SAM operates far from home: eigenvalue regularization as a dynamical phenomenon

Atish Agarwala [*1] **Yann Dauphin** [*1]

Abstract

The Sharpness Aware Minimization (SAM) optimization algorithm has been shown to control large eigenvalues of the loss Hessian and provide generalization benefits in a variety of settings. The original motivation for SAM was a modified loss function which penalized sharp minima; subsequent analyses have also focused on the behavior near minima. However, our work reveals that SAM provides a strong regularization of the eigenvalues throughout the learning trajectory. We show that in a simplified setting, SAM dynamically induces a stabilization related to the edge of stability (EOS) phenomenon observed in large learning rate gradient descent. Our theory predicts the largest eigenvalue as a function of the learning rate and SAM radius parameters. Finally, we show that practical models can also exhibit this EOS stabilization, and that understanding SAM must account for these dynamics far away from any minima.

1. Introduction

Since the dawn of optimization, much effort has gone into developing algorithms which use geometric information about the loss landscape to make optimization more efficient and stable (Nocedal, 1980; Duchi et al., 2011; Lewis & Overton, 2013). In more modern machine learning, control of the large curvature eigenvalues of the loss landscape has been a goal in and of itself (Hochreiter & Schmidhuber, 1997; Chaudhari et al., 2019). There is empirical and theoretical evidence that controlling curvature of the training landscape leads to benefits for generalization (Keskar et al., 2017; Neyshabur et al., 2017), although in general the relationship between the two is complex (Dinh et al., 2017).

[*]Equal contribution [1]Google DeepMind. Correspondence to: Atish Agarwala <thetish@google.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

Recently the *sharpness aware minimization* (SAM) algorithm has emerged as a popular choice for regularizing the curvature during training (Foret et al., 2022). SAM has the advantage of being a tractable first-order method; for the cost of a single extra gradient evaluation, SAM can control the large eigenvalues of the loss Hessian and often leads to improved optimization and generalization (Bahri et al., 2022).

However, understanding the mechanisms behind the effectiveness of SAM is an open question. The SAM algorithm itself is a first-order approximation of SGD on a modified loss function $\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \max_{||\boldsymbol{\delta}||<\rho} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\delta})$. Part of the original motivation was that $\tilde{\mathcal{L}}$ explicitly penalizes sharp minima over flatter ones. However the approximation performs as well or better than running gradient descent on $\tilde{\mathcal{L}}$ directly. SAM often works better with small batch sizes as compared to larger ones (Foret et al., 2022; Andriushchenko & Flammarion, 2022).These stochastic effects suggest that studying the deterministic gradient flow dynamics on $\tilde{\mathcal{L}}$ will not capture key features of SAM, since small batch size induces non-trivial differences from gradient flow defined on the full batch objective (Paquette et al., 2021).

In parallel to the development of SAM, experimental and theoretical work has uncovered some of the curvature-controlling properties of first-order methods due to finite step size - particularly in the full batch setting. At intermediate learning rates, a wide variety of models and optimizers show a tendency for the largest Hessian eigenvalues to stabilize near the *edge of stability* (EOS) for long times (Lewkowycz et al., 2020; Cohen et al., 2022a;b; Jastrzebski et al., 2020). The EOS is the largest eigenvalue which would lead to convergence for a quadratic loss landscape. This effect can be explained in terms of a non-linear feedback between the large eigenvalue and changes in the parameters in that eigendirection (Damian et al., 2022; Agarwala et al., 2022).

We will show that these two areas of research are in fact intimately linked: under a variety of conditions, SAM displays a modified EOS behavior, which leads to stabilization of the largest eigenvalues at a lower magnitude via non-linear, discrete dynamics. These effects highlight the dynamical nature of eigenvalue regularization, and demonstrates that

SAM can have strong effects throughout a training trajectory.

1.1. Related work

Previous experimental work suggested that decreasing batch size causes SAM to display both stronger regularization and better generalization (Andriushchenko & Flammarion, 2022). This analysis also suggested that SAM may induce more sparsity.

A recent theoretical approach studied SAM close to a minimum, where the trajectory oscillates about the minima and provably decreases the largest eigenvalue (Bartlett et al., 2022). A contemporaneous approach studied the SAM algorithm in the limit of small learning rate and SAM radius, and quantified how the implicit and explicit regularization of SAM differs between full batch and batch size 1 dynamics (Wen et al., 2023).

1.2. Our contributions

In contrast to other theoretical approaches, we study the behavior of SAM far from minima. We find that SAM regularizes the eigenvalues throughout training through a dynamical phenomenon and analysis only near convergence cannot capture the full picture. In particular, in simplified models we show:

- Near initialization, full batch SAM provides limited suppression of large eigenvalues (Theorem 2.1).
- SAM induces a modified edge of stability (EOS) (Theorem 2.3).
- For full batch training, the largest eigenvalues stabilize at the SAM-EOS, at a smaller value than pure gradient descent (Section 3).
- As batch size decreases, the effect of SAM is stronger and the dynamics is no longer controlled by the Hessian alone (Theorem 2.2).

We then present experimental results on realistic models which show:

- The SAM-EOS predicts the largest eigenvalue for WideResnet 28-10 on CIFAR10.

Taken together, our results suggest that SAM can operate throughout the learning trajectory, far from minima, and that it can use non-linear, discrete dynamical effects to stabilize large curvatures of the loss function.

2. Quadratic regression model

2.1. Basic model

We consider a *quadratic regression model* (Agarwala et al., 2022) which extends a linear regression model to second order in the parameters. Given a P-dimensional parameter vector $\boldsymbol{\theta}$, the D-dimensional output is given by $\mathbf{f}(\boldsymbol{\theta})$:

$$\mathbf{f}(\boldsymbol{\theta}) = \mathbf{y} + \mathbf{G}^\top \boldsymbol{\theta} + \frac{1}{2}\mathbf{Q}(\boldsymbol{\theta}, \boldsymbol{\theta})\,. \tag{1}$$

Here, $\mathbf{y}$ is a D-dimensional vector, $\mathbf{G}$ is a $D \times P$-dimensional matrix, and $\mathbf{Q}$ is a $D \times P \times P$- dimensional tensor symmetric in the last two indices - that is, $\mathbf{Q}(\cdot,\cdot)$ takes two P-dimensional vectors as input, and outputs a D-dimensional vector $\mathbf{Q}(\boldsymbol{\theta},\boldsymbol{\theta})_\alpha = \boldsymbol{\theta}^\top \mathbf{Q}_\alpha \boldsymbol{\theta}$. If $\mathbf{Q} = \mathbf{0}$, the model corresponds to linear regression. $\mathbf{y}$, $\mathbf{G}$, and $\mathbf{Q}$ are all fixed at initialization.

