SIGN-BASED OPTIMIZERS ARE MASSLESS RELATIVISTIC LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sign-based optimizers are like photons: they travel at a constant speed and only change direction. This suggests a connection with special relativity. We derive Signum and Lion from relativistic learning dynamics with a *maximum speed of adaptation*, c, for each parameter. The sign operation emerges as the massless (photonic) limit, while Lion's gradient mixing emerges from *curvature-aware dissipation* that damps motion more strongly along directions of high positive curvature. We further interpret equal-β Adam as Signum with an adaptive mass determined by gradient variance. Combining this Adam-like adaptation with our finite-mass Lion derivation yields *Massive Lion* (M-Lion), which incorporates Lion's curvature-aware dissipation, variance-based mass adaptation, and a tunable finite rest mass. Its updates interpolate between an approximately linear, SGD-like regime and a saturated, sign-like regime. Experiments reveal the complementary roles of these mechanisms. On CIFAR-10 with ResNet-18, finite rest mass improves both sign-based and adaptive variants, bridging their gap with momentum SGD. Curvature-aware dissipation accelerates optimization on heterogeneous quadratic landscapes and improves language modeling. For a 12-layer transformer trained on 3.25 billion SlimPajama tokens, it consistently improves validation perplexity at high momentum across three learning rates. Our results show how physical reasoning explains relationships among existing optimizers, distinguishes the roles of their constituent mechanisms, and guides the design of new ones.

1 INTRODUCTION

Many machine learning tasks can be cast as an optimization problem, $\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta}} V(\boldsymbol{\theta})$, where $V(\cdot)$ is the objective function and $\boldsymbol{\theta}$ denotes adaptable parameters, such as the weights of a neural network. Closed-form solutions are generally unavailable, so we rely on iterative optimization algorithms to find approximate solutions. This motivates the need for designing effective optimizers, as well as theoretical understanding of successful ones informing this design.

A particularly interesting family of optimizers is sign-based. A canonical example is signSGD (Bernstein et al., 2018), which throws away the gradient magnitude and uses only its sign to update the parameters: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \operatorname{sign}(\boldsymbol{g}_t)$, where $\boldsymbol{g}_t$ is a stochastic estimate of the loss gradient $\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}_t)$. Its momentum variant, Signum, instead applies the sign operation to the momentum buffer: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \operatorname{sign}(\boldsymbol{\mu}_t)$, where $\boldsymbol{\mu}_t = \beta_{\text{mom}} \boldsymbol{\mu}_{t-1} + (1 - \beta_{\text{mom}}) \boldsymbol{g}_t$.

Adam (Kingma & Ba, 2015) also has a connection to sign-based optimization. Balles & Hennig (2018) interpreted Adam's update as a signed momentum direction whose magnitude is adjusted according to an estimate of relative gradient variance. More recently, in their search for Adam's "secret sauce," Orvieto & Gower (2025) found that tying the decay rates of its momentum and second-moment buffers, $\beta_{\text{mom}} = \beta_{\text{var}}$ (Adam's $\beta_1 = \beta_2$), preserves near-optimal performance in language modeling while allowing a precise interpretation of its adaptation beyond signed momentum. Namely, equal-β Adam (which we will call Secret Sauce Adam, or SS-Adam) without bias correction and with $\epsilon = 0$, is exactly a *mollified* Signum update, with the mollification determined by an online estimate of gradient variance (Orvieto & Gower, 2025).

Lion (Chen et al., 2023) is a particularly interesting example. Like Signum, it uses a signed update, but applies the sign operation to a mixture of the current gradient and stored momentum:

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

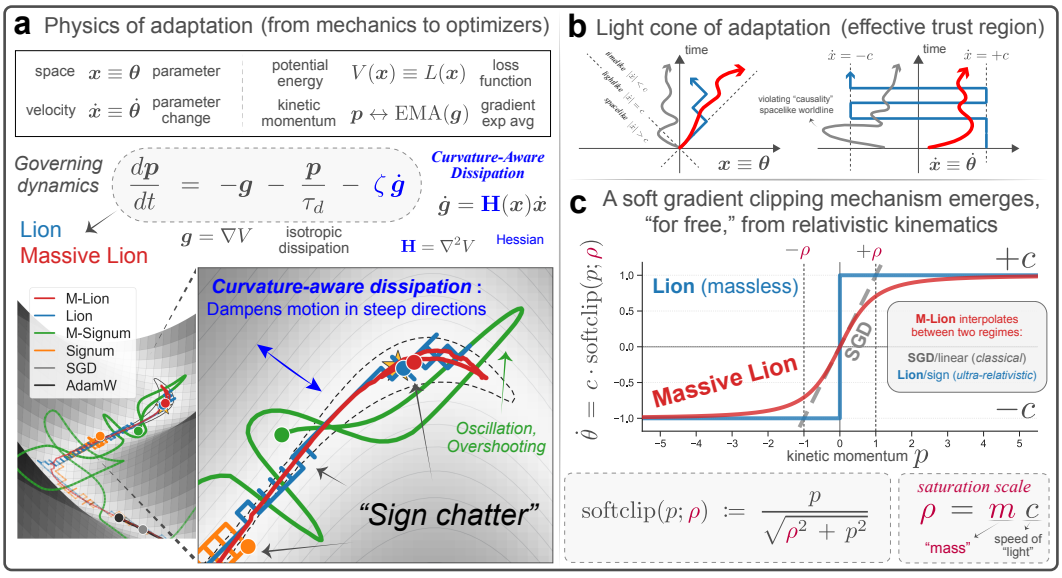

Figure 1: **The physics of adaptation.** We identify neural network parameters $\boldsymbol{\theta}$ with spatial coordinates and the loss with a potential energy, then apply the tools of mechanics to derive learning dynamics. **(a)** Both Lion and M-Lion require a *curvature-aware dissipation* mechanism to damp motion along steep, dangerous directions. As seen in the trajectory inset, sign-based optimizers like Lion suffer from what we call "sign chatter" near minima, a pathology that is eliminated by restoring finite mass. **(b)** In special relativity, particles cannot exceed the speed of light c. We impose the same constraint on parameter velocity $\dot{\boldsymbol{\theta}}$, which acts as an effective trust region. **(c)** Inverting the relativistic momentum yields the exact learner velocity $\dot{\boldsymbol{\theta}} = c \cdot \mathrm{softclip}(\boldsymbol{p}; \rho)$. This soft-clips at $\pm c$ for large momentum (recovering the massless Lion limit) and reduces to a linear response for small momentum (the SGD limit), allowing M-Lion to smoothly interpolate between the two regimes.

$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\,\mathrm{sign}(\boldsymbol{d}_t)$, where $\boldsymbol{d}_t = \beta_{\mathrm{mix}}\boldsymbol{\mu}_{t-1} + (1-\beta_{\mathrm{mix}})\boldsymbol{g}_t$. This mixture has the same form as the momentum update, but uses a separate coefficient, β_{mix}, and reduces to Signum when $\beta_{\mathrm{mix}} = \beta_{\mathrm{mom}}$. The default choice $\beta_{\mathrm{mix}} < \beta_{\mathrm{mom}}$ gives the current gradient more influence over the update while preserving a longer history in the momentum buffer. A theoretical account of Lion was developed by Chen et al. (2024), who introduced the generalized Lion-$\mathcal{K}$ family and used Lyapunov analysis to characterize its convergence and connect weight decay to implicit constraints and regularization. Their analysis also relates gradient mixing to Hessian-dependent damping (Alvarez et al., 2002; Attouch et al., 2016; Shi et al., 2022). However, as Chen et al. (2024) themselves note, a constructive physical derivation of these dynamics remains open. The Lion-$\mathcal{K}$ family also does not establish an adaptive counterpart of Lion analogous to the relationship between Signum and SS-Adam, leaving open how Lion's mixed update direction can be combined with the same variance-adaptation mechanism.

Here, we address these questions through a physics-inspired approach, building on the longstanding connections between optimization and mechanics (Polyak, 1964; Nesterov, 1983; Wibisono et al., 2016; Su et al., 2016; França et al., 2020; 2021). Our starting point identifies parameter space with physical space and the loss with a potential energy (Fig. 1a; Guo & Schölkopf, 2025). Building on this analogy, we formulate learning dynamics using *action principles* (Landau & Lifshitz, 1975), *dissipation* (Rayleigh, 1877), and *relativistic kinematics* (Fig. 1b; Einstein, 1905).

The relativistic response naturally produces soft clipping of the update direction and bounds each parameter update (Fig. 1c; França et al., 2020), connecting this construction to the broader use of clipping for stabilizing optimization (Pascanu et al., 2013). In this physical analogy, we associate each parameter with a mass m which controls how the parameter evolves over time by controlling the steepness of its softclipping function (Fig. 1c).

Our derivation recovers Signum and Lion as massless limits of relativistic learners and gives a mechanical interpretation of SS-Adam's variance adaptation: it is Signum with an adaptive mass. It also derives Lion's gradient injection from a curvature-aware dissipation term that damps motion more strongly along directions of high positive curvature. Building on these insights, we introduce

Algorithm 1 *Massive Lion* (M-Lion).

Require: Loss function $V(\boldsymbol{\theta})$, learning rate η, weight decay λ, rest mass $\rho_0 \geq 0$, momentum decay $\beta_{\text{mom}} \in [0, 1)$, gradient mixing coefficient $\beta_{\text{mix}} \leq \beta_{\text{mom}}$, iterations T

initialize $\boldsymbol{\theta}_0$, $\boldsymbol{\mu}_0 = \boldsymbol{a}_0 = \mathbf{0}$
for $t = 1, \ldots, T$ **do**
  $\boldsymbol{g}_t \leftarrow \nabla V(\boldsymbol{\theta}_{t-1})$
  $\boldsymbol{d}_t \leftarrow \beta_{\text{mix}} \boldsymbol{\mu}_{t-1} + (1 - \beta_{\text{mix}}) \boldsymbol{g}_t$
  if `adaptive_mass` **then**
    $\boldsymbol{e}_t \leftarrow \boldsymbol{g}_t - \boldsymbol{\mu}_{t-1}$
    $\boldsymbol{a}_t^2 \leftarrow \beta_{\text{mom}} \boldsymbol{a}_{t-1}^2 + \beta_{\text{mom}}(1 - \beta_{\text{mom}}) \boldsymbol{e}_t^2$
    $\rho_t \leftarrow \sqrt{\rho_0^2 + \boldsymbol{a}_t^2}$
  else
    $\rho_t \leftarrow \rho_0$
  end if
  $\boldsymbol{u}_t \leftarrow \dfrac{\boldsymbol{d}_t}{\sqrt{\boldsymbol{d}_t^2 + \rho_t^2}}$
  $\boldsymbol{\theta}_t \leftarrow (1 - \eta\lambda)\boldsymbol{\theta}_{t-1} - \eta \boldsymbol{u}_t$
  $\boldsymbol{\mu}_t \leftarrow \beta_{\text{mom}} \boldsymbol{\mu}_{t-1} + (1 - \beta_{\text{mom}}) \boldsymbol{g}_t$
end for
return $\boldsymbol{\theta}_T$

Table 1: The M-Lion family: geometry, rest mass, adaptive mass, and curvature-aware dissipation.

Optimizer	Kinematic geometry	Rest mass	Ada. mass	Curv. diss.	Source
M-Lion	Minkowski	≥ 0	✓	✓	Ours
Curv.-aware SS AdamW	Minkowski	0	✓	✓	Ours
Lion	Minkowski	0	✗	✓	Chen et al. (2023)
M-Signum	Minkowski	≥ 0	✓	✗	Ours
Secret Sauce (SS) AdamW	Minkowski	0	✓	✗	Orvieto & Gower (2025) Balles & Hennig (2018)
Signum	Minkowski	0	✗	✗	Bernstein et al. (2018)
Curv.-aware SGD (QHM)	Euclidean	—	—	✓	Ma & Yarats (2019)
SGD	Euclidean	—	—	✗	Polyak (1964)

Adaptive mass: ✓: on, ✗: off.
Curvature-aware dissipation: ✓ : $\beta_{\text{mix}} < \beta_{\text{mom}}$, ✗ : $\beta_{\text{mix}} = \beta_{\text{mom}}$.
Fixed-mass variants of M-Lion and M-Signum set `adaptive_mass` to `false`.

Massive Lion (M-Lion, Alg. 1), which allows a finite rest mass and combines Lion's gradient injection with SS-Adam's adaptive mass. Its relativistic response interpolates between approximately linear, SGD-like updates in the classical regime and sign updates in the saturated, ultra-relativistic regime. Table 1 shows how M-Lion connects to existing optimizers in the literature.

