THE ROAD TAKEN: THE ROLE OF OPTIMIZERS AT THE EDGE OF STABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The edge of stability refers to a phenomenon in deep learning with gradient-based optimizers where the Hessian eigenvalues of the loss remain stable above a threshold that the classical descent lemma predicts to be unstable. Previous works formulate the edge of stability with respect to the maximum Hessian eigenvalue and the learning rate. However, we observe that many first-order methods, including gradient descent, significantly violate the stability bound predicted by these theories by a factor as large as $\times 21.1$. Moreover, this deviation turns out to be systematic and highly dependent on the underlying optimizer, which is not captured by previous formulations. This calls for a new formulation of the stability threshold, which we derive from the directional Hessian and the gradient-alignment score with respect to the actual update taken by the optimizer, rather than the maximum curvature mode. Our new formulation of the *realized* edge of stability not only removes optimizer-dependent offsets and provides more consistent predictions of the stability threshold, but also introduces new diagnostic tools that reveal the unique role of the optimizer in actively balancing between the temporal and spatial budgets in first-order optimization.

1 INTRODUCTION

The training dynamics of deep learning optimization still possess many mysteries. One of the most intriguing phenomena is the *edge of stability* (EoS) (Cohen et al., 2021), where gradient descent operates in a region where classical theory predicts instability. The good old descent lemma (Ortega & Rheinboldt, 1970; Bertsekas, 1999) suggests that the available step size η for gradient descent is bounded by the curvature of the objective function:

$$\eta \lambda_{\max}(\boldsymbol{H}) \leq 2, \tag{1}$$

where $\lambda_{\max}$ is the maximum eigenvalue of the Hessian matrix $\boldsymbol{H} = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}^\star)$ of the loss $\mathcal{L}(\boldsymbol{\theta})$ at a local minimum $\boldsymbol{\theta}^\star$. Although this bound holds for convex objectives, its behavior becomes more subtle in highly nonconvex settings such as deep learning. In those regimes, it has been reported that the relation (1) behaves more like a stable attractor that draws the curvature $\lambda_{\max}$ towards a value near $2/\eta$, rather than a strict stability threshold (Lewkowycz et al., 2020; Cohen et al., 2021). During the first few hundred iterations, the top Hessian eigenvalue $\lambda_{\max}$ either shrinks rapidly (Lewkowycz et al., 2020; Zhu et al., 2024) or increases gradually (Cohen et al., 2021) depending on the initialization, until it eventually reaches a stable point near $2/\eta$, and then hovers around this value. This mechanism is referred to as the *edge of stability* (EoS) (Cohen et al., 2021).

Rather than being unified into a single definition, previous works have suggested various alternative formulations for the edge of stability (Lyu et al., 2022; Cohen et al., 2023; Damian et al., 2023; Agarwala et al., 2023; Zhu et al., 2023; Andreyev & Beneventano, 2024; Chen et al., 2024; Cohen et al., 2025; Islamov et al., 2026; Andreyev et al., 2026; Litman, 2026). Nevertheless, all share the same core idea and structure as a product of two or three separable quantities: (1) a learning rate scalar η, (2) a *maximum curvature eigenmode* or *sharpness* S with respect to a (possibly preconditioned) Hessian, and optionally (3) a momentum-induced gain Γ if the optimizer is stateful:

$$\eta S \Gamma \leq 2. \tag{2}$$

For instance, the original formulation (Cohen et al., 2021) assigns $S = \lambda_{\max}(\boldsymbol{H})$ and $\Gamma = 1$. Later, Cohen et al. (2023) incorporate preconditioning by taking $S = \lambda_{\max}(\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2})$ and

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.

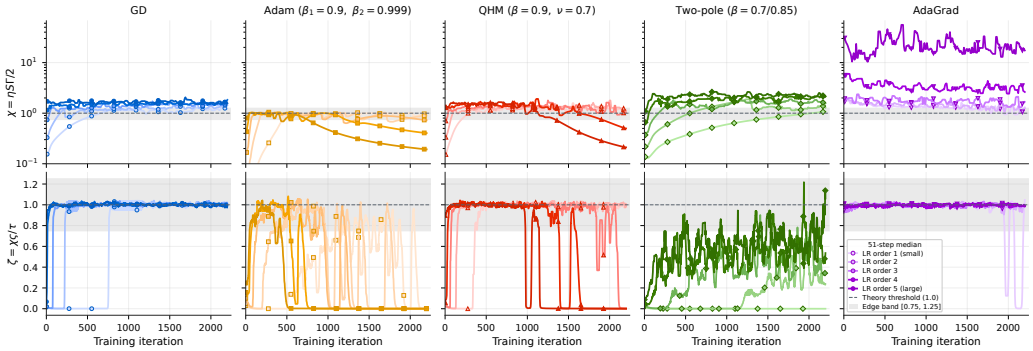


Figure 1: **Edge of stability across different families of optimizers.** We train a fully-connected network on CIFAR-10 using various full-batch gradient-based optimizers: gradient descent (GD), Adam, GD with a pole-zero filter (QHM) and GD with a two-pole filter (Grokfast), and AdaGrad. Theoretical prediction of the edge of stability $\chi := \eta S\Gamma/2 = 1.0$ is occasionally violated in practice, and the offsets χ are heavily dependent on the type of optimizers in use. This systematic deviations motivate a new formulation of the *realized* edge of stability $\zeta := \chi\varsigma/\tau$ with ς and τ being the spatial and temporal calibrations, respectively. This new formulation is more consistent with the empirical observations and removes the optimization-dependent offsets.

introduce a factor $\Gamma = \Gamma(\beta)$ to account for the momentum, e.g., $\Gamma(\beta) = 1/(1+\beta)$ for Heavy Ball (Polyak, 1964) and $\Gamma(\beta) = (1+2\beta)/(1+\beta)$ for Nesterov's momentum (Nesterov, 1983).

In summary, the edge of stability has been understood as a product of the *worst-mode* sharpness and *separable* temporal factors. This formulation provides *available* stability thresholds that are provably marginal for quadratic objectives. However, we find that the actual relative position of the edge, $\chi := \eta S\Gamma/2$, is not consistent with the theoretical prediction $\chi = 1$. As Figure 1 shows, its values range from 0.79 to 21.07 across different types of optimizers and learning rates. Even vanilla GD sits above $\chi = 1$ (top left). The median of the saturated curve grow up to $\chi \approx 1.62$ as the learning rate increases. More complicated optimizers, such as dual-momentum optimizers (Lee et al., 2024; Pagliardini et al., 2025) or pole-zero filters (Ma & Yarats, 2019), show their actual EoS rising to around $\chi = 2.65$ (columns 3 and 4). The most extreme case is AdaGrad (Duchi et al., 2011) (top right), which shows a clear edge starting from $\chi = 1.42$ at $\eta = 0.01$ and reaching $\chi = 21.07$ at $\eta = 0.1$. The deviation of χ from the theoretical prediction of $\chi = 1.0$ depends strongly on the underlying optimization algorithm. This systematic deviation poses a new challenge to the current formulation and interpretation of the edge of stability.

In fact, this formulation (2) is hiding a critical assumption: the worst-mode sharpness S is always realized by the optimizer, which is not always the case. Starting from modeling the actual dynamical system the optimizer operates on, we derive a new formulation of the stability threshold from the directional Hessian (Lee & Jang, 2023; Mishkin et al., 2024; Islamov et al., 2026) and the gradient-update alignment (Lee & Lee, 2025). As shown in the bottom row of Figure 1, this new formulation, called the *realized* edge of stability ζ, largely eliminates optimizer-dependent offsets. Furthermore, it is much more efficient to compute and recovers the original formulation χ by decomposing it into interpretable factors. This provides a richer understanding of the phenomenon itself and reveals the optimizer's role in steering along the edge of stability by actively calibrating the temporal gauge and spatial budget.

Our contributions are:

- We extend the original formulation of the edge of stability (Cohen et al., 2021; 2023) to general gradient-based optimizers with states, which is exact for fixed quadratic objectives.
- We derive a new formulation of the stability threshold, based on the directional Hessian and the gradient-update alignment score (Lee & Lee, 2025), which is more consistent with empirical observations and eliminates optimizer-dependent offsets.
- We factorize the realized edge of stability ζ into interpretable components: the original edge of stability χ, spatial participation ς, and temporal calibration τ, and show that the two EoS formulations coincide, $\zeta = \chi$, if and only if $\varsigma = \tau$.

- This provides a simple yet powerful diagnostic tool for understanding the main factors driving a training trajectory at the edge of stability, revealing how much of the potential loss decrease, $\boldsymbol{g}^\top \boldsymbol{u} = -(1/\eta)\mathrm{d}\mathcal{L}/\mathrm{d}t$, carried out by the optimizer is actually converted into real progress $-\Delta\mathcal{L}/\eta$, and how much is lost to different components.

2 Preliminaries: Edge of Stability

To understand the edge of stability, we start by reviewing its previously established formulations. Following the aforementioned lines of work, we focus on the full-batch settings where the analysis is more tractable and the effect is more pronounced. Let us define a few notations to clarify the following discussion. In gradient-based optimization, we train a parameter $\boldsymbol{\theta}$ by minimizing a loss function $\mathcal{L}(\boldsymbol{\theta})$ through its first-order gradients $\boldsymbol{g} := \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$. We can either use this gradient directly, or filter it by some input-dependent preconditioner $\boldsymbol{P}$ and temporal filter Q to produce a parameter update $\boldsymbol{u}$ which has the same shape as the parameter $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{u}_t\,, \quad \text{where} \quad \boldsymbol{u}_t = \boldsymbol{P}_t^{-1}(Q_t * \boldsymbol{g}_{\leq t})_t\,. \tag{♠}$$

With an input-dependent preconditioner $\boldsymbol{P}_t$ and a causal gradient filter Q_t acting on a history of gradient signal $\boldsymbol{g}_{\leq t}$, this gives a generic formulation for gradient-based optimizers with internal states. This not only includes GD, Heavy Ball (Polyak, 1964), Nesterov's Momentum (Nesterov, 1983), and various preconditioned versions of these (Duchi et al., 2011; Tieleman & Hinton, 2012; Kingma & Ba, 2015; Loshchilov & Hutter, 2019), but also extends to more exotic temporal filters such as dual-momentum (Lee et al., 2024; Pagliardini et al., 2025) or quasi-hyperbolic momentum (QHM) (Ma & Yarats, 2019). For example, GD corresponds to $\boldsymbol{P}_t = I$ and $Q_t = I$, and Heavy Ball with momentum coefficient β is repesented by $Q * \boldsymbol{g}_{\leq t} = \sum_{k=0}^{t} \beta^k \boldsymbol{g}_{t-k}$. Signal processing notation for the gradient filter Q allows us to study this dynamics in frequency domain, i.e., $Q(z) = \sum_{t=0}^{\infty} Q_t z^{-t}$. We also assume its dc gain is normalized to $Q(1) = 1$, so that the learning rate η is solely responsible for the gain of the system. This simplifies the analysis without loss of generality.

From extensive observations, Cohen et al. (2021; 2023) formulated the EoS as a *separable* product of a function of the preconditioner $\boldsymbol{P}$ and a function of the momentum coefficient β,

$$\eta\lambda_{\max}(\boldsymbol{P}_t^{-1/2}\boldsymbol{H}_t\boldsymbol{P}_t^{-1/2})\,\Gamma(\beta) \;<\; 2, \tag{3}$$

where $\boldsymbol{H}_t$ is the Hessian at time t. For example, Heavy Ball gives $\Gamma(\beta) = 1/(1+\beta)$, normalized Heavy Ball gives $\Gamma(\beta) = (1-\beta)/(1+\beta)$, Nesterov's momentum gives $\Gamma(\beta) = (1+2\beta)/(1+\beta)$, and so on. It is proven for each momentum type that this behaves as a marginal stability condition for the frozen quadratic system (Cohen et al., 2023). The following proposition generalizes their results to any gradient-based optimizer of type (♠).

Proposition 2.1 (Edge of Stability for Stateful Optimizers)**.** [proof] *Consider a dynamical system given by the update rule (♠). Let $(\lambda, \boldsymbol{v})$ be any eigenpair of the preconditioned Hessian $\boldsymbol{P}_t^{-1/2}\boldsymbol{H}_t\boldsymbol{P}_t^{-1/2}$. Then the gain $k_\star(Q)$ at the first crossing of the unit circle by the root locus of $Q(z)$ is given by*

$$k_\star(Q) = \min_{\omega\in\Omega_+(Q)} \left|\frac{1-e^{i\omega}}{Q(e^{i\omega})}\right|, \quad \textit{where} \quad \Omega_+(Q) := \left\{\omega\in(0,\pi] : \frac{1-e^{i\omega}}{Q(e^{i\omega})} \in \mathbb{R}_{>0}\right\}, \tag{4}$$

where $\Omega_+(Q)$ is the set of frequencies on the unit circle at which the locus can cross with a real positive gain. If we write $\Gamma := 2/k_\star(Q)$, then the equivalent stability condition for the frozen quadratic system is attained by

$$\eta\lambda\Gamma \;\leq\; 2. \tag{♣}$$

Moreover, if the first unit-circle crossing is at the Nyquist frequency $z = -1$ with $Q(-1) > 0$, i.e., period-2 oscillation, then $\Gamma = Q(-1)$ and the stability condition becomes $\eta\lambda Q(-1) \leq 2$.

Corollary 2.2 (Recovering Cohen et al. (2023)'s formulae)**.** [proof] *The coefficient $\Gamma(\beta)$ in the inequality (3) for momentum-based optimizers is attained at the Nyquist frequency $z = -1$, i.e., $\Gamma(\beta) = Q(-1)$ for (dc gain-normalized) Heavy Ball and Nesterov's momentum.*

The detailed derivation is provided in Appendix B.2. The gauge Γ of each optimizer filter, including the cases $\Gamma \neq Q(-1)$, is derived in Appendix C.1. The sketch of the proof is as follows: Let the

parameter along the i-th eigenvector be $x_i = \boldsymbol{v}_i \cdot \boldsymbol{\theta}$, then the dynamics of the parameter along this direction is given by (omit i for brevity)

$$x_{t+1} \;=\; x_t - \eta\lambda(Q_t * x_{\le t})_t. \tag{5}$$

In z-domain, we have $X(z) = \sum_{t=0}^{\infty} x_t z^{-t}$ and the convolution becomes a multiplication:

$$zX(z) \;=\; X(z) - \eta\lambda Q(z)X(z). \tag{6}$$

Dividing both sides by $X(z)$, we obtain the *characteristic transfer function* along this mode:

$$z - 1 + \eta\lambda Q(z) \;=\; 0. \tag{7}$$

Classical control theory tells us that any linear discrete-time dynamic system must have all its roots z inside the unit circle $|z| < 1$ in order to be stable (Ogata, 1995), which gives the results.

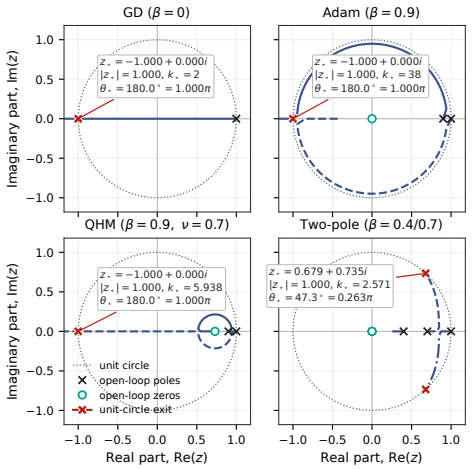


Figure 2: **Root locus of optimizers.** It is possible to exit the unit circle at $z \neq -1$, i.e., $\Gamma \neq Q(-1)$.

From Proposition 2.1 and Corollary 2.2, we see that the original EoS (Cohen et al., 2021) and adaptive EoS (Cohen et al., 2023) are special cases of this general stability condition. The boxed bound (♣) is an *available* stability limit on a frozen quadratic, which is exact in this case, as demonstrated in Appendix C.3: each eigenmode is independent, Γ is a functional of Q alone, and $\chi := \eta S\Gamma/2 \le 1$ is necessary for no mode to be linearly unstable. Using χ as the *observed* hovering edge of a nonlinear trajectory is a different statement, which requires an additional *occupation hypothesis*: over the relevant horizon the update $\boldsymbol{u}$ keeps realizing the worst-case mode of that stability limit. In other words, it stays in $\mathrm{span}(\boldsymbol{v}_{\max})$ of a fixed (preconditioned) Hessian $\boldsymbol{P}_t^{-1/2}\boldsymbol{H}_t\boldsymbol{P}_t^{-1/2}$ over the relevant horizon. Still, the marginal stability condition (♣) remains an exact stability limit for the frozen quadratic system on which the original theory is based. It does not, by itself, imply that a nonlinear deep learning trajectory must hover at $\chi = 1$. It is noteworthy that Q does not play a role in the occupation hypothesis. Section 3 tests the occupation hypothesis first on vanilla GD, where $\boldsymbol{P} = I$, $Q \equiv 1$ and $\Gamma = 1$, so systematic mispredictions $\chi > 1$ cannot be blamed on a missing momentum or on a preconditioning artefact.

3 Optimizer-Dependent Offsets

The extended formulation of the edge of stability (♣) allows us to test Cohen's formulation of EoS in various optimizers. We train a 5-layer MLP with 200 hidden units with $\tanh$ activation on CIFAR-10 (Krizhevsky, 2009) using different full-batch optimizers and learning rates. Specifically, we use gradient descent (GD), GD with Heavy Ball (GDM), GD with Nesterov's momentum (GDN), Adam (Kingma & Ba, 2015), AdamW (Loshchilov & Hutter, 2019), AdaGrad (Duchi et al., 2011), RMSprop (Tieleman & Hinton, 2012), AdaFactor (Shazeer & Stern, 2018), NAdam (Dozat, 2016), PAdam (Chen et al., 2020), AMSGrad (Reddi et al., 2018), GD with a quasi-hyperbolic momentum (QHM) (Ma & Yarats, 2019), and GD with a Grokfast-like two-pole filter (Lee et al., 2024). For each optimizer, we vary the learning rates on a geometrically scaled grid of five values, depending on the optimizer family. The results are summarized in Figure 1 and Figures 8 and 9, and Table 6 in Appendix E. To measure the discrepancy between the realized and theoretical edge of stability, we define the *relative* edge of stability $\chi := \eta S\Gamma/2$ as a unit-free measure. $\chi = 1$ corresponds to theoretical exactness, and is displayed as dashed lines. We additionally summarize the range of χ for different (optimizer, lr) in Figure 3. The range is computed as the median of χ in the interval between first entering $\chi > 0.75$ and last exiting $\chi < 0.75$ during each training trajectory.

