A Stability Analysis of AdamW: Unstable Equilibria and Non-Convergence

Baiyu Su* **Lizhang Chen** **Qiang Liu**
The University of Texas at Austin

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 fixed-point 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 one-dimensional settings, equilibria may be unstable, and trajectories can exhibit persistent oscillations rather than convergence to them. We further extend the analysis to higher dimensions, deriving sufficient conditions for local stability of the continuous-time dynamics, and propose a convergent modification that preserves AdamW's fixed points. These results clarify what optimization problem AdamW is associated with, when its convergence can be expected, and how its dynamics could inspire more reliable optimizers.

1 Introduction

The Adam optimizer with decoupled weight decay (AdamW; [42]) is the default optimizer across modern deep learning systems, including large language models, vision architectures, and generative models [21, 40, 50, 20]. Yet beneath this ubiquity lies a fundamental gap: while the convergence of Adam (without weight decay; [29]) is already subtle to analyze, *no general convergence proof is available for AdamW*[2]. The difficulty arises from the use of decoupled weight decay, which is essential in practice for stability and generalization [30, 15], but remains disconnected from the objective, misaligned with the gradient, and fundamentally alters the algorithm's dynamics.

Our contributions. First, we show that AdamW can fail to converge even on simple static objectives: its fixed points are exactly the stationary points of an implicit objective, but not all are stable. We give exact stability conditions in one dimension and sufficient ones in higher dimensions, with counterexamples for both $\beta_1 > \beta_2$ and $\beta_1 < \beta_2$ (sections 3–5). Second, we propose AdamMe, a rescaled weight decay that preserves AdamW's fixed points and admits a Lyapunov function guaranteeing convergence of bounded continuous-time trajectories (section 4.3).

Implicit objective and stability of AdamW. From the fixed-point conditions of AdamW on an objective f with weight decay λ, we show that its equilibria coincide exactly with the stationary points of a static ℓ_∞-constrained and regularized problem, which we call the *implicit objective* F (Theorem 3.1). This equivalence is only first-order: if AdamW converges, its limit must be a local minimum of F, but a local minimum of F need not be a stable equilibrium of AdamW. Stability also requires a curvature condition on f: in one dimension, an upper or lower bound on f'' depending on the relation between β_1 and β_2 (Proposition 4.3). When $\beta_1 > \beta_2$, even simple convex objectives

*Corresponding author: baiyusu@utexas.edu.
[2]See section 6 for discussion.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

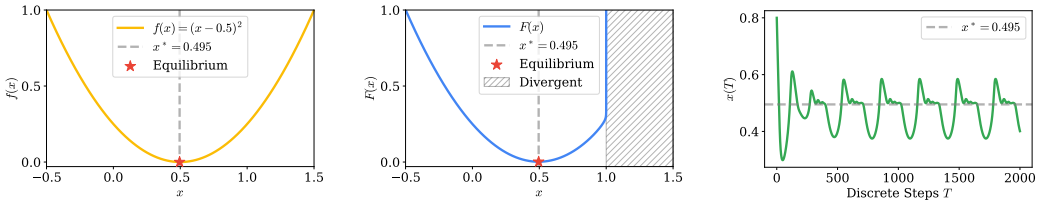


Figure 1: AdamW with $\eta = 10^{-2}$, $\beta_1 = 0.99$, $\beta_2 = 0.9$, $\lambda = 1$, $e = 10^{-2}$. Left: quadratic objective $f(x) = (x - \frac{1}{2})^2$. Middle: implicit objective $F(x)$ whose stationary point coincides with AdamW's equilibrium. Right: AdamW iterates $x(T)$ over time, showing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

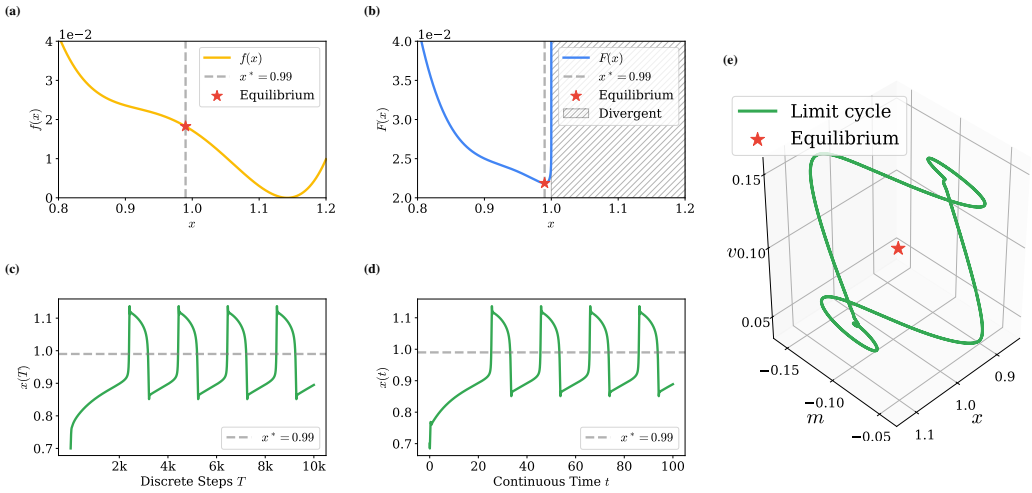


Figure 2: AdamW with $\eta = 10^{-2}$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\lambda = 1$, $e = 10^{-3}$. (a) Target objective $f(x)$ with fixed point at $x^* = 0.99$. (b) Implicit objective $F(x)$ whose stationary point coincides with the AdamW fixed point. (c) Discrete-time trajectory $x(T)$. (d) Continuous-time counterpart $x(t)$. (e) Phase portrait in (x, m, v) space. Together, (c–e) show numerical non-convergence and persistent oscillation around the unstable equilibrium.

can have unstable fixed points in both continuous and discrete time, independent of the learning rate (figure 1). In the common regime $\beta_1 < \beta_2$, the obstruction is subtler but still present: non-convex one-dimensional objectives can yield bounded trajectories with persistent oscillations rather than convergence (figure 2). In higher dimensions, a local Lyapunov analysis gives tractable sufficient conditions for stability (section 5). These behaviors arise from AdamW's update rule, not from adversarial online objectives [cf. 51].

AdamMe: a convergent modification. We trace this instability to a mismatch in AdamW's update: the first-moment estimate $\boldsymbol{m}$ is normalized by $\sqrt{\boldsymbol{v}} + e$, with second-moment estimate $\boldsymbol{v}$ and a small constant $e > 0$, whereas the weight decay $\lambda \boldsymbol{x}$ on the parameters $\boldsymbol{x}$ is not. AdamMe rescales the weight decay by $(|\boldsymbol{m}| + e)/(\sqrt{\boldsymbol{v}} + e)$, a factor that equals one at every equilibrium, so AdamMe has exactly the same fixed points and implicit objective F as AdamW (Proposition 4.7). Unlike AdamW, its continuous-time dynamics admit a Lyapunov function: whenever $1 - \beta_2 \leq 4(1 - \beta_1)$ (e.g., $\beta_1 = 0.9$, $\beta_2 = 0.999$), every bounded trajectory approaches the equilibrium set (Theorem 4.8). The implicit objective is therefore not the source of non-convergence; the dynamics that approach it are. On GPT-2-Small, AdamMe reaches a lower validation loss than AdamW (figure 5).

2 Background

We analyze AdamW through the ODE obtained in its small-step limit, a standard tool for adaptive optimizers including Adam, Lion, and AdamW [3, 4, 44, 8, 14]. The limit is well suited to stability analysis: the fixed points of AdamW coincide with the equilibria of its ODE (section 4), and for small step sizes, their stability carries over to the discrete algorithm [6, 54]. The limit thus isolates

Algorithm 1 AdamW and AdamMe. The AdamW update follows the PyTorch implementation [49].

Require: differentiable objective $f : \mathbb{R}^d \to \mathbb{R}$
Require: initial parameters $\boldsymbol{x}_0$, hyperparameters $\{\eta_t\}_{t=1}^T, \beta_1, \beta_2, e$, weight decay strength $\lambda > 0$
Initialize moments $\boldsymbol{m}_0 \leftarrow \mathbf{0}, \quad \boldsymbol{v}_0 \leftarrow \mathbf{0}$
for $t = 1, 2, \ldots, T$ **do**
  Compute gradient $\boldsymbol{g}_t \leftarrow \nabla f(\boldsymbol{x}_{t-1})$
  Update moment: $\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t, \quad \boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^{\odot 2}$
  Bias correction: $\hat{\boldsymbol{m}}_t \leftarrow \boldsymbol{m}_t/(1-\beta_1^t), \quad \hat{\boldsymbol{v}}_t \leftarrow \boldsymbol{v}_t/(1-\beta_2^t)$
  if AdamW **then**
    $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \eta_t\Big(\dfrac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t} + e} + \lambda \boldsymbol{x}_{t-1}\Big)$
  else if AdamMe **then**
    $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \eta_t\Big(\dfrac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t} + e} + \lambda \dfrac{|\hat{\boldsymbol{m}}_t| + e}{\sqrt{\hat{\boldsymbol{v}}_t} + e} \odot \boldsymbol{x}_{t-1}\Big)$
  end if
end for

the update rule itself from step-size effects such as the edge of stability [12, 13], which it does not capture; Appendix A.6 treats discrete-time AdamW directly.

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.

Definition 2.1 (Lyapunov stability)**.** *Consider the autonomous system $\dot{\boldsymbol{z}}(t) = \boldsymbol{T}(\boldsymbol{z}(t))$, where $\boldsymbol{T} : \mathcal{D} \to \mathbb{R}^{3d}$ is locally Lipschitz with domain $\mathcal{D} \subset \mathbb{R}^{3d}$. An equilibrium $\boldsymbol{z}_* \in \mathcal{D}$ is **Lyapunov stable** if for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|\boldsymbol{z}(0) - \boldsymbol{z}_*\| < \delta$ implies $\|\boldsymbol{z}(t) - \boldsymbol{z}_*\| < \epsilon$ for all $t \geq 0$. It is **asymptotically stable** if, additionally, $\boldsymbol{z}(t) \to \boldsymbol{z}_*$ for all sufficiently close initializations, and **unstable** if it is not Lyapunov stable.*

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

Lyapunov's indirect method gives the following characterization. If $\boldsymbol{J}(\boldsymbol{z}_*)$ is Hurwitz stable, then the equilibrium $\boldsymbol{z}_*$ is *asymptotically stable*. If $\boldsymbol{J}(\boldsymbol{z}_*)$ 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 [28] for further details.

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

Convergence. Throughout the paper, convergence refers to approach to the optimizer's own stationary state. For a continuous-time system $\dot{\boldsymbol{z}} = \boldsymbol{T}(\boldsymbol{z})$, this means $\boldsymbol{z}(t) \to \boldsymbol{z}_*$ for some equilibrium satisfying $\boldsymbol{T}(\boldsymbol{z}_*) = \mathbf{0}$. For a fixed-hyperparameter discrete algorithm $\boldsymbol{z}_{t+1} = \boldsymbol{T}(\boldsymbol{z}_t)$, this means $\boldsymbol{z}_t \to \boldsymbol{z}_*$ for some fixed point satisfying $\boldsymbol{T}(\boldsymbol{z}_*) = \boldsymbol{z}_*$. When discussing diminishing learning-rate schedules, we do not count convergence caused only by finite total movement; the standard non-vacuous setting requires divergent partial sums, $\sum_t \eta_t = \infty$ (equivalently, $t \to \infty$ in continuous time, where $\sum_t \eta_t$ plays the role of elapsed time), so that convergence reflects vanishing fixed-point residual rather than the iterates merely freezing.

Notation We denote the parameter vector by $\boldsymbol{x}$, the first-order momentum by $\boldsymbol{m}$, and the second-order momentum by $\boldsymbol{v}$. The objective function is $f : \mathbb{R}^d \to \mathbb{R}$, defined over the parameter domain.

We use $F(\cdot)$ to denote the *implicit objective*, which will be introduced later. Unless stated otherwise, 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 Fixed-Point Objective of AdamW

We start from the *fixed-point conditions* of AdamW, which characterize the relations that convergence points must satisfy if convergence occurs. Decoupled weight decay changes these conditions: AdamW's fixed points no longer align directly with stationary points of the objective $f(\boldsymbol{x})$, but also depend on the weight decay strength.

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

$$\left.\begin{aligned} \frac{\boldsymbol{m}}{\sqrt{\boldsymbol{v}} + e} + \lambda \boldsymbol{x} = \boldsymbol{0}, \quad \boldsymbol{m} = \nabla f(\boldsymbol{x}), \quad \boldsymbol{v} = \nabla f(\boldsymbol{x})^{\odot 2} \end{aligned}\right\} \Longrightarrow \frac{\nabla f(\boldsymbol{x})}{|\nabla f(\boldsymbol{x})| + e} + \lambda \boldsymbol{x} = \boldsymbol{0}. \tag{1}$$

Equation equation 1 characterizes exactly the fixed points of AdamW. These fixed points 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 $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, and let $\lambda, e > 0$. Then $\boldsymbol{x} \in \mathbb{R}^d$ satisfies the fixed-point condition equation 1 if and only if $\boldsymbol{x}$ is a stationary point of the ℓ_∞-constrained and regularized problem*

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \; F(\boldsymbol{x}) = f(\boldsymbol{x}) + \tfrac{e}{\lambda} \Phi(\lambda \boldsymbol{x}), \quad \textit{subject to} \quad \|\boldsymbol{x}\|_\infty \le \tfrac{1}{\lambda}, \tag{2}$$

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

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

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

This implicit fixed-point objective coincides with the Softsign case $\nabla \mathcal{K} = x/(|x| + e)$ of Lion-$\mathcal{K}$ (Appendix A, Chen et al. [8]), and it reduces to the pure ℓ_∞ constraint of Xie and Li [60] as $e \to 0$. Although e is typically tiny in practice, $e > 0$ is essential for theoretical analysis: without it, convergence to a zero-gradient point of f is not well-defined in AdamW, since the update becomes $0/0$ at any zero-gradient point of f, where $m = v = 0$. The finite-e formulation is what lets us analyze the stability of such candidate limits. This view also contrasts with prior interpretations of AdamW weight decay [42, 65] as an adaptive ℓ_2 penalty that depends on optimizer states and therefore departs from a standard static objective. By contrast, our construction yields a static fixed-point objective $F(\boldsymbol{x})$ that depends only on the parameters.