Summary of contributions and paper outline. We develop and test this framework as follows:

- In section 2, we introduce the physics-based approach to optimization and derive familiar algorithms, including Polyak momentum, from dissipative mechanical dynamics. We then motivate curvature-aware dissipation (section 2.1) and relativistic kinematics (section 2.2) as extensions.
- In section 3, we derive M-Lion's continuous-time dynamics from combining relativistic kinematics with both isotropic and curvature-aware dissipation. We then incorporate variance-based mass adaptation and establish the connections to Lion, Signum, and SS-Adam.
- In section 4, we report empirical results on vision and language modeling tasks. On CIFAR-10 with ResNet-18, a finite rest mass substantially improves performance over SS-AdamW and bridges its gap with momentum SGD. In language modeling, a small amount of curvature-aware dissipation improves on SS-AdamW, especially in the high-momentum settings.
- In section 5, we discuss the distinct roles of mass and dissipation, and how extensions inspired by general relativity could guide the design of further adaptive optimizers.

2 Background, Preliminaries, and Related Work

Notation. Throughout, $\boldsymbol{\theta} \in \mathbb{R}^d$ denotes the adaptable parameters and $V : \mathbb{R}^d \to \mathbb{R}$ is the loss. We write $\boldsymbol{g}(t) = \nabla V(\boldsymbol{\theta}(t))$ for the continuous-time loss gradient, $\mathbf{H}(\boldsymbol{\theta}) = \nabla^2 V(\boldsymbol{\theta})$ for the Hessian, and $\boldsymbol{\mu}$ for the gradient EMA state. In discrete time, $\boldsymbol{g}_t$ may be a stochastic estimate of the loss gradient. We use m to denote physical rest mass and $\rho := mc$ for the momentum saturation scale, where c is the *maximum allowed speed of adaptation* along each coordinate. This plays the role of the speed of light in the physical analogy and sets the maximum update scale after discretization.

Principle of stationary action and Euler–Lagrange equations. In section A, we introduce the principle of stationary action and use it to derive the Euler–Lagrange equations (eq. (22)).

Nonconservative forces, including dissipation, enter through the generalized force $\mathbf{Q}$:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\boldsymbol{\theta}}}\right) - \frac{\partial L}{\partial \boldsymbol{\theta}} \;=\; \mathbf{Q}\,. \tag{1}$$

Here, L is the *Lagrangian*. For a classical particle, it is kinetic minus potential energy: $L = T - V$. We treat the parameter vector $\boldsymbol{\theta}(t)$ as the position of a fictitious particle moving in the potential $V(\boldsymbol{\theta})$.

Defining the canonical momentum $\boldsymbol{p} \coloneqq \partial L/\partial \dot{\boldsymbol{\theta}}$, the generalized Euler-Lagrange equation is:

$$\boxed{\frac{d}{dt}\boldsymbol{p} \;-\; \frac{\partial L}{\partial \boldsymbol{\theta}} \;=\; \mathbf{Q}\,.} \tag{2}$$

Here, the loss V fixes the potential term of $L = T - V$, leaving two choices: the kinetic part of the Lagrangian T, and the dissipation force $\mathbf{Q}$. Different choices lead to different learning dynamics and, after discretization, different optimization algorithms, as we demonstrate below.

Isotropic dissipation yields momentum as a gradient EMA. Consider the classical kinetic energy $T = \frac{1}{2}m\|\dot{\boldsymbol{\theta}}\|^2$, with constant $m > 0$. The canonical momentum is then $\boldsymbol{p} = \partial L/\partial \dot{\boldsymbol{\theta}} = m\dot{\boldsymbol{\theta}}$. We choose an isotropic friction force proportional to momentum, $\mathbf{Q}_{\text{isotropic}} \coloneqq -\boldsymbol{p}/\tau_d$, where $\tau_d > 0$ is the dissipation timescale. Substituting into eq. (1) gives:

$$\dot{\boldsymbol{p}} \;=\; -\boldsymbol{p}/\tau_d \;-\; \boldsymbol{g}\,. \tag{3}$$

For a given gradient history, this is a first-order linear ODE in $\boldsymbol{p}$. Starting from $\boldsymbol{p}(0) = 0$, applying the integrating factor e^{t/τ_d} gives:

$$\boldsymbol{\mu}(t) \;\coloneqq\; -\frac{\boldsymbol{p}(t)}{\tau_d} \;=\; \frac{1}{\tau_d}\int_0^t e^{-(t-s)/\tau_d}\,\boldsymbol{g}(s)\,ds\,. \tag{4}$$

Thus, up to a sign and scale, physical momentum is an exponentially weighted gradient history, or a temporal low-pass filter.

To transition to a discrete time algorithm that can be implemented, we hold the gradient fixed at $\boldsymbol{g}_t$ during each step of duration Δt. Integrating the momentum equation exactly under this frozen-gradient approximation gives the standard EMA:

$$\boldsymbol{\mu}_t \;=\; \beta_{\text{mom}}\boldsymbol{\mu}_{t-1} + (1-\beta_{\text{mom}})\boldsymbol{g}_t\,, \qquad \beta_{\text{mom}} \coloneqq e^{-\Delta t/\tau_d}\,. \tag{5}$$

Using the updated momentum to take a position step gives $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\boldsymbol{\mu}_t$, where $\eta = \tau_d\Delta t/m$. Together, these equations recover heavy-ball momentum in the EMA convention (Polyak, 1964). The coefficient β_{mom} controls momentum retention: it corresponds to β_1 in Adam (Kingma & Ba, 2015) and β_2 in Lion (Chen et al., 2023).

2.1 Motivating curvature-aware dissipation

The isotropic force above is one possible choice of dissipation. More generally, $\mathbf{Q} = -\boldsymbol{\Gamma}\dot{\boldsymbol{\theta}}$, where the friction tensor $\boldsymbol{\Gamma}$ determines how strongly motion is damped in each direction. Here, we rely on two physical considerations to motivate its form (see section A.1 for the full derivations).

Mechanically, motion near a quadratic minimum decomposes into oscillators along Hessian eigenvectors. Critical damping requires friction $2\sqrt{m\lambda_i}$ in a direction with curvature λ_i: sharper directions need stronger damping. Isotropic friction therefore faces a tradeoff between suppressing oscillations in sharp directions and slowing relaxation in flatter ones. A simple approximation is $\boldsymbol{\Gamma} = \gamma_0\mathbf{I} + \zeta\mathbf{H}$, with $\gamma_0 > 0$ and $\zeta > 0$. Its friction coefficients $\gamma_i = \gamma_0 + \zeta\lambda_i$ increase with curvature, selectively damping sharp directions while retaining a baseline in flat ones.

Thermodynamically, damping must also counteract fluctuations from stochastic gradients. Under a local model with gradient-noise covariance proportional to the Hessian, isotropic friction leaves equal parameter variance in every direction. Yet an unlucky displacement in a sharp direction causes a much larger loss increase. The curvature correction suppresses fluctuations precisely in these costly directions: the stationary mean excess loss per mode becomes $D\lambda_i/[2(\gamma_0 + \zeta\lambda_i)]$, where D sets the

noise intensity. This saturates at $D/(2\zeta)$ as curvature increases, instead of growing without bound. Within this local model, stronger damping in sharp directions thus limits noise-induced heating and helps suppress large loss excursions. Setting $\gamma_0 = m/\tau_d$ gives:

$$\boxed{\mathbf{Q} \;=\; -\left(\frac{m}{\tau_d}\mathbf{I} + \zeta\mathbf{H}\right)\dot{\boldsymbol{\theta}} \;=\; -\frac{\boldsymbol{p}}{\tau_d} \;-\; \zeta\dot{\boldsymbol{g}}\,.} \tag{6}$$

Here $\mathbf{H}\dot{\boldsymbol{\theta}} = \dot{\boldsymbol{g}}$ along the smooth loss trajectory. This identity also makes the correction analytically tractable: as we will see below, integration by parts yields the gradient-history representation underlying Lion's gradient mixing.

Curvature-aware dissipation in the literature. Hessian-driven damping has a substantial history in optimization. Alvarez et al. (2002) studied it in inertial dynamics related to regularized continuous Newton methods, and Attouch et al. (2016) established fast convergence properties for dynamics combining isotropic and Hessian-driven damping. The same Hessian-velocity term also emerges as a gradient correction in high-resolution ODE models of Nesterov's accelerated methods (Shi et al., 2022), and its connection to Lion was identified by Chen et al. (2024).

Recall that the generalized Euler–Lagrange equation (eq. (1)) leaves two modeling choices: the dissipative force $\mathbf{Q}$ and the kinetic Lagrangian T. With the former specified by eq. (31), we now turn to the latter. We show that a relativistic kinetic Lagrangian bounds the speed of parameter motion, thereby enforcing an effective trust region.

2.2 Relativistic kinematics and the speed of adaptation

To derive M-Lion, we need one more ingredient that we borrow from special relativity: a finite maximum speed c. In physics, no particle, however energetic, exceeds it. We postulate the same for parameter dynamics on a per-coordinate basis: $|\dot{\theta}_i| \le c$, interpretable as a coordinate-wise *trust region* that no single step may cross. To enforce this saturation independently per coordinate, we use a separable relativistic Lagrangian $L = -mc^2 \sum_i \sqrt{1 - \dot{\theta}_i^2/c^2} - V(\boldsymbol{\theta})$. The canonical momentum is $p_i = \partial L/\partial\dot{\theta}_i = m\gamma_i\dot{\theta}_i$ with a coordinate-specific *Lorentz factor*, $\gamma_i := (1 - \dot{\theta}_i^2/c^2)^{-1/2}$. Inverting this relation yields the exact velocity (section B):

$$\boxed{\dot{\theta}_i \;=\; c \cdot \mathrm{softclip}(p_i;\rho)\,, \qquad \mathrm{softclip}(p_i;\rho) \;:=\; \frac{p_i}{\sqrt{\rho^2 + p_i^2}}\,, \qquad \rho \;:=\; mc\,.} \tag{7}$$

For $|p_i| \ll \rho$ the response is linear in p_i; for $|p_i| \gg \rho$ it saturates at $\pm c$. As $m \to 0$ at fixed p_i, the relation reduces to $\dot{\theta}_i = c \cdot \mathrm{sign}(p_i)$, approaching the "photonic" limit.

This connection between special relativity and optimization has been studied before (França et al., 2020; Vaidhyanathan et al., 2025). Here, we combine this relativistic response with curvature-aware dissipation and variance-based mass adaptation to connect Lion (Chen et al., 2023), Signum (Bernstein et al., 2018), and SS-Adam (Orvieto & Gower, 2025) within a common construction.

3 Massive Lion from relativistic kinematics and curvature-aware dissipation

We now combine the relativistic kinematics of section 2.2 with the dissipation law motivated in section 2.1. We first derive M-Lion with a fixed rest mass, then introduce SS-Adam's variance-based mass adaptation in discrete time.

Continuous-time dynamics with fixed rest mass. Let $m > 0$ be constant. We retain momentum-proportional friction and the curvature-aware force, giving

$$\boxed{\dot{\boldsymbol{p}}(t) \;=\; -\boldsymbol{g}(t) \;-\; \boldsymbol{p}(t)/\tau_d \;-\; \zeta\,\dot{\boldsymbol{g}}(t)\,,} \qquad 0 \le \zeta < \tau_d\,. \tag{8}$$

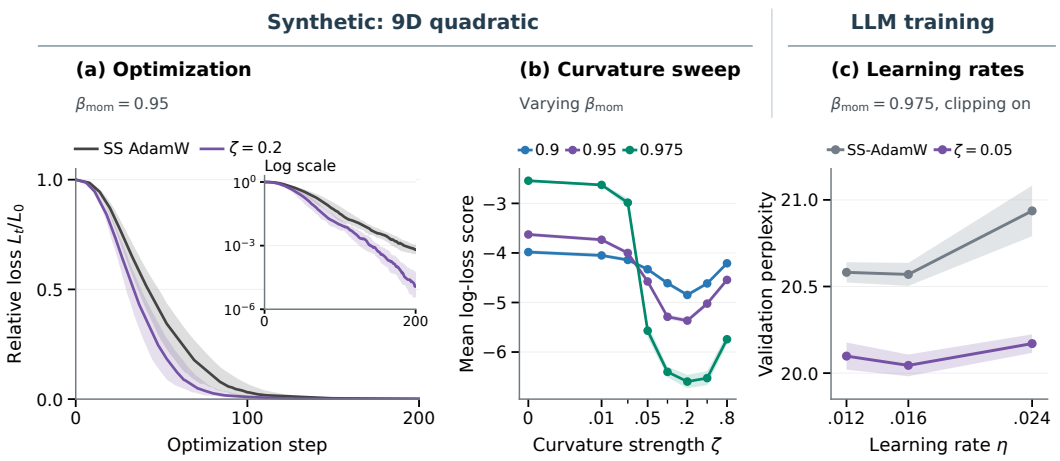

Figure 2: **Curvature-aware dissipation improves optimization and LLM training. (a–b)** Heterogeneous 9D quadratic (Zhang et al., 2024): faster convergence and lower loss with positive dissipation. Shading shows interquartile ranges (a) and pointwise 95% confidence intervals (b) over 128 confirmation seeds. **(c)** LLM validation perplexity at $\beta_{\text{mom}} = 0.975$ with clipping: seven-seed means and pointwise 95% Student-t bands.