We observe that the deviation of χ varies in a very large range $\chi \in [0.79, 21.07]$ in a systematic manner. Different optimizer families exhibit different patterns of deviation, and the offset generally increases with η, as observed in Figure 1 and Figures 8 and 9 in Appendix E. Offsets lower than 1.0 are easy to interpret: either the trajectory remains strictly within the stable region, or the optimizer

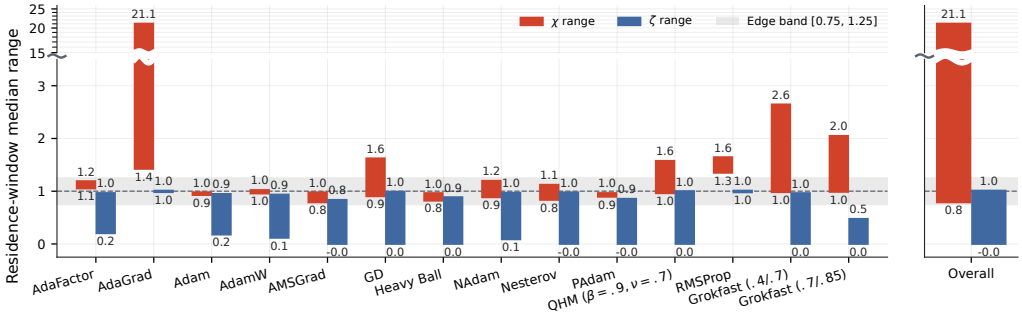


Figure 3: **Family-wise ranges of per-setting interval medians.** Medians of χ taken from the interval between first $\chi > 0.75$ and the final downward exit. The previous formulation of the edge of stability $\chi \leq 1$ (Cohen et al., 2021; 2023) is occasionally violated by large, optimizer-dependent offsets. Our *realized* EoS ζ has its median universally upper bounded by near 1.0, which is more consistent with the name "edge" of stability.

is not able to drive the trajectory to the edge of stability. Those with offsets greater than 1.0 are more interesting. Complex filters, such as QHM and two-pole, raise χ up to 1.576 and 2.646, respectively. At one extreme, AdaGrad severely violates the $\chi \leq 1$ bound to reach $\chi = 21.070$.

Another observation that draws our attention is that even the vanilla full-batch GD violates our expectation to stay near $\chi = 1$: Table 6 reports in-window median χ of 0.904, 1.130, 1.246, 1.478, 1.624 across the five learning rates. Its four of five cases sit above $\chi = 1$ bound, and the observed edge systematically increases with η. Since GD uses $\boldsymbol{P} = I$ and $Q \equiv 1$, hence $\Gamma = 1$ and $\chi = \eta S/2$, this offset cannot be a missing $\Gamma(\beta)$. By contrast, Heavy Ball, Adam, and AdamW stay inside the $\chi \leq 1$ bound, with global median 0.949, 0.960, and 0.971, respectively. The momentum-induced gauge Γ is what those stateful methods relieves from the violation of the edge of stability by GD.

The results not only imply a stabilization mechanism (Damian et al., 2023) that keeps the trajectory stable even outside the guaranteed region of stability by consistently reducing the sharpness, but also suggest that *the attractor of χ itself lies outside the theoretically predicted stable margin*. These optimizer-dependent offsets raise a significant concern that the current formulation of the edge of stability is incomplete, especially in how it identifies the empirical edge with $\lambda_{\max}$ rather than with the curvature along the update that is actually taken.

4 Realized Edge of Stability

The previous observation calls for an update to our formulation of the edge of stability to account for the *actual* training dynamics that are representative of the true stability threshold. Previous works (Lee & Jang, 2023; Mishkin et al., 2024; Islamov et al., 2026) give us a hint to this problem: the worst-case eigenmode $\mathrm{span}(\boldsymbol{v}_{\max})$ that defined the boundary of the region of stability in (♣) may not actually be taken as update directions. This idea encourages us to examine the actual update direction $\boldsymbol{u}$ that the optimizer executes during training. We go back to the generalized dynamical system formulation in equation (♠), and ask:

> Given the update direction $\boldsymbol{u}$ that the optimizer $(\boldsymbol{P}, Q)$ actually takes, where is the local quadratic stability boundary along this particular ray?

To answer this question, fix the iteration index t and consider a ray in parameter space from the current parameter $\boldsymbol{\theta}$ along the update direction $\boldsymbol{u}$, and define the *secant Hessian* along the ray as:

$$\boldsymbol{\theta}(s) \;:=\; \boldsymbol{\theta} - s\boldsymbol{u}, \qquad s \geq 0, \qquad \bar{\boldsymbol{H}}_{s,\boldsymbol{u}} \;:=\; 2\int_0^1 (1-r)\,\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta} - rs\boldsymbol{u})dr. \tag{8}$$

Letting $\phi(s) := \mathcal{L}(\boldsymbol{\theta}(s)) = \mathcal{L}(\boldsymbol{\theta} - s\boldsymbol{u})$, we get its Taylor expansion at $s = \eta$ as

$$\phi(\eta) - \phi(0) \;=\; -\eta \boldsymbol{g}^\top \boldsymbol{u} + \frac{\eta^2}{2}\boldsymbol{u}^\top \bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u} \;=\; -sA + \frac{s^2}{2}B, \tag{9}$$

where we define the *alignment score* $A := \boldsymbol{g}^\top \boldsymbol{u}$ and the *curvature load* $B := \frac{\eta}{2}\boldsymbol{u}^\top \bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}$. Then, the realized update yields the exact expression

$$\Delta\mathcal{L} := \mathcal{L}(\boldsymbol{\theta} - \eta\boldsymbol{u}) - \mathcal{L}(\boldsymbol{\theta}) = -\eta A + \eta B, \quad \text{and} \quad \zeta := \frac{B}{A}. \tag{♢}$$

We define the **realized edge of stability** ζ as the ratio of the curvature load to the alignment score. It is helpful to notice that, by the chain rule, the alignment score is *exactly* the learning rate-compensated loss drop rate for the full-batch update (Lee & Lee, 2025):

$$A = \boldsymbol{g}^\top \boldsymbol{u} = \nabla_{\boldsymbol{\theta}}\mathcal{L}^\top \left(-\frac{\mathrm{d}\boldsymbol{\theta}}{\eta \mathrm{d}t}\right) = -\frac{1}{\eta}\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t} = -\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}s}. \tag{10}$$

In other words, the alignment score A measures the actual *learning work* done by the optimizer to decrease the loss along the update direction $\boldsymbol{u}$, with a time granularity (sampling interval) $\eta \mathrm{d}t = \mathrm{d}s$. Similarly, the curvature load B measures the reduction in the loss drop rate due to the curvature of the loss landscape along the update direction $\boldsymbol{u}$:

$$B = \frac{\eta}{2}\boldsymbol{u}^\top \bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u} = A + \frac{\mathcal{L}(\boldsymbol{\theta} - \eta\boldsymbol{u}) - \mathcal{L}(\boldsymbol{\theta})}{\eta} = A + \frac{\Delta\mathcal{L}}{\eta} = -\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}s} - \left(-\frac{\Delta\mathcal{L}}{\Delta s}\right), \tag{11}$$

where $\Delta s = \eta\Delta t = \eta$ (since $\Delta t = 1$ for the iteration indices). Borrowing a physics analogy, if we casually refer to the loss drop rate A as the *learning power*, we can call the curvature load B the *learning power dissipation* by the loss landscape along the update direction $\boldsymbol{u}$. Therefore, ζ is a unit-free ratio of the learning power dissipation to the learning power output of the optimizer $(\boldsymbol{P}, Q)$. In this sense, we may call the realized EoS ζ the **learning impedance** of the task.

To better understand the relationship between the realized ζ and the relative edge of stability χ, let us define a few more quantities and decompose the realized EoS ζ into factors involving χ. The directional smoothness from previous works (Lee & Jang, 2023; Mishkin et al., 2024; Islamov et al., 2026), $D^{\|\cdot\|}$, is a normalized curvature load, written in our notation as $D^{\|\cdot\|_{\boldsymbol{P}}} := 2B/(\eta\|\boldsymbol{u}\|_{\boldsymbol{P}}^2)$. Recall that the maximum sharpness S in (♣) is defined as $S = \lambda_{\max}(\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2}) = \max_{\boldsymbol{u}\neq 0} \boldsymbol{u}^\top \boldsymbol{H}\boldsymbol{u}/\|\boldsymbol{u}\|_{\boldsymbol{P}}^2$. If the Hessian $\boldsymbol{H}$ remains almost constant along the ray, then $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} \simeq \boldsymbol{H}$ implies $S \simeq \max_{\boldsymbol{u}\neq 0} D^{\|\cdot\|_{\boldsymbol{P}}}$. Hence, $D^{\|\cdot\|_{\boldsymbol{P}}} \leq S$. Therefore, we can define a *spatial participation factor* $\varsigma := D^{\|\cdot\|_{\boldsymbol{P}}}/S$, a unit-free measure of how much of the maximally available sharpness is actuated by the curvature along the update direction $\boldsymbol{u}$ taken by the optimizer. Its value becomes 1 when we take the worst-case direction $\boldsymbol{u} = \mathrm{span}(\boldsymbol{v}_{\max})$, and is lower when we take other directions.

Also, recall that Γ in (♣) is the gain of the gradient filter Q at the escape frequency of its root locus. In other words, at this escape frequency, the optimizer applies the update $\boldsymbol{u} = \boldsymbol{P}^{-1}(\Gamma\boldsymbol{g})$. Equivalently, $\boldsymbol{g} = \Gamma^{-1}\boldsymbol{P}\boldsymbol{u}$. If the update realizes the escape mode of Proposition 2.1, then $\boldsymbol{u} = \boldsymbol{P}^{-1}(\Gamma\boldsymbol{g})$ and the learning power takes the form $A = (\Gamma^{-1}\boldsymbol{P}\boldsymbol{u})^\top\boldsymbol{u} = \|\boldsymbol{u}\|_{\boldsymbol{P}}^2/\Gamma$. Therefore, we can define a *temporal calibration factor* $\tau := A/(\|\boldsymbol{u}\|_{\boldsymbol{P}}^2/\Gamma) = \Gamma A/\|\boldsymbol{u}\|_{\boldsymbol{P}}^2$, a unit-free measure of how closely the actual learning power agrees with the power predicted by that escape mode. Its value is 1 when the filtered stream is in phase with $\Gamma\boldsymbol{g}$ in the $\boldsymbol{P}$-inner product, and is identically 1 for vanilla GD ($\Gamma = 1$, $\boldsymbol{u} = \boldsymbol{g}$).

Using these factors, we can now *exactly* decompose the realized EoS ζ into the following factors:

$$\zeta = \frac{\eta S\Gamma}{2} \cdot \frac{\boldsymbol{u}^\top \bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}}{S\|\boldsymbol{u}\|_{\boldsymbol{P}}^2} \cdot \frac{\|\boldsymbol{u}\|_{\boldsymbol{P}}^2}{\Gamma\boldsymbol{g}^\top\boldsymbol{u}} = \chi\frac{\varsigma}{\tau} = (\text{relative EoS}) \times \frac{(\text{spatial participation})}{(\text{temporal calibration})}, \tag{♡}$$

Therefore $\zeta = \chi$ if and only if $\varsigma = \tau$. The factors ς and τ account for the spatial and temporal discrepancies between the actual update and the binding mode of the available certificate, and their equivalence is not trivially guaranteed. Refer to Corollary 5.1 and Appendix B.5 for the exact condition and the GD reduction $\tau \equiv 1$.

5 Coincidence Theorem

We now delve into the implications of the new formulation (♡) in more detail. We are interested in the conditions under which the two values ζ and χ are equal and when they become different. The following corollary is the exact content of (♡).

Corollary 5.1 (Coincidence). [proof] *Fix an iteration and assume* $\eta, S, \Gamma > 0$, $\boldsymbol{u} \neq \mathbf{0}$, *and* $A = \boldsymbol{g}^\top \boldsymbol{u} \neq 0$. *Then (♡) is an identity, and*

$$\zeta = \chi \iff \varsigma = \tau \iff \boldsymbol{u}^\top \tilde{\boldsymbol{H}}_{\eta,\boldsymbol{u}} \boldsymbol{u} = S\Gamma\, \boldsymbol{g}^\top \boldsymbol{u}. \tag{12}$$

The separate conditions $\varsigma = 1$ *and* $\tau = 1$ *are sufficient for coincidence and not necessary: they impose two scalar constraints where (12) imposes one. In particular, vanilla GD (*$\boldsymbol{P} = I$, $Q \equiv 1$*) has* $\Gamma = 1$ *and* $\tau \equiv 1$ *at every step, so* $\zeta = \chi\varsigma$ *and coincidence reduces to occupation* $\varsigma = 1$. *At the realized edge* $\zeta = 1$ *one must have* $\chi = \tau/\varsigma$.

The proof is the cancellation in Appendix B.5. In plain language, the original edge χ equals the realized edge ζ exactly when spatial participation ς matches temporal calibration τ. For GD, since its τ is fixed to unity, the systematic violation of $\chi > 1$ immediately translates to $\varsigma < 1$. We observe that Adam, AdamW, and Heavy Ball in Section 3 have both χ and ζ near 1. This translates to $\varsigma \approx \tau$.

The following corollary isolates the one special case in which coincidence is structural rather than accidental, namely $\varsigma = \tau = 1$ rather than a mere cancellation $\varsigma = \tau$.

Corollary 5.2 (Coincidence on a mode-locked quadratic system). [proof] *Assume the following over the memory horizon of the filter* Q. *(Q1) The loss is quadratic over the memory horizon:* $\boldsymbol{g}_s = \boldsymbol{H}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_\star)$ *and therefore* $\tilde{\boldsymbol{H}}_{\eta,\boldsymbol{u}} = \boldsymbol{H}$. *(Q2) The preconditioner is frozen:* $\boldsymbol{P}_s \equiv \boldsymbol{P} \succ 0$. *(Q3) The trajectory is locked to the sharpest preconditioned eigenmode:* $\boldsymbol{\theta}_s - \boldsymbol{\theta}_\star = x_s \boldsymbol{P}^{-1/2} \boldsymbol{v}_{\max}$ *for scalars* x_s, *where* $x_t \neq 0$. *(Q4) The root locus of* Q *escapes at the Nyquist frequency:* $\Gamma = Q(-1)$, *and the locked mode is in period-2 oscillation,* $x_{t-k} = (-1)^k x_t$. *Then the alignment score is automatically nondegenerate,* $A = \Gamma S^2 x_t^2 > 0$, *and* $\varsigma = \tau = 1$, *and therefore the realized EoS* ζ *is exactly the relative edge of stability* χ: $\zeta = \chi = \eta S Q(-1)/2$. *Dropping the period-2 part of (Q4), even while keeping* $\Gamma = Q(-1)$, *preserves* $\varsigma = 1$ *but gives*

$$\tau = \frac{\Gamma x_t}{(Q * x_{\leq t})_t}, \tag{13}$$

which equals 1 only when the filtered stream is in phase with the gradient at the escape gain.

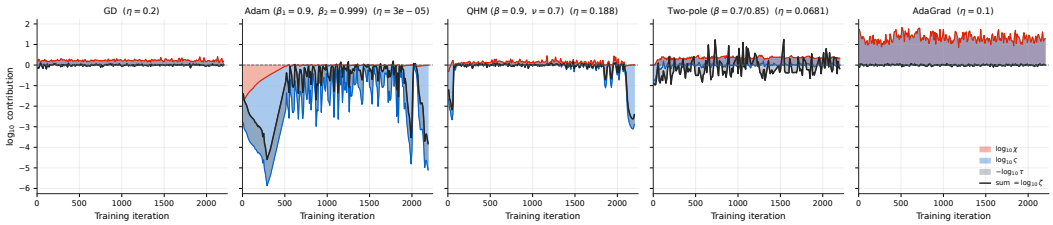
The proof is provided in Appendix B.6. Corollary 5.2 reveals assumptions that are used to derive the original boundary $\chi \leq 1$ in previous works (Cohen et al., 2021; 2023). In this worst-mode-locked quadratic regime, our realized EoS $\zeta \leq 1$ of (♢) is exactly the original EoS $\chi \leq 1$ of (♣). Our formulation generalizes the previous works in this precise sense.

Each of the assumptions (Q1)–(Q4) fails in an identifiable way, and the observations of Section 3 split across those failure types. The decomposition of (♡), therefore, not only serves as an evidence that ζ generalizes χ, but also a useful diagnostic toolkit for analyzing the edge of stability: (Q1) fails whenever the third-order term along the ray is non-negligible, so that $\tilde{\boldsymbol{H}}_{\eta,\boldsymbol{u}} \neq \boldsymbol{H}$; (Q2) fails for adaptive preconditioners, whose $\boldsymbol{P}_t$ drifts over the memory horizon of Q; (Q3) fails whenever the update actuates a non-maximal direction, *including* vanilla GD, where $\boldsymbol{u} = \boldsymbol{g}$; and (Q4) fails for any filter whose root locus escapes away from $z = -1$, or whose locked mode is not in period-2 oscillation. We can categorize the observation results in Section 3 into the following cases: GD is (Q3) with $\tau \equiv 1$ ($\chi = 1.124 > 1$, $\zeta = 0.983 \approx 1$); Adam, AdamW, and Heavy Ball are near coincidence after Γ-scaling (χ at 0.960, 0.971, and 0.949 against ζ at 0.964, 0.919, and 0.983, respectively); AdaGrad is (Q2)–(Q3) ($\chi = 21.070 \gg 1$, yet $\zeta = 1.002 \approx 1$); QHM mixes a moderate χ offset with $\zeta = 1.000$; and the two-pole cascades are (Q4), with $\chi = 2.646 > 1$ and no ζ reaches the edge. This identifies the actual causes of the offsets as elaborated in Section 6.

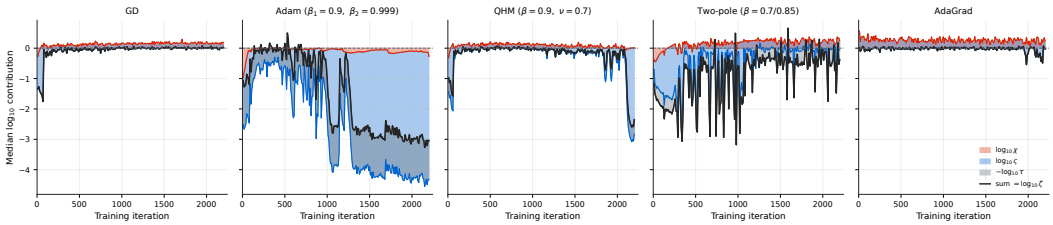
We now record what Corollary 5.1 implies for the available stability limit: χ determines ζ only through the ratio ς/τ, which the gauges (S, Γ) do not control.

Theorem 5.3 (Insufficiency of χ for the realized edge). [proof] *Assume* $\eta, S, \Gamma > 0$, $\boldsymbol{u} \neq \mathbf{0}$, *and* $A \neq 0$. *By Corollary 5.1,* $\zeta = \chi$ *if and only if* $\varsigma = \tau$. *Consequently* (η, S, Γ) *determine* ζ *if and only if they determine the ratio* ς/τ. *For vanilla GD (*$\boldsymbol{P} = I$, $Q \equiv 1$*) one has* $\tau \equiv 1$ *and* $\zeta = \chi\varsigma$, *so* χ *determines* ζ *if and only if occupation* ς *is a function of* (η, S) *alone. This already fails on any frozen quadratic in dimension at least* 2 *whose gradient is not an eigen-direction of* $\boldsymbol{H}$: *then* ς *is the Rayleigh ratio* $\boldsymbol{g}^\top \boldsymbol{H} \boldsymbol{g}/(S\|\boldsymbol{g}\|^2)$ *and is not determined by* $S = \lambda_{\max}(\boldsymbol{H})$.