Consider optimizing the model with under a squared loss. More concretely, let $\mathbf{y}_{tr}$ be a D-dimensional vector of training targets. We focus on the MSE loss

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}||\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}_{tr}||^2 \tag{2}$$

We can write the dynamics in terms of the residuals $\mathbf{z}$ and the Jacobian $\mathbf{J}$ defined by

$$\mathbf{z} \equiv f(\boldsymbol{\theta}) - \mathbf{y}_{tr},\ \mathbf{J} \equiv \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} = \mathbf{G} + \mathbf{Q}(\boldsymbol{\theta}, \cdot)\,. \tag{3}$$

Note that here $\mathbf{Q}(\boldsymbol{\theta},\cdot)$ is the $D \times P$-dimensional matrix with elements given by $[\mathbf{Q}(\boldsymbol{\theta},\cdot)]_{\alpha j} = \sum_i \mathbf{Q}_{\alpha i j}\boldsymbol{\theta}_i$. The loss can be written as $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{z}\cdot\mathbf{z}$. The full batch gradient descent (GD) dynamics of the parameters are given by

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \mathbf{J}_t^\top \mathbf{z}_t \tag{4}$$

which leads to

$$\begin{aligned} \mathbf{z}_{t+1} - \mathbf{z}_t &= -\eta \mathbf{J}_t \mathbf{J}_t^\top \mathbf{z}_t + \frac{1}{2}\eta^2 \mathbf{Q}(\mathbf{J}_t^\top \mathbf{z}_t, \mathbf{J}_t^\top \mathbf{z}_t) \\ \mathbf{J}_{t+1} - \mathbf{J}_t &= -\eta \mathbf{Q}(\mathbf{J}_t^\top \mathbf{z}_t, \cdot)\,. \end{aligned} \tag{5}$$

The $D \times D$-dimensional matrix $\mathbf{J}\mathbf{J}^\top$ is known as the *neural tangent kernel* (NTK) (Jacot et al., 2018), and controls the dynamics for small $\eta||\mathbf{J}^\top \mathbf{z}||$ (Lee et al., 2019).

We now consider the dynamics of un-normalized SAM (Andriushchenko & Flammarion, 2022). That is, given a loss function $\mathcal{L}$ we study the update rule

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = -\eta \nabla \mathcal{L}(\boldsymbol{\theta}_t + \rho \nabla \mathcal{L}(\boldsymbol{\theta}_t)) \tag{6}$$

We will focus our analysis on small learning rate and small SAM radius. The dynamics in $\mathbf{z} - \mathbf{J}$ space are given by

$$\begin{aligned} \mathbf{z}_{t+1} - \mathbf{z}_t = &-\eta \mathbf{J}_t \mathbf{J}_t^\top (1 + \rho \mathbf{J}_t \mathbf{J}_t^\top)\mathbf{z}_t - \eta\rho \mathbf{z}_t \cdot \mathbf{Q}(\mathbf{J}_t^\top \mathbf{z}_t, \mathbf{J}_t^\top \cdot) \\ &+ \eta^2 \frac{1}{2}\mathbf{Q}(\mathbf{J}_t^\top \mathbf{z}_t, \mathbf{J}_t^\top \mathbf{z}_t) + O(\eta\rho(\eta+\rho)||\mathbf{z}_t||^2) \end{aligned} \tag{7}$$

$$\mathbf{J}_{t+1} - \mathbf{J}_t = -\eta \left[\mathbf{Q}((1+\rho \mathbf{J}_t^\top \mathbf{J}_t)\mathbf{J}_t^\top \mathbf{z}_t, \cdot) + \rho \mathbf{Q}(\mathbf{z}_t \cdot \mathbf{Q}(\mathbf{J}_t^\top \mathbf{z}_t, \cdot), \cdot) \right] + O(\eta \rho^2 ||\mathbf{z}_t||^2) \tag{8}$$

to lowest order in η and ρ.

From Equation 7 we see that for small $\eta||\mathbf{z}||$ and $\rho||\mathbf{z}||$, the dynamics of $\mathbf{z}$ is controlled by the modified NTK $(1+\rho \mathbf{J}\mathbf{J}^\top)\mathbf{J}\mathbf{J}^\top$. The factor $1+\rho \mathbf{J}\mathbf{J}^\top$ shows up in the dynamics of $\mathbf{J}$ as well, and we will show that this effective NTK can lead to dynamical stabilization of large eigenvalues. And note that when $\rho = 0$, these dynamics coincide with that of gradient descent.

2.2. Eigenvalue dynamics at early times

A basic question is: how does SAM affect the eigenvalues of the NTK? We can study this directly for early learning dynamics by using randomly initialized quadratic regression models.

2.2.1. Gradient descent theory

For full batch learning, we have the following theorem (proof in Appendix A.2):

Theorem 2.1. *Consider a quadratic regression model, with* $\mathbf{Q}$ *initialized randomly with i.i.d. components with* 0 *mean and variance* 1*. For a model trained with full batch gradient descent, with unnormalized SAM, the change in* $\mathbf{J}$ *at the first step of the dynamics, averaged over* $\mathbf{Q}$ *is*

$$\mathrm{E}_{\mathbf{Q}}[\mathbf{J}_1 - \mathbf{J}_0] = -\rho\eta P \mathbf{z}_0 \mathbf{z}_0^\top \mathbf{J}_0 + O(\rho^2\eta^2||\mathbf{z}_0||^2) + O(\eta^3||\mathbf{z}_0||^3) \tag{9}$$

The α*th singular value* σ_α *of* $\mathbf{J}_0$ *associated with left and right singular vectors* $\mathbf{w}_\alpha$ *and* $\mathbf{v}_\alpha$ *can be approximated as*

$$\begin{aligned}(\sigma_\alpha)_1 - (\sigma_\alpha)_0 &= \mathbf{w}_\alpha^\top \mathrm{E}_{\mathbf{Q}}[\mathbf{J}_1 - \mathbf{J}_0]\mathbf{v}_\alpha + O(\eta^2) \\ &= -\rho\eta P(\mathbf{z}_0 \cdot \mathbf{w}_\alpha)^2 \sigma_\alpha + O(\eta^2)\end{aligned} \tag{10}$$

for small η*.*

Note that the singular vector $\mathbf{w}_\alpha$ is an eigenvector of $\mathbf{J}\mathbf{J}^\mathrm{T}$ associated with the eigenvalue σ_α^2.