3.1 First-order coincidence is not stability

The fact that AdamW's fixed points coincide with the stationary points of equation 2 **does not imply** that AdamW actually **minimizes** $F(\boldsymbol{x})$. Beyond this first-order consistency, we require:

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

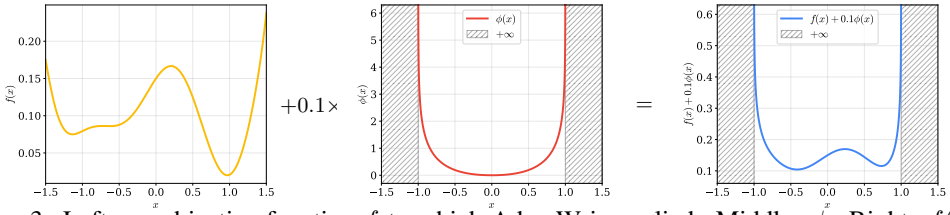


Figure 3: Left: an objective function f to which AdamW is applied. Middle: ϕ. Right: $f(x) + \frac{e}{\lambda}\phi(\lambda x)$ with $e = 0.1$ and $\lambda = 1$.

For the gradient flow $\dot{\boldsymbol{x}} = -\nabla F(\boldsymbol{x})$ this alignment is automatic: F is its own Lyapunov function, so strict local minima are asymptotically stable and nondegenerate saddles are unstable [24, §9.3], and Lee et al. [32] give the discrete-time analogue for gradient descent. AdamW is not a gradient flow of F, so stability is governed by the Jacobian of its own vector field rather than by $\nabla^2 F$, and the alignment must be verified separately. Note that the alignment is **necessary** for AdamW to minimize F but not **sufficient**: it ensures that any equilibrium reached is a local minimum of F, but it does not ensure that trajectories converge at all (section 4); a global guarantee further requires a Lyapunov function and bounded trajectories, as we establish for AdamMe in Theorem 4.8. For many standard optimizers this alignment is known to hold, including the Lion-$\mathcal{K}$ family [43, 8], whose members include SGD with momentum and Lion [11]; for Adam without weight decay under a mild condition on β_1, β_2 [3]; for RMSProp [62] and AdaGrad [17]; and, more recently, for Muon [27, 10]. AdamW, however, is different: the alignment itself fails. 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.

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

Consequently, some local minima of F are unstable under AdamW. Thus the fixed-point equivalence in Theorem 3.1 is only a first-order characterization, not a descent or convergence guarantee for F. It also enables explicit examples where the equilibrium is unstable and numerical trajectories fail to converge to it.

4 Stability Conditions and Non-Convergence of AdamW

Throughout this section we consider exact (batch) gradients and the continuous-time limit of infinitesimal step size; unless stated otherwise, stability refers to equilibria of the resulting ODE. This setting is favorable to convergence: it removes gradient noise and learning-rate schedules, and continuous-time dynamics converge under weaker conditions than their discretizations, since gradient flow reaches local minima without smoothness assumptions [41], while gradient descent also requires smoothness and step-size control [32]. The instabilities we find in this setting therefore reflect the update rule itself.

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

$$\dot{\boldsymbol{x}}_t = -\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t} + e} - \lambda \boldsymbol{x}_t, \quad \dot{\boldsymbol{m}}_t = a(\nabla f(\boldsymbol{x}_t) - \boldsymbol{m}_t), \quad \dot{\boldsymbol{v}}_t = b(\nabla f(\boldsymbol{x}_t)^{\odot 2} - \boldsymbol{v}_t). \tag{3}$$

The discrete-time update of AdamW corresponds to a semi-implicit Euler discretization of this system under the identifications $\eta = \Delta t,\ \beta_1 = 1 - a\Delta t,\ \beta_2 = 1 - b\Delta t$. We ignore the bias correction, i.e., we take $\hat{\boldsymbol{m}}_t = \boldsymbol{m}_t,\ \hat{\boldsymbol{v}}_t = \boldsymbol{v}_t$, which is exact in the large-t limit.

For the active-coordinate stability calculation below, we set $\boldsymbol{h} = \sqrt{\boldsymbol{v}}$ and rewrite the system in the equivalent vector form (suppressing t for clarity):

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{z}), \quad \boldsymbol{z} \coloneqq \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{m} \\ \boldsymbol{h} \end{bmatrix}, \quad \boldsymbol{T}(\boldsymbol{z}) \coloneqq \begin{bmatrix} -\dfrac{\boldsymbol{m}}{\boldsymbol{h} + e} - \lambda \boldsymbol{x} \\ a\big(\nabla f(\boldsymbol{x}) - \boldsymbol{m}\big) \\ \dfrac{b}{2}\Big(\dfrac{\nabla f(\boldsymbol{x})^{\odot 2}}{\boldsymbol{h}} - \boldsymbol{h}\Big) \end{bmatrix}. \tag{4}$$

Let $\boldsymbol{z}_* = [\boldsymbol{x}_*, \boldsymbol{m}_*, \boldsymbol{h}_*]$ denote an equilibrium of the dynamics, i.e. a point satisfying $\boldsymbol{T}(\boldsymbol{z}_*) = \boldsymbol{0}$. Differentiating $\boldsymbol{T}$ and evaluating at $\boldsymbol{z}_*$ gives the Jacobian

$$\boldsymbol{J}(\boldsymbol{z}_*) = \begin{bmatrix} -\lambda \boldsymbol{I} & -\mathrm{Diag}\Big(\dfrac{1}{\boldsymbol{h}_* + e}\Big) & \mathrm{Diag}\Big(\dfrac{\boldsymbol{m}_*}{(\boldsymbol{h}_* + e)^2}\Big) \\ a\nabla^2 f(\boldsymbol{x}_*) & -a\boldsymbol{I} & 0 \\ b\mathrm{Diag}\Big(\dfrac{\nabla f(\boldsymbol{x}_*)}{\boldsymbol{h}_*}\Big)\nabla^2 f(\boldsymbol{x}_*) & 0 & -b\boldsymbol{I} \end{bmatrix}. \tag{5}$$

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 $\boldsymbol{x}$; 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 $\boldsymbol{J}(\boldsymbol{z}_*)$. We study this spectrum through the characteristic polynomial $\chi_{\boldsymbol{J}}(s) \coloneqq \det(s\boldsymbol{I} - \boldsymbol{J}(\boldsymbol{z}_*))$, whose roots are the eigenvalues of $\boldsymbol{J}(\boldsymbol{z}_*)$. In the 1D case, this polynomial reduces to a cubic:

$$\chi_{\boldsymbol{J}}(s) = s^3 + (a+b+\lambda)s^2 + \left[ab + a\lambda + \frac{(a-b)H_*}{\hbar_*} + \frac{be\hat{H}_*}{\hbar_*^2}\right] s + \frac{abe\hat{H}_*}{\hbar_*^2}. \tag{6}$$

with the shorthand

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

where F is the implicit objective introduced in equation 2. Since $\boldsymbol{z}_*$ is an equilibrium, we also use the relations

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

To analyze the signs of the eigenvalues, we recall the 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*

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

Lemma 4.1 involves two sets of conditions, which we treat in turn. The first set, $\alpha_2, \alpha_1, \alpha_0 > 0$, shows that every stable equilibrium of AdamW must be a local minimum of the implicit objective F (Corollary 4.2). The second set, $\alpha_2\alpha_1 > \alpha_0$, shows that the converse fails: local minima of F are stable only under an additional Hessian bound (Proposition 4.3).

Stable equilibria of AdamW $\subseteq$ Local minima of F.

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

Proof. If $\boldsymbol{z}_*$ is Hurwitz stable, then the first set of conditions in Lemma 4.1 holds. Since $\alpha_2 = a + b + \lambda > 0$ always holds and $\alpha_1 > 0$ follows from $\alpha_0 > 0$ and $\alpha_2\alpha_1 > \alpha_0$, the only informative condition is $\alpha_0 = abe\hat{H}_*/\hbar_*^2 > 0$, which implies $\hat{H}_* = F''(x_*) > 0$. □

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 $f''(x_*)$ of the original objective 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 4. Then $[x_*, m_*, h_*]$ is Hurwitz stable if and only if $F''(x_*) > 0$ and*

$$(a+b+\lambda)\left[ab + a\lambda + \frac{(a-b)f''(x_*)}{\hbar_*}\right] + (b+\lambda)be\frac{F''(x_*)}{\hbar_*^2} > 0, \tag{7}$$

which can be rearranged into

$$\left(a(a+\lambda) - b(b+\lambda)|\lambda x_*|\right) f''(x_*) > -\frac{eC}{(1-|\lambda x_*|)}, \tag{8}$$

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

Stability therefore requires two conditions: the implicit Hessian must satisfy $F''(x_*) > 0$, and the inequality equation 8 must hold, which imposes an additional constraint on the Hessian $f''(x_*)$ of the original loss f. We illustrate this constraint in figure 4 for the regime $a < b$: the sign of the coefficient multiplying $f''(x_*)$ changes with $|x_*|$, so the condition can switch between a lower and an upper bound. 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 imposed by equation 8.

Remark 4.4 (When the additional condition vanishes)**.** *The extra requirement in equation 8 disappears in two important cases:*

1. ***Adam limit*** *($\lambda \to 0$). The regularizer and constraint vanish, so the implicit objective reduces to $F(x) = f(x)$. The extra requirement becomes $a^2 f''(x_*) > -eC$, which is automatically implied by $F''(x_*) = f''(x_*) > 0$. Stability then reduces to the standard requirement that x_* be a strict local minimum of f, showing that the additional complication arises entirely from the weight-decay term.*

2. ***Equal momentum coefficients*** *($a = b$). In this case, equation 7 simplifies to $F''(x_*) > -L$ for some $L \geq 0$, which is implied by $F''(x_*) > 0$. Hence stability holds if and only if x_* is a strict local minimum of F (see Appendix A.11).*

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

We present the analysis in continuous time because the limit isolates the mechanism specific to AdamW. The exact discrete-time conditions (Appendix A.6) contain the same requirement $F''(x_*) > 0$ and a version of equation 8 whose coefficients depend on η, together with the familiar large-step-size ceiling, the analogue of $\eta < 2/L$ for gradient descent. The limit removes this ceiling and the η-dependence, leaving equation 8 in closed form.

4.2 Examples of Non-Convergence

Whenever $a \neq b$, one can construct objectives f that violate the stability condition equation 7 while still satisfying $F''(x_*) > 0$. Such objectives yield strict local minima of F, the very points an optimizer is meant to find, that AdamW's own dynamics fail to stabilize. 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 — a step-size-independent obstruction in a regime already known to perform poorly in practice. When $a > b$ (equivalently, $\beta_2 > \beta_1$), which is the regime most common in practice, instability arises only for non-convex objectives and must be constructed more carefully. We now provide simple examples illustrating both cases; in the common regime, the simulations exhibit bounded trajectories with persistent oscillations rather than convergence to the unstable equilibrium.

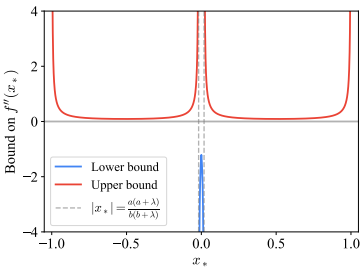


Figure 4: Hessian constraint in equation 8 for $a = 1,\ b = 10$.

Example 4.5. *When $\beta_2 < \beta_1$ (so $a < b$), equation 8 imposes an* upper *bound $f''(x_*) \leq U$, which 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 $\eta = 10^{-2}$ yields $a = 1$, $b = 10$, so equation 8 requires $f''(x_*) < 9.1 \times 10^{-2}$; but $f'' = 2$, hence the equilibrium is unstable. The corresponding discrete fixed point is also unstable at $\eta = 10^{-2}$, as shown in figure 1.

Example 4.6. *When $\beta_2 > \beta_1$ (so $a > b$), the condition imposes a* lower *bound $f''(x_*) \geq -L$, which can be violated by 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 $f''(x_*) \approx -1.20$, violating the required bound $f''(x_*) > -0.22$. The equilibrium is therefore unstable, and the displayed AdamW trajectory again fails to converge to it (see figure 2).

We construct these worst-case counterexamples on controlled objectives to show that AdamW admits no general convergence guarantee without additional assumptions. While our analysis yields theoretical insight into AdamW's dynamics, its practical implications for model training are more subtle: modern training is evaluated by validation performance rather than asymptotic parameter convergence, and LLMs routinely achieve strong downstream performance despite non-convergent or non-monotone training dynamics [7, 12].

4.3 A Convergent Modification

The instability above suggests modifying the weight-decay term while preserving AdamW's fixed-point equation. We replace the fixed weight decay $\lambda \boldsymbol{x}_t$ with an adaptive one that scales with the strength of AdamW's gradient update, yielding the ODE dynamics

$$\dot{\boldsymbol{x}}_t = -\frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t} + e} - \lambda \frac{|\boldsymbol{m}_t| + e}{\sqrt{\boldsymbol{v}_t} + e} \boldsymbol{x}_t, \quad \dot{\boldsymbol{m}}_t = a(\nabla f(\boldsymbol{x}_t) - \boldsymbol{m}_t), \quad \dot{\boldsymbol{v}}_t = b(\nabla f(\boldsymbol{x}_t)^{\odot 2} - \boldsymbol{v}_t). \tag{9}$$

We call this modification AdamMe. At equilibrium, $|\boldsymbol{m}_*| = \sqrt{\boldsymbol{v}_*}$, so the additional multiplier equals one and AdamW's fixed-point structure is preserved. The corresponding discrete update is given in Algorithm 1. AdamMe is designed to change only how the dynamics move, not where they come to rest: as shown in the following proposition, it targets exactly the same solutions as AdamW.