The proof of Theorem 5.3 is the identity of Corollary 5.1 together with the GD reduction $\tau \equiv 1$, detailed in Appendix B.8. It is important to mention that this result does not falsify previous works'



(a) Decomposition of the realized edge of stability ζ from a single experiment.



(b) Decomposition of optimizer-family-wise contributions to ζ, median from an LR sweep of size 5.

Figure 4: **EoS stack plot.** Decomposition of the realized EoS ζ into factors in log-scale, involving the (relative) edge of stability χ (red area and line), the spatial participation factor ς (blue area and line), and the temporal calibration factor τ (grey overlay and black line). Note that $\log \tau$ operates in the opposite direction by subtraction.

contributions on formulating the EoS (Cohen et al., 2021; 2023), but rather we circumscribe the scope of their claims to the case where the update direction is fixed to the fixed maximum curvature, which is condition (Q3) of Corollary 5.2. Violation of this condition implies that the optimizer is more than just a prescribed algorithmic procedure, but rather an active decision-maker that chooses the edge it wants to occupy among the available options provided by the current loss landscape.

6 Discussion

Previous formulations of the edge of stability (Cohen et al., 2021; 2023; Islamov et al., 2026) and its generalized version in Proposition 2.1 predict available worst-case marginality χ. Our realized EoS ζ reveals whether and how this available margin is actually trod upon by the optimizer. This distinction offers a new way to understand the training dynamics of different optimizer families.

6.1 Diagnostic Applications

If we take the logarithm from ($\heartsuit$), we can decompose the realized EoS ζ into additive factors:

$$\log \zeta = \log \chi + \log \varsigma - \log \tau. \tag{14}$$

This simple additive relationship holds for every iteration, allowing us to draw an **EoS stack plot**, as in Figure 4, to visualize how each factor contributes to the realized EoS ζ during the training process. We can use this stack plot as a new diagnostic tool to study training dynamics online, for various instances and families of optimizers.

The first thing we notice from the EoS stack plot in Figure 4 is that the spatial participation factor ς compensates for χ even without a nontrivial preconditioner. Vanilla GD has $\tau \equiv 1$, so the stack is $\log \zeta = \log \chi + \log \varsigma$. The compensation from $\chi = 1.124$ to $\zeta = 0.983$ is solely due to $\varsigma < 1$. The same happens to AdaGrad's extreme offset $\chi \approx 21.1$, which is mostly compensated by ς, resulting in $\zeta \approx 1$. In other words, the optimizer often selects updates $\boldsymbol{u}$ in nonmaximal sharpness directions, and GD already does so. This explains why the previous edge of stability formulation based on the worst-mode update $\boldsymbol{u}_{\max}$ was not sufficient to predict the actual EoS ζ. Another interesting observation is that the temporal calibration factor τ can act in either direction, depending on the type of gradient filter Q. For Adam and QHM, $\log \tau$ is mostly negative, but for more complicated filters like two-pole filters, $\log \tau$ can be positive, pulling the dynamics away from the edge of stability. Adam's near-coincidence $\chi \approx \zeta \approx 1$ is therefore $\varsigma \approx \tau$ after Γ is included in χ, not a restoration of occupation $\varsigma = 1$. This single observation already explains the main cause of the family-dependent offsets in the edge of stability χ and allows us to visualize how different optimizer families compensate for this effect and actively shape their training dynamics online.

Table 1: **Computation requirements.** For one full-batch optimizer step (Adam, params=777k, batch=50k, 8-step Lanczos, 1× RTX 5090). ζ requires no HVPs, reducing computation time by 95.3%.

Component	Calculation formula	HVPs	Time (ms)	Peak mem. (MiB)
$\chi = \eta S\Gamma/2$	k-step Lanczos sharpness	8	110.15 ± 2.56	767.3
$\varsigma = 2B/(\eta S\|\boldsymbol{u}\|_{\boldsymbol{P}}^2)$	k-step Lanczos sharpness	8	114.84 ± 1.15	767.3
$\tau = \Gamma A/\|\boldsymbol{u}\|_{\boldsymbol{P}}^2$	gradient contractions	0	5.22 ± 0.00	305.4
$\zeta = B/A$	gradient contractions	0	5.20 ± 0.01	305.4

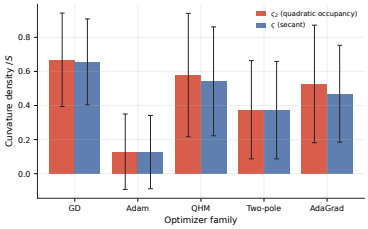


Figure 5: ς vs. ς_2 for optimizers.

6.2 Cost Benefits

One additional benefit of realized EoS ζ is that it does not require heavy calculation. Since $\zeta = \frac{B}{A} = \frac{A+\Delta\mathcal{L}/\eta}{A} = 1 + \frac{\Delta\mathcal{L}}{\eta \boldsymbol{g}^\top \boldsymbol{u}}$, we can compute ζ directly from the current gradient $\boldsymbol{g}$, the chosen update direction $\boldsymbol{u}$, and the resulting change in loss $\Delta\mathcal{L}$ at each training iteration. In comparison, previous definition of the EoS involves a maximizer on preconditioned Hessian (Cohen et al., 2021; 2023; Islamov et al., 2026), which is not only computationally expensive, requiring multiple Hessian-vector products (HVPs), but also forces us to use approximate measures. Table 1 summarizes the computation costs for each component of (14), all measured from the same CIFAR-10 training using the 200-5-tanh model in the initial report (Cohen et al., 2021). We highlight that our formulation is both cost-effective and *exact* for full-batch optimization.

6.3 Effect of Higher-Order Terms

Similar to the original edge of stability (Cohen et al., 2021; 2023; Islamov et al., 2026), Figure 1 shows how actual training trajectories are attracted to, oscillate around, and are eventually pushed away from the edge of stability $\zeta = 1$. As in previous works on EoS (Damian et al., 2023; Mulayoff & Stich, 2026), we can derive sufficient conditions for the stabilization mechanism to take effect. Since our main focus is to resolve the optimizer-dependent offsets found in the empirical EoS, we defer this analysis to Appendix D. In our formulation of ζ, the higher-order terms are absorbed into the weighted secant Hessian $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}$ and thus into the spatial participation factor ς. If we want to isolate the effect of the higher-order terms, we can define the second-order-only spatial participation factor $\varsigma_2 := \boldsymbol{u}^\top \boldsymbol{H}\boldsymbol{u}/(S\|\boldsymbol{u}\|_{\boldsymbol{P}}^2)$ and compare it with ς, as shown in Figure 5. Although higher-order terms are essential for maintaining stability beyond the edge of stability, their contribution to the magnitudes of the realized EoS ζ is small in the studied settings.

6.4 Limitations

This work focuses on full-batch optimization, where the theory is exactly realized in practice. In mini-batch optimization, the terms $A = \boldsymbol{g}^\top \boldsymbol{u}$, $B = \eta \boldsymbol{u}^\top \bar{\boldsymbol{H}}\boldsymbol{u}/2$, and their ratio $\zeta = B/A$ become stochastic due to changes in the instantaneous loss landscape $\mathcal{L}$ along the trajectory. This makes the analysis more challenging and complicates the interpretation of the ratio ζ. Previous works treat the incorporation of stochastic scenarios as an independent problem, which is sometimes referred to as the edge of stochastic stability (EoSS) Andreyev & Beneventano (2024); Andreyev et al. (2026). We likewise leave the extension of our framework to the stochastic setting for future work.

7 Conclusion

The edge of stability has been studied primarily as an effect related to the maximum curvature mode of the loss landscape. This turns out to be only half of the story. The other half is the effect of time-varying stateful optimizers that actually take the updates, given a loss landscape and gradient history. Neglecting this aspect leads to optimizer-dependent systematic offsets in estimating the edge of stability, which are largely eliminated in our framework on the *realized* edge of stability. Decomposing the realized EoS into factors results in diagnostic tools that provide a richer understanding of training dynamics under gradient-based optimizers.

AI USE STATEMENT

In this work, we used AI assistants solely to assist with writing code and in assisting with mathematical proofs. All mathematical claims or proofs augmented with AI assistance were thoroughly reviewed, verified, and rewritten as needed by the human authors to ensure correctness and rigor. We take full responsibility for the final content of this work, including all text, code, and mathematical results.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that our results are reproducible. Sufficient explanations of our methods and setups are provided in Appendix E. All figures presented in the paper can be fully reproduced using the code and data included in the supplementary material.

REFERENCES

Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. In *ICML*, volume 202, pp. 169–195, 2023.

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *ICML*, volume 162, 2022.

Kwangjun Ahn, Sebastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *NeurIPS*, 2023.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for SGD. *arXiv preprint arXiv:2412.20553*, 2024.

Arseniy Andreyev, Advikar Ananthkumar, Marc Walden, Tomaso Poggio, and Pierfrancesco Beneventano. Momentum further constrains sharpness at the edge of stochastic stability. In *ICML*, 2026.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *ICML*, volume 162, pp. 948–1024, 2022.

Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.

Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *IJCAI*, 2020.

Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *ICML*, volume 202, pp. 4330–4391, 2023.

Xuxing Chen, Krishnakumar Balasubramanian, Promit Ghosal, and Bhavya Agrawalla. From stability to chaos: Analyzing gradient descent dynamics in quadratic regression. *Transactions on Machine Learning Research*, 2024.

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 *ICLR*, 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. In *NeurIPS 2023 Workshop on Heavy Tails in Machine Learning: Structure, Stability, and Dynamics*, 2023.

Jeremy M. Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *ICLR*, 2025.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *ICLR*, 2023.

Timothy Dozat. Incorporating nesterov momentum into adam. In *Proceedings of the 4th International Conference on Learning Representations (ICLR) Workshop*, 2016.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 12(61):2121–2159, 2011.

Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. In *NeurIPS*, 2023.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep matrix factorization beyond the edge of stability. In *ICLR*, 2025.

Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instabilities of deep learning models. In *ICLR*, 2022.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *NeurIPS*, 2019.

Thomas Hofmann. The map behind the flow: Finite-step gradient descent as a dynamical system. *arXiv preprint arXiv:2607.04993*, 2026.

Bin Hu and Laurent Lessard. Control interpretations for first-order optimization methods. In *ACC*, pp. 3114–3119, 2017.

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

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *ICLR*, 2019.

Stanisław Jastrzębski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. In *ICLR*, 2020.

Kaiqi Jiang, Jeremy Cohen, and Yuanzhi Li. Understanding the evolution of the neural tangent kernel at the edge of stability. In *NeurIPS*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In *ICML*, volume 202, pp. 17684–17744, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Jaerin Lee and Kyoung Mu Lee. Greedy alignment principle for optimizer selection. *arXiv preprint arXiv:2512.06370*, 2025.

Jaerin Lee, Bong Gyun Kang, Kihoon Kim, and Kyoung Mu Lee. Grokfast: Accelerated grokking by amplifying slow gradients. *arXiv preprint arXiv:2405.20233*, 2024.

Sungyoon Lee and Cheongjae Jang. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *ICLR*, 2023.

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.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the performance analysis of momentum method: A frequency domain perspective. In *ICLR*, 2025.

Elon Litman. The origin of edge of stability. *arXiv preprint arXiv:2604.20446*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *NeurIPS*, 2022.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 1(3):247–267, 2022.

Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and adam for deep learning. *ICLR*, 2019.

Pierre Marion. Edge flow: A tractable and predictive continuous-time model for gradient descent at the edge of stability. *arXiv preprint arXiv:2606.18080*, 2026.

Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert M. Gower. Directional smoothness and gradient methods: Convergence and adaptivity. In *NeurIPS*, 2024.

Rotem Mulayoff and Sebastian U. Stich. On the stability of nonlinear dynamics in GD and SGD: Beyond quadratic potentials. In *COLT*, volume 336, pp. 5210–5243, 2026.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983. English translation: Soviet Mathematics Doklady, 27(2):372–376, 1983.

Katsuhiko Ogata. *Discrete-Time Control Systems*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1995.

James M Ortega and Werner C Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York-London, 1970.

Matteo Pagliardini, Pierre Ablin, and David Grangier. The AdEMAMix optimizer: Better, faster, older. In *ICLR*, 2025.

Boris 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.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *ICLR*, 2018.

Eric Regis and Sinho Chewi. A rod flow model for Adam at the edge of stability. *arXiv preprint arXiv:2605.06821*, 2026a.

Eric Regis and Sinho Chewi. Rod flow: A continuous-time model for gradient descent at the edge of stability. *arXiv preprint arXiv:2602.01480*, 2026b.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *ICML*, 2018.

Minhak Song and Chulhee Yun. Trajectory alignment: Understanding the edge of stability phenomenon via bifurcation theory. In *NeurIPS*, 2023.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5—-RMSProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

Bryan Van Scoy, Randy A. Freeman, and Kevin M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1): 49–54, 2018.

Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. *Journal of Machine Learning Research*, 26(273):1–68, 2025.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along GD trajectory: Progressive sharpening and edge of stability. In *NeurIPS*, 2022.

Jingfeng Wu, Vladimir Braverman, and Jason D. Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. In *NeurIPS*, 2023.

Jingfeng Wu, Peter L. Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: Non-monotonicity of the loss improves optimization efficiency. In *COLT*, volume 247, pp. 5019–5073, 2024.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with SGD. *arXiv preprint arXiv:1802.08770*, 2018.

Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning. In *ICML*, volume 235, pp. 62476–62509, 2024.

Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *ICLR*, 2023.

A Related Work

Edge of stability. It has been known that the learning rate constrains the curvature of the loss landscape during the course of training (Xing et al., 2018; Jastrzębski et al., 2019; Gilmer et al., 2022). Later, Cohen et al. (2021) conducted an extensive study on this phenomenon and first coined the term *edge of stability* to describe the attractive behavior of the maximum Hessian eigenvalue $\lambda_{\max}(\boldsymbol{H})$ towards a value reciprocal to the learning rate $\lambda_{\max} \approx 2/\eta$ under full-batch gradient descent. Similar phenomena have been reported under different optimization settings (Lyu et al., 2022; Cohen et al., 2023; Damian et al., 2023; Agarwala et al., 2023; Zhu et al., 2023; Andreyev & Beneventano, 2024; Chen et al., 2024; Cohen et al., 2025; Islamov et al., 2026; Andreyev et al., 2026; Litman, 2026). This includes full gradient descent (Cohen et al., 2021), mini-batch gradient descent (Andreyev & Beneventano, 2024), momentum-based optimizers (Cohen et al., 2023; Andreyev et al., 2026), and adaptive method (Cohen et al., 2023) such as RMSProp (Tieleman & Hinton, 2012) or Adam (Kingma & Ba, 2015). Following these lines of work, we primarily focus on the full-batch settings where the analysis is more tractable and the effect is more pronounced. In case of stochastic mini-batch gradient descent, Andreyev & Beneventano (2024); Andreyev et al. (2026) empirically found that the edge of stability extends beyond determinism and applies to curvature of the batch statistics.

Mechanisms: catapult, progressive sharpening, self-stabilization, and bifurcation. Existing explanations for the EoS phenomenon can be categorized into four mechanisms: (1) The *catapult* (Lewkowycz et al., 2020), described prior to the formalization of the edge of stability, explains how a large learning rate drives a sudden loss spike that initiates the reduction of the curvature down to the edge of stability. (2) Cohen et al. (2021); Wang et al. (2022) demonstrate the opposite, called *progressive sharpening*, where curvature initially increases up to the EoS bound. The EoS phenomenon also implies the existence of a coincidental stabilization mechanism that consistently reduces the curvature to the edge, causing it to hover just over the theoretical stability bound. (3) Damian et al. (2023) attribute this *self-stabilization* to curvature compensation by third-order derivatives. (4) Jastrzębski et al. (2020); Song & Yun (2023) identify a period-2 *flip bifurcation* in the discrete dynamical system governed by gradient descent. These four mechanisms form the basis of the EoS phenomenon, and many follow-up works have sought to explain them using simpler models. Arora et al. (2022) use a two-timescale decomposition; Ma et al. (2022) employ a one-dimensional model; Chen & Bruna (2023) analyze two-step fixed points that extract period-2 motions; Kreisler et al. (2023) study scalar linear networks; and Hofmann (2026) use a simple quartic model. Continuous-time reductions are also explored, including central flows (Cohen et al., 2025), Edge Flow (Marion, 2026), and Rod Flow (Regis & Chewi, 2026b;a). In structured models, EoS is used as an implicit-bias mechanism for diagonal linear networks (Even et al., 2023), logistic regression (Wu et al., 2023; 2024), threshold neurons (Ahn et al., 2023), and deep matrix factorization (Ghosh et al., 2025); Wang et al. (2025) unify EoS, balancing, and catapult as large-learning-rate

biases resulting from regularity, and Jiang et al. (2025) track the NTK at EoS. None of these works reformulate the dependence of the EoS $\chi = \eta S\Gamma/2$ on the maximum sharpness direction $\boldsymbol{v}_{\max}$ to incorporate actual executed updates carried out *by the optimizers*, which is the gap we address.

Directional smoothness. We are mainly inspired by prior work examining how directional smoothness influences the stability bound of full-batch gradient descent. Ahn et al. (2022) formalize unstable convergence in terms of directional smoothness and relative progress, finding structured oscillations rather than divergence past $2/\eta$. Lee & Jang (2023) replace $\lambda_{\max}$ with an interaction-aware sharpness (batch-gradient–Hessian coupling), and report that this quantity hovers around a concentration measure under mini-batch SGD. Mishkin et al. (2024); Islamov et al. (2026) investigate the directional smoothness of the executed step as a convergence diagnostic. Their works are analogous to our realized EoS $\zeta = B/A$ along an update direction $\boldsymbol{u}$; however, their quantities are not directly representative of the edge of stability, as they resort to the maximum sharpness instead of the executed update. Furthermore, previous works have not accounted for general filtered optimizers. Therefore, our work extends these works rather than competing with them.

Gradient-based optimizers as filters. Momentum methods (Polyak, 1964; Nesterov, 1983) are linear time-invariant (LTI) filters applied to the gradient stream. Qian (1999) identified Heavy Ball (Polyak, 1964) as a discrete-time low-pass filter. Control systems theory makes this viewpoint precise: first-order methods are modeled as LTI plants in feedback with gradient stream signals $\boldsymbol{g}$. These systems can be analyzed using transfer functions, Bode plots, and integral quadratic constraints (Lessard et al., 2016; Hu & Lessard, 2017), and can be designed as higher-order filters such as triple momentum (Van Scoy et al., 2018). Gitman et al. (2019) treat Heavy Ball (Polyak, 1964), Nesterov's momentum (Nesterov, 1983), and quasi-hyperbolic momentum (Ma & Yarats, 2019) as a single IIR family, and map stability regions in (β, ν). Lee et al. (2024) design a two-timescale EMA filter with preconditioner (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). Li et al. (2025) apply the z-transform to momentum methods and interpret the coefficients as a (time-varying) frequency response. Lee & Lee (2025) proposes an adaptive technique to determine the optimal filter coefficients, e.g., the momentum hyperparameter β, further developing this signal processing perspective. In summary, momentum-based optimizers can be understood as filters: GD acts as an identity filter, Heavy Ball and Nesterov's momentum are one-pole filters, QHM is a one-tap pole-zero filter, and two-timescale EMAs (Ghosh et al., 2025) such as Grokfast (Lee et al., 2024) function as cascaded two-pole filters. The latter serve as practical counterparts of Q, whose root locus does not necessarily escape at the Nyquist frequency $z = -1$. We draw extensively on this perspective to study the EoS phenomenon under general filtered optimizers.