This analysis suggests that on average, at early times, the change in the singular value is negative. However, the change also depends linearly on $(\mathbf{w}_\alpha \cdot \mathbf{z}_0)^2$. This suggests that if the component of $\mathbf{z}$ in the direction of the singular vector becomes small, the stabilizing effect of SAM becomes small as well. For large batch size/small learning rate with MSE loss, we in fact expect $\mathbf{z} \cdot \mathbf{w}_\alpha$ to decrease rapidly early in training (Cohen et al., 2022a; Agarwala et al., 2022). Therefore the relative regularizing effect can be *weaker* for larger modes in the GD setting.

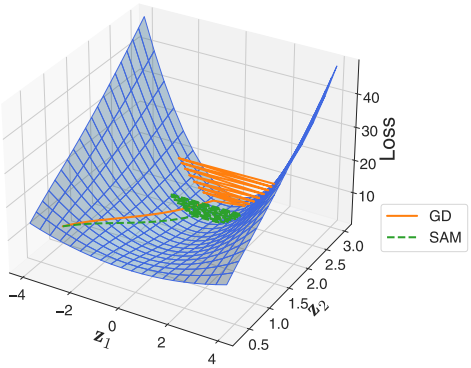

Figure 1. Schematic of SAM-modified EOS. Gradient descent decreases loss until a high-curvature area is reached, where large eigenmode is non-linearly stabilized (orange, solid). SAM causes stabilization to happen earlier, at a smaller value of the curvature (green, dashed).

2.2.2. SGD theory

It has been noted that the effects of SAM have a strong dependence on batch size (Andriushchenko & Flammarion, 2022). While a full analysis of SGD is beyond the scope of this work, we can see some evidence of stronger regularization for SGD in the quadratic regression model.

Consider SGD dynamics, where a random fraction $\beta = B/D$ of the training residuals $\mathbf{z}$ are used to generate the dynamics at each step. We can represent the sampling at each step with a random projection matrix $\mathbf{P}_t$, and replacing all instances of $\mathbf{z}_t$ with $\mathbf{P}_t\mathbf{z}_t$, where $\mathbf{P}_t$ is a random matrix with exactly B entries with value 1 on the diagonal, and 0 everywhere else. Under these dynamics, we can can prove the following:

Theorem 2.2. *Consider a second-order regression model, with* $\mathbf{Q}$ *initialized randomly with i.i.d. components with* 0 *mean and variance* 1*. For a model trained with SGD, sampling* B *datapoints independently at each step, the change in* $\mathbf{z}$ *and* $\mathbf{J}$ *at the first step, averaged over* $\mathbf{Q}$ *and the sampling matrix* $\mathbf{P}_t$*, is given by*

$$\begin{aligned}\mathrm{E}[\mathbf{z}_1 - \mathbf{z}_0]_{\mathbf{Q},\mathbf{P}} &= -\eta\beta \mathbf{J}_0\mathbf{J}_0^\top(1 + \rho[\beta(\mathbf{J}_0\mathbf{J}_0^\top) \\ &+ (1-\beta)\mathrm{diag}(\mathbf{J}_0\mathbf{J}_0^\top)])\mathbf{z}_0 + O(\eta^2||\mathbf{z}||^2) + O(D^{-1})\end{aligned} \tag{11}$$

$$\begin{aligned}\mathrm{E}_{\mathbf{Q},\mathbf{P}}[\mathbf{J}_1 - \mathbf{J}_0] &= -\rho\eta P(\beta^2 \mathbf{z}_0\mathbf{z}_0^\top + \beta(1-\beta)\mathrm{diag}(\mathbf{z}_0\mathbf{z}_0^\top))\mathbf{J}_0 \\ &+ O(\rho^2\eta^2||\mathbf{z}||^2) + O(\eta^3||\mathbf{z}||^3)\end{aligned} \tag{12}$$

where $\beta \equiv B/D$ *is the batch fraction.*

The calculations are detailed in Appendix A.2. This suggests that there are two possible sources of increased regulariza-

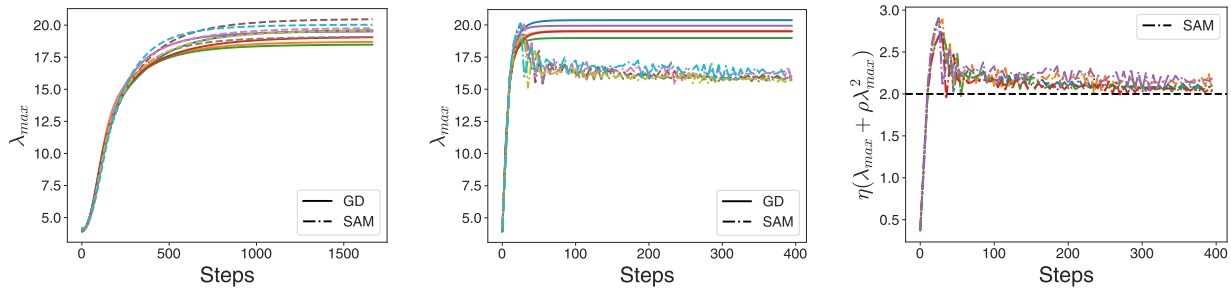

Figure 2. Trajectories of largest eigenvalue λ_{max} of $\mathbf{J}\mathbf{J}^\top$ for quadratic regression model, 5 independent initializations. For gradient descent with small learning rate ($\eta = 3 \cdot 10^{-3}$), SAM ($\rho = 4 \cdot 10^{-2}$) does not regularize the large NTK eigenvalues (left). For larger learning rate ($\eta = 8 \cdot 10^{-2}$), SAM controls large eigenvalues (middle). Largest eigenvalue can be predicted by SAM edge of stability $\eta(\lambda_{max} + \rho\lambda_{max}^2) = 2$ (right).

tion for SGD: the first being the additional terms proportional to $\beta(1-\beta)$. In addition to the fact that $\beta(1-\beta) > \beta^2$ for $\beta < \frac{1}{2}$, we have

$$\mathbf{v}_\alpha \mathrm{diag}(\mathbf{z}_0\mathbf{z}_0^\top)\mathbf{J}_0\mathbf{w}_\alpha = \sigma_\alpha(\mathbf{v}_\alpha \circ \mathbf{z}_0) \cdot (\mathbf{v}_\alpha \circ \mathbf{z}_0) \quad (13)$$

for left and right eigenvectors $\mathbf{v}_\alpha$ and $\mathbf{w}_\alpha$ of $\mathbf{J}_0$, where $\circ$ is the Hadamard (elementwise) product. This term can be large even if $\mathbf{v}_\alpha$ and $\mathbf{z}_t$ have small dot product. This is in contrast to $\beta^2(\mathbf{v}_\alpha \cdot \mathbf{z}_0)^2$, which is small if $\mathbf{z}_0$ does not have a large component in the $\mathbf{v}_\alpha$ direction. This suggests that at short times, where the large eigenmodes decay quickly, the SGD term can still be large. Additionally, the projection onto the largest eigenmode itself decreases more slowly in the SGD setting (Paquette et al., 2021), which also suggests stronger early time regularization for small batch size.