Together with the velocity–momentum relation in eq. (7), this defines the continuous-time dynamics. For a time-independent loss, $\dot{\boldsymbol{g}} = \mathbf{H}(\boldsymbol{\theta})\dot{\boldsymbol{\theta}}$, so the curvature correction damps motion along positive-curvature directions and contributes an energy-injecting force along negative-curvature directions.

Theorem 3.1 (Deriving M-Lion with fixed rest mass)**.** *Let $V \in C^2(\mathbb{R}^d)$, with constant $m > 0$, $c > 0$, $\tau_d > 0$, and $0 \le \zeta < \tau_d$. Consider eq. (8) with the relativistic velocity law in eq. (7). Define the continuous gradient EMA by*

$$\boldsymbol{\mu}(t) \;=\; e^{-t/\tau_d}\boldsymbol{\mu}(0) + \frac{1}{\tau_d}\int_0^t e^{-(t-s)/\tau_d}\boldsymbol{g}(s)\,ds\,. \tag{9}$$

With the consistent initialization $\boldsymbol{p}(0) = -\zeta\boldsymbol{g}(0) - (\tau_d - \zeta)\boldsymbol{\mu}(0)$, the canonical momentum admits the representation

$$-\frac{\boldsymbol{p}(t)}{\tau_d} \;=\; \frac{\zeta}{\tau_d}\boldsymbol{g}(t) + \left(1 - \frac{\zeta}{\tau_d}\right)\boldsymbol{\mu}(t)\,. \tag{10}$$

Freeze the gradient at $\boldsymbol{g}_t$ during each momentum step of duration Δt, integrate the EMA exactly under this approximation, and update position using the resulting velocity. Define:

$$\beta_{mom} := e^{-\Delta t/\tau_d}\,, \qquad \beta_{mix} := \beta_{mom}\left(1 - \frac{\zeta}{\tau_d}\right)\,, \qquad \eta := c\Delta t\,, \qquad \rho_0 := \frac{\rho}{\tau_d} = \frac{mc}{\tau_d}\,. \tag{11}$$

The resulting discrete optimizer is

Update direction (12)	*Parameter update (13)*	*Momentum EMA (14)*
$\boldsymbol{d}_t = \beta_{mix}\boldsymbol{\mu}_{t-1} + (1-\beta_{mix})\boldsymbol{g}_t$	$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\dfrac{\boldsymbol{d}_t}{\sqrt{\rho_0^2 + \boldsymbol{d}_t^2}}$	$\boldsymbol{\mu}_t = \beta_{mom}\boldsymbol{\mu}_{t-1} + (1-\beta_{mom})\boldsymbol{g}_t$

All operations in the position update are coordinatewise. This is M-Lion with fixed rest mass. In units where $\tau_d = 1$, the fixed saturation scale is $\rho_0 = \rho = mc$.

The parameter mapping gives $\zeta > 0 \iff \beta_{mix} < \beta_{mom}$. As $\rho_0 \to 0$, the position update reduces to $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\,\mathrm{sign}(\boldsymbol{d}_t)$, recovering Lion without weight decay (Chen et al., 2023). Setting $\beta_{mix} = \beta_{mom}$ gives $\boldsymbol{d}_t = \boldsymbol{\mu}_t$ and recovers fixed-mass M-Signum. Taking its massless limit gives Signum (Bernstein et al., 2018). At zero direction, the limiting update is defined to be zero.

Proof sketch. Multiplying eq. (8) by the integrating factor e^{t/τ_d} and integrating the $\dot{\boldsymbol{g}}$ term by parts gives: $\boldsymbol{p}(t) = -\zeta\boldsymbol{g}(t) - (\tau_d - \zeta)\boldsymbol{\mu}(t) + e^{-t/\tau_d}\left[\boldsymbol{p}(0) + \zeta\boldsymbol{g}(0) + (\tau_d - \zeta)\boldsymbol{\mu}(0)\right]$. The initialization

cancels the final term, yielding eq. (10). The instantaneous-gradient contribution comes from the integration-by-parts boundary term, and the remaining integral forms the EMA. The full proof is given in section C.

Under the frozen-gradient approximation, the EMA satisfies eq. (14). Substituting it into eq. (10) gives

$$-\frac{\boldsymbol{p}_t}{\tau_d} = \beta_{\text{mom}}\left(1-\frac{\zeta}{\tau_d}\right)\boldsymbol{\mu}_{t-1} + \left[1-\beta_{\text{mom}}\left(1-\frac{\zeta}{\tau_d}\right)\right]\boldsymbol{g}_t = \boldsymbol{d}_t\,.$$

Finally, inserting $\boldsymbol{p}_t = -\tau_d \boldsymbol{d}_t$ into the relativistic velocity law and taking a position step of duration Δt yields eq. (13). The stated reductions follow directly. □

Adding variance-based mass adaptation. The theorem derives M-Lion with a fixed rest mass. We now extend the discrete optimizer using the variance-adaptation mechanism of SS-Adam (Orvieto & Gower, 2025). The connection is direct: SS-Adam applies the same soft-clipping response to Signum's momentum, with a saturation scale determined by an online estimate of gradient variance. We borrow this adaptation rule and apply it to Lion's mixed direction $\boldsymbol{d}_t$.

Starting from $\boldsymbol{\mu}_0 = \boldsymbol{a}_0 = \boldsymbol{0}$, define:

Momentum innovation (15)	*Adaptive mass* (16)	*Total mass scale* (17)
$\boldsymbol{e}_t \coloneqq \boldsymbol{g}_t - \boldsymbol{\mu}_{t-1}$	$\boldsymbol{a}_t^2 = \beta_{\text{mom}}\boldsymbol{a}_{t-1}^2 + \beta_{\text{mom}}(1-\beta_{\text{mom}})\boldsymbol{e}_t^2$	$\rho_{t,i}^2 \coloneqq \rho_0^2 + a_{t,i}^2$

Here, $\boldsymbol{a}_t^2$ is the adaptive contribution to the squared saturation scale, while ρ_0 retains the fixed rest-mass contribution. The variance estimate uses the same decay rate as the gradient EMA.

Replacing the fixed scale in eq. (13) by this coordinatewise adaptive scale, and optionally including decoupled weight decay $\lambda \geq 0$, gives the final M-Lion update:

$$\boxed{\boldsymbol{\theta}_{t+1} = (1-\eta\lambda)\boldsymbol{\theta}_t - \eta\,\frac{\boldsymbol{d}_t}{\sqrt{\boldsymbol{d}_t^2 + \rho_0^2 + \boldsymbol{a}_t^2}}\,.} \tag{18}$$

The direction and momentum updates remain eqs. (12) and (14). All operations are coordinatewise, with $0/0 \coloneqq 0$. Mass adaptation enters only at this discrete stage; the continuous-time derivation retains constant m. Disabling adaptation by setting $\boldsymbol{a}_t = \boldsymbol{0}$ recovers the fixed-mass optimizer, with optional weight decay.

Proposition 3.2 (Adaptive mass and the SS-AdamW reduction)**.** *For any gradient sequence, let $\boldsymbol{\mu}_t$ and $\boldsymbol{a}_t^2$ follow eqs. (14) and (16), with zero initial states. Define the second-moment EMA*

$$\boldsymbol{v}_t = \beta_{mom}\boldsymbol{v}_{t-1} + (1-\beta_{mom})\boldsymbol{g}_t^2\,, \qquad \boldsymbol{v}_0 = \boldsymbol{0}\,. \tag{19}$$

Then, for each coordinate, we have: $\boldsymbol{v}_t = \boldsymbol{\mu}_t^2 + \boldsymbol{a}_t^2$. Consequently, when $\beta_{mix} = \beta_{mom}$, the adaptive M-Lion update becomes

$$\boldsymbol{\theta}_{t+1} = (1-\eta\lambda)\boldsymbol{\theta}_t - \eta\,\frac{\boldsymbol{\mu}_t}{\sqrt{\rho_0^2 + \boldsymbol{v}_t}}\,. \tag{20}$$

With $\rho_0 = 0$, this is exactly SS-AdamW: AdamW with equal momentum and second-moment decay rates, without bias correction and with $\epsilon = 0$. Setting $\lambda = 0$ gives the corresponding SS-Adam update (Orvieto & Gower, 2025).

Proof. Writing $\beta = \beta_{\text{mom}}$, expansion of the two recurrences gives

$$\begin{aligned}\boldsymbol{\mu}_t^2 + \boldsymbol{a}_t^2 &= \left[\beta\boldsymbol{\mu}_{t-1} + (1-\beta)\boldsymbol{g}_t\right]^2 + \beta\boldsymbol{a}_{t-1}^2 + \beta(1-\beta)(\boldsymbol{g}_t - \boldsymbol{\mu}_{t-1})^2 \\ &= \beta(\boldsymbol{\mu}_{t-1}^2 + \boldsymbol{a}_{t-1}^2) + (1-\beta)\boldsymbol{g}_t^2\,.\end{aligned}$$

This is the recurrence for $\boldsymbol{v}_t$, with matching initialization. When $\beta_{\text{mix}} = \beta_{\text{mom}}$, we also have $\boldsymbol{d}_t = \boldsymbol{\mu}_t$, giving the stated update. □

Thus, the same variance-based mass adaptation that connects Signum to SS-AdamW can be combined with Lion's gradient mixing. The resulting M-Lion family retains the finite rest-mass scale derived above while allowing its effective saturation scale to adapt to gradient statistics.

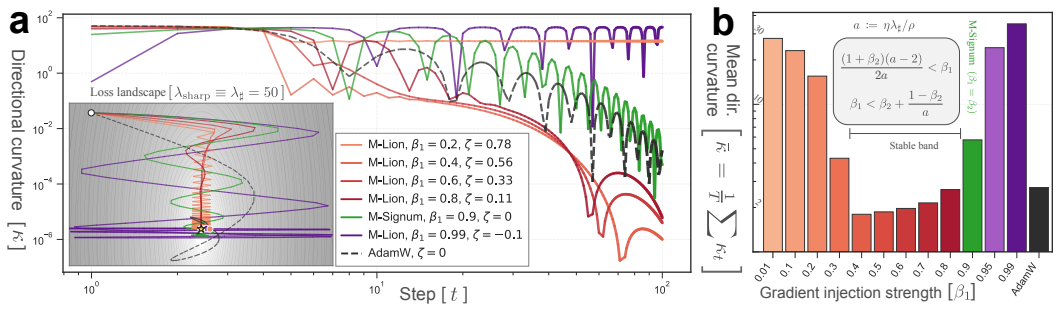

Figure 3: **Local stability band of fixed-mass M-Lion.** Exact-gradient dynamics on $V(\theta_1, \theta_2) = \frac{1}{2}(50\theta_1^2 + \theta_2^2)$, with $\boldsymbol{\theta}_0 = (-1, 3)$, zero initial momentum, $\eta = 0.3$, $\rho_0 = 5$, $\beta_{\text{mom}} = 0.9$, and $T = 100$ steps. Adaptive mass and weight decay are disabled. **(a)** Directional curvature $\kappa_t = \boldsymbol{v}_t^\top \mathbf{H} \boldsymbol{v}_t$ for selected mixing coefficients, where $\boldsymbol{v}_t = (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)/\eta$; the inset shows parameter trajectories. **(b)** Trajectory-averaged directional curvature across the mixing sweep. For the sharp mode, $s_{\max} = 3$ and theorem D.1 gives $19/60 < \beta_{\text{mix}} < 14/15$, approximately $0.317 < \beta_{\text{mix}} < 0.933$. Sampled settings outside this band exhibit sustained oscillations over the plotted interval; intermediate settings approach the minimum. Plot labels use $\beta_1 = \beta_{\text{mix}}$, $\beta_2 = \beta_{\text{mom}}$, and the dimensionless curvature strength $\zeta/\tau_d = 1 - \beta_{\text{mix}}/\beta_{\text{mom}}$ (labeled ζ). The setting $\beta_{\text{mix}} = \beta_{\text{mom}}$ is M-Signum. AdamW uses decay rates $(0.9, 0.999)$ and the same learning rate.

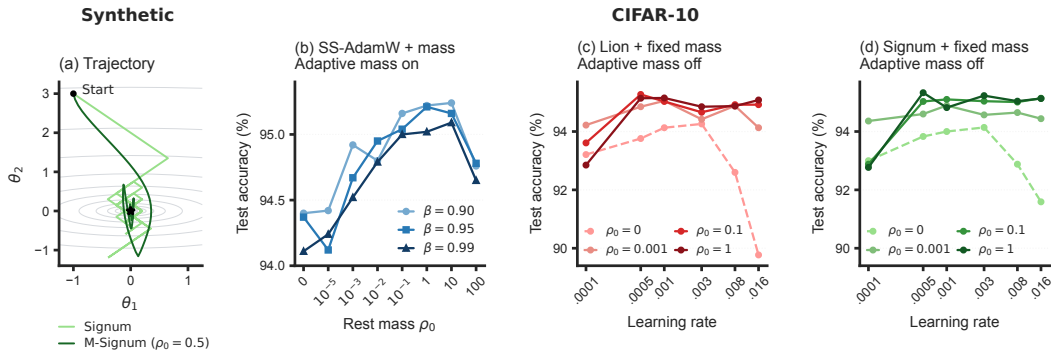

Figure 4: **Rest mass improves CIFAR-10 performance with and without adaptive mass. (a)** On a quadratic, positive rest mass lets Signum's updates shrink and converge. **(b)** CIFAR-10 test accuracy versus rest mass with adaptive mass enabled and tied betas; $\rho_0 = 0$ gives SS-AdamW. **(c–d)** Lion and Signum with fixed mass across learning rates. Each point selects the remaining hyperparameters and checkpoint by validation accuracy. Positive mass improves accuracy across learning rates, although larger masses can hurt at the smallest step sizes.