B Proofs

B.1 The Mode Reduction

Throughout this subsection we work in the *frozen quadratic* regime on which the original edge-of-stability arguments are based: over the memory horizon of the filter the loss is $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\star)^\top \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_\star)$ with $\boldsymbol{H}$ constant and symmetric, the preconditioner is frozen at $\boldsymbol{P}_t \equiv \boldsymbol{P} \succ 0$, and the filter is time invariant, $Q_t \equiv Q$. Let $(\lambda_i, \boldsymbol{v}_i)$ be the eigenpairs of $\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2}$ with $\{\boldsymbol{v}_i\}$ orthonormal, and define the *whitened* coordinates

$$x_t^{(i)} \;:=\; \boldsymbol{v}_i^\top \boldsymbol{P}^{1/2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star). \tag{15}$$

Two remarks on (15). The factor $\boldsymbol{P}^{1/2}$ applies before the mode decomposition, and the offset $\boldsymbol{\theta}_\star$ makes the gradient proportional to the coordinate. The shorthand $x_i = \boldsymbol{v}_i \cdot \boldsymbol{\theta}$ used below (5) incorporates both effects for brevity.

Lemma B.1 (Exact decoupling)**.** *Under the frozen quadratic regime, the update rule (♠) is equivalent to the family of decoupled scalar recursions*

$$x_{t+1}^{(i)} \;=\; x_t^{(i)} - \eta\lambda_i\,(Q * x_{\le t}^{(i)})_t, \qquad i = 1, \dots, d, \tag{16}$$

which is (5) with the coordinates (15).

Proof. Since $\boldsymbol{g}_t = \boldsymbol{H}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star)$, inserting $\boldsymbol{P}^{\pm 1/2}$ gives $\boldsymbol{P}^{-1/2}\boldsymbol{g}_t = (\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2})\,\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star)$, so $\boldsymbol{v}_i^\top \boldsymbol{P}^{-1/2}\boldsymbol{g}_t = \lambda_i x_t^{(i)}$. Applying $\boldsymbol{P}^{1/2}$ to $\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_\star = \boldsymbol{\theta}_t - \boldsymbol{\theta}_\star - \eta \boldsymbol{P}^{-1}(Q * \boldsymbol{g}_{\le t})_t$ and using $\boldsymbol{P}^{-1/2}(Q * \boldsymbol{g}_{\le t})_t = (Q * \boldsymbol{P}^{-1/2}\boldsymbol{g}_{\le t})_t$ (which holds because the convolution acts on the time index and $\boldsymbol{P}$ is constant), we obtain $\boldsymbol{P}^{1/2}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_\star) = \boldsymbol{P}^{1/2}(\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star) - \eta(Q * \boldsymbol{P}^{-1/2}\boldsymbol{g}_{\le t})_t$. Projecting onto $\boldsymbol{v}_i$ yields (16). Since no two modes interact, the reduction is exact rather than an approximation. □

From here on we can fix one mode without loss of generality, drop the index, and write the *loop gain*

$$k \;:=\; \eta\lambda \;>\; 0, \tag{17}$$

assuming $\lambda > 0$. Modes with $\lambda \le 0$ do not contribute to unit-circle crossing for $k > 0$, and need not be considered.

B.2 Proof of Proposition 2.1

Proof. We first record the two standing hypotheses that the statement of the proposition leaves implicit, but generallly holds true for practical filters. Write $Q(z)$ in a rational form $Q(z) = b(\xi)/a(\xi)$ with $\xi := z^{-1}$, where a, b are coprime real polynomials normalized so that $a(0) = 1$.

(H1) Q is causal and internally stable: every pole of Q lies in $|z| < 1$, equivalently every root of a lies in $|\xi| > 1$.

(H2) $Q(1)$ is finite and nonzero; by the dc normalization of Section 2, $Q(1) = 1$ always.

Without (H1) the filter is unstable on its own and no gain k stabilizes the mode, so the phrase "first crossing" has no meaning. Without (H2) the branch of the locus that starts at $z = 1$ does not move into the disc.

The characteristic polynomial. Taking z-transforms of (16) with zero initial data gives $zX(z) = X(z) - kQ(z)X(z)$, hence (7),

$$z - 1 + kQ(z) \;=\; 0. \tag{18}$$

Multiplying (18) by $z^{-1}a(z^{-1})$ and substituting $\xi = z^{-1}$ clears both the pole at $z = 0$ and the denominator of Q, yielding a characteristic polynomial

$$c_k(\xi) \;:=\; (1-\xi)\,a(\xi) \;+\; k\,\xi\,b(\xi) \;=\; 0. \tag{19}$$

Because $|z| < 1 \iff |\xi| > 1$, the mode is asymptotically stable exactly when every root of c_k lies strictly outside the closed unit disc. Roots of c_k move continuously with k on the Riemann sphere, and the exterior $\{|\xi| > 1\} \cup \{\infty\}$ is open, so a root can enter the closed unit disc only by passing through the circle $|\xi| = 1$. A root escaping to $\xi = \infty$, which happens when the leading coefficient of (19) degenerates, is a deadbeat pole at $z = 0$ and never destabilizes.

Step 1: the locus starts inside. At $k = 0$ we have $c_0(\xi) = (1-\xi)a(\xi)$, whose roots are $\xi = 1$ together with the roots of a, which all lie in $|\xi| > 1$ by (H1). The root at $\xi = 1$ is simple because $a(1) \neq 0$ by (H2). Since $\partial_\xi c_k|_{(1,0)} = -a(1)$ and $\partial_k c_k|_{(1,0)} = b(1)$, the implicit function theorem gives a branch $\xi(k)$ with

$$\left.\frac{d\xi}{dk}\right|_{k=0} \;=\; \frac{b(1)}{a(1)} \;=\; Q(1) \;=\; 1 \;>\; 0, \qquad \text{so} \qquad \xi(k) = 1 + k + O(k^2). \tag{20}$$

Equivalently $z(k) = 1 - k + O(k^2)$: the dc pole leaves $z = 1$ along the negative real direction and enters the disc. Hence there is $k_0 > 0$ such that the mode is asymptotically stable for every $k \in (0, k_0)$, and the locus starts inside the disc.

Step 2: the crossings are exactly $\Omega_+(Q)$. For $k > 0$, its corresponding characteristic polynomial c_k has a root on $|\xi| = 1$ if and only if (18) is solved by some $z = e^{i\omega}$ on the unit circle, i.e.

$$k \;=\; \frac{1 - e^{i\omega}}{Q(e^{i\omega})} \;=:\; k(\omega). \tag{21}$$

The frequency $\omega = 0$ forces $k = 0$ and is excluded. Because a, b are real, $k(-\omega) = \overline{k(\omega)}$, so the crossings appear symmetrically, i.e., $\pm\omega$ occur at the same gain, and it suffices to take $\omega \in (0, \pi]$. Finally k must be a real positive number, which is precisely the defining condition of $\Omega_+(Q)$ in (4). On $\Omega_+(Q)$ the value $k(\omega)$ coincides with its modulus, i.e., $k(\omega) = |k(\omega)|$.

Step 3: conclusion. Combining Steps 1 and 2, the set of gains at which some root sits on the unit circle is $\{k(\omega) : \omega \in \Omega_+(Q)\}$, and the mode is stable on the whole interval below its smallest element. Therefore

$$k_\star(Q) \;=\; \min_{\omega \in \Omega_+(Q)} \left| \frac{1 - e^{i\omega}}{Q(e^{i\omega})} \right|, \tag{22}$$

with the convention $k_\star(Q) = +\infty$ and $\Gamma = 0$ when $\Omega_+(Q) = \emptyset$, which is exactly the claim (4). Writing $\Gamma := 2/k_\star(Q)$, the mode is stable for all sufficiently small positive $\eta\lambda < 2/\Gamma$, and marginal at equality, i.e. iff $\eta\lambda\Gamma \le 2$, which recovers (♣). Since Γ depends on Q alone and not on λ (mode decoupling), the binding mode is the sharpest one, and the system is not exponentially unstable iff $\eta\lambda_{\max}(\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2})\,\Gamma \le 2$, which is precisely the boxed condition (♣).

Step 4: the Nyquist case. If the minimum in (22) is attained at the Nyquist frequency $\omega = \pi$, i.e., $z = e^{i\omega} = -1$, and $Q(-1) > 0$, then

$$k_\star(Q) \;=\; \frac{1 - (-1)}{Q(-1)} \;=\; \frac{2}{Q(-1)}, \qquad \text{hence} \qquad \Gamma \;=\; \frac{2}{k_\star(Q)} \;=\; Q(-1), \tag{23}$$

and the stability condition reads $\eta\lambda Q(-1) \le 2$. The marginal root $z = -1$ has $z^2 = 1$, so the marginal motion is a period-2 oscillation. This proves the last sentence of Proposition 2.1. □

B.3 Two Lemmas on the Root Locus Gauge

The gauge Γ is a functional of the filter Q alone, so it can be computed once per optimizer family, and is not dependent on the preconditioner. The following two lemmas consolidates this: the first removes the dc normalization $Q(1)$, and the second identifies a large class of filters for which the escape is provably at Nyquist $z = -1$.

Lemma B.2 (Positive scaling). *Let $c > 0$. Then $\Omega_+(cQ) = \Omega_+(Q)$, $k_\star(cQ) = k_\star(Q)/c$, and*

$$\Gamma(cQ) \;=\; c\,\Gamma(Q). \tag{24}$$

In particular the dc-normalized filter $\tilde{Q} := Q/Q(1)$ satisfies $\Gamma(\tilde{Q}) = \Gamma(Q)/Q(1)$, and $\Gamma(Q) = Q(-1)$ holds for Q if and only if it holds for $\tilde{Q}$.

Proof. $k(\omega)$ in (21) is replaced by $k(\omega)/c$. Multiplication by $c^{-1} > 0$ preserves membership in $\mathbb{R}_{>0}$, so Ω_+ is unchanged, and it scales every modulus by c^{-1}, so the minimum scales by c^{-1} and $\Gamma = 2/k_\star$ by c. The last claim follows since Γ and $Q(-1)$ carry the same factor c. □

Lemma B.3 (Nyquist escape for positive one-pole mixtures). *Let*

$$Q(z) \;=\; \sum_{j=1}^{J} w_j \,\frac{1 - \beta_j}{1 - \beta_j z^{-1}}, \qquad w_j > 0, \quad \sum_{j=1}^{J} w_j = 1, \quad \beta_j \in [0, 1). \tag{25}$$

Then $Q(1) = 1$, hypotheses (H1) and (H2) hold, and

$$\Omega_+(Q) = \{\pi\}, \qquad k_\star(Q) = \frac{2}{Q(-1)}, \qquad \Gamma \;=\; Q(-1) \;=\; \sum_{j=1}^{J} w_j \,\frac{1 - \beta_j}{1 + \beta_j} \;\in\; (0, 1], \tag{26}$$

with $\Gamma = 1$ if and only if every $\beta_j = 0$.

Proof. Each summand equals w_j at $z = 1$, so $Q(1) = 1$, giving (H2). Also, the poles are $z = \beta_j \in [0, 1)$, giving (H1). For $\omega \in (0, \pi]$ the elementary identity

$$1 - e^{i\omega} \;=\; 2\sin\left(\frac{\omega}{2}\right)\, e^{i(\omega - \pi)/2}, \qquad 2\sin\left(\frac{\omega}{2}\right) > 0, \tag{27}$$

shows that $k(\omega) \in \mathbb{R}_{>0}$ holds iff $e^{i\theta}/Q(e^{i\omega}) \in \mathbb{R}_{>0}$ with $\theta := (\omega - \pi)/2$. Writing $e^{i\theta}/Q = e^{i\theta}\overline{Q}/|Q|^2$ and take the conjugate. The reality requirement $e^{i\theta}/Q(e^{i\omega}) \in \mathbb{R}_{>0}$ is exactly the vanishing of the *phase margin*

$$\Psi(\omega) := \operatorname{Im}\Big(e^{-i(\omega-\pi)/2}\, Q(e^{i\omega})\Big). \tag{28}$$

So it suffices to show that $\Psi > 0$ on $(0, \pi)$, which rules out every interior crossing.

Consider one pole, $Q_\beta(z) := (1-\beta)/(1-\beta z^{-1})$, so that $Q_\beta(e^{i\omega}) = (1-\beta)/(1-\beta e^{-i\omega})$. Rationalizing gives

$$\Psi_\beta(\omega) = \frac{1-\beta}{|1-\beta e^{-i\omega}|^2} \operatorname{Im}\big(e^{-i\theta}\,(1-\beta e^{i\omega})\big) = \frac{1-\beta}{|1-\beta e^{-i\omega}|^2}\,(-\sin\theta - \beta\sin(\omega-\theta)). \tag{29}$$

Since $-\theta = \frac{\pi-\omega}{2}$ and $\omega - \theta = \frac{\omega+\pi}{2}$, both sines collapse onto the same value, $-\sin\theta = \cos\frac{\omega}{2} = \sin(\omega-\theta)$, and we have

$$\Psi_\beta(\omega) = \frac{(1-\beta)^2\,\cos(\omega/2)}{1-2\beta\cos\omega+\beta^2} > 0 \qquad \text{for } \omega \in (0,\pi), \tag{30}$$

because $\beta < 1$ and $\cos(\omega/2) > 0$ for $\omega \in (0,\pi)$. Also, the denominator $|1-\beta e^{-i\omega}|^2 \geq (1-\beta)^2 > 0$. The single pole therefore clears the critical line strictly, and it does so by a margin proportional to $(1-\beta)^2$: the closer β is to 1, the closer the filter comes to admitting an interior crossing.

Since Ψ is $\mathbb{R}$-linear in Q, the mixture inherits the sign: $\Psi(\omega) = \sum_j w_j \Psi_{\beta_j}(\omega) > 0$ on $(0,\pi)$, as every $w_j > 0$. Hence $\Omega_+(Q) \cap (0,\pi) = \emptyset$. At $\omega = \pi$ we have $\theta = 0$ and $Q(-1) = \sum_j w_j(1-\beta_j)/(1+\beta_j) > 0$ is real, so $\Psi(\pi) = 0$ and $k(\pi) = 2/Q(-1) \in \mathbb{R}_{>0}$, i.e. $\pi \in \Omega_+(Q)$. Thus $\Omega_+(Q) = \{\pi\}$ and (26) follows from Step 4 of Appendix B.2. Finally $(1-\beta)/(1+\beta) \in (0,1]$ with equality iff $\beta = 0$, and a convex combination of such numbers obeys the same bounds. □

Remark B.4 (Root locus on various filters)**.** *The vast majority of existing gradient-based optimizers rely on one-pole momentum, either in Heavy Ball (Polyak, 1964) or Nesterov's Accelerated Gradient (Nesterov, 1983). As the following section shows, these are the main examples where Lemma B.3 applies. The actual root loci of various optimizers with and without preconditioners are shown in Figure 2 of the main manuscript and in subsequent figures in Appendix C. It is clear that (1) the root locus of optimizers using a single momentum parameter β (or β_1) is identical and independent of their preconditioners, and (2) their crossings always appear at the Nyquist frequency.*

B.4 Proof of Corollary 2.2

All that remains is to write each of the widely-used optimizer filters in the form (25) and read off $Q(-1)$. We use the convention that $\boldsymbol{u}_t$ is the vector multiplying η in (♠) with $\boldsymbol{P} = \boldsymbol{I}$, and $\boldsymbol{m}_t$ denotes the momentum buffer.

Proof. Heavy Ball. From $\boldsymbol{m}_t = \beta\boldsymbol{m}_{t-1} + \boldsymbol{g}_t$ and $\boldsymbol{u}_t = \boldsymbol{m}_t$ we get $\boldsymbol{m}_t = \sum_{k\geq 0}\beta^k \boldsymbol{g}_{t-k}$, the geometric filter quoted below (♠), so

$$Q^{\mathrm{HB}}(z) = \sum_{k\geq 0}\beta^k z^{-k} = \frac{1}{1-\beta z^{-1}}, \qquad Q^{\mathrm{HB}}(1) = \frac{1}{1-\beta}. \tag{31}$$

Its dc-normalized version $Q^{\mathrm{nHB}} = (1-\beta)Q^{\mathrm{HB}}$ is exactly (25) with $J = 1$, $w_1 = 1$, $\beta_1 = \beta$, so Lemma B.3 applies and gives $\Gamma^{\mathrm{nHB}} = (1-\beta)/(1+\beta)$. To recover unnormalized results, we apply Lemma B.2 with $c = 1/(1-\beta)$ yielding $\Gamma^{\mathrm{HB}} = 1/(1+\beta)$. Both equal the respective $Q(-1)$ and recover the results of Cohen et al. (2023).

Nesterov's momentum. With $\boldsymbol{m}_t = \beta\boldsymbol{m}_{t-1} + \boldsymbol{g}_t$ and $\boldsymbol{u}_t = \boldsymbol{g}_t + \beta\boldsymbol{m}_t$,

$$Q^{\mathrm{NAG}}(z) = 1 + \frac{\beta}{1-\beta z^{-1}}, \qquad Q^{\mathrm{NAG}}(1) = \frac{1}{1-\beta}. \tag{32}$$

Multiplying by $(1-\beta)$ and splitting the constant term off as a pole at the origin,

$$Q^{\mathrm{nNAG}}(z) \;=\; (1-\beta)\cdot 1 \;+\; \beta\cdot\frac{1-\beta}{1-\beta z^{-1}}, \tag{33}$$

which is (25) with weights $(1-\beta,\beta)$ and poles $(0,\beta)$. The weights are positive and sum to one for $\beta\in(0,1)$, so Lemma B.3 applies and

$$\Gamma^{\mathrm{nNAG}} \;=\; (1-\beta)\cdot 1+\beta\cdot\frac{1-\beta}{1+\beta} \;=\; \frac{(1-\beta)(1+2\beta)}{1+\beta}, \qquad \Gamma^{\mathrm{NAG}} \;=\; \frac{1+2\beta}{1+\beta}, \tag{34}$$

the second by Lemma B.2. Again, both recover the results reported in Cohen et al. (2023).

This proves Corollary 2.2: for both Heavy Ball and Nesterov's momentum, normalized or not, $\Omega_+=\{\pi\}$ and therefore $\Gamma(\beta)=Q(-1)$, matching the values quoted below (3). □

Similar results on the remaining filters of Section 3, such as QHM (Ma & Yarats, 2019), parallel dual momentum, and cascaded two-pole momenta (Lee et al., 2024; Pagliardini et al., 2025) are derived in Appendix C.1 with their gauges summarized in Table 2. Also refer to the root locus plots in Appendix C for visualization of these results.