Proposition 4.7. *The equilibria of equation 9 coincide with the fixed points of AdamW. Consequently, AdamMe has the same implicit fixed-point objective F as in Theorem 3.1.*

Proof. At equilibrium, $\boldsymbol{m}_* = \nabla f(\boldsymbol{x}_*)$ and $\sqrt{\boldsymbol{v}_*} = |\nabla f(\boldsymbol{x}_*)| = |\boldsymbol{m}_*|$. Substituting these identities into the $\boldsymbol{x}$-equation of equation 9 gives

$$\frac{\nabla f(\boldsymbol{x}_*)}{|\nabla f(\boldsymbol{x}_*)| + e} + \lambda \boldsymbol{x}_* = \boldsymbol{0},$$

which is exactly the AdamW fixed-point condition equation 1. The equivalence with the stationary points of F then follows from Theorem 3.1. □

Theorem 4.8 (Lyapunov convergence of AdamMe)**.** *Assume f is C^1 with locally Lipschitz gradient, $a, b, \lambda, e > 0$, and $b \leq 4a$. Along every solution of equation 9 initialized with $\lambda|x_i(0)| < 1$ and $v_i(0) > 0$, this domain is forward invariant and the Lyapunov function V defined in Appendix A.3 is nonincreasing. Consequently, every bounded trajectory approaches the equilibrium set; if the limiting equilibria are isolated, the trajectory converges to one equilibrium.*

The explicit Lyapunov function and proof are given in Appendix A.3. Thus AdamMe preserves the same implicit objective as AdamW while restoring a convergence guarantee for the modified continuous-time dynamics. This is a constructive counterpart to the negative result: the implicit objective itself is not the source of non-convergence; rather, the issue lies in how AdamW's dynamics approach its fixed points.

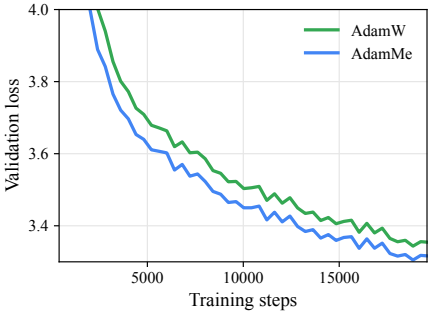


Figure 5: GPT-2-Small validation loss. AdamMe tracks AdamW and achieves better validation loss (3.317 vs. 3.355).

As a sanity check, we evaluate AdamMe on a small language-model run following Amsel et al. [1] with its GPT-2-Small (124M) configuration. As shown in figure 5, AdamMe closely tracks AdamW and reaches a slightly lower validation loss under hyperparameters tuned for each method, suggesting that the correction preserves AdamW-like training behavior with a small performance gain.

5 Multi-Dimensional Stability of AdamW

In the multi-dimensional setting, the Hessian and AdamW's coordinatewise preconditioner interact as matrices, so the scalar cubic analysis no longer gives a compact stability boundary. While local stability can be checked numerically from the full Jacobian spectrum, this gives little theoretical insight into when stability holds. We therefore give a minimal sufficient condition using a Lyapunov function. The linearized dynamics are given by the following lemma.

Lemma 5.1. *Let $\boldsymbol{z}_* = (\boldsymbol{x}_*, \boldsymbol{m}_*, \boldsymbol{h}_*)$ be an equilibrium point of the AdamW ODE equation 4. Then the linearized dynamics $\dot{\boldsymbol{z}}_t = \boldsymbol{J}(\boldsymbol{z}_*)\boldsymbol{z}_t$ imply that $\boldsymbol{x}_t$ satisfies*

$$\dddot{\boldsymbol{x}}_t + \boldsymbol{A}_2 \ddot{\boldsymbol{x}}_t + \boldsymbol{A}_1 \dot{\boldsymbol{x}}_t + \boldsymbol{A}_0 \boldsymbol{x}_t = 0, \tag{10}$$

where

$$\boldsymbol{A}_2 = (a+b+\lambda)\boldsymbol{I} \quad \boldsymbol{A}_0 = ab\left[\operatorname{Diag}(e/(\boldsymbol{h}_*+e)^2)\nabla^2 f(\boldsymbol{x}_*) + \lambda \boldsymbol{I}\right]$$
$$\boldsymbol{A}_1 = (ab+a\lambda+b\lambda)\boldsymbol{I} + (a-b)\operatorname{Diag}(1/(\boldsymbol{h}_*+e))\nabla^2 f(\boldsymbol{x}_*) + b\operatorname{Diag}(e/(\boldsymbol{h}_*+e)^2)\nabla^2 f(\boldsymbol{x}_*),$$

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

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

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

$$V(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) = \tfrac{1}{2}\Big(\|\boldsymbol{A}_2\boldsymbol{x}+\boldsymbol{y}\|_{\boldsymbol{P}}^2 + \|\boldsymbol{A}_2\boldsymbol{y}+\boldsymbol{z}\|_{\boldsymbol{Q}}^2 + \|\boldsymbol{y}\|_{\boldsymbol{K}}^2\Big),$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}V(\boldsymbol{x}_t,\boldsymbol{y}_t,\boldsymbol{z}_t) = -\tfrac{1}{2}\boldsymbol{y}^\top(\boldsymbol{L}+\boldsymbol{L}^\top)\boldsymbol{y}.$$

Thus $\dot{V} \le 0$ whenever $\boldsymbol{L}+\boldsymbol{L}^\top$ 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 $\boldsymbol{Q}$ such that 1. $\boldsymbol{Q}\boldsymbol{A}_0$ and $\boldsymbol{Q}\boldsymbol{A}_1$ are symmetric, and 2. $\boldsymbol{Q}\boldsymbol{A}_0$, $\boldsymbol{Q}\boldsymbol{A}_1$, and $\boldsymbol{Q}(\boldsymbol{A}_1\boldsymbol{A}_2 - \boldsymbol{A}_0)$ are positive definite.*

Here $\boldsymbol{Q}$ specifies the metric used in the Lyapunov function. Once an equilibrium is fixed, $\boldsymbol{A}_0,\boldsymbol{A}_1,\boldsymbol{A}_2$ are fixed by the local Hessian and AdamW preconditioner, and the proposition asks whether some $\boldsymbol{Q}$ satisfies the stated symmetry and positivity conditions. This constitutes a minimal sufficient condition and exact local stability is determined by the spectrum of the full $3d \times 3d$ Jacobian.

6 Related Works

Convergence analysis of Adam. Kingma and Ba [29] introduced Adam with a convergence proof for online convex optimization, which Reddi et al. [51] later showed to contain a gap. Their counterexample, however, relies on carefully constructed adversarial online objectives and does not preclude convergence for a fixed problem, where Adam converges once β_2 is large enough and $\beta_1 < \sqrt{\beta_2}$ [63, 64]. Follow-up work proved convergence for non-convex functions under bounded gradients and consistent gradient signs [16, 17], later relaxed to unbounded gradients [63] and to non-uniform smoothness and affine-variance noise [57, 33, 25, 26], narrowing the gap with SGD; Dereich et al. [19] further derived sharp local rates near strict minimizers. On the negative side, Dereich et al. [18] showed that even on simple stochastic quadratics, Adam converges to the minimizer only when the data are symmetrically distributed. A complementary line of work studies Adam's implicit bias to explain its empirical success [61, 56, 37].

Adam with decoupled weight decay. In practice, AdamW [42], the decoupled weight decay variant of Adam, has become the default choice over the original Adam. Weight decay itself was introduced by Hanson and Pratt [22], and while it is equivalent to L_2 regularization in SGD, Loshchilov and Hutter [42] 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 [30, 15, 58, 5, 23, 9].

Continuous-time optimizer analysis. Continuous-time analysis has long been used to study optimizer stability and acceleration: heavy-ball dynamics give a dissipative ODE model [2], the Nesterov ODE explains acceleration and restart behavior [55], and later work extends this viewpoint to accelerated mirror descent, Bregman/Hamiltonian flows, Lyapunov analyses, and numerical discretizations [31, 59, 52, 53, 45, 38, 39, 46, 10]. For stochastic algorithms, modified-equation and diffusion-limit analyses approximate SGD and adaptive methods in continuous time [35, 36, 47, 44], while Barakat and Bianchi [3] develop an Adam-specific ODE and convergence analysis. Our use of an AdamW ODE is different in focus: we characterize the static fixed-point objective induced by decoupled weight decay, then show that its stationary points need not be stable equilibria of the AdamW dynamics.

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. [65] considered exponentially decaying weight decay $\lambda_k = \lambda(1-\alpha)^k$, in contrast to the standard constant choice, and defined convergence through $\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla F_k(\boldsymbol{x}_k)\|_2^2]$, where F_k is a state-dependent "dynamically regularized" objective. Recently Li et al. [34] established a rate of $\mathcal{O}(d^{1/2}T^{-1/4})$ for the criterion $\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla f(\boldsymbol{x}_k)\|_1] \leq \epsilon^2$, 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 small momentum coefficients ($\beta_1 = \mathcal{O}(T^{-1/2})$). Xie and Li [60] examined AdamW's implicit objective in the limit $e \to 0$, 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 do not conflict with previous analyses.

7 Conclusions

We studied the dynamics of AdamW and showed that, despite admitting an implicit objective, its equilibria can be unstable; on simple problems, numerical trajectories then exhibit persistent oscillations rather than convergence. Extending the analysis to higher dimensions, we derived a sufficient condition for local stability via a Lyapunov function. As a constructive byproduct, an adaptive rescaling of the weight decay (AdamMe) preserves AdamW's fixed points and admits a convergence guarantee. Together, these results reveal that AdamW's theoretical properties are unexpectedly fragile: its stability depends on delicate interactions between curvature, weight decay, and momentum, yet the method remains the default optimizer in practice. This contrast points to room for principled refinements that retain AdamW's empirical strengths while offering stronger guarantees.

Limitations. First, in higher dimensions we give only sufficient conditions for the stability of AdamW (section 5). For any given configuration, the eigenvalues of the $3d \times 3d$ Jacobian equation 5 can still be computed with computer-aided tools to decide Hurwitz stability exactly, but the resulting closed-form expressions are too unwieldy to read, so we present compact sufficient conditions instead. Second, although we show that AdamW fails to converge on some objectives, how severe this is for large-scale neural network training remains unclear. Neural networks need not converge to be useful, and in our experiments training appears robust to this instability; it has not prevented the success of the many foundation models trained with AdamW [21, 40]. Third, the stability conditions for discrete-time AdamW are complicated even in one dimension. We therefore report these conditions, together with visualizations, in Appendix A.6.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Institute for Foundations of Machine Learning (IFML). The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper.

References

[1] Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

[2] Hedy Attouch, Xavier Goudou, and Patrick Redont. The heavy ball with friction method, i. the continuous dynamical system: Global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. *Communications in Contemporary Mathematics*, 2(1):1–34, 2000. doi: 10.1142/S0219199700000025.

[3] Anas Barakat and Pascal Bianchi. Convergence and dynamical behavior of the adam algorithm for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 31(1):244–274, 2021. doi: 10.1137/19M1263443.

[4] André Belotto da Silva and Maxime Gazeau. A general system of differential equations to model first-order adaptive algorithms. *Journal of Machine Learning Research*, 21(129):1–42, 2020.

[5] 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.

[6] Wolf-Jürgen Beyn. On the numerical approximation of phase portraits near stationary points. *SIAM Journal on Numerical Analysis*, 24(5):1095–1113, 1987. doi: 10.1137/0724072.

[7] Nisha Chandramoorthy, Andreas Loukas, Khashayar Gatmiry, and Stefanie Jegelka. On the generalization of learning algorithms that do not converge. In *Advances in Neural Information Processing Systems*, volume 35, pages 34241–34257, 2022.

[8] Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves constrained optimization: As lyapunov predicts. *arXiv preprint arXiv:2310.05898*, 2023. URL `https://arxiv.org/abs/2310.05898v7`. Version 7, revised July 2025.

[9] Lizhang Chen, Jonathan Li, Kaizhao Liang, Baiyu Su, Cong Xie, Nuo Wang Pierse, Chen Liang, Ni Lao, and Qiang Liu. Cautious weight decay. In *International Conference on Learning Representations (ICLR)*, 2026. URL `https://openreview.net/forum?id=Gwe6gbGng5`.

[10] Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research (TMLR)*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=Blz4hjxLwU`.

[11] 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, 2023.

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

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

[14] Enea Monzio Compagnoni, Tianlin Liu, Rustem Islamov, Frank Norbert Proske, Antonio Orvieto, and Aurelien Lucchi. Adaptive methods through the lens of SDEs: Theoretical insights on the role of noise. In *International Conference on Learning Representations*, 2025.

[15] 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.

[16] 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.

[17] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

[18] Steffen Dereich, Thang Do, Arnulf Jentzen, and Philippe von Wurstemberger. Adam symmetry theorem: Characterization of the convergence of the stochastic Adam optimizer. *arXiv preprint arXiv:2511.06675*, 2025. URL https://arxiv.org/abs/2511.06675.

[19] Steffen Dereich, Arnulf Jentzen, and Adrian Riekert. Sharp higher order convergence rates for the Adam optimizer. *arXiv preprint arXiv:2504.19426*, 2025. URL https://arxiv.org/abs/2504.19426.

[20] 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*, 2024.

[21] 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.

[22] Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1, 1988.

[23] 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.

[24] Morris W. Hirsch, Stephen Smale, and Robert L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 3rd edition, 2013.

[25] Yusu Hong and Junhong Lin. On convergence of Adam for stochastic optimization under relaxed assumptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. URL https://arxiv.org/abs/2402.03982.

[26] 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.

[27] 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/.

[28] Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3rd edition, 2002. ISBN 0130673897.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. *arXiv preprint arXiv:2305.17212*, 2023.

[31] Walid Krichene, Alexandre Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[32] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR, 2016.

[33] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36:52166–52196, 2023.

[34] 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.

[35] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2101–2110. PMLR, 2017.

[36] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019.

[37] Xinghan Li, Haodong Wen, and Kaifeng Lyu. Adam reduces a unique form of sharpness: Theoretical insights near the minimizer manifold. *arXiv preprint arXiv:2511.02773*, 2025.

[38] Kaizhao Liang, Bo Liu, Lizhang Chen, and Qiang Liu. Memory-efficient LLM training with online subspace descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 64412–64432, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/760d09bcc06b949f5ac4b6a918739aa8-Abstract-Conference.html.

[39] Kaizhao Liang, Lizhang Chen, Bo Liu, and Qiang Liu. Cautious optimizers: Improving training with one line of code. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=zBPZeRjfgu.

