

THE CURIOUS CASE OF ADAMW

Anonymous authors

Paper under double-blind review

ABSTRACT

AdamW is ubiquitous in deep learning, yet its behavior remains poorly understood. We analyze its dynamics through the lens of dynamical systems and show that AdamW admits an *implicit objective*: its fixed points coincide with the stationary points of a constrained and regularized optimization problem. However, not all of these fixed points are stable under AdamW’s dynamics, and stability depends sensitively on curvature, weight decay, and momentum parameters. Even in simple one-dimensional settings, AdamW can exhibit surprisingly complex behavior: equilibria may be unstable and trajectories can fall into persistent limit cycles. We further extend the analysis to higher dimensions, deriving sufficient conditions for stability, and validate empirically that when AdamW converges in neural network training, it converges to stable equilibria. These results clarify what optimization problem AdamW is associated with, when convergence can be expected, and how its curious dynamics could inspire the development of more reliable optimization algorithms in the future.

1 INTRODUCTION

The Adam optimizer with decoupled weight decay (AdamW; (Loshchilov & Hutter, 2017)) has been the workhorse of modern deep learning. From large language models (Grattafiori et al., 2024; Liu et al., 2024), to vision architectures (Ravi et al., 2024), to generative models (Esser et al.), AdamW powers today’s most influential systems and is widely regarded as the default optimizer for deep learning.

Yet beneath this ubiquity lies a fundamental gap: while the convergence of Adam (without weight decay; (Kingma & Ba, 2014)) is already subtle to analyze, *no general convergence proof is available for AdamW*¹. The difficulty arises from the use of decoupled weight decay, which is essential in practice for stability and generalization (Kosson et al., 2023; D’Angelo et al., 2024), but remains disconnected from the objective, misaligned with the gradient, and fundamentally alters the algorithm’s dynamics.

Our contributions. We present two central findings that clarify AdamW’s behavior: what it implicitly optimizes, and how the stability of its dynamics informs its convergence.

- **Implicit objective of AdamW.** We show that the fixed points of AdamW coincide exactly with the stationary points of a constrained and regularized optimization problem.
- **Stability of AdamW.** We further show that not all stationary points of this implicit objective correspond to stable equilibria of AdamW. This gap enables simple counterexamples where **AdamW is non-convergent**, failing even on benign static objectives.

Implicit objective. Contrary to the common view that weight decay acts as “adaptive regularization,” we argue that an optimizer’s objective should depend only on parameters, not on internal states. By analyzing AdamW’s fixed-point conditions, we identify an implicit objective F whose stationary points coincide AdamW’s equilibria. This construction connects to a special case of the Lion- $\mathcal{K}$ algorithm (Chen et al., 2023a) and extends the conclusions of Xie & Li (2024). However, the equivalence is only first-order: AdamW does not necessarily minimize F , since some local minima of F may be unstable under AdamW’s dynamics, and the algorithm may fail to converge to them.

¹See section 6 for discussion.

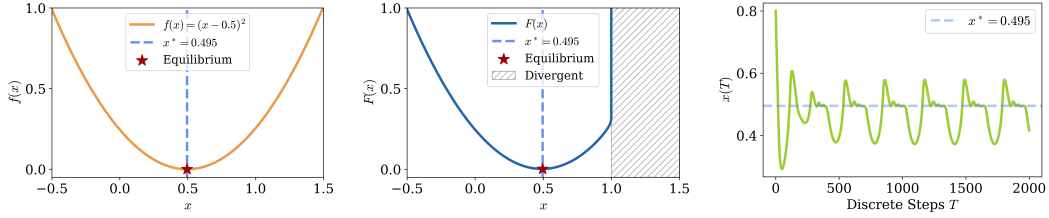


Figure 1: AdamW with $\eta = 10^{-2}$, $\beta_1 = 0.99$, $\beta_2 = 0.9$, $\lambda = 1$, $e = 10^{-2}$. Left: quadratic objective $f(x) = (x - \frac{1}{2})^2$. Middle: implicit objective $F(x)$ whose stationary point coincides with AdamW’s equilibrium. Right: AdamW iterates $x(T)$ over time, showing non-convergence despite the simple convex objective.

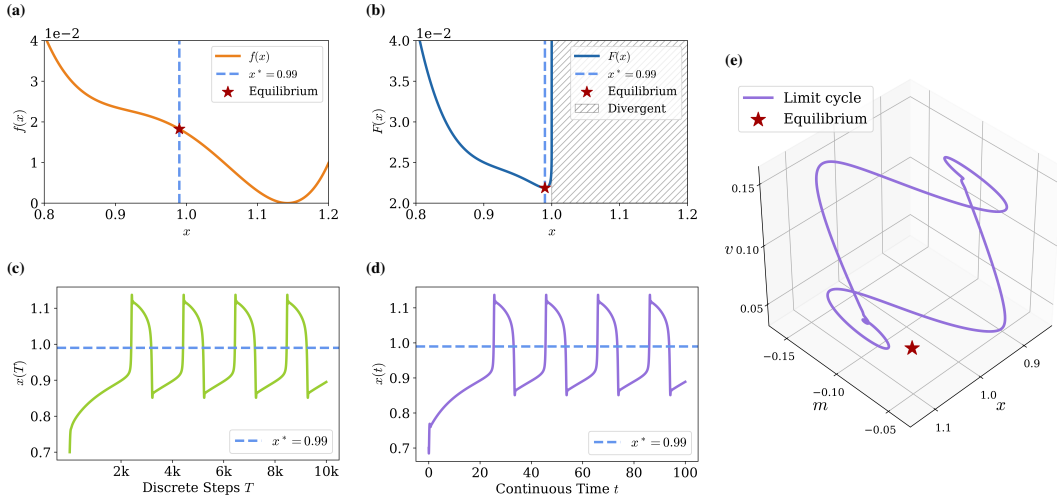


Figure 2: AdamW with $\eta = 10^{-2}$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\lambda = 1$, $e = 10^{-2}$. (a) Target objective $f(x)$ with fixed point at $x^* = 0.99$. (b) Implicit objective $F(x)$ whose stationary point coincides with the AdamW fixed point. (c) Discrete-time trajectory $x(T)$. (d) Continuous-time counterpart $x(t)$. (e) Phase portrait in (x, m, v) space. Together, (c–e) show that AdamW does not converge but instead enters a stable limit cycle around the equilibrium.

Stability of AdamW. Our analysis shows that stability and stationarity need not coincide for AdamW. This misalignment creates a stability gap that produces explicit counterexamples to convergence: when $\beta_1 > \beta_2$, AdamW diverges even on the simple quadratic $(x - \frac{1}{2})^2$, regardless of step size (see figure 1); in the common regime $\beta_1 < \beta_2$, convergence also fails, as iterates on simple one-dimensional polynomials settle into stable limit cycles, orbiting indefinitely rather than converging (see figure 2). These behaviors are not artifacts of adversarial hyperparameters or pathological online objectives (cf. Reddi et al., 2019), but intrinsic to AdamW’s update rule.

We study these dynamics through the lens of stability analysis. If AdamW converges, it necessarily converges to a local minima of the implicit objective, but sufficiency requires additional constraints on the Hessian of the original objective. If the only equilibrium in the domain is unstable, convergence is impossible and trajectories spiral into oscillations. In one dimension, when $\beta_1 > \beta_2$, stability requires an upper bound on curvature, which is often violated even for quadratics; when $\beta_1 < \beta_2$, stability requires a lower bound, so convergence occurs only if the Hessian is not too negative. In higher dimensions the situation is more involved, but local Lyapunov analysis yields tractable sufficient conditions. For neural networks, the Hessian is too large to characterize exactly, yet experiments reveal a consistent pattern: whenever AdamW converges, it does so at a stable equilibrium. Whether this requirement for stability explains AdamW’s empirical success or conceals hidden pitfalls remains open, but what is clear is that convergence is not guaranteed, and the curious dynamics of AdamW open intriguing directions for future research.

Algorithm 1 AdamW Update Rules ²**Require:** initial parameters x_0 , hyperparameters $\{\eta_t\}_{t=1}^T, \beta_1, \beta_2, e$, weight decay strength $\lambda > 0$ 1: Initialize moment $m_0 \leftarrow 0$, $v_0 \leftarrow 0$ 2: **for** $t = 1, 2, \dots, T$ **do**3: Compute gradient $g_t \leftarrow \nabla f_t(x_{t-1})$ 4: Update moment: $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)g_t$, $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)g_t^{\odot 2}$ 5: Bias correction: $\hat{m}_t \leftarrow m_t(1 - \beta_1^t)$, $\hat{v}_t \leftarrow v_t(1 - \beta_2^t)$

6: Parameter update:

$$x_t \leftarrow x_{t-1} - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + e} - \eta_t \cdot \lambda x_{t-1}$$

7: **end for**

2 BACKGROUND

We review the classical notions of Lyapunov stability and the *Lyapunov indirect method* (or *linearization method*), which provide standard tools for analyzing the stability of dynamical systems.

First, we establish a notion of stability of an equilibrium in a dynamical system. Formally, it is characterized with the following definition of Lyapunov stability.

Definition 2.1 (Lyapunov stability). *Consider the autonomous system $\dot{z}(t) = T(z(t))$, where $T : \mathcal{D} \rightarrow \mathbb{R}^d$ is locally Lipschitz with domain $\mathcal{D} \subset \mathbb{R}^d$. Let $z_* \in \mathcal{D}$ be an equilibrium, i.e. $T(z_*) = 0$. Then:*

- z_* is **Lyapunov stable** if for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|z(0) - z_*\| < \delta$ implies $\|z(t) - z_*\| < \epsilon$ for all $t \geq 0$.
- z_* is **asymptotically stable** if it is Lyapunov stable and there exists $\delta' > 0$ such that $\|z(0) - z_*\| < \delta'$ implies $\lim_{t \rightarrow \infty} z(t) = z_*$.
- z_* is **unstable** if it is not Lyapunov stable; i.e., there exists $\epsilon' > 0$ such that for every $\delta > 0$ one can find $\|z(0) - z_*\| < \delta$ but $\|z(t) - z_*\| \geq \epsilon'$ for some $t > 0$.

Lyapunov’s stability establishes the notion of global stability, while the following Hurwitz stability states how the linearized neighborhood near an equilibrium point behaves.

Definition 2.2 (Hurwitz stability). *Let z_* be an equilibrium of $\dot{z} = T(z)$ and let $J(z_*) = [\partial T_i / \partial z_j]_{z_*}$ denote the Jacobian at z_* . Then z_* is **Hurwitz stable** if every eigenvalue s of $J(z_*)$ satisfies $\Re(s) < 0$.*

Lyapunov’s indirect method gives the following characterization. If $J(z_*)$ is Hurwitz stable, then the equilibrium z_* is *asymptotically stable*. If $J(z_*)$ has an eigenvalue with $\Re(s) > 0$, the equilibrium is *unstable*. When all eigenvalues satisfy $\Re(s) \leq 0$ with some lying on the imaginary axis, the test is inconclusive and higher-order terms must be examined. We direct readers to Khalil (2002) for further details.