2.3. Edge of stability and SAM

One of the most dramatic consequences of SAM is the shift of the *edge of stability* (EOS). We begin by reviewing the EOS phenomenology.

The definition of the "edge of stability" comes from considering gradient descent dynamics near a minimum of a twice-differentiable loss. The dynamics of the displacement $\mathbf{x}$ from the minimum (in parameter space) are well-approximated by

$$\mathbf{x}_{t+1} - \mathbf{x}_t = -\eta\mathbf{H}\mathbf{x}_t \quad (14)$$

where $\mathbf{H}$ is the positive semi-definite Hessian at the minimum $\mathbf{x} = 0$. The dynamics converges exponentially iff the largest eigenvalue of $\mathbf{H}$ is bounded by $\eta\lambda_{max} < 2$. We refer to $\eta\lambda_{max}$ as the *normalized eigenvalue*, Otherwise, there is at least one component of $\mathbf{x}$ which is non-decreasing. The value $2/\eta$ is often referred to as the *edge of stability* (EOS) for the dynamics.

Previous work has shown that for many non-linear models, there is a range of learning rates where the largest eigenvalue of the Hessian stabilizes around the edge of stability (Cohen et al., 2022a) - even when models are far from minima. This effect is most robust in full batch training but can show up in SGD training as well. Equivalent phenomenology exists for other gradient-based methods (Cohen et al., 2022b). The stabilization effect is due to feedback between the largest curvature eigenvalue and the displacement in the largest eigendirection (Agarwala et al., 2022; Damian et al., 2022). For MSE loss, EOS behavior occurs for the large NTK eigenvalues as well (Agarwala et al., 2022), as suggested by a low order expansion of $\mathbf{z}_{t+1} - \mathbf{z}_t$ (Equation 5).

We will show that SAM also induces an EOS stabilization effect, but at a smaller eigenvalue than GD. We can understand the shift intuitively by analyzing un-normalized SAM on a loss $\frac{1}{2}\mathbf{x}^\top\mathbf{H}\mathbf{x}$. Direct calculation gives the update rule:

$$\mathbf{x}_{t+1} - \mathbf{x}_t = -\eta(\mathbf{H} + \rho\mathbf{H}^2)\mathbf{x}_t \quad (15)$$

For positive definite $\mathbf{H}$, $\mathbf{x}_t$ converges exponentially to 0 iff $\eta(\lambda_{max} + \rho\lambda_{max}^2) < 2$. Recall from Section 2.1 that the SAM NTK is $(1 + \rho\mathbf{J}\mathbf{J}^\top)\mathbf{J}\mathbf{J}^\top > \mathbf{J}\mathbf{J}^\top$. This suggests that $\eta(\lambda_{max} + \rho\lambda_{max}^2)$ is the *SAM normalized eigenvalue*. This bound gives a critical λ_{max} which is smaller than that in the GD case. This leads to the hypothesis that SAM can cause a stabilization at the EOS in a flatter region of the loss, as schematically illustrated in Figure 1.

We can formalize the *SAM edge of stability* (SAM-EOS) for any differentiable model trained on MSE loss. Equation 7 suggests the matrix $\mathbf{J}\mathbf{J}^\top(1 + \rho\mathbf{J}\mathbf{J}^\top)$ - which has larger eigenvalues for larger ρ - controls the low-order dynamics. We can formalize this intuition in the following theorem (proof in Appendix B.1):

Theorem 2.3. *Consider a $\mathcal{C}^\infty$ model $\mathbf{f}(\boldsymbol{\theta})$ trained using Equation 6 with MSE loss. Suppose that there exists a point $\boldsymbol{\theta}^*$ where $\mathbf{z}(\boldsymbol{\theta}^*) = 0$. Suppose that for some $\epsilon > 0$, we have the lower bound $\epsilon < \eta\lambda_i(1 + \rho\lambda_i)$ for the eigenvalues of the positive definite symmetric matrix $\mathbf{J}(\boldsymbol{\theta}^*)\mathbf{J}(\boldsymbol{\theta}^*)^\top$. Given a bound on the largest eigenvalue, there are two regimes:*

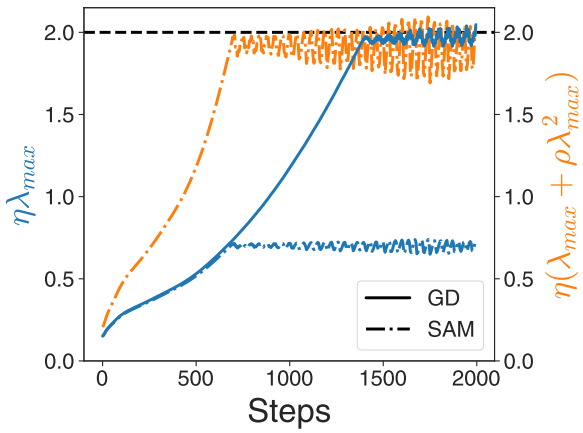

Figure 3. Largest eigenvalues of $\mathbf{J}\mathbf{J}^\top$ for a fully-connected network trained using MSE loss on 2-class CIFAR. For gradient descent ($\eta = 4 \cdot 10^{-3}$) largest eigenvalue stabilizes according to the GD-EOS $\eta\lambda_{max} = 2$ (solid line, blue). SAM ($\rho = 10^{-2}$) stabilizes to a lower value (dashed line, blue), which is well-predicted by the SAM-EOS $\eta(\lambda_{max} + \rho\lambda_{max}^2) = 2$ (dashed line, orange).