[40] 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.

[41] Stanisław Łojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. In *Séminari di Geometria 1982–1983*, pages 115–117, Bologna, Italy, 1984. Istituto di Geometria, Dipartimento di Matematica, Università di Bologna.

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[43] Chris J Maddison, Daniel Paulin, Yee Whye Teh, Brendan O'Donoghue, and Arnaud Doucet. Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*, 2018.

[44] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[45] Michael Muehlebach and Michael I. Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22 (73):1–50, 2021.

[46] Son Nguyen, Lizhang Chen, Bo Liu, and Qiang Liu. Memory-efficient optimization with factorized Hamiltonian descent. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 258 of *Proceedings of Machine Learning Research*, pages 2863–2871, 2025. URL https://proceedings.mlr.press/v258/nguyen25e.html.

[47] Antonio Orvieto and Aurelien Lucchi. Continuous-time models for stochastic optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[48] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

[49] PyTorch Contributors. *torch.optim.AdamW — PyTorch Documentation*. PyTorch, 2025. Accessed: September 19, 2025. https://docs.pytorch.org/docs/stable/generated/torch.optim.AdamW.html.

[50] 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.

[51] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[52] Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[53] Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195:79–148, 2022. doi: 10.1007/s10107-021-01681-8.

[54] A. M. Stuart and A. R. Humphries. *Dynamical Systems and Numerical Analysis*, volume 2 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 1996.

[55] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[56] 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.

[57] Bohan Wang, Yushun Zhang, Huishuai Zhang, Qi Meng, Ruoyu Sun, Zhi-Ming Ma, Tie-Yan Liu, Zhi-Quan Luo, and Wei Chen. Provable adaptivity of Adam under non-uniform smoothness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024. URL https://arxiv.org/abs/2208.09900.

[58] 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.

[59] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016. doi: 10.1073/pnas.1614734113.

[60] 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, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54488–54510. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xie24e.html.

[61] 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.

[62] Manzil Zaheer, Sashank J. Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 9793–9803, 2018.

[63] 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.

[64] Yushun Zhang, Bingran Li, Congliang Chen, Zhi-Quan Luo, and Ruoyu Sun. Adam converges without any modification on update rules. *arXiv preprint arXiv:2603.02092*, 2026. URL https://arxiv.org/abs/2603.02092.

[65] 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 $\boldsymbol{g}_t = \nabla f(\boldsymbol{x}_{t-1})$. The AdamW updates (Algorithm 1) are

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

With the common approximation $\hat{\boldsymbol{m}}_t \approx \boldsymbol{m}_t,\ \hat{\boldsymbol{v}}_t \approx \boldsymbol{v}_t$ and the substitution $\boldsymbol{h}_t = \sqrt{\boldsymbol{v}_t}$, 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:

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

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

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

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

$$\frac{\boldsymbol{v}_t - \boldsymbol{v}_{t-1}}{\Delta t} = -b\boldsymbol{v}_{t-1} + b\boldsymbol{g}_t^{\odot 2} \quad \to \quad \dot{\boldsymbol{h}}_t = \tfrac{b}{2}\left(\frac{\nabla f(\boldsymbol{x}_t)^{\odot 2}}{\boldsymbol{h}_t} - \boldsymbol{h}_t\right).$$

Semi-implicit Euler discretization. Starting from the AdamW ODE

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

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

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

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 (Restatement of Theorem 3.1)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, and let $\lambda, e > 0$. Then $\boldsymbol{x} \in \mathbb{R}^d$ satisfies the fixed-point condition equation 1 if and only if $\boldsymbol{x}$ is a stationary point of the ℓ_∞-constrained and regularized problem*

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \ F(\boldsymbol{x}) = f(\boldsymbol{x}) + \tfrac{e}{\lambda}\,\Phi(\lambda \boldsymbol{x}), \quad \textit{subject to} \quad \|\boldsymbol{x}\|_\infty \le \tfrac{1}{\lambda}, \tag{11}$$

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

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

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

We begin with the following lemma with the regularizer:

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

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

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

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

Proof. The stated derivatives follow by direct differentiation on $(-1, 1)$; convexity holds since $\phi''_\lambda \geq 0$. □

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

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

where the left-hand equation implies $|\lambda x| < 1$, so that $\phi'_\lambda(x)$ on the right-hand side is well-defined. In $\mathbb{R}^d$, the equivalence holds elementwise.

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 $\operatorname{sign}(x) = -\operatorname{sign}(g)$. Simple algebra yields $g + e\,\frac{\lambda x}{1-|\lambda x|} = 0$. The converse is identical in reverse. □

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

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

Proof. Both identities follow by differentiating coordinatewise and invoking Lemma A.2. □

Proof of Theorem 3.1. Fixed points $\iff$ stationary points. For AdamW fixed points (discrete or continuous limit without bias corrections), the stationary condition for $\boldsymbol{x}$ reads

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

By Lemma A.3 this is equivalent to

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

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

Conversely, any stationary point of equation 2 must satisfy $\|\lambda \boldsymbol{x}_*\|_\infty < 1$: since $\Phi(\lambda \boldsymbol{x}) \to +\infty$ 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 $\nabla F(\boldsymbol{x}_*) = 0$, which is equivalent (by Lemma A.3) to the AdamW fixed-point condition. This proves the coincidence of fixed points and stationary points. □

A.3 Lyapunov Function for AdamMe

We prove Theorem 4.8. Throughout this subsection, all vector operations are coordinatewise, and we consider solutions initialized with

$$\lambda |x_i(0)| < 1, \qquad v_i(0) > 0, \qquad i = 1, \ldots, d.$$

On the domain $\{v_i > 0\}$ the right-hand side of equation 9 is locally Lipschitz, since ∇f is locally Lipschitz, $|m_i|$ is Lipschitz, and $\sqrt{v_i}$ is smooth for $v_i > 0$; hence solutions exist, are unique, and are C^1. Lemma A.6 shows that this domain is preserved. For each coordinate, define

$$h_i \coloneqq \sqrt{v_i}, \qquad \hbar_i \coloneqq h_i + e, \qquad \psi_i(x_i, m_i) \coloneqq m_i + \lambda x_i(|m_i| + e),$$

and let

$$r_i(x_i) \coloneqq -\frac{\lambda e x_i}{1-\lambda|x_i|}.$$

Then $r_i(x_i)$ is the unique root of $\psi_i(x_i, r_i(x_i)) = 0$. Define the coordinate barrier induced by the same ϕ as in Theorem 3.1,

$$R_i(x_i) \coloneqq \frac{e}{\lambda}\phi(\lambda x_i),$$

so that $R_i'(x_i) = -r_i(x_i)$, and define

$$J_i(x_i, m_i) \coloneqq \frac{1}{2}\big(m_i|m_i| - r_i(x_i)|r_i(x_i)|\big) + e(m_i - r_i(x_i)).$$

The coordinate energy term is

$$I_i(x_i, m_i) \coloneqq \frac{1}{2}(m_i^2 - r_i(x_i)^2) + \lambda x_i J_i(x_i, m_i).$$

When $b \le 4a$, consider

$$V(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{v}) = f(\boldsymbol{x}) + \sum_{i=1}^{d} R_i(x_i) + \frac{1}{a}\sum_{i=1}^{d} \frac{I_i(x_i, m_i)}{\hbar_i}.$$

Lemma A.5. *Along every solution of the AdamMe ODE equation 9 in the above domain,*

$$\frac{d}{dt}V(\boldsymbol{x}(t), \boldsymbol{m}(t), \boldsymbol{v}(t)) \le 0.$$

Proof. Fix a coordinate i and write the gradient coordinate as

$$\begin{aligned} g_i &\coloneqq \partial_i f(\boldsymbol{x}), & h_i &\coloneqq \sqrt{v_i}, & \hbar_i &\coloneqq h_i + e, \\ \psi_i &\coloneqq \psi_i(x_i, m_i), & r_i &\coloneqq r_i(x_i), & & \\ J_i &\coloneqq J_i(x_i, m_i), & I_i &\coloneqq I_i(x_i, m_i). & & \end{aligned}$$

The coordinate dynamics are

$$\dot{x}_i = -\frac{\psi_i}{\hbar_i}, \qquad \dot{m}_i = a(g_i - m_i), \qquad \dot{v}_i = b(g_i^2 - v_i).$$

Since $\psi_i(x_i, r_i(x_i)) = 0$, direct differentiation of the explicit form of I_i gives

$$\partial_{m_i} I_i = \psi_i, \qquad \partial_{x_i} I_i = \lambda J_i,$$

and

$$\partial_{v_i}\left(\frac{I_i}{\hbar_i}\right) = -\frac{I_i}{2h_i\hbar_i^2}.$$

Therefore

$$\dot{V}_i = g_i\dot{x}_i + R_i'(x_i)\dot{x}_i + \frac{1}{a}\frac{\partial_{x_i} I_i}{\hbar_i}\dot{x}_i + \frac{1}{a}\frac{\partial_{m_i} I_i}{\hbar_i}\dot{m}_i - \frac{1}{a}\frac{I_i}{2h_i\hbar_i^2}\dot{v}_i.$$

Substituting the dynamics and using $R_i'(x_i) = -r_i$ gives

$$\dot{V}_i = -\frac{g_i\psi_i}{\hbar_i} + \frac{r_i\psi_i}{\hbar_i} - \frac{\lambda\psi_i}{a\hbar_i^2}J_i + \frac{\psi_i}{\hbar_i}(g_i - m_i) - \frac{bI_i}{2ah_i\hbar_i^2}(g_i^2 - v_i).$$

The two g_i terms cancel, so

$$\dot{V}_i = -\frac{\psi_i}{\hbar_i}\left[m_i - r_i + \frac{\lambda}{a\hbar_i}J_i\right] - \frac{bI_i}{2ah_i\hbar_i^2}(g_i^2 - v_i).$$

Because $\lambda|x_i| < 1$, the map $m \mapsto \psi_i(x_i, m)$ is strictly increasing. Hence

$$\operatorname{sign}(\psi_i) = \operatorname{sign}(m_i - r_i).$$

Moreover, since $u \mapsto \frac{1}{2}u|u| + eu$ is strictly increasing,

$$\operatorname{sign}(J_i) = \operatorname{sign}(m_i - r_i), \qquad \psi_i J_i \ge 0.$$

The explicit expression for I_i also gives

$$I_i(x_i, m_i) \geq 0, \qquad \psi_i(x_i, m_i)(m_i - r_i) \geq 2I_i(x_i, m_i).$$

Indeed, both inequalities follow by direct inspection of the three cases $m_i, r_i \geq 0$, $m_i, r_i \leq 0$, and m_i and r_i on opposite sides of zero. Combining these facts,

$$-\frac{\psi_i}{\hbar_i}\left[m_i - r_i + \frac{\lambda}{a\hbar_i}J_i\right] \leq -\frac{2I_i}{\hbar_i}.$$

Therefore

$$\dot{V}_i \leq -\frac{2I_i}{\hbar_i} - \frac{bI_i}{2ah_i\hbar_i^2}(g_i^2 - v_i).$$

If $g_i^2 \geq v_i$, the second term is nonpositive. If $g_i^2 < v_i$, then

$$-\frac{bI_i}{2ah_i\hbar_i^2}(g_i^2 - v_i) = \frac{bI_i}{2ah_i\hbar_i^2}(v_i - g_i^2) \leq \frac{bI_ih_i}{2a\hbar_i^2},$$

and hence

$$\dot{V}_i \leq -\frac{I_i}{\hbar_i}\left[2 - \frac{bh_i}{2a\hbar_i}\right].$$

Since $0 < h_i/\hbar_i < 1$ and $b \leq 4a$, the bracket is nonnegative. Thus $\dot{V}_i \leq 0$ in both cases. Summing over coordinates proves the claim. □

Lemma A.6. *The domain $\{\boldsymbol{x} : \lambda|x_i| < 1,\ i = 1, \ldots, d\}$ is forward invariant. Moreover, if $v_i(0) > 0$, then $v_i(t) > 0$ for all $t \geq 0$.*

Proof. At the boundary $x_i = 1/\lambda$,

$$\psi_i = m_i + |m_i| + e > 0,$$

so $\dot{x}_i = -\psi_i/(\sqrt{v_i} + e) < 0$. At the boundary $x_i = -1/\lambda$,

$$\psi_i = m_i - |m_i| - e < 0,$$

so $\dot{x}_i = -\psi_i/(\sqrt{v_i} + e) > 0$. Thus the vector field points inward at the boundary of the box.

For the second claim, the v_i equation gives

$$v_i(t) = e^{-bt}v_i(0) + b\int_0^t e^{-b(t-s)}\left(\partial_i f(\boldsymbol{x}(s))\right)^2 ds > 0.$$

□

Proof of Theorem 4.8. Lemmas A.5 and A.6 show that V is a Lyapunov function on the stated forward-invariant domain. For any bounded trajectory, the remaining argument follows the standard LaSalle routine, as in the final step of Theorem B.3 of Chen et al. [8]. Let the limiting set denote the set of all subsequential limits along times $t_k \to \infty$. LaSalle's invariance principle implies that this limiting set is contained in the largest invariant subset of $\dot{V} = 0$, with our vector field and Lyapunov function replacing the Lion-$\mathcal{K}$ dynamics and Lyapunov function there.

From the proof of Lemma A.5, equality requires $I_i(x_i, m_i) = 0$ for every coordinate, hence $m_i = r_i(x_i)$ and $\psi_i(x_i, m_i) = 0$. Therefore $\dot{\boldsymbol{x}} = 0$. On any invariant subset of this set, we must also have $\dot{\boldsymbol{m}} = 0$, so $\boldsymbol{m} = \nabla f(\boldsymbol{x})$. Finally, $\dot{\boldsymbol{v}} = 0$ gives $\boldsymbol{v} = \nabla f(\boldsymbol{x})^{\odot 2}$. These are exactly the equilibrium conditions for equation 9. Hence every bounded trajectory approaches the equilibrium set. If the limiting equilibria are isolated, the trajectory converges to a single equilibrium.

□

A.4 Discrete-Time Convergence of AdamMe

We adapt the energy-increment argument of Theorem B.8 and the convex-combination argument of Theorem B.10 of Chen et al. [8]. Those results concern Lion-$\mathcal{K}$ with a fixed convex potential; the additional step for AdamMe is to control its changing second-moment denominator. The functions F, I_i, ψ_i, and r_i below are those defined above.

