \in \{-1, 1\}\},$$

$$\mathcal{C}_\beta(t) := \left\{ \left(\chi r_\beta(t) v_1(t), 1, -\chi \frac{1-\beta}{1+\beta} \theta_1(t) r_\beta(t) v_1(t) \right) : \chi \in \{-1, 1\} \right\} \quad (\beta > 0). \tag{74}$$

These sets lie in $\mathbb{R}^{d-1}$ for plain descent and $\mathbb{R}^{2d-1}$ for momentum; the latter includes both position and buffer coordinates. The spectral condition ensures the existence of these cycles, while δ controls variation of the local model over one update. For a fixed loss, reducing η decreases both δ and Θ, so the small-variation and spectral conditions must be checked together. The contraction estimates below provide the stability needed to control actual updates under the errors just bounded.

Theorem F.4 (Local period-two orbits and uniform contraction)**.** *Under Assumption F.1, fix $\beta \in [0, 1)$ and a compact interval $J \subseteq I$ on which the top eigenvalue is simple and $\inf_J \theta_1 > 2(1+\beta)/(1-\beta)$. The reference states $\mathcal{C}_\beta(t)$ in Equation (74) form a period-two orbit of $F_{\beta,t}$ with normalized physical progress $a_{\beta,*}(t) = 2(1+\beta)/[(1-\beta)\theta_1(t)]$. There exist $C \ge 1$, $R > 0$, $\sigma_\beta \in (0, 1/2]$, and positive-definite matrices $M_{\beta,\chi}(t)$, with $M_{0,\chi}(t) = I_{d-1}$, such that the following holds uniformly for $t \in J$ and $\chi \in \{-1, 1\}$. The side norms satisfy $\|v\|_2 \le \|v\|_{M_{\beta,\chi}(t)} \le C\|v\|_2$. For a cycle point w on side χ and $\|Z - w\|_{M_{\beta,\chi}(t)} \le R$,*

$$\|F_{\beta,t}(Z) - F_{\beta,t}(w)\|_{M_{\beta,-\chi}(t)} \le (1-\sigma_\beta)\|Z - w\|_{M_{\beta,\chi}(t)}.$$

The cycle coordinates and the matrices $M_{\beta,\chi}$ are CL_Θ-Lipschitz with respect to $|\ell(t,t')|$, in Euclidean and operator norm, respectively. On these neighborhoods, $a_\beta(t,\cdot)$ is positive and C-Lipschitz in Euclidean norm. The constants depend only on d, β, θ_-, θ_+, $\inf_J \theta_1 - 2(1+\beta)/(1-\beta)$, and $\inf_J(\theta_1 - \theta_2)$, with the last dependence omitted when $d = 2$.

Proof of Theorem F.4. Write $Z_\chi(t)$ for the state on side χ of $\mathcal{C}_\beta(t)$, and set $\theta_{\mathrm{lo}} := \inf_J \theta_1 > 2(1+\beta)/(1-\beta)$. At $\beta = 0$, the normalization denominator at $Z_\chi = \chi r_0 v_1$ is $\sqrt{1 + \theta_1^2 r_0^2} = \theta_1/2$. For $\beta > 0$, substitution into the buffer update gives $\widehat{h} = 1$, $\widehat{p} = \chi(1-\beta)\theta_1 r_\beta v_1/(1+\beta)$, and

$$\widehat{r}^2 = 1 + \frac{(1-\beta)^2\theta_1^2 r_\beta^2}{(1+\beta)^2} = \frac{(1-\beta)^2\theta_1^2}{4(1+\beta)^2}.$$

In either case, the position update subtracts $2\chi r_\beta v_1$; for momentum, the buffer also changes to that of $Z_{-\chi}$. Thus $F_{\beta,t}(Z_\chi) = Z_{-\chi}$ and $a_\beta(t, Z_\chi) = a_{\beta,*}(t)$, with

$$0 < \frac{2(1+\beta)}{(1-\beta)\theta_+} \le a_{\beta,*}(t) \le \frac{2(1+\beta)}{(1-\beta)\theta_{\mathrm{lo}}} < 1.$$

For regularity, set $g_1 := \inf_J \min\big(\{\theta_1\} \cup \{\theta_1 - \theta_i : 2 \le i \le d-1\}\big) > 0$; the singleton term covers $d = 2$. Differentiating the simple eigenpair and the amplitude gives

$$\frac{1}{v(t)} \left| \frac{\mathrm{d}\theta_1(t)}{\mathrm{d}t} \right| \le L_\Theta, \qquad \frac{1}{v(t)} \Big\| \frac{\mathrm{d}v_1(t)}{\mathrm{d}t} \Big\|_2 \le L_\Theta/g_1,$$

$$\frac{\mathrm{d}r_\beta(t)}{\mathrm{d}t} = \frac{(1+\beta)^2}{(1-\beta)^2 r_\beta \theta_1^3} \frac{\mathrm{d}\theta_1(t)}{\mathrm{d}t}.$$

The strict threshold bounds r_β away from zero, so the cycle coordinates are CL_Θ-Lipschitz with respect to path length, with C depending only on the stated data.

For $\beta = 0$, differentiating $F_{0,t}(u) = u - f(\Theta(t)u)$, where $f(y) := y/\sqrt{1+\|y\|^2}$, gives the same symmetric Jacobian A_t at both cycle points. Its multipliers are $1 - 8/\theta_1(t)^2$ in the top direction and $1 - 2\theta_i(t)/\theta_1(t)$ for $i > 1$. The threshold, positive spectral floor, and simple-top gap imply $\sup_J \|A_t\|_{\mathrm{op}} < 1$. Take $\sigma_0 := [1 - \sup_J \|A_t\|_{\mathrm{op}}]/2$, $M_{0,\chi} := I_{d-1}$, $C_{M,0} := 1$, and $L_{M,0} := 0$. Direct differentiation gives

$$\|D^2 f(y)\|_{\mathrm{op}} \le \frac{3\|y\|}{(1+\|y\|^2)^{3/2}} + \frac{3\|y\|^3}{(1+\|y\|^2)^{5/2}} \le 6,$$
$$\|D^2 F_{0,t}(u)\|_{\mathrm{op}} \le 6\theta_+^2.$$

Choose $R > 0$ with $6\theta_+^2 R \le \sigma_0$. Taylor's formula then gives, for $\|v\|_2 \le R$,

$$\|F_{0,t}(Z_\chi + v) - Z_{-\chi}\|_2 \le (1 - 2\sigma_0)\|v\|_2 + 3\theta_+^2\|v\|_2^2 \le (1-\sigma_0)\|v\|_2.$$

The speed a_0 is everywhere positive and has state derivative bounded by $L_{a,0} := \theta_+$. For the later tracking estimates, take $L_{\mathrm{cyc},0} := \sup_{t\in J} v(t)^{-1}(|\mathrm{d}r_0(t)/\mathrm{d}t| + r_0(t)\|\mathrm{d}v_1(t)/\mathrm{d}t\|_2) \le CL_\Theta$. The radius choice also gives $L_{a,0}R \le (2/\theta_+)/4$, completing Equations (75), (77) and (78) for NGD.

Now let $\beta > 0$. On the invariant slice $h = 1$, the change $w = \Theta(t)u$ turns the update into $y \mapsto \Phi_\beta(y;1)$ in Equation (39), with $\rho_i = \theta_i(t)$. The transformation and its inverse are uniformly bounded by θ_+ and θ_-^{-1}. Set $d_{1|1} := \{2(1+\beta)/[(1-\beta)\theta_1]\}^3$ and $d_{i|1} := 2(1+\beta)/[(1-\beta)\theta_1]$ for $i > 1$. The Jacobian block $J_{i|1}^{(w,p)}$ is the matrix in Equation (43) with $\rho_i = \theta_i$ and $c_i = d_{i|1}$, and

$$J_{i|1}^{(u,p)} = S_i^{-1} J_{i|1}^{(w,p)} S_i, \qquad S_i := \mathrm{diag}(\theta_i, 1).$$

For $a_{i|1} := \theta_i d_{i|1}$, the spectral margins give

$$a_{1|1} \in \left[\frac{8(1+\beta)^3}{(1-\beta)^3\theta_+^2}, \frac{8(1+\beta)^3}{(1-\beta)^3\theta_{\mathrm{lo}}^2}\right],$$
$$a_{i|1} \in \left[\frac{2(1+\beta)\theta_-}{(1-\beta)\theta_+}, \frac{2(1+\beta)(\theta_+ - g_1)}{(1-\beta)\theta_+}\right] \quad (i > 1).$$

These are compact subintervals of $(0, 2(1+\beta)/(1-\beta))$; the second is absent when $d = 2$. The strict Jury inequalities in the proof of Theorem E.4 therefore give uniform Schur stability of all these blocks.

Define the Euclidean isometry $\mathcal{R}(u,h,p) := (-u,h,-p)$. The update satisfies $F_{\beta,t}\mathcal{R} = \mathcal{R}F_{\beta,t}$, and $\mathcal{R}Z_\chi = Z_{-\chi}$. In the ordering of normal pairs followed by h, the full Jacobian has diagonal blocks $J_{i|1}^{(u,p)}$ and β. Its only off-diagonal coupling to h is $\partial_h F_{\beta,t}^u(Z_\chi) = \beta\widehat{p}/\widehat{r}^3$, which changes sign with χ; the normal variables do not feed back into h. Consequently, $A_t := \mathcal{R}DF_{\beta,t}(Z_1(t))$ has diagonal blocks $-J_{i|1}^{(u,p)}$ and β, so the spectral radius of A_t is less than one, uniformly in J, on the full state space.

On Euclidean balls of radius $R \le 1$ with $\beta R \le 1/2$ around the cycle points, $\widehat{h} \ge 1/2$. The normalization denominator therefore stays at least $1/2$, including along straight state interpolations. For $n(y) := y/\|y\|$, the bounds $\|Dn(y)\| \le \|y\|^{-1}$ and $\|D^2 n(y)\| \le 6\|y\|^{-2}$ control the first two state derivatives of $F_{\beta,t}$ and the state derivative of a_β. At fixed state, $v(t)^{-1}\partial_t(\widehat{h},\widehat{p}) = (0, (1-\beta)v(t)^{-1}(\mathrm{d}\Theta(t)/\mathrm{d}t)u)$. Together with the cycle derivative bound above, the chain rule gives a constant $C_0 \ge 1$, depending only on the stated data, such that

$$\|A_t\|_{\mathrm{op}} + \|D_Z^2 F_{\beta,t}\|_{\mathrm{op}} + \|D_Z a_\beta(t,\cdot)\|_{\mathrm{op}} \le C_0,$$
$$\|A_{t'} - A_t\|_{\mathrm{op}} + \|Z_1(t') - Z_1(t)\|_2 \le C_0 L_\Theta |\ell(t,t')|.$$

Take $L_{a,\beta} := C_0$. Schur stability need not give Euclidean contraction at each step, so define the convergent Lyapunov series

$$M_t := M_{\beta,1}(t) := \sum_{n=0}^{\infty} (A_t^\top)^n A_t^n, \qquad M_{\beta,-1}(t) := \mathcal{R}^\top M_t \mathcal{R}.$$

Then $M_t \succeq I_{2d-1}$ and one linearized step removes the current squared Euclidean error:

$$\|A_t v\|_{M_t}^2 = \sum_{n=1}^{\infty} \|A_t^n v\|_2^2 = \|v\|_{M_t}^2 - \|v\|_2^2.$$

For $\mathcal{L}_t(X) := X - A_t^\top X A_t$, this is $\mathcal{L}_t(M_t) = I_{2d-1}$. The operator is invertible: $\mathcal{L}_t(X) = 0$ implies $X = (A_t^\top)^n X A_t^n \to 0$. Set $C_M^2 := \sup_J \|M_t\|_{\mathrm{op}}$ and $M_{\mathcal{L}} := \sup_J \|\mathcal{L}_t^{-1}\|$, using the operator norm induced by the matrix spectral norm. Since the spectral matrices and their unit top eigenvectors range over a compact set determined by the stated data, continuity of inversion makes both constants finite and uniform in those data. Thus $\|v\|_2 \le \|v\|_{M_{\beta,\chi}(t)} \le C_M \|v\|_2$. Choose

$$L_M := M_{\mathcal{L}} C_0^2 C_M^2 L_\Theta,$$
$$\sigma := \frac{1 - \sqrt{1 - C_M^{-2}}}{2}.$$

The decrease identity gives $\|A_t v\|_{M_t} \le (1 - 2\sigma)\|v\|_{M_t}$. Choose $R > 0$ with $R \le 1$, $\beta R \le 1/2$, $L_{a,\beta} R \le (1+\beta)/[2(1-\beta)\theta_+]$, and $C_M C_0 R \le \sigma$. For $Z = Z_1 + v$ with $\|v\|_{M_t} \le R$, Taylor's formula yields

$$\begin{aligned}\|F_{\beta,t}(Z) - Z_{-1}\|_{M_{\beta,-1}(t)} &= \|\mathcal{R} F_{\beta,t}(Z_1 + v) - Z_1\|_{M_t} \\ &\le \|A_t v\|_{M_t} + \frac{C_M C_0}{2}\|v\|_2^2 \le (1 - \sigma)\|v\|_{M_t}.\end{aligned}$$

The equality uses $\mathcal{R} Z_{-1} = Z_1$; the last inequality uses $\|v\|_2 \le \|v\|_{M_t} \le R$ and the radius choice. Applying this estimate to $\mathcal{R} Z$ and using equivariance proves the reverse-side contraction.

To compare norms at nearby reference times, the inverse-difference identity gives

$$M_{t'} - M_t = \mathcal{L}_{t'}^{-1}(\mathcal{L}_t - \mathcal{L}_{t'}) M_t,$$
$$\|M_{t'} - M_t\|_{\mathrm{op}} \le 2 M_{\mathcal{L}} C_0^2 C_M^2 L_\Theta |\ell(t,t')| = 2 L_M |\ell(t,t')|.$$

Taking quadratic forms, using $M_t \succeq I_{2d-1}$ and $\sqrt{1+2x} \le 1 + x$, gives Equation (77); conjugation by $\mathcal{R}$ gives the same bound on the other side. Restore the method index on C_M, L_M, σ and take $L_{\mathrm{cyc},\beta} := C_{M,\beta} C_0 L_\Theta$. These choices supply Equations (75) and (78), with $L_{M,\beta} + L_{\mathrm{cyc},\beta} \le C L_\Theta$. The speed bound and radius choice give $a_\beta(t, Z) \ge 3 \inf_J a_{\beta,*}/4 > 0$. Enlarge the final C to dominate the norm-equivalence and regularity coefficients in both cases. □

Actual tracking and iteration counts. For a supplied initial state sufficiently close to a reference cycle, the local contraction estimates and the Fermi error bounds control the distance of the actual iterates to the moving cycle. Together with positive arclength progress, this gives finite crossing of the endpoint; summing the arclength errors then gives the reference-time and iteration bounds. The reference curve $z(t)$ is the curvature-adaptive flow in Equation (9), with endpoint time T_*. For an iterate, its projected time is $t_k = \pi(x_k)$; a superscript β distinguishes the momentum trajectory when both methods are compared. The proofs also give quantitative bounds for separately prescribed state and arclength defects, with entry condition Equation (81) and accumulated error Equation (84).

Proposition F.5 (Actual tracking and finite crossing)**.** *Under Assumption F.1, fix $\beta \in [0,1)$ and suppose that the top normal eigenvalue is simple on I and $\theta_{1,-} > 2(1+\beta)/(1-\beta)$. There exist $r \in (0, 1/2)$ and $c, C, \delta_0 > 0$ with the following property. Suppose that $\rho > 2\eta$, $[0, T_*]^{+4\eta} \subset \operatorname{int} I$, and $\delta \le \delta_0$. Run Equation (47) from $x_0 \in \mathcal{X}$ with $0 \le t_0 < T_*$, retaining the supplied momentum buffer when $\beta > 0$. For each generated index, let $D_k := \mathrm{dist}_2(Z_k, \mathcal{C}_\beta(t_k))$. If $D_0 < r$, the first crossing $N := \inf\{k \ge 0 : t_k \ge T_*\}$ is finite, all updates through it are defined in $\mathcal{X}$, and*

$$D_k \le C(e^{-ck} D_0 + \delta) < 1/2 \quad (0 \le k \le N).$$

The constants depend only on d, β, B_+/B_-, θ_-, θ_+, the positive margin $\theta_{1,-} - 2(1+\beta)/(1-\beta)$, and $\inf_I(\theta_1 - \theta_2)$, with the last dependence omitted when $d = 2$. They are uniform in the learning rate and section length when these dimensionless data are fixed.