B.5 Proof of Corollary 5.1

This subsection proves Corollary 5.1 and the GD reduction used in Theorem 5.3. Fix an iteration and recall the definitions $A=\boldsymbol{g}^\top\boldsymbol{u}$, $B=\eta\boldsymbol{u}^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}/2$, $\|\boldsymbol{u}\|_{\boldsymbol{P}}^2=\boldsymbol{u}^\top\boldsymbol{P}\boldsymbol{u}$,

$$\zeta \;=\; \frac{B}{A}, \qquad \chi \;=\; \frac{\eta S\Gamma}{2}, \qquad \varsigma \;=\; \frac{D^{\|\cdot\|_{\boldsymbol{P}}}}{S} \;=\; \frac{\boldsymbol{u}^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}}{S\|\boldsymbol{u}\|_{\boldsymbol{P}}^2}, \qquad \tau \;=\; \frac{\Gamma A}{\|\boldsymbol{u}\|_{\boldsymbol{P}}^2}. \tag{35}$$

The assumptions $\eta,S,\Gamma>0$, $\boldsymbol{u}\neq\boldsymbol{0}$, and $A\neq 0$ ensure all four quantities are finite, with $\chi>0$ and $\tau\neq 0$. This identity provides additional insight into when the worst-case circumscription χ is fully realized along the actual optimization trajectory, which is summarized in the following proposition.

Proof. The decomposition is exact. Substituting (35) directly,

$$\chi\frac{\varsigma}{\tau} \;=\; \frac{\eta S\Gamma}{2}\cdot\frac{\boldsymbol{u}^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}}{S\|\boldsymbol{u}\|_{\boldsymbol{P}}^2}\cdot\frac{\|\boldsymbol{u}\|_{\boldsymbol{P}}^2}{\Gamma A} \;=\; \frac{\eta}{2}\frac{\boldsymbol{u}^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}}{\boldsymbol{g}^\top\boldsymbol{u}} \;=\; \frac{B}{A} \;=\; \zeta. \tag{36}$$

Every occurrence of S, Γ, and $\|\boldsymbol{u}\|_{\boldsymbol{P}}^2$ cancels, so (♡) is an identity: neither $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\simeq\boldsymbol{H}$ nor occupation of $\boldsymbol{v}_{\max}$ is used anywhere in (36). This is precisely why the decomposition constrains only the ratio ς/τ.

Coincidence condition. From (36) and $\chi>0$, it is trivial that $\zeta=\chi$ if and only if $\varsigma=\tau$. Clearing the common denominator $\|\boldsymbol{u}\|_{\boldsymbol{P}}^2>0$ and multiplying by $S>0$,

$$\zeta=\chi \iff \varsigma=\tau \iff \frac{\boldsymbol{u}^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}}{S}=\Gamma A \iff \boldsymbol{u}^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}=S\Gamma\,\boldsymbol{g}^\top\boldsymbol{u}, \tag{37}$$

which gives the coincidence conditio of (36).

Conditions on the individual factors. Because $S>0$, the definition of ς gives $\varsigma=1 \iff \boldsymbol{u}^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}/\|\boldsymbol{u}\|_{\boldsymbol{P}}^2=S$. Because $\|\boldsymbol{u}\|_{\boldsymbol{P}}^2>0$, the definition of τ gives $\tau=1 \iff \Gamma A=\boldsymbol{u}^\top\boldsymbol{P}\boldsymbol{u} \iff \boldsymbol{u}^\top(\Gamma\boldsymbol{g}-\boldsymbol{P}\boldsymbol{u})=0$. Sufficiency of $\varsigma=\tau=1$ is immediate from (36). Necessity fails, and Appendix B.7 exhibits an admissible instance with $\varsigma=\tau\neq 1$.

Form at the realized edge. Setting $\zeta=1$ in (36) and solving for χ gives $\chi=\tau/\varsigma$. If $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\simeq\boldsymbol{H}$, then $D^{\|\cdot\|_{\boldsymbol{P}}}\le S$ by the variational characterization of S recalled in Section 4, hence $\varsigma\le 1$ and $\chi\ge\tau$.

GD reduction. Take $\boldsymbol{P}=I$ and $Q\equiv 1$. Then $\Gamma=Q(-1)=1$ and $\boldsymbol{u}=\boldsymbol{g}$, so $A=\boldsymbol{g}^\top\boldsymbol{u}=\|\boldsymbol{g}\|^2$ and $\|\boldsymbol{u}\|_{\boldsymbol{P}}^2=\|\boldsymbol{g}\|^2$, hence $\tau=\Gamma A/\|\boldsymbol{u}\|_{\boldsymbol{P}}^2=1$. The identity collapses to $\zeta=\chi\varsigma$ with $\chi=\eta S/2$. On a frozen quadratic, $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}=\boldsymbol{H}$ and $\varsigma=\boldsymbol{g}^\top\boldsymbol{H}\boldsymbol{g}/(S\|\boldsymbol{g}\|^2)$, which equals 1 if and only if $\boldsymbol{g}$ is an eigen-direction for $\lambda_{\max}(\boldsymbol{H})$. This is the occupation hypothesis of Section 2, and is the content of the GD clause in Corollary 5.1. □

B.6 Proof of Corollary 5.2

Proof. Notations. Write $\boldsymbol{v} = \boldsymbol{v}_{\max}$ for the unit top eigenvector of $\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2}$, so that

$$\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2}\boldsymbol{v} \;=\; S\boldsymbol{v} \qquad\Longleftrightarrow\qquad \boldsymbol{H}\boldsymbol{P}^{-1/2}\boldsymbol{v} \;=\; S\boldsymbol{P}^{1/2}\boldsymbol{v}. \tag{38}$$

This will be used throughout the proof.

The locked mode is invariant. By (Q1) the loss is quadratic over the horizon, so $\boldsymbol{g}_s = \boldsymbol{H}(\boldsymbol{\theta}_s - \boldsymbol{\theta}_\star)$. With (Q3), we have $x_s\boldsymbol{H}\boldsymbol{P}^{-1/2}\boldsymbol{v} = Sx_s\boldsymbol{P}^{1/2}\boldsymbol{v}$ using (38). By (Q2) the preconditioner may be pulled out of the convolution, so the update (♠) is

$$\boldsymbol{u}_t \;=\; \boldsymbol{P}^{-1}(Q * \boldsymbol{g}_{\le t})_t \;=\; S(Q * x_{\le t})_t\,\boldsymbol{P}^{-1}\boldsymbol{P}^{1/2}\boldsymbol{v} \;=\; S(Q * x_{\le t})_t\,\boldsymbol{P}^{-1/2}\boldsymbol{v}. \tag{39}$$

The update therefore stays in $\mathrm{span}(\boldsymbol{P}^{-1/2}\boldsymbol{v})$, which is the locked mode.

Spatial factor. Abbreviate $q := (Q * x_{\le t})_t$ and use $\boldsymbol{v}^\top\boldsymbol{v} = 1$ together with (38) to get:

$$\begin{aligned} A \;&=\; \boldsymbol{g}_t^\top\boldsymbol{u}_t \;=\; S^2x_tq, \\ \|\boldsymbol{u}_t\|_{\boldsymbol{P}}^2 \;&=\; \boldsymbol{u}_t^\top\boldsymbol{P}\boldsymbol{u}_t \;=\; S^2q^2, \\ \boldsymbol{u}_t^\top\boldsymbol{H}\boldsymbol{u}_t \;&=\; S^2q^2\,\boldsymbol{v}^\top\boldsymbol{P}^{-1/2}\boldsymbol{H}\boldsymbol{P}^{-1/2}\boldsymbol{v} \;=\; S^3q^2. \end{aligned} \tag{40}$$

Since $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} = \boldsymbol{H}$ by (Q1), the spatial participation factor is

$$\varsigma \;=\; \frac{\boldsymbol{u}_t^\top\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}_t}{S\|\boldsymbol{u}_t\|_{\boldsymbol{P}}^2} \;=\; \frac{S^3q^2}{S\cdot S^2q^2} \;=\; 1. \tag{41}$$

Therefore, the conditions (Q1)–(Q3) are sufficient for the spatial factor to be $\varsigma = 1$. The temporal factor is $\tau = \Gamma A/\|\boldsymbol{u}_t\|_{\boldsymbol{P}}^2 = \Gamma S^2x_tq/(S^2q^2) = \Gamma x_t/q$, which recovers (13).

Temporal factor. By (Q4) the escape is at $z = -1$, so $\Gamma = Q(-1)$ by Proposition 2.1, and the period-2 history $x_{t-k} = (-1)^k x_t$ gives

$$q \;=\; \sum_{k\ge0} Q_k x_{t-k} \;=\; \Big(\sum_{k\ge0} Q_k(-1)^k\Big)x_t \;=\; Q(-1)\,x_t \;=\; \Gamma x_t. \tag{42}$$

Hence $\tau = \Gamma x_t/q = 1$, and $A = S^2x_tq = \Gamma S^2x_t^2 > 0$ because $\Gamma > 0$ and $x_t \neq 0$, so $A \neq 0$ and the identity (36) applies. Substituting $\varsigma = \tau = 1$ into the identity (36) gives $\zeta = \chi = \eta S\Gamma/2 = \eta SQ(-1)/2$. Note that (42) is what requires the Nyquist frequency: a mode oscillating at any other escape frequency $\omega_\star$ contributes a phase shift, so no real gain satisfies $q = \Gamma x_t$ at every step. □

B.7 A Counterexample to the Necessity of $\varsigma = \tau = 1$

The following instance lies entirely inside the framework of (♠) and is exactly computable. Take the quadratic objective $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{H}\boldsymbol{\theta}$ on $\mathbb{R}^2$ with $\boldsymbol{H} = \mathrm{diag}(4,1)$, so that $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} = \boldsymbol{H}$ *exactly* and $\boldsymbol{v}_{\max} = \boldsymbol{e}_1$. Take $\boldsymbol{P} = \boldsymbol{I}$, so $S = 4$, and take the dc-normalized Heavy Ball filter $Q_k = (1-\beta)\beta^k$ with $\beta = 1/2$, so $Q(1) = 1$ and $\Gamma = Q(-1) = (1-\beta)/(1+\beta) = 1/3$. Let $\eta = 1/10$ and start from a zero momentum buffer, so that $\boldsymbol{u}_0 = (1-\beta)\boldsymbol{g}_0$ and

$$\boldsymbol{g}_1 \;=\; (\boldsymbol{I} - \eta(1-\beta)\boldsymbol{H})\boldsymbol{g}_0 \;=\; \mathrm{diag}(\tfrac{4}{5},\tfrac{19}{20})\boldsymbol{g}_0, \qquad \boldsymbol{u}_1 \;=\; \mathrm{diag}(\tfrac{13}{20},\tfrac{29}{40})\boldsymbol{g}_0. \tag{43}$$

All four quantities are invariant to the scale of $\boldsymbol{g}_0$, so only its direction matters. Reading them at $t = 1$:

$$\begin{aligned} \boldsymbol{g}_0 = \boldsymbol{e}_1 :&\quad \varsigma = 1, \quad \tau = \tfrac{16}{39}, \quad \chi = \tfrac{1}{15}, \quad \zeta = \tfrac{13}{80}; \\ \boldsymbol{g}_0 = \boldsymbol{e}_2 :&\quad \varsigma = \tfrac{1}{4}, \quad \tau = \tfrac{38}{87}, \quad \chi = \tfrac{1}{15}, \quad \zeta = \tfrac{29}{760}. \end{aligned} \tag{44}$$

Two conclusions can made from this. First, the case $\boldsymbol{g}_0 = \boldsymbol{e}_1$ satisfies every geometric condition one would attach to coincidence, namely $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} = \boldsymbol{H}$ exactly and $\boldsymbol{u}_1 \in \mathrm{span}(\boldsymbol{v}_{\max})$, hence $\varsigma = 1$; yet $\zeta/\chi = 39/16 \neq 1$, because Γ is the loop gain at the root-locus escape frequency while the trajectory at $t = 1$ is not at that frequency. Second, writing $\boldsymbol{g}_0(\phi) = (\cos\phi, \sin\phi)$, the map $\phi \mapsto \varsigma - \tau$ is continuous on $[0, \pi/2]$ with $\varsigma - \tau = 23/39 > 0$ at $\phi = 0$ and $\varsigma - \tau = -65/348 < 0$ at $\phi = \pi/2$, so it has a root $\phi_\star \in (0, \pi/2)$. At $\phi_\star$ we have $\varsigma = \tau$ and therefore $\zeta = \chi$, while $\varsigma < 1$ because $\varsigma = 1$ forces $\boldsymbol{u}_1 \in \mathrm{span}(\boldsymbol{e}_1)$, i.e. $\phi = 0$. Numerically $\phi_\star = 1.01024954$ and $\varsigma = \tau = 0.43040137$, giving $\zeta = \chi = 1/15$ to ten significant digits. Hence $\varsigma = \tau = 1$ is not necessary for $\zeta = \chi$.

Table 2: Root-locus gauge of the *one-pole and parallel* optimizer filters. Every dc-normalized entry is a positive one-pole mixture (25), so Lemma B.3 certifies $\Omega_+ = \{\pi\}$ and $\Gamma = Q(-1)$. Cascaded two-pole momenta are *not* in this table. The last column recovers the results from Cohen et al. (2023).

Optimizer	$Q(z)$	$Q(1)$	mixture $(w_j;\beta_j)$	$\Gamma = Q(-1)$
GD	1	1	$(1;0)$	1
Heavy Ball	$\dfrac{1}{1-\beta z^{-1}}$	$\dfrac{1}{1-\beta}$	—	$\dfrac{1}{1+\beta}$
normalized HB	$\dfrac{1-\beta}{1-\beta z^{-1}}$	1	$(1;\beta)$	$\dfrac{1-\beta}{1+\beta}$
Nesterov	$1+\dfrac{\beta}{1-\beta z^{-1}}$	$\dfrac{1}{1-\beta}$	—	$\dfrac{1+2\beta}{1+\beta}$
normalized Nesterov	$(1-\beta)+\dfrac{\beta(1-\beta)}{1-\beta z^{-1}}$	1	$(1-\beta,\beta;0,\beta)$	$\dfrac{(1-\beta)(1+2\beta)}{1+\beta}$
QHM, $\nu \in [0,1]$	$(1-\nu)+\dfrac{\nu(1-\beta)}{1-\beta z^{-1}}$	1	$(1-\nu,\nu;0,\beta)$	$\dfrac{1+\beta(1-2\nu)}{1+\beta}$
parallel dual momentum, $\alpha > 0$	$\displaystyle\sum_{j=1,2}\dfrac{w_j(1-\beta_j)}{1-\beta_j z^{-1}}$	1	$\dfrac{(1,\alpha)}{1+\alpha};(\beta_1,\beta_2)$	$\displaystyle\sum_{j=1,2} w_j\dfrac{1-\beta_j}{1+\beta_j}$

B.8 Proof of Theorem 5.3

Proof. From the identity. Corollary 5.1 gives $\zeta = \chi\varsigma/\tau$ and $\zeta = \chi$ if and only if $\varsigma = \tau$. The scalar $\chi = \eta S\Gamma/2$ is a function of (η, S, Γ) alone. Therefore ζ is a function of (η, S, Γ) if and only if the ratio ς/τ is.

GD reduction. For $\boldsymbol{P} = I$ and $Q \equiv 1$, the GD clause of Corollary 5.1 gives $\Gamma = 1$, $\tau \equiv 1$, and $\zeta = \chi\varsigma$ with $\chi = \eta S/2$. Hence χ determines ζ if and only if occupation ς is a function of (η, S) alone.

Occupation is not a function of S. On a frozen quadratic, $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} = \boldsymbol{H}$ and $\boldsymbol{u} = \boldsymbol{g}$, so $\varsigma = \boldsymbol{g}^\top\boldsymbol{H}\boldsymbol{g}/(S\|\boldsymbol{g}\|^2)$. This equals 1 if and only if $\boldsymbol{g}$ lies in the top eigenspace of $\boldsymbol{H}$. In dimension at least 2 that is not implied by $S = \lambda_{\max}(\boldsymbol{H})$. The instance $\boldsymbol{H} = \mathrm{diag}(4,1)$ has $S = 4$ independently of $\boldsymbol{g}$, but $\varsigma = 1$ along $\boldsymbol{e}_1$ and $\varsigma = 1/4$ along $\boldsymbol{e}_2$. The same $(\eta, S, \Gamma = 1)$ therefore yields two values of ζ. □

C Gauge Derivations

Proposition 2.1 assigns a scalar gauge $\Gamma := 2/k_\star(Q)$ to every causal filter Q. It is the loop gain at the first positive-gain unit-circle crossing of the root locus, abbreviated as the *escape gain* in the main text. Corollary 2.2 is the special case $\Gamma = Q(-1)$, which holds for Heavy Ball and Nesterov's momentum and is proved in Appendix B.4 from Lemmas B.2 and B.3. This additional section extends this to include $Q(z)$ and Γ for every optimizer that appears in the main text, especially those with $\Gamma \neq Q(-1)$, which are the filters that violate hypothesis (Q4) of Corollary 5.2 on the filter side.

C.1 QHM and Parallel Dual Momentum

We use the convention that $\boldsymbol{u}_t$ is the vector multiplying η in (♠) with $\boldsymbol{P} = \boldsymbol{I}$, and $\boldsymbol{m}_t$ denotes the momentum buffer. Every dc-normalized filter in this subsection is a positive one-pole mixture (25), so Lemma B.3 applies and gives $\Omega_+ = \{\pi\}$ and $\Gamma = Q(-1)$.

QHM. QHM (Ma & Yarats, 2019) uses the normalized buffer $\boldsymbol{m}_t = \beta\boldsymbol{m}_{t-1} + (1-\beta)\boldsymbol{g}_t$ and $\boldsymbol{u}_t = (1-\nu)\boldsymbol{g}_t + \nu\boldsymbol{m}_t$, which is (25) with weights $(1-\nu, \nu)$ and poles $(0, \beta)$; it is already dc-normalized, i.e., $Q(1) = 1$. For $\nu \in (0,1)$ the weights are strictly positive. Lemma B.3 gives

$$\Gamma^{\mathrm{QHM}}(\beta,\nu) \;=\; \frac{1+\beta(1-2\nu)}{1+\beta}. \tag{45}$$

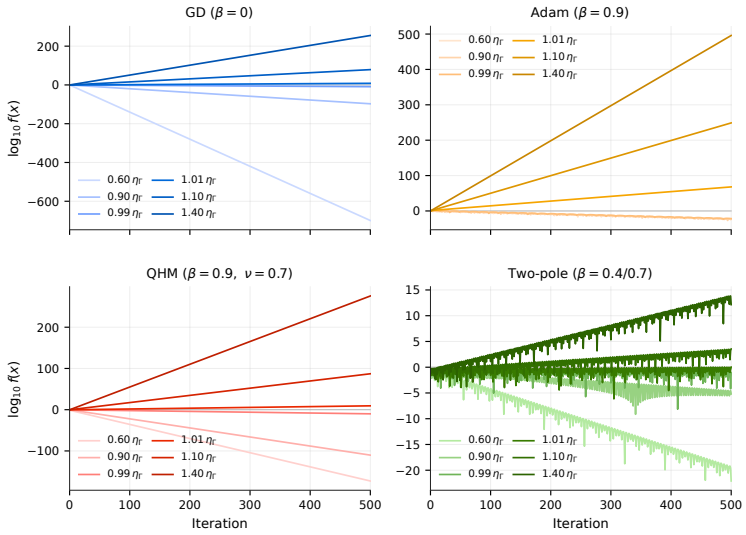


Figure 6: We optimize a quadratic loss $f(x) = \frac{1}{2}x^2$ with four types of optimizers: GD, Adam, QHM, and cascade two-pole at various learning rates η. Observe that the stability bound of $\chi = \eta S\Gamma/2 \leq 1$ is always satisfied for this quadratic loss.