Local stability of finite-mass dynamics. Finite rest mass allows updates to vanish near a minimum, but local convergence also depends on gradient mixing and the step size relative to curvature. In section D, we derive an exact local linear stability band for fixed-mass M-Lion on a quadratic (theorem D.1), governed by $s_i = \eta\lambda_i/\rho_0$ for each Hessian eigenvalue λ_i. Too much instantaneous-gradient injection can cause alternating overshoot, while sufficiently large $\beta_{\text{mix}} > \beta_{\text{mom}}$ destabilizes the dynamics through negative curvature-aware dissipation. The physically motivated ordering $\beta_{\text{mix}} \leq \beta_{\text{mom}}$ avoids the upper instability boundary, but a curvature-dependent step-size bound still applies. The quadratic experiment in Fig. 3 illustrates these two failure modes and the decaying oscillations between them.

4 Experiments

We test two consequences of the proposed dynamics: finite rest mass permits updates to shrink, and curvature-aware dissipation improves progress on heterogeneous losses. Adaptive mass remains

Table 2: **LLM confirmation across seven initialization seeds.** Final validation loss and perplexity after 3.25B tokens; entries show mean ± sample SD. The $n = 6$ column excludes exploratory seed 100. Bold marks improvements over matched $\zeta = 0$ controls using paired loss tests, Holm-adjusted across ten comparisons. Full settings and selection procedures appear in section E.

Optimizer	β_{mom}	η	ζ	β_{mix}	Loss ($n = 7$)	PPL ($n = 7$)	PPL ($n = 6$)
SS AdamW	0.95	0.008	0	0.95	$3.0128_{\pm 0.0034}$	$20.344_{\pm 0.069}$	$20.344_{\pm 0.076}$
M-Lion			0.0125	0.938125	$3.0074_{\pm 0.0097}$	$20.236_{\pm 0.197}$	$20.273_{\pm 0.187}$
M-Lion			0.025	0.92625	$3.0108_{\pm 0.0097}$	$20.304_{\pm 0.199}$	$20.230_{\pm 0.047}$
SS AdamW	0.975	0.012	0	0.975	$3.0244_{\pm 0.0031}$	$20.582_{\pm 0.064}$	$20.566_{\pm 0.053}$
M-Lion			0.025	0.950625	$\mathbf{3.0007}_{\pm 0.0018}$	$\mathbf{20.099}_{\pm 0.036}$	$\mathbf{20.107}_{\pm 0.034}$
M-Lion			0.05	0.92625	$\mathbf{3.0006}_{\pm 0.0042}$	$\mathbf{20.098}_{\pm 0.085}$	$\mathbf{20.123}_{\pm 0.059}$
SS AdamW	0.9	0.016	0	0.9	$3.0286_{\pm 0.0065}$	$20.668_{\pm 0.134}$	$20.668_{\pm 0.147}$
M-Lion			0.025	0.8775	$3.0226_{\pm 0.0079}$	$20.544_{\pm 0.162}$	$20.572_{\pm 0.158}$
M-Lion			0.05	0.855	$3.0191_{\pm 0.0025}$	$20.473_{\pm 0.051}$	$20.487_{\pm 0.040}$
SS AdamW	0.975	0.016	0	0.975	$3.0238_{\pm 0.0035}$	$20.570_{\pm 0.072}$	$20.562_{\pm 0.076}$
M-Lion			0.05	0.92625	$\mathbf{2.9980}_{\pm 0.0034}$	$\mathbf{20.045}_{\pm 0.068}$	$\mathbf{20.050}_{\pm 0.073}$
SS AdamW	0.975	0.024	0	0.975	$3.0415_{\pm 0.0075}$	$20.936_{\pm 0.158}$	$20.936_{\pm 0.173}$
M-Lion			0.05	0.92625	$\mathbf{3.0042}_{\pm 0.0029}$	$\mathbf{20.170}_{\pm 0.058}$	$\mathbf{20.164}_{\pm 0.061}$

tied to momentum decay; curvature strength is $\zeta = 1 - \beta_{\text{mix}}/\beta_{\text{mom}}$. Full experimental settings, hyperparameter grids, and statistical procedures are provided in section E.

Rest mass on CIFAR-10. For ResNet-18 trained for 80,000 steps, positive rest mass improves test accuracy over the massless Lion, Signum, and SS-AdamW reductions (Fig. 4). Hyperparameters and checkpoints are selected by validation accuracy. Seven-seed comparisons reach approximately 95.2% test accuracy with both fixed and adaptive mass, comparable to momentum SGD (95.29%). The learning-rate curves show that the benefit extends across multiple step sizes.

Curvature-aware dissipation. Positive dissipation accelerates optimization on the heterogeneous 9D quadratic (Fig. 2a–b). For the 12-layer language model trained on 3.25B tokens, $\beta_{\text{mom}} = 0.975$ and $\zeta = 0.05$ improve perplexity across three learning rates (panel c; Table 2). Each comparison wins on all six fresh seeds and passes Holm correction across ten planned tests. The lowest replicated mean is 20.045 ± 0.068, versus 20.344 ± 0.069 for the selected SS-AdamW baseline: a 1.47% reduction (1.44% excluding exploratory seed 100). Values are mean ± sample SD; two held-out seeds confirmed the interim winner.

5 Conclusions and Discussion

We derive sign-based optimization from relativistic kinematics and dissipative dynamics. The massless limit yields saturated updates, finite rest mass allows their magnitude to shrink, and curvature-aware dissipation introduces the gradient mixing that distinguishes Lion from Signum. Combining these mechanisms with SS-Adam's variance-based mass adaptation yields *Massive Lion* (M-Lion): an adaptive optimizer that interpolates between SGD-like and sign-like responses while adjusting to gradient statistics. This construction unifies Lion, Signum, and SS-AdamW within a family whose rest mass, adaptation, and dissipation can be controlled separately.

The experiments illustrate what these complementary mechanisms accomplish. On CIFAR-10, positive rest mass improves both sign-based and adaptive reductions across learning rates, bringing their accuracy close to momentum SGD. In language modeling, curvature-aware dissipation improves on SS-AdamW at high momentum, with gains that persist across learning rates and fresh initialization seeds. Controlled quadratic experiments clarify the corresponding dynamics. Together, these results demonstrate the value of a unifying theoretical framework: it explains existing optimizers and guides the design of an adaptable family with distinct mechanisms for different optimization demands.

AI USE STATEMENT

We used generative AI tools to assist with software implementation and debugging, experimental analysis and visualization, literature searches, mathematical derivations, and manuscript drafting and editing. The authors are responsible for the final text, results, and claims.

REPRODUCIBILITY STATEMENT

We have uploaded our code as a supplementary zip file in this submission. To ensure reproducibility, we plan to publicly release the full codebase, our model checkpoints, and generated data upon acceptance.

REFERENCES

Felipe Alvarez, Hédy Attouch, Jérôme Bolte, and Patrick Redont. A second-order gradient-like dissipative dynamical system with Hessian-driven damping. application to optimization and mechanics. *Journal de Mathématiques Pures et Appliquées*, 81(8):747–779, 2002. doi: 10.1016/S0021-7824(01)01253-3. URL https://doi.org/10.1016/S0021-7824(01)01253-3.

Karl J Astrom and Tore Hägglund. Advanced pid control. *IEEE Control Systems*, 26(1):98–101, 2006. doi: 10.1109/MCS.2006.1580160.

Hédy Attouch, Juan Peypouquet, and Patrick Redont. Fast convex optimization via inertial dynamics with Hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783, 2016. doi: 10.1016/j.jde.2016.08.020.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 7 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 7 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves constrained optimization: As Lyapunov predicts. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.05898.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ne6zeqLFCZ.

Albert Einstein. Zur elektrodynamik bewegter körper. 322(10):891–921, 1905. doi: 10.1002/andp.19053221004.

Guilherme França, Jeremias Sulam, Daniel P. Robinson, and René Vidal. Conformal symplectic and relativistic optimization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16916–16926, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c4b108f53550f1d5967305a9a8140ddd-Abstract.html.

Guilherme França, Michael I Jordan, and René Vidal. On dissipative symplectic integration with applications to gradient-based optimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(4):043402, 2021. doi: 10.1088/1742-5468/abf5d4.

Siyuan Guo and Bernhard Schölkopf. Physics of learning: A lagrangian perspective to different learning paradigms, 2025. URL https://arxiv.org/abs/2509.21049.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

L.D. Landau and E.M. Lifshitz. *The Classical Theory of Fields*, volume 2 of *Course of Theoretical Physics*. Pergamon Press, 4th edition, 1975. URL https://archive.org/details/classicaltheoryo0000land_k6k2.

Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and adam for deep learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1fUpoR5FQ.

Nicolas Minorsky. Directional stability of automatically steered bodies. *Journal of the American Society of Naval Engineers*, 34(2):280–309, 1922. doi: 10.1111/j.1559-3584.1922.tb04958.x.

Yu. E. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983. URL https://www.mathnet.ru/eng/dan46009. Translated from Doklady Akademii Nauk SSSR, 269(3):543–547, 1983.

Antonio Orvieto and Robert M. Gower. In search of adam's secret sauce. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=CH72XyZs4y.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318. PMLR, 2013. URL https://proceedings.mlr.press/v28/pascanu13.html.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-5553(64)90137-5. URL https://doi.org/10.1016/0041-5553(64)90137-5.

Lord Rayleigh. *The Theory of Sound*, volume 1. Macmillan, London, 1877.

Jian Rong, Chenhao Ma, Qinghui Zhang, Yong Cao, and Weili Kou. A refined lion optimizer for deep learning. *Scientific reports*, 15(1):23082, 2025. doi: 10.1038/s41598-025-07112-4.

Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195(1–2):79–148, 2022. doi: 10.1007/s10107-021-01681-8. URL https://doi.org/10.1007/s10107-021-01681-8.

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. URL https://jmlr.org/papers/v17/15-084.html.

Pranav Vaidhyanathan, Lucas Schorling, Natalia Ares, and Michael A. Osborne. A physics-inspired optimizer: Velocity regularized adam, 2025. URL https://arxiv.org/abs/2505.13196.

John Vastola, Samuel J. Gershman, and Kanaka Rajan. Gradient descent as loss landscape navigation: a normative framework for deriving learning rules. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=oMi4uyNOlL.

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.

Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima, 2021. URL https://arxiv.org/abs/2002.03495.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=X6rqEpbnj3.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7654–7663. PMLR, 6 2019. URL https://proceedings.mlr.press/v97/zhu19e.html.

John G. Ziegler and Nathaniel B. Nichols. Optimum settings for automatic controllers. *Transactions of the ASME*, 64(11):759–768, 1942.

APPENDIX: SIGN-BASED OPTIMIZERS ARE MASSLESS RELATIVISTIC LEARNERS

A A physics approach to deriving optimization dynamics **13**

A.1 Two motivations for curvature-aware dissipation: mechanical and thermodynamical arguments . . . 14

B Derivation of the relativistic softclip relation **15**

C Proof of theorem 3.1 **16**

D Stability band analysis **18**

D.1 Linearization and the stability band . . . 18

D.2 Proof of the stability criterion . . . 19

D.3 Illustration on an anisotropic quadratic . . . 20

E Experimental details and extended results **20**

E.1 Controlled quadratic experiments . . . 20

E.2 CIFAR-10 . . . 21

E.3 Language modeling . . . 21

E.4 Rest mass across training cutoffs . . . 23

F Relating the curvature-aware dissipation to PID control **23**

A A PHYSICS APPROACH TO DERIVING OPTIMIZATION DYNAMICS

The principle of stationary action. Physics is fundamentally about dynamics. The most successful theories in physics—including Newtonian mechanics, special and general relativity, electromagnetism, and quantum field theory—share two core attributes. They all: (1) identify some dynamical objects of interest, and (2) derive their dynamics via the *principle of stationary action*:

Principle of stationary action: The closest physics has come to a theory of everything

Physical systems evolve to extremize *action* S (time integral of the *Lagrangian* L):

$$S = \int L\,dt, \quad \xrightarrow{\text{real physical systems follow trajectories that extremize action}} \quad \boxed{\delta S = 0} \tag{21}$$

Extremizing the action leads directly to the Euler–Lagrange equations:

$$\delta S = 0 \quad \Longrightarrow \quad \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\boldsymbol{x}}}\right) - \frac{\partial L}{\partial \boldsymbol{x}} = \mathbf{0}\,. \tag{22}$$

Physics of adaptation. In this paper, we show that the *dynamics of adaptation* are no different. We treat the parameter vector $\boldsymbol{\theta}(t)$ as the position of a fictitious particle moving in a potential $V(\boldsymbol{\theta})$. We then identify Lagrangians of learning systems and derive optimization algorithms from action principles. However, settling at a minimum requires an important additional component: *energy dissipation*. Dissipation and other non-conservative forces, denoted by the generalized force vector $\mathbf{Q}$, augment the Euler–Lagrange equations as follows:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\boldsymbol{\theta}}}\right) - \frac{\partial L}{\partial \boldsymbol{\theta}} = \mathbf{Q}\,. \tag{23}$$

A.1 Two motivations for curvature-aware dissipation: mechanical and thermodynamical arguments

The isotropic force above is one possible choice of dissipation. More generally, we can write

$$\mathbf{Q} \;=\; -\boldsymbol{\Gamma}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}})\, \dot{\boldsymbol{\theta}}\,, \tag{24}$$

where the friction tensor $\boldsymbol{\Gamma}$ determines how motion is damped in different directions. What should this tensor look like? We approach this question through two physical considerations: the relaxation of deterministic motion and the response to stochastic forcing.