While Lyapunov’s indirect method addresses continuous-time systems, a parallel criterion applies to discrete dynamics of the form $z_{t+1} = T(z_t)$. In this case, an equilibrium z_* is *Schur stable* if all eigenvalues of $J(z_*)$ lie strictly inside the unit disk, i.e. $|\lambda| < 1$ for every eigenvalue λ . This condition ensures asymptotic stability. If any eigenvalue satisfies $|\lambda| > 1$, the equilibrium is unstable. Thus, stability analysis for discrete-time optimization algorithms is analogous to the continuous-time setting.

Notation We denote the parameter vector by x , the first-order momentum by m , and the second-order momentum by v . The objective function is $f : \mathbb{R}^d \rightarrow \mathbb{R}$, defined over the parameter domain. We use $F(\cdot)$ to denote the *implicit objective*, which will be introduced later. Unless stated otherwise,

²We use the default AdamW implementation from PyTorch (PyTorch Contributors, 2025), which computes the denominator as $\sqrt{\hat{v}_t} + e$ rather than $\sqrt{\hat{v}_t + e^2}$. While the alternative will lead to a slightly different implicit objective, our main conclusions remain unaffected.

all operations are elementwise. To avoid ambiguity, we use the following terminology: *equilibria* are stationary points of continuous-time dynamical systems, *fixed points* are update-invariant states of discrete-time algorithms, and *stationary points* are points where the gradient of a function vanishes.

3 THE IMPLICIT OBJECTIVE OF ADAMW

We start from the *fixed-point conditions* of AdamW, which characterize the relations that convergence points must satisfy if convergence were to occur. For gradient descent or Adam, these fixed points coincide with the stationary points of the objective $f(\mathbf{x})$. In contrast, the decoupled weight decay alters this relationship: equilibria no longer align directly with stationary points of $f(\mathbf{x})$ but also depend on the weight decay strength.

Formally, let $[\mathbf{x}_t, \mathbf{m}_t, \mathbf{v}_t]$ denote the AdamW iterates. As $t \rightarrow \infty$, bias corrections vanish so that $\hat{\mathbf{m}}_t \rightarrow \mathbf{m}_t$ and $\hat{\mathbf{v}}_t \rightarrow \mathbf{v}_t$. At a fixed point, the updates are stationary, and the iterates satisfy

$$\frac{\mathbf{m}}{\sqrt{\mathbf{v} + \mathbf{e}}} + \lambda \mathbf{x} = \mathbf{0}, \quad \mathbf{m} = \nabla f(\mathbf{x}), \quad \mathbf{v} = \nabla f(\mathbf{x})^{\odot 2}. \quad (1)$$

Eliminating $\mathbf{m}$ and $\mathbf{v}$ yields the necessary condition

$$\frac{\nabla f(\mathbf{x})}{|\nabla f(\mathbf{x})| + e} + \lambda \mathbf{x} = \mathbf{0}. \quad (2)$$

Equation 2 exactly describes the fixed-point set of AdamW. These equilibria turn out to coincide with the stationary points of a constrained and regularized optimization problem, which we refer to as the *implicit objective* of AdamW.

Theorem 3.1. *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be continuously differentiable, and let $\lambda, e > 0$. Then $\mathbf{x} \in \mathbb{R}^d$ satisfies the fixed-point condition equation 2 if and only if $\mathbf{x}$ is a stationary point of the ℓ_∞ -constrained and regularized problem*

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) = f(\mathbf{x}) + \frac{e}{\lambda} \Phi(\lambda \mathbf{x}), \quad \text{subject to} \quad \|\mathbf{x}\|_\infty \leq \frac{1}{\lambda}, \quad (3)$$

where $\Phi(\mathbf{x}) = \sum_{i=1}^d \phi(x_i)$ with

$$\phi(x) = -|x| - \log(1 - |x|), \quad |x| < 1.$$

Moreover, since $\phi(x) \rightarrow +\infty$ as $|x| \rightarrow 1$, the regularizer Φ serves as a barrier function that naturally enforces the constraint $\|\mathbf{x}\|_\infty < \frac{1}{\lambda}$.

This implicit objective coincides with the Softsign case $\nabla \mathcal{K} = x/(|x| + e)$ of Lion- $\mathcal{K}$ (Appendix A, Chen et al. (2023a)), and it reduces to the pure ℓ_∞ constraint of Xie & Li (2024) when $e \rightarrow 0$. Additionally, this view of AdamW’s implicit objective contrasts with prior interpretations (Loshchilov & Hutter, 2017; Zhou et al., 2024), which describe weight decay as an adaptive ℓ_2 penalty that depends on optimizer states and therefore deviates from the standard notion of an optimization objective. In contrast, our construction produces a static objective $F(\mathbf{x})$ that depends only on the parameters.

3.1 FIXED-POINT COINCIDENCE IS INSUFFICIENT

The fact that AdamW’s fixed points coincide with the stationary points of equation 3 does not imply that AdamW actually minimizes $F(\mathbf{x})$. Beyond this first-order consistency, we require:

The stable equilibria of AdamW must coincide with the local minima of F .

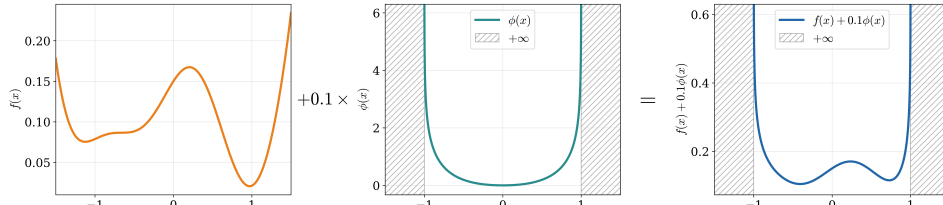


Figure 3: Left: an objective function f that AdamW is applied on. Middle: ϕ . Right: $f(\mathbf{x}) + \frac{e}{\lambda} \phi(\lambda \mathbf{x})$ with $e = 0.1$ and $\lambda = 1$.

For many standard optimization algorithms this alignment holds, often guaranteed by a global Lyapunov function: stable equilibria coincide exactly with local minima. As proved in previous works (Maddison et al., 2018; Chen et al., 2023a), this property is shared across the Lion- $\mathcal{K}$ family, including SGD with momentum, Lion (Chen et al., 2023b), and Muon (Jordan et al., 2024). AdamW, however, is different. A global Lyapunov function *may not exist*, and the equivalence between stable equilibria and local minima holds unconditionally only in the degenerate case $\beta_1 = \beta_2$. When $\beta_1 \neq \beta_2$, as will be shown in the next section, being a local minimum of F is *necessary but not sufficient* for stability.

$$\text{Stable equilibria of AdamW} \subseteq \text{Local minima of } F$$

Consequently, some local minima of F are unstable under AdamW. Since Theorem 3.1 shows that any implicit objective minimized by AdamW must be F (up to composition with a strictly monotone C^1 function), this misalignment implies that AdamW may not minimize any general objective function. It also enables explicit counterexamples where the dynamics fail to converge to a fixed point and instead evolve into stable limit cycles.

4 STABILITY CONDITIONS AND NON-CONVERGENCE OF ADAMW

We work in the deterministic (batch) gradient setting and analyze AdamW in the continuous-time limit, corresponding to infinitesimal step size. This removes the intervention from gradient noise and learning rate schedules. Importantly, *a sound optimization method should at least converge in this idealized regime*; instability here indicates a fundamental issue in the design. Continuous-time dynamics also require weaker conditions for convergence: gradient flow converges to local minima even without smoothness assumptions (Łojasiewicz, 1984), whereas discrete-time methods such as gradient descent require smoothness and step-size control (Lee et al., 2016). While an analogous discrete-time analysis is possible, the presence of learning rates introduces additional complications that render the resulting expressions far less interpretable; we therefore defer this analysis to Appendix A.4. In practice, learning rates are often small, so continuous-time analysis provides a useful proxy for understanding the behavior of discrete algorithms.

We model the continuous-time dynamics of AdamW by the ODE

$$\dot{x}_t = -\frac{m_t}{h_t + e} - \lambda x_t, \quad \dot{m}_t = a(\nabla f(x_t) - m_t), \quad \dot{h}_t = \frac{b}{2} \left(\frac{\nabla f(x_t)^{\odot 2}}{h_t} - h_t \right). \quad (4)$$

The discrete-time update of AdamW corresponds to a semi-implicit Euler discretization of this system with $v_t = h_t^{\odot 2}$ and the identifications $\eta = \Delta t$, $\beta_1 = 1 - a\Delta t$, $\beta_2 = 1 - b\Delta t$. For large t we adopt the standard approximation $\hat{m}_t \approx m_t$, $\hat{v}_t \approx v_t$.

It is convenient to rewrite the system in vector form (suppressing t for clarity):

$$\frac{d}{dt} z = T(z), \quad z := \begin{bmatrix} x \\ m \\ h \end{bmatrix}, \quad T(z) := \begin{bmatrix} -m/(h + e) - \lambda x \\ a(\nabla f(x) - m) \\ \frac{b}{2}(\nabla f(x)^{\odot 2}/h - h) \end{bmatrix}. \quad (5)$$

Let $z_* = [x_*, m_*, h_*]$ denote an equilibrium point of the dynamics, that is, a point satisfying $T(z_*) = 0$. Differentiating T and evaluating at z_* gives the Jacobian

$$J(z_*) = \begin{bmatrix} -\lambda I & -\text{Diag}(1/(h_* + e)) & \text{Diag}(m_*/(h_* + e)^2) \\ a\nabla^2 f(x_*) & -aI & 0 \\ b\text{Diag}(\nabla f(x_*)/h_*)\nabla^2 f(x_*) & 0 & -bI \end{bmatrix}. \quad (6)$$

4.1 LOCAL STABILITY OF ADAMW IN 1D

We first analyze the one-dimensional (1D) case, deferring the higher-dimensional analysis to section 5. Here “1D” refers only to the dimension of x , while the full dynamics remain three-dimensional in $[x, m, h]$. As discussed in section 2, the local stability of an equilibrium is determined by the spectrum of $J(z_*)$. To study this spectrum, we use the characteristic polynomial

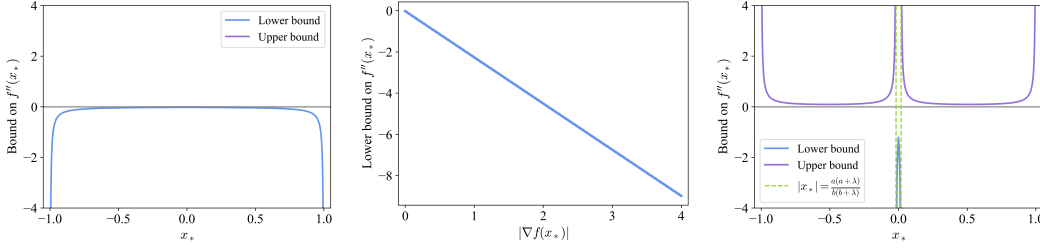


Figure 4: Left: lower bound on $f''(x_*)$ (equation 9) as a function of x_* when $a = 10, b = 1$. Middle: the bound as a function of gradient norm $|\nabla f(x_*)|$. Right: case $a = 1, b = 10$, where the condition yields either a lower or an upper bound depending on $|x_*|$. We use $\lambda = 1, e = 10^{-2}$ in both cases.

