ADAPTIVE PRECONDITIONERS TRIGGER LOSS SPIKES IN ADAM

Anonymous authors
Paper under double-blind review

ABSTRACT

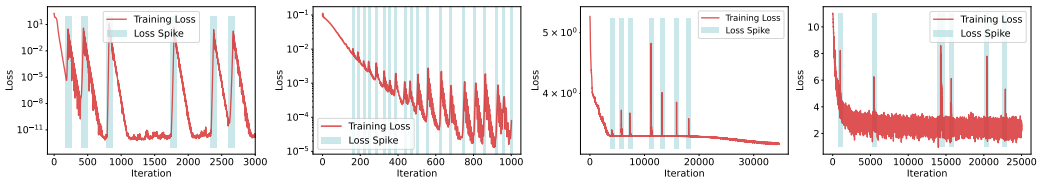
Loss spikes commonly emerge during neural network training with the Adam optimizer across diverse architectures and scales, yet their underlying mechanisms remain poorly understood. In this work, we investigate the fundamental causes of Adam spikes. While previous explanations attribute these phenomena to sharper loss landscapes at lower loss values, our analysis reveals that it is Adam's adaptive preconditioners that trigger spikes during training. We identify a key mechanism where the second moment estimate becomes insensitive to current gradients when using large β_2 values. This insensitivity can push the maximum eigenvalue of the preconditioned Hessian beyond the stability threshold $2/\eta$ for sustained periods, manifesting as dramatic loss spikes. We theoretically and experimentally characterize five distinct stages of spike evolution and propose a predictor for anticipating spikes based on gradient-directional curvature. We further validate our mechanism and demonstrate practical mitigation strategies from small fully connected networks to large-scale Transformers. These findings provide new theoretical insights for understanding and controlling loss spike behavior in Adam optimization.

1 INTRODUCTION

Neural network optimization remains a complex and sometimes unpredictable process despite significant advances in training methodologies. One particularly intriguing phenomenon that practitioners frequently encounter but rarely explore systematically is the "loss spike" — a sudden and sharp surge in the loss function that subsequently subsides. As illustrated in Fig. 1, these spikes differ markedly from normal fluctuations, resembling systematic instabilities rather than random noise. While observed across diverse architectures and datasets, their underlying mechanisms remain poorly understood. This creates a critical dilemma for practitioners: should they intervene to eliminate these apparent anomalies, or might loss spikes actually benefit the optimization process? Answering this question requires deeper theoretical understanding of when, how, and why loss spikes occur.

Previous research has tried to explain loss spikes through the geometry of loss landscapes (Ma et al., 2022a; Li et al., 2025). The lower-loss-as-sharper (LLAS) hypothesis (Li et al., 2025) suggests that regions of lower loss correspond to sharper curvature in the loss landscape, potentially causing instability. While this explanation provides some intuition, it fails to explain the specific behavior of adaptive optimizers like Adam (Kingma & Ba, 2014) that consistently exhibit spikes even in simple scenarios where landscape geometry is well-understood. For instance, as shown in Fig. 2(a), Adam produces loss spikes on a simple quadratic function even with learning rates well below theoretical stability thresholds, while gradient descent converges smoothly. This behavior can not be explained by loss landscape alone, since quadratic functions have constant curvature. Furthermore, although previous research has identified the Edge of Stability (EoS) phenomenon, where loss decreases non-monotonically while the largest Hessian eigenvalue hovers around $2/\eta$ (η is the learning rate) (Cohen et al., 2021; Wu et al., 2018; Xing et al., 2018; Ahn et al., 2022; Lyu et al., 2022; Arora et al., 2022; Wang et al., 2022; Cohen et al., 2023), loss spikes appear to represent more dramatic instabilities than typical EoS behavior. In particular, the precise relationship between these instabilities and observed spikes remains unclear—instability may sometimes manifest as oscillations and sometimes as spikes (Ma et al., 2022b), the specific mechanism under which spikes occur is not well understood.

In this work, we present a detailed mechanistic explanation for loss spikes in Adam optimization. Our key finding is that these spikes arise not primarily from the complex geometry of the loss landscape,



(a) FNN (b) CNN (c) Transformer (10M) (d) Transformer (187M)

Figure 1: Loss spikes across architectures: (a) FNNs for function approximation. (b) CNNs on CIFAR10. (c-d) Transformers on language tasks. See experimental details in Appendix G.

but rather from the intrinsic dynamics of Adam's adaptive preconditioners. Specifically, we demonstrate both theoretically and experimentally that Adam's stability is governed by a preconditioned Hessian. When using large values of β_2 (as is common in practice), the second moment estimate becomes insensitive to current gradients, causing the maximum eigenvalue of the preconditioned Hessian to exceed the stability threshold $2/\eta$ for sustained periods. This creates severe instability that manifests as dramatic loss spikes. The instability further induces alignment between the gradient and the maximum eigendirection, with loss spikes occurring precisely when the gradient-directional curvature exceeds $2/\eta$. We find that directly reducing β_2 is effective in mitigating loss spikes.

Our main contributions are summarized as follows:

(i) We show that it is Adam's adaptive preconditioners that causes spikes in practical Adam training. The five stages of spike evolution are clearly characterized, both theoretically and experimentally. This mechanism is distinct from previous lower-loss-as-sharper (LLAS) landscape hypothesis (Li et al., 2025) (please refer to Sec. 3, Sec. 4.1 and Sec. 5).

(ii) We identify a key mechanism whereby the second moment estimate becomes insensitive to current gradients when employing a relatively large β_2. This causes the maximum eigenvalue of the preconditioned Hessian to **persistently** exceed the classical stability threshold $2/\eta$, manifesting as dramatic loss spikes. (please refer to Sec. 4.1, Sec. 4.2, and Sec. 6).

(iii) We propose a predictor, $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ for anticipating spikes based on the curvature in the gradient direction. We empirically show that this predictor is highly accurate in forecasting spike onset, and we further validate practical strategies for mitigating spikes. (please refer to Sec. 4.3 and Sec. 6).

2 Related Works

Edge of Stability (EoS). Various works (Cohen et al., 2021; Wu et al., 2018; Xing et al., 2018; Ahn et al., 2022; Lyu et al., 2022; Arora et al., 2022; Jastrzebski et al., 2020; Jastrzębski et al., 2019; Lewkowycz et al., 2020) have investigated the *Edge of Stability* (EoS), a phenomenon where gradient descent progressively increases the sharpness of the loss landscape—a process known as *progressive sharpening*—until the maximum Hessian eigenvalue stabilizes near the threshold $2/\eta$, while the loss continues to decrease non-monotonically. Ma et al. (2022a) proposed a subquadratic structure near local minima, where sharpness increases when the loss decreases along the gradient direction, providing a theoretical account of this behavior. Other studies (Damian et al., 2023; Wang et al., 2022) show that when $\lambda_{\max} > 2/\eta$, self-stabilization mechanisms can reduce sharpness and restore stability. More recently, Cohen et al. (2023) extended the EoS framework to adaptive optimizers, introducing the concept of *Adaptive Edge of Stability* (AEoS). Furthermore, Cohen et al. (2025) also developed the concept of central flow to study the average trajectory of oscillatory dynamics during EoS. While EoS has been widely explored, its direct association with loss spikes has yet to be thoroughly investigated.

Convergence Analysis of Adam. Numerous works have analyzed the convergence behavior of adaptive gradient methods (Chen et al., 2019; Li & Orabona, 2019; Xie et al., 2020; Défossez et al., 2022; Da Silva & Gazeau, 2020; Shi et al., 2021; Zou et al., 2019; Zhou et al., 2024). In particular, Reddi et al. (2018) demonstrated that Adam may fail to converge even in simple convex settings, prompting a series of variants (Liu et al., 2019; Taniguchi et al., 2024). Zhang et al. (2022) showed

that in the case of learning rate decay Adam can converge to a neighborhood of critical points when β_2 is large, and this convergence is guaranteed if $\beta_1 < \sqrt{\beta_2}$.

Loss Spike Analysis. Chowdhery et al. (2023) reported that restarting training from an earlier checkpoint and skipping the spiking data batch can mitigate spikes in large models. Molybog et al. (2023) found that the gradient and second-moment estimates of shallow layer parameters can decay to near-zero and then spike upon encountering a large gradient. Li et al. (2025) argued that spikes occur in sharp regions of the loss landscape with a lower-loss-as-sharper (LLAS) structure. Ma et al. (2022b) qualitatively demonstrated that Adam's hyperparameters impact the occurrence of spikes or oscillations. Although previous studies have uncovered parts of the puzzle surrounding spikes, this work provides a more detailed and comprehensive understanding of the spike formation.

3 Distinct Loss Spike Mechanism in Adam and Gradient Descent

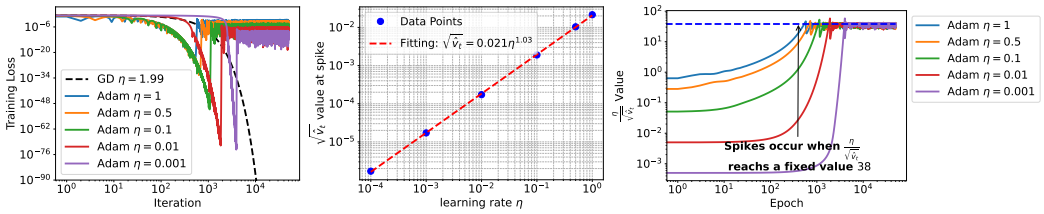
Adam Algorithm. The Adam algorithm is widely used in training Transformer models and is widely observed to be more prone to cause loss spikes. Adam maintains exponential moving averages of gradients (first moment) and squared gradients (second moment) to speed up training:

$$\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t, \quad \boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2, \tag{1}$$

where $\boldsymbol{g}_t := \nabla L(\boldsymbol{\theta}_t)$ is the gradient, and $\beta_1, \beta_2 \in [0, 1)$ are hyperparameters controlling the exponential decay rates (default values: $\beta_1 = 0.9, \beta_2 = 0.999$). To counteract the initialization bias toward zero, these moments are corrected: $\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-\beta_1^t}, \quad \hat{\boldsymbol{v}}_t = \frac{\boldsymbol{v}_t}{1-\beta_2^t}$. The parameter update rule is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon}, \tag{2}$$

where $\eta > 0$ is the learning rate and $\varepsilon > 0$ is a small constant (default 10^{-8} in PyTorch).



(a) Loss (b) $\sqrt{\hat{v}_t} \propto \eta$ (c) Evolution of $\frac{\eta}{\sqrt{\hat{v}_t}}$

Figure 2: Optimization of $f(\theta) = \frac{1}{2}\theta^2$. (a) Loss trajectories during Adam and GD training across various learning rates. Curves of different colors represent Adam's training loss, which initially decreases steadily before abruptly spiking to significantly higher values. (b) The relationship between learning rate and $\sqrt{\hat{v}_t}$ value at spike occurrence follows a power law, appearing as a straight line with a slope of approximately 1 in log-log scale. (c) Under different learning rates, the ratio $\eta/\sqrt{\hat{v}_t}$ consistently reaches a nearly identical threshold value immediately before the loss begins to spike.

Differences in Spike Behavior Between GD and Adam. Adaptive methods like Adam exhibit fundamentally different behavior compared to standard gradient descent (GD). A notable distinction is that Adam can encounter convergence difficulties even with simple quadratic functions and very small learning rates. For the quadratic function $f(\theta) = \frac{1}{2}\theta^2$, it is well established that gradient descent converges when the learning rate $\eta < 2/\lambda_{\max} = 2$ (depicted by the black dashed line in Fig. 2(a)). However, Adam displays more intricate dynamics. As illustrated in Fig. 2(a), Adam with a learning rate $\eta \ll 2$ (using hyperparameters $\beta_1 = 0.9, \beta_2 = 0.99, \varepsilon = 10^{-8}$) still fails to converge. This non-convergence manifests in the distinctive colored curves in Fig. 2(a), where the training loss initially decreases steadily before abruptly spiking to a substantially higher magnitude. Fig. 2(b) further examines the relationship between Adam's second moment $\sqrt{\hat{v}_t}$ at spike occurrence and learning rate. From Fig. 2(b), we observe that smaller learning rates correspond to smaller $\sqrt{\hat{v}_t}$ values when spikes occur, with the relationship appearing linear in log-log scale with a slope near 1. For one-dimensional quadratic optimization, $\eta/\sqrt{\hat{v}_t}$ can be interpreted as the effective learning rate

and it increases as training progresses because $\sqrt{\hat{v}_t}$ diminishes alongside the gradient g_t according to Eq. (1). Experimentally, Fig. 2(c) confirms that this ratio increases until reaching a nearly consistent threshold value 38 (see Prop. 2 for a theoretical explanation), at which point the loss spike invariably occurs. While straightforward, this analysis provides valuable intuition for the emergence of spikes. However, it is important to note that in high-dimensional optimization scenarios, $\sqrt{\hat{\boldsymbol{v}}_t}$ becomes a vector rather than a scalar, rendering the notion of an effective learning rate inapplicable. In the following section, we will quantitatively characterize Adam's spike behavior in more general settings.

4 Loss Spike Analysis of Adam

Quadratic Approximation. To understand the mechanics behind loss spikes, we begin with a linear stability analysis that connects optimization dynamics to the geometry of the loss landscape. Consider optimizing a loss function $L(\boldsymbol{\theta})$ with respect to parameters $\boldsymbol{\theta} \in \mathbb{R}^M$. Around any point $\boldsymbol{\theta}_0$, we can approximate the loss using a second-order Taylor expansion:

$$L(\boldsymbol{\theta}_0 + \delta\boldsymbol{\theta}) \approx \tilde{L}(\delta\boldsymbol{\theta}) := L(\boldsymbol{\theta}_0) + \nabla L(\boldsymbol{\theta}_0)^\top \delta\boldsymbol{\theta} + \tfrac{1}{2}\delta\boldsymbol{\theta}^\top \boldsymbol{H}\delta\boldsymbol{\theta}, \tag{3}$$

where $\nabla L(\boldsymbol{\theta}_0)$ is the gradient and $\boldsymbol{H} := \boldsymbol{H}(\boldsymbol{\theta}_0) = \nabla^2 L(\boldsymbol{\theta}_0)$ is the Hessian matrix at $\boldsymbol{\theta}_0$.

Stability Analysis. For GD with learning rate η, the parameter update is: $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\nabla L(\boldsymbol{\theta}_t)$. Using the quadratic approximation from Eq. (3), the displacement $\delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_0$ evolves as:

$$\delta\boldsymbol{\theta}_{t+1} \approx \delta\boldsymbol{\theta}_t - \eta\nabla\tilde{L}(\delta\boldsymbol{\theta}_t) = \delta\boldsymbol{\theta}_t - \eta(\nabla L(\boldsymbol{\theta}_0) + \boldsymbol{H}\delta\boldsymbol{\theta}_t) = (\boldsymbol{I} - \eta\boldsymbol{H})\delta\boldsymbol{\theta}_t - \eta\nabla L(\boldsymbol{\theta}_0).$$

The optimization becomes unstable along the maximum eigendirection when $\lambda_{\max}(\boldsymbol{H}) > 2/\eta$.

Practical Stability Condition. In neural network optimization, the loss landscape—and consequently the Hessian matrix—evolves continuously as parameters are updated. The local Hessian stability condition ensures stable loss decrease at each iteration, as formalized below.

Proposition 1 (see Appendix D Prop. D.1 for proof). *Let $L : \mathbb{R}^M \to \mathbb{R}$ be twice continuously differentiable. For any iterate $\boldsymbol{\theta}_t$ define the gradient $\boldsymbol{g}_t := \nabla L(\boldsymbol{\theta}_t)$ and, for a fixed learning rate $\eta > 0$, define the local directional maximum Hessian $\bar{\lambda}_t := \sup_{s\in[0,1]} \lambda_{\max}\big(\nabla^2 L(\boldsymbol{\theta}_t - s\eta\boldsymbol{g}_t)\big)$, the maximum eigenvalue of the Hessian along the line segment from $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\boldsymbol{g}_t$. If $\eta < \frac{2}{\bar{\lambda}_t}$, the gradient descent step $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\boldsymbol{g}_t$ satisfies the descent estimate:*

$$L(\boldsymbol{\theta}_{t+1}) \le L(\boldsymbol{\theta}_t) - \eta\Big(1 - \frac{\eta\bar{\lambda}_t}{2}\Big)\|\boldsymbol{g}_t\|^2.$$

In particular, whenever $\eta \in \big(0, 2/\bar{\lambda}_t\big)$ and $\boldsymbol{g}_t \neq 0$ we have strict decrease $L(\boldsymbol{\theta}_{t+1}) < L(\boldsymbol{\theta}_t)$.

In practice, since learning rates are typically small, we can monitor the step-wise stability condition $\lambda_{\max}(\boldsymbol{H}_t) \le 2/\eta$ as a proxy. When this condition is persistently violated, there is likely a loss spike.

4.1 Adam's Preconditioned Hessian and Stability

Stability Analysis of Adaptive Mechanism. To analyze Adam's stability conditions, we first examine the adaptive mechanism by setting $\beta_1 = 0$, ignoring momentum effects. Following the Taylor expansion approach from Eq. (3), we have:

$$\delta\boldsymbol{\theta}_{t+1} \approx \delta\boldsymbol{\theta}_t - \eta\frac{\nabla\tilde{L}(\delta\boldsymbol{\theta}_t)}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon} = \left(\boldsymbol{I} - \eta\mathrm{diag}\left(\frac{1}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon}\right)\boldsymbol{H}\right)\delta\boldsymbol{\theta}_t - \eta\frac{\nabla L(\boldsymbol{\theta}_0)}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon}.$$

Stability requires the spectral radius $\rho\left(\boldsymbol{I} - \eta\hat{\boldsymbol{H}}\right) < 1$, where $\hat{\boldsymbol{H}} = \mathrm{diag}((\sqrt{\hat{v}_t} + \varepsilon)^{-1})\boldsymbol{H}$ is the "adaptive preconditioned Hessian". Although asymmetric, $\hat{\boldsymbol{H}}$ can be diagonalized with real eigenvalues (see Appendix D Lem. D.1), yielding the stability condition $\lambda_{\max}(\hat{\boldsymbol{H}}) < 2/\eta$.