Proof of Proposition F.5. We first prove a quantitative version for any supplied contracting period-two orbit with positive arclength speed. Assume Assumption F.1, fix $\beta \in [0,1)$, and let $I_\Delta :=$

$[0, T_*]^{+\Delta} \subset \operatorname{int} I$ for some $\Delta > 0$. Index the two supplied cycle sides by $\chi \in \{-1, 1\}$ and give them positive-definite metrics $M_{\beta,\chi}(t)$. Suppose that $F_{\beta,t}$ interchanges the sides and, for some $R > 0$ and $\sigma_\beta \in (0, 1/2]$, satisfies

$$\|F_{\beta,t}(Z) - F_{\beta,t}(w)\|_{M_{\beta,-\chi}(t)} \le (1 - \sigma_\beta)\|Z - w\|_{M_{\beta,\chi}(t)}, \tag{75}$$

whenever $t \in I_\Delta$, w is the cycle point on side χ, and $\|Z - w\|_{M_{\beta,\chi}(t)} \le R$. At either cycle point require $a_\beta(t, w) = a_{\beta,*}(t)$, where $a_{\beta,-} := \inf_{I_\Delta} a_{\beta,*} > 0$ and $\sup_{I_\Delta} a_{\beta,*} < 1$. Let $d_\beta(t, Z)$ be the smaller of the two distances from Z to the cycle points, each measured in its associated side norm. Reindex a supplied trajectory of Equation (47), retaining its momentum buffer, with $x_0 \in \mathcal{X}$ and $t_0 \in [0, T_*)$. Define $N := \inf\{k \ge 0 : t_k \ge T_*\}$ and $d_k := d_\beta(t_k, Z_k)$ for each generated index. All bounds below are uniform on I_Δ.

For the reference sets in Equation (74), the corresponding distances are

$$\begin{aligned} d_0(t, u) &= \min_{\chi=\pm1} \|u - \chi r_0(t) v_1(t)\|_2, \\ d_\beta(t, (u, h, p)) &= \min_{\chi=\pm1} \|\left(u - \chi r_\beta v_1,\ h - 1,\ p + \frac{1-\beta}{1+\beta}\chi r_\beta \theta_1 v_1\right)\|_{M_{\beta,\chi}(t)} \quad (\beta > 0), \end{aligned} \tag{76}$$

where the cycle coordinates on the second line are evaluated at t. The two momentum sides have the same tangential buffer component $h = 1$. Plain descent uses $M_{0,\pm1} = I_{d-1}$; the momentum side norm acts on the full state in $\mathbb{R}^{2d-1}$.

Fix $C_{M,\beta} \ge 1$ and $L_{M,\beta}, L_{\mathrm{cyc},\beta}, L_{a,\beta} \ge 0$ satisfying

$$\begin{aligned} I \preceq M_{\beta,\chi}(t) &\preceq C_{M,\beta}^2 I, \\ \|v\|_{M_{\beta,\chi}(t')} &\le (1 + L_{M,\beta}|\ell(t, t')|)\|v\|_{M_{\beta,\chi}(t)}. \end{aligned} \tag{77}$$

For same-side cycle points w, w' at t, t', the cycle-motion and speed assumptions read

$$\begin{aligned} \|w' - w\|_{M_{\beta,\chi}(t)} &\le L_{\mathrm{cyc},\beta}|\ell(t, t')|, \\ |a_\beta(t, Z) - a_{\beta,*}(t)| &\le L_{a,\beta} d_\beta(t, Z), \\ L_{a,\beta} R &\le a_{\beta,-}/4. \end{aligned} \tag{78}$$

The speed estimate holds on the radius-R cycle neighborhood. Assume every physical update from these neighborhoods is defined in $\mathcal{X}$ and obeys the Fermi recurrence, with uniform separate bounds $\|\xi_\beta\|_2 \le \varepsilon_Z$ and $|\zeta_\beta| \le \varepsilon_{\mathrm{ph}}$. One sufficient condition, preserving the full quantitative Fermi regime, is that each neighborhood satisfy Equation (61), including its complete-step ball condition, with uniform bounds $\|u\| \le U$ and, for momentum, $|h| + \|p\| \le M$ and $\widehat{r} \ge r_* > 0$, for fixed U, M, r_* and $q_0 \in (0, 1]$. Here the normal radius in the Fermi conditions is U, while R measures distance to the cycle in the side norm. For the concrete cycles in Equation (76), this neighborhood satisfies $\|u\| \le 1/2 + R$. For momentum, $I \preceq M_{\beta,\chi}$ also gives $|h| + \|p\| \le M := 1 + (1-\beta)\theta_+/[2(1+\beta)] + 2R$ and $\widehat{r} \ge \widehat{h} \ge r_* := 1 - \beta R$ whenever $\beta R < 1$; the local-cycle construction uses $\beta R \le 1/2$. Under these conditions, Lemma F.3 supplies the errors Equations (69) and (70). Require

$$\begin{aligned} \frac{3}{2}\eta &< \Delta, \\ \varepsilon_{\mathrm{ph}} &\le a_{\beta,-}/4, \\ \frac{3}{2} L_{M,\beta}\eta &\le \sigma_\beta/2. \end{aligned} \tag{79}$$

The first condition provides room for the crossing update; the others preserve forward motion and contraction. Define the one-step tracking error by

$$\delta_{\beta,\eta} := \frac{3}{2}\left(C_{M,\beta}\varepsilon_Z + \frac{3}{2} L_{\mathrm{cyc},\beta}\eta\right). \tag{80}$$

For an entry with $t_0 \in [0, T_*)$ and $d_0 = d_\beta(t_0, Z_0)$, require the reserve

$$d_0 + \frac{2\delta_{\beta,\eta}}{\sigma_\beta} < R. \tag{81}$$

The displayed reserve retains the full quantitative entry range. Fix the chosen method and suppress its index on the constants, writing $q := 1 - \sigma_\beta/2$. Stop at the first crossing of the endpoint or the first failure of $d_k < R$ or $t_k \in [0, T_*)$. Also stop if the operational chart or a normalization fails, and call this index H. At entry, the strict reserve places the state in the radius-R neighborhood. The supplied complete-step update condition makes the first normalization valid. When verified through Equation (61), this follows from the predicted-buffer floor for momentum and the nonzero-gradient bound for plain descent in the Fermi proof. Thus $H > 0$.

At each $k < H$, the complete-step update condition makes the next physical update well defined, with endpoint in $\mathcal{X}$. The arclength-speed error and forward increment satisfy

$$|a_\beta(t_k, Z_k) - a_{\beta,*}(t_k)| + |\zeta_k| \le L_a R + \varepsilon_{\mathrm{ph}} \le a_-/2,$$
$$\frac{1}{2}\eta a_{\beta,*}(t_k) \le \ell(t_k, t_{k+1}) \le \eta[a_{\beta,*}(t_k) + a_-/2] < \frac{3}{2}\eta. \tag{82}$$

The first line uses $d_k < R$ and the arclength-speed regularity bound; the second follows from the arclength recurrence and $0 < a_- \le a_{\beta,*}(t_k) < 1$. The endpoint lies in $[0, T_*]^{+3\eta/2} \subset I_\Delta$ even if it crosses T_*. Choose a sign χ attaining the minimum that defines d_k. By Equation (52), the algorithm state update differs from $F_{\beta,t_k}(Z_k)$ by a Euclidean error of at most ε_Z. Measure its error relative to the cycle point on side $-\chi$ at t_{k+1}, then minimize over the next sign. The contraction, matrix variation, and cycle-motion bounds give, for $k < H$,

$$\begin{aligned} d_{k+1} &\le (1 + 3L_M\eta/2)[(1-\sigma)d_k + C_M\varepsilon_Z + 3L_{\mathrm{cyc}}\eta/2] \\ &\le q d_k + \delta_{\beta,\eta}. \end{aligned}$$

The first line accounts for contraction toward the opposite cycle state at t_k, its motion to t_{k+1}, and the change of its Mahalanobis norm. The bounds $3L_M\eta/2 \le \sigma/2$ and $0 < \sigma \le 1/2$ imply $(1+\sigma/2)(1-\sigma) \le 1 - \sigma/2 = q$ and $1 + \sigma/2 \le 3/2$, giving the scalar recurrence. Iterating through H and applying the strict entry reserve gives

$$\begin{aligned} d_k &\le q^k d_0 + \frac{\delta_{\beta,\eta}(1-q^k)}{1-q} \\ &\le q^k d_0 + \frac{\delta_{\beta,\eta}}{1-q} \\ &\le d_0 + \frac{\delta_{\beta,\eta}}{1-q} < R. \end{aligned}$$

If $H < \infty$ and $t_H < T_*$, forward motion gives $t_H \in [0, T_*)$, and the tracking bound gives $d_H < R$. The preceding complete-step condition places x_H in the operational chart, and the supplied update condition permits the next normalization. Thus the trajectory cannot stop before crossing. If $H = \infty$, the lower bound on the arclength increment $\eta a_-/2 > 0$ contradicts $t_k < T_*$ for all k. It follows that $H = N < \infty$. The distance estimate includes the crossing endpoint. In particular, the exact tracking bound is

$$d_\beta(t_k, Z_k) \le (1 - \sigma_\beta/2)^k d_0 + \frac{2\delta_{\beta,\eta}}{\sigma_\beta}[1 - (1 - \sigma_\beta/2)^k] < R. \tag{83}$$

This bound and crossing hold for any supplied positive arclength speed under the same entry and quantitative conditions, without requiring a constant reference-time increment.

For the cycles in Equation (76), a minimizing cycle point has normal-coordinate norm $r_\beta(t_k) < 1/2$, with $r_\beta = r_0$ when $\beta = 0$. The Euclidean lower bound on the Mahalanobis norm and Equation (45) give, before crossing,

$$\|x_k - z(t_k)\| = \eta\|u_k\| \le \eta[r_\beta(t_k) + d_k] < \eta(1/2 + R).$$

For the geometric statement, use the cycle constructions in Theorem F.4 on I_Δ with $\Delta = 2\eta$. They give the supplied metrics, contraction, and positive speed, with $C_{M,\beta}$, $\sigma_\beta > 0$, and $L_{a,\beta}$ depending only on the listed spectral data. Decrease R so that $R < 1/4$, $L_{a,\beta}R \le a_{\beta,-}/4$, and $\beta R \le 1/2$. Let C_0 depend only on the stated data, increasing it as needed below. The cycle constructions give $L_{M,\beta} + L_{\mathrm{cyc},\beta} \le C_0 L_\Theta$. Since $\|u\| \le 1/2 + R < 1$ on these neighborhoods, apply the geometric

part of Lemma F.3 with its normal radius set to 1 to obtain $\eta L_\Theta \le C_0\delta$ and $\varepsilon_Z + \varepsilon_{\mathrm{ph}} \le C_0\delta$. For momentum use the established bounds $|h| + \|p\| \le M = 1 + (1-\beta)\theta_+/[2(1+\beta)] + 2R$ and $\widehat{r} \ge r_* = 1 - \beta R \ge 1/2$. The hypotheses $\rho > 2\eta$ and $[0, T_*]^{+4\eta} \subset \operatorname{int} I$ give the complete-step balls throughout I_Δ. After decreasing the allowable δ, these neighborhoods therefore satisfy the exact Fermi conditions with $q_0 = 1/2$. The Fermi bounds add only the listed dependence on B_+/B_-. After enlarging C_0 once more, Equation (80) gives $\delta_{\beta,\eta} \le C_0\delta$.

The metric comparison gives $D_k \le d_k \le C_{M,\beta} D_k$. Take $c := -\log(1 - \sigma_\beta/2)$ and $C := \max\{C_{M,\beta}, 2C_0/\sigma_\beta, 1\}$. Choose $r \in (0, 1/2)$ so that $C_{M,\beta} r < R/4$ and $Cr < 1/4$. Choose $\delta_0 \le 1$ small enough to satisfy the Fermi conditions and Equation (79), and also require $2C_0\delta_0/\sigma_\beta \le R/4$ and $C\delta_0 \le 1/4$. Then $D_0 < r$ implies the strict reserve Equation (81), so the quantitative argument gives legality and finite crossing. For $0 \le k \le N$, Equation (83) and the metric comparison now give

$$\begin{aligned} D_k \le d_k &\le e^{-ck} d_0 + \frac{2\delta_{\beta,\eta}}{\sigma_\beta} \\ &\le C_{M,\beta} e^{-ck} D_0 + \frac{2C_0}{\sigma_\beta}\delta \\ &\le C(e^{-ck} D_0 + \delta) < 1/2. \end{aligned}$$

□

Proposition F.6 (Reference time and iteration count)**.** *Under the hypotheses of Proposition F.5, use its trajectory and crossing index N. There is $C > 0$, depending only on the data listed there, such that*

$$\left| t_k - t_0 - \frac{1+\beta}{1-\beta} k \right| \le C(1 + k\delta) \quad (0 \le k \le N).$$

For any fixed $C_{\mathrm{ent}} \ge 0$, if also $\ell(0, t_0) \le C_{\mathrm{ent}}\eta$, then, after enlarging C to allow dependence on C_{ent},

$$\left| \frac{1+\beta}{1-\beta} N - T_* \right| \le C(1 + \delta T_*).$$

The constant is uniform in the learning rate and section length when the listed dimensionless data are fixed.

Proof of Proposition F.6. Use the quantitative setting established in the proof of Proposition F.5, including its separate errors, complete-step conditions, and strict entry reserve. In addition, suppose that for every $t \in I_\Delta$ and either cycle point w,

$$\begin{aligned} F_{\beta,t}^2(w) &= w, \\ a_\beta(t, w) &= a_{\beta,*}(t) = \frac{2K}{\theta_1(t)}, \end{aligned}$$

where $K > 0$. The preceding proof supplies finite crossing, Equations (82) and (83), and their scalar tracking recurrence with $q := 1 - \sigma_\beta/2$. We use these conclusions directly to bound the accumulated arclength error and convert it to reference time.