$\chi_J(s) := \det(sI - J(z_*))$, whose roots coincide with the eigenvalues of $J(z_*)$. In the 1D case, this polynomial simplifies to a cubic:

$$\chi_J(s) = s^3 + (a + b + \lambda)s^2 + \left[ab + a\lambda + \frac{(a - b)H_*}{\hat{h}_*} + \frac{be\hat{H}_*}{\hat{h}_*^2}\right]s + \frac{abe\hat{H}_*}{\hat{h}_*^2}. \quad (7)$$

Here the coefficients are expressed in terms of the following quantities:

$$\hat{h}_* := h_* + e, \quad H_* := f''(x_*), \quad \hat{H}_* := F''(x_*),$$

where F is the implicit objective introduced in equation 3. Since z_* is an equilibrium point, we also use the relations

$$h_* = |\nabla f(x_*)| = |m_*|, \quad \hat{H}_* = H_* + \frac{\lambda}{e}h_*^2.$$

To analyze the signs of the eigenvalues, we recall this following fact.

Lemma 4.1 (Routh–Hurwitz criterion for cubics). *All roots of the cubic equation $s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0 = 0$ have negative real part (i.e., the system is Hurwitz stable) if and only if*

$$\alpha_2, \alpha_1, \alpha_0 > 0 \quad \text{and} \quad \alpha_2 \alpha_1 > \alpha_0.$$

Stable equilibria of AdamW $\subseteq$ Local minima of F We first show that any stable equilibrium of the AdamW dynamics must correspond to a local minimum of the implicit objective F .

Corollary 4.2. *Assume $a, b, \lambda, e > 0$ and consider the one-dimensional case. If z_* is a Hurwitz stable equilibrium of the AdamW ODE, then it must be a strict local minimum of the implicit objective F , i.e. $F''(x_*) > 0$.*

Proof. If z_* is Hurwitz stable, then $\alpha_0 = abe\hat{H}_*/\hat{h}_*^2 > 0$, which implies $\hat{H}_* = F''(x_*) > 0$. $\square$

The subtlety, however, is that $F''(x_*) > 0$ is necessary but *not sufficient*. Even strict local minima of F may correspond to unstable equilibria of AdamW. An additional bound on the Hessian of the original objective $f''(x_*)$ is required.

Proposition 4.3. *Assume $a, b, e > 0$ and $\lambda \geq 0$. Let $[x_*, m_*, h_*]$ be an equilibrium of the AdamW ODE in equation 5. Then $[x_*, m_*, h_*]$ is Hurwitz stable if and only if $F''(x_*) > 0$ and*

$$(a + b + \lambda) \left[ab + a\lambda + \frac{(a - b)f''(x_*)}{\hat{h}_*} \right] + (b + \lambda)be \frac{F''(x_*)}{\hat{h}_*^2} > 0, \quad (8)$$

which can be rearranged into

$$(a(a + \lambda) - b(b + \lambda)|\lambda x_*|) f''(x_*) > -\frac{eC}{(1 - |\lambda x_*|)}, \quad (9)$$

where $C = (a + b + \lambda)(ab + a\lambda + b\lambda) - ab\lambda > 0$.

Thus, stability requires two conditions: first, the implicit Hessian must satisfy $F''(x_*) > 0$; second, the inequality in equation 9 must hold, which imposes an additional constraint on the Hessian $f''(x_*)$ of the original loss f . We illustrate this constraint in figure 4, showing how the bound behaves for both $a < b$ and $a > b$, and how it depends on the gradient norm $|\nabla f(x_*)|$. In this sense, one may say

AdamW is a faithful but selective optimizer of $F(x)$.

Equivalently, AdamW can be interpreted as minimizing $F(x)$, but only under the additional constraint specified by equation 9.

Remark 4.4 (When the additional condition vanishes). *The extra requirement in equation 9 disappears in two important cases:*

1. **Adam limit** ($\lambda \rightarrow 0$). *The implicit objective reduces to $F(x) = f(x)$ as the regularizer and constraint vanishes. The extra requirements becomes $a^2 f''(x_*) > -eC$, which is automatically implied by $F''(x_*) = f''(x_*) \geq 0$. Thus stability reduces to the usual requirement that x_* be a strict local minimum of f , explaining that the additional complication arises entirely from the weight-decay term.*
2. **Equal momentum coefficients** ($a = b$). *In this case, equation 8 simplifies into $F''(x_*) > -L$ for $L \geq 0$, which is implied by $F''(x_*) > 0$. Hence stability holds if and only if x_* is a strict local minimum of F . (See more in appendix A.8.)*

In both cases stability reduces to $F''(x_) > 0$, so the extra bound arises only from weight decay combined with unequal momentum coefficients a and b .*

4.2 EXAMPLES OF NON-CONVERGENCE

Whenever $a \neq b$, one can construct objectives f that violate the stability condition equation 8 while still satisfying $F''(x_*) > 0$. This creates local minima of F that are stable in the usual optimization sense but unstable under AdamW dynamics. If such a point is the only minimum of F in the domain, the system has no stable fixed point; since trajectories remain bounded, this necessarily yields non-convergent behavior such as limit cycles or other recurrent dynamics. The type of violation depends on the relation between a and b : when $a < b$ (equivalently, $\beta_2 < \beta_1$), the condition fails for sufficiently large f'' , so even strongly convex functions can break stability, which explains this regime is rarely used in practice. By contrast, when $a > b$ (equivalently, $\beta_2 > \beta_1$), which is the regime most common in practice, non-convergence arises only for non-convex objectives. We now provide simple counterexamples illustrating both cases.

Example 4.5. *When $\beta_2 < \beta_1$ (so $a < b$), the extra condition in equation 9 imposes an upper bound $f''(x_*) \leq U$. This can fail even for simple convex functions. Consider the quadratic*

$$f(x) = (x - 0.5)^2.$$

With $\lambda = 1$ and $e = 10^{-2}$, the equilibrium is $x_ \approx 0.49$. Choosing $\beta_1 = 0.99$, $\beta_2 = 0.9$, and step size $\eta = 10^{-2}$ gives $a = 1$, $b = 10$. Then equation 9 requires $f''(x_*) < 9.1 \times 10^{-2}$, while in fact $f'' = 2$. Hence the only equilibrium is unstable, and AdamW fails to converge (see figure 1).*

Example 4.6. *When $\beta_2 > \beta_1$ (so $a > b$), the condition imposes a lower bound $f''(x_*) \geq -L$. This can be violated with a carefully chosen non-convex quartic. Consider*

$$f(x) = 20(x - 0.99)^4 - 0.60(x - 0.99)^2 - 0.099(x - 0.99) + 0.018.$$

With $a = 10$, $b = 1$ (corresponding to $\beta_1 = 0.90$, $\beta_2 = 0.99$), $\eta = 10^{-2}$, $e = 10^{-3}$, and $\lambda = 1$, the equilibrium at $x_ = 0.99$ has $f''(x_*) \approx -0.60$, violating the required bound $f''(x_*) > -0.22$. Thus the equilibrium is unstable, and AdamW again fails to converge (see figure 2).*

We emphasize that these counterexamples are primarily of theoretical interest, meant to highlight the intricate dynamics of AdamW. In practice, realizing such cases in neural networks would typically require extreme settings, such as unusually large weight decay (e.g., $\lambda > 20$ to drive some $|\lambda x|$ close enough to 1) or deliberately enforcing the atypical regime $\beta_1 > \beta_2$. Neither choice reflects standard training practice. This is unsurprising: modern recipes for network training have been shaped by years of trial and error and by the coevolution of optimizers and architectures, which has naturally filtered out such pathological configurations.

5 MULTI-DIMENSIONAL STABILITY OF ADAMW

In the multi-dimensional setting, the analysis becomes more involved and there is no simple stability criterion. We adopt an alternative approach: we transform the linearized system into a third-order ODE and then derive a sufficient condition for convergence using a Lyapunov function, which in turn yields a sufficient condition for stability of equilibria. The linearized dynamics are given by the following lemma.

Lemma 5.1. *Let $\mathbf{z}_* = (\mathbf{x}_*, \mathbf{m}_*, \mathbf{h}_*)$ be an equilibrium point of the AdamW ODE equation 5. If $a \neq b$, then the linearized dynamics $\dot{\mathbf{z}}_t = \mathbf{J}(\mathbf{z}_*)\mathbf{z}_t$ imply that $\mathbf{x}_t$ satisfies*

$$\ddot{\mathbf{x}}_t + \mathbf{A}_2\ddot{\mathbf{x}}_t + \mathbf{A}_1\dot{\mathbf{x}}_t + \mathbf{A}_0\mathbf{x}_t = 0, \quad (10)$$

where

$$\begin{aligned} \mathbf{A}_2 &= (a + b + \lambda)\mathbf{I} & \mathbf{A}_0 &= ab [\text{Diag}(e/(\mathbf{h}_* + e)^2)\nabla^2 f(\mathbf{x}_*) + \lambda\mathbf{I}] \\ \mathbf{A}_1 &= (ab + a\lambda + b\lambda)\mathbf{I} + (a - b) \text{Diag}(1/(\mathbf{h}_* + e))\nabla^2 f(\mathbf{x}_*) + b \text{Diag}(e/(\mathbf{h}_* + e)^2)\nabla^2 f(\mathbf{x}_*), \end{aligned}$$

We can further rewrite equation 10 as a 3d-dimensional first-order ODE system:

$$\dot{\mathbf{x}}_t = \mathbf{y}_t, \quad \dot{\mathbf{y}}_t = \mathbf{z}_t, \quad \dot{\mathbf{z}}_t = -\mathbf{A}_0\mathbf{x}_t - \mathbf{A}_1\mathbf{y}_t - \mathbf{A}_2\mathbf{z}_t.$$

To analyze stability, we construct a Lyapunov function of the form

$$V(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{1}{2} \left(\|\mathbf{A}_2\mathbf{x} + \mathbf{y}\|_P^2 + \|\mathbf{A}_2\mathbf{y} + \mathbf{z}\|_Q^2 + \|\mathbf{y}\|_K^2 \right),$$

where $\mathbf{P}, \mathbf{Q}, \mathbf{K}$ are defined as $\mathbf{P} = \mathbf{Q}\mathbf{A}_0\mathbf{A}_2^{-1}$, $\mathbf{K} = \mathbf{Q}(\mathbf{A}_1 - \mathbf{A}_0\mathbf{A}_2^{-1})$, and $\mathbf{L} = \mathbf{A}_2^\top \mathbf{Q}(\mathbf{A}_1 - \mathbf{A}_0\mathbf{A}_2^{-1})$, with $\mathbf{Q}$ chosen so that $\mathbf{P}, \mathbf{Q}, \mathbf{K}$ are positive definite. With these definitions, we obtain

$$\frac{d}{dt}V(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) = -\frac{1}{2}\mathbf{y}^\top(\mathbf{L} + \mathbf{L}^\top)\mathbf{y}.$$

Thus $\dot{V} \leq 0$ whenever $\mathbf{L} + \mathbf{L}^\top$ is positive definite, leading to the following result.