Stability Analysis of Momentum Mechanism. With momentum ($\beta_1 > 0$), we analyze the update rule $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\boldsymbol{m}_t$. Following the same Taylor expansion approach: $\delta\boldsymbol{\theta}_{t+1} \approx \delta\boldsymbol{\theta}_t - \eta(\beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)(\nabla L(\boldsymbol{\theta}_0) + \boldsymbol{H}\delta\boldsymbol{\theta}_t))$. Substituting $\eta\boldsymbol{m}_{t-1} = \delta\boldsymbol{\theta}_{t-1} - \delta\boldsymbol{\theta}_t$ gives:

$$\delta\boldsymbol{\theta}_{t+1} \approx [(1+\beta_1)\boldsymbol{I} - \eta(1-\beta_1)\boldsymbol{H}]\,\delta\boldsymbol{\theta}_t - \beta_1\delta\boldsymbol{\theta}_{t-1} - \eta(1-\beta_1)\nabla L(\boldsymbol{\theta}_0). \tag{4}$$

Proposition 2 (see Appendix D Prop. D.2 for proof). *Consider the three-term recursive iteration*

$$\delta\boldsymbol{\theta}_{t+1} = \left[(1+\beta_1)\boldsymbol{I} - \eta(1-\beta_1)\boldsymbol{H}(\boldsymbol{\theta}_0)\right]\delta\boldsymbol{\theta}_t - \beta_1\delta\boldsymbol{\theta}_{t-1} - \eta(1-\beta_1)\nabla L(\boldsymbol{\theta}_0),$$

with learning rate $\eta > 0$ and momentum parameter $\beta_1 \in [0, 1)$. Then the linearized system at $\boldsymbol{\theta}_0$ is asymptotically stable in all positive-curvature eigendirections (i.e., for every eigenvalue $\lambda_i > 0$ the characteristic roots lie strictly inside the unit disk) if and only if

$$\lambda_{\max}\left(\frac{1-\beta_1}{1+\beta_1}\boldsymbol{H}(\boldsymbol{\theta}_0)\right) < \frac{2}{\eta},$$

where $\lambda_{\max}(\cdot)$ denotes the largest positive eigenvalue.

Comprehensive Stability Analysis of Adam. Integrating both mechanisms and the momentum bias correction $\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1-\beta_1^t}$, the comprehensive "Adam preconditioned Hessian[1]" becomes:

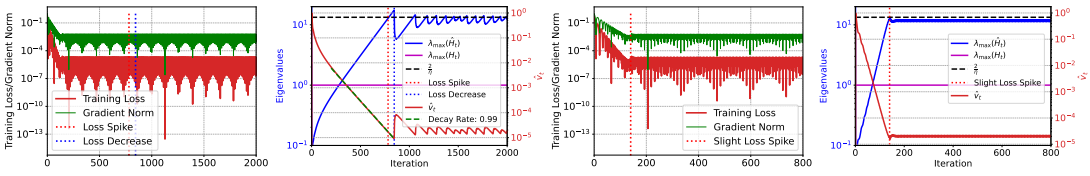
$$\hat{\boldsymbol{H}}_t = \frac{1}{1-\beta_1^t}\frac{1-\beta_1}{1+\beta_1}\mathrm{diag}\left(\frac{1}{\sqrt{\hat{\boldsymbol{v}}_t}+\varepsilon}\right)\boldsymbol{H}_t. \tag{5}$$

In Sec. 4.2, we experimentally validate that this modified step-wise instability criterion $\lambda_{\max}(\hat{\boldsymbol{H}}_t) > 2/\eta$ accurately predicts loss spikes in one-dimensional scenarios.

4.2 Sustained Decay of Second-Order Moment Triggers Loss Spikes

The key difference between gradient descent and Adam stability lies in Adam's adaptive preconditioners $\boldsymbol{v}_t$. To investigate how the decay behavior of $\boldsymbol{v}_t$ affects loss spikes, we conducted controlled experiments on a simple quadratic function $f(\theta) = \frac{1}{2}\theta^2$.

Large β_2 Causes Sustained Instability and Spikes. Fig. 3(a–b) shows results with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Initially, loss decreases gradually until epoch 782, when a spike occurs precisely as $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ exceeds the threshold $2/\eta$. The mechanism works as follows: Before the spike, the gradient norm (green line, $\approx 10^{-15}$) becomes much smaller than $\sqrt{\hat{v}_t}$ (red line, $\approx 10^{-1}$). According to Eq. (1), this causes v_t to decay exponentially as $v_t \approx \beta_2 v_{t-1}$. The green dashed line in Fig. 3(b) fits this decay with $\hat{v}_t = A\alpha^t$, confirming $\alpha \approx \beta_2 = 0.99$. When $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ surpasses $2/\eta$, the loss spikes and gradient norm increases. However, due to the large β_2, v_t responds sluggishly to current gradients, allowing the exponential decay to continue. This maintains $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ above the stability threshold, sustaining the spike until epoch 845, when the gradient grows large enough to increase $\hat{v}_t$. This causes $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ to drop back below $2/\eta$, and the loss begins to decrease again at epoch 845.



(a) $\beta_2 = 0.99$ (b) $\beta_2 = 0.99$ (c) $\beta_2 = 0.9$ (d) $\beta_2 = 0.9$

Figure 3: Adam optimization on $f(\theta) = \frac{1}{2}\theta^2$ with different β_2 values. (a, c) Evolution of training loss and gradient norm. (b, d) Evolution of the second moment estimate $\hat{\boldsymbol{v}}_t$ and the maximum eigenvalue of the preconditioned Hessian. The red dotted line marks the onset of the loss spike, while the blue dotted line indicates the point where the loss begins to decrease. The green dashed lines fit $\hat{\boldsymbol{v}}_t$ decay using $\hat{\boldsymbol{v}}_t = A\alpha^t$ with decay rate shown in the labels.

Small β_2 Prevents Sustained Instability. Fig. 3(c–d) shows results with $\beta_1 = 0.9$ and $\beta_2 = 0.9$—a configuration less commonly used in practice due to its inferior convergence guarantees (Shi et al., 2021; Zhang et al., 2022). Here, the gradient remains non-negligible relative to $\sqrt{v_t}$ throughout

[1]This preconditioner jointly incorporates the effects of β_1 and β_2, unifying the stability threshold at $\frac{2}{\eta}$. While the formulation differs slightly from that in Cohen et al. (2023), the two definitions are essentially equivalent.

training, preventing pure β_2-exponential decay (the observed decay rate $\alpha \approx 0.93$ exceeds $\beta_2 = 0.9$). As training progresses and gradients diminish, $\hat{v}_t$ decreases and $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ gradually increases. However, when $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ reaches $2/\eta$, the responsive v_t immediately adjusts to the rising gradient, preventing sustained instability. Instead of large spikes, we observe minor oscillations (Fig. 3(c)). An extreme case is to set $\beta_1 = \beta_2 = 0$, then Adam becomes "signGD" and spike never occurs. This helps explain why Adam training, as empirically observed by Ma et al. (2022b), sometimes results in sudden spikes in loss and sometimes in oscillatory behavior.

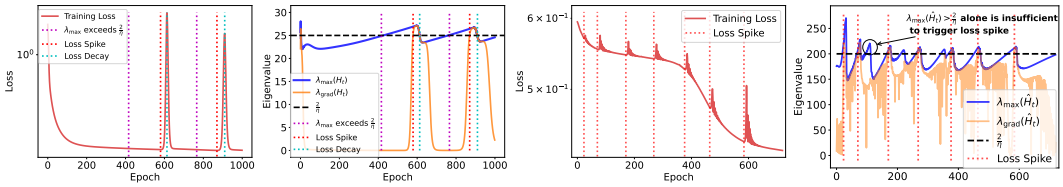
4.3 Precise Loss Spike Prediction via Gradient-Directional Curvature

In high-dimensional optimization, when $\lambda_{\max}(\boldsymbol{H}_t) > 2/\eta$, instability occurs primarily along the corresponding unstable eigendirection, while other directions may remain stable. As shown in our 2-dimension experiments with gradient descent (Figures D4 and D5), even when $\lambda_{\max}(\boldsymbol{H}_t)$ exceeds $2/\eta$, the loss can continue decreasing normally for some time until oscillations along the unstable direction grow sufficiently large to cause the loss to increase. Consequently, exceeding $\lambda_{\max}(\boldsymbol{H}_t) > 2/\eta$ does not immediately trigger a spike. To precisely predict loss spikes, we analyze the loss change between consecutive steps using a second-order Taylor expansion: $L(\boldsymbol{\theta}_{t+1}) \approx L(\boldsymbol{\theta}_t) + \nabla L(\boldsymbol{\theta}_t)^\top(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)^\top \boldsymbol{H}_t(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t)$. Substituting the gradient descent update $\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = -\eta \nabla L(\boldsymbol{\theta}_t)$: $L(\boldsymbol{\theta}_{t+1}) - L(\boldsymbol{\theta}_t) \approx -\eta\|\nabla L(\boldsymbol{\theta}_t)\|^2 + \frac{1}{2}\eta^2 \nabla L(\boldsymbol{\theta}_t)^\top \boldsymbol{H}_t \nabla L(\boldsymbol{\theta}_t)$. A loss increase (necessary for a loss spike) occurs when this expression is positive, yielding the condition (see Theorem D.1 for a general result and rigorous proof):

$$\lambda_{\text{grad}}(\boldsymbol{H}_t) := \frac{\nabla L(\boldsymbol{\theta}_t)^\top \boldsymbol{H}_t \nabla L(\boldsymbol{\theta}_t)}{\|\nabla L(\boldsymbol{\theta}_t)\|^2} > \frac{2}{\eta}. \tag{6}$$

Here, $\lambda_{\text{grad}}(\boldsymbol{H}_t)$ represents the curvature along the gradient. For Adam, we define the analogous predictor as $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t) := \frac{\nabla L(\boldsymbol{\theta}_t)^\top \hat{\boldsymbol{H}}_t \nabla L(\boldsymbol{\theta}_t)}{\|\nabla L(\boldsymbol{\theta}_t)\|^2}$, where $\hat{\boldsymbol{H}}_t$ is the preconditioned Hessian from Eq. (5).

Experimental Verification of Loss Spike Predictor. We validate our predictor using a two-layer network trained on 20 data points to fit $f(x) = \sin(x) + \sin(4x)$. We track both $\lambda_{\max}(\boldsymbol{H}_t)$ and $\lambda_{\text{grad}}(\boldsymbol{H}_t)$ during training. For gradient descent (Fig. 4(a–b)), two loss spikes occur. At epoch 416, although $\lambda_{\max}(\boldsymbol{H}_t)$ exceeds $2/\eta$, loss continues decreasing. The spike occurs only when $\lambda_{\text{grad}}(\boldsymbol{H}_t)$ also exceeds $2/\eta$. For Adam (Fig. 4(c–d)), 7 distinct spikes occur, while $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ exceeds $2/\eta$ at 10 time steps. Crucially, spikes occur only when $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t) > 2/\eta$, confirming that $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ alone is insufficient for spike prediction.



(a) GD (loss) (b) GD (eigenvalues) (c) Adam (loss) (d) Adam (eigenvalues)

Figure 4: Experimental validation of the gradient-directional loss spike predictor. A two-layer fully connected neural network (width 20) is trained on 200 randomly sampled data points to fit $f(x) = \sin(x) + \sin(4x)$. (a–b) Gradient descent with learning rate $\eta = 0.08$. (c–d) Adam with learning rate $\eta = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

5 Five-Stage Characterization for Loss Spike Mechanics in Adam

Building on our theoretical and empirical findings, we conjecture a five-stage progression that characterizes how loss spikes form and resolve during Adam optimization (Fig. 5).

Stage 1: Stable Loss Decrease. Training loss decreases steadily with no abnormalities observed.

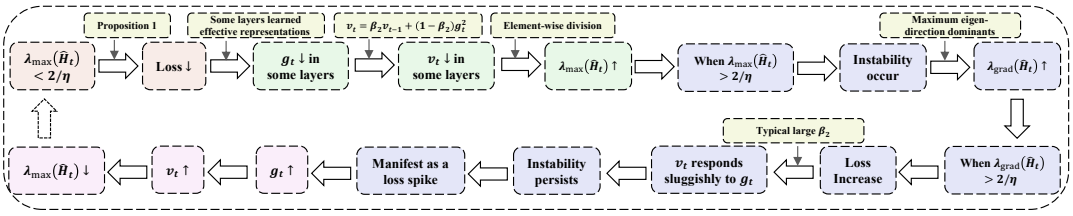


Figure 5: Five-stage progression for loss spike mechanics in Adam.

Stage 2: Preconditioner Decay. As training progresses, gradients in some layers diminish as effective representations are learned. The corresponding second moment estimates v_t also decrease. Due to the element-wise division in Eq. (5), this causes $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ to gradually increase.

Stage 3: Spike Onset. Instability begins when $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ exceeds the stability threshold $2/\eta$. Initially localized, the instability intensifies as the gradient aligns with max eigen-direction. A loss increase occurs only when the gradient curvature $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ also exceeds $2/\eta$. With typical large values $\beta_2 \in [0.95, 0.9999]$, the second moment $\boldsymbol{v}_t$ responds sluggishly to gradient information, causing $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ to persistently exceed $2/\eta$ and thus manifesting as a dramatic loss spike.

Stage 4: Preconditioner Growth. As the spike intensifies, gradients grow larger. When gradients become sufficiently large to influence $\boldsymbol{v}_t$, the decay of $\boldsymbol{v}_t$ halts and reverses. This growth in $\boldsymbol{v}_t$ reduces $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$, helping restore stability.

Stage 5: Loss Decrease. When $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ falls below $2/\eta$, the optimizer regains stability. Loss resumes decreasing, completing the spike cycle and returning to Stage 1.

These five stages provide an intuitive understanding of the Adam loss spike phenomenon. We also provide a rigorous mathematical five-stage characterization for quadratic optimization:

Theorem 1 (**Five Stages of Adam for Quadratic Optimization** (see Appendix D Thm. D.2 and Fig. D1 for details and proof)). *Consider the 1D loss $L(\theta) = \frac{1}{2}\theta^2$, optimized using Adam with $\beta_1 = 0$, $\beta_2 \in (0,1)$, and $\eta > 0$. The update rules are: $\theta_{t+1} = \left(1 - \frac{\eta}{\sqrt{v_t}}\right)\theta_t, \quad v_{t+1} = \beta_2 v_t + (1-\beta_2)\theta_t^2$. Assume $v_0 = \theta_0^2$ and $|\theta_0| > \frac{\eta}{2}$. Then there exist integers $t_0 < t_1 < t_2 < t_3 < t_4 < t_5 < \infty$ such that the iterates (θ_t, v_t) exhibit the five stages described above in intervals $[t_i, t_{i+1})$, respectively.*

Furthermore, we show that common learning rate decay strategies are insufficient to avoid this unstable behavior for sufficiently large β_2, suggesting its inevitability:

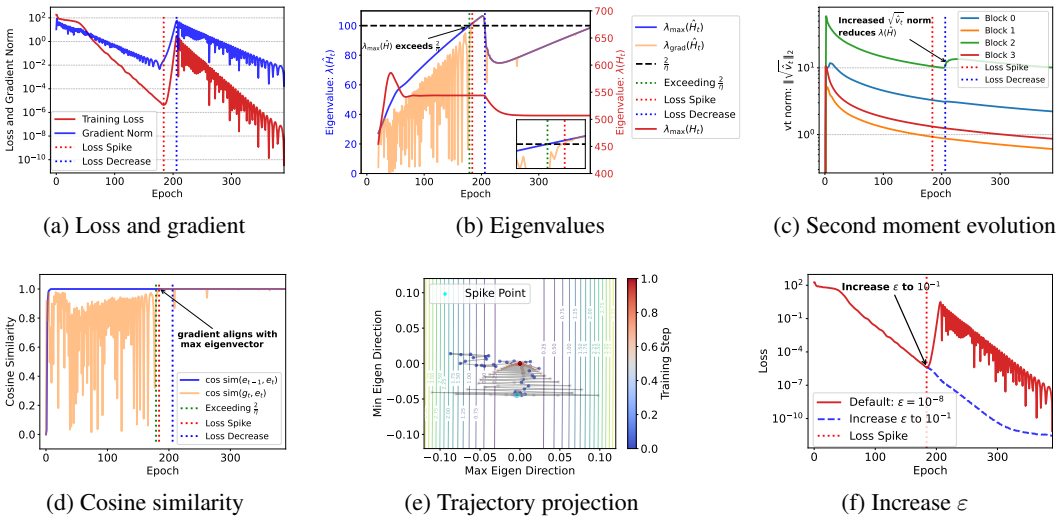
Theorem 2 (**Decaying Learning Rate Scheduler** (see Appendix D Thm. D.3 for proof)). *Consider the same setup as Thm. 1 with decaying learning rate $\eta_t = \eta_0(t+1)^{-\alpha}$ where $\alpha \in (0,1)$. Assume the initialization satisfies $v_0 = \theta_0^2$ and $|\theta_0| > 2\eta_0 > 0$. Assume β_2 is sufficiently close to 1. Then the stability condition $|1 - \frac{\eta_t}{\sqrt{v_t}}| < 1$ cannot hold for all $t \in \mathbb{N}^+$.*

6 Empirical Validation of Loss Spike Mechanics in Adam

To empirically validate the proposed loss spike mechanics in realistic, high-dimensional settings, we conduct comprehensive experiments across various neural network architectures and optimization tasks. We implement efficient Hessian-vector products for eigenvalue computation to track the theoretical indicators proposed in our conjecture. This method allows us to calculate $\lambda_{\max}$ and λ_{grad} by obtaining the product of the Hessian and a vector, without computing the full Hessian matrix. Detailed experimental configurations are provided in Appendix G, with additional validation experiments (including CNNs model) in Appendix F.