Convergent regime. *If $\eta\lambda_i(1 + \rho\lambda_i) < 2 - \epsilon$ for all for all eigenvalues λ_i of $\mathbf{J}(\boldsymbol{\theta}^*)\mathbf{J}(\boldsymbol{\theta}^*)^\top$, there exists a neighborhood U of $\boldsymbol{\theta}^*$ such that $\lim_{t\to\infty} \mathbf{z}_t = 0$ with exponential convergence for any trajectory initialized at $\boldsymbol{\theta}_0 \in U$.*

Divergent regime. *If $\eta\lambda_i(1 + \rho\lambda_i) > 2 + \epsilon$ for some eigenvector $\mathbf{v}_i$ of $\mathbf{J}(\boldsymbol{\theta}^*)\mathbf{J}(\boldsymbol{\theta}^*)^\top$, then there exists some q_{min} such that for any $q < q_{min}$, given $B_q(\boldsymbol{\theta}^*)$, the ball of radius q around $\boldsymbol{\theta}^*$, there exists some initialization $\boldsymbol{\theta}_0 \in B_q(\boldsymbol{\theta}^*)$ such that the trajectory $\{\boldsymbol{\theta}_t\}$ leaves $B_q(\boldsymbol{\theta}^*)$ at some time t.*

Note that the theorem is proven for the NTK eigenvalues, which also show EOS behavior for MSE loss in the GD setting (Agarwala et al., 2022).

This theorem gives us the modified edge of stability (EOS) condition:

$$\eta\lambda_{max}(1 + \rho\lambda_{max}) \approx 2 \tag{16}$$

This SAM-EOS condition is derived near a minimum like the original EOS condition; however, as we will see, the largest eigenvalue stabilizes according to SAM-EOS even *away* from a minimum, in a similar manner to the original EOS phenomenology.

For larger ρ, a smaller λ_{max} is needed to meet the edge of stability condition. In terms of the normalized eigenvalue $\tilde{\lambda} = \eta\lambda$, the modified EOS can be written as $\tilde{\lambda}(1 + r\tilde{\lambda}) = 2$ with the ratio $r = \rho/\eta$. Larger values of r lead to stronger regularization effects, and for the quadratic regression model specifically η can be factored out leaving r as the key dimensionless parameter (Appendix A.1).

3. Experiments on basic models

3.1. Quadratic regression model

We can explore the effects of SAM and show the SAM-EOS behavior via numerical experiments on the quadratic regression model. We use the update rule in Equation 6, working directly in $\mathbf{z}$ and $\mathbf{J}$ space as in (Agarwala et al., 2022). Experimental details can be found in Appendix A.3.

For small learning rates, we see that SAM does not reduce the large eigenvalues of $\mathbf{J}\mathbf{J}^\top$ in the dynamics (Figure 2, left). In fact in some cases the final eigenvalue is *larger* with SAM turned on. The projection onto the largest eigenmodes of $\mathbf{J}\mathbf{J}^\top$ exponentially decreases to 0 quicker than any other mode; as suggested by Theorem 2.1, this leads to only a small decreasing pressure from SAM. The primary dynamics of the large eigenvalues is due to the progressive sharpening phenomenology studied in (Agarwala et al., 2022), which tends to increase the eigenmodes.

However, for larger learning rates, SAM has a strong suppressing effect on the largest eigenvalues (Figure 2, middle). The overall dynamics are more non-linear than in the small learning rate case. The eigenvalues stabilize at the modified EOS boundary $\eta(\lambda_{max} + \rho\lambda_{max}^2) = 2$ (Figure 2, right), suggesting non-linear stabilization of the eigenvalues. In Appendix A.3 we conduct additional experiments which confirm that the boundary predicts the largest eigenvalue for a range of ρ, and that consequently generally increasing ρ leads to decreased λ_{max}.

3.2. CIFAR-2 with MSE loss

We can see this phenomenology in more general non-linear models as well. We trained a fully-connected network on the first 2 classes of CIFAR with MSE loss, with both full batch gradient descent and SAM. We then computed the largest eigenvalues of $\mathbf{J}\mathbf{J}^\top$ along the trajectory. We can see that in both GD and SAM the largest eigenvalues stabilize, and the stabilization threshold is smaller for SAM (Figure 3). The threshold is once again well predicted by the SAM-EOS.

4. Connection to realistic models

In this section, we show that our analysis of quadratic models can bring insights into the behavior of more realistic models.

4.1. Setup

Sharpness For MSE loss, edge of stability dynamics can be shown in terms of either the NTK eigenvalues *or* the Hessian eigenvalues (Agarwala et al., 2022). For more general loss functions, EOS dynamics takes place with respect to the largest Hessian eigenvalues (Cohen et al., 2022a; Damian

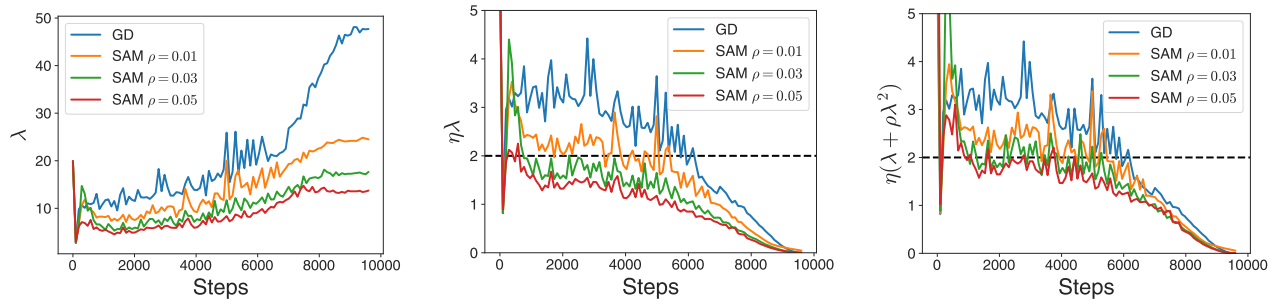

Figure 4. Largest Hessian eigenvalues for CIFAR10 trained with MSE loss (averaged over 10 seeds). Left: largest eigenvalues increase at late times. Larger SAM radius mitigates eigenvalue increase. Middle: eigenvalues normalized by learning rate decrease at late times, and SGD shows edge of stability (EOS) behavior. Right: For larger ρ, SAM-normalized eigenvalues show modified EOS behavior.

et al., 2022). Following these results and the analysis in Equation 15, we chose to measure the largest eigenvalue of the Hessian rather than the NTK. We used a Lanczos method (Ghorbani et al., 2019) to approximately compute λ_{max}. Any reference to λ_{max} in this section refers to eigenvalues computed in this way.

CIFAR-10 We conducted experiments on the popular CIFAR-10 dataset (Krizhevsky et al., 2009) using the WideResnet 28-10 architecture (Zagoruyko & Komodakis, 2016). We report results for both the MSE loss and the cross-entropy loss. In the case of the MSE loss, we replace the softmax non-linearity with Tanh and rescale the one-hot labels $\mathbf{y} \in \{0, 1\}^{10}$ to $\{-1, 1\}^{10}$. In both cases, the loss is averaged across the number of elements in the batch and the number of classes. For each setting, we report results for a single configuration of the learning rate η and weight decay μ found from an initial cross-validation sweep. For MSE, we use $\eta = 0.3, \mu = 0.005$ and $\eta = 0.4, \mu = 0.005$ for cross-entropy. We use the cosine learning rate schedule (Loshchilov & Hutter, 2016) and SGD instead of Nesterov momentum (Sutskever et al., 2013) to better match the theoretical setup. Despite the changes to the optimizer and the loss, the test error for the models remains in a reasonable range (4.4% for SAM regularized models with MSE and 5.3% with SGD). In accordance with the theory, we use unnormalized SAM in these experiments. We keep all other hyper-parameters to the default values described in the original WideResnet paper.