Proposition 5.2. *Consider the system in equation 10. The dynamics are asymptotically stable if there exists a symmetric positive definite matrix $\mathbf{Q}$ such that 1. $\mathbf{Q}\mathbf{A}_0$ and $\mathbf{Q}\mathbf{A}_1$ are symmetric, and 2. $\mathbf{Q}\mathbf{A}_0$, $\mathbf{Q}\mathbf{A}_1$, and $\mathbf{Q}(\mathbf{A}_1\mathbf{A}_2 - \mathbf{A}_0)$ are positive definite.*

To complement the multidimensional analysis, we provide a small empirical check on a neural network. We train a three-layer network with hidden dimension 256 and GELU activations (Hendrycks & Gimpel, 2016) on a 5000-sample subset of MNIST, using batch gradients and a continuous-time simulation with Runge–Kutta 4 (step size 2×10^{-3} , 100k steps). We vary the weight decay λ and report the update size $\|\Delta\theta\|_2$ between consecutive steps (Fig. 5, left). For small λ ($\lambda = 0$ reducing to Adam or 0.1), updates remain steady and bounded; for large λ ($\lambda = 100$), updates fluctuate strongly, indicating non-convergence; for an intermediate $\lambda = 5$, updates quickly decay to machine precision, suggesting convergence.

We further repeat this $\lambda = 5$ setting with 20 random seeds and compute the spectral abscissa of the Jacobian equation 6 at the final equilibrium, i.e. the eigenvalue with largest real part. The histogram in Fig. 5 (right) shows all values strictly negative, indicating that the equilibrium is Hurwitz and consistent with AdamW reaching a stable equilibrium when convergence occurs.

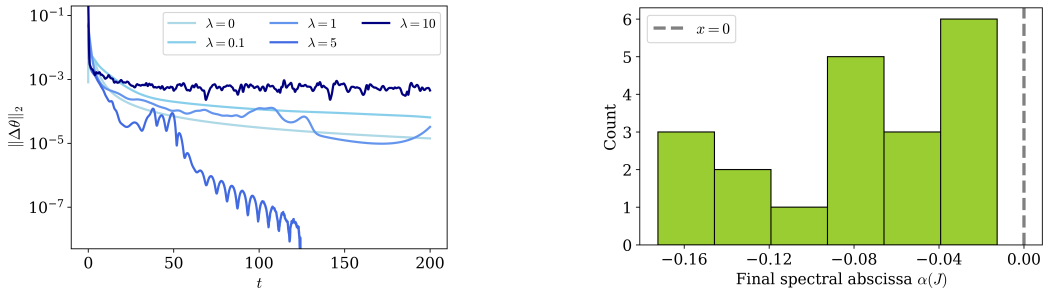


Figure 5: Left: Update size $\|\Delta\theta\|_2$ during training with varying weight decay λ . Right: Histogram of the spectral abscissa of $\mathbf{J}(\mathbf{z}_*)$ at the final equilibrium for $\lambda = 5$ over 20 runs.

6 RELATED WORKS

Convergence analysis of Adam. Adam was introduced by Kingma & Ba (2014) with a convergence proof for convex functions in online convex optimization, but this argument was later shown to contain a gap by Reddi et al. (2019). Their counterexample, however, relies on carefully constructed adversarial online objectives, and thus does not preclude convergence under standard settings (Zhang et al., 2022). Follow-up work established several positive results: De et al. (2018) and Défossez et al. (2020) proved convergence for non-convex functions under bounded gradients and consistent gradient signs, while Zhang et al. (2022) relaxed the bounded-gradient requirement. More recently, Li et al. (2023); Jin et al. (2024) demonstrated convergence under weaker assumptions, narrowing the gap with SGD. A complementary line of work studies Adam’s implicit bias and seeks to explain its empirical success (Xie et al., 2024; Vasudeva et al., 2025; Li et al.).

Adam with decoupled weight decay. In practice, AdamW (Loshchilov & Hutter, 2017), the decoupled weight decay variant of Adam, has become the default choice over the original Adam. Weight decay itself was introduced by Hanson & Pratt (1988), and while it is equivalent to L_2 regularization in SGD, Loshchilov & Hutter (2017) showed this equivalence breaks for adaptive methods and that decoupling improves performance. Since then, the effectiveness of weight decay has been confirmed across architectures and scales (Kosson et al., 2023; D’Angelo et al., 2024; Wang & Aitchison, 2024; Bergsma et al., 2025; He et al., 2025).

No conflicts with previous studies of AdamW’s convergence. Several works have analyzed AdamW’s convergence, but under assumptions that depart from the algorithm as commonly used. Zhou et al. (2024) considered exponentially decaying weight decay $\lambda_k = \lambda(1 - \alpha)^k$, in contrast to the standard constant choice, and defined convergence through $\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}[\|\nabla F_k(\mathbf{x}_k)\|_2^2]$, where F_k is a state-dependent “dynamically regularized” objective. Recently Li et al. (2025) established a rate of $\mathcal{O}(d^{1/2}T^{-1/4})$ for the criterion $\frac{1}{T} \sum_{k=0}^{T-1} \mathbb{E}[\|\nabla f(\mathbf{x}_k)\|_1] \leq \epsilon^2$, but their result requires that the iterates already satisfy the ℓ_∞ constraint and assumes small weight decay ($\lambda \leq \mathcal{O}(d^{1/2}T^{-3/4})$) together with large momentum coefficients ($\beta_1 = \mathcal{O}(T^{-1/2})$). Xie & Li (2024) examined AdamW’s implicit objective in the limit $e \rightarrow 0$, showing convergence to KKT points under suitable learning-rate conditions, but only conditional on AdamW itself converging. In contrast, our results demonstrate that AdamW fails to converge under standard hyperparameters without additional assumptions, and thus are in no conflicts with previous analyses.

7 CONCLUSIONS

We studied the dynamics of AdamW and showed that, despite admitting an implicit objective, its equilibria can be unstable, leading to persistent limit cycles on simple problems. Extending the analysis to higher dimensions, we derived sufficient conditions for stability and validated that AdamW converges only to stable equilibria when it converges at all. These results reveal that AdamW’s theoretical properties are unexpectedly fragile. Its stability hinges on delicate interactions between curvature, weight decay, and momentum, yet the method remains the default optimizer in practice. This contrast suggests that while AdamW succeeds empirically, its dynamics are theoretically unsatisfying, and points to the need for more natural optimizers that preserve its effectiveness while offering stronger guarantees.

Limitations This work provides concrete examples of AdamW’s non-convergence and empirical evidence of limit-cycle behavior, but we do not give a formal proof characterizing the dynamics (e.g., cycles versus chaos). In the multi-dimensional case, eigenanalysis is too complex for interpretable necessary-and-sufficient conditions, and we offer only preliminary sufficient ones. Neural network training adds further complications: high dimensionality, discrete learning-rate schedules, stochasticity, and the AdamW’s preference of fixed points along manifolds of stationary points, and the expense of large-scale experiments. We regard these as important directions for future study.

REFERENCES

- Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in llm pre-training. *arXiv preprint arXiv:2505.13738*, 2025.
- Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves constrained optimization: As lyapunov predicts. *arXiv preprint arXiv:2310.05898*, 2023a.
- Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023b.
- Francesco D’Angelo, Maksym Andriushchenko, Aditya Vardhan Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? *Advances in Neural Information Processing Systems*, 37:23191–23223, 2024.
- Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.
- Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1, 1988.
- Di He, Ajay Jaiswal, Songjun Tu, Li Shen, Ganzhao Yuan, Shiwei Liu, and Lu Yin. Alphadecay: Module-wise weight decay for heavy-tailed balancing in llms. *arXiv preprint arXiv:2506.14562*, 2025.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- Ruinan Jin, Xiao Li, Yaoliang Yu, and Baoxiang Wang. A comprehensive framework for analyzing the convergence of adam: Bridging the gap with sgd. *arXiv preprint arXiv:2410.04458*, 2024.
- Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL <https://kellerjordan.github.io/posts/muon/>.
- Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2002. ISBN 0130673897.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. *arXiv preprint arXiv:2305.17212*, 2023.
- Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.
- Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023.

- Huan Li, Yiming Dong, and Zhouchen Lin. On the $o\left(\frac{\sqrt{d}}{K^{1/4}}\right)$ convergence rate of adamw measured by ℓ_1 norm. *arXiv preprint arXiv:2505.11840*, 2025.
- Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold. In *High-dimensional Learning Dynamics 2025*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- Stanisław Łojasiewicz. Sur les trajectoires du gradient d’une fonction analytique. In *Séminari di Geometria 1982–1983*, pp. 115–117, Bologna, Italy, 1984. Istituto di Geometria, Dipartimento di Matematica, Università di Bologna.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- Chris J Maddison, Daniel Paulin, Yee Whye Teh, Brendan O’Donoghue, and Arnaud Doucet. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.
- PyTorch Contributors. *torch.optim.AdamW — PyTorch Documentation*. PyTorch, 2025. Accessed: September 19, 2025. <https://docs.pytorch.org/docs/stable/generated/torch.optim.AdamW.html>.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.
- Bhavya Vasudeva, Jung Whan Lee, Vatsal Sharan, and Mahdi Soltanolkotabi. The rich and the simple: On the implicit bias of adam and sgd. *arXiv preprint arXiv:2505.24022*, 2025.
- Xi Wang and Laurence Aitchison. How to set adamw’s weight decay as you scale model and dataset size. *arXiv preprint arXiv:2405.13698*, 2024.
- Shuo Xie and Zhiyuan Li. Implicit bias of AdamW: ℓ_∞ -norm constrained optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54488–54510. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/xie24e.html>.
- Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞ geometry of loss landscape via coordinate-wise adaptivity. *arXiv preprint arXiv:2410.08198*, 2024.
- Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. *Advances in neural information processing systems*, 35:28386–28399, 2022.
- Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of adamw. *IEEE transactions on pattern analysis and machine intelligence*, 46(9):6486–6493, 2024.

A DEFERRED PROOFS

A.1 DERIVATION OF THE ADAMW ODE

We now sketch how the AdamW ODE arises as the continuous-time limit of the discrete update, and how the discrete dynamics can be interpreted as a semi-implicit Euler scheme applied to this ODE.