Proposition A.7. *Assume f is C^1 with L-Lipschitz gradient in the Euclidean norm on a neighborhood of $\mathcal{C} = \{\|\lambda \boldsymbol{x}\|_\infty \le 1\}$, $\lambda, e > 0$, and $\beta_1, \beta_2 \in (0,1)$ satisfy*

$$s \coloneqq \sqrt{\beta_2} > \frac{\beta_1}{2-\beta_1}. \tag{12}$$

Given an initialization with $\|\lambda \boldsymbol{x}_0\|_\infty < 1$ and $\boldsymbol{v}_0 \ge 0$, write $\alpha = 1-\beta_1$ and define

$$M = \max\left\{\|\boldsymbol{m}_0\|_\infty,\ \|\nabla f(\boldsymbol{0})\|_\infty + \frac{L\sqrt{d}}{\lambda}\right\},$$
$$\rho = \min\left\{1 - \|\lambda \boldsymbol{x}_0\|_\infty,\ \frac{e}{M+e}\right\}, \qquad \kappa = 2 - \frac{\beta_1(1-s)}{\alpha s} > 0,$$
$$L_0 = L + \frac{\lambda e}{\rho^2}, \qquad L_1 = \frac{\beta_1 \lambda^2 e}{\alpha s \rho^3}.$$

For every constant step size

$$0 < \eta \le \eta_{\max} \coloneqq \min\left\{\frac{e}{\lambda(M+e)},\quad \frac{\kappa e}{2\left(L_0 + \sqrt{L_0^2 + \kappa e L_1}\right)}\right\}, \tag{13}$$

deterministic AdamMe without bias correction remains in $\{\|\lambda \boldsymbol{x}\|_\infty \le 1-\rho\}$ and satisfies

$$\frac{1}{T}\sum_{k=0}^{T-1} \|\nabla F(\boldsymbol{x}_k)\|^2 = O(T^{-1}), \qquad \nabla F(\boldsymbol{x}_k) \longrightarrow 0.$$

If the stationary points in this compact set are isolated, the full state converges to an equilibrium. The same conclusions hold with the bias corrections and zero moment initialization in Algorithm 1.

Proof. Write $d_{k,i} = \sqrt{v_{k,i}} + e$ and $\Delta m_{k,i} = m_{k+1,i} - m_{k,i}$. The updates are

$$\boldsymbol{m}_{k+1} = \beta_1 \boldsymbol{m}_k + \alpha \nabla f(\boldsymbol{x}_k), \qquad \boldsymbol{v}_{k+1} = \beta_2 \boldsymbol{v}_k + (1-\beta_2)\nabla f(\boldsymbol{x}_k)^{\odot 2},$$
$$x_{k+1,i} = x_{k,i} - \eta \frac{\psi_i(x_{k,i}, m_{k+1,i})}{d_{k+1,i}}.$$

Smoothness gives $\|\nabla f(\boldsymbol{x})\|_\infty \le \|\nabla f(\boldsymbol{0})\|_\infty + L\|\boldsymbol{x}\|_2 \le M$ on $\mathcal{C}$. Set $q = 1-\rho = \max\{\|\lambda \boldsymbol{x}_0\|_\infty, M/(M+e)\} < 1$. The first bound in equation 13 ensures $\eta\lambda(M+e) \le e$, so the identity

$$\lambda x_{k+1,i} = (1-\theta_{k,i})\lambda x_{k,i} - \theta_{k,i}\frac{m_{k+1,i}}{|m_{k+1,i}| + e}, \qquad \theta_{k,i} = \eta\lambda\frac{|m_{k+1,i}| + e}{d_{k+1,i}} \in [0,1]$$

proves inductively that $\|\lambda \boldsymbol{x}_k\|_\infty \le q$, as in the proof of Theorem B.10 of Chen et al. [8]. Both moments are then bounded, and $e \le d_{k,i} \le D$ for some $D < \infty$.

Define the discrete energy by

$$W_k = F(\boldsymbol{x}_k) + c_\eta \sum_i \frac{I_i(x_{k,i}, m_{k,i})}{d_{k,i}}, \qquad c_\eta = \frac{\eta\beta_1}{\alpha s}.$$

The second-moment update gives $d_{k+1,i} \ge s d_{k,i}$. Moreover, $I_i(x,\cdot)$ is ρ-strongly convex with derivative $\psi_i(x,\cdot)$, whose Lipschitz constant is at most 2. Thus, abbreviating $I_i^+ = I_i(x_{k,i}, m_{k+1,i})$ and $\psi_i^+ = \psi_i(x_{k,i}, m_{k+1,i})$, we obtain the required denominator estimate:

$$\frac{I_i^+}{d_{k+1,i}} - \frac{I_i(x_{k,i}, m_{k,i})}{d_{k,i}} \le \frac{(1-s)I_i^+ + s\psi_i^+ \Delta m_{k,i} - \frac{s\rho}{2}(\Delta m_{k,i})^2}{d_{k+1,i}}.$$

For the position update, hold the updated moments $(\boldsymbol{m}_{k+1}, \boldsymbol{v}_{k+1})$ fixed and regard the energy as a function of $\boldsymbol{x}$ alone:

$$P_k(\boldsymbol{x}) \coloneqq F(\boldsymbol{x}) + c_\eta \sum_i \frac{I_i(x_i, m_{k+1,i})}{d_{k+1,i}}.$$

The full energy increment splits into a position increment and a moment increment,

$$W_{k+1} - W_k = P_k(\boldsymbol{x}_{k+1}) - P_k(\boldsymbol{x}_k) + c_\eta \sum_i \left(\frac{I_i^+}{d_{k+1,i}} - \frac{I_i(x_{k,i}, m_{k,i})}{d_{k,i}} \right).$$

The preceding denominator estimate controls the second term. To bound the first, note that the barrier in F contributes at most $\lambda e/\rho^2$ to its gradient Lipschitz constant. Since $\partial_{x_i} I_i = \lambda J_i$, $\partial_{x_i}^2 I_i = \lambda^2 e^2/(1-\lambda|x_i|)^3$, and $d_{k+1,i} \geq e$, the function P_k is L_η-smooth on the invariant cube, where

$$L_\eta = L + \frac{\lambda e}{\rho^2} + \frac{c_\eta \lambda^2 e}{\rho^3} = L_0 + L_1 \eta.$$

Write $g_{k,i} = \nabla_i f(\boldsymbol{x}_k)$, $r_{k,i} = r_i(x_{k,i})$, and $J_i^+ = J_i(x_{k,i}, m_{k+1,i})$. Then $\partial_i P_k(\boldsymbol{x}_k) = g_{k,i} - r_{k,i} + c_\eta \lambda J_i^+/d_{k+1,i}$. Substituting $\Delta x_{k,i} = x_{k+1,i} - x_{k,i} = -\eta \psi_i^+/d_{k+1,i}$ into the smoothness inequality gives

$$\begin{aligned} P_k(\boldsymbol{x}_{k+1}) - P_k(\boldsymbol{x}_k) &\leq \sum_i \partial_i P_k(\boldsymbol{x}_k) \Delta x_{k,i} + \frac{L_\eta}{2} \|\Delta \boldsymbol{x}_k\|^2 \\ &= -\eta \sum_i \frac{(g_{k,i} - r_{k,i})\psi_i^+}{d_{k+1,i}} - c_\eta \eta \lambda \sum_i \frac{\psi_i^+ J_i^+}{d_{k+1,i}^2} + \frac{L_\eta \eta^2}{2} \sum_i \frac{(\psi_i^+)^2}{d_{k+1,i}^2}. \end{aligned}$$

Adding the moment-increment bound makes the cancellation explicit: because $g_{k,i} = m_{k+1,i} + \beta_1 \Delta m_{k,i}/\alpha$ and $c_\eta s = \eta \beta_1/\alpha$,

$$-\eta \frac{(g_{k,i} - r_{k,i})\psi_i^+}{d_{k+1,i}} + c_\eta s \frac{\psi_i^+ \Delta m_{k,i}}{d_{k+1,i}} = -\eta \frac{(m_{k+1,i} - r_{k,i})\psi_i^+}{d_{k+1,i}}.$$

Finally, $\psi_i^+ J_i^+ \geq 0$ makes the second term in the position bound nonpositive, while $\psi_i^+(m_{k+1,i} - r_{k,i}) \geq 2I_i^+$ bounds the remaining linear term. The quadratic remainder satisfies

$$\frac{L_\eta \eta^2}{2} \sum_i \frac{(\psi_i^+)^2}{d_{k+1,i}^2} \leq \frac{2 L_\eta \eta^2}{e} \sum_i \frac{I_i^+}{d_{k+1,i}},$$

using $(\psi_i^+)^2 \leq 4 I_i^+$ and $d_{k+1,i} \geq e$. Combining these bounds with $2\eta - c_\eta(1-s) = \eta\kappa$ yields

$$W_{k+1} - W_k \leq -\eta \left(\kappa - \frac{2\eta L_\eta}{e} \right) S_k - \frac{\eta \beta_1 \rho}{2\alpha} Q_k, \tag{14}$$

where $S_k = \sum_i I_i^+/d_{k+1,i}$ and $Q_k = \sum_i (\Delta m_{k,i})^2/d_{k+1,i}$. The second bound in equation 13 is the positive root bound for $L_1\eta^2 + L_0\eta \leq \kappa e/4$. Since $L_\eta = L_0 + L_1\eta$, this ensures $2\eta L_\eta/e \leq \kappa/2$, leaving a strictly positive coefficient for both S_k and Q_k in the descent estimate. Let $F_{\min} = \min_{\|\lambda \boldsymbol{x}\|_\infty \leq q} F(\boldsymbol{x})$, which is finite by continuity on the compact invariant cube. Because $I_i \geq 0$, we have $W_k \geq F_{\min}$. Summing equation 14 from $k = 0$ to $T-1$ therefore gives

$$\frac{\eta\kappa}{2} \sum_{k=0}^{T-1} S_k + \frac{\eta \beta_1 \rho}{2\alpha} \sum_{k=0}^{T-1} Q_k \leq W_0 - W_T \leq W_0 - F_{\min}.$$

The right-hand side is independent of T, so both nonnegative sequences are summable; in particular, $S_k \to 0$ and $Q_k \to 0$.

To translate these energy residuals into stationarity, use the strong convexity of $I_i(x_{k,i}, \cdot)$ around its minimizer $r_i(x_{k,i})$:

$$I_i^+ \geq \frac{\rho}{2} |m_{k+1,i} - r_i(x_{k,i})|^2, \qquad \nabla_i F(\boldsymbol{x}_k) = m_{k+1,i} - r_i(x_{k,i}) + \frac{\beta_1}{\alpha} \Delta m_{k,i}.$$

Applying $\|u+w\|^2 \le 2\|u\|^2 + 2\|w\|^2$ and the upper bound $d_{k+1,i} \le D$ yields

$$\|\nabla F(\boldsymbol{x}_k)\|^2 \le \frac{4D}{\rho} S_k + \frac{2\beta_1^2 D}{\alpha^2} Q_k.$$

Combining this inequality with the preceding telescoping bound gives, for every T,

$$\sum_{k=0}^{T-1} \|\nabla F(\boldsymbol{x}_k)\|^2 \le \frac{D(W_0 - F_{\min})}{\eta\rho} \max\left\{\frac{8}{\kappa}, \frac{4\beta_1}{\alpha}\right\}.$$

Thus the squared gradients are summable and $\nabla F(\boldsymbol{x}_k) \to 0$. Dividing the same bound by T proves the claimed $O(T^{-1})$ average stationarity rate.

For convergence of the iterates, the position update also gives

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\|^2 = \eta^2 \sum_i \frac{(\psi_i^+)^2}{d_{k+1,i}^2} \le \frac{4\eta^2}{e} S_k \longrightarrow 0.$$

Since the iterates remain in a compact set and $\nabla F(\boldsymbol{x}_k) \to 0$, these vanishing increments imply convergence to a stationary point $\boldsymbol{x}_*$ whenever the stationary points are isolated.

Since $S_k \to 0$, strong convexity yields $\boldsymbol{m}_{k+1} - \boldsymbol{r}(\boldsymbol{x}_k) \to 0$. Thus $\boldsymbol{m}_k \to \boldsymbol{r}(\boldsymbol{x}_*) = \boldsymbol{g}_*$, where $\boldsymbol{g}_* = \nabla f(\boldsymbol{x}_*)$ by stationarity. The second-moment error obeys

$$\boldsymbol{v}_{k+1} - \boldsymbol{g}_*^{\odot 2} = \beta_2(\boldsymbol{v}_k - \boldsymbol{g}_*^{\odot 2}) + (1-\beta_2)\big(\nabla f(\boldsymbol{x}_k)^{\odot 2} - \boldsymbol{g}_*^{\odot 2}\big).$$

Because $0 < \beta_2 < 1$ and the forcing term vanishes, this recursion gives $\boldsymbol{v}_k \to \boldsymbol{g}_*^{\odot 2}$. Hence the full state converges to an equilibrium.

Bias-corrected moments are bounded convex averages of the past gradients and their squared values, preserving domain invariance. Their position update differs from the uncorrected update by $O(\eta(\beta_1^{k+1} + \beta_2^{k+1}))$ on this compact set. This adds a summable remainder to equation 14, preserving the convergence conclusions. □

The condition in equation 12 is the discrete counterpart of $b < 4a$: substituting $\beta_1 = 1 - a\eta$ and $\beta_2 = 1 - b\eta$ gives $\kappa \to 2 - b/(2a)$ as $\eta \to 0$.