4.2. Results

As shown in Figure 4 (left), the maximum eigenvalue increases significantly throughout training for all approaches considered. However, the normalized curvature $\eta\lambda_{max}$, which sets the edge of stability in GD, remains relatively stable early on in training when the learning rate is high, but necessarily decreases as the cosine schedule drives the learning rate to 0 (Figure 4, middle). Note that the results in Figure 4 are averaged over 10 seeds to reduce the noise from SGD and to improve visual clarity.

SAM radius drives curvature below GD-EOS. As we increase the SAM radius, the largest eigenvalue is more controlled (Figure 4, left) - falling below the gradient descent edge of stability (Figure 4, middle). The stabilizing effect of SAM on the large eigenvalues is evident even early on in training.

Eigenvalues stabilize around SAM-EOS. If we instead plot the SAM-normalized eigenvalue $\eta(\lambda_{max} + \rho\lambda_{max}^2)$, we see that the eigenvalues stay close to (and often slightly above) the critical value of 2, as predicted by theory (Figure 4, right). This suggests that there are settings where the control that SAM has on the large eigenvalues of the Hessian comes, in part, from a modified EOS stabilization effect. This is true over a range of learning rates; see Appendix C.1 for more details. The results are similar for alternative characterizations of the EOS (Appendix C.3).

Altering SAM radius during training can successfully move us between GD-EOS and SAM-EOS. Further evidence from EOS stabilization comes from using a *SAM schedule*. We trained the model with two settings: early SAM, where SAM is used for the first 2500 steps (50 epochs), after which the training proceeds with SGD ($\rho = 0$), and late SAM, where SAM is used for the first 2500 steps, after which only SGD is used. The maximum eigenvalue is lower for early SAM before 2500 steps, after which there is a quick crossover and late SAM gives better control (Figure 5). Both SAM schedules give improvement over SGD-only training. Generally, turning SAM on later or for the full trajectory gave better generalization than turning SAM on early, consistent with the earlier work of Andriushchenko & Flammarion (2022).

Plotting the eigenvalues for the early SAM and late SAM schedules, we see that when SAM is turned off, the normalized eigenvalues lie above the gradient descent EOS (Figure 5, right, blue curves). However when SAM is turned

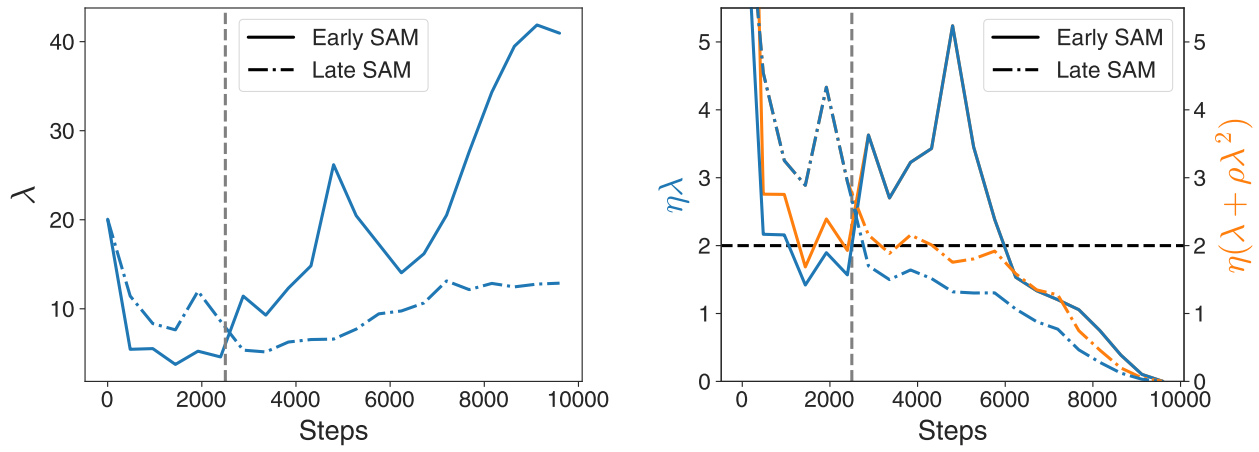

Figure 5. Maximum eigenvalues for CIFAR-10 model trained on MSE loss with a SAM schedule. Starting out with SAM ($\rho = 0.05$, solid lines) and turning it off at 2500 steps leads to initial suppression and eventual increase of λ_{max}; starting out with SGD and turning SAM on after 2500 steps leads to the opposite behavior (left). Eigenvalues cross over quickly after the switch. Plotting GD normalized eigenvalues (blue, right) shows GD-EOS behavior in SGD phase; plotting SAM normalized eigenvalues (orange, right) shows SAM EOS behavior in SAM phase.

on, $\eta\lambda_{max}$ is usually below the edge of stability value of 2; instead, the SAM-normalized value $\eta(\lambda_{max} + \rho\lambda_{max}^2)$ lies close to the critical value of 2 (Figure 5, right, orange curves). This suggests that turning SAM on or off during the intermediate part of training causes the dynamics to quickly reach the appropriate edge of stability.

Networks with cross-entropy loss behave similarly. We found similar results for cross-entropy loss as well, which we detail in Appendix C.2. The mini-batch gradient magnitude and eigenvalues vary more over the learning trajectories; this may be related to effects of logit magnitudes which have been previously shown to affect curvature and general training dynamics (Agarwala et al., 2020; Cohen et al., 2022a).

Minibatch gradient norm varies little. Another quantity of interest is the magnitude of the mini-batch gradients. For SGD, the gradient magnitudes were steady during the first half of training and dropped by a factor of 4 at late times (Figure 6). Gradient magnitudes were very stable for SAM, particularly for larger ρ. This suggests that in practice, there may not be much difference between the normalized and un-normalized SAM algorithms. This is consistent with previous work which showed that the generalization of the two approaches is similar (Andriushchenko & Flammarion, 2022).