Table 3: Predicted and observed quadratic thresholds. Γ_{exp} is the experiment-report gauge. The observed η is the last stable and first unstable η on a 0.005-relative grid. 605 sweep runs in total.

Filter	Γ_{exp} / 2	$2/\Gamma_{\text{exp}}$	$2/Q(-1)$	$\omega_\star/\pi$	observed η
GD	0.5	2	2	1.000	[2, 2.01]
EMA-HB $\beta = 0.9$	0.02632	38	38	1.000	[37.81, 38]
QHM $\beta = 0.9, \nu = 0.7$	0.1684	5.9375	5.9375	1.000	[5.908, 5.938]
two-pole 0.4/0.7	0.3889	2.571	26.44	0.263	[2.571, 2.584]
two-pole 0.7/0.85	1.469	0.681	139.8	0.088	[0.681, 0.684]

The endpoints $\nu = 0$ and $\nu = 1$ are one-pole reductions (GD and normalized Heavy Ball), both still with $\Gamma = Q(-1)$. Two special cases are immediate from (45) and (33): $\nu = 1$ recovers normalized Heavy Ball, and $\nu = \beta$ recovers *normalized Nesterov*, since $1 + \beta - 2\beta^2 = (1 - \beta)(1 + 2\beta)$. The experimental QHM arm of Section 3 uses $\nu = 0.7 \in (0, 1)$, so it remains in this Nyquist class with $\Gamma = Q(-1)$.

Parallel Dual Momentum. Dual momentum in its *parallel* realization, two buffers at $\beta_1 \neq \beta_2$ summed as $\boldsymbol{u}_t = (\boldsymbol{m}_t^{(1)} + \alpha \boldsymbol{m}_t^{(2)})/(1 + \alpha)$, is again (25), now with $J = 2$, weights $(1, \alpha)/(1 + \alpha)$ and poles (β_1, β_2), so for $\alpha > 0$ Lemma B.3 applies verbatim, with no restriction on the number of poles. Table 2 collects these results. In our text, we did not use this construction for two-pole momenta (see the next section), but we included them in the table for completeness.

C.2 Two-Pole Momentum: the Gauge is No Longer $Q(-1)$

Everything so far has been a *parallel* construction: gradients are filtered by several one-pole blocks and the outputs are added. The two-pole filters (Lee et al., 2024) of Section 3 are instead built by *cascading* a momentum buffer with a second one-pole block. Define the two-pole momentum family, dc-normalized, with $\beta_1, \beta_2 \in (0, 1)$ and $\alpha \geq 0$,

$$Q_\alpha(z) \;=\; \underbrace{\frac{1 - \beta_1}{1 - \beta_1 z^{-1}}}_{\text{momentum buffer}} \cdot \underbrace{\frac{1}{1 + \alpha}\left[1 + \frac{\alpha(1 - \beta_2)}{1 - \beta_2 z^{-1}}\right]}_{\text{second one-pole block}}, \tag{46}$$

so that $Q_\alpha(1) = 1$, the poles are β_1 and β_2, and there is one finite zero. The parameter α interpolates between the two limits of interest: at $\alpha = 0$ the zero cancels the second pole and Q_0 is normalized

Table 4: Temporal filters of the executed families, in the paper-normalized convention $\chi = \eta S\Gamma/2$ of Appendix C.1. $k_\star$ is the unit-curvature learning-rate edge. $\omega_\star$ is the first unit-circle crossing in radians. Adam, AdamW, Adafactor, AMSGrad and PAdam share one EMA filter; RMSProp, AdaGrad and GD share the memoryless filter. Note that two-pole momentum cases have $\Gamma \neq Q(-1)$.

Filter	Families	$Q(1)$	$Q(-1)$	Γ	$k_\star$	$\omega_\star/\pi$
GD	GD, RMSProp, AdaGrad	1	1	1	2	1
EMA $\beta = 0.9$	Adam, AdamW, Adafactor, AMSGrad, PAdam	1	0.05263	0.05263	38	1
NAdam $\beta = 0.9$	NAdam	1	0.1474	0.1474	13.57	1
HB $\beta = 0.9$	Heavy Ball	10	0.5263	0.5263	3.8	1
NAG $\beta = 0.9$	Nesterov	10	1.474	1.474	1.357	1
QHM $\beta = 0.9, \nu = 0.7$	QHM	1	0.3368	0.3368	5.938	1
two-pole 0.4/0.7	Grokfast-style	1	0.07563	0.7778	2.571	0.263
two-pole 0.7/0.85	Grokfast-style	1	0.01431	2.938	0.681	0.088

Heavy Ball, while as $\alpha \to \infty$ the zero moves to the origin and Q_∞ is the pure double-EMA cascade $Q_\infty(z) = \prod_{j=1,2}(1-\beta_j)/(1-\beta_j z^{-1})$.

One EMA pole cannot reach a $90°$ lag according to (30). A parallel sum stays inside the same cone, and no interior crossing is possible. A cascade, by contrast, *adds* lags, and two poles can jointly exceed the critical line. Making this quantitative near $\omega = \pi$ already gives the exact threshold: writing $\omega = \pi - \epsilon$ and $\beta_1 = \beta_2 = \beta$, the cascade lag is $2\varphi_\beta(\pi - \epsilon) = 2\beta\epsilon/(1+\beta) + O(\epsilon^2)$ while the required lag is $\epsilon/2$, so an interior crossing detaches from $\omega = \pi$ if and only if

$$\frac{2\beta}{1+\beta} > \frac{1}{2} \qquad \Longleftrightarrow \qquad \beta > \tfrac{1}{3}. \tag{47}$$

For two-pole momenta, we numerically evaluate the gauge and plot it in Figure 7.

C.3 Stability Bound for Fixed Quadratic Loss

To demonstrate that Proposition 2.1 yields the exact stability bound for a fixed quadratic loss, we conduct simple numerical experiments. Four types of optimizers—GD, Adam, QHM, and cascade two-pole—are trained on a quadratic loss $f(x) = \frac{1}{2}x^2$ with various learning rates η. Figure 6 summarizes the results. In the legend, we use the relative learning rate η/η_{r}, where $\eta_{\mathrm{r}} = 2/(S\Gamma)$ is determined by the stability condition $\chi = \eta S\Gamma/2 \leq 1$ in Proposition 2.1. For the two-pole momentum optimizer, we calculate the numerically found gauge, as shown in the root locus plot in Figure 7, which is not the Nyquist gain, i.e., $\Gamma \neq Q(-1)$. Regardless of optimizer type, the stability bound is always satisfied on this fixed quadratic loss.

C.4 Complete Root Locus

Complete root loci for the filters studied in this paper are shown in Figure 7, extending the demonstration in Figure 2 in the main text to all filters. See also Remark B.4 in the previous section for a more detailed interpretation of the results.

D Stabilization from High-Order Terms

We begin by noting that the realized EoS ($\diamondsuit$) is already exact. Section 4 expresses the residual quantity in terms of a quadratic load $\boldsymbol{u}^\top \bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} \boldsymbol{u}$ and occasionally interprets this as an approximation of the second Taylor term $\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} \simeq \boldsymbol{H}$. However, this simplification is only for our conceptual understanding of the quantity. It does *not* affect the exactness of our realized EoS ζ. This section provides a more detailed analysis to support this simplification. Higher-order geometry does not explain *why* the edge is occupied. It determines *how* an occupied quadratic load is realized as a finite step, and whether that realization rescues or amplifies a one-step crossing of $\zeta = 1$.

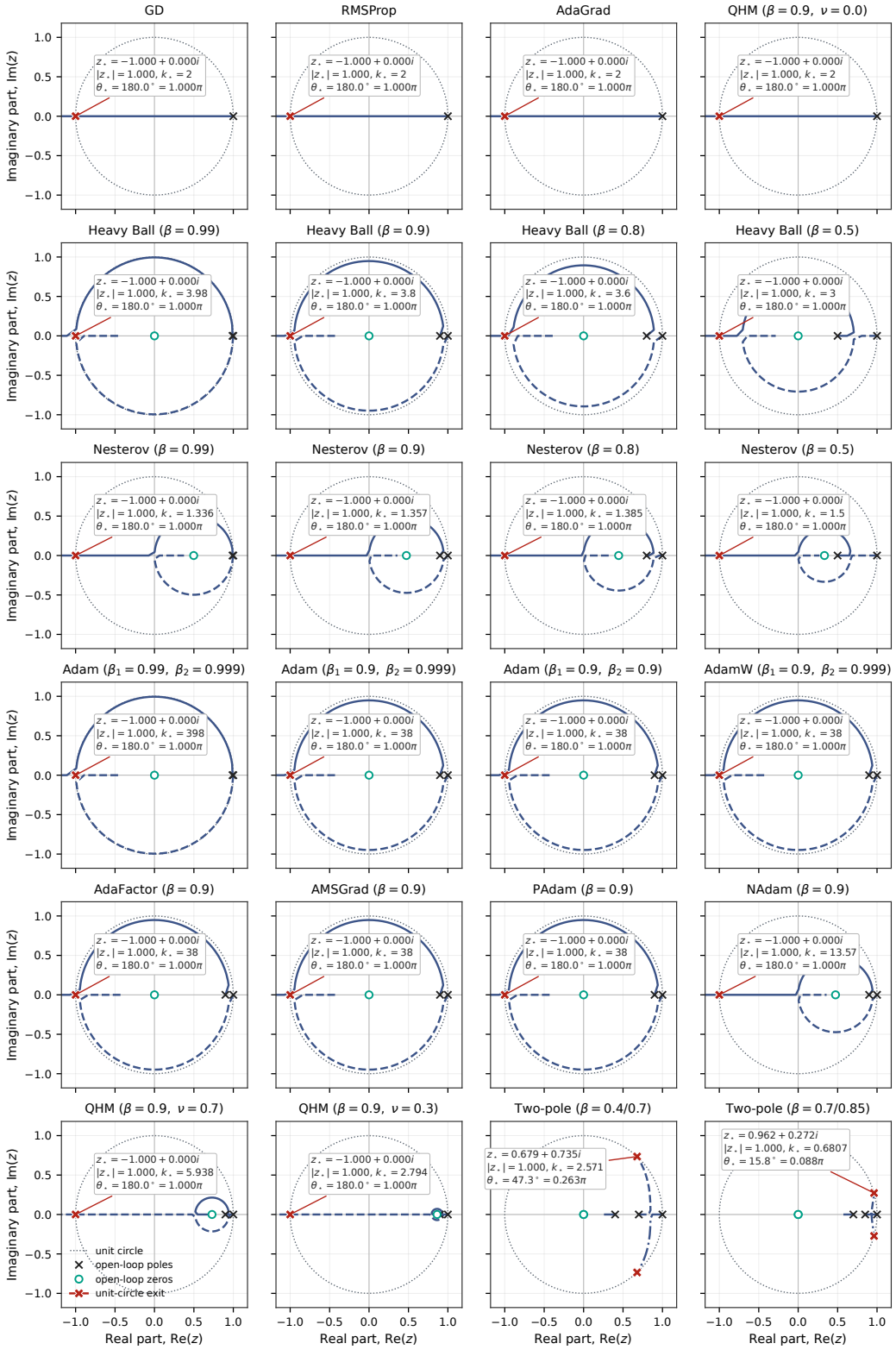


Figure 7: Complete root locus for the filters studied in this paper, extending Figure 2 in the main text to all filters.

D.1 EXACT NONLINEAR REALIZATION

Fix a step $\boldsymbol{\theta}^+ := \boldsymbol{\theta} - \eta \boldsymbol{u}$ with $A := \boldsymbol{g}^\top \boldsymbol{u} > 0$ and $H_{\boldsymbol{u}} := \boldsymbol{u}^\top \boldsymbol{H} \boldsymbol{u} > 0$. The quadratic and exact secant expenditures are defined as

$$B_2 \;:=\; \frac{\eta}{2} H_{\boldsymbol{u}}, \qquad B \;:=\; A + \frac{\mathcal{L}(\boldsymbol{\theta}^+) - \mathcal{L}(\boldsymbol{\theta})}{\eta} \;=\; \frac{\eta}{2} \boldsymbol{u}^\top \bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} \boldsymbol{u}, \tag{48}$$

the second equality being (11). We also define the corresponding loads $\zeta_2 := B_2/A$ and $\zeta := B/A$. Let the *nonlinear realization factor* be the ratio of the two,

$$h \;:=\; \frac{\zeta}{\zeta_2} \;=\; \frac{B}{B_2}, \tag{49}$$

so that, identically,

$$\zeta \;=\; \zeta_2 \, h. \tag{50}$$

Note that the spatial factor ς of (♡) is built from the *secant* Hessian, hence already contains h. Write

$$\varsigma_2 \;:=\; \frac{\boldsymbol{u}^\top \boldsymbol{H} \boldsymbol{u}}{S \|\boldsymbol{u}\|_{\boldsymbol{P}}^2} \tag{51}$$

for the quadratic occupancy of the same ray. Then $\varsigma = \varsigma_2 h$, and the main-text factorization splits as

$$\zeta_2 \;=\; \chi \, \frac{\varsigma_2}{\tau}, \qquad \zeta \;=\; \chi \, \frac{\varsigma_2}{\tau} \, h. \tag{52}$$

Thus we can further interpret that (1) χ is *available load*, (2) ς_2/τ occupies a *quadratic-only load* along $\boldsymbol{u}$, and (3) h *realizes* that load as a finite step. We need not apply the correction factor h to the main-text ς again, which is already exact as $\varsigma = \varsigma_2 h$. Multiplying the main-text ς by a further h would double-count the same correction.

Since we are dealing with directional derivatives along the update direction $\boldsymbol{u}$ actuated by the optimizer, the factor h is a Taylor series of one *scalar*, not a collection of tensors. Write $T_{\boldsymbol{u}} := \nabla^3 \mathcal{L}[\boldsymbol{u}]^3$, $F_{\boldsymbol{u}} := \nabla^4 \mathcal{L}[\boldsymbol{u}]^4$, $P_{\boldsymbol{u}} := \nabla^5 \mathcal{L}[\boldsymbol{u}]^5$, and

$$\theta_3 := \frac{\eta T_{\boldsymbol{u}}}{3 H_{\boldsymbol{u}}}, \qquad \theta_4 := \frac{\eta^2 F_{\boldsymbol{u}}}{12 H_{\boldsymbol{u}}}, \qquad \theta_5 := \frac{\eta^3 P_{\boldsymbol{u}}}{60 H_{\boldsymbol{u}}}. \tag{53}$$

Lemma D.1 (Secant Hessian expansion)**.** *Let $\mathcal{L}$ be C^{m+2} on a neighbourhood of the executed segment $\{\boldsymbol{\theta} - s\eta\boldsymbol{u} : s \in [0,1]\}$. (If the activation is piecewise linear, this requires that the segment not cross a kink.) Then the weighted secant Hessian (**??**) satisfies*

$$\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} \;=\; 2 \sum_{k=0}^{m} \frac{(-\eta)^k}{(k+2)!} \nabla^{k+2} \mathcal{L}(\boldsymbol{\theta})[\boldsymbol{u}]^k \;+\; O(\eta^{m+1}), \tag{54}$$

where $\nabla^j \mathcal{L}[\boldsymbol{u}]^k$ denotes the j-th derivative tensor contracted k times with $\boldsymbol{u}$.

Proof. Expand $\nabla^2 \mathcal{L}(\boldsymbol{\theta} - s\eta\boldsymbol{u}) = \sum_{k \geq 0} \frac{(-s\eta)^k}{k!} \nabla^{k+2} \mathcal{L}(\boldsymbol{\theta})[\boldsymbol{u}]^k$ and integrate against the kernel $2(1-s)$ using $2 \int_0^1 (1-s) s^k \, ds = \frac{2}{(k+1)(k+2)}$. The coefficient of $\nabla^{k+2} \mathcal{L}[\boldsymbol{u}]^k$ is therefore $\frac{2(-\eta)^k}{k!\,(k+1)(k+2)} = \frac{2(-\eta)^k}{(k+2)!}$. □

Proposition D.2 (Impedance series and its loss-only form)**.** *With $A \neq 0$,*

$$\zeta \;=\; \sum_{j \geq 2} \frac{(-1)^j \eta^{j-1}}{j!\, A} \nabla^j \mathcal{L}(\boldsymbol{\theta})[\boldsymbol{u}]^j \;=\; \zeta_2 \big(1 - \theta_3 + \theta_4 - \theta_5 + O(\eta^4)\big), \tag{55}$$

and therefore $h = 1 - \theta_3 + \theta_4 - \theta_5 + O(\eta^4)$. Moreover, with $\Delta\mathcal{L} := \mathcal{L}(\boldsymbol{\theta} - \eta\boldsymbol{u}) - \mathcal{L}(\boldsymbol{\theta})$,

$$\zeta - 1 \;=\; \frac{\Delta\mathcal{L}}{\eta A} \;=\; \frac{\mathcal{L}(\boldsymbol{\theta} - \eta\boldsymbol{u}) - \mathcal{L}(\boldsymbol{\theta})}{\eta \, \boldsymbol{g}^\top \boldsymbol{u}}, \tag{56}$$

so (55) expands a quantity that is exactly the one-step loss change normalized by the learning power.

Proof. The first equality follows from Lemma D.1 and $\zeta = \frac{\eta}{2A}\boldsymbol{u}^\top \bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}}\boldsymbol{u}$. The factorization is (53) collected term by term. For (56), expand $\Delta\mathcal{L} = \sum_{j\geq 1}\frac{(-\eta)^j}{j!}\nabla^j\mathcal{L}[\boldsymbol{u}]^j = -\eta A + \sum_{j\geq 2}\frac{(-\eta)^j}{j!}\nabla^j\mathcal{L}[\boldsymbol{u}]^j$ and divide by ηA. Equivalently, it is (11) restated. □

Equation (56) is Hessian-free, which was mentioned in the main text as well. Only two forward passes and the inner product A are needed, which are already computed by the optimizer. Only the quadratic load ζ_2 requires a Hessian-vector product. The series (55) is simply the Taylor expansion of the exact ratio h, not an independent mechanism. The following corollary upper bounds the approximation error when we treat the secant Hessian as the true Hessian.