6.1 Fully Connected Neural Networks for Function Approximation

We trained a two-layer fully connected network on a 50-dimensional function approximation task using Adam with $\beta_1 = 0.9, \beta_2 = 0.999$. The optimization dynamics mirror our quadratic analysis: both loss and gradient norm decrease rapidly before experiencing a sharp spike (Fig. 6(a)).



(a) Loss and gradient (b) Eigenvalues (c) Second moment evolution

(d) Cosine similarity (e) Trajectory projection (f) Increase ε

Figure 6: (a) Training loss and gradient norm over time. (b) Evolution of critical eigenvalues: original Hessian maximum eigenvalue $\lambda_{\max}(\boldsymbol{H}_t)$, preconditioned Hessian maximum eigenvalue $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ and gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ relative to $2/\eta$. (c) L_2-norm of second moment $||\sqrt{\hat{\boldsymbol{v}}_t}||_2$ of different parameter blocks during training. (d) Cosine similarity between maximum eigenvectors in two consecutive epochs (blue) and between gradient and current maximum eigenvector (orange). (e) Training trajectory projected onto maximum and minimum Hessian eigenvectors at epoch 390. The colorbar for training steps is normalized to the range $[0, 1]$, where 0 corresponds to epoch 28 and 1 corresponds to epoch 390. (f) Increase the default ε in Eq. (2) to 0.1 at epoch 184.

Eigenvalue Evolution and Spike Timing: Fig. 6(b) shows that $\lambda_{\max}(\boldsymbol{H}_t)$ stabilizes quickly while $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ continues increasing due to decreasing $\boldsymbol{v}_t$ (Fig. 6(c)). Crucially, although $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ surpasses the stability threshold $2/\eta$ at epoch 179, the spike occurs precisely at epoch 184 when $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ exceeds $2/\eta$, confirming our directional stability analysis in Sec. 4.3.

Second Moment $\boldsymbol{v}_t$ Dynamics: Fig. 6(c) shows the evolution of second-moment norms $\sqrt{\hat{\boldsymbol{v}}_t}$ for each parameter block. Before the spike, the gradient norm $\|\boldsymbol{g}_t\| \approx 10^{-2}$ becomes much smaller than $\|\sqrt{\hat{\boldsymbol{v}}_t}\|$, causing $\boldsymbol{v}_t$ to decay exponentially at rate β_2. During the spike, gradient norms increase while $\hat{\boldsymbol{v}}_t$ continues decreasing due to its sluggish response. Once gradients become sufficiently large, $\boldsymbol{v}_t$ rises rapidly, driving $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ below $2/\eta$ and allowing loss descent to resume at epoch 206.

Validation of Quadratic Analysis. The cosine similarity between maximum eigenvectors of $\boldsymbol{H}_t$ across consecutive steps approaches 1 early in training (Fig. 6(d)), validating our quadratic analysis. Fig. 6(e) confirms that spikes occur when gradients align with the maximum curvature direction by projecting the trajectory onto maximum and minimum eigenvectors. To suppress the spike, a straightforward method involves increasing ε in Eq. (2). As demonstrated in Fig. 6(f), increasing ε to 0.1 at spike onset effectively eliminates the instability.

6.2 Transformer Models for Language Tasks

We trained an 8-layer Transformer (approximately 10 million parameters) on a synthetic dataset of 900k sequences (batch size 2048) for compositional rule learning under the next-token prediction paradigm. Fig. 7(a) shows seven distinct loss spikes (blue regions). Prior to each spike, the norm of the second-moment estimate $\hat{\boldsymbol{v}}_t$ for the embedding and $\boldsymbol{W}_V$ parameters across attention layers decays at a rate of approximately 0.999003 (close to β_2), followed by a sudden increase in $\|\hat{\boldsymbol{v}}_t\|$ and a sharp drop in loss. Fig. 7(b) describes a typical case where $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ exceeds $2/\eta$ causing a spike. However, it is important to note that stochastic batching introduces significant noise, making precise spike prediction challenging. To address this, we define a "sustained spike predictor" as: $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)(\text{sustained}) = \min(\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_{t-1}), \lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t), \lambda_{\text{grad}}(\hat{\boldsymbol{H}}_{t+1}))$. This refined predictor

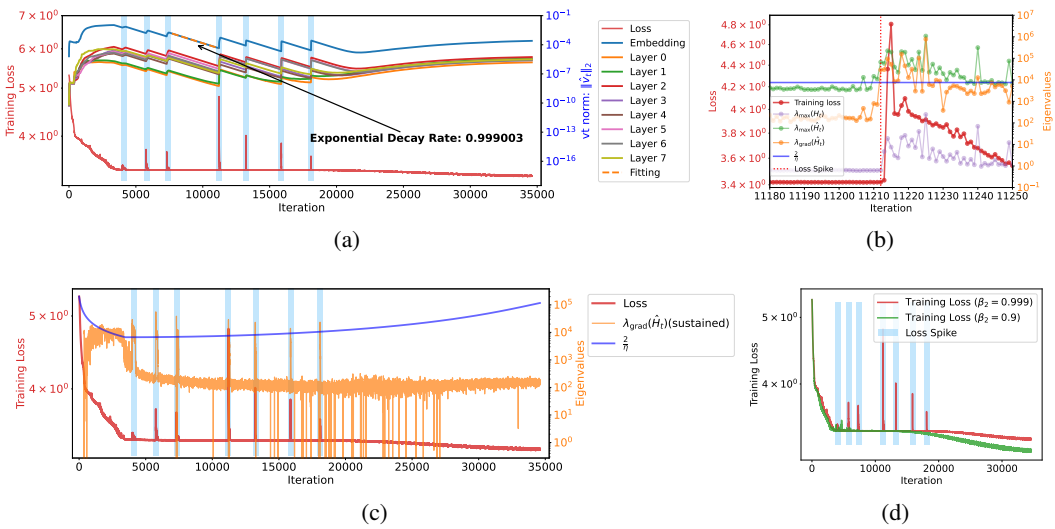
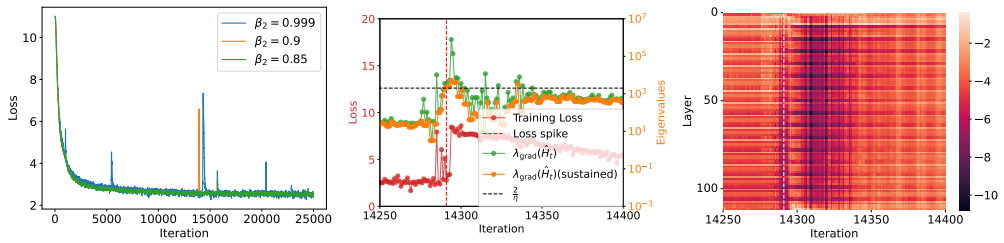


Figure 7: (a) Evolution of training loss and second moment $\|\hat{\boldsymbol{v}}_t\|$, with seven spikes highlighted. (b) Eigenvalue analysis near a typical spike. (c) Sustained gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)(\text{sustained})$ (orange) versus stability threshold $2/\eta$. The raw $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ is shown in Fig. D9. The $2/\eta$ line is plotted against the secondary y-axis on the right for for comparison with the eigenvalues. (d) Reduce the hyperparameter β_2 in Adam to 0.9 and retrain.



(a) Loss among different β_2 (b) Eigenvalues (c) Gradient norm

Figure 8: (a) Training loss evolution for a 187M parameter LLaMA transformer with different β_2 values. Loss curves show time-weighted EMA smoothing; raw loss appears in Fig. D11. (b) Gradient-directional eigenvalues $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ and sustained version $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)(\text{sustained})$ during a representative spike (iterations 14,250-14,400) with $\beta_2 = 0.999$. (c) Layer-wise gradient norms during the spike period. Layer indices on y-axis; gradient magnitudes shown in log-scale colorbar.

(Fig. 7(b), orange line) demonstrates perfect correspondence with all seven loss spike occurrences. Sustained periods above threshold trigger loss spikes, which is consistent with the findings in Fig. 3. In addition, we find that directly reducing β_2 is effective to mitigate loss spikes (Fig. 7(d)).

Large-Scale Language Model Validation: We trained a 187M parameter LLaMA-structured transformer on 100B tokens from SlimPajama to validate our mechanics in realistic large-scale settings. With the default $\beta_2 = 0.999$, training exhibits multiple loss spikes (Fig. 8(a)). Fig. 8(b) examines a representative spike occurring between iterations 14,250-14,400. We observe that the gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ exceeds the stability threshold $2/\eta$, signaling the spike onset. Consistent with our proposed mechanism (Sec. 5), gradient norms in certain layers diminish before this spike (Fig. 8(c)). As expected, reducing β_2 consistently decreases spike frequency during training (Fig. 8(a)), confirming the key role of second-moment in spike formation.

7 Conclusion and Discussion

In this work, we provide a detailed mechanistic analysis of loss spikes in Adam, showing that these spikes are triggered by Adam's adaptive preconditioners. By identifying a critical response delay between the second-moment and the current gradients, we reveal the mechanism underlying the persistence of these instabilities. Our theory suggests a simple remedy—reducing β_2—and we experimentally confirm its effectiveness. Encouragingly, many recent large-scale language model studies (Touvron et al., 2023; Dubey et al., 2024; Orvieto & Gower, 2025) have already adopted lower values of β_2 (e.g., 0.95 or lower), further underscoring the practical relevance of our analysis.

In addition, loss spikes represent more than mere optimization phenomena; they may signify transitions between distinct attractor basins in the landscape. Our supplementary experiments in Appendix E identify four spike types (**neutral**, **benign**, **malignant**, and **catastrophic**) in Transformer training—highlighting the importance of context-specific decisions on whether to suppress or preserve them. Precisely distinguishing between these spike types remains an unresolved challenge.

Beyond hyperparameter adjustments to Adam, alternative spike mitigation techniques include sandwich normalization (Ding et al., 2021; Yin et al., 2025), σ-Reparam (Zhai et al., 2023), and scaled-decouple distribution (Wang et al., 2025). While some studies (Lyu et al., 2022; Mueller et al., 2023) attribute normalization's effectiveness to sharpness reduction, a deeper understanding of how to leverage or control spikes remains a promising avenue for future research.

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International conference on machine learning*, pp. 247–257. PMLR, 2022.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024. PMLR, 2022.

Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1x-x309tm.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Jeremy 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. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Jeremy Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, Zachary Nado, George E Dahl, and Justin Gilmer. Adaptive gradient methods at the edge of stability. In *NeurIPS 2023 Workshop Heavy Tails in Machine Learning*, 2023.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J Zico Kolter, and Jason D. Lee. Understanding optimization in deep learning with central flows. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sIE2rI3ZPs.

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.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nhKHA59gXz.

Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=ZPQhzTSWA7.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Saber Elaydi. *An Introduction to Difference Equations*. Undergraduate Texts in Mathematics. Springer Science & Business Media, 3rd edition, 2005. ISBN 9780387230598.

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho*, and Krzysztof Geras*. The break-even point on optimization trajectories of deep neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1g87C4KwB.

Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amost Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkgEaj05t7.

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

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

Xiaolong Li, Zhi-Qin John Xu, and Zhongwang Zhang. Loss spike in training neural networks. *Journal of Computational Mathematics*, 2025.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd international conference on artificial intelligence and statistics*, pp. 983–992. PMLR, 2019.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2019.

Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35: 34689–34708, 2022.

Chao Ma, Daniel Kunin, Lei Wu, and Lexing Ying. Beyond the quadratic approximation: The multiscale structure of neural network loss landscapes. *Journal of Machine Learning*, 1(3): 247–267, 2022a. ISSN 2790-2048. doi: https://doi.org/10.4208/jml.220404. URL http://global-sci.org/intro/article_detail/jml/21028.html.

Chao Ma, Lei Wu, and weinan E. A qualitative study of the dynamic behavior for adaptive gradient algorithms. In *Mathematical and scientific machine learning*, pp. 671–692. PMLR, 2022b.

Igor Molybog, Peter Albert, Moya Chen, Zachary DeVito, David Esiobu, Naman Goyal, Punit Singh Koura, Sharan Narang, Andrew Poulton, Ruan Silva, et al. A theory on adam instability in large-scale machine learning. *arXiv preprint arXiv:2304.09871*, 2023.

Maximilian Mueller, Tiffany Joyce Vlaar, David Rolnick, and Matthias Hein. Normalization layers are all that sharpness-aware minimization needs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lArwl3y9x6.

Antonio Orvieto and Robert Gower. In search of adam's secret sauce. *arXiv preprint arXiv:2505.21829*, 2025.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSprop converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3UDSdyIcBDA.

Shohei Taniguchi, Keno Harada, Gouki Minegishi, Yuta Oshima, Seong Cheol Jeong, Go Nagahara, Tomoshi Iiyama, Masahiro Suzuki, Yusuke Iwasawa, and Yutaka Matsuo. Adopt: Modified adam can converge with any β_2 with the optimal rate. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Ya Wang, Zhijian Zhuo, Yutao Zeng, Xun Zhou, Jian Yang, and Xiaoqing Li. Scale-distribution decoupling: Enabling stable and effective training of large language models. *arXiv preprint arXiv:2502.15499*, 2025.

Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 35:9983–9994, 2022.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

Yuege Xie, Xiaoxia Wu, and Rachel Ward. Linear convergence of adaptive stochastic gradient descent. In *International conference on artificial intelligence and statistics*, pp. 1475–1485. PMLR, 2020.

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

Yichun Yin, Wenyong Huang, Kaikai Song, Yehui Tang, Xueyu Wu, Wei Guo, Peng Guo, Yaoyuan Wang, Xiaojun Meng, Yasheng Wang, Dong Li, Can Chen, Dandan Tu, Yin Li, Fisher Yu, Ruiming Tang, Yunhe Wang, Baojun Wang, Bin Wang, Bo Wang, Boxiao Liu, Changzheng Zhang, Duyu Tang, Fei Mi, Hui Jin, Jiansheng Wei, Jiarui Qin, Jinpeng Li, Jun Zhao, Liqun Deng, Lin Li, Minghui Xu, Naifu Zhang, Nianzu Zheng, Qiang Li, Rongju Ruan, Shengjun Cheng, Tianyu Guo, Wei He, Wei Li, Weiwen Liu, Wulong Liu, Xinyi Dai, Yonghan Dong, Yu Pan, Yue Li, Yufei Wang, Yujun Li, Yunsheng Ni, Zhe Liu, Zhenhe Zhang, and Zhicheng Liu. Pangu ultra: Pushing the limits of dense large language models on ascend npus, 2025. URL https://arxiv.org/abs/2504.07866.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pp. 40770–40803. PMLR, 2023.

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.

Zhongwang Zhang, Zhiwei Wang, Junjie Yao, Zhangchen Zhou, Xiaolong Li, Weinan E, and Zhi-Qin John Xu. Anchor function: a type of benchmark functions for studying language models. In *ICLR 2025 Workshop Bridging the Gap Between Practice and Theory in Deep Learning*, 2025. URL https://arxiv.org/abs/2401.08309.

Dongruo Zhou, Jinghui Chen, Yuan Cao, Ziyan Yang, and Quanquan Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=Gh0cxhbz3c. Featured Certification.

Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11127–11135, 2019.

A THE USE OF LARGE LANGUAGE MODELS(LLMS)

We acknowledge the use of large language models in the preparation of this manuscript. Specifically, we employed LLMs (including but not limited to GPT-4, Claude, and similar models) solely for language polishing and writing enhancement purposes. The LLMs were used to: (i) Improve sentence structure and clarity; (ii) Enhance grammatical accuracy and flow; (iii) Refine technical writing style and consistency; and (iv) Polish language expression while preserving original meaning.

B ETHICS AND REPRODUCIBILITY STATEMENT

Ethics Statement. This work involves theoretical analysis and empirical studies of Adam optimization algorithm using standard neural network architectures and publicly available datasets. All experiments were conducted following established ethical guidelines for machine learning research. No human subjects, sensitive data, or potentially harmful applications were involved in this study.

Reproducibility Statement. To ensure reproducibility, we provide detailed experimental configurations in Appendix G and supplementary experiments in Appendix F. Our theoretical analysis includes complete mathematical derivations and proofs in Appendix D. All hyperparameters, network architectures, and training procedures are fully specified. The synthetic datasets and training procedures can be reproduced following the provided specifications. Code and additional implementation details are made available in the supplementary materials.

C LIMITATION AND FUTURE WORK

Our detailed analysis of loss spikes in Adam optimization reveals that adaptive preconditioners can themselves trigger these phenomena and we verify this mechanism in certain neural network architectures. However, we acknowledge that in more complex scenarios, both the intrinsic geometry of the loss landscape and the applied preconditioners likely interact to jointly produce loss spikes. Disentangling these individual contributions and accurately attributing different spike mechanisms in large-scale models remains a significant challenge for future research.

While we have developed efficient Hessian-vector products to compute gradient-directional eigenvalues without full Hessian computation, computational cost remains a key constraint for scaling this analysis to larger models. Developing efficient algorithms to approximate maximum Hessian eigenvalues and gradient-directional eigenvalues represents a critical direction for future work.

Furthermore, as discussed in Appendix E, the precise categorization of loss spikes into our proposed taxonomy (**neutral**, **benign**, **malignant**, and **catastrophic** types) presents ongoing challenges. Developing robust, computationally efficient criteria to distinguish between these categories would significantly enhance our ability to detect and appropriately respond to different spike types during training.