Let $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$. The AdamW updates are

$$\begin{aligned}\mathbf{m}_t &= \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t, \\ \mathbf{v}_t &= \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^{\odot 2}, \\ \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta \left(\frac{\mathbf{m}_t}{\sqrt{\mathbf{v}_t} + e} + \lambda \mathbf{x}_t \right).\end{aligned}$$

With the common approximation $\hat{\mathbf{m}}_t \approx \mathbf{m}_t$, $\hat{\mathbf{v}}_t \approx \mathbf{v}_t$ and the substitution $\mathbf{h}_t = \sqrt{\mathbf{v}_t}$, we set

$$\eta = \Delta t, \quad \beta_1 = 1 - a\Delta t, \quad \beta_2 = 1 - b\Delta t.$$

Continuous-time limit. Consider the first-moment update:

$$\frac{\mathbf{m}_t - \mathbf{m}_{t-1}}{\Delta t} = \frac{-(a\Delta t)\mathbf{m}_{t-1} + (a\Delta t)\mathbf{g}_t}{\Delta t} = -a\mathbf{m}_{t-1} + a\mathbf{g}_t.$$

Letting $\Delta t \rightarrow 0$, the left-hand side becomes $\dot{\mathbf{m}}_t$ and $\mathbf{m}_{t-1} \rightarrow \mathbf{m}_t$, giving $\dot{\mathbf{m}}_t = a(\mathbf{g}_t - \mathbf{m}_t)$. In the same way,

$$\frac{\mathbf{x}_{t+1} - \mathbf{x}_t}{\Delta t} = -\frac{\mathbf{m}_t}{\mathbf{h}_t + e} - \lambda \mathbf{x}_t \quad \rightarrow \quad \dot{\mathbf{x}}_t = -\frac{\mathbf{m}_t}{\mathbf{h}_t + e} - \lambda \mathbf{x}_t,$$

and for $\mathbf{v}_t = \mathbf{h}_t^{\odot 2}$,

$$\frac{\mathbf{v}_t - \mathbf{v}_{t-1}}{\Delta t} = -b\mathbf{v}_{t-1} + b\mathbf{g}_t^{\odot 2} \quad \rightarrow \quad \dot{\mathbf{h}}_t = \frac{b}{2} \left(\frac{\mathbf{g}_t^{\odot 2}}{\mathbf{h}_t} - \mathbf{h}_t \right).$$

Semi-implicit Euler discretization. Starting from the AdamW ODE

$$\dot{\mathbf{x}}_t = -\frac{\mathbf{m}_t}{\mathbf{h}_t + e} - \lambda \mathbf{x}_t, \quad \dot{\mathbf{m}}_t = a(\nabla f(\mathbf{x}_t) - \mathbf{m}_t), \quad \dot{\mathbf{h}}_t = \frac{b}{2} \left(\frac{\nabla f(\mathbf{x}_t)^{\odot 2}}{\mathbf{h}_t} - \mathbf{h}_t \right),$$

a semi-implicit Euler step of size Δt updates $(\mathbf{m}, \mathbf{v})$ explicitly at t and then advances $\mathbf{x}$ using the new $(\mathbf{m}_{t+1}, \mathbf{v}_{t+1})$. Writing $\mathbf{v} = \mathbf{h}^{\odot 2}$ and $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$:

$$\begin{aligned}\mathbf{m}_{t+1} &= \mathbf{m}_t + \Delta t a(\mathbf{g}_t - \mathbf{m}_t) = (1 - a\Delta t)\mathbf{m}_t + (a\Delta t)\mathbf{g}_t, \\ \mathbf{v}_{t+1} &= \mathbf{v}_t + \Delta t b(\mathbf{g}_t^{\odot 2} - \mathbf{v}_t) = (1 - b\Delta t)\mathbf{v}_t + (b\Delta t)\mathbf{g}_t^{\odot 2}, \quad \mathbf{h}_{t+1} = \sqrt{\mathbf{v}_{t+1}}, \\ \mathbf{x}_{t+1} &= \mathbf{x}_t + \Delta t \left(-\frac{\mathbf{m}_{t+1}}{\mathbf{h}_{t+1} + e} - \lambda \mathbf{x}_t \right) = \mathbf{x}_t - \eta \left(\frac{\mathbf{m}_{t+1}}{\sqrt{\mathbf{v}_{t+1}} + e} + \lambda \mathbf{x}_t \right),\end{aligned}$$

with the identifications $\eta = \Delta t$, $\beta_1 = 1 - a\Delta t$, $\beta_2 = 1 - b\Delta t$. This is exactly the AdamW update without bias correction, confirming that AdamW implements a semi-implicit Euler discretization of the AdamW ODE.

A.2 IMPLICIT OBJECTIVE OF ADAMW

Theorem A.1. Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be continuously differentiable, and let $\lambda, e > 0$. Then $\mathbf{x} \in \mathbb{R}^d$ satisfies the fixed-point condition equation 2 if and only if $\mathbf{x}$ is a stationary point of the ℓ_∞ -constrained and regularized problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) = f(\mathbf{x}) + \frac{e}{\lambda} \Phi(\lambda \mathbf{x}), \quad \text{subject to} \quad \|\mathbf{x}\|_\infty \leq \frac{1}{\lambda}, \quad (11)$$

where $\Phi(\mathbf{x}) = \sum_{i=1}^d \phi(x_i)$ with

$$\phi(x) = -|x| - \log(1 - |x|), \quad |x| < 1.$$

Moreover, since $\phi(x) \rightarrow +\infty$ as $|x| \rightarrow 1$, the regularizer Φ serves as a barrier function that naturally enforces the constraint $\|\mathbf{x}\|_\infty < \frac{1}{\lambda}$.

We begin with the following lemma with the regularizer:

Lemma A.2. Let $\phi : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ be

$$\phi(x) = \begin{cases} -|x| - \log(1 - |x|), & |x| < 1, \\ +\infty, & |x| \geq 1, \end{cases} \quad \Phi(\mathbf{x}) = \sum_{i=1}^d \phi(x_i),$$

and for $\lambda > 0$ define the rescaled barrier $\phi_\lambda(x) := \lambda^{-1}\phi(\lambda x)$ and $\Phi_\lambda(\mathbf{x}) := \sum_i \phi_\lambda(x_i)$. Then ϕ_λ is convex and, for $|x| < 1/\lambda$,

$$\phi'_\lambda(x) = \frac{\lambda x}{1 - |\lambda x|}, \quad \phi''_\lambda(x) = \frac{\lambda}{(1 - |\lambda x|)^2}.$$

Moreover, the conjugate of ϕ is $\phi^*(y) = |y| - e \log(|y| + e)$ with $(\phi^*)'(y) = \frac{y}{|y| + e}$ and $(\phi^*)''(y) = \frac{e}{(|y| + e)^2}$.

Proof. The stated derivatives follow by direct differentiation on $(-1, 1)$; convexity holds since $\phi''_\lambda \geq 0$. The conjugate and its derivatives are obtained by a standard one-dimensional Legendre transform; see the expressions above. $\square$

Lemma A.3. For $e, \lambda > 0$ and any $g, x \in \mathbb{R}$,

$$\frac{g}{|g| + e} + \lambda x = 0 \iff g + e \phi'_\lambda(x) = 0,$$

and hence the left-hand side is well-defined only if $|\lambda x| < 1$. In $\mathbb{R}^d$, the equivalence holds element-wise.

Proof. Using Lemma A.2, $\phi'_\lambda(x) = \frac{\lambda x}{1 - |\lambda x|}$. If $\frac{g}{|g| + e} + \lambda x = 0$, then $|\lambda x| < 1$ and $\text{sign}(x) = -\text{sign}(g)$. Simple algebra yields $g + e \frac{\lambda x}{1 - |\lambda x|} = 0$. The converse is identical in reverse. $\square$

Lemma A.4. Let $F(\mathbf{x}) := f(\mathbf{x}) + \frac{e}{\lambda} \Phi(\lambda \mathbf{x}) = f(\mathbf{x}) + e \Phi_\lambda(\mathbf{x})$. On $\{\mathbf{x} : \|\lambda \mathbf{x}\|_\infty < 1\}$,

$$\nabla F(\mathbf{x}) = \nabla f(\mathbf{x}) + e \nabla \Phi_\lambda(\mathbf{x}), \quad \nabla^2 F(\mathbf{x}) = \nabla^2 f(\mathbf{x}) + \text{Diag}\left(\frac{\lambda e}{(1 - |\lambda \mathbf{x}|)^{\odot 2}}\right).$$

Proof. Both identities follow by differentiating coordinatewise and invoking Lemma A.2. $\square$

Proof of Theorem. Fixed points $\iff$ stationary points. For AdamW fixed points (discrete or continuous limit without bias corrections), the stationary condition for $\mathbf{x}$ reads

$$\frac{\nabla f(\mathbf{x}_*)}{|\nabla f(\mathbf{x}_*)| + e} + \lambda \mathbf{x}_* = 0 \quad (\text{elementwise}).$$

By Lemma A.3 this is equivalent to

$$\nabla f(\mathbf{x}_*) + e \nabla \Phi_\lambda(\mathbf{x}_*) = \nabla F(\mathbf{x}_*) = 0, \quad \|\lambda \mathbf{x}_*\|_\infty < 1.$$

Thus fixed points lie in the open cube $\{\|\lambda \mathbf{x}\|_\infty < 1\}$ and are precisely the unconstrained stationary points of F there.

Conversely, any stationary point of equation 3 must satisfy $\|\lambda \mathbf{x}_*\|_\infty < 1$: since $\Phi(\lambda \mathbf{x}) \rightarrow +\infty$ as $|\lambda x_i| \uparrow 1$, no optimality condition can be met on the boundary for a continuous f . Hence the inequality constraints are inactive, the Lagrange multipliers vanish, and the stationary conditions reduce to $\nabla F(\mathbf{x}_*) = 0$, which is equivalent (by Lemma A.3) to the AdamW fixed-point condition. This proves the coincidence of fixed points and stationary points. $\square$

A.3 STABILITY CONDITION OF ADAMW ODE

Writing the AdamW ODE in a vectorized form, we have:

$$\frac{d}{dt}z = f(z), \quad z := \begin{bmatrix} x \\ m \\ h \end{bmatrix}, \quad f(z) := \begin{bmatrix} -m/(h+e) - \lambda x \\ a(\nabla f(x) - m) \\ \frac{b}{2}(\nabla f(x)^{\odot 2}/h - h) \end{bmatrix}. \quad (12)$$

By differentiating the system $J(z) := \partial f(z)/\partial z$, we obtain:

$$J(z) = \begin{bmatrix} -\lambda I & -\text{Diag}(1/(h+e)) & \text{Diag}(m/(h+e)^2) \\ a\nabla^2 f(x) & -aI & 0 \\ b\text{Diag}(\nabla f(x)/h)\nabla^2 f(x) & 0 & -\frac{b}{2}(\text{Diag}(\nabla f(x)^{\odot 2}/h^2) + I) \end{bmatrix}.$$

Noting that at equilibrium point $z_* = [x_*, m_*, h_*]$, we have $\nabla f(x)^{\odot 2} = h_*^{\odot 2}$, this yields:

$$J(z_*) = \begin{bmatrix} -\lambda I & -\text{Diag}(1/(h_*+e)) & \text{Diag}(m_*/(h_*+e)^2) \\ a\nabla^2 f(x_*) & -aI & 0 \\ b\text{Diag}(\nabla f(x_*)/h_*)\nabla^2 f(x_*) & 0 & -bI \end{bmatrix}.$$

This following lemma is useful for simplifying the characteristic polynomial:

Lemma A.5. *If $[x_*, m_*, h_*]$ is an fixed point of AdamW in Eq. equation 1, then*

$$\nabla^2 F(x_*) = \nabla^2 f(x_*) + \text{Diag}\left(\frac{\lambda e}{(1-|\lambda x_*|)^2}\right) = \nabla^2 f(x_*) + \text{Diag}\left(\frac{\lambda}{e}\right)(h_*+e)^2.$$

Proof. At equilibrium, we have $\frac{m_*}{h_*+e} + \lambda x_* = 0$ and $h_* = |\nabla f(x_*)| = |m_*|$. Hence

$$|\lambda x_*| = \frac{|m_*|}{h_*+e} = \frac{h_*}{h_*+e}.$$

Therefore,

$$\frac{\lambda e}{(1-|\lambda x_*|)^2} = \frac{\lambda e}{(e/(h_*+e))^2} = \frac{\lambda}{e}(h_*+e)^2.$$

The claim then follows from lemma A.2. $\square$

Lemma A.6 (Routh–Hurwitz criterion for cubics). *All roots of the cubic equation $s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0 = 0$ have negative real part (i.e., the system is Hurwitz stable) if and only if*

$$\alpha_2, \alpha_1, \alpha_0 > 0 \quad \text{and} \quad \alpha_2 \alpha_1 > \alpha_0.$$

Proof. ($\Rightarrow$) Suppose all roots satisfy $\Re(s_i) < 0$. Writing the polynomial as

$$(s - s_1)(s - s_2)(s - s_3) = s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0,$$

Vieta's relations give

$$\alpha_2 = -(s_1 + s_2 + s_3), \quad \alpha_1 = s_1 s_2 + s_1 s_3 + s_2 s_3, \quad \alpha_0 = -s_1 s_2 s_3.$$

Since $\Re(s_i) < 0$, each coefficient is strictly positive. Moreover,

$$\alpha_2 \alpha_1 - \alpha_0 = (-(s_1 + s_2 + s_3))(s_1 s_2 + s_1 s_3 + s_2 s_3) + s_1 s_2 s_3 \geq 0,$$

where non-negativity follows whether the roots are all real or include a complex-conjugate pair.