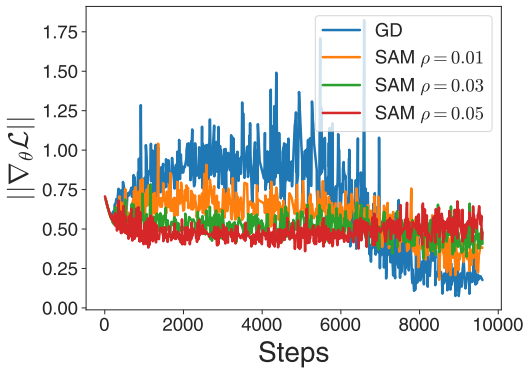

Figure 6. Minibatch gradient magnitudes for CIFAR-10 model trained on MSE loss. Magnitudes are steady early on in SGD training, but decrease at the end of training. Eigenvalue variation is smaller for increasing sam radius ρ.

5. Discussion

5.1. SAM as a dynamical phenomenon

Much like the study of EOS before it, our analysis of SAM suggests that sharpness dynamics near minima are insufficient to capture relevant phenomenology. Our analysis of the quadratic regression model suggests that SAM already regularizes the large eigenmodes at early times, and the EOS analysis shows how SAM can have strong effects even in the large-batch setting. Our theory also suggested that SGD has additional mechanisms to control curvature early on in training as compared to full batch gradient descent.

The SAM schedule experiments provided further evidence that multiple phases of the optimization trajectory are important for understanding the relationship between SAM and generalization. If the important effect was the convergence to a particular minimum, then only late SAM would improve generalization. If instead some form of "basin selection" was key, then only early SAM would improve generaliza-

tion. The fact that both are important (Andriushchenko & Flammarion, 2022) suggests that the entire optimization trajectory matters.

We note that while EOS effects are *necessary* to understand some aspects of SAM, they are certainly not *sufficient*. As shown in Appendix A.3, the details of the behavior near the EOS have a complex dependence on ρ (and the model). Later on in learning, especially with a loss like cross entropy, the largest eigenvalues may decrease even without SAM (Cohen et al., 2022a) - potentially leading the dynamics away from the EOS. Small batch size may add other effects, and EOS effects become harder to understand if multiple eigenvalues are at the EOS. Nonetheless, even in more complicated cases the SAM-EOS gives a good approximation to the control SAM has on the eigenvalues, particularly at early times.

5.2. Optimization and regularization are deeply linked

This work provides additional evidence that understanding some regularization methods may in fact require analysis of the optimization dynamics - especially those at early or intermediate times. This is in contrast to approaches which seek to understand learning by characterizing minima, or analyzing behavior near convergence only. A similar phenomenology has been observed in evolutionary dynamics - the basic 0th order optimization method - where the details of optimization trajectories are often more important than the statistics of the minima to understand long-term dynamics (Nowak & Krug, 2015; Park & Krug, 2016; Agarwala & Fisher, 2019).

6. Future work

Our main theoretical analysis focused on the dynamics $\mathbf{z}$ and $\mathbf{J}$ under squared loss; additional complications arise for non-squared losses like cross-entropy. Providing a detailed quantitative characterization of the EOS dynamics under these more general conditions is an important next step.

Another important open question is the analysis of SAM (and the EOS effect more generally) under SGD. There are alternative formulations of the EOS for SGD (such as Lee & Jang (2023)) which give consistent results with our experiments on λ_{max} (Appendix C.3). While Theorem 2.2 provides some insight into the differences, a full understanding would require an analysis of $\mathrm{E}_{\mathbf{P}}[(\mathbf{z} \cdot \mathbf{v}_i)^2]$ for the different eigenmodes $\mathbf{v}_i$ - which has only recently been analyzed for a quadratic loss function (Paquette et al., 2021; 2022a;b; Lee et al., 2022). Our analysis of the CIFAR10 models showed that the SGD gradient magnitude does not change much over training. Further characterization of the SGD gradient statistics will also be useful in understanding the interaction of SAM and SGD.

More detailed theoretical and experimental analysis of more complex settings may allow for improvements to the SAM algorithm and its implementation in practice. A more detailed theoretical understanding could lead to proposals for ρ-schedules, or improvements to the core algorithm itself - already a field of active research (Zhuang et al., 2022).

Finally, our work focuses on optimization and training dynamics; linking these properties to generalization remains a key goal of any further research into SAM and other optimization methods.

References

Agarwala, A. and Fisher, D. S. Adaptive walks on high-dimensional fitness landscapes and seascapes with distance-dependent statistics. *Theoretical Population Biology*, 130:13–49, December 2019. ISSN 0040-5809. doi: 10.1016/j.tpb.2019.09.011.

Agarwala, A., Pennington, J., Dauphin, Y., and Schoenholz, S. Temperature check: Theory and practice for training models with softmax-cross-entropy losses, October 2020.

Agarwala, A., Pedregosa, F., and Pennington, J. Second-order regression models exhibit progressive sharpening to the edge of stability, October 2022.

Andriushchenko, M. and Flammarion, N. Towards Understanding Sharpness-Aware Minimization, June 2022.

Bahri, D., Mobahi, H., and Tay, Y. Sharpness-Aware Minimization Improves Language Model Generalization, March 2022.

Bartlett, P. L., Long, P. M., and Bousquet, O. The Dynamics of Sharpness-Aware Minimization: Bouncing Across Ravines and Drifting Towards Wide Minima, October 2022.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, December 2019. ISSN 1742-5468. doi: 10.1088/1742-5468/ab39d9.

Cohen, J., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In *International Conference on Learning Representations*, February 2022a.

Cohen, J. M., Ghorbani, B., Krishnan, S., Agarwal, N., Medapati, S., Badura, M., Suo, D., Cardoze, D., Nado, Z., Dahl, G. E., and Gilmer, J. Adaptive Gradient Methods at the Edge of Stability, July 2022b.

Damian, A., Nichani, E., and Lee, J. D. Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability, September 2022.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1019–1028, Sydney, NSW, Australia, August 2017. JMLR.org.

Duchi, J., Hazan, E., and Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(61): 2121–2159, 2011. ISSN 1533-7928.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware Minimization for Efficiently Improving Generalization. In *International Conference on Learning Representations*, April 2022.

Ghorbani, B., Krishnan, S., and Xiao, Y. An Investigation into Neural Net Optimization via Hessian Eigenvalue Density. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2232–2241. PMLR, May 2019.

Hochreiter, S. and Schmidhuber, J. Flat Minima. *Neural Computation*, 9(1):1–42, January 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1.

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pp. 8571–8580. Curran Associates, Inc., 2018.

Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho*, K., and Geras*, K. 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.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima, February 2017.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems 32*, pp. 8570–8581. Curran Associates, Inc., 2019.