D PROOFS OF THEORETICAL RESULTS

Proposition D.1. *Let $L : \mathbb{R}^M \to \mathbb{R}$ be twice continuously differentiable. For any iterate $\boldsymbol{\theta}_t$ define the gradient $\boldsymbol{g}_t := \nabla L(\boldsymbol{\theta}_t)$ and, for a fixed learning rate $\eta > 0$, define the local directional maximum Hessian $\bar{\lambda}_t := \sup_{s\in[0,1]} \lambda_{\max}\big(\nabla^2 L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t)\big)$, the maximum eigenvalue of the Hessian along the line segment from $\boldsymbol{\theta}_t$ to $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{g}_t$. If $\eta < \frac{2}{\bar{\lambda}_t}$, then we have the descent estimate:*

$$L(\boldsymbol{\theta}_{t+1}) \le L(\boldsymbol{\theta}_t) - \eta\Big(1 - \frac{\eta\bar{\lambda}_t}{2}\Big)\|\boldsymbol{g}_t\|^2.$$

In particular, whenever $\eta \in \big(0, 2/\bar{\lambda}_t\big)$ and $\boldsymbol{g}_t \neq 0$ we have strict decrease $L(\boldsymbol{\theta}_{t+1}) < L(\boldsymbol{\theta}_t)$.

Proof. Apply the one-dimensional Taylor expansion of the scalar function $\phi(s) := L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t)$ around $s = 0$ up to second order with the remainder written using the Hessian at some point along the segment. Equivalently, use the multivariate Taylor expansion along the direction $-\eta \boldsymbol{g}_t$:

$$L(\boldsymbol{\theta}_t - \eta \boldsymbol{g}_t) \;=\; L(\boldsymbol{\theta}_t) - \eta \boldsymbol{g}_t^\top \boldsymbol{g}_t + \frac{\eta^2}{2}\, \boldsymbol{g}_t^\top \Big(\nabla^2 L(\boldsymbol{\theta}_t - s^* \eta \boldsymbol{g}_t)\Big) \boldsymbol{g}_t$$

for some $s^* \in (0, 1)$. Since the symmetric matrix $\nabla^2 L(\boldsymbol{\theta}_t - s^* \eta \boldsymbol{g}_t)$ has largest eigenvalue at most $\bar{\lambda}_t$, we get

$$\boldsymbol{g}_t^\top \Big(\nabla^2 L(\boldsymbol{\theta}_t - s^* \eta \boldsymbol{g}_t)\Big) \boldsymbol{g}_t \le \bar{\lambda}_t \, \|\boldsymbol{g}_t\|^2.$$

Hence

$$L(\boldsymbol{\theta}_{t+1}) \le L(\boldsymbol{\theta}_t) - \eta \|\boldsymbol{g}_t\|^2 + \frac{\eta^2}{2} \bar{\lambda}_t \|\boldsymbol{g}_t\|^2 = L(\boldsymbol{\theta}_t) - \eta \Big(1 - \frac{\eta \bar{\lambda}_t}{2}\Big) \|\boldsymbol{g}_t\|^2.$$

If $\eta < 2/\bar{\lambda}_t$, then $1 - \frac{\eta \bar{\lambda}_t}{2} > 0$, so the right-hand side is strictly less than $L(\boldsymbol{\theta}_t)$ whenever $\boldsymbol{g}_t \neq 0$. □

Lemma D.1. *Let $\boldsymbol{H}$ be a real symmetric matrix and $\hat{\boldsymbol{H}} = \mathrm{diag}\left(\frac{1}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon}\right) \boldsymbol{H}$. Then $\hat{\boldsymbol{H}}$ is diagonalizable in the field of real numbers.*

Proof. While $\mathrm{diag}\left(\frac{1}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon}\right) \boldsymbol{H}$ is generally asymmetric, we can demonstrate that it is similar to a symmetric matrix and therefore has real eigenvalues. Let $\boldsymbol{D}_t = \mathrm{diag}\left(\frac{1}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon}\right)$, which is positive definite. We can express:

$$\boldsymbol{D}_t \boldsymbol{H} = \boldsymbol{D}_t^{1/2} \cdot (\boldsymbol{D}_t^{1/2} \boldsymbol{H} \boldsymbol{D}_t^{1/2}) \cdot \boldsymbol{D}_t^{-1/2}$$

Since $\boldsymbol{D}_t^{1/2} \boldsymbol{H} \boldsymbol{D}_t^{1/2}$ is symmetric, $\boldsymbol{D}_t \boldsymbol{H}$ is similar to a symmetric matrix. This confirms that $\boldsymbol{D}_t \boldsymbol{H}$ has real eigenvalues and is diagonalizable. □

Proposition D.2. *Consider the three-term recursive iteration*

$$\delta \boldsymbol{\theta}_{t+1} = \big[(1 + \beta_1) \boldsymbol{I} - \eta(1 - \beta_1) \boldsymbol{H}(\boldsymbol{\theta}_0)\big] \, \delta \boldsymbol{\theta}_t - \beta_1 \delta \boldsymbol{\theta}_{t-1} - \eta(1 - \beta_1) \nabla L(\boldsymbol{\theta}_0),$$

with learning rate $\eta > 0$ and momentum parameter $\beta_1 \in [0, 1)$. Then the linearized system at $\boldsymbol{\theta}_0$ is asymptotically stable in all positive-curvature eigendirections (i.e., for every eigenvalue $\lambda_i > 0$ the characteristic roots lie strictly inside the unit disk) if and only if

$$\lambda_{\max}\left(\frac{1 - \beta_1}{1 + \beta_1} \boldsymbol{H}(\boldsymbol{\theta}_0)\right) < \frac{2}{\eta},$$

where $\lambda_{\max}(\cdot)$ denotes the largest positive eigenvalue.

Proof. We analyze the stability of the vector recurrence by decomposing it along the eigenspace of the Hessian matrix. Since the Hessian $\boldsymbol{H} := \boldsymbol{H}(\boldsymbol{\theta}_0)$ is symmetric, it admits an eigen-decomposition $\boldsymbol{H} = \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^\top$, where $\boldsymbol{Q}$ is an orthogonal matrix and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ contains the eigenvalues of $\boldsymbol{H}$.

Define the change of variables $\delta \boldsymbol{\theta}_t = \boldsymbol{Q} \boldsymbol{z}_t$. Substituting into the recurrence yields

$$\boldsymbol{z}_{t+1} = [(1 + \beta_1) \boldsymbol{I} - \eta(1 - \beta_1) \boldsymbol{\Lambda}] \, \boldsymbol{z}_t - \beta_1 \boldsymbol{z}_{t-1} - \eta(1 - \beta_1) \boldsymbol{Q}^\top \nabla L(\boldsymbol{\theta}_0).$$

Since this is a decoupled system in the eigenbasis, for each positive-curvature eigendirections with $\lambda_i > 0$, the i-th component $z_t^{(i)}$ satisfies a scalar second-order linear nonhomogeneous recurrence:

$$z_{t+1}^{(i)} = \alpha_i z_t^{(i)} - \beta_1 z_{t-1}^{(i)} + c_i,$$

where

$$\alpha_i := (1 + \beta_1) - \eta(1 - \beta_1) \lambda_i, \quad c_i := -\eta(1 - \beta_1) g^{(i)}, \quad g^{(i)} := \big[\boldsymbol{Q}^\top \nabla L(\boldsymbol{\theta}_0)\big]_i.$$

The general solution to this nonhomogeneous recurrence is the sum of the homogeneous solution and a particular solution. The homogeneous part is governed by the characteristic equation:

$$r^2 - \alpha_i r + \beta_1 = 0.$$

It is well known (e.g., see Elaydi, *An Introduction to Difference Equations* (Elaydi, 2005)) that the solution $z_t^{(i)}$ that the homogeneous solution is (asymptotically) stable (both characteristic roots lie strictly inside the unit disk, so perturbations in this direction decay and do not grow exponentially) if and only if both roots of the characteristic equation lie strictly inside the unit circle in the complex plane. This is equivalent to the following three conditions:

$$\begin{cases} 1 + \alpha_i + \beta_1 > 0, \\ 1 - \alpha_i + \beta_1 > 0, \\ |\beta_1| < 1. \end{cases}$$

Since $\beta_1 \in [0, 1)$ by assumption, the third condition always holds. The first two inequalities can be rewritten as:

$$|\alpha_i| < 1 + \beta_1.$$

Substituting the expression for α_i, we obtain:

$$|(1 + \beta_1) - \eta(1 - \beta_1)\lambda_i| < 1 + \beta_1.$$

Solving this inequality gives:

$$\eta(1 - \beta_1)\lambda_i < 2(1 + \beta_1) \quad \Longleftrightarrow \quad \lambda_i < \frac{2}{\eta} \cdot \frac{1 + \beta_1}{1 - \beta_1}.$$

Therefore, the recurrence stabilize in all eigendirections with $\lambda_i > 0$ if and only if

$$\lambda_{\max}\left(\frac{1 - \beta_1}{1 + \beta_1}\boldsymbol{H}\right) < \frac{2}{\eta}.$$

This completes the proof. □

Theorem D.1 (Exact Necessary and Sufficient Condition for Loss Spike Onset). *Let $L : \mathbb{R}^M \to \mathbb{R}$ be twice continuously differentiable. At iterate $\boldsymbol{\theta}_t$, denote the gradient $\boldsymbol{g}_t := \nabla L(\boldsymbol{\theta}_t) \neq 0$, and consider a gradient descent update*

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{g}_t, \qquad \eta > 0.$$

Define the weighted averaged Hessian along the update direction by

$$\bar{\boldsymbol{H}}_t := 2 \int_0^1 (1 - s)\, \nabla^2 L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t)\, ds,$$

and the corresponding directional curvature by

$$\lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t) := \frac{\boldsymbol{g}_t^\top \bar{\boldsymbol{H}}_t \boldsymbol{g}_t}{\|\boldsymbol{g}_t\|^2}.$$

Then the update exhibits a loss increase (necessary for a loss spike) if and only if $\lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t) > \frac{2}{\eta}$, i.e.,

$$L(\boldsymbol{\theta}_{t+1}) > L(\boldsymbol{\theta}_t) \quad \Longleftrightarrow \quad \lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t) > \frac{2}{\eta}.$$

Proof. Consider the univariate function

$$\phi(s) := L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t), \qquad s \in [0, 1].$$

By the chain rule,

$$\phi'(s) = -\eta\, \boldsymbol{g}_t^\top \nabla L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t), \qquad \phi'(0) = -\eta \|\boldsymbol{g}_t\|^2.$$

Differentiating once more yields

$$\phi''(s) = \eta^2 \, \boldsymbol{g}_t^\top \nabla^2 L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t) \, \boldsymbol{g}_t.$$

Since L is twice continuously differentiable, ϕ is C^2 on $[0, 1]$. The second-order Taylor theorem with *integral remainder* gives the exact identity

$$\phi(1) = \phi(0) + \phi'(0) + \int_0^1 (1 - s) \, \phi''(s) \, ds.$$

Substituting the expressions for $\phi(0)$, $\phi'(0)$, and $\phi''(s)$ yields

$$L(\boldsymbol{\theta}_{t+1}) - L(\boldsymbol{\theta}_t) = -\eta \|\boldsymbol{g}_t\|^2 + \eta^2 \int_0^1 (1 - s) \, \boldsymbol{g}_t^\top \nabla^2 L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t) \, \boldsymbol{g}_t \, ds.$$

Introduce the weighted averaged Hessian

$$\bar{\boldsymbol{H}}_t := \frac{\int_0^1 (1 - s) \, \nabla^2 L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t) \, ds}{\int_0^1 (1 - s) \, ds} = 2 \int_0^1 (1 - s) \, \nabla^2 L(\boldsymbol{\theta}_t - s\eta \boldsymbol{g}_t) \, ds.$$

Then the previous equality becomes

$$L(\boldsymbol{\theta}_{t+1}) - L(\boldsymbol{\theta}_t) = -\eta \|\boldsymbol{g}_t\|^2 + \frac{\eta^2}{2} \, \boldsymbol{g}_t^\top \bar{\boldsymbol{H}}_t \boldsymbol{g}_t.$$

Dividing both sides by $\|\boldsymbol{g}_t\|^2 > 0$ shows that the sign of the loss change is exactly the sign of

$$-\eta + \frac{\eta^2}{2} \lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t),$$

where

$$\lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t) := \frac{\boldsymbol{g}_t^\top \bar{\boldsymbol{H}}_t \boldsymbol{g}_t}{\|\boldsymbol{g}_t\|^2}.$$

Therefore,

$$L(\boldsymbol{\theta}_{t+1}) > L(\boldsymbol{\theta}_t) \quad \Longleftrightarrow \quad -\eta + \frac{\eta^2}{2} \lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t) > 0 \quad \Longleftrightarrow \quad \lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t) > \frac{2}{\eta}.$$

This proves the claimed necessary and sufficient condition for a loss spike onset. □

Practical proxy for loss spike onset. The exact loss–spike condition in Theorem D.1 depends on the directional curvature $\lambda_{\text{grad}}(\bar{\boldsymbol{H}}_t)$, where $\bar{\boldsymbol{H}}_t$ is the weighted line–segment average of the true Hessian. Computing $\bar{\boldsymbol{H}}_t$ is intractable in modern deep networks, as it requires access to second-order information along the entire update path. In practice, since learning rates are typically small, we can monitor the step-wise curvature as a proxy:

$$\lambda_{\text{grad}}(\boldsymbol{H}_t) := \frac{\boldsymbol{g}_t^\top \boldsymbol{H}_t \boldsymbol{g}_t}{\|\boldsymbol{g}_t\|^2}.$$

Our central theoretical insight is that Adam can be understood as applying a preconditioning transformation to the Hessian, as expressed in our Equation 5:

$$\hat{\boldsymbol{H}}_t = \frac{1}{1 - \beta_1^t} \frac{1 - \beta_1}{1 + \beta_1} \text{diag}\left(\frac{1}{\sqrt{\hat{\boldsymbol{v}}_t} + \varepsilon}\right) \boldsymbol{H}_t.$$

Therefore, a natural extension for Adam is to replace $\boldsymbol{H}_t$ with the preconditioned Hessian $\hat{\boldsymbol{H}}_t$. This yields our predictor:

$$\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t) := \frac{\nabla L(\boldsymbol{\theta}_t)^\top \hat{\boldsymbol{H}}_t \nabla L(\boldsymbol{\theta}_t)}{\|\nabla L(\boldsymbol{\theta}_t)\|^2} > \frac{2}{\eta}.$$

Empirically, we observe that this curvature proxy aligns closely with the onset of loss spikes across architectures and datasets, suggesting that it provide a robust approximation to the underlying directional curvature governing spike formation.

Theorem D.2 (Five Stages of Adam for Optimizing Quadratic Loss). *Consider the 1-d quadratic loss $L(\theta) = \frac{1}{2}\theta^2$, optimized using Adam with hyper-parameters $\beta_1 = 0$, $\beta_2 \in (0, 1)$, and learning rate $\eta > 0$. The update rules are:*

$$\theta_{t+1} = \left(1 - \frac{\eta}{\sqrt{v_t}}\right)\theta_t, \quad v_{t+1} = \beta_2 v_t + (1-\beta_2)\theta_t^2.$$

Assume the initialization satisfies $v_0 = \theta_0^2$ and $|\theta_0| > \frac{\eta}{2}$. Assume $\frac{1}{\ln(1/\beta_2)} > \frac{1}{\ln(\frac{2|\theta_0|}{\eta})} + \frac{1}{\ln 2}$. Then there exist integers $t_0 < t_1 < t_2 < t_3 < t_4 < t_5 < \infty$ such that the iterates (θ_t, v_t) exhibit the five stages described above in intervals $[t_i, t_{i+1})$, respectively. Specifically,

(i) ***Stable Loss Decrease****. Define $t_0 = 0$, then for all $t_0 \le t < t_1$, where*

$$t_1 := \frac{2\ln\left(\frac{|\theta_0|}{\eta} + \frac{1}{2}\right)}{\ln\frac{1}{\beta_2}},$$

the sequence $|\theta_t|$ decreases exponentially, and $v_t \in [\beta_2^t\theta_0^2, \theta_0^2]$. In particular, there exists $s \in (0, 1)$ such that

$$|\theta_t| \le s^t|\theta_0|, \quad \textit{and} \quad |\theta_{t_1}| \le \delta := s^{t_1}|\theta_0|.$$

(ii) ***Preconditioners Decay.*** *For $t_1 \le t < t_2$, where*

$$t_2 := \inf\left\{t > t_1 \mid \sqrt{v_t} < \tfrac{\eta}{2}\right\},$$

the momentum v_t decays exponentially as

$$v_t \le (v_{t_1+1} - \delta^2)\beta_2^{t-t_1-1} + \delta^2.$$

(iii) ***Spike Onset.*** *Define*

$$t_3 := \inf\left\{t > t_2 \mid v_{t+1} > v_t\right\}.$$

For $t_2 \le t < t_3$, the preconditioner v_t continues to decay, and the update multiplier $\left|1 - \frac{\eta}{\sqrt{v_t}}\right|$ grows, causing $|\theta_t|$ to increase exponentially.

(iv) ***Preconditioners Growth.*** *Define*

$$t_4 := \inf\{\, t > t_3 \mid \sqrt{v_t} > \tfrac{\eta}{2}\,\}.$$

For $t_3 \le t < t_4$, the growing gradient magnitude forces the preconditioner v_t to increase. Consequently, the update multiplier $\left|1 - \frac{\eta}{\sqrt{v_t}}\right|$ shrinks steadily, preparing the transition from explosive growth to contraction.

(v) ***Loss Decrease.*** *Define*

$$t_5 := \inf\left\{t > t_4 : \ \sqrt{v_t} < \tfrac{\eta}{2}\right\}.$$

If no such t exists, we simply take $t_5 > t_4$ to be any larger index. For $t_4 \le t < t_5$, the preconditioner has grown sufficiently so that $\frac{\eta}{\sqrt{v_t}} < 1$. In this regime, the update multiplier satisfies $\left|1 - \frac{\eta}{\sqrt{v_t}}\right| < 1$, ensuring that $|\theta_t|$ contracts and the loss $L(\theta_t) = \frac{1}{2}\theta_t^2$ decreases once again.