($\Leftarrow$) Conversely, assume $\alpha_2, \alpha_1, \alpha_0 > 0$ and $\alpha_2 \alpha_1 > \alpha_0$. - If $s \in \mathbb{R}$ were positive, then

$$s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0 > 0,$$

contradiction. Hence all real roots satisfy $s \leq 0$. Now let $s = u + iv$ with $v \neq 0$. Substituting and separating real and imaginary parts yields

$$3u^2 + 2\alpha_2 u + \alpha_1 = v^2, \quad u^3 + \alpha_2 u^2 + \alpha_1 u + \alpha_0 = (3u + \alpha_2)v^2.$$

Eliminating v^2 gives

$$8u^3 + 8\alpha_2 u^2 + 2(\alpha_2^2 + \alpha_1)u + (\alpha_2 \alpha_1 - \alpha_0) = 0.$$

If $u > 0$, every term on the left-hand side is strictly positive, a contradiction. Thus $u \leq 0$, i.e., all complex roots lie in the closed left half-plane.

Therefore, the given conditions are necessary and sufficient for Hurwitz stability. $\square$

Corollary A.7. Assume $a, b, e > 0$ and $\lambda \geq 0$. Let $[x_*, m_*, h_*]$ be an equilibrium of the AdamW ODE. Then $[x_*, m_*, h_*]$ is Hurwitz stable if and only if $F''(x_*) > 0$ and

$$[a(a + \lambda) - b(b + \lambda)|\lambda x_*|]f''(x_*) > -\frac{eC}{(1 - |\lambda x_*|)}, \quad (13)$$

where $C = (a + b + \lambda)(ab + a\lambda + b\lambda) - ab\lambda > 0$.

Proof. Assume $a, b, e > 0$ and $\lambda \geq 0$. Recall from the Routh–Hurwitz criterion (Lemma A.6) that all roots of a cubic $s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0$ have negative real part if and only if

$$\alpha_2, \alpha_1, \alpha_0 > 0 \quad \text{and} \quad \alpha_2 \alpha_1 > \alpha_0.$$

For the characteristic polynomial of $\mathbf{J}(\mathbf{z}_*)$,

$$\chi_{\mathbf{J}}(s) = s^3 + (a + b + \lambda)s^2 + \left[ab + a\lambda + \frac{(a - b)H_*}{h_*} + \frac{beh_*}{h_*^2}\right]s + \frac{abe\hat{H}_*}{h_*^2},$$

we identify

$$\alpha_2 = a + b + \lambda, \quad \alpha_1 = ab + a\lambda + \frac{(a - b)H_*}{h_*} + \frac{beh_*}{h_*^2}, \quad \alpha_0 = \frac{abe\hat{H}_*}{h_*^2}.$$

Since $a, b, e > 0$ and $\lambda \geq 0$, clearly $\alpha_2 > 0$. Moreover, $\alpha_0 > 0$ if and only if $\hat{H}_* = F''(x_*) > 0$, which shows that any Hurwitz stable equilibrium must be a strict local minimum of F .

It remains to impose $\alpha_2 \alpha_1 > \alpha_0$. Substituting the definitions and rearranging, one obtains

$$[(a + b + \lambda)(a\hat{h}_* - bh_*) - abe]H_* > [(a + b + \lambda)(ab + a\lambda + b\lambda) - ab\lambda] \frac{e}{1 - |\lambda x_*|}.$$

Using $h_*/\hat{h}_* = |\lambda x_*|$ and $\hat{h}_* = e/(1 - |\lambda x_*|)$, this inequality is equivalent to

$$[a(a + \lambda) - b(b + \lambda)|\lambda x_*|]f''(x_*) > -\frac{eC}{1 - |\lambda x_*|},$$

where

$$C := (a + b + \lambda)(ab + a\lambda + b\lambda) - ab\lambda > 0.$$

Thus the equilibrium (x_*, m_*, h_*) is Hurwitz stable if and only if $F''(x_*) > 0$ and the above inequality holds. $\square$

Corollary A.8. Assume $a, b, e > 0$, and $\lambda \geq 0$. Let $[x_*, m_*, h_*]$ is an equilibrium point of AdamW ODE. If $a = b$, or $\lambda = 0$, then x_* is Hurwitz stable iff $F''(x_*) > 0$.

Proof. By Lemma A.6, Hurwitz stability of an equilibrium $\mathbf{z}_* = [x_*, m_*, h_*]$ requires $F''(x_*) > 0$ and

$$(a + b + \lambda) \left[ab + a\lambda + \frac{(a - b)H_*}{h_*} + \frac{beF''(x_*)}{h_*^2} \right] > \frac{abeF''(x_*)}{h_*^2}.$$

This can be rearranged as

$$(a + b + \lambda) \left[ab + a\lambda + \frac{(a - b)H_*}{h_*} \right] + (b + \lambda) \frac{beF''(x_*)}{h_*^2} > 0.$$

If $a = b$, the term involving $(a - b)$ vanishes, and the inequality simplifies to

$$(a + b + \lambda)(ab + a\lambda) + (b + \lambda) \frac{beF''(x_*)}{h_*^2} > 0.$$

Since $a = b > 0$, $\lambda \geq 0$, and $e > 0$, this condition is automatically satisfied whenever $F''(x_*) > 0$.

If $\lambda = 0$, the dynamics reduce to the unconstrained Adam system. In this case $\bar{h}_* = e$, $h_* = 0$, and $F''(x_*) = H_*$. The inequality then becomes

$$(a+b) \left[ab + \frac{(a-b)H_*}{e} \right] + \frac{b^2 H_*}{e} > 0,$$

which simplifies further to

$$(a+b)ab + \frac{a^2}{e} H_* > 0.$$

Because $a, b, e > 0$, this condition is equivalent to requiring $H_* > 0$, i.e. $F''(x_*) > 0$.

In both cases, stability is therefore equivalent to $F''(x_*) > 0$, completing the proof. $\square$

A.4 THE DISCRETE ADAMW SYSTEM

Definition A.9 (Lyapunov stability for discrete systems). *Consider the autonomous system $\mathbf{z}_{t+1} = T(\mathbf{z}_t)$, where $T : \mathcal{D} \rightarrow \mathbb{R}^d$ with domain $\mathcal{D} \subset \mathbb{R}^d$. Let $\mathbf{z}_* \in \mathcal{D}$ be a fixed point, i.e. $T(\mathbf{z}_*) = \mathbf{z}_*$. Then:*

- $\mathbf{z}_*$ is **Lyapunov stable** if for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|\mathbf{z}_0 - \mathbf{z}_*\| < \delta$ implies $\|\mathbf{z}_t - \mathbf{z}_*\| < \epsilon$ for all integers $t \geq 0$.
- $\mathbf{z}_*$ is **asymptotically stable** if it is Lyapunov stable and there exists $\delta' > 0$ such that $\|\mathbf{z}_0 - \mathbf{z}_*\| < \delta'$ implies $\lim_{t \rightarrow \infty} \mathbf{z}_t = \mathbf{z}_*$.
- $\mathbf{z}_*$ is **unstable** if it is not Lyapunov stable; i.e., there exists $\epsilon' > 0$ such that for every $\delta > 0$ one can find $\|\mathbf{z}_0 - \mathbf{z}_*\| < \delta$ but $\|\mathbf{z}_t - \mathbf{z}_*\| \geq \epsilon'$ for some integer $t > 0$.

Lyapunov's stability characterizes robustness of discrete trajectories, while the following Schur stability criterion describes how the linearized neighborhood near a fixed point behaves.

Definition A.10 (Schur stability). *Let $\mathbf{z}_*$ be a fixed point of $\mathbf{z}_{t+1} = T(\mathbf{z}_t)$ and let $\mathbf{J}(\mathbf{z}_*) = [\partial T_i / \partial z_j]_{\mathbf{z}_*}$ denote the Jacobian at $\mathbf{z}_*$. Then $\mathbf{z}_*$ is **Schur stable** if every eigenvalue $\lambda \in \mathbb{C}$ of $\mathbf{J}(\mathbf{z}_*)$ satisfies $|\lambda| < 1$.*

Lyapunov's indirect method (discrete form) states: If $\mathbf{J}(\mathbf{z}_*)$ is Schur stable, then the fixed point $\mathbf{z}_*$ is asymptotically stable. If $\mathbf{J}(\mathbf{z}_*)$ has an eigenvalue with $|\lambda| > 1$, the fixed point is unstable.