For the exact budget, define $\mathcal{E}_\beta(0; d) := 0$ for every $d \ge 0$ and, for $n \ge 1$,

$$\begin{aligned} \mathcal{E}_\beta(n; d) := \frac{\theta_+ L_{a,\beta}}{\sigma_\beta} d \\ + n\left[\frac{\theta_+ L_{a,\beta} \delta_{\beta,\eta}}{\sigma_\beta} + \frac{\theta_+}{2}\varepsilon_{\mathrm{ph}} + \frac{9\eta L_{\theta_1}}{16} \right]. \end{aligned} \tag{84}$$

For $k < N$, define $e_k := a_\beta(t_k, Z_k) - a_{\beta,*}(t_k) + \zeta_k$. The arclength recurrence and the arclength-speed regularity bound give

$$\begin{aligned} \frac{\ell(t_k, t_{k+1})}{\eta} &= a_{\beta,*}(t_k) + e_k, \\ |e_k| &\le L_{a,\beta} d_k + \varepsilon_{\mathrm{ph}}. \end{aligned} \tag{85}$$

For $1 \le n \le N$, summing the tracking recurrence and the arclength errors gives

$$\begin{aligned}\sum_{k=0}^{n-1} |e_k| &\le L_{a,\beta} \sum_{k=0}^{n-1} d_k + n\varepsilon_{\mathrm{ph}} \\ &\le \frac{L_{a,\beta}}{1-q}(d_0 - d_n + n\delta_{\beta,\eta}) + n\varepsilon_{\mathrm{ph}} \\ &\le \frac{2L_{a,\beta}}{\sigma_\beta}(d_0 + n\delta_{\beta,\eta}) + n\varepsilon_{\mathrm{ph}}. \end{aligned} \tag{86}$$

The first inequality uses Equation (85); the second sums $d_{k+1} - d_k \le -(1-q)d_k + \delta_{\beta,\eta}$, and the last uses $d_n \ge 0$ and $1 - q = \sigma_\beta/2$. The final bound also holds at $n = 0$. When $a_{\beta,*}(t) = 2K/\theta_1(t)$, Equation (82) gives $\ell(t_k, t_{k+1}) \ge K\eta/\theta_+$ before crossing. Consequently,

$$N \le 1 + \left\lceil \frac{\theta_+ S}{K\eta} \right\rceil.$$

For $k < N$, the reference-speed identity $\theta_1(t)v(t) = 2\eta$, the arclength recurrence, and the cycle identity give

$$\begin{aligned} t_{k+1} - t_k - K &= \frac{1}{2\eta} \int_{t_k}^{t_{k+1}} \theta_1(q)v(q)\,\mathrm{d}q - K \\ &= \frac{\theta_1(t_k)}{2} e_k + \frac{1}{2\eta} \int_{t_k}^{t_{k+1}} [\theta_1(q) - \theta_1(t_k)]v(q)\,\mathrm{d}q, \\ |t_{k+1} - t_k - K| &\le \frac{\theta_+}{2}|e_k| + \frac{L_{\theta_1}}{4\eta}\ell(t_k, t_{k+1})^2 \\ &\le \frac{\theta_+}{2}|e_k| + \frac{9\eta L_{\theta_1}}{16}. \end{aligned}$$

The inequalities use Lipschitz continuity of θ_1 with respect to path length and $\ell(t_k, t_{k+1}) < 3\eta/2$ from Equation (82). These estimates include the crossing update $k = N - 1$. Summing and using Equations (84) and (86) gives

$$\begin{aligned} |t_n - t_0 - Kn| &\le \frac{\theta_+}{2} \sum_{k=0}^{n-1} |e_k| + \frac{9n\eta L_{\theta_1}}{16} \\ &\le \mathcal{E}_\beta(n; d_0) \quad (0 \le n \le N). \end{aligned} \tag{87}$$

For the iteration-count estimate, additionally assume $0 \le \ell(0, t_0) \le C_{\mathrm{ent}}\eta$, where $C_{\mathrm{ent}} \ge 0$. The terminal offset satisfies $0 \le \ell(T_*, t_N) < 3\eta/2$ by Equation (82). The endpoint decomposition gives

$$\begin{aligned} |KN - T_*| &\le |KN - (t_N - t_0)| + |t_N - T_*| + |t_0| \\ &\le \mathcal{E}_\beta(N; d_0) + \frac{\theta_+}{2}\left(\frac{3}{2} + C_{\mathrm{ent}}\right). \end{aligned} \tag{88}$$

Here $v(t)^{-1} = \theta_1(t)/(2\eta) \le \theta_+/(2\eta)$ converts the initial and terminal arclength offsets into reference-time offsets.

For the cycles in Equation (74), the local calculations in Theorem F.4 give $K = (1+\beta)/(1-\beta)$. The geometric specialization in the proof of Proposition F.5 gives $\varepsilon_{\mathrm{ph}} + \delta_{\beta,\eta} + \eta L_{\theta_1} \le C_1\delta$, while $d_0 < R$ and all remaining coefficients in Equation (84) are bounded in terms of the stated data. Substituting these estimates in the exact budget gives

$$\mathcal{E}_\beta(n; d_0) \le C_2(d_0 + n\delta) \le C_2(1 + n\delta) \quad (0 \le n \le N).$$

For $1 \le n \le N$, this also gives $\mathcal{E}_\beta(n; d_0)/n \le C_2(\delta + 1/n)$. Taking $C = C_2$ in Equation (87) proves the reference-time bound without any dependence on C_{ent}. For the endpoint bound, the lower forward increment gives $N \le 2 + (1-\beta)\theta_+ S/[(1+\beta)\eta]$, and $T_* \ge \theta_{1,-} S/(2\eta)$ gives $N \le C_3(1 + T_*)$. Since $\delta \le 1$, inserting this estimate into Equation (88) proves the iteration bound after enlarging C to include the entry offset. □

NGM near the reference two-cycle. For $\beta \in (0,1)$, apply Propositions F.5 and F.6 to the joint state $Z_k^\beta = (u_k^\beta, h_k^\beta, p_k^\beta)$, with $K_\beta = (1+\beta)/(1-\beta)$. Their hypotheses require the position and both buffer components to start jointly close to the same cycle state; they do not assert entry from zero momentum. The crossing index $n_{\mathrm{P2}} := \inf\{k \geq 0 : t_k^\beta \geq T_*\}$ is finite, and the position component of $D_k < 1/2$ gives $\|x_k^\beta - z(t_k^\beta)\| \leq \eta(r_\beta(t_k^\beta) + D_k) < \eta$ through crossing. The two-sided time bound and $v \leq 2\eta/\theta_{1,-}$ therefore imply

$$\|x_k^\beta - z(t_0^\beta + K_\beta k)\| \leq C_\beta \eta(1 + \delta k),$$

for $0 \leq k \leq n_{\mathrm{P2}}$ whenever $t_0^\beta + K_\beta k \in I$, with C_β depending only on the data of those propositions. In the oscillatory phase of Equation (7), this reference is $X_\beta(k)$ with $X_\beta(0) = z(t_0^\beta)$. If $0 \leq \ell(0, t_0^\beta) \leq C_{\mathrm{ent}}\eta$ for fixed $C_{\mathrm{ent}} \geq 0$, the count clause of Proposition F.6 applies with $N = n_{\mathrm{P2}}$ and retains the stated dependence on this entry offset.

F.2 Proof for NGD dynamics

The results below concern `NormGD` updates on L, given by Equation (47) with $\beta = 0$. When the top normal eigenvalue is simple and $\theta_{1,-} > 2$, use the amplitude r_0 from Equation (74). The two positions $\pm r_0(t)v_1(t)$ form the reference period-two orbit of the local model with coefficients fixed at that reference time. The local contraction and regularity were established in Theorem F.4. We first prove finite-time entry from the prescribed initial set, then bound the iterations from entry to the endpoint. The proofs retain the quantitative conditions expressed through the Fermi defect before specializing to small geometric variation.

Finite-time entry.

Theorem F.7 (Finite-time NGD warm-up)**.** *Under Assumption F.1, suppose that the top normal eigenvalue is simple on I and $\theta_{1,-} > 2$. Fix a target accuracy $\epsilon > 0$ and a nonempty compact initial set*

$$\mathcal{K}_{\mathrm{prep}} \subset \big\{(t,u) \in \operatorname{int} I \times \mathbb{R}^{d-1} : \langle v_1(t), u\rangle \neq 0,\ \sqrt{1 + \|\Theta(t)u\|^2} < \theta_1(t)\big\}.$$

Choose a compact interval $I_{\mathrm{prep}} \subset \operatorname{int} I$ whose interior contains all time coordinates represented in $\mathcal{K}_{\mathrm{prep}}$. For the fixed initial set, target accuracy, and normal-sharpness family, there exist a horizon $N_{\mathrm{prep}} \in \mathbb{N}$, a tube-width coefficient $C > 1$, and a geometric threshold $\delta_0 > 0$. They depend only on d, $\mathcal{K}_{\mathrm{prep}}$, I_{prep}, Θ on I, B_+/B_-, and ϵ. Suppose that the geometric variation in Equation (51) *satisfies $\delta \leq \delta_0$, that $C\eta < \rho$, and that $I_{\mathrm{prep}}^{+2\eta} \subset \operatorname{int} I$. Every NGD trajectory initialized by $(t_0, u_0) \in \mathcal{K}_{\mathrm{prep}}$ with*

$$\{t_0\}^{+2N_{\mathrm{prep}}\eta} \subset I_{\mathrm{prep}}$$

has an index $k_{\mathrm{prep}} \leq N_{\mathrm{prep}}$ and a sign $\chi \in \{-1, 1\}$ such that

$$\|u_{k_{\mathrm{prep}}} - \chi r_0(t_{k_{\mathrm{prep}}})v_1(t_{k_{\mathrm{prep}}})\|_2 < \epsilon. \tag{89}$$

All updates through k_{prep} are defined in $\mathcal{X}$, and their time coordinates remain in I_{prep}.

Proof of Theorem F.7. We prove a quantitative version at the same time. In place of the geometric smallness, tube-width, and endpoint-margin conditions in the statement, suppose that the Fermi conditions in Equation (61) hold uniformly for $t \in I_{\mathrm{prep}}$ and normal coordinates in an open neighborhood of the trajectories of the local model from u, $(t,u) \in \mathcal{K}_{\mathrm{prep}}$, with $\|u\| \leq R$ for a fixed $R > 0$ and fixed $q_0 \in (0,1]$ satisfying $\eta K_\kappa(R+1) \leq 1 - q_0$. There exists $\varepsilon_{\mathrm{prep}} > 0$, depending on these uniform bounds, the warm-up data listed in the statement, and ϵ, such that every NGD trajectory with $(t_0, u_0) \in \mathcal{K}_{\mathrm{prep}}$ and

$$\varepsilon_{0,\eta} + \eta L_\Theta \leq \varepsilon_{\mathrm{prep}},$$
$$\{t_0\}^{+N_{\mathrm{prep}}\eta/q_0} \subset I_{\mathrm{prep}}$$

satisfies Equation (89) and the same validity conclusions, with the normal coordinates also remaining in the specified neighborhood through k_{prep}. The construction below chooses N_{prep} from the local dynamics before choosing either error threshold.

For $(t,u) \in \mathcal{K}_{\mathrm{prep}}$, set $\widehat{u}_0(t,u) := u$ and generate a trajectory of the local model with coefficients fixed at this reference time by

$$\widehat{u}_{j+1}(t,u) := F_{0,t}\big(\widehat{u}_j(t,u)\big).$$

Fix (t,u) and write $w_j := \Theta(t)\widehat{u}_j(t,u)$. The local map gives

$$w_{j+1} = \Theta(t)\widehat{u}_j - \frac{\Theta(t)^2\widehat{u}_j}{\sqrt{1+\|\Theta(t)\widehat{u}_j\|^2}} = \left[I_{d-1} - \frac{\Theta(t)}{\sqrt{1+\|w_j\|^2}}\right] w_j.$$

Apply the recurrence conclusion of Theorem E.1, identifying the action of Θ on its range with $\Theta(t)$ and setting $\rho_i = \theta_i(t)$. The initial set supplies its small-ball condition and nonzero top projection, since $\langle v_1(t), w_0\rangle = \theta_1(t)\langle v_1(t), u\rangle \neq 0$; also $\theta_1(t) > 2$ by the spectral assumption. Thus the top mode survives, and simplicity gives opposite parity limits with amplitude $\sqrt{\theta_1(t)^2/4 - 1} = \theta_1(t) r_0(t)$. The sign of the initial top projection is absorbed into χ. Applying the fixed inverse $\Theta(t)^{-1}$ gives

$$\begin{aligned} \widehat{u}_{2j}(t,u) &\longrightarrow \chi r_0(t) v_1(t), \\ \widehat{u}_{2j+1}(t,u) &\longrightarrow -\chi r_0(t) v_1(t). \end{aligned}$$

Hence, for each (t,u), some index j and sign χ satisfy

$$\|\widehat{u}_j(t,u) - \chi r_0(t) v_1(t)\|_2 < \epsilon/4.$$

This strict inequality persists on a neighborhood of (t,u) because each finite iterate of the local model and its reference period-two orbit are continuous in their inputs. A finite subcover of $\mathcal{K}_{\mathrm{prep}}$ therefore gives one uniform upper bound N_{prep} for the entrance time. Increase this bound to at least one if necessary.

The normal coordinates of all trajectories of the local model through N_{prep} form a compact set. Choose $U > 2$ so that every normal coordinate in these prefixes has norm at most $U - 2$. Under the geometric conditions, choose the tube-width constant $C > U + 1$; then $C\eta < \rho$ gives $\eta(U+1) < \rho$. Initially require $\delta_0 \leq \min\{1, [2(U+1)]^{-1}, [2U + U^2]^{-1}\}$, which verifies the geometric step conditions of Lemma F.3 with $R = U$. Its complete-ball conclusion applies on I_{prep} by the stated arclength margin, and its Fermi estimate gives $q_0 = 1/2$ and

$$\begin{aligned} \varepsilon_{0,\eta} + \eta L_\Theta &\leq C_1 \delta, \\ C_1 &:= 8U + 3U^2 + \theta_+ + 6, \end{aligned}$$

by Equations (72) and (73). Under the quantitative alternative, the uniform Fermi conditions are supplied directly. In either case, choose a smaller open neighborhood of the compact prefix set, contained in $\{u : \|u\| < U\}$, whose compact closure lies in the applicable uniform Fermi neighborhood. Let $\nu > 0$ be the distance of the compact set of local-model prefixes from the boundary of this smaller neighborhood. Let $H \leq N_{\mathrm{prep}}$ be the first exit index of u_j from this smaller neighborhood, capped at N_{prep}. Before this stopped index, the assumed arclength margin and the chart bound below inductively keep $t_j \in I_{\mathrm{prep}}$, where the uniform Fermi hypotheses apply. Thus the NGD recurrence in Fermi coordinates is

$$\begin{aligned} \ell(t_j, t_{j+1}) &= \eta[a_0(t_j, u_j) + \zeta_j], \\ u_{j+1} &= F_{0,t_j}(u_j) + \xi_j, \end{aligned} \tag{90}$$

where Equation (71) gives $|\zeta_j| + \|\xi_j\| \leq \varepsilon_{0,\eta}$. Summing the bound on each arclength increment in Equation (62) also gives

$$|\ell(t_0, t_j)| \leq \frac{j\eta}{q_0}, \qquad 0 \leq j \leq H. \tag{91}$$

For the NGD initial pair (t_0, u_0), compare u_j with the trajectory $\widehat{u}_j(t_0, u_0)$ of the local model and write $D_j := \|u_j - \widehat{u}_j\|_2$, so $D_0 = 0$. The map $f(y) = y/\sqrt{1+\|y\|^2}$ is one-Lipschitz. Comparing

Equation (90) with the local recurrence and using Equation (91) gives

$$\begin{aligned} D_{j+1} &\le (1+\theta_+)D_j + UL_\Theta|\ell(t_0,t_j)| + \varepsilon_{0,\eta}, \\ D_j &\le \sum_{i<j}(1+\theta_+)^{j-1-i}\left(\frac{UL_\Theta i\eta}{q_0} + \varepsilon_{0,\eta}\right), \\ \max_{0\le j\le H} D_j &\le C_{\mathrm{prep}}(\varepsilon_{0,\eta} + \eta L_\Theta), \\ C_{\mathrm{prep}} &:= \sum_{i=0}^{N_{\mathrm{prep}}-1}(1+\theta_+)^{N_{\mathrm{prep}}-1-i}\left(1+\frac{Ui}{q_0}\right). \end{aligned}$$

The first inequality holds for $j < H$, and iteration from $D_0 = 0$ gives the second for $j \le H$. The last bound is uniform because the horizon is fixed. Requiring $\varepsilon_{\mathrm{prep}} \le \nu/(2C_{\mathrm{prep}})$ under the quantitative conditions, or $C_1\delta_0 \le \nu/(2C_{\mathrm{prep}})$ under the geometric conditions, makes the last bound smaller than the neighborhood margin. Together with the respective arclength margin, this rules out $H < N_{\mathrm{prep}}$ and proves the bound through N_{prep}.