The first motivation: a mechanical argument. Consider classical motion near a quadratic minimum with constant Hessian $\mathbf{H} \succ 0$. To let friction respond to the local curvature, consider tensors of the form $\boldsymbol{\Gamma} = f(\mathbf{H})$, with friction coefficient $\gamma_i = f(\lambda_i)$ along a Hessian eigenvector of eigenvalue λ_i. The displacement x_i along this direction obeys

$$m\ddot{x}_i + \gamma_i \dot{x}_i + \lambda_i x_i = 0\,. \tag{25}$$

For a fixed mode, the fastest asymptotic decay occurs at critical damping, $\gamma_i = 2\sqrt{m\lambda_i}$. Thus, sharper directions require stronger damping. Isotropic friction cannot critically damp modes with different curvatures: choosing enough friction to suppress oscillations in sharp directions can overdamp flatter ones. A related square-root dependence on curvature appears in the ballistic learning rule of Vastola et al. (2025), who obtain $\mathbf{H}^{-1/2}$ preconditioning from a different optimal-control argument.

Critical damping suggests $\boldsymbol{\Gamma} \propto \mathbf{H}^{1/2}$, but it also motivates a simpler approximation. Expanding the critical-damping coefficient around a reference curvature $\lambda_* > 0$ gives

$$2\sqrt{m\lambda} \;\approx\; \sqrt{m\lambda_*} + \sqrt{\frac{m}{\lambda_*}}\,\lambda\,. \tag{26}$$

This has the form of an isotropic baseline plus a linear curvature correction. It suggests considering the family

$$\boldsymbol{\Gamma} \;=\; \gamma_0 \mathbf{I} + \zeta \mathbf{H}\,, \qquad \gamma_0 > 0, \quad \zeta \geq 0\,, \tag{27}$$

with independently adjustable coefficients. This approximation does not critically damp every mode, but a suitable positive correction can reduce persistent oscillations in sharp directions while adding relatively little friction in flatter ones.

The second motivation: a thermodynamic argument. Consider the fluctuations introduced by stochastic gradients. Previous work has found that gradient-noise covariance can be approximately proportional to the Hessian near minima (Zhu et al., 2019; Xie et al., 2021). Taking this as a local model, we represent the stochastic forcing in each mode by zero-mean white noise:

$$m\ddot{x}_i + \gamma_i \dot{x}_i + \lambda_i x_i \;=\; \xi_i(t)\,, \qquad \mathbb{E}[\xi_i(t)\xi_j(s)] \;=\; 2D\lambda_i \delta_{ij}\delta(t-s)\,, \tag{28}$$

where $D > 0$ sets the noise intensity.

For these fixed local coefficients, the stationary second moments give

$$\mathbb{E}[x_i^2] \;=\; \frac{D}{\gamma_i}\,, \qquad \mathbb{E}[V_i] \;=\; \frac{\lambda_i}{2}\mathbb{E}[x_i^2] \;=\; \frac{D\lambda_i}{2\gamma_i}\,. \tag{29}$$

With isotropic friction, $\gamma_i = \gamma_0$, parameter fluctuations have the same variance in every mode, even though a displacement in a sharp direction is more costly in loss. Consequently, the expected excess loss due to fluctuations grows linearly with curvature.

The linear correction in eq. (27) mitigates this effect:

$$\mathbb{E}[x_i^2] \;=\; \frac{D}{\gamma_0 + \zeta\lambda_i}\,, \qquad \mathbb{E}[V_i] \;=\; \frac{D\lambda_i}{2(\gamma_0 + \zeta\lambda_i)}\,. \tag{30}$$

Sharper directions now have smaller parameter fluctuations. For $\zeta > 0$, their expected loss contribution approaches $D/(2\zeta)$ as curvature increases, instead of growing without bound. The isotropic baseline retains finite momentum damping in flat directions, while the curvature correction selectively suppresses fluctuations in sharper ones.

These two considerations motivate an isotropic friction tensor with a linear Hessian correction. This choice also has a useful analytical property: for the smooth loss gradient along a trajectory, $\mathbf{H}(\boldsymbol{\theta})\dot{\boldsymbol{\theta}} = \dot{\boldsymbol{g}}$. Setting $\gamma_0 = m/\tau_d$ therefore gives:

$$\mathbf{Q} \;=\; -\left(\frac{m}{\tau_d}\mathbf{I} + \zeta\mathbf{H}\right)\dot{\boldsymbol{\theta}} \;=\; \underbrace{-\frac{\boldsymbol{p}}{\tau_d}}_{\text{isotropic}} \;-\; \underbrace{\zeta\mathbf{H}\dot{\boldsymbol{\theta}}}_{\text{curvature-aware}} \;=\; -\frac{\boldsymbol{p}}{\tau_d} \;-\; \zeta\dot{\boldsymbol{g}}\,. \tag{31}$$

The curvature correction becomes a time derivative of the gradient, allowing the momentum equation to be integrated by parts into a closed-form gradient-history representation. As we show below, this representation produces Lion's gradient mixing. Deterministic damping, control of noise-induced loss, and analytical tractability thus jointly motivate our choice of dissipation.

B Derivation of the relativistic softclip relation

In section 2, we derived the canonical momentum for a single coordinate i under the separable relativistic Lagrangian as:

$$p_i \;=\; m\gamma_i\dot{\theta}_i, \tag{32}$$

where m is the physical rest mass, and γ_i is the coordinate-specific Lorentz factor:

$$\gamma_i \;=\; \frac{1}{\sqrt{1-\frac{\dot{\theta}_i^2}{c^2}}}. \tag{33}$$

Our goal is to invert this relationship to express the velocity $\dot{\theta}_i$ strictly as a function of the momentum p_i. We begin by substituting the definition of γ_i into the momentum equation:

$$p_i \;=\; \frac{m\dot{\theta}_i}{\sqrt{1-\frac{\dot{\theta}_i^2}{c^2}}}. \tag{34}$$

To isolate $\dot{\theta}_i$, we square both sides. This is a mathematically safe operation without loss of sign information because $\gamma_i > 0$ and $m > 0$, meaning p_i and $\dot{\theta}_i$ always share the same sign.

$$p_i^2 \;=\; \frac{m^2\dot{\theta}_i^2}{1-\frac{\dot{\theta}_i^2}{c^2}}. \tag{35}$$

Next, we multiply both sides by the denominator to clear the fraction:

$$p_i^2\left(1-\frac{\dot{\theta}_i^2}{c^2}\right) \;=\; m^2\dot{\theta}_i^2. \tag{36}$$

Distributing p_i^2 on the left side yields:

$$p_i^2 - p_i^2\frac{\dot{\theta}_i^2}{c^2} \;=\; m^2\dot{\theta}_i^2. \tag{37}$$

We now group all terms containing $\dot{\theta}_i^2$ on the right side of the equation:

$$p_i^2 \;=\; m^2\dot{\theta}_i^2 + p_i^2\frac{\dot{\theta}_i^2}{c^2}. \tag{38}$$

Factoring out $\dot{\theta}_i^2$ gives:

$$p_i^2 \;=\; \dot{\theta}_i^2\left(m^2 + \frac{p_i^2}{c^2}\right). \tag{39}$$

To connect this to the optimizer hyperparameters, we factor out $\frac{1}{c^2}$ from the parenthetical term:

$$p_i^2 \;=\; \frac{\dot{\theta}_i^2}{c^2}\left(m^2c^2 + p_i^2\right). \tag{40}$$

At this stage, we substitute our physical definition for the saturation scale, $\rho := mc$. This yields:

$$p_i^2 \;=\; \frac{\dot\theta_i^2}{c^2}\left(\rho^2 + p_i^2\right). \tag{41}$$

Dividing by the parenthetical term isolates $\dot\theta_i^2$:

$$\frac{p_i^2 c^2}{\rho^2 + p_i^2} \;=\; \dot\theta_i^2. \tag{42}$$

Finally, we take the square root of both sides. As established earlier, $\dot\theta_i$ and p_i must carry the same sign. Therefore, we can safely write the exact continuous velocity:

$$\dot\theta_i \;=\; c\frac{p_i}{\sqrt{\rho^2 + p_i^2}}. \tag{43}$$

Recognizing the fractional term as our defined $\mathrm{softclip}$ function, we arrive at the exact kinematic relation:

$$\dot\theta_i \;=\; c \cdot \mathrm{softclip}(p_i; \rho). \tag{44}$$

This completes the derivation, showing that the bounded-influence update in M-Lion arises from inverting the geometry of relativistic momentum.

C Proof of Theorem 3.1

Throughout this proof, the rest mass $m > 0$ is constant. We assume $V \in C^2(\mathbb{R}^d)$, $c > 0$, $\tau_d > 0$, and $0 \le \zeta < \tau_d$, as in theorem 3.1.

Integrating the momentum equation. We start from the curvature-aware momentum equation,

$$\dot{\boldsymbol{p}}(t) + \frac{\boldsymbol{p}(t)}{\tau_d} \;=\; -\boldsymbol{g}(t) - \zeta\dot{\boldsymbol{g}}(t)\,.$$

For a given gradient history, this is a first-order linear ODE in $\boldsymbol{p}$. Multiplying by the integrating factor e^{t/τ_d} and integrating from 0 to t gives

$$\boldsymbol{p}(t) = e^{-t/\tau_d}\boldsymbol{p}(0) - \int_0^t e^{-(t-s)/\tau_d}\boldsymbol{g}(s)\,ds - \zeta\int_0^t e^{-(t-s)/\tau_d}\dot{\boldsymbol{g}}(s)\,ds\,. \tag{45}$$

To eliminate the gradient derivative from the second integral, we integrate by parts. Taking $u = e^{-(t-s)/\tau_d}$ and $dv = \dot{\boldsymbol{g}}(s)\,ds$ yields

$$\begin{aligned}\int_0^t e^{-(t-s)/\tau_d}\dot{\boldsymbol{g}}(s)\,ds &= \Big[e^{-(t-s)/\tau_d}\boldsymbol{g}(s)\Big]_0^t - \frac{1}{\tau_d}\int_0^t e^{-(t-s)/\tau_d}\boldsymbol{g}(s)\,ds \\ &= \boldsymbol{g}(t) - e^{-t/\tau_d}\boldsymbol{g}(0) - \frac{1}{\tau_d}\int_0^t e^{-(t-s)/\tau_d}\boldsymbol{g}(s)\,ds\,.\end{aligned}$$

Substituting into eq. (45), we obtain

$$\boldsymbol{p}(t) = -\zeta\boldsymbol{g}(t) - \left(1 - \frac{\zeta}{\tau_d}\right)\int_0^t e^{-(t-s)/\tau_d}\boldsymbol{g}(s)\,ds + e^{-t/\tau_d}\left[\boldsymbol{p}(0) + \zeta\boldsymbol{g}(0)\right]\,.$$

Now introduce the continuous gradient EMA,

$$\boldsymbol{\mu}(t) = e^{-t/\tau_d}\boldsymbol{\mu}(0) + \frac{1}{\tau_d}\int_0^t e^{-(t-s)/\tau_d}\boldsymbol{g}(s)\,ds\,.$$

The momentum can therefore be written as

$$\boldsymbol{p}(t) = -\zeta\boldsymbol{g}(t) - (\tau_d - \zeta)\boldsymbol{\mu}(t) + e^{-t/\tau_d}\left[\boldsymbol{p}(0) + \zeta\boldsymbol{g}(0) + (\tau_d - \zeta)\boldsymbol{\mu}(0)\right]\,. \tag{46}$$

Under the consistent initialization

$$\boldsymbol{p}(0) = -\zeta\boldsymbol{g}(0) - (\tau_d - \zeta)\boldsymbol{\mu}(0)\,,$$

the final term vanishes. Dividing by $-\tau_d$ gives

$$-\frac{\boldsymbol{p}(t)}{\tau_d} = \frac{\zeta}{\tau_d}\boldsymbol{g}(t) + \left(1 - \frac{\zeta}{\tau_d}\right)\boldsymbol{\mu}(t) \,,$$

which is eq. (10). For other initial conditions, eq. (46) gives the additional exponentially decaying contribution explicitly. In particular, choosing $\boldsymbol{\mu}(0) = \mathbf{0}$ in the theorem corresponds to $\boldsymbol{p}(0) = -\zeta\boldsymbol{g}(0)$.