We first write the standard discrete time AdamW update in a vectorized form:

$$\mathbf{z}_{t+1} = T(\mathbf{z}_t), \quad \mathbf{z} := \begin{bmatrix} \mathbf{x} \\ \mathbf{m} \\ \mathbf{h} \end{bmatrix}, \quad T(\mathbf{z}) := \begin{bmatrix} (1 - \lambda\eta)\mathbf{x} - \eta \frac{\beta_1 \mathbf{m} + (1 - \beta_1) \nabla f(\mathbf{x})}{\beta'_2 \mathbf{h} + (1 - \beta'_2) \nabla f(\mathbf{x})^{\odot 2} / \mathbf{h} + e} \\ \beta_1 \mathbf{m}_t + (1 - \beta_2) \nabla f(\mathbf{x}) \\ \beta'_2 \mathbf{h} + (1 - \beta'_2) \nabla f(\mathbf{x})^{\odot 2} / \mathbf{h} \end{bmatrix}. \quad (14)$$

where $\beta'_2 = \frac{1 + \beta_2}{2}$. Note that the term corresponding to $\mathbf{x}_{t+1}$ involves more terms than the continuous counterpart. This is because AdamW update corresponds to the semi-implicit Euler discretization of the AdamW ODE and we need to write $\mathbf{x}_{t+1}$ in terms of $[\mathbf{x}_t, \mathbf{m}_t, \mathbf{h}_t]$ only for analysis. Nevertheless, we will use $\mathbf{m}_{t+1}$ and $\mathbf{h}_{t+1}$ to simplify expressions in the following. Differentiating this system $\mathbf{J}(\mathbf{z}) := \partial T(\mathbf{z}) / \partial \mathbf{z}$, we obtain:

$$\mathbf{J}(\mathbf{z}_t) = \begin{bmatrix} \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} & -\eta \beta_1 \text{Diag}\left(\frac{1}{\mathbf{h}_{t+1} + e}\right) & \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{h}_t} \\ (1 - \beta_1) \nabla^2 f(\mathbf{x}_t) & \beta_1 \mathbf{I} & 0 \\ (1 - \beta_2) \text{Diag}\left(\frac{\nabla f(\mathbf{x}_t)}{\mathbf{h}_t}\right) \nabla^2 f(\mathbf{x}_t) & 0 & (1 + \beta_2) \mathbf{I} - \text{Diag}\left(\frac{\mathbf{h}_{t+1}}{\mathbf{h}_t}\right) \end{bmatrix}.$$

where

$$\begin{aligned} \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} &= (1 - \lambda\eta) \mathbf{I} - \eta(1 - \beta_1) \text{Diag}\left(\frac{1}{\mathbf{h}_{t+1} + e}\right) \nabla^2 f(\mathbf{x}_t) \\ &\quad + \eta(1 - \beta_2) \text{Diag}\left(\frac{\mathbf{m}_{t+1} \nabla f(\mathbf{x}_t)}{\mathbf{h}_t(\mathbf{h}_{t+1} + e)^2}\right) \nabla^2 f(\mathbf{x}_t) \\ \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{h}_t} &= \eta \text{Diag}\left(\frac{\mathbf{m}_{t+1}}{(\mathbf{h}_{t+1} + e)^2}\right) \left(\beta'_2 \mathbf{I} - (1 - \beta'_2) \text{Diag}\left(\frac{\nabla f(\mathbf{x}_t)^{\odot 2}}{\mathbf{h}_t^2}\right) \right) \end{aligned}$$

Similarly, at fixed point $\mathbf{z}_* = [\mathbf{x}_*, \mathbf{m}_*, \mathbf{h}_*]$, we have $\nabla f(\mathbf{x})^{\odot 2} = \mathbf{h}_*^{\odot 2}$, this yields:

$$J(\mathbf{z}_*) = \begin{bmatrix} \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} & -\eta\beta_1 \text{Diag}\left(\frac{1}{\mathbf{h}_* + e}\right) & \eta\beta_2 \text{Diag}\left(\frac{\mathbf{m}_*}{(\mathbf{h}_* + e)^2}\right) \\ (1 - \beta_1)\nabla^2 f(\mathbf{x}_*) & \beta_1 \mathbf{I} & 0 \\ (1 - \beta_2)\text{Diag}\left(\frac{\nabla f(\mathbf{x}_*)}{\mathbf{h}_*}\right)\nabla^2 f(\mathbf{x}_*) & 0 & \beta_2 \mathbf{I} \end{bmatrix}.$$

where

$$\begin{aligned} \frac{\partial \mathbf{x}_{t+1}}{\partial \mathbf{x}_t} &= (1 - \lambda\eta)\mathbf{I} - \eta(1 - \beta_1)\text{Diag}\left(\frac{1}{\mathbf{h}_* + e}\right)\nabla^2 f(\mathbf{x}_*) \\ &\quad + \eta(1 - \beta_2)\text{Diag}\left(\frac{\mathbf{h}_*}{(\mathbf{h}_* + e)^2}\right)\nabla^2 f(\mathbf{x}_*) \end{aligned}$$

Consider the one-dimensional case. The characteristic polynomial of the Jacobian at the fixed point is

$$\chi_J(s) = \det(s\mathbf{I} - J(\mathbf{z}_*)) = s^3 + as^2 + bs + c,$$

where

$$\begin{aligned} a &= \frac{\eta H_*}{h_*^2} \left[(1 - \beta_1)h_* + (\beta_2 - 1)h_* \right] - (\beta_1 + \beta_2 + 1) + \frac{\eta\lambda}{e}(h_* - h_*), \\ b &= \frac{\eta H_*}{h_*^2} \left[\beta_1\beta_2(h_* - h_*) + \beta_1h_* - \beta_2h_* \right] + (\beta_1\beta_2 + \beta_1 + \beta_2) - \eta\lambda(\beta_1 + \beta_2), \\ c &= \beta_1\beta_2(\eta\lambda - 1). \end{aligned}$$

To analyze the stability criterion given the polynomial, we can take advantage of the following lemma.

Lemma A.11 (Schur–Jury criterion for cubics). *All roots of the cubic polynomial $s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0 = 0$ lie strictly inside the unit disk (i.e., the system is Schur stable) if and only if the following four inequalities hold:*

$$|\alpha_0| < 1, \quad 1 + \alpha_2 + \alpha_1 + \alpha_0 > 0, \quad 1 - \alpha_2 + \alpha_1 - \alpha_0 > 0, \quad 1 - \alpha_1 + \alpha_2\alpha_0 - \alpha_0^2 > 0.$$

Applying Lemma A.11 reveals the stability condition for the discrete system.

$$|\beta_1\beta_2(\eta\lambda - 1)| < 1, \quad F'''(x_*) > 0,$$

$$\begin{aligned} &\frac{\eta}{e} (1 - |\lambda x_*|) f''(x_*) [(1 - |\lambda x_*|)(\beta_1 - 1)(\beta_2 + 1) + 2|\lambda x_*|(\beta_1 - \beta_2)] \\ &\quad + (\beta_1\beta_2 + \beta_1 + \beta_2 + 1)(2 - \eta\lambda) > 0, \\ &\frac{\eta}{e} (1 - |\lambda x_*|) f''(x_*) [\beta_1\beta_2(\eta\lambda - 1)(|\lambda x_*|(\beta_2 - 1) - (\beta_1 - 1)) + \beta_1|\lambda x_*|(\beta_2 - 1) - \beta_2(\beta_1 - 1)] \\ &\quad - (\beta_1\beta_2 - 1)(\beta_1\eta\lambda - \beta_1 + 1)(\beta_2\eta\lambda - \beta_2 + 1) > 0. \end{aligned}$$

A.5 MULTIDIMENSIONAL CASE

Lemma A.12. *Let $\mathbf{z}_* = (\mathbf{x}_*, \mathbf{m}_*, \mathbf{h}_*)$ be an equilibrium point of the AdamW ODE equation 5. If $a \neq b$, then the linearized dynamics $\dot{\mathbf{z}}_t = \mathbf{J}(\mathbf{z}_*)\mathbf{z}_t$ imply that $\mathbf{x}_t$ satisfies*

$$\ddot{\mathbf{x}}_t + \mathbf{A}_2\dot{\mathbf{x}}_t + \mathbf{A}_1\dot{\mathbf{x}}_t + \mathbf{A}_0\mathbf{x}_t = 0, \quad (15)$$

where

$$\begin{aligned} \mathbf{A}_2 &= (a + b + \lambda)\mathbf{I} \\ \mathbf{A}_1 &= (ab + a\lambda + b\lambda)\mathbf{I} + (a - b)\text{Diag}(1/(\mathbf{h}_* + e))\nabla^2 f(\mathbf{x}_*) + b\text{Diag}(e/(\mathbf{h}_* + e)^2)\nabla^2 f(\mathbf{x}_*), \\ \mathbf{A}_0 &= ab[\text{Diag}(e/(\mathbf{h}_* + e)^2)\nabla^2 f(\mathbf{x}_*) + \lambda\mathbf{I}], \end{aligned}$$

Proof. Linearize the AdamW ODE equation 5 at the equilibrium $\mathbf{z}_* = (\mathbf{x}_*, \mathbf{m}_*, \mathbf{h}_*)$. Let $(\mathbf{x}, \mathbf{m}, \mathbf{h})$ denote infinitesimal perturbations (we drop δ notation for readability). Writing $H := \nabla^2 f(\mathbf{x}_*)$ and

$$U := \text{Diag}\left(\frac{1}{\mathbf{h}_* + e}\right), V := \text{Diag}\left(\frac{\mathbf{m}_*}{(\mathbf{h}_* + e)^2}\right), D := \text{Diag}\left(\frac{\nabla f(\mathbf{x}_*)}{\mathbf{h}_*}\right), S := \text{Diag}\left(\frac{e}{(\mathbf{h}_* + e)^2}\right),$$

the linearized dynamics are

$$\dot{\mathbf{x}} = -\lambda\mathbf{x} - U\mathbf{m} + V\mathbf{h}, \quad \dot{\mathbf{m}} = a(H\mathbf{x} - \mathbf{m}), \quad \dot{\mathbf{h}} = b(DH\mathbf{x} - \mathbf{h}). \quad (16)$$

Differentiate the first equation and substitute equation 16:

$$\ddot{\mathbf{x}} = -\lambda\dot{\mathbf{x}} - U\dot{\mathbf{m}} + V\dot{\mathbf{h}} = -\lambda\dot{\mathbf{x}} - aUH\mathbf{x} + aU\mathbf{m} + bVDH\mathbf{x} - bV\mathbf{h}.$$

Using the identity

$$VD = \text{Diag}\left(\frac{\mathbf{m}_*}{(\mathbf{h}_* + e)^2}\right)\text{Diag}\left(\frac{\nabla f(\mathbf{x}_*)}{\mathbf{h}_*}\right) = \text{Diag}\left(\frac{\mathbf{h}_*}{(\mathbf{h}_* + e)^2}\right) = U - S,$$

where we used $\mathbf{m}_* = \nabla f(\mathbf{x}_*)$ and $\mathbf{h}_* = |\nabla f(\mathbf{x}_*)|$ (elementwise) at equilibrium, we obtain

$$\ddot{\mathbf{x}} + \lambda\dot{\mathbf{x}} + (a - b)UH\mathbf{x} + bSH\mathbf{x} = aU\mathbf{m} - bV\mathbf{h}.$$

Eliminate $U\mathbf{m}$ using the first line of equation 16: $U\mathbf{m} = -\dot{\mathbf{x}} - \lambda\mathbf{x} + V\mathbf{h}$, and thus

$$\ddot{\mathbf{x}} + (\lambda + a)\dot{\mathbf{x}} + (a - b)UH\mathbf{x} + bSH\mathbf{x} + a\lambda\mathbf{x} = (a - b)V\mathbf{h}. \quad (17)$$

When $a \neq b$, equation 17 gives

$$V\mathbf{h} = \frac{1}{a - b} \left[\ddot{\mathbf{x}} + (\lambda + a)\dot{\mathbf{x}} + (a - b)UH\mathbf{x} + bSH\mathbf{x} + a\lambda\mathbf{x} \right]. \quad (18)$$

Differentiate equation 17, substitute $\dot{\mathbf{h}} = b(DH\mathbf{x} - \mathbf{h})$ from equation 16, and use equation 18 to eliminate $V\mathbf{h}$:

$$\ddot{\mathbf{x}} + (\lambda + a)\ddot{\mathbf{x}} + [(a - b)UH + bSH + a\lambda\mathbf{I}]\dot{\mathbf{x}} - b(a - b)VDH\mathbf{x} + b(a - b)V\mathbf{h} = 0.$$