Stochastic gradients. The conditional-energy argument parallels Theorem B.14 of Chen et al. [8] (version 7). Assume additionally that the sampled gradients satisfy $\|\boldsymbol{g}_k\|_\infty \le G$ on the cube, are conditionally unbiased, and have conditional variance at most σ^2/B_k, where B_k is the batch size. Set $M = \max\{\|\boldsymbol{m}_0\|_\infty, G\}$ and compute $\rho, L_0, L_1, \eta_{\max}$ as in Proposition A.7. Writing $\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k) + \boldsymbol{\xi}_k$, the only additional term in equation 14 is $\eta \sum_i \psi_i^+ \xi_{k,i}/d_{k+1,i}$. Young's inequality and $(\psi_i^+)^2 \le 4I_i^+$ bound this by $\eta\kappa S_k/4 + 4\eta\|\boldsymbol{\xi}_k\|^2/(\kappa e)$. Therefore, for every $0 < \eta \le \eta_{\max}$,

$$\mathbb{E}_k W_{k+1} \le W_k - \frac{\eta\kappa}{4}\mathbb{E}_k S_k - \frac{\eta\beta_1\rho}{2\alpha}\mathbb{E}_k Q_k + \frac{4\eta\sigma^2}{\kappa e B_k},$$

$$\frac{1}{T}\sum_{k<T} \mathbb{E}\|\nabla F(\boldsymbol{x}_k)\|^2 \le \frac{C_0}{\eta T} + \frac{C_1\sigma^2}{T}\sum_{k<T} B_k^{-1},$$

where $\mathbb{E}_k$ conditions on the history before sampling $\boldsymbol{g}_k$ and C_0, C_1 are independent of T and the batch-size schedule. The second bound follows by telescoping and using $\nabla F(\boldsymbol{x}_k) = \boldsymbol{m}_{k+1} - \boldsymbol{r}(\boldsymbol{x}_k) + \beta_1(\boldsymbol{m}_{k+1} - \boldsymbol{m}_k)/\alpha - \boldsymbol{\xi}_k$. Thus fixed batches yield an $O(T^{-1} + \sigma^2/B)$ stationarity bound. If $\sum_k B_k^{-1} < \infty$, telescoping the first bound gives $\sum_k (S_k + Q_k + \|\boldsymbol{\xi}_k\|^2) < \infty$ almost surely, hence $\nabla F(\boldsymbol{x}_k) \to 0$ almost surely, and convergence to one equilibrium when stationary points are isolated. Bias correction again contributes only a summable remainder.

A.5 Stability Condition of AdamW ODE

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

$$\frac{d}{dt}\boldsymbol{z} = \boldsymbol{T}(\boldsymbol{z}), \quad \boldsymbol{z} \coloneqq \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{m} \\ \boldsymbol{h} \end{bmatrix}, \quad \boldsymbol{T}(\boldsymbol{z}) \coloneqq \begin{bmatrix} -\boldsymbol{m}/(\boldsymbol{h} + e) - \lambda\boldsymbol{x} \\ a(\nabla f(\boldsymbol{x}) - \boldsymbol{m}) \\ \frac{b}{2}(\nabla f(\boldsymbol{x})^{\odot 2}/\boldsymbol{h} - \boldsymbol{h}) \end{bmatrix}. \tag{15}$$

By differentiating the system $\boldsymbol{J}(\boldsymbol{z}) \coloneqq \partial \boldsymbol{T}(\boldsymbol{z})/\partial \boldsymbol{z}$, we obtain:

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

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

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

This following lemma is useful for simplifying the characteristic polynomial:

Lemma A.8. *If $[\boldsymbol{x}_*, \boldsymbol{m}_*, \boldsymbol{h}_*]$ is a fixed point of AdamW in Eq. equation 1, then*

$$\nabla^2 F(\boldsymbol{x}_*) = \nabla^2 f(\boldsymbol{x}_*) + \mathrm{Diag}\tfrac{\lambda e}{(1-|\lambda \boldsymbol{x}_*|)^2} = \nabla^2 f(\boldsymbol{x}_*) + \mathrm{Diag}\tfrac{\lambda}{e}\,(\boldsymbol{h}_* + e)^2.$$

Proof. At equilibrium, we have $\frac{\boldsymbol{m}_*}{\boldsymbol{h}_*+e} + \lambda \boldsymbol{x}_* = 0$ and $\boldsymbol{h}_* = |\nabla f(\boldsymbol{x}_*)| = |\boldsymbol{m}_*|$. Hence

$$|\lambda \boldsymbol{x}_*| = \frac{|\boldsymbol{m}_*|}{\boldsymbol{h}_* + e} = \frac{\boldsymbol{h}_*}{\boldsymbol{h}_* + e}.$$

Therefore,

$$\frac{\lambda e}{(1-|\lambda \boldsymbol{x}_*|)^2} = \frac{\lambda e}{(e/(\boldsymbol{h}_*+e))^2} = \frac{\lambda}{e}\,(\boldsymbol{h}_* + e)^2.$$

The claim then follows from Lemma A.4. □

Lemma A.9 (Restatement of Lemma 4.1)**.** *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*

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

Proof of Lemma 4.1. ($\Rightarrow$) 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

$$\alpha_2 = -(s_1+s_2+s_3), \quad \alpha_1 = s_1s_2 + s_1s_3 + s_2s_3, \quad \alpha_0 = -s_1s_2s_3.$$

The polynomial factors into terms $s - r$ with $-r > 0$ for real roots r, and $(s-p)(s-\bar{p}) = s^2 - 2\Re(p)\,s + |p|^2$ with $-2\Re(p) > 0$ for complex-conjugate pairs; all of these have positive coefficients, so $\alpha_2, \alpha_1, \alpha_0 > 0$. Moreover, direct expansion gives

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

If all roots are real, each factor $s_i + s_j$ is negative, so the right-hand side is positive. If $s_1 = p$, $s_2 = \bar{p}$, and $s_3 = r$ is real, then $s_1 + s_2 = 2\Re(p) < 0$ and $(s_1+s_3)(s_2+s_3) = |p+r|^2 > 0$, where $p + r \neq 0$ because $\Re(p+r) < 0$; so the right-hand side is again positive. Hence $\alpha_2\alpha_1 > \alpha_0$.

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

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

so s is not a root. Hence all real roots satisfy $s < 0$. Now let $s = u + iv$ with $v \neq 0$. Substituting and separating real and imaginary parts yields

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

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 $u \geq 0$, the left-hand side is at least $\alpha_2\alpha_1 - \alpha_0 > 0$, a contradiction. Thus $u < 0$, i.e., all complex roots lie in the open left half-plane.

Therefore, the given conditions are necessary and sufficient for Hurwitz stability. □

Proposition A.10 (Restatement of Proposition 4.3)**.** *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*

$$[a(a+\lambda) - b(b+\lambda)|\lambda x_*|]f''(x_*) > -\frac{eC}{(1-|\lambda x_*|)}, \tag{16}$$

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

Proof of Proposition 4.3. Assume $a, b, e > 0$ and $\lambda \geq 0$. Recall from the Routh–Hurwitz criterion (Lemma 4.1) 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 $\boldsymbol{J}(\boldsymbol{z}_*)$,

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

we identify

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

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 $\alpha_2\alpha_1 > \alpha_0$; together with $\alpha_0 > 0$ and $\alpha_2 > 0$, this gives $\alpha_1 > \alpha_0/\alpha_2 > 0$, so the condition $\alpha_1 > 0$ holds automatically. Substituting the definitions and rearranging, one obtains

$$\big[(a+b+\lambda)(a\hbar_* - bh_*) - abe\big]H_* > -\big[(a+b+\lambda)(ab+a\lambda+b\lambda) - ab\lambda\big]\hbar_*^2.$$

Dividing by $\hbar_* > 0$ and using $h_*/\hbar_* = |\lambda x_*|$ and $\hbar_* = e/(1-|\lambda x_*|)$, this inequality is equivalent to

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

where

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

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

Corollary A.11. *Assume $a, b, e > 0$, and $\lambda \geq 0$. Let $[x_*, m_*, h_*]$ be an equilibrium point of AdamW ODE. If $a = b$, or $\lambda = 0$, then $[x_*, m_*, h_*]$ is Hurwitz stable iff $F''(x_*) > 0$.*

Proof. By Lemma 4.1, Hurwitz stability of an equilibrium $\boldsymbol{z}_* = [x_*, m_*, h_*]$ requires $F''(x_*) > 0$ and

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

This can be rearranged as

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

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

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

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 $\hbar_* = e$, $h_* = 0$, 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 implied by $H_* = F''(x_*) > 0$.

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

A.6 The Discrete AdamW System

We consider exact gradients and the autonomous AdamW update without bias correction, as in section 4. The calculation in $h = \sqrt{v}$ coordinates applies at active fixed points, where $h_* = |f'(x_*)| > 0$. We first derive the Jacobian and then express its Schur conditions as curvature bounds that recover the continuous-time analysis.

Definition A.12 (Lyapunov stability for discrete systems)**.** *Consider the autonomous system $\boldsymbol{z}_{t+1} = \boldsymbol{T}(\boldsymbol{z}_t)$, where $\boldsymbol{T} : \mathcal{D} \to \mathbb{R}^{3d}$ with domain $\mathcal{D} \subset \mathbb{R}^{3d}$. Let $\boldsymbol{z}_* \in \mathcal{D}$ be a fixed point, i.e. $\boldsymbol{T}(\boldsymbol{z}_*) = \boldsymbol{z}_*$. Then:*

- *$\boldsymbol{z}_*$ is **Lyapunov stable** if for every $\epsilon > 0$ there exists $\delta > 0$ such that $\|\boldsymbol{z}_0 - \boldsymbol{z}_*\| < \delta$ implies $\|\boldsymbol{z}_t - \boldsymbol{z}_*\| < \epsilon$ for all integers $t \geq 0$.*
- *$\boldsymbol{z}_*$ is **asymptotically stable** if it is Lyapunov stable and there exists $\delta' > 0$ such that $\|\boldsymbol{z}_0 - \boldsymbol{z}_*\| < \delta'$ implies $\lim_{t\to\infty} \boldsymbol{z}_t = \boldsymbol{z}_*$.*
- *$\boldsymbol{z}_*$ is **unstable** if it is not Lyapunov stable; i.e., there exists $\epsilon' > 0$ such that for every $\delta > 0$ one can find $\|\boldsymbol{z}_0 - \boldsymbol{z}_*\| < \delta$ but $\|\boldsymbol{z}_t - \boldsymbol{z}_*\| \geq \epsilon'$ 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.13 (Schur stability)**.** *Let $\boldsymbol{z}_*$ be a fixed point of $\boldsymbol{z}_{t+1} = \boldsymbol{T}(\boldsymbol{z}_t)$ and let $\boldsymbol{J}(\boldsymbol{z}_*) = \left[\partial T_i/\partial z_j\right]_{\boldsymbol{z}_*}$ denote the Jacobian at $\boldsymbol{z}_*$. Then $\boldsymbol{z}_*$ is **Schur stable** if every eigenvalue $s \in \mathbb{C}$ of $\boldsymbol{J}(\boldsymbol{z}_*)$ satisfies $|s| < 1$.*

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

For the Schur-stability calculation, we work with the first-order local representation of the $h = \sqrt{v}$ dynamics around a fixed point:

$$\boldsymbol{z}_{t+1} = \boldsymbol{T}(\boldsymbol{z}_t), \quad \boldsymbol{z} \coloneqq \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{m} \\ \boldsymbol{h} \end{bmatrix}, \quad \boldsymbol{T}(\boldsymbol{z}) \coloneqq \begin{bmatrix} (1-\lambda\eta)\boldsymbol{x} - \eta \frac{\beta_1 \boldsymbol{m} + (1-\beta_1)\nabla f(\boldsymbol{x})}{\beta_2' \boldsymbol{h} + (1-\beta_2')\nabla f(\boldsymbol{x})^{\odot 2}/\boldsymbol{h} + e} \\ \beta_1 \boldsymbol{m} + (1-\beta_1)\nabla f(\boldsymbol{x}) \\ \beta_2' \boldsymbol{h} + (1-\beta_2')\nabla f(\boldsymbol{x})^{\odot 2}/\boldsymbol{h} \end{bmatrix}. \tag{17}$$

where $\beta_2' = \frac{1+\beta_2}{2}$. This representation has the same fixed point and Jacobian as the exact AdamW v-update. Note that the term corresponding to $\boldsymbol{x}_{t+1}$ 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 $\boldsymbol{x}_{t+1}$ in terms of $[\boldsymbol{x}_t, \boldsymbol{m}_t, \boldsymbol{h}_t]$ only for analysis. Nevertheless, we will use $\boldsymbol{m}_{t+1}$ and $\boldsymbol{h}_{t+1}$ to simplify expressions in the following. Differentiating this system $\boldsymbol{J}(\boldsymbol{z}) \coloneqq \partial \boldsymbol{T}(\boldsymbol{z})/\partial \boldsymbol{z}$, we obtain:

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

where

$$\frac{\partial \boldsymbol{x}_{t+1}}{\partial \boldsymbol{x}_t} = (1-\lambda\eta)\boldsymbol{I} - \eta(1-\beta_1)\mathrm{Diag}\left(\frac{1}{\boldsymbol{h}_{t+1}+e}\right)\nabla^2 f(\boldsymbol{x}_t) + \eta(1-\beta_2)\mathrm{Diag}\left(\frac{\boldsymbol{m}_{t+1}\nabla f(\boldsymbol{x}_t)}{\boldsymbol{h}_t(\boldsymbol{h}_{t+1}+e)^2}\right)\nabla^2 f(\boldsymbol{x}_t)$$

$$\frac{\partial \boldsymbol{x}_{t+1}}{\partial \boldsymbol{h}_t} = \eta\mathrm{Diag}\left(\frac{\boldsymbol{m}_{t+1}}{(\boldsymbol{h}_{t+1}+e)^2}\right)\left(\beta_2'\boldsymbol{I} - (1-\beta_2')\mathrm{Diag}\left(\frac{\nabla f(\boldsymbol{x}_t)^{\odot 2}}{\boldsymbol{h}_t^2}\right)\right)$$

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

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

where

$$\frac{\partial \boldsymbol{x}_{t+1}}{\partial \boldsymbol{x}_t} = (1-\lambda\eta)\boldsymbol{I} - \eta(1-\beta_1)\mathrm{Diag}\left(\frac{1}{\boldsymbol{h}_*+e}\right)\nabla^2 f(\boldsymbol{x}_*) + \eta(1-\beta_2)\mathrm{Diag}\left(\frac{\boldsymbol{h}_*}{(\boldsymbol{h}_*+e)^2}\right)\nabla^2 f(\boldsymbol{x}_*)$$