Corollary D.3 (Approximation of the edge of stability)**.** *If the Hessian $\boldsymbol{H} = \nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$ is $L_{\boldsymbol{H}}$-Lipschitz on the segment $\{\boldsymbol{\theta} - s\eta\boldsymbol{u} : s \in [0,1]\}$, then*

$$\left|\boldsymbol{u}^\top(\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} - \boldsymbol{H})\boldsymbol{u}\right| \leq \frac{L_{\boldsymbol{H}}\eta\|\boldsymbol{u}\|^3}{3}. \tag{57}$$

Writing

$$\bar{\zeta} := \frac{\eta\boldsymbol{u}^\top\boldsymbol{H}\boldsymbol{u}}{2\boldsymbol{g}^\top\boldsymbol{u}} \tag{58}$$

for the quadratic estimator (i.e. ζ_2 of (55)), the approximation error is bounded by

$$\left|\zeta - \bar{\zeta}\right| \leq \frac{L_{\boldsymbol{H}}\eta^2\|\boldsymbol{u}\|^3}{6\,|\boldsymbol{g}^\top\boldsymbol{u}|}. \tag{59}$$

Proof. By (**??**), $\boldsymbol{u}^\top(\bar{\boldsymbol{H}}_{\eta,\boldsymbol{u}} - \boldsymbol{H})\boldsymbol{u} = 2\int_0^1(1-s)\,\boldsymbol{u}^\top\big(\nabla^2\mathcal{L}(\boldsymbol{\theta}-s\eta\boldsymbol{u}) - \boldsymbol{H}\big)\boldsymbol{u}\,ds$. Lipschitz continuity of $\boldsymbol{H}$ bounds the integrand by $L_{\boldsymbol{H}}(s\eta\|\boldsymbol{u}\|)\,\|\boldsymbol{u}\|^2$, so the integral is at most $2L_{\boldsymbol{H}}\eta\|\boldsymbol{u}\|^3\int_0^1(1-s)s\,ds = L_{\boldsymbol{H}}\eta\|\boldsymbol{u}\|^3/3$. Dividing by $2|A|/\eta$ yields (59). This is the $m = 0$ remainder of Lemma D.1 under a Lipschitz rather than a C^3 hypothesis. □

All higher-order content in ζ is this one scalar. It is the ratio of the averaged directional curvature over the step to the instantaneous directional curvature at its start. The coefficients $-\theta_3 + \theta_4 - \cdots$ are its Taylor coefficients. This scalar can act in either side of the edge.

Let $\hat{\boldsymbol{u}} := \boldsymbol{u}/\|\boldsymbol{u}\|$, let $r := \eta\|\boldsymbol{u}\|$ be the step length, and define the *directional curvature profile*

$$\psi(s) := \hat{\boldsymbol{u}}^\top\nabla^2\mathcal{L}(\boldsymbol{\theta} - s\hat{\boldsymbol{u}})\,\hat{\boldsymbol{u}}, \qquad g_{\boldsymbol{u}} := \boldsymbol{g}^\top\hat{\boldsymbol{u}} > 0. \tag{60}$$

Proposition D.4 (Profile form)**.** *The impedance is the $(r-s)$-weighted mean of the directional curvature over the executed step,*

$$\zeta = \frac{1}{r\,g_{\boldsymbol{u}}}\int_0^r (r-s)\,\psi(s)\,ds = \frac{r\,\bar{\psi}_r}{2\,g_{\boldsymbol{u}}}, \qquad \bar{\psi}_r := \frac{2}{r^2}\int_0^r (r-s)\,\psi(s)\,ds, \tag{61}$$

and therefore, with $\zeta_2 = r\psi(0)/(2g_{\boldsymbol{u}})$,

$$h = \frac{\zeta}{\zeta_2} = \frac{\bar{\psi}_r}{\psi(0)}. \tag{62}$$

Proof. Let $\phi(s) := \mathcal{L}(\boldsymbol{\theta} - s\hat{\boldsymbol{u}})$, so $\phi'(0) = -g_{\boldsymbol{u}}$ and $\phi''(s) = \psi(s)$. Taylor's theorem with integral remainder gives $\phi(r) - \phi(0) = -rg_{\boldsymbol{u}} + \int_0^r(r-s)\psi(s)\,ds$, and $\eta A = rg_{\boldsymbol{u}}$, so by (56) $\zeta = 1 + \frac{\phi(r)-\phi(0)}{rg_{\boldsymbol{u}}}$, which is (61). □

Corollary D.5 (Monotone softening and stiffening)**.** *If ψ is nonincreasing on $[0,r]$ then $h \leq 1$, and if ψ is nondecreasing then $h \geq 1$. Both inequalities are strict if the monotonicity is strict on a set of positive measure. More quantitatively, if $\psi(s) \leq (1-\mu)\psi(0)$ for $s \geq \varrho r$ with $\mu \in (0,1)$ and $\varrho \in (0,1)$, then*

$$h \leq 1 - \mu(1-\varrho)^2. \tag{63}$$

Proof. The kernel $2(r-s)/r^2$ is a probability density on $[0,r]$, so $\bar{\psi}_r$ is an average of ψ and (62) gives the two monotone statements. For the bound, $\int_{\varrho r}^{r} \frac{2(r-s)}{r^2}\, ds = (1-\varrho)^2$, and on that sub-interval ψ is below $(1-\mu)\psi(0)$ while elsewhere it is at most $\psi(0)$. □

Thus we can call $h < 1$ *directional softening* and $h > 1$ *directional stiffening*.

D.2 Entry, return, and maintenance

Because $\zeta = \zeta_2 h$, a step lies below, on, or above the realized edge according as h lies below, on, or above $1/\zeta_2$:

$$\zeta < 1 \iff h < \frac{1}{\zeta_2}, \qquad \zeta = 1 \iff h = \frac{1}{\zeta_2}, \qquad \zeta > 1 \iff h > \frac{1}{\zeta_2}. \tag{64}$$

These are simple identities. When we are discussing the higher-order term h, we are mainly interested in the condition whether this term rescues or destabilizes a one-step crossing of $\zeta = 1$, i.e., whether h is below, on, or above $1/\zeta_2$.

Theorem D.6 (Nonlinear rescue and destabilization)**.** *Assume $A > 0$ and $H_{\boldsymbol{u}} > 0$.*

1. *If $\zeta_2 = 1 + \delta$ with $\delta > 0$, the step is quadratically supercritical. It is rescued to $\zeta \le 1$ if and only if*
$$h \;\le\; \frac{1}{1+\delta} \qquad \textit{i.e.} \qquad 1 - h \;\ge\; \frac{\delta}{1+\delta}. \tag{65}$$
To the order retained in (55), the same comparison reads $\theta_3 - \theta_4 + O(\eta^3) \ge \delta/(1+\delta)$, with remainder controlled by Corollary D.3 when only the Lipschitz modulus is assumed.

2. *If $\zeta_2 = 1 - \delta$ with $\delta \in (0,1)$, the step is quadratically subcritical. Nonlinear stiffening pushes it above the edge if and only if $h > 1/(1-\delta)$.*

3. *If ψ is nonincreasing and $\zeta_2 \le 1$, then $\zeta \le \zeta_2 \le 1$. If ψ is nondecreasing and $\zeta_2 \ge 1$, then $\zeta \ge \zeta_2 \ge 1$.*

Proof. Claims 1–2 are (64) rewritten in δ. Claim 3 is Corollary D.5. □

A quadratically supercritical step is therefore rescued precisely when the fractional drop in path-averaged directional curvature exceeds the fractional linear excess. A quadratically subcritical step is destabilized when stiffening consumes more than the quadratic margin. We only need to consider two event indicators, both measurable from (ζ_2, ζ) alone: *nonlinear rescue* is $\zeta_2 > 1$ and $\zeta \le 1$; *nonlinear destabilization* is $\zeta_2 \le 1$ and $\zeta > 1$. An ϵ-band $1 - \epsilon \le \zeta \le 1 + \epsilon$ is equivalent to $(1-\epsilon)/\zeta_2 \le h \le (1+\epsilon)/\zeta_2$. That is a one-step constraint on h, not a statement that a period-2 orbit remains in $1 \le \zeta < 1 + \epsilon$ (a nontrivial two-cycle with $A_t > 0$ must change the sign of $\zeta - 1$).

The factorization (52) also splits *changes* of ζ. On the positive domain,

$$\log \zeta \;=\; \log \chi + \log \varsigma_2 - \log \tau + \log h, \tag{66}$$

hence

$$\Delta \log \zeta \;=\; \Delta \log \chi + \Delta \log \varsigma_2 - \Delta \log \tau + \Delta \log h. \tag{67}$$

Basically, this is the same as the one in the main text, but with one additional granularity, decomposing the spatial participation ς into the quadratic participation ς_2 and higher-order maintenance h. Each summand is a channel of diagnostics, we may not treat them as causes while analyzing the dynamics. Therefore, we can read the diagnostics by following interpretations:

Channel	raises ζ	lowers ζ
χ	spectral sharpening	spectral relief
ς_2	occupancy of a sharper direction	rotation toward lower curvature
τ	loss of temporal calibration	recovery of alignment/calibration
h	pathwise stiffening	pathwise softening

Table 5: Optimizer grids on used in the paper. Learning rates are five geometrically spaced values. The displayed endpoints are the grid minima and maxima. "BC" is bias correction. p is the preconditioner exponent on the second moment. wd is decoupled weight decay.

Family	η grid	β_1 / β	β_2	ε	BC / p	wd
Adam	3×10^{-5}–10^{-3}	0.9	0.999	10^{-7}	yes / 0.5	0
AdamW	3×10^{-5}–10^{-3}	0.9	0.999	10^{-7}	yes / 0.5	0.01
Adafactor	10^{-5}–10^{-3}	0.9	0.8	10^{-7}	no / 0.5	0
AMSGrad	10^{-5}–10^{-3}	0.9	0.999	10^{-7}	no / 0.5	0
PAdam	10^{-3}–10^{-1}	0.9	0.999	10^{-7}	no / 0.25	0
NAdam	10^{-5}–10^{-3}	0.9	0.999	10^{-7}	no / 0.5	0
RMSProp	10^{-4}–10^{-3}	0	0.995	10^{-7}	no / 0.5	0
AdaGrad	10^{-2}–10^{-1}	0	1	10^{-10}	no / 0.5	0
GD	0.02–0.2	—	—	—	—	0
Heavy Ball	0.002–0.02	0.9	—	—	—	0
Nesterov	0.002–0.02	0.9	—	—	—	0
QHM	0.059375–0.59375	0.9	—	—	$\nu = 0.7$	0
two-pole $0.4/0.7$	0.0257–0.257	—	—	—	cascade	0
two-pole $0.7/0.85$	0.00681–0.0681	—	—	—	cascade	0

Entry from below ($\zeta_t < 1$ and $\Delta\log\zeta_t > 0$) is $\Delta\log\chi + \Delta\log\varsigma_2 + \Delta\log h > \Delta\log\tau$, return from above is the opposite inequality $\Delta\log\chi + \Delta\log\varsigma_2 + \Delta\log h < \Delta\log\tau$, and maintenance near $\zeta \simeq 1$ is approximate balance of the four increments. These are measurable channel-balance conditions. Note that these are not an attractor theorem: the quantities do not assert that higher-order terms should drive ζ toward 1, nor that h is the dominant channel.

E Additional Experiments

This section presents the experimental setup and all experimental results.

E.1 Implementation Details

Dataset. We use CIFAR-10 (Krizhevsky, 2009) dataset, all training samples are used. Every update uses one deterministic 50,000-example full batch without shuffling.

Architecture. The network flattens the $3\times32\times32$ input, applies five fully connected hidden layers of width 200 with $\tanh$ activations, and produces 10 logits. The objective is mean categorical cross-entropy on the full training set. This is exactly the same as the 5×200 architecture of Cohen et al. (2023), not the 2×200 network of Cohen et al. (2021).

Optimizers and their hyperparameters. Table 5 displays the learning rate grids and optimizer hyperparameters. AdamW weight decay is 0.01 and AdaGrad ε is 10^{-10}. Adam uses bias correction. Adafactor, AMSGrad, PAdam, NAdam, RMSProp and AdaGrad do not. PAdam uses partial-adaptivity exponent 0.25. QHM uses $\beta = 0.9$, $\nu = 0.7$. The two-pole filters are cascaded EMA poles at $(0.4, 0.7)$ and $(0.7, 0.85)$, i.e., Grokfast-style (Lee et al., 2024) cascades rather than parallel mixtures (Appendix C.2).

Temporal filters Q. The 14 optimizer names collapse to eight distinct causal filters Q, whose transfer functions are calculated in Table 4. Heavy Ball and Nesterov are the classical (not dc-normalized) recurrences, so $Q(1) = 1/(1-\beta) = 10$ at $\beta = 0.9$. Adam-family first moments are dc-normalized and share the EMA filter of the dc-normalized Heavy Ball. The two-pole rows are the only filters whose first root-locus exit is off Nyquist. Refer to Figure 7 for the full root locus.

E.2 Full Comparison of Worst-Case EoS χ and Realized EoS ζ

This section shows a full comparison of worst-case EoS χ and realized EoS ζ, extending Figure 1 in the main text. Figures 8 and Figures 9 show the results. Table 6 is the complete numerical

Table 6: Measured relative edge $\chi = \eta S\Gamma/2$ offsets for different optimizers. Each cell is the median of χ between first entry above 0.75 and last exit below 0.75 (open windows remain open at the horizon). A dash means the trajectory never exceeded 0.75. Extreme cases $\chi > 1.2$ are highlighted in boldface. *The manuscript marks AdaGrad and RMSProp as sparse edge-touch statistics.

Family	η_1	η_2	η_3	η_4	η_5	Center	n	χ range	ζ range
Adafactor	—	1.052	1.103	1.102	1.191	1.084	204/350	1.052–1.191	0.202442–0.965959
AdaGrad	1.424	1.556	1.845	4.047	21.070	**1.248***	3/350	1.424–21.070	0.982240–1.011240
Adam	0.927	0.980	0.971	0.976	0.974	0.960	259/350	0.927–0.980	0.177652–0.948673
AdamW	0.956	0.981	0.975	1.023	0.971	0.971	247/350	0.956–1.023	0.115807–0.936750
AMSGrad	0.789	0.931	0.979	0.946	0.949	0.960	203/350	0.789–0.979	−0.000206–0.836529
GD	0.904	1.130	1.246	1.478	1.624	1.124	47/140	0.904–1.624	0.000101–0.991144
Heavy Ball	—	0.820	0.960	0.934	0.964	0.949	165/400	0.820–0.964	0.000313–0.887029
NAdam	1.036	1.050	0.884	1.183	1.198	1.070	259/350	0.884–1.198	0.087206–0.971799
Nesterov	0.836	1.078	1.088	1.123	1.121	1.060	206/400	0.836–1.123	−0.000020–0.977996
PAdam	0.895	0.959	0.966	0.927	0.950	0.956	236/350	0.895–0.966	−0.000206–0.857368
QHM (0.9, 0.7)	0.961	1.238	1.309	1.454	1.576	1.100	116/350	0.961–1.576	0.000287–1.001553
RMSProp	1.349	1.452	1.568	1.579	1.646	**1.222***	12/350	1.349–1.646	0.978233–1.011893
Grokfast (0.4/0.7)	0.982	1.432	1.819	2.294	2.646	1.021	52/350	0.982–2.646	0.000152–0.964556
Grokfast (0.7/0.85)	—	0.988	1.354	1.625	2.049	1.040	61/350	0.988–2.049	0.000365–0.476646

results, measured as the median of χ on the closed interval from first $\chi > 0.75$ to last $\chi < 0.75$, or to the terminal checkpoint if the trajectory never returns below 0.75. Two settings never exceed 0.75 (Adafactor $\eta = 10^{-5}$, Heavy Ball $\eta = 0.002$, two-pole $0.7/0.85$ at $\eta = 0.00681$) and are recorded as dashes. Besides the discussion in Section 3, Table 6 also suggests that the offsets are not an artefact of decoupled weight decay. Adam ($\lambda = 0$) and AdamW ($\lambda = 0.01$) share the same temporal filter, the same η grid, and the same architecture; their edge-conditioned family centers are 0.960 and 0.971.

E.3 Full Decomposition of Realized EoS ζ

This section shows a full decomposition of realized EoS ζ for each individual setting of (optimizer, learning rate), extending Figure 4 in the main text. Full results are visualized in Figures 10 through Figure 13.

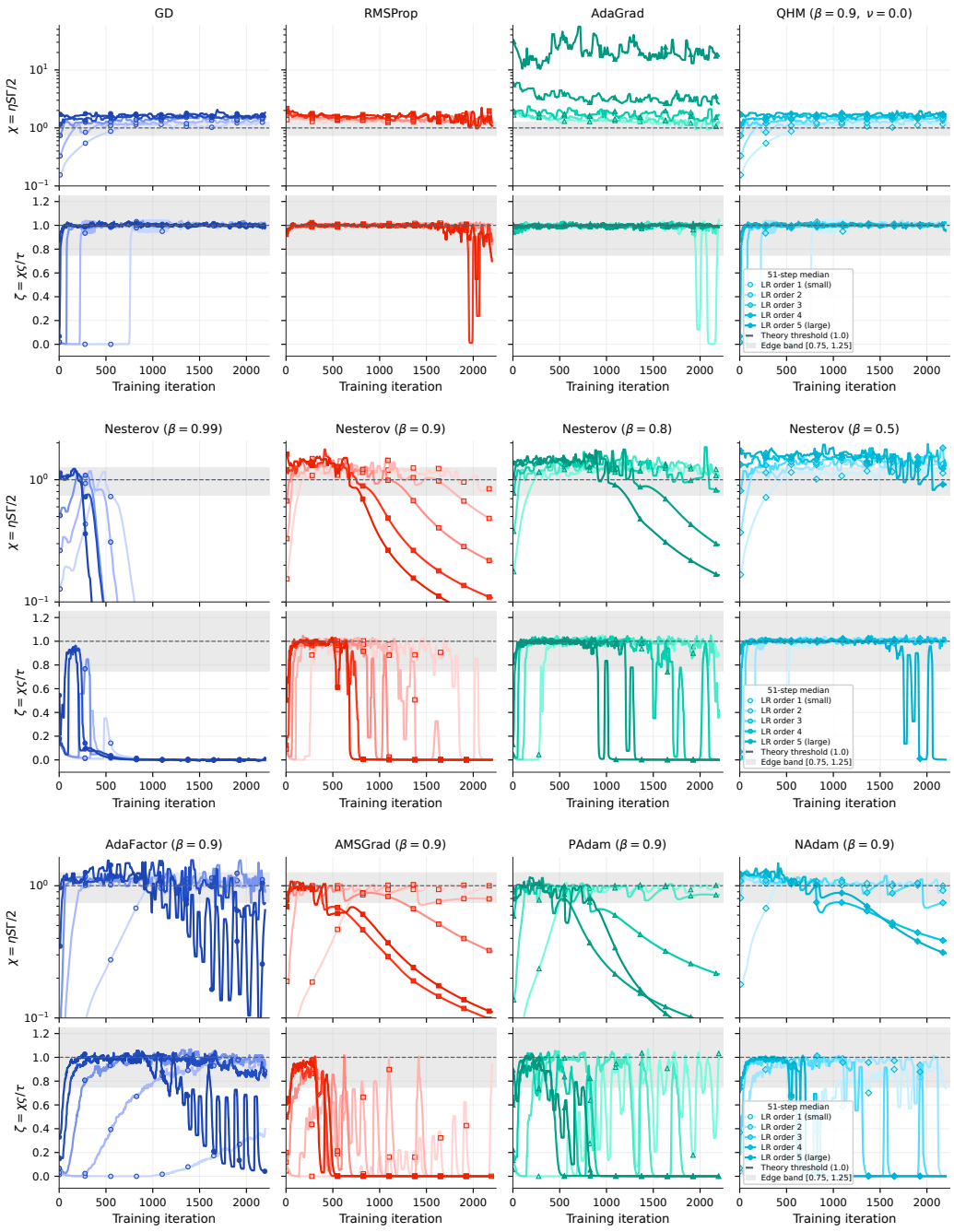


Figure 8: Comparison of worst-case EoS χ and realized EoS ζ on the primary CIFAR-10 experiments.