This decomposition identifies two contributions to the effective update direction $-\boldsymbol{p}/\tau_d$: an instantaneous gradient term, arising from the integration-by-parts boundary term, and an exponentially weighted gradient history. The curvature-aware force changes their relative weights.

Discretizing the exponential memory. We discretize this gradient-history representation. At iteration t, the available state is $(\boldsymbol{\theta}_t, \boldsymbol{\mu}_{t-1})$. We evaluate the gradient $\boldsymbol{g}_t$ at $\boldsymbol{\theta}_t$ and hold it fixed during the momentum step of duration Δt. For stochastic optimization, the same construction uses a stochastic gradient estimate.

The continuous EMA satisfies $\tau_d\dot{\boldsymbol{\mu}} = \boldsymbol{g} - \boldsymbol{\mu}$. Integrating this equation exactly under the frozen-gradient approximation gives

$$\begin{aligned}\boldsymbol{\mu}_t &= e^{-\Delta t/\tau_d}\boldsymbol{\mu}_{t-1} + \frac{1}{\tau_d}\int_0^{\Delta t} e^{-(\Delta t - s)/\tau_d}\boldsymbol{g}_t \, ds \\ &= \beta_{\text{mom}}\boldsymbol{\mu}_{t-1} + (1-\beta_{\text{mom}})\boldsymbol{g}_t \,, \qquad \beta_{\text{mom}} \coloneqq e^{-\Delta t/\tau_d} \,.\end{aligned}$$

This is eq. (14). The integration is exact for the frozen-gradient memory equation; the coupled parameter dynamics are discretized separately below.

Substituting this EMA update into the momentum representation gives

$$\begin{aligned}-\frac{\boldsymbol{p}_t}{\tau_d} &= \frac{\zeta}{\tau_d}\boldsymbol{g}_t + \left(1 - \frac{\zeta}{\tau_d}\right)\left[\beta_{\text{mom}}\boldsymbol{\mu}_{t-1} + (1-\beta_{\text{mom}})\boldsymbol{g}_t\right] \\ &= \beta_{\text{mom}}\left(1 - \frac{\zeta}{\tau_d}\right)\boldsymbol{\mu}_{t-1} + \left[1 - \beta_{\text{mom}}\left(1 - \frac{\zeta}{\tau_d}\right)\right]\boldsymbol{g}_t \,.\end{aligned}$$

Defining

$$\beta_{\text{mix}} \coloneqq \beta_{\text{mom}}\left(1 - \frac{\zeta}{\tau_d}\right) \,, \qquad \boldsymbol{d}_t \coloneqq -\frac{\boldsymbol{p}_t}{\tau_d} \,,$$

we recover Lion's mixed update direction:

$$\boldsymbol{d}_t = \beta_{\text{mix}}\boldsymbol{\mu}_{t-1} + (1-\beta_{\text{mix}})\boldsymbol{g}_t \,.$$

This establishes eq. (12). No numerical differentiation of the gradient is needed.

The position update and finite rest mass. The relativistic velocity law is

$$\dot{\theta}_i = c\frac{p_i}{\sqrt{\rho^2 + p_i^2}} \,, \qquad \rho = mc \,.$$

Using the updated momentum $\boldsymbol{p}_t = -\tau_d\boldsymbol{d}_t$ in a position step of duration Δt gives

$$\begin{aligned}\theta_{t+1,i} &= \theta_{t,i} + c\Delta t\,\frac{-\tau_d d_{t,i}}{\sqrt{\rho^2 + \tau_d^2 d_{t,i}^2}} \\ &= \theta_{t,i} - \eta\,\frac{d_{t,i}}{\sqrt{\rho_0^2 + d_{t,i}^2}} \,,\end{aligned}$$

where

$$\eta \coloneqq c\Delta t \,, \qquad \rho_0 \coloneqq \frac{\rho}{\tau_d} = \frac{mc}{\tau_d} \,.$$

The factor τ_d appears because $\boldsymbol{d}_t$ is the normalized momentum $-\boldsymbol{p}_t/\tau_d$. In units where $\tau_d = 1$, the two saturation scales coincide: $\rho_0 = \rho = mc$.

Applying the position update coordinatewise establishes eq. (13). Together with the direction and EMA recurrences, it gives M-Lion with fixed rest mass.

The ordering of the decay coefficients. Since $\Delta t > 0$ and $\tau_d > 0$, we have $0 < \beta_{\text{mom}} < 1$. The assumed range $0 \le \zeta < \tau_d$ gives

$$0 < \beta_{\text{mix}} \le \beta_{\text{mom}} < 1 \,, \qquad \zeta > 0 \iff \beta_{\text{mix}} < \beta_{\text{mom}} \,.$$

Thus, within this construction, Lion's usual ordering of its two coefficients corresponds to a positive coefficient of the curvature-aware correction.

Massless and zero-mixing reductions. At fixed c and τ_d, taking $m \to 0$ sends $\rho_0 \to 0$. For each nonzero coordinate of the update direction,

$$\lim_{\rho_0 \to 0} \frac{d_{t,i}}{\sqrt{\rho_0^2 + d_{t,i}^2}} = \operatorname{sign}(d_{t,i}) \,.$$

At $d_{t,i} = 0$, the finite-mass response is zero for every $\rho_0 > 0$, and we retain this value in the limit. The limiting position step is therefore

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \operatorname{sign}(\boldsymbol{d}_t) \,,$$

recovering Lion's signed update without weight decay.

Setting $\zeta = 0$ gives $\beta_{\text{mix}} = \beta_{\text{mom}}$, so

$$\boldsymbol{d}_t = \beta_{\text{mom}} \boldsymbol{\mu}_{t-1} + (1 - \beta_{\text{mom}}) \boldsymbol{g}_t = \boldsymbol{\mu}_t \,.$$

For $\rho_0 > 0$, this is fixed-mass M-Signum. Taking $\rho_0 \to 0$ as well recovers Signum.

For finite ρ_0, each coordinate has a linear response when $|d_{t,i}| \ll \rho_0$ and approaches a signed step of magnitude η when $|d_{t,i}| \gg \rho_0$. These are the classical and saturated regimes of the same fixed-mass update.

This completes the proof of theorem 3.1. Variance-based mass adaptation is introduced subsequently in discrete time; its SS-AdamW reduction is established separately in proposition 3.2.

D Stability Band Analysis

Finite rest mass makes the update smooth at zero direction, allowing M-Lion to settle at a minimum. Whether it does so locally also depends on the learning rate, curvature, and gradient mixing. Here we derive the exact stability band of the linearized, fixed-mass dynamics and compare it with trajectories on an anisotropic quadratic. Throughout this section, gradients are exact, $\eta > 0$ and $\rho_0 > 0$ are constant, and both adaptive mass and weight decay are disabled.

D.1 Linearization and the Stability Band

Let V be twice continuously differentiable, and let $\boldsymbol{\theta}_\star$ be a strict local minimum with positive-definite Hessian $\mathbf{H}_\star = \nabla^2 V(\boldsymbol{\theta}_\star)$. The fixed-mass response has the expansion

$$\frac{d}{\sqrt{\rho_0^2 + d^2}} = \frac{d}{\rho_0} + \mathcal{O}\left(\frac{d^3}{\rho_0^3}\right) \,.$$

Consequently, the linearization of the parameter and momentum updates decouples along the eigenvectors of $\mathbf{H}_\star$. For a mode with eigenvalue $\lambda_i > 0$, write x_t and μ_{t-1} for the projections of $\boldsymbol{\theta}_t - \boldsymbol{\theta}_\star$ and $\boldsymbol{\mu}_{t-1}$ onto that eigenvector. The linearized recurrences are

$$x_{t+1} = x_t - \frac{\eta}{\rho_0}\left[\beta_{\text{mix}} \mu_{t-1} + (1 - \beta_{\text{mix}}) \lambda_i x_t\right] \,,$$
$$\mu_t = \beta_{\text{mom}} \mu_{t-1} + (1 - \beta_{\text{mom}}) \lambda_i x_t \,.$$

Defining $s_i \coloneqq \eta \lambda_i / \rho_0$ and $\boldsymbol{z}_t \coloneqq (x_t, \mu_{t-1})^\top$ gives

$$\boldsymbol{z}_{t+1} = \mathbf{A}_i \boldsymbol{z}_t \,, \qquad \mathbf{A}_i = \begin{pmatrix} 1 - s_i(1 - \beta_{\text{mix}}) & -\eta \beta_{\text{mix}} / \rho_0 \\ (1 - \beta_{\text{mom}}) \lambda_i & \beta_{\text{mom}} \end{pmatrix} \,. \tag{47}$$

Here $\rho_0 = \rho / \tau_d$ is the discrete saturation scale, rather than the physical momentum scale $\rho = mc$.

Theorem D.1 (Quadratic stability band). *Assume $0 < \beta_{mom} < 1$, $0 \le \beta_{mix} < 1$, and $s_i > 0$. The linearized dynamics of mode i are asymptotically stable if and only if every eigenvalue of $\mathbf{A}_i$ has modulus strictly less than one. Equivalently,*

$$\boxed{\frac{(1+\beta_{mom})(s_i-2)}{2s_i} < \beta_{mix} < \beta_{mom} + \frac{1-\beta_{mom}}{s_i}\,.} \tag{48}$$

The full linearization is asymptotically stable exactly when this condition holds for every Hessian eigenmode. In that case, the nonlinear fixed-mass optimizer is locally asymptotically stable at $(\boldsymbol{\theta}_\star, \mathbf{0})$.

Both bounds become more restrictive as s_i increases, so it suffices to check the sharpest mode, $s_{\max} = \eta\lambda_{\max}(\mathbf{H}_\star)/\rho_0$. For $s_i < 2$, the lower bound is automatically satisfied by $\beta_{\text{mix}} \ge 0$; at $s_i = 2$, strict stability requires $\beta_{\text{mix}} > 0$. Equality at either edge of the band is excluded from strict linear stability and requires a separate nonlinear analysis.

D.2 Proof of the stability criterion

Proof of theorem D.1. The trace and determinant of the mode matrix are

$$T_i := \operatorname{tr}(\mathbf{A}_i) = 1 + \beta_{\text{mom}} - s_i(1-\beta_{\text{mix}})\,,$$
$$D_i := \det(\mathbf{A}_i) = \beta_{\text{mom}} + s_i(\beta_{\text{mix}} - \beta_{\text{mom}})\,.$$

Its characteristic polynomial is $P_i(r) = r^2 - T_i r + D_i$. The second-order Schur criterion states that both roots lie strictly inside the unit disk if and only if

$$P_i(1) > 0\,, \qquad P_i(-1) > 0\,, \qquad D_i < 1\,.$$

To see sufficiency directly, complex-conjugate roots have squared modulus D_i. For real roots, positivity at 1 and -1 places either both roots inside $(-1, 1)$ or both beyond the same endpoint; the latter possibilities have product greater than one and are excluded by $D_i < 1$. Necessity follows by evaluating the factored polynomial at the two endpoints and taking the product of its roots.

Substituting the expressions for T_i and D_i gives

$$P_i(1) = s_i(1-\beta_{\text{mom}}) > 0\,,$$
$$P_i(-1) = (2-s_i)(1+\beta_{\text{mom}}) + 2s_i\beta_{\text{mix}}\,,$$
$$1 - D_i = 1 - \beta_{\text{mom}} - s_i(\beta_{\text{mix}} - \beta_{\text{mom}})\,.$$

The first condition holds by assumption. The second and third are respectively the lower and upper inequalities in eq. (48). Applying the criterion to every mode establishes stability of the full linearization. Since $\rho_0 > 0$ makes the update map continuously differentiable near the equilibrium, strict linear stability also implies local asymptotic stability of the nonlinear map. □

Interpretation of the two boundaries. At the lower boundary, $P_i(-1) = 0$: an eigenvalue reaches -1, marking the onset of an alternating instability. Thus, increasing the instantaneous-gradient contribution by reducing β_{mix} can destabilize the discrete update at a fixed value of $\eta\lambda_i/\rho_0$, even though it corresponds to stronger curvature-aware dissipation in continuous time. At the upper boundary, $D_i = 1$; beyond it, the product of the eigenvalue moduli exceeds one. This boundary lies above β_{mom}, in the extension of the update rule corresponding to negative curvature-aware dissipation.

In the physically motivated range $0 \le \beta_{\text{mix}} \le \beta_{\text{mom}}$, the upper bound is automatic, but the lower bound still imposes

$$\frac{\eta\lambda_{\max}(\mathbf{H}_\star)}{\rho_0} < \frac{2(1+\beta_{\text{mom}})}{1+\beta_{\text{mom}} - 2\beta_{\text{mix}}}\,. \tag{49}$$

The ordering $\beta_{\text{mix}} < \beta_{\text{mom}}$ therefore does not by itself ensure stability. At $\beta_{\text{mix}} = \beta_{\text{mom}}$, the result reduces to the fixed-mass M-Signum condition $s_{\max} < 2(1+\beta_{\text{mom}})/(1-\beta_{\text{mom}})$. These are local results for positive fixed mass; the linearization does not apply to the nonsmooth massless limit.