Proof. We proceed in stages and make all inequalities explicit. The corresponding schematic diagrams of the five stages are shown in Fig. D1.

Stage 1 (Stable loss decrease). For the given initialization $v_0 = \theta_0^2$ and $0 < \beta_2 < 1$ we have the trivial lower bound (single-step recurrence gives a simple monotone inequality)

$$v_t \ge \beta_2^t v_0 = \beta_2^t\theta_0^2, \qquad \forall t \ge 0.$$

Also note $v_t \ge 0$ for all t.

Construction of t_1 and δ. Define

$$t_1 := \frac{2\ln\left(\frac{|\theta_0|}{\eta} + \frac{1}{2}\right)}{\ln(1/\beta_2)}.$$

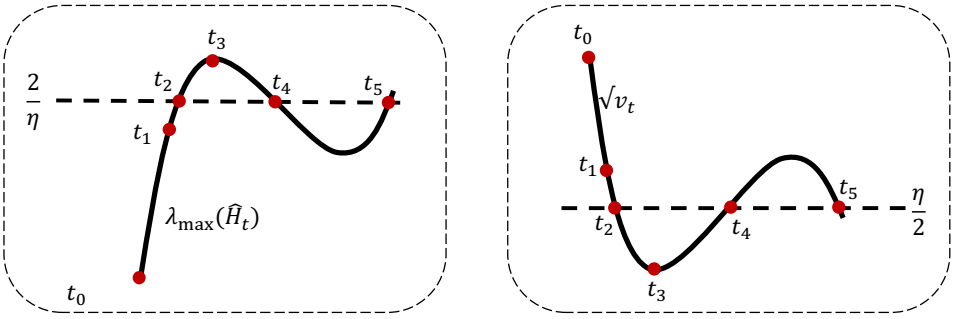


Figure D1: The five stages are illustrated schematically.

Because $0 < \beta_2 < 1$, $\ln(1/\beta_2) > 0$ and t_1 is well defined. Set

$$s := \max\Big\{\tfrac{1}{2}\frac{\eta}{|\theta_0|},\ \big|1 - \tfrac{\eta}{|\theta_0|}\big|\Big\}.$$

By the hypothesis $|\theta_0| > \eta/2$ we have $s \in (0, 1)$. Define

$$\delta := s^{\lfloor t_1 \rfloor}|\theta_0|.$$

Here, $\lfloor\cdot\rfloor$ is the floor function. The choice of t_1 ensures the following inequality chain for all integers t with $t_0 \le t < t_1$. Using the lower bound $v_t > \beta_2^t\theta_0^2$ and the definition of t_1, one obtains

$$\sqrt{v_t} \ge \beta_2^{t/2}|\theta_0| \ge \beta_2^{t_1/2}|\theta_0| \quad \text{and by the definition of } t_1, \quad \beta_2^{t_1/2}|\theta_0| = \frac{|\theta_0|}{\frac{|\theta_0|}{\eta} + \frac{1}{2}},$$

so in particular $\sqrt{v_t} > \frac{|\theta_0|}{\frac{|\theta_0|}{\eta}+\frac{1}{2}}$ and hence

$$1 - \frac{\eta}{\sqrt{v_t}} > -\frac{1}{2}\frac{\eta}{|\theta_0|}.$$

Therefore

$$-1 < -\frac{1}{2}\frac{\eta}{|\theta_0|} < 1 - \frac{\eta}{\sqrt{v_t}} < 1, \forall 0 \le t < t_1.$$

This indicates that $|\theta_t|$ is monotonically decreasing for $0 < t < t_1$. Thus, $\sqrt{v_t} \le |\theta_0|$ for all $0 < t < t_1$. This completes the upper bound of $1 - \frac{\eta}{\sqrt{v_t}}$ as follows:

$$-\frac{1}{2}\frac{\eta}{|\theta_0|} < 1 - \frac{\eta}{\sqrt{v_t}} < 1 - \frac{\eta}{|\theta_0|}, \forall 0 \le t < t_1.$$

By definition of s, we get

$$\Big|1 - \frac{\eta}{\sqrt{v_t}}\Big| \le s < 1.$$

Therefore for $0 < t < t_1$,

$$|\theta_t| \le s^t|\theta_0|.$$

In particular $|\theta_{\lfloor t_1 \rfloor}| \le \delta$, establishing the intended bound at the end of Stage 1. This proves Stage 1.

Stage 2 (Preconditioner decay). Define

$$t_2 := \inf\Big\{t \in \mathbb{N}^+ :\ 1 - \frac{\eta}{\sqrt{v_t}} < -1\Big\}.$$

For integers $t_1 \le t \le t_2$, we have $|\theta_t| \le |\theta_{t_1}| \le \delta$. The recurrence for v implies

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)\theta_t^2 \le \beta_2 v_t + (1-\beta_2)\delta^2.$$

This is an affine linear inequality in v_t. Iterating this inequality forward from $t = t_1+1$ yields, for any integer $t_1 + 1 \le t \le t_2$,

$$v_t \le (v_{t_1+1} - \delta^2)\beta_2^{t-t_1-1} + \delta^2, \tag{7}$$

which shows v_t decays geometrically toward δ^2 with factor β_2 so long as $|\theta_t| \le \delta$. Because $|\theta_t| \le \delta$ on the time window following Stage 1 by construction, we have established the Stage 2 statement.

Note also the obvious lower bound obtained by ignoring the additive $(1 - \beta_2)\theta_t^2$ term:

$$v_t \ge v_{t_1+1}\beta_2^{t-t_1-1},$$

so v_t is squeezed between two geometric forms until $|\theta_t|$ leaves the small region.

Existence and finiteness of t_2: Suppose by contradiction that $t_2 = +\infty$. Then $1 - \frac{\eta}{\sqrt{v_t}} \ge -1$, which simplifies to $v_t \ge \frac{\eta^2}{4}, \forall t \in \mathbb{N}^+$. In Eq. (7) let $t \to +\infty$, it follows that $\limsup_{t\to\infty} v_t \le \delta^2$. So $\delta^2 \ge \frac{\eta^2}{4}$, which indicates that $\delta \ge \frac{\eta}{2}$. Since $\delta := s^{\lfloor t_1 \rfloor}|\theta_0|$, we have $\lfloor t_1 \rfloor \le \frac{\ln(\frac{2|\theta_0|}{\eta})}{\ln(1/s)}$. By definition, $s \ge \frac{1}{2}$, so $\lfloor t_1 \rfloor \le \frac{\ln(\frac{2|\theta_0|}{\eta})}{\ln 2}$. By definition of t_1, it follows that

$$\frac{\ln(\frac{2|\theta_0|}{\eta})}{\ln(1/\beta_2)} - 1 \le \frac{2\ln(\frac{|\theta_0|}{\eta} + \frac{1}{2})}{\ln(1/\beta_2)} - 1 \le \lfloor t_1 \rfloor \le \frac{\ln(\frac{2|\theta_0|}{\eta})}{\ln 2}.$$

Therefore we have

$$\frac{1}{\ln(1/\beta_2)} \le \frac{1}{\ln(\frac{2|\theta_0|}{\eta})} + \frac{1}{\ln 2},$$

which contradicts the assumption. So t_2 is finite.

Stage 3 (Spike onset). By definition of t_2, at $t = t_2$ we have $\sqrt{v_{t_2}} < \eta/2$. Consequently

$$\left|1 - \frac{\eta}{\sqrt{v_{t_2}}}\right| > 1,$$

so passing from t_2 to $t_2 + 1$ yields

$$|\theta_{t_2+1}| = \left|1 - \frac{\eta}{\sqrt{v_{t_2}}}\right| |\theta_{t_2}| > |\theta_{t_2}|.$$

Thus $|\theta_t|$ grows for t just after t_1.

Finiteness of t_3. To capture when the second-moment estimate v_t ceases to decay, define

$$t_3 := \inf\{\, t > t_2 : v_{t+1} > v_t \,\}.$$

If no such t exists we set $t_3 = +\infty$. Suppose, for contradiction, that $t_3 = \infty$. Then $v_{t+1} \le v_t$ for all $t \ge t_2$, so v_t is monotonically decreasing and bounded below by 0. Thus the limit

$$v_\infty := \lim_{t\to\infty} v_t$$

exists. Since $v_t \le v_{t_2}$ for all $t \ge t_2$, we obtain

$$\frac{\eta}{\sqrt{v_t}} \ge \frac{\eta}{\sqrt{v_{t_2}}} > 2,$$

hence there exists a constant $q := \frac{\eta}{\sqrt{v_{t_2}}} - 1 > 1$ such that

$$\left|1 - \frac{\eta}{\sqrt{v_t}}\right| \ge q > 1, \qquad \forall t \ge t_2.$$

By recursion,

$$|\theta_{t_2+k}| \;\ge\; q^k|\theta_{t_2}| \;\to\; \infty \quad \text{as } k \to \infty.$$

However, the recurrence for v_t is

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2)\theta_t^2.$$

Since $|\theta_t| \to \infty$ and $1 - \beta_2 > 0$, the term $(1 - \beta_2)\theta_t^2 \to \infty$, forcing $v_{t+1} \to \infty$. This contradicts the assumption that v_t is monotonically decreasing with a finite limit v_∞. Therefore, $t_3 < \infty$. The larger β_2 is, the more slowly v_t responds to g_t, and the later the index t_3 of the monotonic change will occur.

Exponential growth in loss for $t_2 \le t < t_3$. For any $t_2 \le t < t_3$, we have $v_{t+1} \le v_t \le v_{t_2}$. Hence

$$\frac{\eta}{\sqrt{v_t}} \ge \frac{\eta}{\sqrt{v_{t_2}}} > 2,$$

and so

$$\left|1 - \frac{\eta}{\sqrt{v_t}}\right| \ge q > 1,$$

where $q = \frac{\eta}{\sqrt{v_{t_2}}} - 1$. By induction,

$$|\theta_t| \;\ge\; q^{\,t-t_2}|\theta_{t_2}|, \qquad \forall t_2 \le t < t_3.$$

Thus $|\theta_t|$ grows at least exponentially on the interval $[t_2, t_3)$, and the loss

$$l(\theta_t) = \tfrac{1}{2}\theta_t^2$$

increases dramatically, capturing the onset of the spike.

Stage 4 (Preconditioner growth). Define

$$t_4 := \inf\{\, t > t_3 \mid \sqrt{v_t} > \tfrac{\eta}{2} \,\}.$$

Finiteness of t_4. We first show that $t_4 < +\infty$. Suppose, for contradiction, that $t_4 = +\infty$. By the definition of t_3, we have $v_{t_3+1} > v_{t_3}$. Since

$$v_{t_3+1} = \beta_2 v_{t_3} + (1 - \beta_2)\theta_{t_3}^2,$$

this inequality implies $\theta_{t_3}^2 > v_{t_3}$. On the other hand,

$$\theta_{t_3+1} = \Big(1 - \tfrac{\eta}{\sqrt{v_{t_3}}}\Big)\theta_{t_3}.$$

If $\theta_{t_3} > 0$, then

$$\theta_{t_3+1} < (1 - \frac{\eta}{\theta_{t_3}})\theta_{t_3} = \theta_{t_3} - \eta,$$

so either $\theta_{t_3} > \frac{\eta}{2}$ or $\theta_{t_3+1} < -\frac{\eta}{2}$. Thus, in either case, there exists some $t \in \{t_3, t_3 + 1\}$ such that

$$|\theta_t| > \tfrac{\eta}{2}.$$

Now assume $t_4 = +\infty$. Then by definition we must have $\sqrt{v_t} \le \frac{\eta}{2}$ for all $t > t_3$. Hence

$$\left|1 - \tfrac{\eta}{\sqrt{v_t}}\right| \ge 1,$$

implying that $|\theta_t|$ is monotonically non-decreasing. Since at least one of $|\theta_{t_3}|$ or $|\theta_{t_3+1}|$ already exceeds $\frac{\eta}{2}$, it follows that

$$|\theta_t| \ge a := \max\{|\theta_{t_3}|, |\theta_{t_3+1}|\} > \tfrac{\eta}{2}, \qquad \forall\, t > t_3.$$

Thus $|\theta_t|$ converges to a limit (possibly $+\infty$) with

$$\lim_{t\to\infty} |\theta_t| \;\ge\; a > \tfrac{\eta}{2}.$$

But then, since

$$v_{t+1} = \beta_2 v_t + (1 - \beta_2)\theta_t^2,$$

we must have

$$\lim_{t\to\infty} v_t = a^2,$$

so that

$$\lim_{t\to\infty} \sqrt{v_t} = a > \tfrac{\eta}{2}.$$

This contradicts the assumption that $\sqrt{v_t} \le \frac{\eta}{2}$ for all $t > t_3$. Therefore t_4 must be finite.

During the interval $t_3 < t \le t_4$, the preconditioner $\sqrt{v_t}$ evolves from being strictly below $\frac{\eta}{2}$ to exceeding it. We refer to this regime as the "preconditioner growth stage".

Stage 5 (Loss decrease). Define

$$t_5 := \inf\Big\{t > t_4 :\ 1 - \frac{\eta}{\sqrt{v_t}} < -1\Big\}.$$

If no such t exists, we simply set $t_5 > t_4$ to be any larger index for convenience. At time t_4, the preconditioner satisfies $\sqrt{v_{t_4}} > \frac{\eta}{2}$. Hence, for $t \ge t_4$,

$$\Big|1 - \tfrac{\eta}{\sqrt{v_t}}\Big| < 1.$$

This ensures that, during the interval $t_4 \le t \le t_5$, the multiplicative factor falls strictly within $(-1, 1)$, so $|\theta_t|$ no longer grows but instead contracts. Consequently, the loss $L(\theta_t) = \frac{1}{2}\theta_t^2$ decreases over this period.

Thus the trajectory transitions from exponential growth (Stage 3) and preconditioner growth (Stage 4) into a contraction regime. In this way, the cycle closes and the dynamics return to behavior of the same type as in Stage 1.

This completes the proof of the five-stage behavior for the quadratic optimization. □

Theorem D.3 (Analysis of decaying learning rate scheduler). *Consider the same setup as Thm. 1 with decaying learning rate $\eta_t = \eta_0(t+1)^{-\alpha}$ where $\alpha \in (0, 1)$. Assume the initialization satisfies $v_0 = \theta_0^2$ and $|\theta_0| > 2\eta_0 > 0$. Assume β_2 is sufficiently close to 1. Then the stability condition $|1 - \frac{\eta_t}{\sqrt{v_t}}| < 1$ cannot hold for all $t \in \mathbb{N}^+$.*

Proof. Assume by contradiction that $|1 - \frac{\eta_t}{\sqrt{v_t}}| < 1$ holds for all $t \in \mathbb{N}^+$.

Stage 1 (Loss Decay Stage). For all t, $\beta_2^t v_0 \le v_t \le \theta_0^2$. Define $t_0 = \frac{\log 2}{\log \frac{1}{\beta_2}}$. Then for all $t \le t_0$, $v_t \ge \frac{1}{2}v_0$. Since $|\theta_0| > 2\eta_0$, we have $\frac{\eta_t}{\sqrt{v_t}} < \frac{\eta_0}{\sqrt{\frac{1}{2}v_0}} < \frac{2\eta_0}{|\theta_0|} < 1$ for all $0 \le t \le t_0$. Therefore , $\prod_{k=0}^{t_0}(1 - \frac{\eta_k}{\sqrt{v_k}}) = e^{\sum_{k=0}^{t_0}\log(1-\frac{\eta_k}{\sqrt{v_k}})} \le e^{-\sum_{k=0}^{t_0}\frac{\eta_k}{\sqrt{v_k}}} \le e^{-\frac{1}{|\theta_0|}\sum_{k=0}^{t_0}\eta_k} \le e^{-\frac{\eta_0}{(1-\alpha)|\theta_0|}\left((t_0+2)^{1-\alpha}-1\right)} \le e^{-\frac{\eta_0}{(1-\alpha)|\theta_0|}\left(t_0^{1-\alpha}-1\right)}$. Therefore $|\theta_t| \le |\theta_0| e^{-\frac{\eta_0}{(1-\alpha)|\theta_0|}\left(t_0^{1-\alpha}-1\right)}$. By assumption, $s := t_0^{1-\alpha}$ is sufficiently large. Therefore $|\theta_{t_0}| := \delta$ is sufficiently small, whereas $\frac{1}{2}|\theta_0|^2 \le v_{t_0} \le |\theta_0^2|$.