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

$$\chi_{\boldsymbol{J}}(s) = \det\big(s\boldsymbol{I} - \boldsymbol{J}(\boldsymbol{z}_*)\big) = s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0,$$

where

$$\alpha_2 = \frac{\eta H_*}{\hbar_*^2}\Big[(1-\beta_1)\hbar_* + (\beta_2-1)h_*\Big] - (\beta_1+\beta_2+1) + \eta\lambda,$$
$$\alpha_1 = \frac{\eta H_*}{\hbar_*^2}\Big[\beta_1\beta_2(\hbar_*-h_*) + \beta_1 h_* - \beta_2\hbar_*\Big] + (\beta_1\beta_2+\beta_1+\beta_2) - \eta\lambda(\beta_1+\beta_2),$$
$$\alpha_0 = \beta_1\beta_2(\eta\lambda - 1).$$

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

Lemma A.14 (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:*

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

Stability conditions as curvature bounds. We express the momentum parameters through the ODE rates $a = (1-\beta_1)/\eta$ and $b = (1-\beta_2)/\eta$, that is, $\beta_1 = 1-a\eta$ and $\beta_2 = 1-b\eta$. For each step size $\eta > 0$, the map $(\beta_1, \beta_2) \mapsto (a, b)$ is a bijection from $(0,1)^2$ onto $(0, 1/\eta)^2$, so Proposition A.15 below covers every choice of η, β_1, and β_2. To compare different step sizes, we hold a, b fixed and let β_1, β_2 vary with η. This is the setting of the limit $\eta \to 0$, which recovers the AdamW ODE, of the dependence of r_{crit} on η, and of Corollary A.16. At a fixed point, let

$$r = |\lambda x_*|, \qquad H_* = f''(x_*), \qquad \hbar_* = \frac{e}{1-r}, \qquad F''(x_*) = H_* + \frac{\lambda e}{(1-r)^2}.$$

Define

$$\begin{aligned} A_\eta &= a + \lambda(1-a\eta), \qquad B_\eta = b + \lambda(1-b\eta), \\ C_\eta &= (a+b-ab\eta)A_\eta B_\eta, \\ D_\eta &= a(1-b\eta)A_\eta - rb(1-a\eta)B_\eta, \\ E_\eta &= a(2-b\eta) - rb(2-a\eta), \qquad T_\eta = (2-\lambda\eta)(2-a\eta)(2-b\eta). \end{aligned} \tag{18}$$

Each coefficient is a polynomial in η. As $\eta \to 0$, C_η and D_η tend to the constants of the continuous-time condition equation 8, whereas E_η and T_η belong to a condition that only finite steps impose. With these coefficients, the four Jury inequalities reduce to three bounds on the curvature $H_* = f''(x_*)$, an exact discrete counterpart of Proposition 4.3.

Proposition A.15 (Discrete curvature conditions)**.** *Assume $a, b, \lambda, e > 0$, let f be C^2 near an active one-dimensional fixed point, and suppose*

$$0 < \eta < \min\{a^{-1}, b^{-1}, 2\lambda^{-1}\}.$$

The fixed point is Schur stable if and only if

$$F''(x_*) > 0, \tag{19}$$

$$D_\eta H_* > -\frac{eC_\eta}{1-r}, \tag{20}$$

$$\eta^2 E_\eta H_* < \frac{eT_\eta}{1-r}. \tag{21}$$

Thus these conditions ensure local asymptotic stability of the exact nonlinear AdamW update.

Proof. Set $p = \beta_1$, $q = \beta_2$, $d = \eta\lambda$, and $k = \eta H_*/\hbar_*$. The characteristic polynomial above factors as

$$\begin{aligned} P(s) &= (s - 1 + d)(s - p)(s - q) \\ &\quad + ks\big[(1-p)(s-q) - r(1-q)(s-p)\big]. \end{aligned}$$

Its coefficients satisfy

$$\begin{aligned} |\alpha_0| &= pq|1-d| < 1, \\ P(1) &= (1-p)(1-q)\big[d + k(1-r)\big], \\ -P(-1) &= T_\eta - \eta^2 E_\eta H_*/\hbar_*, \\ 1 - \alpha_1 + \alpha_2\alpha_0 - \alpha_0^2 &= \eta^3\big(C_\eta + D_\eta H_*/\hbar_*\big). \end{aligned}$$

The first inequality follows from $p, q \in (0,1)$ and $0 < d < 2$. Since $d + k(1-r) = \eta e F''(x_*)/\hbar_*^2$, positivity of the remaining three expressions gives the three stated inequalities. Lemma A.14 completes the proof. □

What each condition rules out. Each condition in Proposition A.15 is the positivity of one Jury expression, and it fails when an eigenvalue of $\boldsymbol{J}(\boldsymbol{z}_*)$ leaves the unit disk at a specific point of the unit circle. That point determines how a small perturbation of the fixed point grows (Table 1). We name the three bounds accordingly.

- **Implicit-minimum floor (19).** The fixed point must be a strict local minimum of the implicit objective F, exactly as in continuous time (Corollary 4.2). Otherwise a real eigenvalue exceeds $+1$, and perturbations grow without changing sign.
- **Moment-mismatch bound (20).** This is the discrete form of the Hessian constraint equation 8. Its direction depends on the sign of D_η: it is a floor on H_* when $D_\eta > 0$ and a ceiling when $D_\eta < 0$. At equality, with the other Jury inequalities strict, a complex-conjugate pair reaches the unit circle, so perturbations rotate while they grow. This is the oscillatory instability caused by unequal moment timescales, consistent with the persistent oscillations in figure 2.
- **Step-flip ceiling (21).** This bound has no continuous-time counterpart. When $E_\eta > 0$, it caps the curvature at $eT_\eta/(\eta^2(1-r)E_\eta)$. Above this cap, an eigenvalue passes -1: each step overshoots x_*, and the perturbation flips sign from one iteration to the next. This is the analogue of the familiar $\eta < 2/L$ step-size limit of gradient descent.

Recovery of the continuous-time bound. As $\eta \to 0$ with a, b fixed,

$$D_\eta \to a(a+\lambda) - rb(b+\lambda), \qquad C_\eta \to (a+b)(a+\lambda)(b+\lambda) = C,$$

where C is the constant in equation 8. Hence the moment-mismatch bound equation 20 recovers the continuous-time constraint of Proposition 4.3 exactly. The implicit-minimum floor equation 19

Table 1: The three conditions of Proposition A.15. Each fails when an eigenvalue of $\boldsymbol{J}(\boldsymbol{z}_*)$ crosses the unit circle at the listed point.

Condition	Crossing	Perturbation if violated	ODE limit ($\eta \to 0$)
Implicit-minimum floor (19)	real, at $+1$	grows monotonically	$F''(x_*) > 0$
Moment-mismatch bound (20)	complex pair	grows while oscillating	condition (8)
Step-flip ceiling (21)	real, at -1	flips sign every step	vacuous

does not depend on η at all, since neither the fixed-point equation nor F involves η, β_1, or β_2. The step-flip ceiling equation 21 disappears in the limit: its left side is $O(\eta^2)$ while $T_\eta \to 8$, so for any fixed H_* it holds once η is small enough. Finite step sizes therefore leave the implicit-minimum floor unchanged, make the moment-mismatch bound interact with η through C_η and D_η, and add one new obstruction, the step-flip ceiling.

The first two conditions also keep their continuous-time instability modes. Let $\boldsymbol{J}_{\mathrm{ODE}}(\boldsymbol{z}_*)$ denote the ODE Jacobian in (5). Substituting $\beta_1 = 1 - a\eta$ and $\beta_2 = 1 - b\eta$ into the discrete Jacobian at the fixed point gives, entrywise,

$$\boldsymbol{J}(\boldsymbol{z}_*) = \boldsymbol{I} + \eta\, \boldsymbol{J}_{\mathrm{ODE}}(\boldsymbol{z}_*) + O(\eta^2).$$

For instance, the diagonal entries are $1 - \lambda\eta + O(\eta^2)$, $\beta_1 = 1 - a\eta$, and $\beta_2 = 1 - b\eta$, matching $-\lambda$, $-a$, and $-b$ in (5). Each eigenvalue μ of $\boldsymbol{J}_{\mathrm{ODE}}(\boldsymbol{z}_*)$ thus yields an eigenvalue $s = 1 + \eta\mu + O(\eta^2)$ of $\boldsymbol{J}(\boldsymbol{z}_*)$, and $|s| < 1$ corresponds to $\Re\mu < 0$ for small η. The implicit-minimum floor fails through a real eigenvalue at $s = +1$, the counterpart of the ODE eigenvalue $\mu = 0$; both occur exactly when $F''(x_*) = 0$. The moment-mismatch bound fails through a complex pair on the unit circle, the counterpart of the ODE pair crossing the imaginary axis; for small η the two crossings agree up to $O(\eta)$, while at larger η the crossing angle and the bound itself depend on η. The step-flip crossing at $s = -1$ has no counterpart, because it requires $\eta\mu \approx -2$, far from $s \approx 1$.

Curvature windows in the three momentum regimes. Figure 6 plots, for each fixed-point location x_*, the interval of curvatures $f''(x_*)$ for which the fixed point is Schur stable. Unstable curvatures are hatched, and the dashed grey curves show the continuous-time bounds. Since $x_* = 0$ is not an active fixed point, the curves there are limits from $x_* \neq 0$. We take $\eta = 0.07$, larger than the step size $\eta = 0.01$ of the examples in section 4, so that the effects of the discrete steps are visible. We keep the examples' rates a, b, so their momentum parameters differ from those in section 4 (e.g., $\beta_1 = 0.3$ and $\beta_2 = 0.93$ for the quartic). The three panels correspond to the three possible orderings of the momentum rates.

1. **Reversed momenta ($a < b$, i.e., $\beta_1 > \beta_2$); panel (a).** The coefficient D_η is positive for small r and negative for large r. The moment-mismatch bound therefore becomes a ceiling once the fixed point is far enough from the origin:

$$r > r_{\mathrm{crit}}(\eta) \coloneqq \frac{a(1-b\eta)A_\eta}{b(1-a\eta)B_\eta}. \tag{22}$$

   For $r > r_{\mathrm{crit}}(\eta)$, this ceiling lies just above zero (about 0.087 at the quadratic example, against 0.091 for the ODE), so even mildly convex minima are unstable. This includes the quadratic, which has $f'' = 2$. Larger steps make this obstruction slightly more widespread. The threshold $r_{\mathrm{crit}}(\eta)$ decreases from its ODE value $a(a+\lambda)/(b(b+\lambda))$ (here 0.018, the dashed lines in figure 4) to 0.006 at $\eta = 0.07$, so the ceiling applies to more fixed-point locations. In general,

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \log r_{\mathrm{crit}}(\eta) = (b-a)\left[-\frac{1}{(1-a\eta)(1-b\eta)} + \frac{\lambda^2}{A_\eta B_\eta}\right] < 0,$$

   because $A_\eta B_\eta > \lambda^2(1-a\eta)(1-b\eta)$.

2. **Usual momenta ($a > b$, i.e., $\beta_1 < \beta_2$); panel (b).** Both D_η and E_η are positive for $0 \le r < 1$, so stability requires H_* to lie in the window

$$\max\left\{-\frac{\lambda e}{(1-r)^2}, -\frac{eC_\eta}{(1-r)D_\eta}\right\} < H_* < \frac{eT_\eta}{\eta^2(1-r)E_\eta}. \tag{23}$$

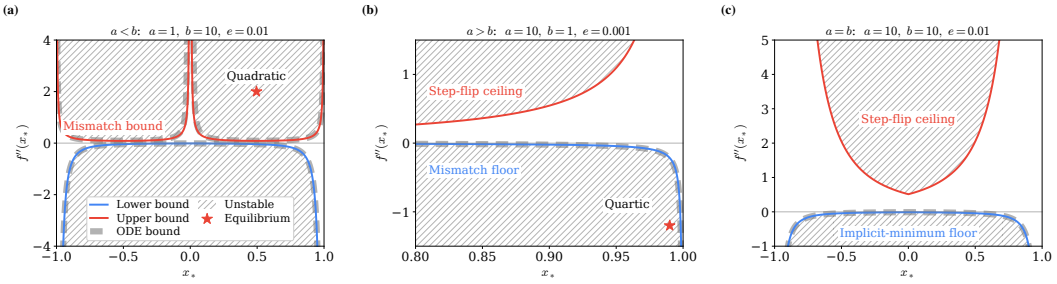


Figure 6: Schur-stability windows (Proposition A.15) with $\lambda = 1$, $\eta = 0.07$, $\beta_1 = 1 - a\eta$, $\beta_2 = 1 - b\eta$. Fixed points are stable between the lower (blue) and upper (red) bounds; hatched regions are unstable, and dashed grey curves are the ODE bounds. Red stars mark the examples of section 4. (a) $a < b$: a moment-mismatch ceiling excludes the convex quadratic. (b) $a > b$: the moment-mismatch floor (blue, below zero) excludes the quartic, whose $f''(x_*) < 0$; for $f'' > 0$ only the step-flip ceiling applies. (c) $a = b$: only the implicit-minimum floor and step-flip ceiling remain.

The lower end is the larger of the implicit-minimum floor and the moment-mismatch floor. It nearly coincides with the ODE floor (-0.215 versus -0.224 at the quartic example), so negative curvature is still excluded even when $F''(x_*) > 0$. This floor excludes the quartic, which has $f''(x_*) = -1.2$. The upper end is the step-flip ceiling, which the ODE does not have. It scales as $e/(\eta^2(1-r))$, so it only constrains large steps or sharp minima. To verify the signs, write $S = a + b - ab\eta$, $p = 1 - a\eta$, and $q = 1 - b\eta$. At $r = 1$, $D_\eta = (a-b)(S + \lambda pq) > 0$ and $E_\eta = 2(a-b) > 0$, and both increase as r decreases.

3. **Equal momenta ($a = b$); panel (c).** The moment-mismatch bound is now implied by the implicit-minimum floor, so only the floor and the step-flip ceiling remain:

$$-\frac{\lambda e}{(1-r)^2} < H_* < \frac{e(2-\lambda\eta)(2-a\eta)}{a\eta^2(1-r)^2}. \tag{24}$$

Indeed, for $p = q$ the polynomial factors as

$$P(s) = (s-p)\big[(s-1+d)(s-p) + k(1-p)(1-r)s\big].$$

The quadratic factor is Schur stable exactly when $d + k(1-r) > 0$ and $(2-d)(1+p) > k(1-p)(1-r)$, because its determinant bound holds automatically in the stated step range. These two inequalities give equation 24. As $\eta \to 0$ the ceiling disappears, and stability reduces to $F''(x_*) > 0$, recovering Corollary A.11 for $a = b$.