D.3 Illustration on an anisotropic quadratic

We use the exact gradients of $V(\theta_1, \theta_2) = \frac{1}{2}(50\theta_1^2 + \theta_2^2)$, starting from $\boldsymbol{\theta}_0 = (-1, 3)$ with zero momentum. The experiment runs for $T = 100$ steps with $\eta = 0.3$, $\rho_0 = 5$, and $\beta_{\text{mom}} = 0.9$, without gradient noise, weight decay, or adaptive mass. We sweep $\beta_{\text{mix}} \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$. The settings above β_{mom} deliberately probe the extension to negative curvature-aware dissipation. AdamW is included as a reference trajectory, with the same learning rate, decay rates $(0.9, 0.999)$, and zero weight decay.

For the sharp mode, $s_{\max} = (0.3)(50)/5 = 3$, so the stability band is

$$\frac{19}{60} < \beta_{\text{mix}} < \frac{14}{15}\,, \qquad \text{or approximately} \qquad 0.317 < \beta_{\text{mix}} < 0.933\,.$$

The flat mode has $s_i = 0.06$ and adds no restriction within $0 \le \beta_{\text{mix}} < 1$.

The plotted diagnostic uses the update per unit learning rate, $\boldsymbol{v}_t := (\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)/\eta$, and measures

$$\kappa_t = \boldsymbol{v}_t^\top \mathbf{H} \boldsymbol{v}_t\,, \qquad \bar{\kappa} = \frac{1}{T}\sum_{t=0}^{T-1} \kappa_t\,.$$

Under the physical parameter map, $\boldsymbol{v}_t$ is the discretized velocity in units of c, so this diagnostic is $\dot{\boldsymbol{\theta}}^\top \mathbf{H} \dot{\boldsymbol{\theta}}/c^2$. It depends on both curvature and update magnitude. In Fig. 3, sampled settings within the predicted band show decaying oscillations and approach the minimum, whereas settings outside it sustain oscillations over the observation window. This illustrates the local criterion; it does not establish global convergence or compare independently tuned optimizers. In particular, local instability does not imply unbounded directional curvature: for M-Lion, $|v_{t,j}| \le 1$, so $\kappa_t = 50v_{t,1}^2 + v_{t,2}^2 \le 51$ on this landscape.

E Experimental details and extended results

Optimizer conventions. Adaptive mass uses the momentum-difference recurrence in eq. (16), with its decay tied to β_{mom}; disabling adaptation sets $\boldsymbol{a}_t = \mathbf{0}$. In the implementation, these settings are $\beta_3 = \kappa = \beta_{\text{mom}}$ and $\beta_3 = \kappa = 0$, respectively. SS-AdamW denotes the equal-beta, zero-rest-mass reduction without bias correction, rather than conventional AdamW with independently chosen decay rates. We parameterize discrete curvature strength by $\zeta = 1 - \beta_{\text{mix}}/\beta_{\text{mom}}$ (the continuous curvature coefficient divided by τ_d). Thus $\zeta = 0$ retains isotropic momentum dissipation, while $\zeta > 0$ adds the curvature-aware dissipation.

E.1 Controlled quadratic experiments

The stability band. Figure 3 uses exact gradients of $V(\theta_1, \theta_2) = \frac{1}{2}(50\theta_1^2 + \theta_2^2)$, starting at $(-1, 3)$ with zero optimizer state. We run 100 steps with constant learning rate 0.3, fixed rest mass 5, and $\beta_{\text{mom}} = 0.9$, without noise, weight decay, or adaptive mass. The mixing sweep is $\beta_{\text{mix}} \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99\}$; the time-series panel shows a subset for readability. AdamW uses the same learning rate, decay rates $(0.9, 0.999)$, and zero weight decay. The measured quantity is directional curvature along the update, $\dot{\boldsymbol{\theta}}^\top \mathbf{H} \dot{\boldsymbol{\theta}}$, and its trajectory average.

Rest mass and shrinking updates. The trajectory in Fig. 4a uses $V(\theta_1, \theta_2) = \frac{1}{2}(\theta_1^2 + 10\theta_2^2)$ and the same initial position $(-1, 3)$. Signum and fixed-mass M-Signum differ only in $\rho_0 \in \{0, 0.5\}$: both use $\eta = 0.15$, $\beta_{\text{mix}} = \beta_{\text{mom}} = 0.9$, zero initial momentum, and no noise or weight decay. We display the first 100 steps at constant learning rate. Positive rest mass allows small-momentum updates to shrink, whereas the massless update retains its magnitude.

Heterogeneous curvature. Figure 2a–b uses a nine-dimensional quadratic with three curvature blocks, whose spectra are $\{1, 2, 3\}$, $\{99, 100, 101\}$, and $\{4998, 4999, 5000\}$. The displayed experiment uses minibatches of size 3 and Gaussian initialization. Rest mass is zero and adaptive mass remains tied to momentum decay. Learning rates are selected on 32 tuning seeds; the displayed

results use 128 fresh confirmation seeds. Panel (a) compares $\zeta = 0$ and 0.2 at $\beta_{\text{mom}} = 0.95$, showing relative loss L_t/L_0 over the first 200 steps of the 500-step experiment. The inset shows the same trajectories on a logarithmic scale, and shading gives the interquartile range. Panel (b) varies $\beta_{\text{mom}} \in \{0.9, 0.95, 0.975\}$, retuning the learning rate at each curvature setting. It reports the source experiment's mean log-loss score, with a scoring floor of 10^{-12} and pointwise 95% intervals for the mean over confirmation seeds.

E.2 CIFAR-10

Data and training. We train ResNet-18 on 45,000 CIFAR-10 images, using a fixed 5,000-image validation split and the standard 10,000-image test set. Training uses random 32×32 crops after four-pixel padding and random horizontal flips; evaluation uses unaugmented images. Inputs are normalized channelwise. All CIFAR-10 results reported here use batch size 256 and 80,000 optimizer steps, with cosine learning-rate decay to 10^{-5}, no warmup, and no gradient clipping. Validation and test evaluation occur every 5,000 steps.

Rest-mass sweep. In Fig. 4b, adaptive mass is enabled and $\beta_{\text{mix}} = \beta_{\text{mom}} \in \{0.90, 0.95, 0.99\}$. We sweep $\rho_0 \in \{0, 10^{-5}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 100\}$, learning rates $\{0.0001, 0.0005, 0.001, 0.003, 0.008, 0.016\}$, and weight decays $\{0.01, 0.1, 1, 10\}$. Each mass/beta point selects learning rate and weight decay by validation. The plotted snapshot contains 574 completed runs of the 576-point grid. Horizontal positions denote discrete tested masses. The zero-mass endpoint is SS-AdamW; positive mass gives adaptive M-Signum.

Learning-rate response. Figure 4c–d disables adaptive mass and displays $\rho_0 \in \{0, 0.001, 0.1, 1\}$ across the same six learning rates. For each mass/learning-rate pair, we independently select betas and weight decay by validation. Lion uses nine beta pairs: $\beta_{\text{mix}} \in \{0.5, 0.7, 0.8\}$ at $\beta_{\text{mom}} = 0.9$, and $\beta_{\text{mix}} \in \{0.5, 0.7, 0.9\}$ at each of $\beta_{\text{mom}} \in \{0.95, 0.99\}$; Signum uses tied betas in $\{0.7, 0.9, 0.95, 0.99\}$. Crossing these with the four weight decays gives 36 candidates per Lion point and 16 per Signum point. Search support is shared across cells within each panel. These curves measure performance across learning rates after retuning the remaining parameters.

Repeated-seed ablations. Table 3 selects one configuration per optimizer variant using the seed-0 validation sweep, then retains that configuration for seeds 1–6. Rest mass is tuned jointly with the remaining parameters in positive-mass rows. SGD and QHM use EMA momentum and decoupled weight decay under the same training recipe. The paired tests described below exclude the tuning seed and use only seeds 1–6.

Statistical comparisons. We compared test accuracy at the best-validation-accuracy checkpoint for every pair of rows within each of the Lion, Signum, and SGD sections (6, 6, and 1 comparisons, respectively). For each pair, we matched runs by seed and applied a two-sided paired t-test to the six differences from seeds 1–6 (5 degrees of freedom); seed 0 was used to select the configurations and was omitted from inference. Holm–Bonferroni adjustment controlled the familywise error rate at 0.05 across all 13 comparisons. Boldface identifies rows with at least one adjusted-significant win and no adjusted-significant loss in their section; it does not assert equivalence between the two bold rows in either the Lion or Signum section. The table's descriptive means and standard deviations still use all seven seeds. These tests quantify variation over training seeds on the fixed CIFAR-10 test split and assume approximately normal paired differences.

E.3 Language modeling

Data and training. Figure 2c uses a 12-layer decoder-only transformer with width 768, 12 attention heads, a gated MLP with expansion $8/3$, rotary position embeddings, and untied input/output embeddings. Query/key normalization and embedding normalization are enabled. We train on SlimPajama tokenized with the GPT-NeoX tokenizer, using sequence length 2,048 and effective batch size 256 (microbatch 32 with eight accumulation steps). Training lasts 6,200 steps, or approximately 3.25 billion tokens, with 620 linear warmup steps from zero followed by cosine decay to 10^{-5}. All runs use bfloat16 and weight decay 0.1. Confirmation runs use initialization seeds 100–106; the main figure uses gradient-norm clipping at 1. The sampler traverses the tokenized training sequences

Table 3: CIFAR-10 ablations (ResNet-18). Each row uses the top configuration from the seed-0 validation sweep. Entries are mean ± sample standard deviation over seeds 0–6, using each seed's best-validation-accuracy checkpoint. Bold optimizer names and test-accuracy means mark section winners under the corrected comparisons described below. All configurations have $0 < \beta_{\text{mom}} < 1$ and retain isotropic momentum dissipation. Curvature-aware dissipation is on when $\beta_{\text{mix}} < \beta_{\text{mom}}$ and off when $\beta_{\text{mix}} = \beta_{\text{mom}}$. η: learning rate, λ: weight decay.

	Ablation			Validation	Test		Selected hyperparameters			
Optimizer	Rest mass ρ_0	Ada. mass	Curv. diss.	Accuracy (%) ↑	Accuracy (%) ↑	Loss ↓	η	λ	β_{mix}	β_{mom}
M-Lion	1	✓	✓	$95.11_{\pm 0.22}$	$\mathbf{95.22}_{\pm 0.14}$	$.186_{\pm .004}$	0.001	1	0.7	0.95
M-Lion	10	✗	✓	$95.16_{\pm 0.21}$	$\mathbf{95.22}_{\pm 0.12}$	$.185_{\pm .003}$	0.008	0.1	0.8	0.9
Curv.-aware SS AdamW	0	✓	✓	$94.39_{\pm 0.17}$	$94.47_{\pm 0.20}$	$.515_{\pm .020}$	0.008	0.1	0.7	0.9
Lion	0	✗	✓	$93.41_{\pm 0.84}$	$93.26_{\pm 0.99}$	$.639_{\pm .069}$	0.003	0.1	0.8	0.9
M-Signum	1	✓	✗	$95.01_{\pm 0.20}$	$\mathbf{95.17}_{\pm 0.07}$	$.187_{\pm .002}$	0.0005	1	—	0.9
M-Signum	10	✗	✗	$95.11_{\pm 0.21}$	$\mathbf{95.22}_{\pm 0.21}$	$.182_{\pm .005}$	0.008	0.1	—	0.7
SS AdamW	0	✓	✗	$94.46_{\pm 0.18}$	$94.43_{\pm 0.13}$	$.517_{\pm .012}$	0.008	0.1	—	0.9
Signum	0	✗	✗	$94.04_{\pm 0.33}$	$93.99_{\pm 0.18}$	$.398_{\pm .018}$	0.001	1	—	0.99
SGD	—	—	✗	$95.17_{\pm 0.09}$	$\mathbf{95.29}_{\pm 0.09}$	$.181_{\pm .003}$	0.001	1	—	0.7
Curv.-aware SGD (QHM)	—	—	✓	$95.00_{\pm 0.19}$	$95.02_{\pm 0.11}$	$.184_{\pm .002}$	0.003	1	0.5	0.9

sequentially. Final validation cross-entropy is evaluated on 100 million tokens; perplexity is its exponential.

Confirmation design and evaluation. We fix $\rho_0 = 0$ and momentum-difference mass adaptation with $\beta_3 = \kappa = \beta_{\text{mom}}$. The confirmation set contains 16 fixed configurations, each run with seeds 100–106: 96 fresh runs and 16 exploratory references. Seed 100 selected the configurations; seeds 101–106 are fresh initialization replicates with the same prepared data order. All metrics use the final 6,200-step checkpoint. We average per-run perplexities directly rather than exponentiating mean cross-entropy. At $\beta_{\text{mom}} = 0.975$, we compare $\zeta = 0$ and 0.05 at learning rates $\{0.012, 0.016, 0.024\}$, with an additional $\zeta = 0.025$ setting at $\eta = 0.012$; all use clipping at norm 1. The remaining settings in the table complete the fixed 16-configuration design.