Stage 2 (Decay of the Adaptive Preconditioners). With the same argument of Theorem D.2(ii), we have

$$v_t \le (v_{t_0+1} - \delta^2)\beta_2^{t-t_0-1} + \delta^2,$$

Solving $\eta_T = 3\delta$, we have $T = (\frac{\eta_0}{3\delta})^{\alpha} - 1$. Then $v_T \le (v_{t_0+1} - \delta^2)\beta_2^{T-t_0-1} + \delta^2$. Therefore

$$\frac{\eta_T}{\sqrt{v_T}} \ge \frac{3\delta}{\sqrt{(v_{t_0+1} - \delta^2)\beta_2^{T-t_0-1} + \delta^2}} = \frac{3}{\sqrt{(v_{t_0+1} - \delta^2)\frac{\beta_2^{T-t_0-1}}{\delta^2} + 1}}.$$

By calculation,

$$\frac{\beta_2^{T-t_0-1}}{\delta^2} = e^{\left((\frac{\eta_0}{3\delta})^{\alpha} - \frac{\log 2}{\log\frac{1}{\beta_2}} - 2\right)\log\beta_2 - 2\log\delta}.$$

When $\beta_2 \to 1$, $\log\beta_2 \to 0, \delta \to 0$, but δ is of the form $e^{(\frac{c_1}{\log\beta_2})^{c_2}}$ with $c_1, c_2 > 0$. Intuitively, $\delta << \log\beta_2$. From $e^{(\frac{c_1}{\log\beta_2})^{c_2}}$ with $c_1, c_2 > 0$, one may verify that

$$\left((\frac{\eta_0}{3\delta})^{\alpha} - \frac{\log 2}{\log\frac{1}{\beta_2}} - 2\right)\log\beta_2 - 2\log\delta \to -\infty.$$

So $\frac{\beta_2^{T-t_0-1}}{\delta^2} \to 0$. Thus, $\frac{\eta_T}{\sqrt{v_T}} > 2$ when β_2 is sufficiently close to 1. This breaks the stability condition. □

E Discussion: The Pros and Cons of Loss Spikes

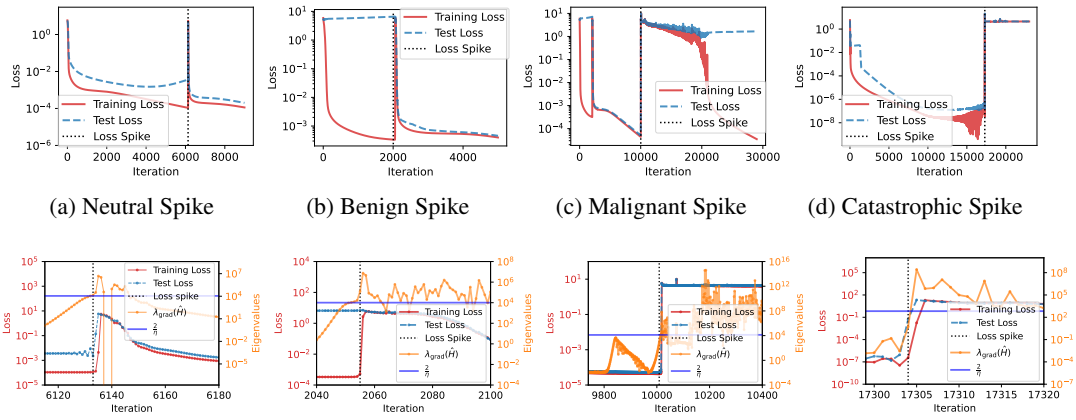
Connection to Generalization Transitions. Loss spikes represent more than mere optimization phenomena; they may signify transitions between distinct attractor basins in the optimization landscape. To systematically investigate the relationship between loss spikes and generalization, we conducted controlled experiments using a Transformer model. The model was trained to identify specific anchors within sequences, using a dataset of 2,000 samples (1,800 training, 200 test). We employed full-batch Adam optimization for training (detailed experimental setups and dataset specifications are provided in Appendix F). By analyzing the differential impacts on training and test losses before and after spike occurrences, we identified four distinct categories of loss spikes:

(i) Neutral Spikes (Fig. D2(a)): Both training and test losses resume their normal declining trajectory following the spike, suggesting minimal impact on the overall optimization process.

(ii) Benign Spikes (Fig. D2(b)): Prior to the spike, training loss reaches very low values while test loss remains elevated, indicating overfitting. After the spike, test loss decreases rapidly, suggesting improved generalization performance.

(iii) Malignant Spikes (Fig. D2(c)): Before the spike, both training and test losses achieve low values. After the spike, while training loss continues to decrease normally, test loss plateaus, indicating deteriorated generalization.

(iv) Catastrophic Spikes (Fig. D2(d)): Both training and test losses are low before the spike but neither recovers afterward, signifying a complete breakdown of the optimization process. These findings demonstrate that loss spikes can have context-dependent effects on generalization—sometimes enhancing model performance while in other cases degrading performance.



(a) Neutral Spike (b) Benign Spike (c) Malignant Spike (d) Catastrophic Spike

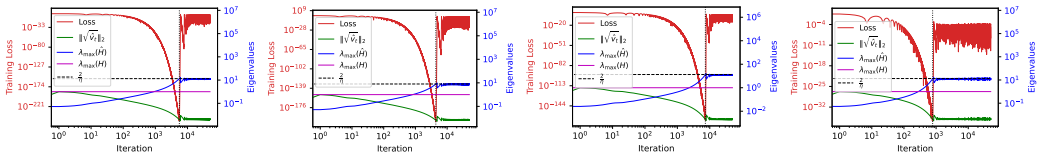
(e) Neutral Spike (f) Benign Spike (g) Malignant Spike (h) Catastrophic Spike

Figure D2: The Transformer model was trained to identify specific anchors within sequences. (a–d) Evolution of the training and test losses over the course of training. (e-h) Evolution of the eigenvalues in the gradient direction $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ near the spike.

As shown in Fig. D2(e–h), all four types of spikes correspond to our proposed indicator, $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$, exceeding the classical stability threshold $2/\eta$. Despite this commonality, their effects on generalization differ significantly. While our study uncovers the underlying mechanism that triggers these spikes, determining the precise conditions under which a spike becomes benign or malignant remains an open question for future research.

F Supplementary Experiments

Optimization of Quadratic Function with Varying Hyper-parameters. For the optimization of a one-dimensional quadratic function, Fig. D3 illustrates the precise location of the spike under various hyperparameter configurations, where $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ exceeds the stability threshold $\frac{2}{\eta}$.



(a) $\eta = 0.15, \beta_1 = 0.9, \beta_2 = 0.999$ (b) $\eta = 0.25, \beta_1 = 0.9, \beta_2 = 0.999$ (c) $\eta = 0.15, \beta_1 = 0.95, \beta_2 = 0.999$ (d) $\eta = 0.15, \beta_1 = 0.9, \beta_2 = 0.99$

Figure D3: Optimization of $f(\theta) = \frac{1}{2}\theta^2$ using the Adam algorithm with different hyperparameter settings. The solid red line denotes the training loss. The dashed black line indicates the stability threshold $\frac{2}{\eta}$. The blue, purple, and green solid lines represent $\lambda_{\max}(\boldsymbol{H}_t)$, $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$, and the bias-corrected $\|\sqrt{\hat{\boldsymbol{v}}_t}\|_2$, respectively, at each training step.

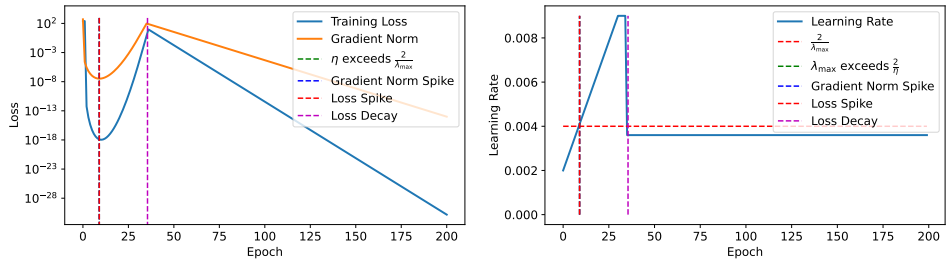
Delay Mechanism in Gradient Descent

To verify that in high-dimensional cases, when $\lambda_{\max} > \frac{2}{\eta}$, the maximum eigenvalue direction oscillates while other eigenvalue directions steadily decrease (resulting in overall loss reduction), we conducted experiments on one and two-dimensional quadratic functions with varying learning rates.

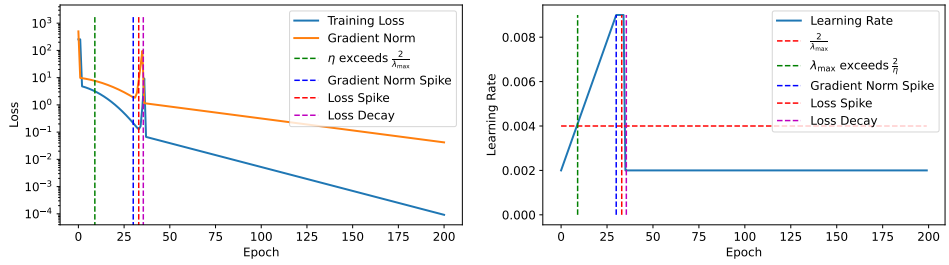
For a one-dimensional quadratic function, the loss landscape curvature remains constant. In this setting, the learning rate initially produces linear improvement over time, followed by gradual decay. When the instability condition is met—as illustrated in Fig. D4(a)—the loss increases immediately.

In contrast, for the two-dimensional case, instability primarily emerges along the dominant eigendirection, while other directions continue to descend stably. As shown in Fig. D4(b), this leads to a delayed onset of the loss spike.

To further validate this mechanism, we visualize the training trajectories in Fig. D5(a–b). In gradient descent (GD), the component along the maximum eigenvalue direction is learned rapidly at first, resulting in a small magnitude. However, once the instability condition is triggered, this component requires significant time to grow and eventually dominate the dynamics.

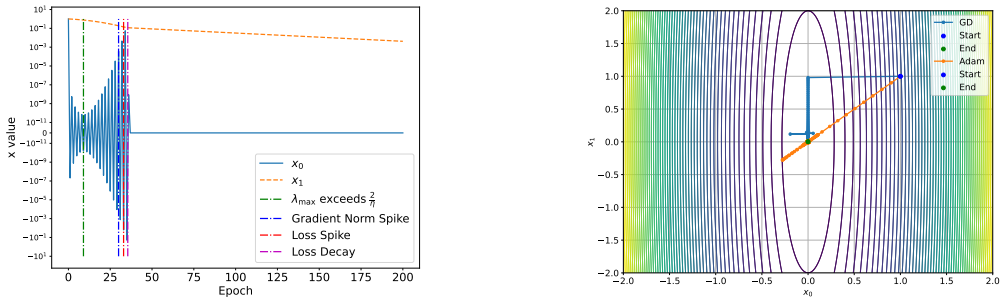


(a) 1d-quadratic



(b) 2d-quadratic

Figure D4: Delay mechanism in gradient descent: Comparison of loss dynamics for 1D and 2D quadratic functions. The learning rate varies over the course of training.



(a) Parameter value (b) Trajectory

Figure D5: Training dynamics for the 2D quadratic function under gradient descent. (a) Evolution of the solution components along different eigendirections. (b) Optimization trajectory in parameter space.

Gradient-direction Curvature vs. Update-direction Curvature for Loss Spike Prediction

For Adam, where the Hessian is preconditioned, we define the predictor as

$$\lambda_{\text{grad}}(\hat{\boldsymbol{H}}) := \frac{\nabla L(\boldsymbol{\theta}_t)^\top \hat{\boldsymbol{H}} \nabla L(\boldsymbol{\theta}_t)}{\|\nabla L(\boldsymbol{\theta}_t)\|^2},$$

where $\hat{\boldsymbol{H}}$ denotes the preconditioned Hessian in Eq. (5).

We also define

$$\lambda_{\text{update}}(\hat{\boldsymbol{H}}) := \frac{\boldsymbol{u}_t^\top \hat{\boldsymbol{H}} \boldsymbol{u}_t}{\|\boldsymbol{u}_t\|^2},$$

where $\boldsymbol{u}_t = \frac{\hat{\boldsymbol{m}}_t}{\sqrt{\hat{\boldsymbol{v}}_t}+\varepsilon}$ is the update vector.

To validate our quadratic approximation-based predictor, we tracked the eigenvalue evolution of the preconditioned Hessian throughout training. Fig. D6(b) reveals that while $\lambda_{\max}(\boldsymbol{H}_t)$ quickly stabilizes, $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ continues to increase steadily. Notably, $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ surpasses the stability threshold $\frac{2}{\eta}$ at epoch 179, yet no immediate spike occurs. At epoch 184, precisely when $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ exceeds $\frac{2}{\eta}$, we observe the loss spike depicted in Fig. D6(a). Subsequently, the eigenvalue $\lambda_{\text{update}}(\hat{\boldsymbol{H}}_t)$ in the parameter update direction also exceeds $\frac{2}{\eta}$.

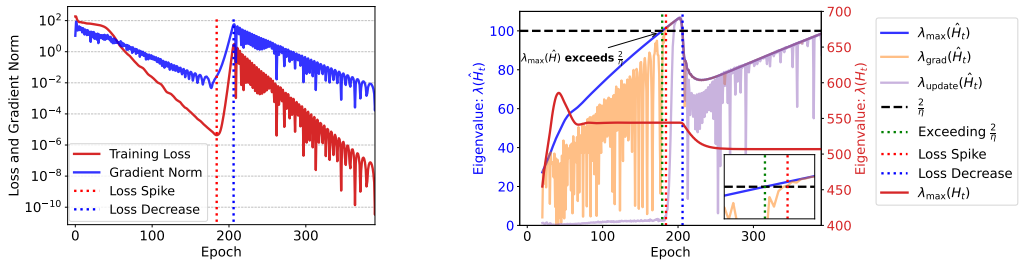
This demonstrates that the eigenvalue in the gradient direction more accurately predicts the onset of the actual spike. The update direction requires time to respond to changes in the gradient. When λ_{update} exceeds $2/\eta$, the loss spike has already occurred.

CIFAR-10 Experiments

We trained a convolutional neural network on CIFAR10 using Adam hyperparameters $\beta_1 = 0.9, \beta_2 = 0.999$. As shown in Fig. D7(a), the optimization follows a pattern similar to FNN, with an initial loss decrease followed by three distinct spikes. Analysis of the preconditioned Hessian's eigenvalues (Fig. D7(b)) shows $\lambda_{\max}(\boldsymbol{H}_t)$ remaining below the stability threshold $2/\eta$, while $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ increases until exceeding it. Loss spikes occur precisely when $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ surpasses $2/\eta$. Figs. D7(c-d) show the evolution of squared gradients and second-order moments $\sqrt{\hat{\boldsymbol{v}}_t}$ across parameter blocks. Before spikes, $\|\boldsymbol{g}_t\|$ is much smaller than $\|\sqrt{\hat{\boldsymbol{v}}_t}\|$, with $\hat{\boldsymbol{v}}_t$ decaying exponentially at rate $\approx \beta_2$. During spikes, while $\hat{\boldsymbol{v}}_t$ continues decreasing, the gradient norm increases until substantially impacting $\boldsymbol{v}_t$. Subsequently, $\hat{\boldsymbol{v}}_t$ rises, causing $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ to fall below $2/\eta$ and allowing loss descent to resume.

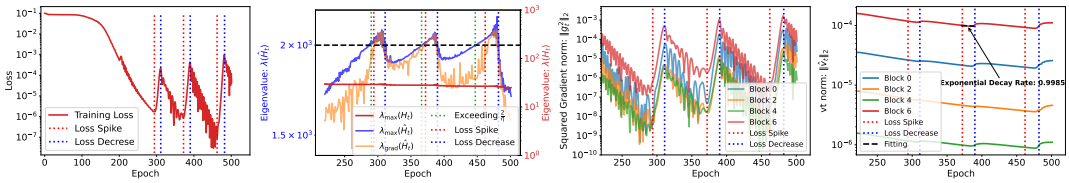
Transformer Models for Sequence Learning

For the experiment illustrated in Fig. 7, Fig. D9 presents the complete evolution of all eigenvalues, along with detailed views of each spike in Fig. 7(c-e) and Fig. D10(a-d).



(a) Loss (b) Eigenvalues

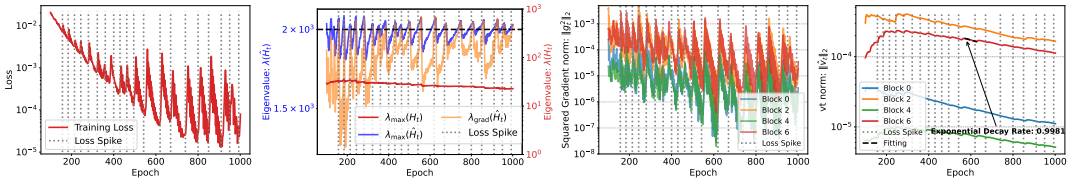
Figure D6: (a) Training loss and gradient norm over time. (b) Evolution of critical eigenvalues: original Hessian maximum eigenvalue $\lambda_{\max}(\boldsymbol{H}_t)$, preconditioned Hessian maximum eigenvalue $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$, gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ and update-directional eigenvalue $\lambda_{\text{update}}(\hat{\boldsymbol{H}}_t)$ relative to $2/\eta$.



(a) Loss (b) Eigenvalues (c) Squared gradient (d) Second moment

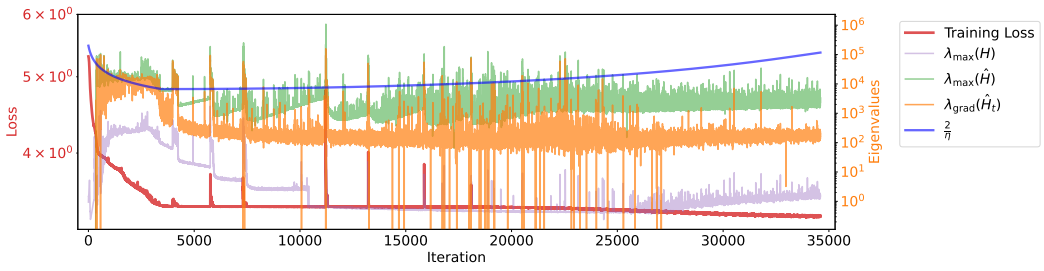
Figure D7: Training a CNN on 50 randomly selected CIFAR-10 images to illustrate the detailed spikes (see similar result for larger datasets in Appendix F Fig. D8). (a) Training loss over time. (b) Evolution of eigenvalues: original Hessian maximum eigenvalue $\lambda_{\max}(\boldsymbol{H}_t)$, preconditioned Hessian maximum eigenvalue $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$, and gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ relative to $2/\eta$ (black dashed line). (c) Gradient norm evolution across parameter blocks. (d) L_2-norm of second moment estimate $\|\hat{\boldsymbol{v}}_t\|$ of different parameter blocks.