In summary, the moment-mismatch obstructions of the continuous-time analysis survive discretization almost unchanged, and finite steps add only the step-flip ceiling on overly sharp minima.

Persistence of the non-convergence examples. In Figure 6, both main-text examples (red stars) lie outside their stability windows. The next result shows that this does not depend on the choice $\eta = 0.07$: with their ODE rates held fixed, both counterexamples remain unstable throughout the admissible step range.

Corollary A.16. *Consider the quadratic and quartic in section 4, with respective settings $(a, b, e, \lambda) = (1, 10, 0.01, 1)$ and $(10, 1, 0.001, 1)$. Hold these ODE rates a, b fixed and let $\beta_1 = 1 - a\eta$ and $\beta_2 = 1 - b\eta$ vary with η. Then both fixed points are unstable for every $0 < \eta < 0.1$.*

Proof. Let $K = H_*/\hbar_*$. In both cases, $C_\eta \le 242$ throughout $0 < \eta < 0.1$. For the quadratic,

$$x_* = \frac{3.01 - \sqrt{3.01^2 - 8}}{4} \in (0.495, 0.496), \qquad K = 200(1 - x_*) > 100.8,$$

and $D_\eta \le 2 - 0.495 \cdot 10 \cdot 0.9 \cdot 10 = -42.55$. Hence $C_\eta + D_\eta K < 242 - 42.55 \cdot 100.8 < 0$. For the quartic as specified in the main text, $x_* = 0.99$, $H_* = -1.2$, and $\hbar_* = 0.1$, so $K = -12$. Moreover, $D_\eta \ge 10 \cdot 0.9 \cdot 10 - 0.99 \cdot 2 = 88.02$, giving $C_\eta + D_\eta K \le 242 - 12 \cdot 88.02 < 0$. Thus the oscillation Jury expression is strictly negative in both cases. Since polynomials with all roots in the closed unit disk satisfy the non-strict Jury inequalities, each Jacobian has an eigenvalue of modulus greater than one. The fixed points are therefore unstable. □

A.7 Multidimensional Case

Lemma A.17 (Restatement of Lemma 5.1)**.** *Let $\boldsymbol{z}_* = (\boldsymbol{x}_*, \boldsymbol{m}_*, \boldsymbol{h}_*)$ be an equilibrium point of the AdamW ODE equation 4. Then the linearized dynamics $\dot{\boldsymbol{z}}_t = \boldsymbol{J}(\boldsymbol{z}_*)\boldsymbol{z}_t$ imply that $\boldsymbol{x}_t$ satisfies*

$$\dddot{\boldsymbol{x}}_t + \boldsymbol{A}_2\ddot{\boldsymbol{x}}_t + \boldsymbol{A}_1\dot{\boldsymbol{x}}_t + \boldsymbol{A}_0\boldsymbol{x}_t = 0, \tag{25}$$

where

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

Proof. Linearize the AdamW ODE equation 4 at the equilibrium $\boldsymbol{z}_* = (\boldsymbol{x}_*, \boldsymbol{m}_*, \boldsymbol{h}_*)$. Let $(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{h})$ denote infinitesimal perturbations (we drop δ notation for readability). Writing $H \coloneqq \nabla^2 f(\boldsymbol{x}_*)$ and

$$U \coloneqq \mathrm{Diag}\Big(\frac{1}{\boldsymbol{h}_*+e}\Big),\ V \coloneqq \mathrm{Diag}\Big(\frac{\boldsymbol{m}_*}{(\boldsymbol{h}_*+e)^2}\Big),\ D \coloneqq \mathrm{Diag}\Big(\frac{\nabla f(\boldsymbol{x}_*)}{\boldsymbol{h}_*}\Big),\ S \coloneqq \mathrm{Diag}\Big(\frac{e}{(\boldsymbol{h}_*+e)^2}\Big),$$

the linearized dynamics are

$$\dot{\boldsymbol{x}} = -\lambda\boldsymbol{x} - U\boldsymbol{m} + V\boldsymbol{h}, \quad \dot{\boldsymbol{m}} = a(H\boldsymbol{x} - \boldsymbol{m}), \quad \dot{\boldsymbol{h}} = b(DH\boldsymbol{x} - \boldsymbol{h}). \tag{26}$$

Differentiate the first equation and substitute equation 26:

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

Using the identity

$$VD = \mathrm{Diag}\Big(\frac{\boldsymbol{m}_*}{(\boldsymbol{h}_*+e)^2}\Big)\mathrm{Diag}\Big(\frac{\nabla f(\boldsymbol{x}_*)}{\boldsymbol{h}_*}\Big) = \mathrm{Diag}\Big(\frac{\boldsymbol{h}_*}{(\boldsymbol{h}_*+e)^2}\Big) = U - S,$$

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

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

Eliminate $U\boldsymbol{m}$ using the first line of equation 26: $U\boldsymbol{m} = -\dot{\boldsymbol{x}} - \lambda\boldsymbol{x} + V\boldsymbol{h}$, and thus

$$\ddot{\boldsymbol{x}} + (\lambda+a)\dot{\boldsymbol{x}} + (a-b)UH\boldsymbol{x} + bSH\boldsymbol{x} + a\lambda\boldsymbol{x} = (a-b)V\boldsymbol{h}. \tag{27}$$

Multiplying equation 27 by b gives

$$b(a-b)V\boldsymbol{h} = b\Big[\ddot{\boldsymbol{x}} + (\lambda+a)\dot{\boldsymbol{x}} + (a-b)UH\boldsymbol{x} + bSH\boldsymbol{x} + a\lambda\boldsymbol{x}\Big]. \tag{28}$$

Differentiate equation 27 and substitute $\dot{\boldsymbol{h}} = b(DH\boldsymbol{x} - \boldsymbol{h})$ from equation 26:

$$\dddot{\boldsymbol{x}} + (\lambda+a)\ddot{\boldsymbol{x}} + \big[(a-b)UH + bSH + a\lambda\boldsymbol{I}\big]\dot{\boldsymbol{x}} - b(a-b)VD\,H\boldsymbol{x} + b(a-b)V\boldsymbol{h} = 0.$$

Using $VD = U - S$, this becomes

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

Substituting equation 28 to eliminate $b(a-b)V\boldsymbol{h}$ gives

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

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

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

which completes the proof. □

Lemma A.18 (First-order reduction)**.** *Let $\boldsymbol{x}_t$ solve the third-order ODE*

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

Define $\boldsymbol{y}_t \coloneqq \dot{\boldsymbol{x}}_t$ and $\boldsymbol{z}_t \coloneqq \dot{\boldsymbol{y}}_t$. Then $(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{z}_t)$ satisfies

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

Proof. Immediate by substitution: $\dot{\boldsymbol{x}} = \boldsymbol{y}$, $\dot{\boldsymbol{y}} = \boldsymbol{z}$, and using $\dddot{\boldsymbol{x}} = \dot{\boldsymbol{z}}$ in the original ODE. □

Lemma A.19 (Derivative identities)**.** *Let $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ evolve according to Lemma A.18. For any symmetric matrices $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{K}$ define*

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

with $\|u\|_{\boldsymbol{M}}^2 \coloneqq u^\top \boldsymbol{M} u$. Then

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

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

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{A}_2\boldsymbol{x} + \boldsymbol{y}) = \boldsymbol{A}_2\boldsymbol{y} + \boldsymbol{z},$$

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

For W_2, note

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

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

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

Lemma A.20 (Lyapunov cancellation)**.** *Let $\boldsymbol{Q}$ be symmetric positive definite and define*

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

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

$$V(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \coloneqq \tfrac{1}{2}\Big(\|\boldsymbol{A}_2\boldsymbol{x} + \boldsymbol{y}\|_{\boldsymbol{P}}^2 + \|\boldsymbol{A}_2\boldsymbol{y} + \boldsymbol{z}\|_{\boldsymbol{Q}}^2 + \|\boldsymbol{y}\|_{\boldsymbol{K}}^2\Big).$$

Then along trajectories of Lemma A.18,

$$\frac{\mathrm{d}}{\mathrm{d}t}V(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{z}_t) = -\tfrac{1}{2}\,\boldsymbol{y}^\top(\boldsymbol{L} + \boldsymbol{L}^\top)\boldsymbol{y}.$$

Proof. Combine Lemma A.19:

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

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

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

□

Lemma A.21 (Lyapunov stability for the third-order system)**.** *Consider* $\dddot{\boldsymbol{x}}_t + \boldsymbol{A}_2\ddot{\boldsymbol{x}}_t + \boldsymbol{A}_1\dot{\boldsymbol{x}}_t + \boldsymbol{A}_0\boldsymbol{x}_t = 0$*. If there exists a symmetric positive definite* $\boldsymbol{Q}$ *such that*

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

then the dynamics are asymptotically stable.

Proof. By Lemma A.18, stability of the third-order system is equivalent to stability of its first-order form. Define $\boldsymbol{P}, \boldsymbol{K}, \boldsymbol{L}$ as in Lemma A.20. Assumptions (1)–(2) ensure $\boldsymbol{P}, \boldsymbol{K} \succ 0$ and $\boldsymbol{L} + \boldsymbol{L}^\top \succ 0$. Thus V in Lemma A.20 is positive definite and radially unbounded.

Along trajectories,

$$\frac{\mathrm{d}}{\mathrm{d}t}V(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{z}_t) = -\tfrac{1}{2}\,\boldsymbol{y}_t^\top(\boldsymbol{L} + \boldsymbol{L}^\top)\boldsymbol{y}_t \le 0,$$

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

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

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

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

a scalar multiple of the identity. □

B GPT-2-Small Experiment Configuration

For the GPT-2-Small experiment in figure 5, we followed the GPT-opt training setup of Amsel et al. [1]. The model is a GPT-2-Small architecture trained from scratch with 12 layers, 12 attention heads, hidden width 768, vocabulary size 50,257, context length 1024, and flash attention. Training uses the FineWeb10B GPT-2-tokenized shards with batch size 512, one epoch, gradient clipping at 1.0, bfloat16 autocast, `torch.compile`, and a flat-then-linear-decay learning rate schedule: the learning rate stays at its base value for the first 40% of optimizer steps and then linearly decays to 10% of the base value. Each optimizer step accumulates 2^{19} tokens, and validation is run every 400 optimizer steps on 2^{19} tokens. A single tuning run, including the validation passes during training, costs approximately 10 H100 hours. For both AdamW and AdamMe, we tuned learning rate over $\{2 \cdot 10^{-4}, 5 \cdot 10^{-4}, 10^{-3}, 2 \cdot 10^{-3}\}$, weight decay over $\{0.005, 0.02, 0.1\}$, and optimizer ϵ over $\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$. The best setting used learning rate $5 \cdot 10^{-4}$ and weight decay 0.1 for both methods, with $\epsilon = 10^{-8}$ for AdamW and $\epsilon = 10^{-5}$ for AdamMe.

Asset credit and licenses. The training pipeline reuses the publicly available GPT-opt codebase associated with Amsel et al. [1],[3] with the polar-branch fork.[4] The training corpus is the 10B-token GPT-2-tokenized subset (`sample/10BT`) of the FineWeb dataset [48],[5] which is released under the Open Data Commons Attribution License (ODC-By) v1.0; we use it in accordance with that license.

C 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, including symbolic and numerical checks of the proposed Lyapunov function for AdamMe during its development. The research questions, theoretical analysis, and main contributions were formulated, derived, and written by the authors.

[3] `https://github.com/gowerrobert/GPT-opt`

[4] `https://github.com/modichirag/GPT-opt/tree/polar`

[5] `https://huggingface.co/datasets/HuggingFaceFW/fineweb`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The two contributions stated in the abstract and Section 1—the implicit fixed-point objective (Theorem 3.1) and the stability/non-convergence analysis (Proposition 4.3 and the examples)—are exactly what the paper establishes.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper analyzes an idealized continuous-time deterministic regime (defers discrete-time to Appendix A.6), and explicitly notes that the non-convergence examples are worst-case constructions that need not imply detrimental behavior in large-scale training (Section 7 and discussion after the examples).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems, propositions, and lemmas state their assumptions explicitly and are numbered and cross-referenced; full proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main results are theoretical and are supported by deterministic low-dimensional simulations; the objectives, optimizer hyperparameters $(\eta, \beta_1, \beta_2, \lambda, e)$, and equilibria are stated in the figure captions and examples. For the GPT-2-Small run, the model, data, schedule, validation protocol, tuning grid, and selected hyperparameters are summarized in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide AdamMe reference code; Appendix B specifies the public GPT-opt/FineWeb10B setup.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All optimizer hyperparameters $(\eta, \beta_1, \beta_2, \lambda, e)$ used in the deterministic figures are listed in the captions and example statements. Appendix B gives the GPT-2-Small model size, dataset, schedule, validation protocol, optimizer grid, and best AdamW/AdamMe hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: The deterministic ODE and iteration trajectories have no sampling variability. The GPT-2-Small experiment is reported as a single run without error bars or significance tests; this is common for LLM pretraining comparisons because repeated runs are computationally expensive, and we do not use this experiment as statistical evidence for the paper's main theoretical claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The deterministic numerical illustrations run in seconds on a standard laptop CPU. The GPT-2-Small experiment follows the GPU training setup described in Appendix B; each reported run uses the stated token budget, batch size, sequence length, mixed precision, validation cadence, and approximately 10 H100 hours per tuning run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is purely theoretical with no human subjects, data collection, or deployed system, and conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational theoretical work on optimizer dynamics with no direct societal application or deployment path.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no data, pretrained models, or other artifacts with misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The GPT-opt codebase and FineWeb-10B dataset used in the GPT-2-Small experiment are credited and their license status is documented in Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release AdamMe reference code; Algorithm 1 documents the optimizer.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper involves no human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for writing polish and auxiliary visualization checks, not as a component of the core methodology; this is disclosed in the appendix ("The Use of Large Language Models").

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.