Using $VD = U - S$, this becomes

$$\ddot{\mathbf{x}} + (\lambda + a)\ddot{\mathbf{x}} + [(a - b)UH + bSH + a\lambda\mathbf{I}]\dot{\mathbf{x}} - b(a - b)(U - S)H\mathbf{x} + b(a - b)V\mathbf{h} = 0.$$

Substituting equation 18 to eliminate $V\mathbf{h}$ gives

$$\ddot{\mathbf{x}} + (a + b + \lambda)\ddot{\mathbf{x}} + \left[(ab + a\lambda + b\lambda)\mathbf{I} + (a - b)UH + bSH \right] \dot{\mathbf{x}} + ab \left[SH + \lambda\mathbf{I} \right] \mathbf{x} = 0.$$

Recognizing $U = \text{Diag}(1/(\mathbf{h}_* + e))$ and $S = \text{Diag}(e/(\mathbf{h}_* + e)^2)$ yields the claimed coefficients

$$\mathbf{A}_2 = (a + b + \lambda)\mathbf{I},$$

$$\mathbf{A}_1 = (ab + a\lambda + b\lambda)\mathbf{I} + (a - b)\text{Diag}(1/(\mathbf{h}_* + e))\nabla^2 f(\mathbf{x}_*) + b\text{Diag}(e/(\mathbf{h}_* + e)^2)\nabla^2 f(\mathbf{x}_*),$$

$$\mathbf{A}_0 = ab \left[\text{Diag}(e/(\mathbf{h}_* + e)^2)\nabla^2 f(\mathbf{x}_*) + \lambda\mathbf{I} \right],$$

which completes the proof. $\square$

Lemma A.13 (First-order reduction). *Let $\mathbf{x}_t$ solve the third-order ODE*

$$\ddot{\mathbf{x}}_t + \mathbf{A}_2\dot{\mathbf{x}}_t + \mathbf{A}_1\mathbf{x}_t + \mathbf{A}_0\mathbf{x}_t = 0.$$

Define $\mathbf{y}_t := \dot{\mathbf{x}}_t$ and $\mathbf{z}_t := \dot{\mathbf{y}}_t$. Then $(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t)$ satisfies

$$\dot{\mathbf{x}} = \mathbf{y}, \quad \dot{\mathbf{y}} = \mathbf{z}, \quad \dot{\mathbf{z}} = -\mathbf{A}_0\mathbf{x} - \mathbf{A}_1\mathbf{y} - \mathbf{A}_2\mathbf{z}.$$

Proof. Immediate by substitution: $\dot{\mathbf{x}} = \mathbf{y}$, $\dot{\mathbf{y}} = \mathbf{z}$, and using $\ddot{\mathbf{x}} = \dot{\mathbf{z}}$ in the original ODE. $\square$

Lemma A.14 (Derivative identities). *Let $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ evolve according to Lemma A.13. For any symmetric matrices $\mathbf{P}, \mathbf{Q}, \mathbf{K}$ define*

$$W_1(\mathbf{x}, \mathbf{y}) := \|\mathbf{A}_2\mathbf{x} + \mathbf{y}\|_{\mathbf{P}}^2, \quad W_2(\mathbf{y}, \mathbf{z}) := \|\mathbf{A}_2\mathbf{y} + \mathbf{z}\|_{\mathbf{Q}}^2, \quad W_3(\mathbf{y}) := \|\mathbf{y}\|_{\mathbf{K}}^2,$$

with $\|u\|_M^2 := u^\top M u$. Then

$$\begin{aligned} \frac{d}{dt} W_1 &= 2(\mathbf{A}_2\mathbf{x} + \mathbf{y})^\top \mathbf{P}(\mathbf{A}_2\mathbf{y} + \mathbf{z}), \\ \frac{d}{dt} W_2 &= -2(\mathbf{A}_2\mathbf{y} + \mathbf{z})^\top (\mathbf{Q}\mathbf{A}_0\mathbf{x} + \mathbf{Q}\mathbf{A}_1\mathbf{y}), \\ \frac{d}{dt} W_3 &= 2\mathbf{y}^\top \mathbf{K}\mathbf{z}. \end{aligned}$$

Proof. For W_1 , use $\dot{\mathbf{x}} = \mathbf{y}$ and $\dot{\mathbf{y}} = \mathbf{z}$:

$$\frac{d}{dt}(\mathbf{A}_2\mathbf{x} + \mathbf{y}) = \mathbf{A}_2\mathbf{y} + \mathbf{z},$$

hence $\frac{d}{dt} W_1 = 2(\mathbf{A}_2\mathbf{x} + \mathbf{y})^\top \mathbf{P}(\mathbf{A}_2\mathbf{y} + \mathbf{z})$.

For W_2 , note

$$\frac{d}{dt}(\mathbf{A}_2\mathbf{y} + \mathbf{z}) = \mathbf{A}_2\mathbf{z} + \dot{\mathbf{z}} = -\mathbf{A}_0\mathbf{x} - \mathbf{A}_1\mathbf{y},$$

so $\frac{d}{dt} W_2 = 2(\mathbf{A}_2\mathbf{y} + \mathbf{z})^\top \mathbf{Q}(-\mathbf{A}_0\mathbf{x} - \mathbf{A}_1\mathbf{y})$.

For W_3 , $\dot{\mathbf{y}} = \mathbf{z}$, hence $\frac{d}{dt} W_3 = 2\mathbf{y}^\top \mathbf{K}\mathbf{z}$. □

Lemma A.15 (Lyapunov cancellation). *Let $\mathbf{Q}$ be symmetric positive definite and define*

$$\mathbf{P} := \mathbf{Q}\mathbf{A}_0\mathbf{A}_2^{-1}, \quad \mathbf{K} := \mathbf{Q}(\mathbf{A}_1 - \mathbf{A}_0\mathbf{A}_2^{-1}), \quad \mathbf{L} := \mathbf{A}_2^\top \mathbf{Q}(\mathbf{A}_1 - \mathbf{A}_0\mathbf{A}_2^{-1}).$$

Assume $\mathbf{Q}\mathbf{A}_0$ and $\mathbf{Q}\mathbf{A}_1$ are symmetric. Consider

$$V(\mathbf{x}, \mathbf{y}, \mathbf{z}) := \frac{1}{2} \left(\|\mathbf{A}_2\mathbf{x} + \mathbf{y}\|_{\mathbf{P}}^2 + \|\mathbf{A}_2\mathbf{y} + \mathbf{z}\|_{\mathbf{Q}}^2 + \|\mathbf{y}\|_{\mathbf{K}}^2 \right).$$

Then along trajectories of Lemma A.13,

$$\frac{d}{dt} V(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) = -\frac{1}{2} \mathbf{y}^\top (\mathbf{L} + \mathbf{L}^\top) \mathbf{y}.$$

Proof. Combine Lemma A.14:

$$\frac{d}{dt} V = (\mathbf{A}_2\mathbf{x} + \mathbf{y})^\top \mathbf{P}(\mathbf{A}_2\mathbf{y} + \mathbf{z}) - (\mathbf{A}_2\mathbf{y} + \mathbf{z})^\top (\mathbf{Q}\mathbf{A}_0\mathbf{x} + \mathbf{Q}\mathbf{A}_1\mathbf{y}) + \mathbf{y}^\top \mathbf{K}\mathbf{z}.$$

With $\mathbf{P} = \mathbf{Q}\mathbf{A}_0\mathbf{A}_2^{-1}$ we have $\mathbf{Q}\mathbf{A}_0 - \mathbf{P}\mathbf{A}_2 = 0$ and $\mathbf{Q}\mathbf{A}_1 - \mathbf{P} = \mathbf{K}$. Substituting yields

$$\begin{aligned} \frac{d}{dt} V &= -(\mathbf{A}_2\mathbf{y} + \mathbf{z})^\top \mathbf{K}\mathbf{y} + \mathbf{y}^\top \mathbf{K}\mathbf{z} \\ &= -\mathbf{y}^\top \mathbf{A}_2^\top \mathbf{K}\mathbf{y} \\ &= -\mathbf{y}^\top \mathbf{A}_2^\top \mathbf{Q}(\mathbf{A}_1 - \mathbf{A}_0\mathbf{A}_2^{-1})\mathbf{y} \\ &= -\frac{1}{2} \mathbf{y}^\top (\mathbf{L} + \mathbf{L}^\top) \mathbf{y}. \end{aligned}$$

□

Lemma A.16 (Lyapunov stability for the third-order system). *Consider $\ddot{\mathbf{x}}_t + \mathbf{A}_2\dot{\mathbf{x}}_t + \mathbf{A}_1\dot{\mathbf{x}}_t + \mathbf{A}_0\mathbf{x}_t = 0$. If there exists a symmetric positive definite $\mathbf{Q}$ such that*

1. $\mathbf{Q}\mathbf{A}_0$ and $\mathbf{Q}\mathbf{A}_1$ are symmetric;
2. $\mathbf{Q}\mathbf{A}_0$, $\mathbf{Q}\mathbf{A}_1$, and $\mathbf{Q}(\mathbf{A}_1\mathbf{A}_2 - \mathbf{A}_0)$ are positive definite,

then the dynamics are asymptotically stable.

Proof. By Lemma A.13, stability of the third-order system is equivalent to stability of its first-order form. Define $\mathbf{P}, \mathbf{K}, \mathbf{L}$ as in Lemma A.15. Assumptions (1)–(2) ensure $\mathbf{P}, \mathbf{K} \succ 0$ and $\mathbf{L} + \mathbf{L}^\top \succ 0$. Thus V in Lemma A.15 is positive definite and radially unbounded.

Along trajectories,

$$\frac{d}{dt}V(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) = -\frac{1}{2}\mathbf{y}_t^\top(\mathbf{L} + \mathbf{L}^\top)\mathbf{y}_t \leq 0,$$

so V is nonincreasing. Hence trajectories are bounded and Lyapunov stable.

By LaSalle’s invariance principle, the Ω -limit set lies inside $\mathcal{E} = \{(\mathbf{x}, \mathbf{y}, \mathbf{z}) : \mathbf{y} = 0\}$. On $\mathcal{E}$, $\dot{\mathbf{x}} = 0$, so $\mathbf{x}$ is constant; also $\dot{\mathbf{y}} = \mathbf{z} = 0$. Finally $\dot{\mathbf{z}} = -\mathbf{A}_0\mathbf{x} = 0$ implies $\mathbf{x} = 0$ because $\mathbf{Q}\mathbf{A}_0 \succ 0$. Thus the only invariant set is $(0, 0, 0)$, so the system is asymptotically stable.

In this derivation we used the fact that $\mathbf{A}_2$ is symmetric positive definite and commutes with every matrix. This holds automatically because

$$\mathbf{A}_2 = (a + b + \lambda)\mathbf{I},$$

a scalar multiple of the identity. □

B THE USE OF LARGE LANGUAGE MODELS (LLMs)

We used large language models (LLMs) as a general-purpose assistive tool in preparing this paper. Specifically, LLMs were employed to polish the writing for clarity and readability, and to provide stylistic suggestions. In addition, we interacted with LLMs during the process of generating plots and conducting numerical checks to validate some of the theoretical results. No part of the research questions, theoretical analysis, or main contributions was produced by LLMs; their role was limited to writing assistance and auxiliary support in experimentation and visualization.