As depicted in Fig. D10(a-d), we found that transient periods where $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ and $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ exceed $2/\eta$ are insufficient to induce a spike. Loss spikes only materialize when $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ remains above the threshold for a sustained duration. This observation aligns with stability analysis principles, which suggest that loss increases exponentially only after persistent instability, with isolated threshold violations being insufficient to trigger rapid loss elevation. Based on this insight, we formulated a

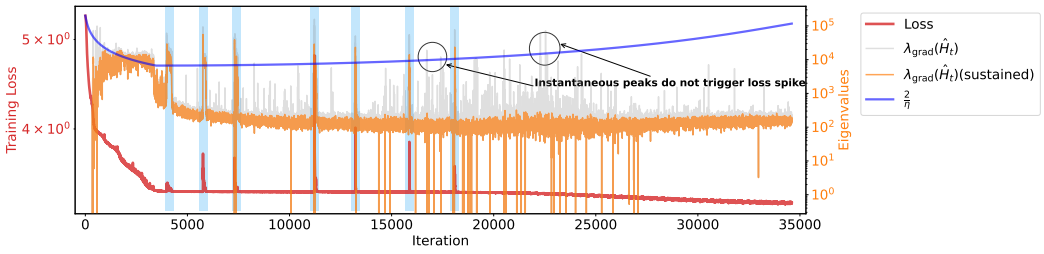


(a) Loss (b) Eigenvalues (c) Squared gradient (d) Second moment

Figure D8: Loss spike in CNNs on CIFAR10 for randomly sampled 1000 images. (a) Temporal evolution of training loss. (b) Progression of critical eigenvalue metrics: original Hessian maximum eigenvalue $\lambda_{\max}(\boldsymbol{H}_t)$, preconditioned Hessian maximum eigenvalue $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$, and gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ relative to the stability threshold $\frac{2}{\eta}$ (black dashed line). (c) Temporal evolution of gradient norm of different parameter blocks. (d) L_2-norm of second moment $\|\hat{\boldsymbol{v}}_t\|$ of different parameter blocks.



(a) Eigenvalues



(b) $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$

Figure D9: (a) Evolution of critical eigenvalues: original Hessian maximum eigenvalue $\lambda_{\max}(\boldsymbol{H}_t)$, preconditioned Hessian maximum eigenvalue $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ and gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ relative to $2/\eta$. (b) Gradient-directional eigenvalues $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ (gray) and sustained predictor $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)(\text{sustained})$ (orange) vs. $2/\eta$.

"sustained spike predictor" defined as:

$$\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)(\text{sustained}) = \min(\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_{t-1}), \lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t), \lambda_{\text{grad}}(\hat{\boldsymbol{H}}_{t+1})).$$

This refined predictor demonstrates perfect correspondence with loss spike occurrences, as shown by the orange line in Fig. D9(b).

Controlling Adaptive Preconditioners to Eliminate Spikes

We discovered that the epsilon parameter (ε) in Adam plays a critical role in modulating loss spike behavior. Specifically, using a larger ε significantly reduces spike severity by effectively imposing an upper bound on the preconditioned eigenvalues. Additionally, we experimented with component-wise clipping of $\boldsymbol{v}_t$, where elements falling below a specified threshold are clipped to that threshold value.

As shown in Fig. D12(a), locally increasing ε during training can effectively suppress loss spikes. Fig. D12(b) further demonstrates that increasing ε or applying $\boldsymbol{v}_t$ clipping from the beginning of training can also mitigate spike behavior, although this may come at the cost of slower convergence.

G Experimental Setup

All experiments were conducted on 1 NVIDIA RTX 4080 GPU. The runtime varied across tasks, ranging from a few minutes for smaller models to several days for large-scale training.

Computing the full Hessian matrix for large-scale neural networks is computationally prohibitive due to its quadratic memory complexity. To address this challenge, we employ an efficient power iteration method combined with Hessian-vector products that leverages automatic differentiation, circumventing the explicit construction of the complete Hessian matrix.

Setup for Fig. 6 and Fig. 1(a). We trained two-layer fully connected neural network applied to a high-dimensional function approximation task. The target function is defined as $f^*(\boldsymbol{x}) = \boldsymbol{w}^{*\top}\boldsymbol{x} + \boldsymbol{x}^\top \text{diag}(\boldsymbol{v}^*)\boldsymbol{x}$, where $\boldsymbol{w}^*, \boldsymbol{v}^* \in \mathbb{R}^{50}$ are the ground-truth parameters and $\boldsymbol{x} \in \mathbb{R}^{50}$ denotes the input features. A total of $n = 200$ data points are sampled, with inputs drawn from a standard

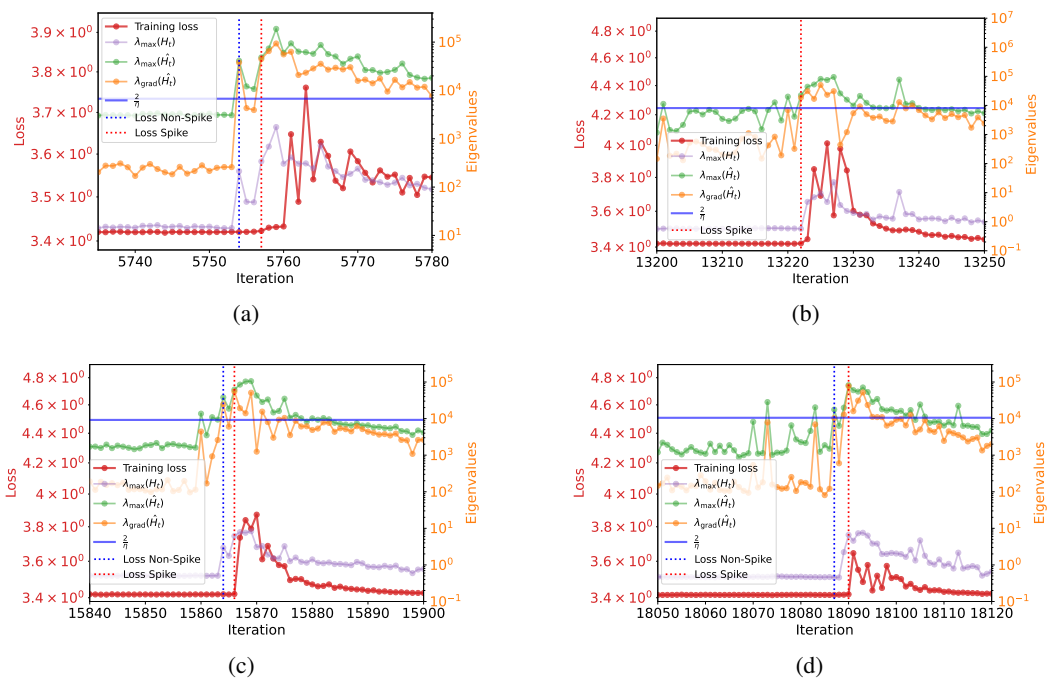


Figure D10: Detailed inspection of loss spike intervals showing the maximum eigenvalues of the original Hessian $\lambda_{\max}(\boldsymbol{H}_t)$, preconditioned Hessian $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$, and $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$.

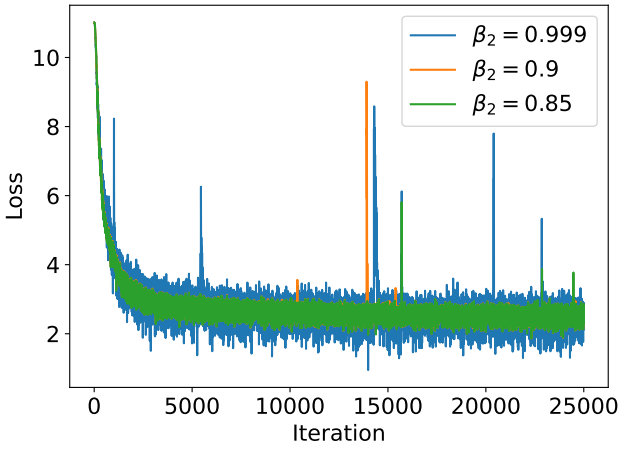
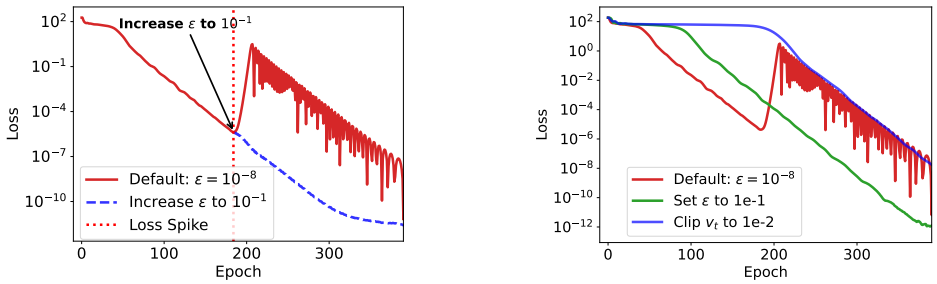


Figure D11: The raw loss of the Fig. 8(a).

Gaussian distribution. Gaussian noise with standard deviation $\varepsilon = 0.1$ is added to the outputs. The network has a hidden layer width of $m = 1000$, placing it in the over-parameterized regime. All weights are initialized from a Gaussian distribution $\mathcal{N}(0, \frac{1}{m})$. Training is performed using full-batch Adam with a learning rate of $\eta = 0.02$, and momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Setup for Fig. D7 and Fig. 1(b). We trained a convolutional neural network on the CIFAR-10 dataset. For computational tractability in computing Hessian eigenvalues, we restricted the training set to 50 randomly sampled images. The network contains approximately $500,000$ parameters and is trained using Mean Squared Error (MSE) loss with one-hot encoded labels. Optimization is performed



(a) (b)

Figure D12: The training loss with the same experiment settings as Fig. 6. (a) The only difference of the blue solid line is that we change the ε in Adam to 0.1 at epoch 184 where the loss in the original training process begin to spike. (b) The green solid line is the training loss that we change the ε to 0.1 at the beginning of the training. The blue solid line is the training loss that we clip the v_t in Adam to 0.01.

using full-batch Adam with a learning rate of $\eta = 0.001$ and default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Setup for Fig. D17(a,b). We trained a ViT on the CIFAR-10 dataset. The ViT consists of 4 layers and 8 heads. The embedding dimension is 64. The network is trained using Mean Squared Error (MSE) loss with one-hot encoded labels. Optimization is performed using full-batch Adam with a learning rate of $\eta = 0.001$ and default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Setup for Fig. D17(c,d). We trained a ResNet on the CIFAR-10 dataset. The network is trained using Mean Squared Error (MSE) loss with one-hot encoded labels. Optimization is performed using full-batch Adam with a learning rate of $\eta = 0.001$ and default momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Setup for Fig. 7 and Fig. 1(d). We implemented an 8-layer standard Transformer with approximately 10 million parameters. The model is trained on a synthetic dataset designed to learn compositional rules from sequences (Zhang et al., 2025), consisting of $900,000$ sequences. Training uses a batch size of 2048 and follows the next-token prediction paradigm with cross-entropy loss. The learning rate follows a linear warm-up stage followed by cosine decay. Optimization is performed using Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Setup for Fig. 8 and Fig. D11 We implemented a LLaMA structure Transformer with 187M non-embedding parameters and trained on 100B data split from SlimPajama. The detailed hyperparameters are shown in Table 1.

Setup for Fig. D2, Fig. D13 and Fig. 1(c). We further evaluate our theoretical insights using 4-layer (Fig. D2, Fig. D13) and 12-layer ((Fig. D2, Fig. 1(c))) standard Transformers trained on a synthetic classification task. The dataset is constructed to learn a specific anchor rule ($3x \to x$) from sequences (Zhang et al., 2025), comprising $2,000$ sequences. The model is trained using cross-entropy loss. The learning rate follows a linear warm-up followed by cosine decay. Adam is used for optimization with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Setup for Fig. D20 We trained two-layer fully connected neural network applied to a high-dimensional function approximation task. The target function is defined as $f^*(\boldsymbol{x}) = \boldsymbol{w}^{*\top}\boldsymbol{x} + \boldsymbol{x}^\top \mathrm{diag}(\boldsymbol{v}^*)\boldsymbol{x}$, where $\boldsymbol{w}^*, \boldsymbol{v}^* \in \mathbb{R}^{50}$ are the ground-truth parameters and $\boldsymbol{x} \in \mathbb{R}^{50}$ denotes the input features. A total of $n = 200$ data points are sampled, with inputs drawn from a standard Gaussian

Hyperparameter	Value
Number of Layers	16
Hidden Size	1280
FFN Inner Hidden Size	1280
Attention Heads	16
Attention Head Size	80
Batch Size	512
Learning Rate Scheduler	10% Warmup + Cosine Annealing
Adam β_1	0.9
Adam β_2	0.999; 0.9; 0.85
Adam ϵ	10^{-8}
Gradient Clipping	1.00

Table 1: Detailed Hyperparameters for the 187M Transformer.

distribution. Gaussian noise with standard deviation $\varepsilon = 0.1$ is added to the outputs. The network has a hidden layer width of $m = 1000$, placing it in the over-parameterized regime. All weights are initialized from a Gaussian distribution $\mathcal{N}(0, \frac{1}{m})$. Training is performed using full-batch Adam with a learning rate of $\eta = 0.002$, momentum parameter $\beta_1 = 0.0$, and different variations of β_2.

G.1 Practical feasibility of our monitoring approach.

For $\lambda_{\max}(\hat{H}_t)$: We do not need the full spectral information or all eigenvalues. To estimate the maximum eigenvalue, we employ the power iteration method, which requires only multiple Hessian-vector products. Specifically, starting from a random vector $\mathbf{v}_0$, power iteration performs:

$$\mathbf{v}_{k+1} = \frac{\hat{H}_t \mathbf{v}_k}{\|\hat{H}_t \mathbf{v}_k\|},$$

and the largest eigenvalue is approximated by $\mathbf{v}_k^\top \hat{H}_t \mathbf{v}_k$. This converges rapidly (typically 5-10 iterations) and each iteration costs only $O(n)$ via automatic differentiation, requiring no explicit Hessian construction. The total cost is $O(kn)$ where $k \ll n$ is the number of power iterations—entirely tractable even for large models.

For our predictor $\lambda_{\text{grad}}(\hat{H}_t)$: The computational cost is even lower. By definition,

$$\lambda_{\text{grad}}(\hat{H}_t) = \frac{g_t^\top \hat{H}_t g_t}{\|g_t\|^2},$$

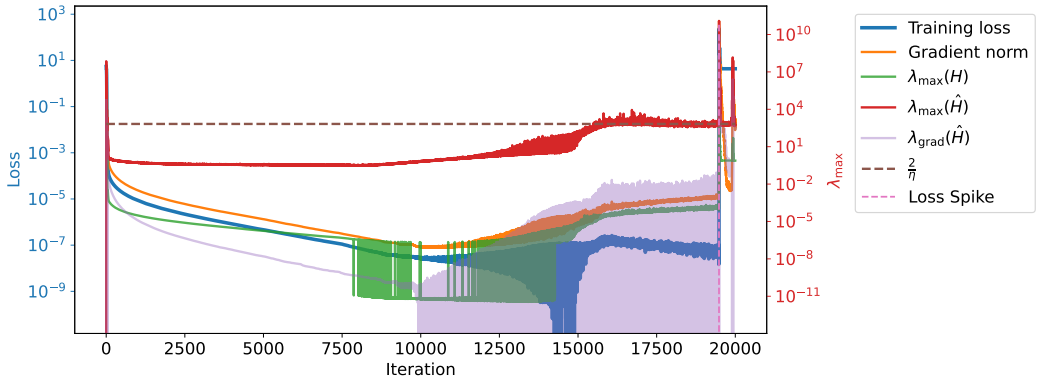
which requires only a single Hessian-vector product $\hat{H}_t g_t$ in the gradient direction. This is precisely "a single projection", but this is not a limitation—it is exactly the relevant information for predicting loss spikes. We do not need full spectral information; we only need the curvature in the direction the optimizer is moving, which is captured by this single directional derivative.

H New Supplementary Experiments

Compared to research on Edge of Stability (EoS). Several papers on EoS have noted the close relationship between η and $2/\lambda_{\max}(H)$ in modern deep learning as discussed in the main text. However, these phenomena are typically characterized as edge-of-stability behavior, which differs from the large, pronounced loss spikes we observe. The precise relationship between these instabilities and observed spikes remains unclear—instability may manifest as oscillations or as spikes, but the

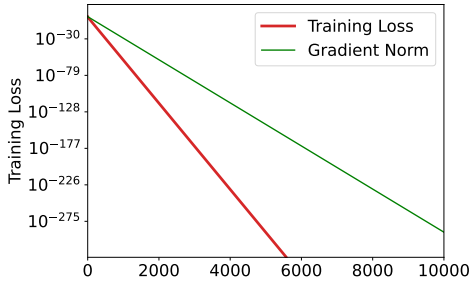
specific mechanism under which spikes occur is not well understood. As shown in our experiment (Figure D13), the system can remain in the EoS region for extended periods, but spikes occur specifically when the curvature in the gradient direction $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ exceeds $2/\eta$. Our work reveals how $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ increases, how larger β_2 leads to persistent instability and identifies that spikes occur precisely when the curvature in the gradient direction $\lambda_{\text{grad}}(\hat{H}_t)$ exceeds $2/\eta$, rather than $\lambda_{\max}(H)$ as discussed in EoS literature. To our best knowledge, no prior work has explicitly identified these mechanisms.

Why understanding quantitative mechanisms matters: Loss spikes are notoriously difficult to study due to their strong correlations with numerous factors, leading to many seemingly plausible but ambiguous explanations without causal understanding. We emphasize that mechanistic understanding and quantitative prediction are crucial because they typically indicate causality.