Put $r_- := \sqrt{1/4 - \theta_{1,-}^{-2}}$, $r_+ := \sqrt{1/4 - \theta_+^{-2}}$, and $\Delta_{\mathrm{sp}} := \inf_I(\theta_1 - \theta_2) > 0$ when $d > 2$. The simple-eigenpair and amplitude derivatives in the proof of Theorem F.4 give

$$\begin{aligned} L_{\mathrm{cyc},0} &\le L_\Theta\left(\frac{1}{r_-\theta_{1,-}^3} + \frac{r_+}{\Delta_{\mathrm{sp}}}\right), \\ C'_{\mathrm{prep}} &:= C_{\mathrm{prep}} + \frac{N_{\mathrm{prep}}}{q_0}\left(\frac{1}{r_-\theta_{1,-}^3} + \frac{r_+}{\Delta_{\mathrm{sp}}}\right), \end{aligned}$$

with the terms involving Δ_{sp} omitted when $d = 2$. At the entrance index $k = k_{\mathrm{prep}}$ of the local model, the triangle inequality, bound on arclength drift, and regularity construction give

$$\begin{aligned} \|u_k - \chi r_0(t_k)v_1(t_k)\|_2 &\le D_k + \epsilon/4 + L_{\mathrm{cyc},0}|\ell(t_0,t_k)| \\ &\le \epsilon/4 + C'_{\mathrm{prep}}(\varepsilon_{0,\eta} + \eta L_\Theta). \end{aligned}$$

Thus the explicit choices

$$\begin{aligned} \varepsilon_{\mathrm{prep}} &:= \min\left\{\frac{\nu}{2C_{\mathrm{prep}}}, \frac{\epsilon}{2C'_{\mathrm{prep}}}\right\}, \\ \delta_0 &:= \min\left\{1, \frac{1}{2(U+1)}, \frac{1}{2U+U^2}, \frac{\varepsilon_{\mathrm{prep}}}{C_1}\right\} \end{aligned}$$

make the last expression at most $3\epsilon/4 < \epsilon$. For the geometric choice, evaluate the preparation constants at $q_0 = 1/2$. This gives the claimed entry and closes the finite-horizon bootstrap. □

The compact initial set excludes a missing top component and the boundary at which that component can be eliminated in one update of the local model. The number of warm-up steps and required tube width may increase as the target accuracy decreases. The horizon is uniform over the compact initial set for the fixed normal-sharpness family; the learning-rate and geometric conditions are checked using this horizon. The proof does not give an explicit warm-up rate.

Iterations needed to reach the endpoint. The local model has normalized physical progress $2/\theta_1(t)$, which corresponds to one unit of reference time per update because $v(t) = 2\eta/\theta_1(t)$. The next result gives explicit dependence on the spectral contraction margin. Theorem F.7 supplies its position entry by taking $\epsilon \le \sigma/9$ and reindexing at k_{prep}, provided the passage hypotheses also hold at that new starting point. The preparation steps remain separate from the count below.

Theorem F.8 (NGD tracking of the curvature-adapted flow)**.** *Under Assumption F.1, suppose that the top normal eigenvalue is simple on I and $\theta_{1,-} > 2$. Write $\Delta_{\mathrm{sp}} := \inf_I(\theta_1 - \theta_2)$ when $d > 2$. Define the spectral margin*

$$\sigma := \min\left\{\frac{4}{\theta_+^2}, \quad 1 - \frac{4}{\theta_{1,-}^2}, \quad \frac{\theta_-}{\theta_+}, \quad \frac{\Delta_{\mathrm{sp}}}{\theta_+}\right\} > 0, \tag{92}$$

with the last two terms omitted when $d = 2$. Suppose that

$$\frac{3}{2}\eta < \rho, \qquad [0, T_*]^{+2\eta} \subset \operatorname{int} I, \qquad \delta \le \frac{\sigma^3}{36}. \tag{93}$$

Fix $C_{\mathrm{ent}} \ge 0$ and run `NormGD` *from $x_0 = \Psi(t_0, u_0)$ with $0 \le \ell(0, t_0) \le C_{\mathrm{ent}}\eta$ and $t_0 < T_*$. If*

$$d_0 := \min_{\chi=\pm 1} \|u_0 - \chi r_0(t_0) v_1(t_0)\|_2 \le \frac{\sigma}{9},$$

then the first crossing $N = n_{\mathrm{NGD}} := \inf\{k \ge 0 : t_k \ge T_*\}$ is finite and all updates through it are defined in $\mathcal{X}$. For every $0 \le n \le N$,

$$\|x_n - z(t_n)\| < \frac{\eta}{2},$$
$$|t_n - t_0 - n| \le \frac{4d_0}{\sigma} + \frac{20}{\sigma^2}\delta n. \tag{94}$$

For every such n with $t_0 + n \in I$, the error at the corresponding flow time satisfies

$$\|x_n - z(t_0 + n)\| \le \eta + \frac{40\delta n\eta}{\theta_{1,-}\sigma^2}. \tag{95}$$

The count from this supplied entry satisfies

$$|n_{\mathrm{NGD}} - T_*| \le \frac{40}{\sigma^2}\delta T_* + \left[\frac{8}{9} + \frac{5\theta_+}{4} + \frac{\theta_+ C_{\mathrm{ent}}}{2}\right]. \tag{96}$$

These bounds are uniform in the learning rate and section length whenever the displayed dimensionless margins and geometric conditions hold.

Proof of Theorem F.8. The first two terms defining σ give $0 < \sigma \le 1/2$ and $\theta_+ \le 2/\sqrt{\sigma}$. For the reference positions $U_\chi(t) := \chi r_0(t) v_1(t)$, set $R = 2\sigma/9 \le 1/9$. At fixed t, write $\theta = \theta_1(t)$ and $F_t(u) = u - f(\Theta(t)u)$, where $f(y) = y/\sqrt{1 + \|y\|^2}$. The Jacobian at either reference position is symmetric, with eigenvalues

$$1 - 8/\theta^2, \qquad 1 - 2\theta_i(t)/\theta \quad (2 \le i \le d-1).$$

Consequently $\|DF_t(U_\chi)\|_{\mathrm{op}} \le 1 - 2\sigma$; the transverse terms are absent when $d = 2$.

A normalization calculation bounds the nonlinear remainder without a factor θ_+^2. For $n(y) = y/\|y\|$, a unit vector w, and $b = \langle y/\|y\|, w\rangle$, differentiation gives

$$D^2 n(y)[w, w] = \frac{-2bw + (3b^2 - 1)y/\|y\|}{\|y\|^2},$$
$$\|D^2 n(y)[w, w]\|^2 = \frac{(1 - b^2)(1 + 3b^2)}{\|y\|^4} \le \frac{4}{3\|y\|^4}.$$

The same bound controls the bilinear operator norm: dualize the output against a unit vector and use the equality of the diagonal and bilinear norms for a real symmetric scalar form. Since f is the projection of $n((1, y))$ onto the coordinates of y, it follows that

$$\|D^2 f(y)\|_{\mathrm{op}} \le \frac{2}{\sqrt{3}(1 + \|y\|^2)},$$
$$\sqrt{1 + \|\Theta(t)u\|^2} \ge \frac{\theta}{2} - \theta R \qquad \text{if } \|u - U_\chi(t)\| \le R,$$
$$\|D^2 F_t(u)\|_{\mathrm{op}} \le \frac{2}{\sqrt{3}(1/2 - R)^2} \le \frac{648}{49\sqrt{3}} < 9,$$
$$\|F_t(U_\chi + w) - U_{-\chi}\| \le (1 - 2\sigma)\|w\| + \frac{9}{2}\|w\|^2 \le (1 - \sigma)\|w\| \qquad (\|w\| \le R).$$

This is contraction of the distance to the reference position, which is the estimate used below. The same neighborhood lies in the fixed normal-radius domain $\|u\| < 1/2$, because

$$\frac{1}{2} - r_0(t) = \frac{1}{\theta^2[1/2 + r_0(t)]} > \frac{1}{\theta^2}, \qquad R \le \frac{8}{9\theta_+^2} \le \frac{8}{9\theta^2}.$$

For the frozen phase speed $a_0(t,u) = (1 + \|\Theta(t)u\|^2)^{-1/2}$, the derivative bound on this neighborhood is

$$\frac{\theta}{2}\|\nabla_u a_0(t,u)\| \le \frac{\theta^2}{2[1+\|\Theta(t)u\|^2]} \le \frac{1}{2(1/2-R)^2} \le \frac{162}{49} < 4. \tag{97}$$

Use Equations (72) and (73) with normal radius $U = 1/2$. The tube and endpoint conditions in Equation (93) supply each complete step from a preterminal state in this domain. Indeed, $\delta \le \sigma^3/36 \le 1/288$ implies the chart margin $1 - 3\delta/2 > 1/2$ and the raw gradient-error bound $(1/2 + 1/8)\delta < 1/2$. The separate defects and coefficient variation therefore satisfy

$$\|\xi_k^u\| \le \frac{9}{4}\delta, \qquad |\zeta_k| \le \frac{13}{2}\delta, \qquad \eta \max\{L_\Theta, L_{\theta_1}\} \le (\theta_+ + 2)\delta. \tag{98}$$

Put $r_- := \sqrt{1/4 - \theta_{1,-}^{-2}}$ and $r_+ := \sqrt{1/4 - \theta_+^{-2}}$. Differentiation of the simple top eigenpair and its amplitude gives

$$\frac{\eta}{v(t)}\|\frac{\mathrm{d}U_\chi(t)}{\mathrm{d}t}\| \le (\theta_+ + 2)\left(\frac{1}{r_-\theta_{1,-}^3} + \frac{r_+}{\Delta_{\mathrm{sp}}}\right)\delta \le \frac{2\delta}{\sigma}.$$

Omit the term involving Δ_{sp} when $d = 2$. For the last inequality, use $r_- \ge \sqrt{\sigma}/2$, $\theta_{1,-} > 2$, $r_+ < 1/2$, $\Delta_{\mathrm{sp}} \ge \theta_+\sigma$, and $\theta_+ \le 2/\sqrt{\sigma}$ to bound each of the two coefficients by $1/\sigma$. The small-variation hypothesis also gives

$$13\theta_+\delta \le \frac{13}{18}\sigma^{5/2} < 1,$$
$$\frac{9 + 10/\sigma}{4\sigma}\delta \le \frac{9\sigma^2 + 10\sigma}{144} \le \frac{29\sigma}{288} < \frac{\sigma}{9}.$$

Choose the sign χ attaining d_0 and track its alternating sides. As long as the current distance $d_k := \|u_k - U_{(-1)^k\chi}(t_k)\|$ is at most R, Equations (97) and (98) give

$$\frac{\ell(t_k,t_{k+1})}{\eta} \ge \frac{2 - 8R}{\theta_1(t_k)} - \frac{13}{2}\delta \ge \frac{1}{2\theta_+} > 0,$$
$$\frac{\ell(t_k,t_{k+1})}{\eta} \le 1 + \frac{13}{2}\delta \le \frac{5}{4},$$
$$d_{k+1} \le (1-\sigma)d_k + \frac{9}{4}\delta + \frac{5}{2\sigma}\delta = (1-\sigma)d_k + \frac{9 + 10/\sigma}{4}\delta,$$
$$d_k \le (1-\sigma)^k d_0 + \frac{9 + 10/\sigma}{4\sigma}\delta[1 - (1-\sigma)^k] \le \frac{\sigma}{9} < R.$$

This closes the induction, including the crossing update. It gives $\|u_k\| < 1/2$ and finite crossing with $N \le \lceil 2\theta_+\ell(t_0, T_*)/\eta \rceil$; the endpoint overshoot obeys $0 \le \ell(T_*, t_N) \le 5\eta/4$.

For every complete step, the reference identity $\theta_1(t_k)a_0(t_k, U_{(-1)^k\chi})/2 = 1$ and a one-step quadrature estimate give

$$t_{k+1} - t_k = \frac{1}{2\eta}\int_{t_k}^{t_{k+1}} \theta_1(q)v(q)\,\mathrm{d}q,$$
$$\left|t_{k+1} - t_k - \frac{\theta_1(t_k)}{2\eta}\ell(t_k,t_{k+1})\right| \le \frac{L_{\theta_1}}{4\eta}\ell(t_k,t_{k+1})^2 \le \frac{25}{64}(\theta_+ + 2)\delta,$$
$$|t_{k+1} - t_k - 1| \le 4d_k + \frac{\theta_+}{2}\frac{13}{2}\delta + \frac{25}{64}(\theta_+ + 2)\delta,$$
$$\sum_{k=0}^{n-1} d_k \le \frac{d_0}{\sigma} + \frac{9 + 10/\sigma}{4\sigma}\delta n,$$
$$|t_n - t_0 - n| \le \frac{4d_0}{\sigma} + \left[\frac{9 + 10/\sigma}{\sigma} + \frac{233\theta_+ + 50}{64}\right]\delta n$$
$$\le \frac{4d_0}{\sigma} + \frac{20}{\sigma^2}\delta n.$$

For the last simplification, $\theta_+ \le 2/\sqrt{\sigma}$ and $\sigma \le 1/2$ give

$$\sigma^2\left[\frac{9+10/\sigma}{\sigma}+\frac{233\theta_+ +50}{64}\right] \le 10+9\sigma+\frac{233}{32}\sigma^{3/2}+\frac{25}{32}\sigma^2 \le \frac{2347}{128} < 20.$$

This proves Equation (94). Since $\|\mathrm{d}z(t)/\mathrm{d}t\| \le 2\eta/\theta_{1,-}$ on I, every $0 \le n \le N$ with $t_0+n \in I$ satisfies

$$\begin{aligned}\|x_n - z(t_0+n)\| &\le \|x_n - z(t_n)\| + \frac{2\eta}{\theta_{1,-}}|t_n - t_0 - n| \\ &\le \frac{\eta}{2}+\frac{8d_0\eta}{\theta_{1,-}\sigma}+\frac{40\delta n\eta}{\theta_{1,-}\sigma^2} \\ &\le \eta+\frac{40\delta n\eta}{\theta_{1,-}\sigma^2},\end{aligned}$$

where the last inequality uses $d_0 \le \sigma/9$ and $\theta_{1,-} > 2$. This gives Equation (95).

Finally, put $B_G := 20\delta/\sigma^2 \le 5/18$ and $A_G := 4d_0/\sigma \le 4/9$. The phase overshoot, the initial offset, and the prefix-clock estimate imply

$$\begin{aligned}0 \le t_N - T_* \le \frac{5\theta_+}{8}, \qquad 0 \le t_0 \le \frac{\theta_+ C_{\mathrm{ent}}}{2}, \\ (1-B_G)N &\le T_* + (A_G + 5\theta_+/8), \\ |N - T_*| &\le A_G + B_G N + 5\theta_+/8 + t_0 \\ &\le \frac{B_G T_* + (A_G + 5\theta_+/8)}{1-B_G} + t_0 \\ &\le \frac{40}{\sigma^2}\delta T_* + \left[\frac{8}{9}+\frac{5\theta_+}{4}+\frac{\theta_+ C_{\mathrm{ent}}}{2}\right].\end{aligned}$$

This proves Equation (96), counting from the stated entry. □

From warm-up to flow tracking. The two results give an NGD trajectory that first enters a neighborhood of the reference two-cycle and then tracks the flow to the endpoint. Apply Theorem F.7 with $\epsilon = \sigma/9$ and retain all its hypotheses, where σ is defined in Equation (92). In addition, require $\delta \le \sigma^3/36$, $\rho > 3\eta/2$, and

$$[\inf I_{\mathrm{prep}}, T_*]^{+2\eta} \subset \operatorname{int} I, \qquad \ell(t_0, T_*) > 2N_{\mathrm{prep}}\eta.$$

Write $k_{\mathrm{p}} = k_{\mathrm{prep}} \le N_{\mathrm{prep}}$ and $t_{\mathrm{p}} = t_{k_{\mathrm{p}}}$. The preparation bound Equation (91), with $q_0 = 1/2$, gives $|\ell(t_0, t_j)| \le 2j\eta$ for $j \le k_{\mathrm{p}}$, so no warm-up iterate crosses the endpoint. Since $t_{\mathrm{p}} \in I_{\mathrm{prep}}$ and $t_{\mathrm{p}} < T_*$, the interval $[t_{\mathrm{p}}, T_*]$ has the required passage margin. Reindex at k_{p} and translate the reference time by t_{p}. Then Theorem F.8 applies with $C_{\mathrm{ent}} = 0$ and endpoint time $T_* - t_{\mathrm{p}}$, giving

$$\begin{aligned}\|x_{k_{\mathrm{p}}+n} - z(t_{\mathrm{p}}+n)\| &\le \eta + \frac{40\delta n\eta}{\theta_{1,-}\sigma^2}, \\ \left|N_{\mathrm{tot}} - k_{\mathrm{p}} - (T_* - t_{\mathrm{p}})\right| &\le \frac{40\delta}{\sigma^2}(T_* - t_{\mathrm{p}}) + \frac{8}{9} + \frac{5\theta_+}{4}.\end{aligned}$$

Here N_{tot} is the first crossing counted from the original initialization, and the tracking bound holds for $0 \le n \le N_{\mathrm{tot}} - k_{\mathrm{p}}$ whenever $t_{\mathrm{p}} + n \in I$. Moreover, $v^{-1} \le \theta_+/(2\eta)$ gives $|t_{\mathrm{p}} - t_0| \le \theta_+ N_{\mathrm{prep}}$. Thus the count relative to the original reference time has the explicit startup correction

$$\left|N_{\mathrm{tot}} - (T_* - t_0)\right| \le \frac{40\delta}{\sigma^2}(T_* - t_{\mathrm{p}}) + \frac{8}{9} + \frac{5\theta_+}{4} + (1+\theta_+)N_{\mathrm{prep}}.$$

The warm-up horizon retains its dependence on the prescribed compact initial set; this combination does not supply a uniform rate as the initial top component tends to zero.