Uncertainty and selection. Figure 2c shows arithmetic mean perplexity and pointwise 95% Student-t intervals over seven seeds ($\bar{x} \pm t_{0.975,6} s/\sqrt{7}$); these intervals are descriptive and do not adjust for configuration selection. Table entries give mean ± sample standard deviation. Inference uses two-sided paired tests of final validation-loss differences on seeds 101–106, with Holm correction over all ten planned contrasts. All four $\beta_{\text{mom}} = 0.975$ contrasts improve on every fresh seed and pass this correction (largest adjusted $p = 0.00104$). These tests assume independent initialization replicates and approximately normal paired differences.

The selected SS-AdamW baseline uses $\beta_{\text{mom}} = 0.95$ and $\eta = 0.008$, with clipping enabled. The M-Lion candidate chosen before replication uses $\beta_{\text{mom}} = 0.975$, $\eta = 0.012$, and $\zeta = 0.05$; it reaches 20.098 ± 0.085 perplexity and improves over the selected baseline on every fresh seed (1.08% lower fresh-seed mean). The $\eta = 0.016$ candidate was identified as the leader after seeds 101–104; seeds 105 and 106 both confirmed its ranking. Its gain over the selected baseline is 1.47% over seven seeds and 1.44% over the six fresh seeds. The gray curve in the main figure instead uses matched $\beta_{\text{mom}} = 0.975$ controls at each learning rate. Thus its larger paired gains measure the effect of dissipation at fixed momentum and learning rate, rather than the gain over the separately selected baseline. Results concern final validation performance under this model, data order, and training budget.

Table 4: Kinematics ablation for adaptive-mass M-Lion on CIFAR-10 (ResNet-18). We test whether the particular choice of relativistic kinematics, and hence the smooth clipping map, changes performance. We find no clear systematic difference: averaging matched configurations within each of five seeds gives 95% confidence intervals containing zero for all pairwise test-accuracy gaps. The coordinatewise update is $u_t = \phi(z)$, where $z = d_t/\rho_t$ is the normalized direction and $\rho_t = \sqrt{\rho_0^2 + a_t^2}$ includes rest and adaptive mass. All three maps have unit slope at the origin and saturate at ± 1. Each row selects the best of 12 repeated configurations by mean validation accuracy. Entries are mean $\pm$ sample standard deviation over seeds 0–4; all metrics use each seed's best-validation-accuracy checkpoint. η: peak learning rate; λ: weight decay.

		Validation		Test		Selected hyperparameters				
Kinematics	**Clipping** $\phi(z)$	Accuracy (%) ↑	Loss ↓	Accuracy (%) ↑	Loss ↓	ρ_0	η	λ	β_{mix}	β_{mom}
Minkowski	$\frac{z}{\sqrt{1+z^2}}$	$95.22_{\pm 0.18}$	$.191_{\pm .006}$	$95.23_{\pm 0.15}$	$.186_{\pm .004}$	1	0.001	1	0.7	0.95
Arctan	$\frac{2}{\pi}\arctan\left(\frac{\pi z}{2}\right)$	$95.14_{\pm 0.17}$	$.192_{\pm .002}$	$95.25_{\pm 0.08}$	$.183_{\pm .003}$	1	0.001	1	0.5	0.9
Tanh	$\tanh(z)$	$95.26_{\pm 0.21}$	$.188_{\pm .003}$	$95.14_{\pm 0.32}$	$.186_{\pm .006}$	1	0.001	1	0.5	0.9

Tables 4 and 5 compare Minkowski, arctan, and tanh maps with unit slope at the origin and saturation at ± 1, keeping adaptive mass enabled. The grid table specifies the two broad searches and seed extension, totaling 2,736 distinct runs. The repeated subset contains 12 configurations per map, each evaluated on seeds 0–4. Each table row selects a configuration by mean validation accuracy over those five seeds; these are descriptive results on the selection seeds, not fresh-seed confirmation of the chosen row. The separate paired comparison averages matched-configuration differences within each seed before forming the interval, so configurations from one seed are not treated as independent replicates. Intervals containing zero do not establish equivalence between maps.

These runs use the CIFAR-10 data and training recipe in section E.2: ResNet-18, batch size 256, 80,000 steps, cosine decay to 10^{-5} without warmup or gradient clipping, and evaluation every 5,000 steps. Updates are coordinatewise, with momentum-difference mass adaptation and no bias correction.

E.4 Rest mass across training cutoffs

Figure 5 uses 16 fixed-mass runs: $\rho_0 \in \{0, 10^{-5}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 100\}$ for each of M-Lion and M-Signum, with adaptive mass disabled, all at $\eta = 0.008$, weight decay 0.1, $\beta_{\text{mom}} = 0.9$, and seed 0. Mixing is 0.8 for M-Lion and 0.9 for M-Signum. For each mass and training cutoff, we select the highest-validation-accuracy checkpoint available so far and display its test accuracy. White outlines identify the validation-selected mass; exact ties favor the earlier checkpoint and then the smaller mass. Epochs are approximated by steps $\times 256/45{,}000$. These are prefixes of the original 80,000-step cosine schedule, not independently trained shorter schedules. The change in selected mass is a one-seed, fixed-setting diagnostic and does not by itself identify overfitting as its cause.

F Relating the curvature-aware dissipation to PID control

Proportional-Integral-Derivative (PID) control is one of the most widely used feedback control strategies in engineering practice (Minorsky, 1922; Ziegler & Nichols, 1942; Astrom & Hägglund, 2006). It is used to drive a dynamical system toward a desired reference state by applying a corrective control signal based on the tracking error. A standard PID controller computes:

$$u(t) \;=\; K_p e(t) \;+\; K_i \int_0^t e(s)\, ds \;+\; K_d \frac{de(t)}{dt}\,, \tag{50}$$

where the error signal $e(t)$ denotes the difference between the current state and the target.

In optimization, the analogue of the control objective is to reach a stationary point. Thus the "error" we seek to eliminate is the gradient itself, suggesting the identification $e(t) \equiv \boldsymbol{g}(t)$. And the analogue

Table 5: Hyperparameter search for the CIFAR-10 kinematics ablation. Each stage includes all three kinematics: Minkowski, arctan (equivalent to Refined Lion of Rong et al. (2025)), and tanh. Entries are fully crossed within each stage, except parameters explicitly grouped into pairs or triples. All runs use adaptive-mass M-Lion with coordinatewise updates, momentum-difference adaptation, and no bias correction. The search comprises 2,736 distinct training runs. The seed extension repeats two selected mass/learning-rate/weight-decay settings: its 144 runs combine with 36 matching seed-0 runs from the broad grids to give 36 configurations (12 per kinematics) across seeds 0–4, or 180 runs already included in the total. All runs use ResNet-18, batch size 256, 80,000 iterations, cosine learning-rate decay to 10^{-5} without warmup, and evaluation every 5,000 iterations.

Parameter	**Broad grid I**	**Broad grid II**	**Seed extension**
Mixing/decay pairs $(\beta_{\text{mix}}, \beta_{\text{mom}})$	$(0.5, 0.99)$, $(0.7, 0.99)$, $(0.9, 0.99)$	$(0.5, 0.95)$, $(0.7, 0.95)$, $(0.9, 0.95)$, $(0.5, 0.90)$, $(0.7, 0.90)$, $(0.8, 0.90)$	$(0.5, 0.90)$, $(0.7, 0.90)$, $(0.5, 0.95)$, $(0.7, 0.95)$, $(0.5, 0.99)$, $(0.7, 0.99)$
Rest mass ρ_0	$\{0, 10^{-5}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$	$\{0.1, 1, 10, 100\}$	*Paired* (ρ_0, η, λ): $(1, 0.001, 1)$, $(10, 0.008, 0.1)$
Peak learning rate η	$\{0.0001, 0.0005, 0.001, 0.003, 0.008, 0.016\}$	$\{0.001, 0.003, 0.008, 0.016\}$	
Weight decay λ	$\{0.01, 0.1, 1, 10\}$	$\{0.1, 1, 10\}$	
Training seeds	$\{0\}$	$\{0\}$	$\{1, 2, 3, 4\}$
New training runs	$3 \times 3 \times 8 \times 6 \times 4 = 1{,}728$	$3 \times 6 \times 4 \times 4 \times 3 = 864$	$3 \times 6 \times 2 \times 4 = 144$

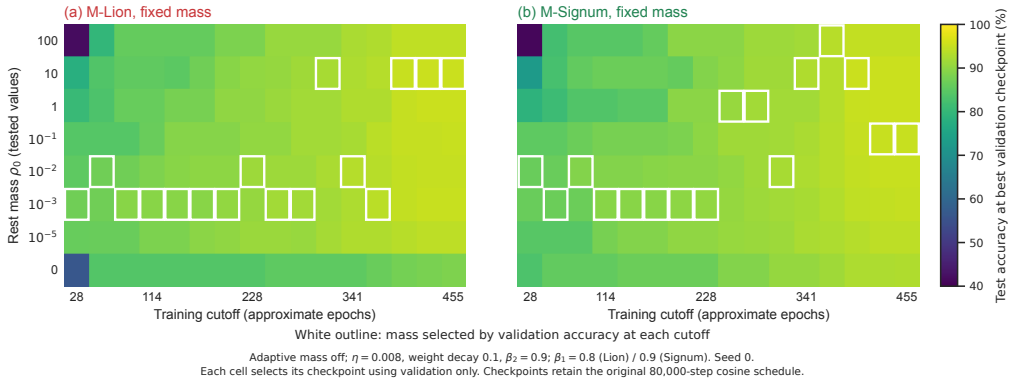

Figure 5: **Rest mass and training duration on CIFAR-10.** Colors show test accuracy at the checkpoint with the highest validation accuracy available up to each training cutoff, separately for each rest mass. White outlines mark the mass with the highest such validation accuracy in each column; test accuracy is never used for selection. Exact checkpoint ties retain the earlier checkpoint, and ties between masses favor the smaller mass. Panels show fixed-mass M-Lion (left) and M-Signum (right), with adaptive mass disabled ($\beta_3 = \kappa = 0$). All runs use ResNet-18, learning rate 0.008, weight decay 0.1, $\beta_{\text{mom}} = 0.9$, and seed 0; $\beta_{\text{mix}} = 0.8$ for M-Lion and 0.9 for M-Signum. Rows are the eight tested masses, with categorical spacing. Checkpoints are evaluated every 5,000 steps; approximate epochs equal steps $\times 256/45{,}000$. These are prefixes of the original 80,000-step cosine schedule, not separately trained shorter schedules. The figure is a single-seed diagnostic at fixed hyperparameters, without uncertainty estimates.

of the control signal is the applied force on the optimization dynamics, which in our formulation is the rate of change of momentum, $u(t) \equiv \dot{\boldsymbol{p}}(t)$.

The continuous master dynamics equation of M-Lion (eq. (8)) mirrors the PID control law:

$$\underbrace{\dot{\boldsymbol{p}}}_{\text{control } u(t)} = \underbrace{-\boldsymbol{g}}_{\text{Proportional (P)}} - \underbrace{\frac{1}{\tau_d}\boldsymbol{p}}_{\text{Integral (I)}} - \underbrace{\zeta\dot{\boldsymbol{g}}}_{\text{Derivative (D)}} . \tag{51}$$

Table 6: Mapping classical PID control (eq. (50)) to the master dynamics (eq. (8)).

PID concept	**Optimization counterpart**	**Physical and algorithmic role**
Error signal $e(t)$	Negative gradient $-\boldsymbol{g}(t)$	The discrepancy between the current slope and the target flat minimum.
Proportional (P)	Direct gradient force $-\boldsymbol{g}$	The immediate, memoryless reaction pushing the parameter state down the current slope.
Integral (I)	Isotropic friction $-\boldsymbol{p}/\tau_d$	Since momentum $\boldsymbol{p}$ integrates past forces, this term provides a leaky accumulation of past gradients, overcoming local noise and flat regions.
Derivative (D)	Curvature-aware dissipation $-\zeta\dot{\boldsymbol{g}}$	Anticipates changes in the gradient via the Hessian ($\mathbf{H}\dot{\boldsymbol{\theta}}$). It acts as an active braking mechanism, damping oscillations along steep valley walls.

This equivalence shows that the continuous-time formulation of M-Lion is not an ad hoc collection of moving averages, but a structurally complete PID controller operating directly on the momentum state (Table 6). Given this correspondence, no wonder Lion and M-Lion perform so well: through the combination of standard isotropic (I) and curvature-aware (D) dissipation, our derivation unifies all the theoretical siblings of M-Lion/Lion, including Hessian-driven damping (Alvarez et al., 2002; Attouch et al., 2016) and high-resolution ODEs (Shi et al., 2022), with control theory.

Thus, what began as a surprising artifact of automated search is now demystified: *Lion is simply a canonical PID controller operating in the massless limit within a relativistic Minkowski geometry*.