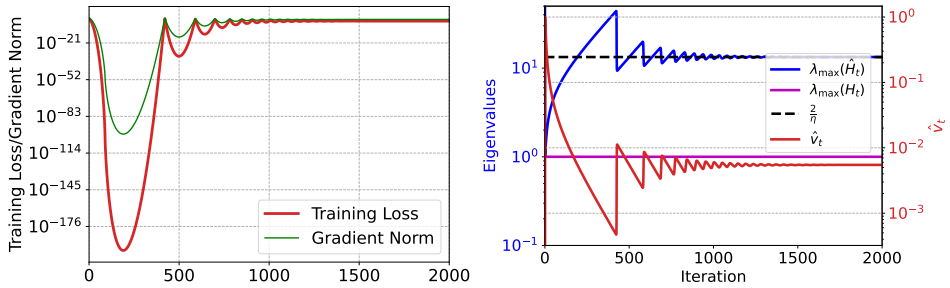
(a)

Figure D13: (a) Evolution of critical eigenvalues of a $3x \to x$ task (Zhang et al., 2025): original Hessian maximum eigenvalue $\lambda_{\max}(\boldsymbol{H}_t)$, preconditioned Hessian maximum eigenvalue $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$ and gradient-directional eigenvalue $\lambda_{\text{grad}}(\hat{\boldsymbol{H}}_t)$ relative to $2/\eta$.

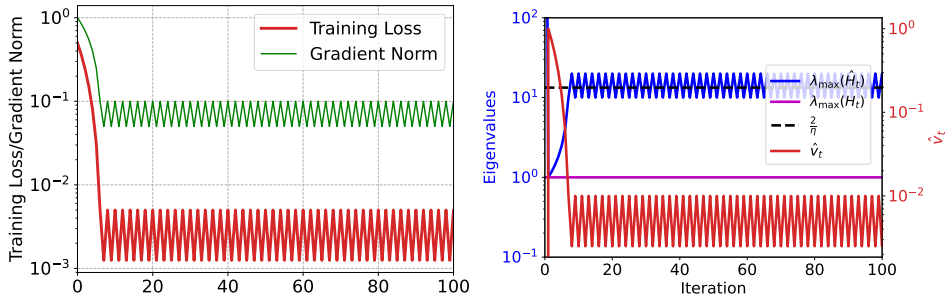


(a)

Figure D14: Adagrad optimization on $f(\theta) = \frac{1}{2}\theta^2$. AdaGrad's second-moment estimate follows $v_t = v_{t-1} + g_t^2$, which is a strict accumulation. This ensures the effective learning rate $\eta/\sqrt{v_t}$ can only decrease monotonically over time, precluding the possibility of preconditioner decay that our theory identifies as the root cause of spikes.

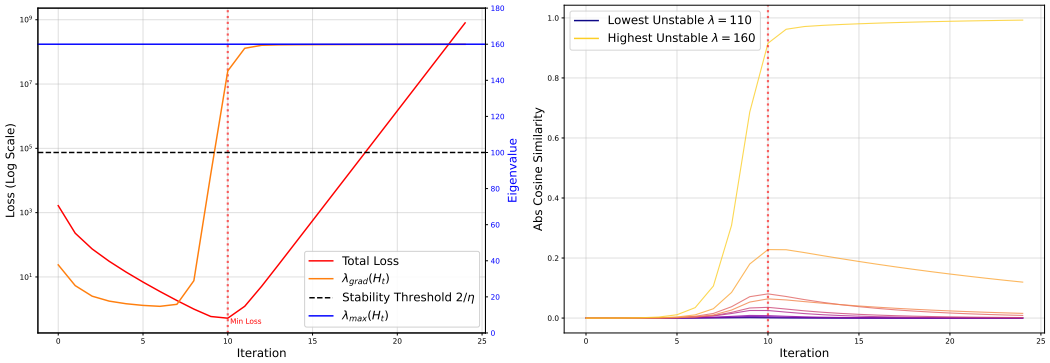


(a) $\beta_2 = 0.99$ (b) $\beta_2 = 0.99$



(c) $\beta_2 = 0$ (d) $\beta_2 = 0$

Figure D15: RMSProp optimization ($\beta_1 = 0$ in Adam) on $f(\theta) = \frac{1}{2}\theta^2$ with $\beta_2 = 0.99$ and 0.00. (a, c) Evolution of training loss and gradient norm. (b, d) Evolution of the second moment estimate $\hat{\boldsymbol{v}}_t$ and the maximum eigenvalue of the preconditioned Hessian.



(a) (b)

Figure D16: The optimization for a 100-dimensional quadratic function with gradient descent. $\eta = 0.02$ and there are 90 stable direction that $\lambda < 100$ and 10 unstable direction that $\lambda > 100$. (a) Evolution of loss and critical eigenvalues: Hessian maximum eigenvalue $\lambda_{\max}(\boldsymbol{H}_t)$ and gradient-directional eigenvalue $\lambda_{\text{grad}}(\boldsymbol{H}_t)$ relative to $2/\eta$. (b) Cosine similarity between gradient direction and 10 unstable directions. When spikes occur, the gradient direction aligns predominantly with the most unstable eigendirection (i.e., the one corresponding to $\lambda_{\max}(\boldsymbol{H}_t)$), as this direction dominates the optimization dynamics.

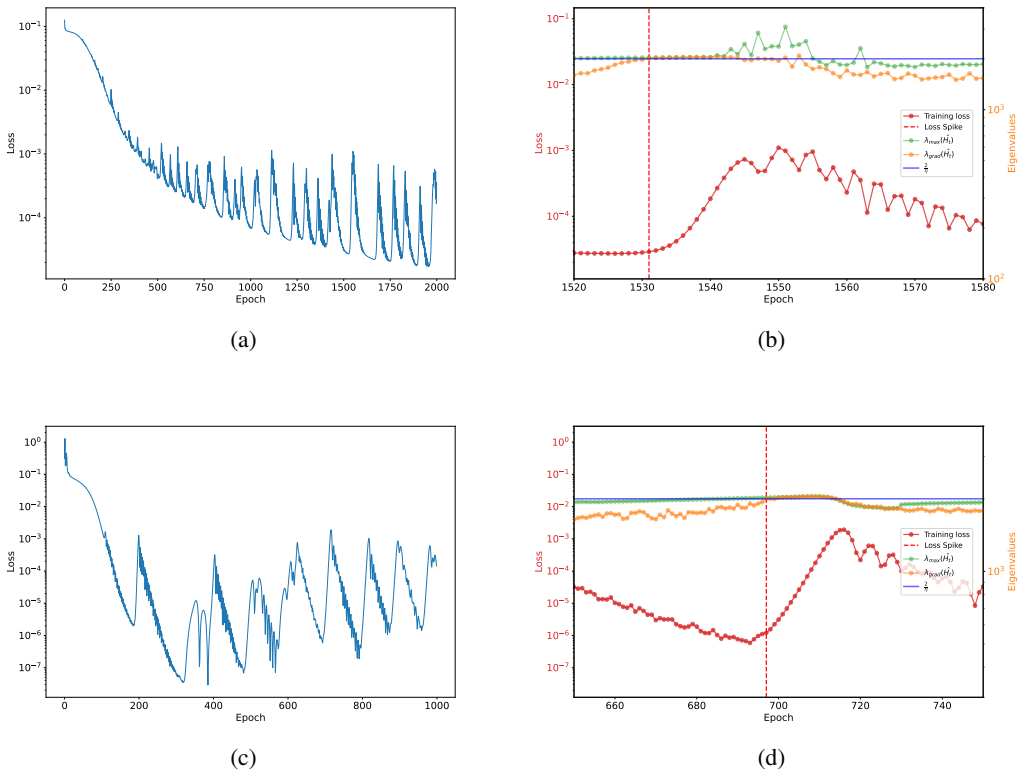


Figure D17: (a,c) The training loss of ViT and ResNet18 model on randomly selected 1000 CIFAR-10 images respectively. (b,d) Detailed inspection of loss spike intervals showing the maximum eigenvalues of the preconditioned Hessian $\lambda_{\max}(\hat{\boldsymbol{H}}_t)$, and gradient-directional eigenvalue $\lambda_{\text{grad}}(\boldsymbol{H}_t)$ relative to $2/\eta$.

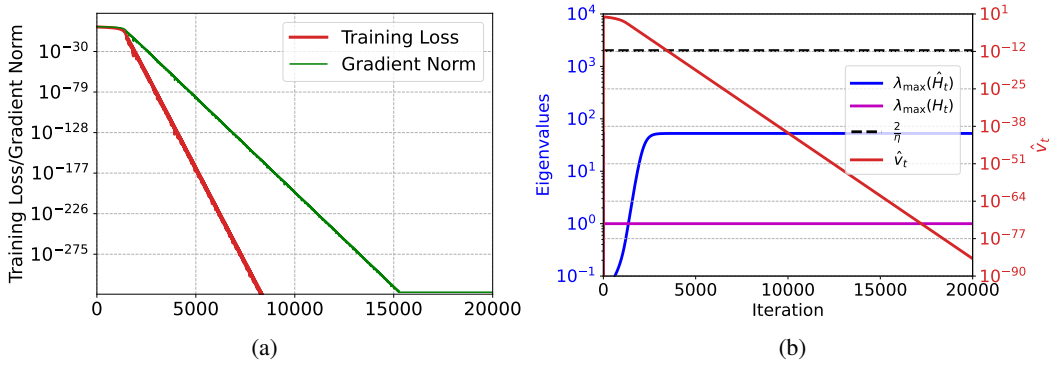
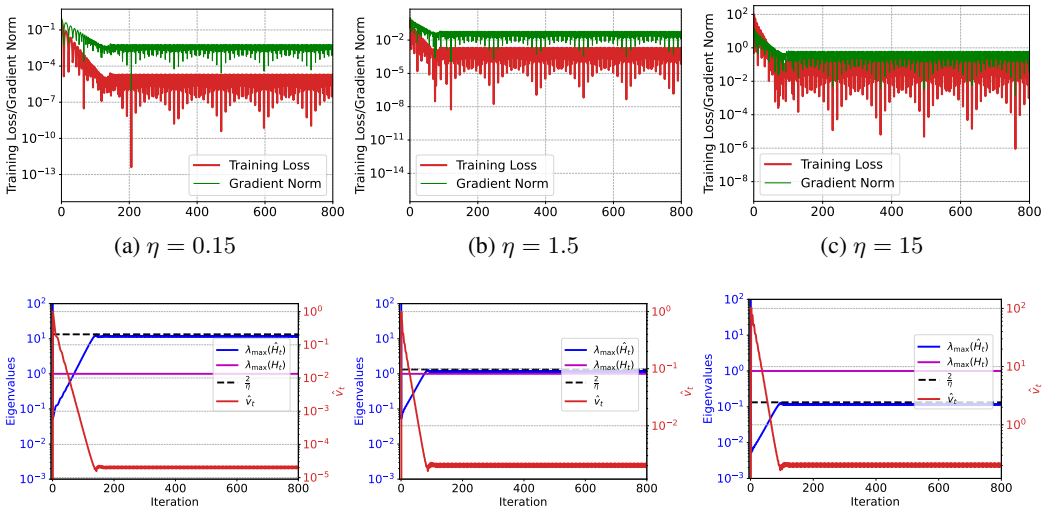


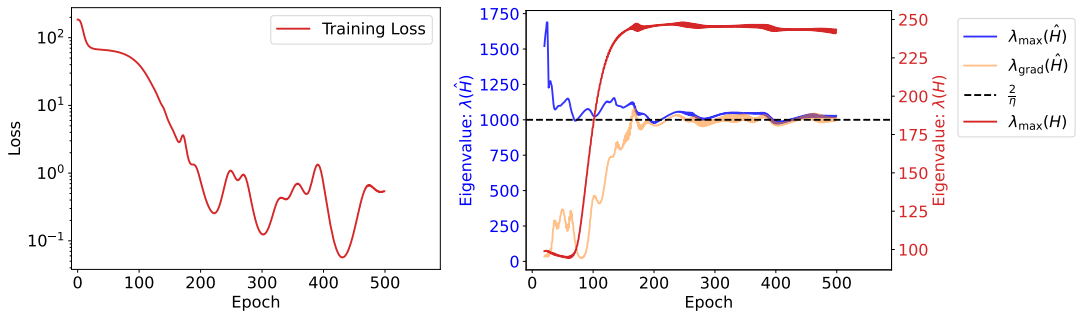
Figure D18: The loss spike in Figure 2 is not caused by rounding errors. Adam optimization on $f(\theta) = \frac{1}{2}\theta^2$ with a large $\epsilon = 10^{-3}$ values and learning rate 0.001. (a) Evolution of training loss and gradient norm. (b) Evolution of the second moment estimate $\hat{\boldsymbol{v}}_t$ and the maximum eigenvalue of the preconditioned Hessian. We increase Adam's ϵ parameter to 10^{-3} to ensure that $\lambda_{\text{grad}}(\hat{H}_t)$ can not exceed $2/\eta$, Adam can converge to loss values as low as 10^{-300}.



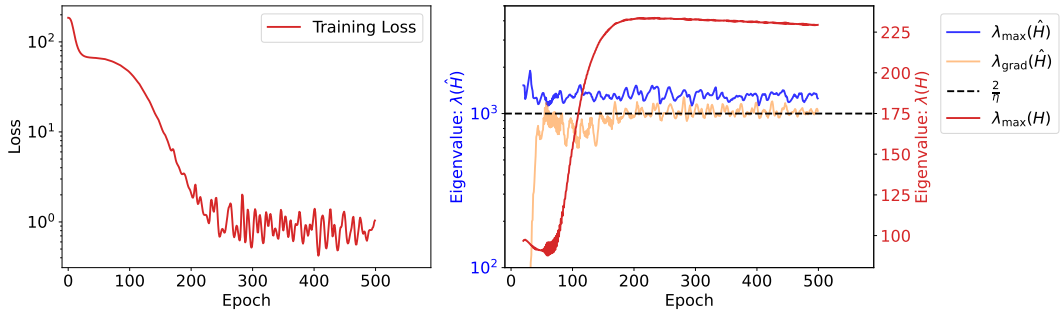
(a) $\eta = 0.15$ (b) $\eta = 1.5$ (c) $\eta = 15$

(d) $\eta = 0.15$ (e) $\eta = 1.5$ (f) $\eta = 15$

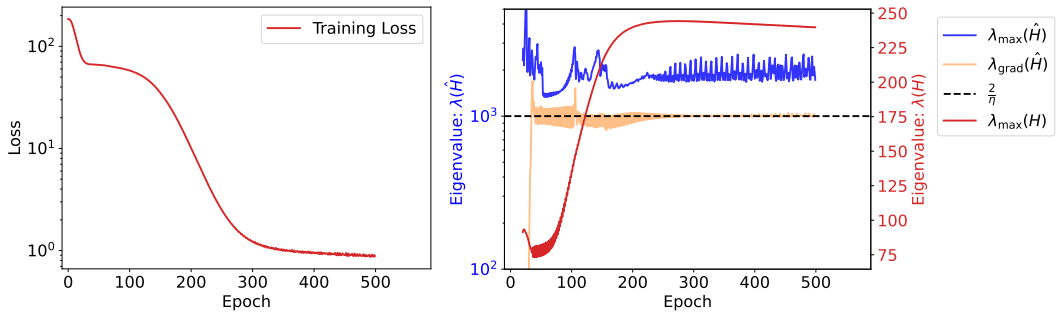
Figure D19: Stable loss decrease is still observed initially even with larger learning rates in the case of $\beta_2 = 0.9$. Our results show that when the learning rate is particularly large, v_t grows rapidly in the early stages of optimization. This rapid growth of v_t effectively reduces the preconditioned step size $\eta/\sqrt{v_t}$, which allows the loss to decrease stably at the beginning even under large nominal learning rates.



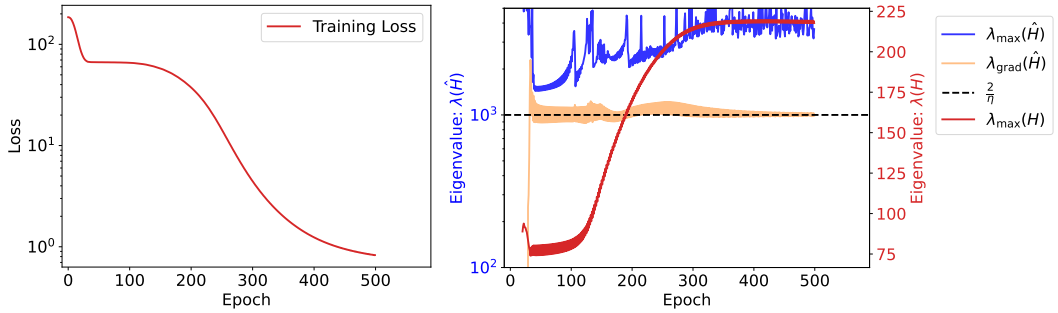
(a) Loss, $\beta_2 = 0.999$ (b) Eigenvalues, $\beta_2 = 0.999$



(c) Loss, $\beta_2 = 0.9$ (d) Eigenvalues, $\beta_2 = 0.9$



(e) Loss, $\beta_2 = 0.5$ (f) Eigenvalues, $\beta_2 = 0.5$



(g) Loss, $\beta_2 = 0.1$ (h) Eigenvalues, $\beta_2 = 0.1$

Figure D20: Training trajectories and eigenvalue evolution for varying β_2 values with $\beta_1 = 0$ (to isolate the effect of adaptive learning rate from momentum). Each row shows the loss curve and corresponding evolution of $\lambda_{\max}(\hat{H}_t)$ and $\lambda_{\text{grad}}(\hat{H}_t)$ for a different β_2 setting. Larger β_2 values produce more pronounced spikes in the loss, while smaller β_2 values lead to denser oscillations, mirroring the behavior observed for the quadratic function in Fig. 3. Notably, loss spikes and oscillations correlate with λ_{grad} approaching $2/\eta$, rather than with $\lambda_{\max}(\hat{H}_t)$, providing empirical validation for the practical utility of our proposed λ_{grad} metric.