F.3 Proof for branch-free NGM

The progress bounds in this section do not assume periodic behavior and allow bounded normal coordinates at initialization with zero momentum, followed by one entry update. Theorem F.10 treats both a supplied joint state with $0 < h < 2$ and zero momentum, without requiring proximity to a cycle in normal coordinates. We first bound the normal coordinates, then estimate progress using the normal metric at each actual step. The resulting clock bound has a fixed endpoint deficit and an error linear in geometric variation, with separate initialization clauses.

For the closure and speed results, use Assumption F.1 and fix $\beta \in (0, 1)$. Let I_{BF} denote the compact source interval specified in each result.

For a fixed defect budget $\varepsilon_{\mathrm{F}} \geq 0$, define admissible transitions by the perturbed recurrence

$$\begin{aligned} \ell(t, t^+) &= \eta\big[a_\beta(t, Z) + \zeta\big], \\ Z^+ &= F_{\beta,t}(Z) + \xi, \\ |\zeta| + \|\xi\| &\leq \varepsilon_{\mathrm{F}}, \end{aligned} \tag{99}$$

where ζ is scalar and ξ is in the full momentum state space. The next lemma concerns sequences satisfying this controlled recurrence. The passage theorem verifies the recurrence uniformly for complete NGM updates on a containing domain with a positive predicted-buffer norm.

Robust closure. Within proofs containing no plain trajectory, we suppress the superscript β on Fermi coordinates.

Lemma F.9 (Containing domain and continuation for controlled transitions)**.** *For the controlled recurrence Equation* (99) *and every nonempty compact entry set $\mathcal{K}_{\mathrm{in}} \subset \mathbb{R}^{d-1} \times (0, 2) \times \mathbb{R}^{d-1}$, there are a compact domain $\mathcal{K} \subset \mathbb{R}^{d-1} \times (0, 2) \times \mathbb{R}^{d-1}$ with $\mathcal{K}_{\mathrm{in}} \subset \operatorname{int} \mathcal{K}$ and constants $C \geq 1$, $\varepsilon_{\mathrm{cl}}, a_{\mathrm{cl}} > 0$, depending only on β, $\mathcal{K}_{\mathrm{in}}$, and θ_-, θ_+, with the following property. Every $Z = (u, h, p) \in \mathcal{K}$ satisfies $\|u\| \leq C$, and the predicted local buffer satisfies $\inf_{t \in I,\, Z \in \mathcal{K}} \widehat{r}(t, Z) > 0$.*

Suppose that $[0, T_*] \subset I_{\mathrm{BF}} \subset \operatorname{int} I$, $\max\{\varepsilon_{\mathrm{F}}, \eta L_\Theta\} \leq \varepsilon_{\mathrm{cl}}$, and $t_0^\beta < T_*$. Require $[t_0^\beta, t_+] \subset I_{\mathrm{BF}}$, where $\ell(T_*, t_+) = \eta(1 + \varepsilon_{\mathrm{F}})$. Every sequence of transitions Equation (99) *starting from $Z_0^\beta \in \mathcal{K}_{\mathrm{in}}$ remains in $I_{\mathrm{BF}} \times \operatorname{int} \mathcal{K}$ and satisfies $\|u_k^\beta\| \leq C$ up to and including its first crossing of T_*. At every preterminal state, $\widehat{r} \geq a_{\mathrm{cl}}$ and $\ell(t, t^+) \geq \eta a_{\mathrm{cl}}$. Every such sequence continued while $t < T_*$ crosses T_* after finitely many updates.*

Proof of Lemma F.9. By Assumption F.1(i), Θ is a continuous positive-definite family on compact I, so $\theta_- > 0$. Set $H := (1 + \max_{(u,h,p) \in \mathcal{K}_{\mathrm{in}}} |h - 1|)/2$. Compactness of $\mathcal{K}_{\mathrm{in}}$ and its restriction $0 < h < 2$ give $1/2 \leq H < 1$. Choose

$$R := \max\left\{1 + \frac{4}{\theta_-} \sup_{(u,h,p) \in \mathcal{K}_{\mathrm{in}}} [\theta_+ \|u\| + \|p\|], \quad \frac{8(\theta_+ + 1)}{(1 - \beta)\theta_-}\right\}.$$

Define the containing domain by

$$\mathcal{K} := \left\{(u, h, p) : \|u\| \leq 2R, \quad |h - 1| \leq H, \quad \|p\| \leq 2\theta_+ R\right\}.$$

This choice gives $\mathcal{K}_{\mathrm{in}} \subset \operatorname{int} \mathcal{K}$ and fixes the domain before the defect tolerance is chosen. Set $C := 2R \geq 1$. On the entire domain, the predicted tangential buffer satisfies $\widehat{h} \geq 1 - \beta H > 0$, so the local map and physical progress are defined. Set

$$a_{\mathrm{cl}} := \frac{1 - \beta H}{2\sqrt{(1 - \beta H)^2 + (5\theta_+ R/4)^2}}.$$

For the moving-coordinate error, use

$$e_w := \theta_+ \varepsilon_{\mathrm{F}} + \eta L_\Theta (1 + \varepsilon_{\mathrm{F}})(2R + 1).$$

Choose the explicit tolerance

$$\varepsilon_{\mathrm{cl}} := \min\left\{\frac{1}{4}, \frac{(1 - \beta)H}{2}, a_{\mathrm{cl}}, \frac{\theta_-}{12(\theta_+ + 1)}, \frac{1}{\theta_+ + 4R + 3}\right\}.$$

Whenever $\max\{\varepsilon_{\mathrm{F}}, \eta L_\Theta\} \le \varepsilon_{\mathrm{cl}}$, one has

$$\begin{aligned} e_w + \varepsilon_{\mathrm{F}} &\le 1, \\ \varepsilon_{\mathrm{F}} &\le \frac{\theta_-}{12(\theta_+ + 1)} < \frac{1}{12}, \\ \varepsilon_{\mathrm{F}} &< (1-\beta)H, \\ \varepsilon_{\mathrm{F}} &\le a_{\mathrm{cl}}. \end{aligned} \tag{100}$$

These choices depend only on the stated data. Since $R \ge 8$ and $\theta_- \le \theta_+$, the chosen a_{cl} is no larger than $1 - \beta H$. The scalar recurrence gives, throughout every admissible prefix,

$$|h_k - 1| \le \beta^k (2H - 1) + \frac{1-\beta^k}{1-\beta}\varepsilon_{\mathrm{F}} < H.$$

Write $\Theta_k := \Theta(t_k)$, $w_k := \Theta_k u_k$, and $d_k := w_k - p_k$. We propagate the two bounds

$$\|u_k\| \le R, \tag{101}$$

$$\|d_k\| \le \frac{\theta_- R}{4}. \tag{102}$$

They hold initially by the choice of R. They imply $\|p_k\| \le 5\theta_+ R/4$, since $p_k = \Theta_k u_k - d_k$ and $\theta_- \le \theta_+$. Together with the tangential bound, these estimates place the state strictly inside $\mathcal{K}$. At a preterminal state, $\widehat{p}_k = \Theta_k u_k - \beta d_k$ gives $\|\widehat{p}_k\| \le 5\theta_+ R/4$, while $\widehat{h}_k \ge 1 - \beta H$. Consequently, $\widehat{r}_k \ge 1 - \beta H$ and

$$\begin{aligned} a_\beta(t_k, Z_k) &= \frac{\widehat{h}_k}{\sqrt{\widehat{h}_k^2 + \|\widehat{p}_k\|^2}} \ge 2a_{\mathrm{cl}}, \\ \eta a_{\mathrm{cl}} &\le \ell(t_k, t_{k+1}) \le \eta(1 + \varepsilon_{\mathrm{F}}). \end{aligned}$$

Thus the next projected time lies in the forward buffer before its coefficients are used. Set $\phi_k := \widehat{p}_k/\widehat{r}_k$. The robust recurrence gives $u_{k+1} = u_k - \phi_k + \xi_k^u$ and $p_{k+1} = \widehat{p}_k + \xi_k^p$. Multiplying the first identity by Θ_{k+1} yields

$$\begin{aligned} w_{k+1} &= w_k - \Theta_k \phi_k + e_k^w, \\ e_k^w &:= (\Theta_{k+1} - \Theta_k)(u_k - \phi_k) + \Theta_{k+1}\xi_k^u, \\ \|e_k^w\| &\le \eta L_\Theta (1 + \varepsilon_{\mathrm{F}})(2R + 1) + \theta_+ \varepsilon_{\mathrm{F}} = e_w. \end{aligned}$$

The lag recurrence and Equation (100) give

$$\begin{aligned} d_{k+1} &= \beta d_k - \Theta_k \phi_k + e_k^w - \xi_k^p, \\ \|d_{k+1}\| &\le \frac{\beta\theta_- R}{4} + \theta_+ + e_w + \varepsilon_{\mathrm{F}} \\ &\le \frac{\beta\theta_- R}{4} + \theta_+ + 1 \le \frac{\theta_- R}{4}. \end{aligned}$$

The last inequality follows from the choice of R. This propagates Equation (102).

For the normal bound, work directly with $r := \|u_k\|$ and first suppose that $r \ge R/2$. Then

$$\begin{aligned} u_k^\top \widehat{p}_k &\ge \theta_- r^2 - \beta r \|d_k\| \\ &\ge \theta_- r^2 - \frac{\theta_- R r}{4} \ge \frac{\theta_- r^2}{2}. \end{aligned}$$

Since $R \ge 8$, $|\widehat{h}_k| < 2$, and $\|\widehat{p}_k\| \le \theta_+ r + \theta_- R/4$, one has

$$\widehat{r}_k \le 2 + \theta_+ r + \theta_- R/4 \le 2 + (\theta_+ + \theta_-/2) r \le 2(\theta_+ + 1) r.$$

Expanding $u_{k+1} = u_k - \widehat{p}_k/\widehat{r}_k + \xi_k^u$ now gives

$$\begin{aligned} \|u_{k+1}\|^2 - r^2 &\le -2\frac{u_k^\top \widehat{p}_k}{\widehat{r}_k} + 1 + 2\varepsilon_{\mathrm{F}}(r+1) + \varepsilon_{\mathrm{F}}^2 \\ &\le -\left(\frac{\theta_-}{2(\theta_+ + 1)} - 2\varepsilon_{\mathrm{F}}\right) r + (1 + \varepsilon_{\mathrm{F}})^2 \\ &\le -\frac{\theta_- R}{6(\theta_+ + 1)} + \frac{169}{144} \le -\frac{23}{144} < 0. \end{aligned}$$

The last two inequalities use $\varepsilon_{\mathrm{F}} \le \theta_-/[12(\theta_+ + 1)] < 1/12$ and the choice of R. When $r < R/2$, the one-step bound instead gives

$$\|u_{k+1}\| \le r + 1 + \varepsilon_{\mathrm{F}} < R/2 + 13/12 < R.$$

The two cases propagate Equation (101). The endpoint therefore satisfies $\|u_{k+1}\| \le R$, $\|p_{k+1}\| \le 5\theta_+ R/4$, and $|h_{k+1} - 1| < H$, so it remains in $\operatorname{int} \mathcal{K}$, including at the crossing update. Since every preterminal physical increment is at least $\eta a_{\mathrm{cl}} > 0$, continued updates reach T_* in finitely many steps. □

Progress and iterations needed to reach the endpoint. The next proof uses the actual normal metric at every step. Its elementary gain is explicit in the normal spectrum and momentum parameter; the error depends on one-step variation rather than on the length of a frozen comparison block.

Theorem F.10 (Branch-free NGM progress and iteration bounds)**.** *Under Assumption F.1, suppose that $\theta_{1,-} > 2$ and fix $\beta \in (0,1)$. Define*

$$\mu := \min\left\{ \frac{\theta_{1,-}}{2} - 1, \frac{2\beta}{\theta_+ + 1 - \beta} \right\} > 0.$$

The following initialization clauses hold independently.

(a) Supplied joint state. *For every nonempty compact set $\mathcal{K}_{\mathrm{in}} \subset \mathbb{R}^{d-1} \times (0,2) \times \mathbb{R}^{d-1}$, there are $C \ge 1$ and $C_M, D, \delta_0 > 0$, depending only on β, $\theta_-, \theta_+, \theta_{1,-}$, and $\mathcal{K}_{\mathrm{in}}$, with the following property. Suppose that $C\eta < \rho$, $\ell(\inf I, 0) > 2\eta$, $\ell(T_*, \sup I) > 4\eta$, and $\delta \le \delta_0$. Run Equation* (47) *from a supplied $x_0^\beta \in \mathcal{X}$ with $t_0^\beta \in [0, T_*)$ and $Z_0^\beta \in \mathcal{K}_{\mathrm{in}}$, retaining its buffer. Every update through the first crossing of T_* is well defined, its index n_{BF} is finite, and*

$$\|x_j^\beta - z(t_j^\beta)\| \le C\eta, \qquad 0 \le j \le n_{\mathrm{BF}}.$$

Writing $v_{\mathrm{BF}} := 1 + 3\mu/4 - C_M\delta$, one has $v_{\mathrm{BF}} \ge 1 + 5\mu/8$ and

$$t_n^\beta - t_0^\beta \ge v_{\mathrm{BF}} n - D, \qquad 0 \le n \le n_{\mathrm{BF}}, \tag{103}$$

$$\frac{t_n^\beta - t_0^\beta}{n} \ge 1 + \frac{\mu}{2}, \qquad \frac{8D}{\mu} \le n \le n_{\mathrm{BF}}, \tag{104}$$

$$n_{\mathrm{BF}} \le \left\lceil \frac{T_* - t_0^\beta + D}{v_{\mathrm{BF}}} \right\rceil.$$

(b) Zero momentum. *For every $r > 0$, there are $C \ge 1$ and $C_M, D, \delta_0 > 0$, depending only on the same data with r replacing $\mathcal{K}_{\mathrm{in}}$. Suppose that $S > 2\eta$, $C\eta < \rho$, $\ell(\inf I, 0) > 2\eta$, $\ell(T_*, \sup I) > 4\eta$, and $\delta \le \delta_0$. Start from $x_0^\beta = \Psi(0, u_0^\beta)$ with $\|u_0^\beta\| \le r$ and $m_0^\beta = 0$. The first update gives $0 < t_1^\beta < T_*$ and a joint state to which part* (a) *applies with its carried buffer. Every update through crossing is well defined, and its index n_{BF}, counted from zero-momentum initialization, is finite. With $v_{\mathrm{BF}} = 1 + 3\mu/4 - C_M\delta \ge 1 + 5\mu/8$,*

$$\begin{aligned} &\|x_j^\beta - z(t_j^\beta)\| \le C\eta, \qquad 0 \le j \le n_{\mathrm{BF}}, \\ &t_{n+1}^\beta - t_1^\beta \ge v_{\mathrm{BF}} n - D, \qquad 0 \le n \le n_{\mathrm{BF}} - 1, \\ &\frac{t_{n+1}^\beta - t_1^\beta}{n} \ge 1 + \frac{\mu}{2}, \qquad \frac{8D}{\mu} \le n \le n_{\mathrm{BF}} - 1, \\ &n_{\mathrm{BF}} \le 1 + \left\lceil \frac{T_* - t_1^\beta + D}{v_{\mathrm{BF}}} \right\rceil. \end{aligned} \tag{105}$$

The proof constructs all constants explicitly.