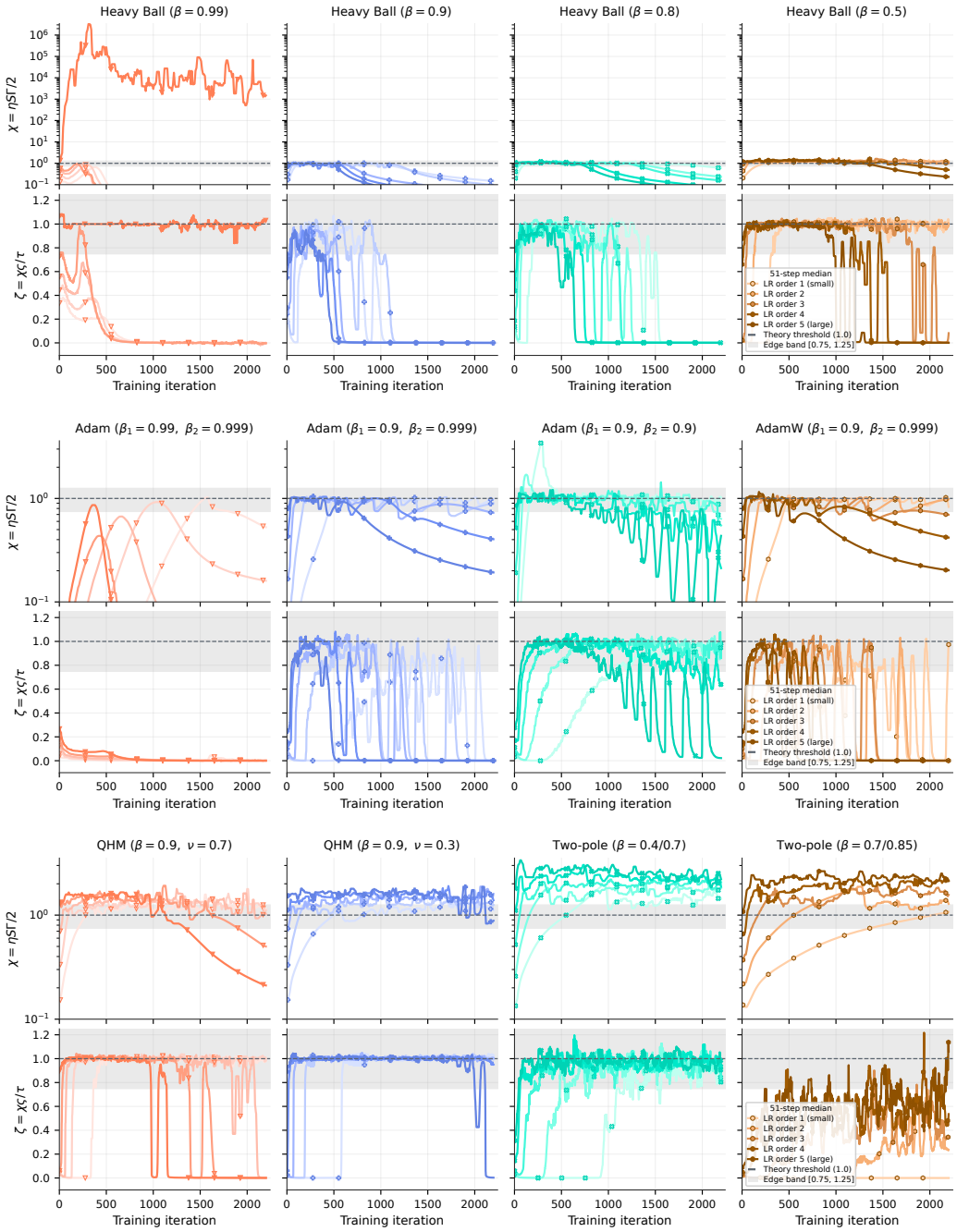


Figure 9: Comparison of worst-case EoS χ and realized EoS ζ on the primary CIFAR-10 experiments.

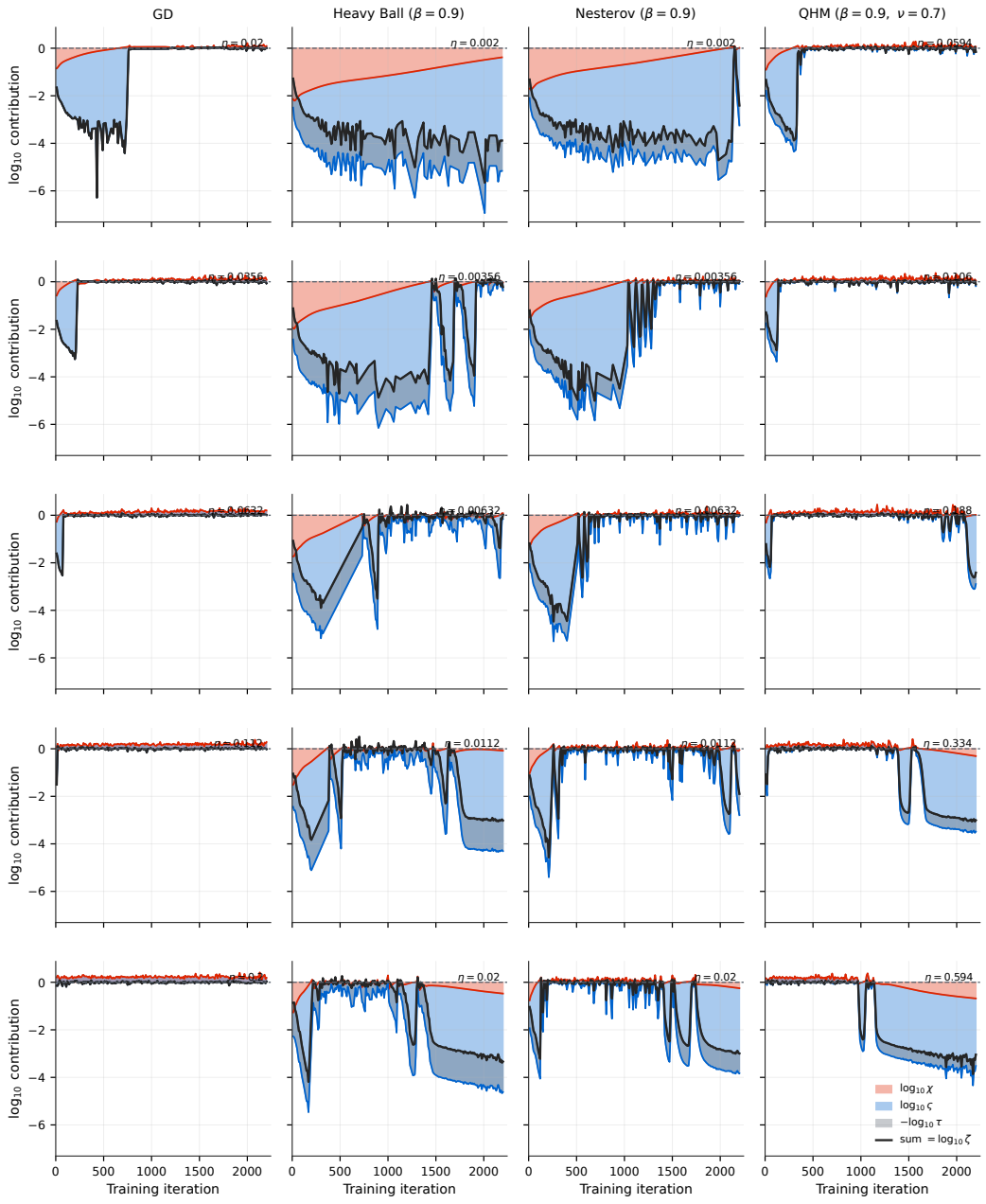


Figure 10: Decomposition of the realized EoS ζ into factors in log-scale, involving the (relative) edge of stability χ (red area and line), the spatial participation factor ς (blue area and line), and the temporal calibration factor τ (grey overlay and black line). Note that $\log \tau$ operates in the opposite direction by subtraction.

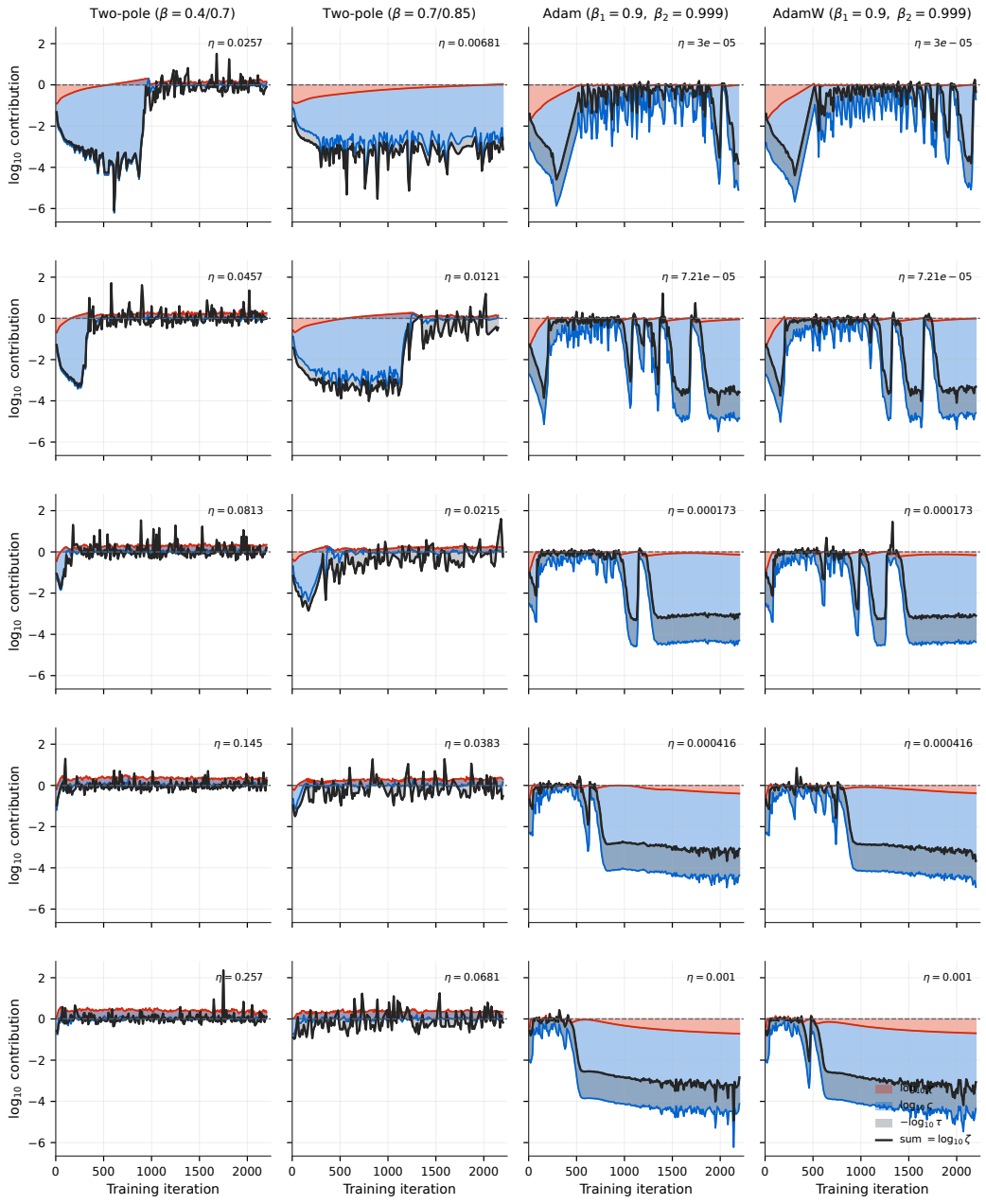


Figure 11: Decomposition of the realized EoS ζ into factors in log-scale, involving the (relative) edge of stability χ (red area and line), the spatial participation factor ς (blue area and line), and the temporal calibration factor τ (grey overlay and black line). Note that $\log \tau$ operates in the opposite direction by subtraction.

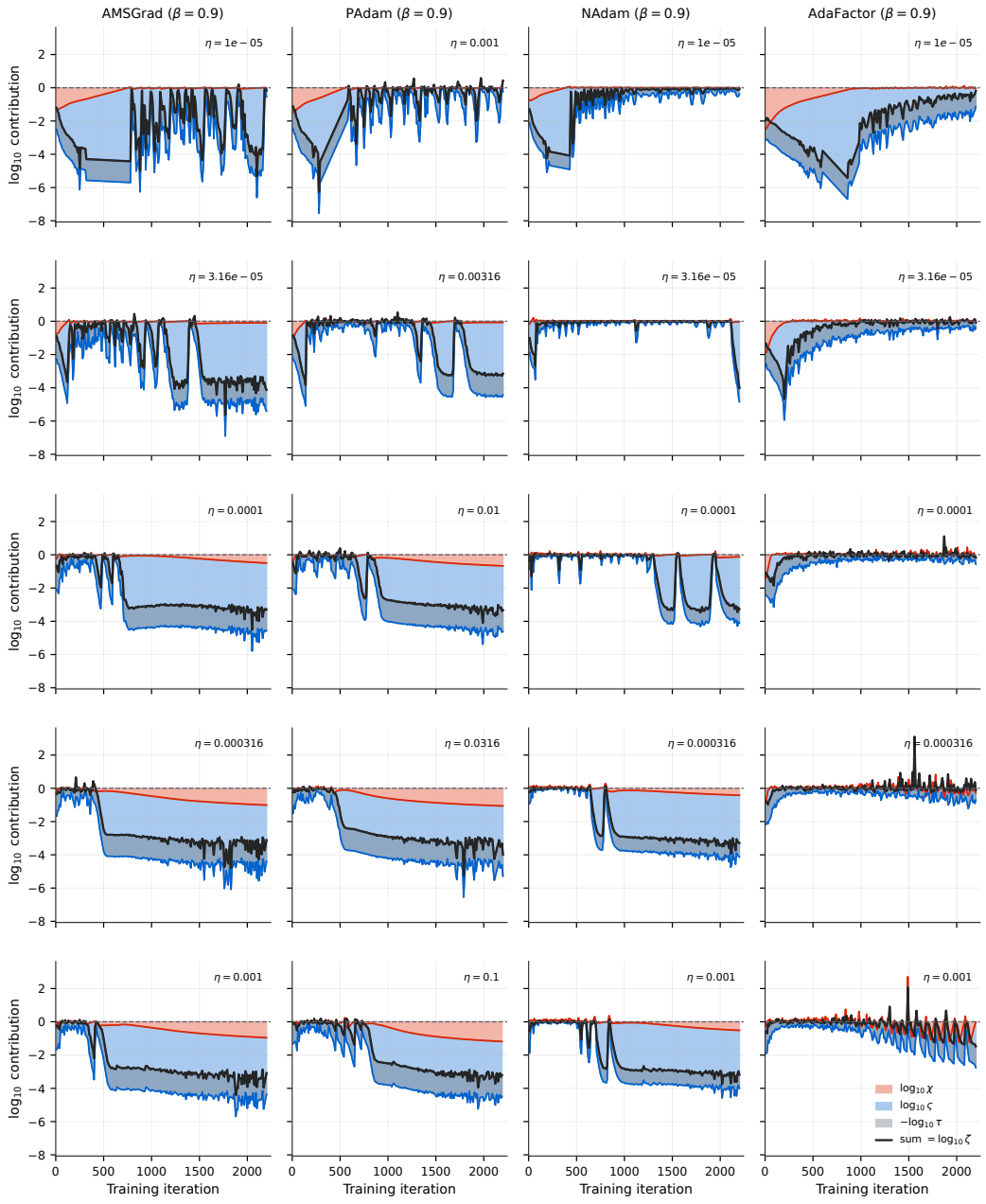


Figure 12: Decomposition of the realized EoS ζ into factors in log-scale, involving the (relative) edge of stability χ (red area and line), the spatial participation factor ς (blue area and line), and the temporal calibration factor τ (grey overlay and black line). Note that $\log \tau$ operates in the opposite direction by subtraction.

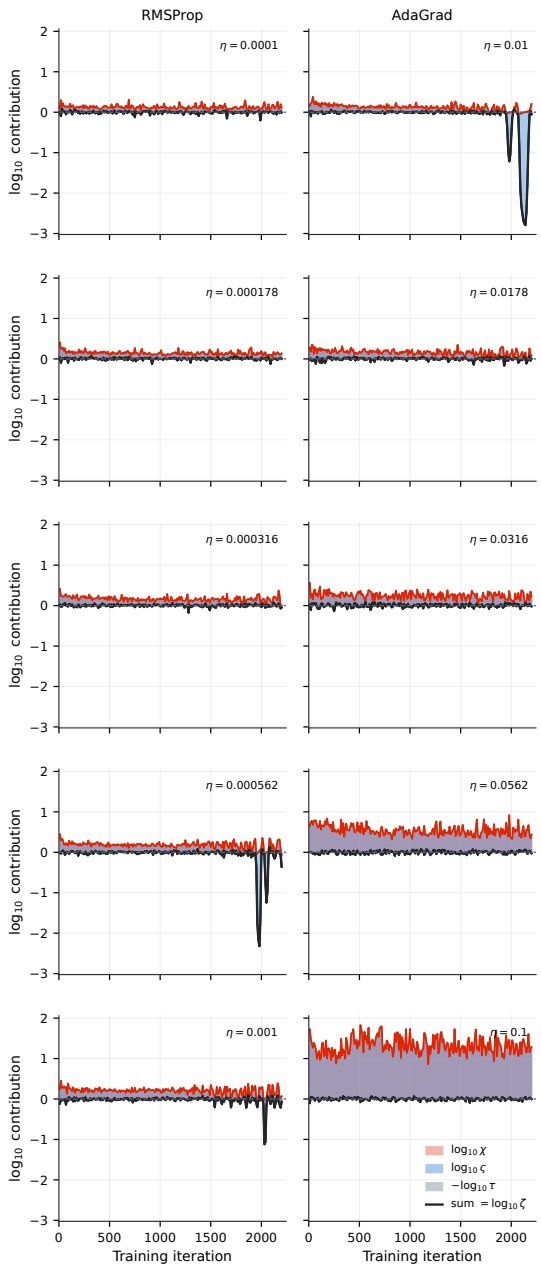


Figure 13: Decomposition of the realized EoS ζ into factors in log-scale, involving the (relative) edge of stability χ (red area and line), the spatial participation factor ς (blue area and line), and the temporal calibration factor τ (grey overlay and black line). Note that $\log \tau$ operates in the opposite direction by subtraction.