Proof of Theorem F.10. Domain and constants. Use the explicit choices $R, H, \mathcal{K}, \varepsilon_{\mathrm{cl}}$ in the proof of Lemma F.9 for the prescribed entry set. Set

$$U = 2R, \qquad C = U + 1, \qquad P = \theta_+ U,$$
$$h_* = 1 - \beta H, \qquad \widehat{M} = \sqrt{(1+\beta H)^2 + P^2}, \qquad F_* = \sqrt{1+P^2}.$$

On the entire domain $\mathcal{K}$, one has $\|u\| \le U$, $\|p\| \le P$, $\widehat{h} \ge h_*$, and $\|\widehat{p}\| \le P$. Let c_u, c_ℓ, c_b be the explicit coefficients in Equation (72), evaluated with $R = U$, $r_* = h_*$, and the displayed $\widehat{M}$. For the moving metric, choose

$$\alpha = 1 - \frac{4}{\theta_{1,-}^2}, \qquad r_{\mathrm{tar}} = 2 + \frac{3\mu}{2}, \qquad A = \theta_+ + \frac{1}{3}, \qquad Z_* = \theta_+ U + \frac{1}{24},$$
$$\mathcal{E} = \frac{Z_*^2}{2\theta_-} + \frac{4\beta F_*}{1-\beta}, \qquad \mathcal{G} = \frac{16}{\alpha}\left(\frac{Z_*}{\theta_-} + \frac{3}{4}\right),$$
$$\mathcal{V} = \frac{32\mathcal{E}}{\alpha^2} + \frac{8Z_*^2}{\alpha\theta_-^2}. \tag{106}$$

The definition of μ gives $0 < \mu < 1$ and $2 < r_{\mathrm{tar}} < \theta_{1,-}$, so $\alpha > 0$. These coefficients give the following explicit sufficient speed and endpoint constants:

$$\begin{aligned} C_M &= \frac{\theta_+}{2}(c_\ell + \mathcal{G}c_u) + \frac{\theta_+\beta c_b}{1-\beta}(\mathcal{G}+1) \\ &\quad + (\theta_+ + 2)(\mathcal{G}U + \mathcal{V} + 1), \\ b &= \max\left\{0, \left\lceil \frac{\log(8P)}{\log(1/\beta)} \right\rceil - 1\right\}, \\ D &= \left(1 + \frac{3\mu}{4}\right)(b+1) + \frac{8\mathcal{E}}{\alpha} + \frac{\theta_+(\mathcal{G}+1)H\beta^{b+1}}{2(1-\beta)}. \end{aligned} \tag{107}$$

Choose the geometric threshold before generating the trajectory:

$$\begin{aligned} \delta_0 := \min\Bigg\{ &\frac{1}{2\max\{U+1, U+U^2/2\}}, \frac{\varepsilon_{\mathrm{cl}}}{c_\ell + c_u + c_b + \theta_+ + 2}, \\ &\frac{1}{12[\theta_+ c_u + 2U(\theta_+ + 2) + 2\beta c_b/(1-\beta)]}, \frac{\mu}{8C_M} \Bigg\} > 0. \end{aligned} \tag{108}$$

All quantities depend only on the stated data, and every denominator is positive.

Take $I_{\mathrm{BF}} = [0, t_+]$, where $\ell(T_*, t_+) = 2\eta$. The conditions $C\eta < \rho$, $\ell(\inf I, 0) > 2\eta$, $\ell(T_*, \sup I) > 4\eta$, and $\delta(U+1) \le 1/2$ verify the complete-ball condition of Lemma F.3 on $I_{\mathrm{BF}} \times \mathcal{K}$. The first term in Equation (108) also gives $\delta(U + U^2/2) \le 1/2$. Since $h_* > 1 - \beta$, this suffices for the predicted-buffer perturbation margin in that lemma. Its componentwise estimates and the normal-sharpness derivative bound give

$$|\zeta_k| \le c_\ell \delta, \qquad \|\xi_k^u\| \le c_u \delta, \qquad \sqrt{|\xi_k^h|^2 + \|\xi_k^p\|^2} \le c_b \delta,$$
$$\varepsilon_{\mathrm{F}} = (c_\ell + c_u + c_b)\delta \le \varepsilon_{\mathrm{cl}}, \qquad \eta L_\Theta \le (\theta_+ + 2)\delta \le \varepsilon_{\mathrm{cl}}. \tag{109}$$

Here $\|\xi\| \le \|\xi^u\| + \sqrt{|\xi^h|^2 + \|\xi^p\|^2}$ verifies the joint norm in Equation (99). At every preterminal state, the closure lemma places the state in $\operatorname{int}\mathcal{K}$, and the complete-ball condition supplies its next actual update. Induction therefore constructs the actual trajectory through its finite first crossing, with positive phase increments and

$$0 < \ell(t_k, t_{k+1}) \le 2\eta, \qquad \|x_k^\beta - z(t_k^\beta)\| = \eta\|u_k^\beta\| \le C\eta.$$

This proves the geometric assertion; the remainder of the proof bounds its clock.

Predicted buffers and the moving metric. Suppress the superscript β. Put $(\widetilde{h}_{k+1}, \widetilde{p}_{k+1}) = (\widehat{h}_k, \widehat{p}_k)$, $(\widetilde{h}_0, \widetilde{p}_0) = (h_0, p_0)$, and $\xi_{-1}^h = \xi_{-1}^p = 0$. With $\Theta_k = \Theta(t_k)$ and $F_k = \sqrt{\widetilde{h}_{k+1}^2 + \|\widetilde{p}_{k+1}\|^2}$, the

exact recurrences are

$$\begin{aligned}
u_{k+1} &= u_k - \widetilde{p}_{k+1}/F_k + \xi^u_k,\\
\widetilde{p}_{k+1} &= \beta\widetilde{p}_k + (1-\beta)\Theta_k u_k + \beta\xi^p_{k-1},\\
\widetilde{h}_{k+1} &= \beta\widetilde{h}_k + 1 - \beta + \beta\xi^h_{k-1},\\
w_k &:= \frac{\widetilde{p}_{k+1} - \beta\widetilde{p}_k}{1-\beta} = \Theta_k u_k + \frac{\beta}{1-\beta}\xi^p_{k-1},\\
w_{k+1} - w_k &= -\Theta_k\frac{\widetilde{p}_{k+1}}{F_k} + e_k,\\
e_k &= \Theta_k\xi^u_k + (\Theta_{k+1}-\Theta_k)u_{k+1} + \frac{\beta}{1-\beta}(\xi^p_k - \xi^p_{k-1}).
\end{aligned}$$

The third term of Equation (108) gives

$$\begin{aligned}
\|e_k\| &\le \left[\theta_+ c_u + 2U(\theta_+ + 2) + \frac{2\beta c_b}{1-\beta}\right]\delta \le \frac{1}{12},\\
\|w_k\| &\le \theta_+ U + \frac{\beta c_b \delta}{1-\beta} \le Z_*,\\
2P\beta^{b+1} &\le \frac{1}{4}.
\end{aligned}$$

Consequently the second-order buffer recurrence yields

$$\begin{aligned}
\|\widetilde{p}_{k+2} - \widetilde{p}_{k+1}\| &\le \beta\|\widetilde{p}_{k+1} - \widetilde{p}_k\| + (1-\beta)(\theta_+ + 1/12)\\
&\le 2P\beta^{k+1} + \theta_+ + 1/12 \le A, \qquad k \ge b.
\end{aligned} \tag{110}$$

For $h > 0$, the function $h/\sqrt{h^2 + \|p\|^2}$ is increasing and concave, and its secants from $h = 1$ have magnitude at most $1/\sqrt{1+\|p\|^2}$. The case $p = 0$ has zero difference. Thus, with $\widehat{F}_k = \sqrt{1 + \|\widetilde{p}_{k+1}\|^2}$ and $\phi_k = \widetilde{p}_{k+1}/\widehat{F}_k$,

$$\begin{aligned}
\left|\frac{\widetilde{h}_{k+1}}{F_k} - \frac{1}{\widehat{F}_k}\right| &\le |\widetilde{h}_{k+1} - 1|,\\
\|\widetilde{p}_{k+1}\|\left|F_k^{-1} - \widehat{F}_k^{-1}\right| &\le |\widetilde{h}_{k+1} - 1|,\\
w_{k+1} - w_k &= -\Theta_k\phi_k + g_k, \qquad g_k = e_k + \Theta_k\widetilde{p}_{k+1}(\widehat{F}_k^{-1} - F_k^{-1}).
\end{aligned} \tag{111}$$

For the second inequality, use $|F_k - \widehat{F}_k| \le |\widetilde{h}_{k+1} - 1|$, $\|\widetilde{p}_{k+1}\|/F_k \le 1$, and $\widehat{F}_k \ge 1$. In particular no inverse tangential floor is needed in these storage errors.

Fix $n \ge b + 2$ complete updates and sum below over $b \le k \le n-2$. This leaves $\widetilde{p}_n$, already used by the final actual update, available for the endpoint terms. Set $\theta_k = \lambda_{\max}(\Theta_k)$ and

$$\omega(\theta) = \frac{r_{\mathrm{tar}}^2\theta}{\theta^2 - r_{\mathrm{tar}}^2}, \qquad \omega_k = \omega(\theta_k), \qquad c = \frac{A + 1 - \beta}{2(A + 1 + \beta)}.$$

Since $\mu \le 2\beta/(\theta_+ + 1 - \beta)$, the fixed choice $A = \theta_+ + 1/3$ satisfies

$$\frac{1}{c} - r_{\mathrm{tar}} = \frac{4\beta}{\theta_+ + 1 - \beta + 1/3} - \frac{3\mu}{2} \ge \frac{3\mu}{14} > 0.$$

Also $r_{\mathrm{tar}} - 2 \le 3(\theta_{1,-} - 2)/4$ and $r_{\mathrm{tar}} < 7/2$. Therefore, for every $\theta \ge \theta_{1,-}$,

$$\begin{aligned}
\theta^2 - r_{\mathrm{tar}}^2 &\ge \frac{\theta^2 - 4}{4}, & 0 < \omega(\theta) &\le \frac{16}{\alpha},\\
|\omega'(\theta)| &= \frac{r_{\mathrm{tar}}^2(\theta^2 + r_{\mathrm{tar}}^2)}{(\theta^2 - r_{\mathrm{tar}}^2)^2} \le \frac{32}{\alpha^2}.
\end{aligned} \tag{112}$$

Write $V_\Theta = \sum_{k=b}^{n-2} \|\Theta_{k+1} - \Theta_k\|_{\mathrm{op}}$ and $V_\omega = \sum_{k=b+1}^{n-2} |\omega_k - \omega_{k-1}|$. Eigenvalue perturbation and Equation (112) give $V_\omega \le 32 V_\Theta / \alpha^2$.

The two weighted storage identities. Define $\widehat{F}_{-1} = \sqrt{1 + \|p_0\|^2}$ and

$$\begin{aligned} B_k &= \widehat{F}_{k-1} - \widehat{F}_k - \phi_k^\top(\widetilde{p}_k - \widetilde{p}_{k+1}), && 0 \le B_k \le 2F_*, \\ X_k &= \phi_k^\top \Theta_k \phi_k, && X_k \le \theta_k(1 - \widehat{F}_k^{-2}), \\ \mathcal{H}_k &= w_k^\top \Theta_k^{-1} w_k + \frac{2\beta}{1-\beta}\widehat{F}_{k-1}, && 0 \le \mathcal{H}_k \le H_* := \frac{Z_*^2}{\theta_-} + \frac{2\beta F_*}{1-\beta}. \end{aligned}$$

Convexity of $p \mapsto \sqrt{1 + \|p\|^2}$ gives the bounds on B_k. The inverse-matrix identity and the recurrence give

$$\begin{aligned} \Theta_{k+1}^{-1} - \Theta_k^{-1} &= \Theta_{k+1}^{-1}(\Theta_k - \Theta_{k+1})\Theta_k^{-1}, \\ w_{k+1}^\top \Theta_{k+1}^{-1} w_{k+1} - w_k^\top \Theta_k^{-1} w_k &= -2 w_k^\top \phi_k + X_k + R_k, \\ R_k &= (w_{k+1} + w_k)^\top \Theta_k^{-1} g_k - \phi_k^\top g_k \\ &\qquad + w_{k+1}^\top (\Theta_{k+1}^{-1} - \Theta_k^{-1}) w_{k+1}, \\ |R_k| &\le \left(\frac{2Z_*}{\theta_-} + 1\right)\|g_k\| + \frac{Z_*^2}{\theta_-^2}\|\Theta_{k+1} - \Theta_k\|_{\mathrm{op}}, \\ w_k^\top \phi_k &= \widehat{F}_k - \widehat{F}_k^{-1} + \frac{\beta}{1-\beta}(\widehat{F}_k - \widehat{F}_{k-1} + B_k), \\ \mathcal{H}_{k+1} - \mathcal{H}_k &= -2(\widehat{F}_k - \widehat{F}_k^{-1}) + X_k - \frac{2\beta}{1-\beta} B_k + R_k. \end{aligned}$$

For any scalar sequence q_k, the required summation by parts is

$$\sum_{k=b}^{n-2} \omega_k (q_{k+1} - q_k) = \omega_{n-2} q_{n-1} - \omega_b q_b + \sum_{k=b+1}^{n-2} (\omega_{k-1} - \omega_k) q_k. \tag{113}$$

Applying it to $\mathcal{H}_k$ yields

$$\begin{aligned} \sum_{k=b}^{n-2} \omega_k \left(\widehat{F}_k - \widehat{F}_k^{-1} - \frac{X_k}{2} + \frac{\beta B_k}{1-\beta}\right) &\le \frac{H_*}{2}\left(\frac{16}{\alpha} + V_\omega\right) \\ &\quad + \frac{8}{\alpha}\left[\left(\frac{2Z_*}{\theta_-} + 1\right)\sum_{k=b}^{n-2}\|g_k\| + \frac{Z_*^2}{\theta_-^2} V_\Theta\right]. \end{aligned}$$

For the second identity put $\nu_k = \widehat{F}_k - \beta \widehat{F}_{k-1}$ and $\varrho_k = \widehat{F}_{k+1}/\widehat{F}_k$. Then $|\nu_k| \le F_*$ and, by Equation (110), $\varrho_k \le 1 + A$. The Bregman identity and the second-order buffer recurrence give

$$\begin{aligned} \nu_{k+1} - \nu_k &= \phi_k^\top [\widetilde{p}_{k+2} - (1+\beta)\widetilde{p}_{k+1} + \beta \widetilde{p}_k] + \varrho_k B_{k+1} + \beta B_k \\ &= -(1-\beta) X_k + \varrho_k B_{k+1} + \beta B_k + (1-\beta) g_k^\top \phi_k. \end{aligned}$$

Using Equation (113) for ν_k and B_k gives

$$\begin{aligned} &\left|\sum_{k=b}^{n-2} \omega_k(\nu_{k+1} - \nu_k)\right| \le F_*\left(\frac{32}{\alpha} + V_\omega\right), \\ &\sum_{k=b}^{n-2} \omega_k (B_{k+1} - B_k) \le 2F_*\left(\frac{16}{\alpha} + V_\omega\right), \\ &(1-\beta)\sum_{k=b}^{n-2} \omega_k X_k \le (1 + A + \beta)\sum_{k=b}^{n-2} \omega_k B_k + F_*\left(\frac{32}{\alpha} + V_\omega\right) \\ &\qquad + 2(1+A)F_*\left(\frac{16}{\alpha} + V_\omega\right) + \frac{16(1-\beta)}{\alpha}\sum_{k=b}^{n-2}\|g_k\|. \end{aligned}$$

Substitute this last inequality into the weighted energy bound to eliminate B_k. The resulting coefficient of X_k is $1/2 - \beta/(1 + A + \beta) = c$. The coefficients of 1, $\sum\|g_k\|$, and V_Θ are bounded, respectively, by $16\mathcal{E}/\alpha$, $\mathcal{G}$, and $\mathcal{V}$:

$$\sum_{k=b}^{n-2} \omega_k \left[\widehat{F}_k - \widehat{F}_k^{-1} - c\theta_k(1 - \widehat{F}_k^{-2})\right] \le \frac{16\mathcal{E}}{\alpha} + \mathcal{G}\sum_{k=b}^{n-2}\|g_k\| + \mathcal{V}V_\Theta. \tag{114}$$

For completeness, these coefficient bounds use

$$\frac{1}{2} + \frac{\beta}{1 + A + \beta} \le \frac{3}{4},$$
$$1 + \frac{2(A+2)}{1 + A + \beta} \le 4, \qquad 1 + \frac{3 + 2A}{1 + A + \beta} \le 4,$$

and $X_k \le \theta_k(1 - \widehat{F}_k^{-2})$.

The clock follows from an exact supporting square. For any $F \ge 1$ and $\chi = \theta/r_{\mathrm{tar}} > 1$,

$$(F - F^{-1}) - \frac{\theta}{r_{\mathrm{tar}}}(1 - F^{-2}) - \left(\frac{\theta}{r_{\mathrm{tar}}^2} - \frac{1}{\theta}\right)(r_{\mathrm{tar}} - \theta/F)$$
$$= \frac{(F - \chi)^2(\chi F + 1)}{\chi F^2} \ge 0.$$

Since $c < 1/r_{\mathrm{tar}}$, multiplying by $\omega(\theta)$ and using Equation (114) gives

$$\sum_{k=b}^{n-2} \frac{\theta_k}{\widehat{F}_k} \ge r_{\mathrm{tar}}(n - b - 1) - \frac{16\mathcal{E}}{\alpha} - \mathcal{G}\sum_{k=b}^{n-2}\|g_k\| - \mathcal{V}V_\Theta. \tag{115}$$

Actual clock and endpoint. The passive recurrence retains the influence of every preceding defect:

$$\sum_{k=b}^{n-2} |\widetilde{h}_{k+1} - 1| \le \frac{\beta^{b+1}H}{1 - \beta} + \frac{\beta}{1 - \beta}\sum_{j=0}^{n-1} |\xi_j^h|,$$
$$\sum_{k=b}^{n-2}\|g_k\| \le \theta_+ \sum_{j=0}^{n-1}\|\xi_j^u\| + UV_\Theta + \frac{2\beta}{1 - \beta}\sum_{j=0}^{n-1}\|\xi_j^p\|$$
$$+ \theta_+ \sum_{k=b}^{n-2} |\widetilde{h}_{k+1} - 1|.$$

The sums over j include defects before b. Insert these bounds and Equation (111) in Equation (115) and discard only nonnegative normalized-progress terms outside $b \le k \le n - 2$. One obtains

$$\frac{1}{2}\sum_{k=0}^{n-1} \theta_k \frac{\widetilde{h}_{k+1}}{F_k} \ge \left(1 + \frac{3\mu}{4}\right) n - D - \frac{\theta_+\mathcal{G}}{2}\sum_{j=0}^{n-1}\|\xi_j^u\|$$
$$- \frac{\beta}{1 - \beta}\sum_{j=0}^{n-1}\left[\mathcal{G}\|\xi_j^p\| + \frac{\theta_+(\mathcal{G} + 1)}{2}|\xi_j^h|\right] - \frac{\mathcal{G}U + \mathcal{V}}{2}V_\Theta.$$

Since $\theta_+ > 2$, Cauchy–Schwarz bounds the bracket by $\theta_+(\mathcal{G} + 1)\sqrt{\|\xi_j^p\|^2 + |\xi_j^h|^2}$. Finally, the stepwise clock quadrature and $\ell(t_k, t_{k+1}) \le 2\eta$ give

$$t_{k+1} - t_k = \frac{\theta_k}{2}\left(\frac{\widetilde{h}_{k+1}}{F_k} + \zeta_k\right) + \frac{1}{2\eta}\int_{t_k}^{t_{k+1}} [\theta_1(q) - \theta_k]v(q)\,dq,$$
$$\left|\frac{1}{2\eta}\int_{t_k}^{t_{k+1}} [\theta_1(q) - \theta_k]v(q)\,dq\right| \le \frac{\eta L_\Theta}{4}\left(\frac{\ell(t_k, t_{k+1})}{\eta}\right)^2 \le (\theta_+ + 2)\delta,$$
$$V_\Theta \le 2n(\theta_+ + 2)\delta.$$

Combining these bounds with Equation (109) proves Equation (103) for $n \geq b+2$ with exactly the C_M, D in Equation (107). For $0 \leq n \leq b+1$, positive phase gives $t_n - t_0 \geq 0$, whereas $D \geq (1+3\mu/4)(b+1)$ makes the asserted lower bound nonpositive. Thus the bound holds for every prefix. The threshold gives $C_M\delta \leq \mu/8$, and $n \geq 8D/\mu$ therefore implies Equation (104). At the last preterminal iterate,

$$v_{\mathrm{BF}}(n_{\mathrm{BF}}-1) - D \leq t_{n_{\mathrm{BF}}-1} - t_0 < T_* - t_0,$$
$$n_{\mathrm{BF}} \leq \left\lceil \frac{T_* - t_0 + D}{v_{\mathrm{BF}}} \right\rceil,$$

which proves the supplied-entry count without discarding an endpoint term.

Entry from zero momentum. For the prescribed radius r, use part (a) with the fixed compact entry set

$$\mathcal{K}_{\mathrm{in}} := \big\{(u,h,p) : \|u\| \leq r+2, \quad |h-1| \leq (1+\beta)/2, \quad \|p\| \leq (1-\beta)\theta_+ r + 1\big\}.$$

Increase C to at least $r+1$. At a zero-buffer state, the predicted tangential buffer is $1-\beta$ and its full norm is at most $(1-\beta)\sqrt{1+(\theta_+ r)^2}$. The componentwise Fermi calculation at radius r therefore bounds the initial joint defect by $C_{\mathrm{init}}\delta$, where

$$C_{\mathrm{init}} := 3r^2 + 8r + 4 + e(1-\beta)\left[r + \frac{r^2}{2} + 4\sqrt{1+(\theta_+ r)^2}\right].$$

Replace δ_0 by the minimum of its previous value and

$$\frac{1}{2\max\{r+1, r+r^2/2\}}, \qquad \frac{1-\beta}{4C_{\mathrm{init}}}, \qquad \frac{1}{2C_{\mathrm{init}}\sqrt{1+(\theta_+ r)^2}}.$$

These are explicit positive choices with the claimed dependencies. The unchanged tube and interval hypotheses supply a complete first step. Writing $\varepsilon_{\mathrm{init}} = C_{\mathrm{init}}\delta$, one has

$$\begin{aligned} |h_1 - 1| &\leq \beta + \varepsilon_{\mathrm{init}} < (1+\beta)/2, \\ \|p_1\| &\leq (1-\beta)\theta_+ r + \varepsilon_{\mathrm{init}} < (1-\beta)\theta_+ r + 1, \\ \|u_1\| &\leq r + 1 + \varepsilon_{\mathrm{init}} < r + 2, \\ 0 < \eta &\left([1+(\theta_+ r)^2]^{-1/2} - \varepsilon_{\mathrm{init}}\right) \leq \ell(0,t_1) \leq \eta(1+\varepsilon_{\mathrm{init}}) < 2\eta < S. \end{aligned}$$

The endpoint belongs to $(0,T_*) \times \operatorname{int}\mathcal{K}_{\mathrm{in}}$ with its actual carried buffer. Applying part (a) from that endpoint proves the zero-momentum clock and count, with exactly one additional update. The initial normal distance is at most $r\eta \leq C\eta$, so its geometric assertion also includes initialization. □

The all-prefix clock and crossing count do not require a minimum trajectory length. The displayed gain $\mu/2$ applies once the explicit endpoint deficit has been amortized over $8D/\mu$ updates. No normal direction or cycle phase is prescribed, and the first zero-momentum update is included in the total count.

The conditions are nonempty: for $\eta = 1$ and $\theta > 2$, the loss $L(x,y) = -x + \theta\|y\|^2/2 + (x-b)_+^4$, with $b > 2\max I/\theta$, is C^3 and bounded below. Here $y \in \mathbb{R}^{d-1}$ and $(a)_+ = \max\{a, 0\}$. Along $z(t) = (2t/\theta, 0)$ before $2t/\theta = b$, one has $B = 1$, $\Theta = \theta I_{d-1}$, and $\delta = 0$. Choose the tube radius and endpoint margins to contain the complete steps from the constructed domain. The one-step entry and the all-prefix clock then apply with the explicit positive gain above.

F.4 Post-entry comparisons

We compare NGD with momentum on the same reference path segment, using the same learning rate, iteration time, and endpoint. The cycle comparison retains its supplied joint entry. The branch-free comparison starts both methods at the same prepared position and initializes momentum at zero; its count includes the first momentum update. Any earlier NGD warm-up is outside these counts.

Remark F.11 (Acceleration from a zero buffer, informal). Start `NormGD` and `NormGD-M` with $m_0 = 0$ from the same point, where `NormGD` already oscillates, on a part of the flow path of arclength S where $\theta > 2$. By Corollary F.12(b), `NormGD-M` needs fewer updates to traverse it, $n_{\mathrm{BF}} \leq (1 -$

$a\beta)\, n_{\mathrm{NGD}} + C/(1-\beta)^2$, where $a, C > 0$ do not depend on η or β and the last term accounts for the warm-up of the momentum buffer. At fixed spectral bounds, this holds when the geometry changes slowly over the memory of the buffer and the path is long enough to amortize the warm-up:

$$\delta \lesssim \beta(1-\beta)^4, \tag{116}$$

$$\frac{S}{\eta} \gtrsim \frac{1}{\beta(1-\beta)^2}. \tag{117}$$

The gain is smaller than the factor K_β of the period-two orbit, but it needs neither the orbit nor its entry.

Corollary F.12 (Conditional comparison on one reference path segment)**.** *Suppose that Theorem F.8 applies to a given* `NormGD` *trajectory. The following momentum clauses are independent.*

(a) Momentum near the reference period-two orbit. *If Propositions F.5 and F.6 apply to NGM on the same reference path segment, with $0 \le \ell(0, t_0^\beta) \le C_{\mathrm{ent}}\eta$ for fixed $C_{\mathrm{ent}} \ge 0$, there is a constant C, depending only on the fixed data of these results and the NGD passage, such that*

$$\left| n_{\mathrm{P2}} - \frac{1-\beta}{1+\beta} n_{\mathrm{NGD}} \right| \le C(1+\delta T_*),$$

$$n_{\mathrm{NGD}} - n_{\mathrm{P2}} \ge \frac{2\beta}{1+\beta} T_* - C(1+\delta T_*).$$

(b) Zero momentum from the same prepared position. *Let $t_0 = 0$ and initialize* `NormGD-M` *at $x_0^\beta = x_0$ with $m_0^\beta = 0$. Suppose that the zero-momentum clause of Theorem F.10 applies, with constants C_M, D and gain parameter μ from that theorem. Let σ be the NGD spectral margin in Equation* (92)*. If*

$$\delta \le \frac{\mu}{4(20/\sigma^2 + C_M)}, \tag{118}$$

then the total crossing counts from this common position satisfy

$$n_{\mathrm{BF}} \le \left(1 - \frac{\mu}{4}\right) n_{\mathrm{NGD}} + 2 + \frac{D + 4/9}{1 + \mu/2}. \tag{119}$$

In particular, momentum uses strictly fewer updates if

$$n_{\mathrm{NGD}} > \frac{4}{\mu}\left(2 + \frac{D + 4/9}{1 + \mu/2}\right).$$

For the common prepared position, assume $S > 2\eta$, $\ell(\inf I, 0) > 2\eta$, and $\ell(T_*, \sup I) > 4\eta$. The full spectral dependence underlying Remark F.11 is as follows. It suffices to require $\rho/\eta \gtrsim \kappa_\perp/(1-\beta)$ and

$$\delta \lesssim \min\Bigg\{ \sigma^2 \min\{\sigma, \beta/\theta_+\},\ \frac{(1-\beta)^4}{\theta_+ \kappa_\perp^2 [\kappa_\perp + (1-\beta)\theta_+]},\ \frac{\alpha^2 (1-\beta)^3 \min\{\alpha, \beta/\theta_+\}}{\theta_+ \kappa_\perp^3 [\alpha\kappa_\perp + (\alpha + 1 - \beta)\theta_+]} \Bigg\}. \tag{120}$$

These tube and variation bounds suffice for both passage results and give

$$n_{\mathrm{BF}} \le \left(1 - \frac{1}{4}\min\left\{1 - \frac{2}{\theta_{1,-}}, \frac{\beta}{\theta_+}\right\}\right) n_{\mathrm{NGD}} + O\left(\frac{\theta_+^4}{\theta_-^3 (1 - 2/\theta_{1,-})(1-\beta)^2}\right).$$

The implicit constant is absolute numerical. With $g_\beta := \min\{1 - 2/\theta_{1,-}, \beta/\theta_+\}$ and $D_\beta := \theta_+ \kappa_\perp^3 / [(1 - 2/\theta_{1,-})(1-\beta)^2]$, strict improvement $n_{\mathrm{BF}} < n_{\mathrm{NGD}}$ holds whenever

$$T_* \gtrsim \frac{D_\beta}{g_\beta}. \tag{121}$$

All implicit factors in these sufficient conditions are compatible absolute numerical constants. The proof supplies simultaneous numerical choices for every factor in these sufficient conditions.

All constants and smallness conditions are uniform in the segment length when the declared geometric and entry bounds are fixed.

Proof of Corollary F.12. For part (a), the two passage results give nonnegative count budgets Δ_{NGD} and Δ_{P2} such that

$$|n_{\mathrm{NGD}} - T_*| \le \Delta_{\mathrm{NGD}},$$
$$\left|\frac{1+\beta}{1-\beta} n_{\mathrm{P2}} - T_*\right| \le \Delta_{\mathrm{P2}}.$$

The budgets are the endpoint estimates in Equations (88) and (96), evaluated with the respective entries. Subtraction gives

$$\left|n_{\mathrm{P2}} - \frac{1-\beta}{1+\beta} n_{\mathrm{NGD}}\right| \le \frac{1-\beta}{1+\beta}(\Delta_{\mathrm{NGD}} + \Delta_{\mathrm{P2}}),$$
$$n_{\mathrm{NGD}} - n_{\mathrm{P2}} \ge \frac{2\beta}{1+\beta} T_* - \Delta_{\mathrm{NGD}} - \frac{1-\beta}{1+\beta}\Delta_{\mathrm{P2}}.$$

Both budgets are bounded by their stated constants times $1 + \delta T_*$, proving this clause.

For part (b), put $v_{\mathrm{BF}} := 1 + 3\mu/4 - C_M\delta$. The first momentum update has positive arclength and retains its resulting buffer. By Equation (105) and $t_1^\beta > 0$,

$$n_{\mathrm{BF}} \le 1 + \left\lceil \frac{T_* - t_1^\beta + D}{v_{\mathrm{BF}}} \right\rceil \le 2 + \frac{T_* + D}{v_{\mathrm{BF}}}.$$

At the NGD crossing, Equation (94) gives

$$T_* \le t_{n_{\mathrm{NGD}}} \le \left\lceil \left(1 + \frac{20\delta}{\sigma^2}\right) n_{\mathrm{NGD}} + \frac{4}{9} \right\rceil.$$

Combining these inequalities retains the full finite-step comparison:

$$n_{\mathrm{BF}} \le \frac{1 + 20\delta/\sigma^2}{1 + 3\mu/4 - C_M\delta} n_{\mathrm{NGD}} + 2 + \frac{D + 4/9}{1 + 3\mu/4 - C_M\delta}.$$

Since $0 < \mu < 1$, condition (118) implies

$$v_{\mathrm{BF}} \ge 1 + \mu/2,$$
$$1 - \frac{1 + 20\delta/\sigma^2}{v_{\mathrm{BF}}} = \frac{3\mu/4 - (20/\sigma^2 + C_M)\delta}{v_{\mathrm{BF}}} \ge \frac{\mu/2}{1 + 3\mu/4} \ge \frac{\mu}{4}.$$

Substitution proves Equation (119). Thus its additive term counts the initial momentum update, the integer crossing error, and the two clock transients. For a separately verified preparation from an earlier common position, its actual number of updates and clock displacement must be included before applying this comparison. The preparation theorem does not furnish a geometry-only uniform rate over arbitrarily small initial top components.

Simultaneous bounds from the geometry. For this calculation write $q = 1-\beta$, $Q = \theta_+$, and $m = \theta_-$, so $\kappa_\perp = Q/m$, and retain $\alpha = 1 - 4/\theta_{1,-}^2$. Introduce the proof constants

$$U_\beta := \frac{16(Q+1)}{qm}, \qquad E_\beta := \frac{Q^2 U_\beta^2}{\alpha m},$$
$$A_\beta := 64 E_\beta \left(U_\beta + \frac{4\beta Q}{q} + \frac{Q}{\alpha}\right).$$

We first verify the precise sufficient conditions

$$(U_\beta + 1)\eta < \rho,$$
$$\delta \le \min\left\{\frac{\sigma^3}{36}, \frac{q}{32 Q U_\beta^2 (U_\beta + Q)}, \frac{\mu}{8(20/\sigma^2 + A_\beta)}\right\}. \tag{122}$$

Set $U = U_\beta$ and $P = QU$. Thus $Q > 2$, $m \le Q$, $U \ge 16/q$, and $P > 32$. The NGD passage proof gives $\|u_j\| < 1/2$ through crossing. The first zero-buffer momentum update is exactly the first NGD update, so the entry calculation in Theorem F.10 can use the smaller supplied set

$$\mathcal{K}_{\mathrm{in}} = \{(u, h, p) : \|u\| \le 1/2, \quad |h - 1| \le (1 + \beta)/2, \quad \|p\| \le qQ/2 + 1\}.$$

The closure construction then gives $H = (3 + \beta)/4$ and $R = 8(Q + 1)/(qm) = U/2$, because

$$1 + \frac{4}{m} \sup_{\mathcal{K}_{\mathrm{in}}} (Q\|u\| + \|p\|) \le 1 + \frac{4(Q + 1)}{m} < \frac{8(Q + 1)}{qm}.$$

Its entire containing domain has normal radius U and buffer bound P. The componentwise coefficients in Equation (72) satisfy

$$c_u \le \frac{5}{4}U^2, \qquad c_\ell \le \frac{5}{2}U^2, \qquad c_b \le 2qU^2 + 12P.$$

Indeed, $h_* \ge q$, $U + U^2/2 \le 9U^2/16$, and $\widehat{M} \le \sqrt{4 + P^2} \le 65P/64$; use $e < 11/4$ in the buffer coefficient. The closure tolerance and the coefficient that controls the full Fermi defect obey

$$\varepsilon_{\mathrm{cl}} \ge \frac{q}{2QU}, \qquad c_\ell + c_u + c_b + Q + 2 \le 13U(U + Q).$$

For the first inequality, each entry in the minimum defining $\varepsilon_{\mathrm{cl}}$ is at least $q/(2QU)$; in particular, $a_{\mathrm{cl}} \ge q/(2QU)$ follows from $h_* \le 1$ and $R = U/2$. Hence the second variation bound in Equation (122) implies the closure and complete-step conditions, as well as

$$\left[Qc_u + 2U(Q + 2) + \frac{2\beta c_b}{q}\right] \delta \le 5QU^2\delta < \frac{1}{12}.$$

It also implies all first-update conditions: at $r = 1/2$ the entry proof has $C_{\mathrm{init}} \le 12(1 + qQ)$ and $\sqrt{1 + Q^2/4} \le Q$, so it is enough that

$$\delta \le \min\left\{\frac{1}{3}, \frac{q}{48(1 + qQ)}, \frac{1}{24Q(1 + qQ)}\right\}.$$

Each of these bounds follows from $\delta \le q/[32QU^2(U + Q)]$ and $U \ge 16/q$. Thus the smaller supplied set is reached after one counted update, and all subsequent transitions are valid.

For the storage coefficients in Equation (106), $Z_* \le 9P/8$, $F_* \le 9P/8$, and $qP/m \ge 16$ give

$$\mathcal{E} \le \frac{117P^2}{128m}, \qquad \mathcal{G} \le \frac{20P}{\alpha m}, \qquad \mathcal{G} + 1 \le \frac{21P}{\alpha m},$$
$$\mathcal{V} \le \frac{32P^2}{\alpha^2 m} + \frac{11P^2}{\alpha m^2}.$$

Substitute these estimates into Equation (107) and use $E_\beta = P^2/(\alpha m)$ to obtain

$$\begin{aligned}\frac{C_M}{E_\beta} &\le \frac{5}{4} + \frac{25U}{2} + 42\beta U + \frac{252\beta Q}{q} + 42 + \frac{32(Q + 2)}{\alpha} + \frac{11(Q + 2)}{m} \\ &\le 64\left(U + \frac{4\beta Q}{q} + \frac{Q}{\alpha}\right).\end{aligned}$$

Here $Q + 2 \le 2Q$ and $(Q + 2)/m \le qU/12$ give the last line. For the transient, $\log(1/\beta) \ge q$, $\log(8P) \le P$, and $\beta^{b+1} \le 1/(8P)$ imply

$$\begin{aligned}D &\le 2 + \frac{2P}{q} + \frac{117}{16}E_\beta + \frac{21Q}{16q\alpha m} \\ &\le \left(\frac{1}{256} + \frac{1}{8} + \frac{117}{16} + \frac{21}{8192}\right) E_\beta < 8E_\beta.\end{aligned}$$

We have therefore constructed

$$C = U_\beta + 1, \qquad C_M \le A_\beta, \qquad D \le 8E_\beta.$$

The final variation bound in Equation (122) now gives $C_M\delta \le \mu/8$ and Equation (118), simultaneously. The clock comparison and Equation (119) therefore give the claimed clock and count bounds with no remaining constant choices. Finally, the NGD clock estimate and $20\delta/\sigma^2 \le \mu/8 < 1/4$ yield

$$n_{\mathrm{NGD}} \ge \frac{4}{5}\left(T_* - \frac{4}{9}\right) > \frac{4}{\mu}(8E_\beta + 3) \quad \text{if } T_* > \frac{6(8E_\beta+3)}{\mu}.$$

This absorbs the additive count term and proves strict improvement.

Numerical factors for the leading-order bounds. Since $Q > 2$, the inequality $Q + 1 \le 3Q/2$ gives

$$\begin{aligned} U_\beta + 1 &\le \frac{26\kappa_\perp}{q}, \\ 8E_\beta + 3 &\le \frac{5000Q\kappa_\perp^3}{\alpha q^2}, \\ A_\beta &\le \frac{36864Q^5}{\alpha^2 q^3 m^4}(24\alpha + 4\alpha\beta m + qm). \end{aligned}$$

For the second inequality, use $E_\beta \le 576Q^4/(\alpha q^2 m^3)$ and $Q^4/(\alpha q^2 m^3) > 2$. The radius bound also controls the tracking distance, and the same factor 5000 bounds the momentum clock deficit $8E_\beta$. The closure threshold satisfies

$$\begin{aligned} \frac{q}{32QU_\beta^2(U_\beta + Q)} &= \frac{q^4 m^3}{8192Q(Q+1)^2[16(Q+1) + qmQ]} \\ &\ge \frac{q^4 m^3}{442368Q^4(1+qm)}. \end{aligned}$$

For the remaining speed threshold, it suffices to impose

$$\begin{aligned} \delta &\le \frac{\mu\sigma^2}{320}, \\ \delta &\le \frac{\mu\alpha^2 q^3 m^4}{14155776Q^5(\alpha + \alpha\beta m + qm)}. \end{aligned}$$

Indeed, these imply $\delta \le \mu/[16(20/\sigma^2)]$ and $\delta \le \mu/(16A_\beta)$, whose sum gives the last bound in Equation (122). The identities

$$\begin{aligned} \frac{q^4 m^3}{Q^4(1+qm)} &= \frac{q^4}{Q\kappa_\perp^2(\kappa_\perp + qQ)}, \\ \frac{\mu\alpha^2 q^3 m^4}{Q^5(\alpha + \alpha\beta m + qm)} &= \frac{\mu\alpha^2 q^3}{Q\kappa_\perp^3[\alpha\kappa_\perp + (\alpha\beta + q)Q]} \end{aligned}$$

express the preceding sufficient bounds in terms of $\kappa_\perp$. Since 36, 320, 442368, and 14155776 are smaller than 2^{24}, the precise certificate follows whenever

$$\delta \le 2^{-24}\min\left\{\sigma^2\min\{\sigma,\mu\},\ \frac{q^4}{Q\kappa_\perp^2(\kappa_\perp + qQ)},\ \frac{\mu\alpha^2 q^3}{Q\kappa_\perp^3[\alpha\kappa_\perp + (\alpha\beta+q)Q]}\right\}.$$

To simplify the acceleration formulas, put $\varepsilon = 1 - 2/\theta_{1,-}$, $b = \beta/Q$, and $\nu = \min\{\varepsilon, b\}$ within this calculation. Then $0 < \varepsilon < 1$, $0 < b < 1/2$, and

$$\begin{aligned} &\alpha = \varepsilon(2-\varepsilon), \qquad \varepsilon \le \alpha \le 2\varepsilon, \\ &\mu = \min\left\{\frac{\varepsilon}{1-\varepsilon}, \frac{2bQ}{Q+q}\right\}, \qquad \nu \le \mu \le 2\nu. \end{aligned}$$

The lower bound follows from $\varepsilon/(1-\varepsilon) \geq \varepsilon$ and $2Q/(Q+q) \geq 1$. For the upper bound, if $\varepsilon \leq b$ then $\varepsilon < 1/2$ gives $\varepsilon/(1-\varepsilon) \leq 2\varepsilon$; otherwise the second term is at most $2b$. Consequently,

$$\begin{aligned}
\frac{\mu}{4} &\geq \frac{1}{4}\min\left\{1-\frac{2}{\theta_{1,-}}, \frac{\beta}{Q}\right\},\\
\frac{Q\kappa_\perp^3}{\alpha q^2} &\leq \frac{Q^4}{m^3\varepsilon q^2},\\
\frac{Q\kappa_\perp^3}{\mu\alpha q^2} &\leq \frac{Q^4}{m^3\varepsilon q^2}\max\left\{\frac{\theta_{1,-}}{\theta_{1,-}-2}, \frac{Q}{\beta}\right\}.
\end{aligned}$$

The same comparison replaces the clock gain $1+5\mu/8$ by $1+5\nu/8$. Thus the constants 5000 for the count and clock remainders and 30000 for the sufficient length remain valid after these relaxations.

For the variation condition, $\sigma \leq \alpha \leq 2\varepsilon$ and $\mu \geq \nu$ imply

$$\begin{aligned}
\sigma^2\min\{\sigma, b\} &\leq 2\sigma^2\min\{\sigma, \mu\},\\
\frac{\alpha^2 q^3\min\{\alpha, b\}}{Q\kappa_\perp^3[\alpha\kappa_\perp + (\alpha+q)Q]} &\leq \frac{2\mu\alpha^2 q^3}{Q\kappa_\perp^3[\alpha\kappa_\perp + (\alpha\beta+q)Q]}.
\end{aligned}$$

For the first inequality, use $\min\{\sigma, b\} \leq 2\min\{\sigma, \varepsilon, b\}$. For the second inequality, use $\min\{\alpha, b\} \leq 2\nu \leq 2\mu$ and the larger denominator on the left. The closure term is unchanged. Multiplying the minimum in Equation (120) by 2^{-25} therefore implies the preceding sharper certificate. Thus one simultaneous choice of the numerical factors for the tube condition, variation condition, count and clock remainders, and sufficient length is

$$C_{\mathrm{tube}} = 26, \qquad c_{\mathrm{var}} = 2^{-25}, \qquad C_{\mathrm{rem}} = 5000, \qquad C_{\mathrm{length}} = 30000. \tag{123}$$

Concretely, require $\rho/\eta > 26\kappa_\perp/q$, use the factor 2^{-25} in the variation bound, and take $T_* > 30000Q^4/(m^3\varepsilon\nu q^2)$ for strict improvement. These choices prove the bounds above with absolute numerical constants; none of them depend on the geometry or momentum. □

The conditions hold on nonquadratic parameter families, not only on a single fixed example. To see this directly, fix $\theta_0 > 2$, $0 < a < \theta_0 - 2$, and $\beta \in (0,1)$, take $\eta = 1$ and $\ell_0 > 0$, and consider

$$L(x,y) = -x + \frac{1}{2}[\theta_0 + a\sin(x/\ell_0)]y^2.$$

Along $z(t) = (q(t), 0)$, where $q(0) = 0$ and $\frac{\mathrm{d}q}{\mathrm{d}t} = 2/[\theta_0 + a\sin(q/\ell_0)]$, one has $B = 1$, $K_\kappa = L_B = 0$, and $\theta_1(t) \in [\theta_0 - a, \theta_0 + a]$. Evaluate the spectral margins using the uniform bounds $m = \theta_0 - a$, $Q = \theta_0 + a$, and $\kappa_\perp = Q/m$. Choose $\rho > 26\kappa_\perp/(1-\beta)$ and use the common initial position $(0, \sqrt{1/4 - \theta_0^{-2}})$ with zero momentum for NGM. On this tube,

$$\Gamma_3 \leq a\left(\frac{3}{\ell_0} + \frac{3\rho}{\ell_0^2} + \frac{\rho^2}{2\ell_0^3}\right) \leq \frac{7a}{\ell_0} \qquad \text{if } \ell_0 \geq \max\{1, \rho\}.$$

All spectral margins, containing radii, and error coefficients above can be chosen uniformly in ℓ_0 and S. Let $\delta_{\mathrm{adm}} > 0$ be the minimum in Equation (120) multiplied by 2^{-25}, as justified in Equation (123). Taking $\ell_0 \geq \max\{1, \rho, 7a/\delta_{\mathrm{adm}}\}$ verifies all variation conditions simultaneously. Extend the straight reference beyond both endpoints to provide the required clearance. Finally, choose

$$S > \max\left\{2, \frac{60000Q^4}{m^4(1-2/m)(1-\beta)^2}\max\left\{\frac{m}{m-2}, \frac{Q}{\beta}\right\}\right\}.$$

Then $T_* \geq (\theta_0 - a)S/2$ verifies Equation (121), while all preceding bounds remain unchanged. This proves strict finite-step improvement on a nonempty family with nonzero geometric variation, without changing the learning rate or fixing one value of β.

At fixed spectral geometry, these sufficient bounds imply Remark F.11, with the conditions in Equations (116) and (117). For all $\beta \in (0,1)$,

$$\beta\min\left\{1-\frac{2}{\theta_{1,-}}, \frac{1}{Q}\right\} \leq \min\left\{1-\frac{2}{\theta_{1,-}}, \frac{\beta}{Q}\right\} \leq \frac{\beta}{Q}.$$

Hence the guaranteed fractional gain is of order β, with constants depending only on the fixed geometry. The three terms in Equation (120) are bounded below by positive geometry-dependent multiples of β, $(1-\beta)^4$, and $\beta(1-\beta)^3$, respectively. Together with the tube-width condition and prepared initialization, it is therefore sufficient that

$$\delta \le c_{\mathrm{geo}}\beta(1-\beta)^4, \qquad\qquad T_* > \frac{C_{\mathrm{geo}}}{\beta(1-\beta)^2},$$

for suitable positive constants $c_{\mathrm{geo}}, C_{\mathrm{geo}}$ depending only on that fixed spectral geometry; the tube width is a fixed multiple of $\eta/(1-\beta)$. Since $\theta_{1,-}S/(2\eta) \le T_* \le \theta_+ S/(2\eta)$, the same fixed-geometry scaling gives the sufficient arclength condition in Equation (117). This last specialization absorbs geometric factors into its constants; the factors in Equation (123) remain absolute numerical.