Lee, K., Cheng, A. N., Paquette, C., and Paquette, E. Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions, June 2022.

Lee, S. and Jang, C. A new characterization of the edge of stability based on a sharpness measure aware of batch gradient distribution. In *The Eleventh International Conference on Learning Representations*, February 2023.

Lewis, A. S. and Overton, M. L. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, 141(1):135–163, October 2013. ISSN 1436-4646. doi: 10.1007/s10107-012-0514-2.

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: The catapult mechanism. March 2020.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems 30*, pp. 5947–5956. Curran Associates, Inc., 2017.

Nocedal, J. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151): 773–782, 1980. ISSN 0025-5718, 1088-6842. doi: 10.1090/S0025-5718-1980-0572855-7.

Nowak, S. and Krug, J. Analysis of adaptive walks on NK fitness landscapes with different interaction schemes. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(6):P06014, 2015.

Paquette, C., Lee, K., Pedregosa, F., and Paquette, E. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 3548–3626. PMLR, July 2021.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties, May 2022a.

Paquette, C., Paquette, E., Adlam, B., and Pennington, J. Implicit Regularization or Implicit Conditioning? Exact Risk Trajectories of SGD in High Dimensions, June 2022b.

Park, S.-C. and Krug, J. δ-exceedance records and random adaptive walks. *Journal of Physics A: Mathematical and Theoretical*, 49(31):315601, 2016.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Wen, K., Ma, T., and Li, Z. How Does Sharpness-Aware Minimization Minimize Sharpness?, January 2023.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhuang, J., Gong, B., Yuan, L., Cui, Y., Adam, H., Dvornek, N., Tatikonda, S., Duncan, J., and Liu, T. Surrogate Gap Minimization Improves Sharpness-Aware Training, March 2022.

A. Quadratic regression model

A.1. Rescaled dynamics

The learning rate can be rescaled out of the quadratic regression model. In previous work, the the rescaling

$$\tilde{\mathbf{z}} = \eta \mathbf{z},\ \tilde{\mathbf{J}} = \eta^{1/2}\mathbf{J} \tag{17}$$

which gave a universal representation of the dynamics. The same rescaling in the SAM case gives us:

$$\begin{aligned}\tilde{\mathbf{z}}_{t+1} - \tilde{\mathbf{z}}_t = &-(\tilde{\mathbf{J}}_t\tilde{\mathbf{J}}_t^\top + r(\tilde{\mathbf{J}}_t\tilde{\mathbf{J}}_t^\top)^2)\tilde{\mathbf{z}}_t - r[(1 + r\tilde{\mathbf{J}}_t\tilde{\mathbf{J}}_t^\top)\tilde{\mathbf{z}}_t]^\top \mathbf{Q}(\tilde{\mathbf{J}}_t^\top\tilde{\mathbf{z}}_t, \tilde{\mathbf{J}}_t^\top\cdot) \\ &+ \frac{1}{2}\mathbf{Q}[\tilde{\mathbf{J}}_t^\top(1 + r\tilde{\mathbf{J}}_t\tilde{\mathbf{J}}_t^\top)\tilde{\mathbf{z}}_t, \tilde{\mathbf{J}}_t^\top(1 + r\tilde{\mathbf{J}}_t\tilde{\mathbf{J}}_t^\top)\tilde{\mathbf{z}}_t] + O(||\tilde{\mathbf{z}}^3||)\end{aligned} \tag{18}$$

$$\begin{aligned}\tilde{\mathbf{J}}_{t+1} - \tilde{\mathbf{J}}_t = &-\mathbf{Q}(\tilde{\mathbf{J}}_t^\top(1 + r\tilde{\mathbf{J}}_t\tilde{\mathbf{J}}_t^\top)\tilde{\mathbf{z}}_t, \cdot) - r\mathbf{Q}([(1 + r\tilde{\mathbf{J}}_t\tilde{\mathbf{J}}_t^\top)\tilde{\mathbf{z}}_t]^\top \mathbf{Q}(\tilde{\mathbf{J}}_t^\top\tilde{\mathbf{z}}_t, \cdot), \cdot) \\ &- \frac{1}{2}r^2\mathbf{Q}\left[\tilde{\mathbf{J}}_t^\top\mathbf{Q}(\tilde{\mathbf{J}}_t^\top\tilde{\mathbf{z}}_t, \tilde{\mathbf{J}}_t^\top\tilde{\mathbf{z}}_t), \cdot\right] + O(||\tilde{\mathbf{z}}^3||)\end{aligned} \tag{19}$$

where r is the rescaled SAM radius $r = \rho/\eta$.

This suggests that, at least for gradient descent, the *ratio* of the SAM radius to the learning rate determines the amount of regularization which SAM provides.

A.2. Average values, one step SGD

We will prove Theorem 2.2 first, and then derive Theorem 2.1 is as a special case.

Theorem 2.2. Consider a second-order regression model, with $\mathbf{Q}$ initialized randomly with i.i.d. components with 0 mean and variance 1. For a model trained with SGD, sampling B datapoints independently at each step, the change in $\mathbf{z}$ and $\mathbf{J}$ at the first step, averaged over $\mathbf{Q}$ and the sampling matrix $\mathbf{P}_t$, is given by

$$\mathrm{E}[\mathbf{z}_1 - \mathbf{z}_0]_{\mathbf{Q},\mathbf{P}} = -\eta\beta\mathbf{J}_0\mathbf{J}_0^\top(1 + \rho[\beta(\mathbf{J}_0\mathbf{J}_0^\top) + (1-\beta)\mathrm{diag}(\mathbf{J}_0\mathbf{J}_0^\top)])\mathbf{z}_0 + O(\eta^2||\mathbf{z}||^2) + O(D^{-1}) \tag{20}$$

$$\mathrm{E}_{\mathbf{Q},\mathbf{P}}[\mathbf{J}_1 - \mathbf{J}_0] = -\rho\eta P(\beta^2\mathbf{z}_0\mathbf{z}_0^\top + \beta(1-\beta)\mathrm{diag}(\mathbf{z}_0\mathbf{z}_0^\top))\mathbf{J}_0 + O(\rho^2\eta^2||\mathbf{z}||^2) + O(\eta^3||\mathbf{z}||^3) \tag{21}$$

where $\beta \equiv B/D$ is the batch fraction.

Proof. We can write the SGD dynamics of the quadratic regression model as:

$$\mathbf{z}_{t+1} - \mathbf{z}_t = -\eta\mathbf{J}_t\mathbf{J}_t^\top\mathbf{P}_t\mathbf{z}_t + \frac{1}{2}\eta^2\mathbf{Q}(\mathbf{J}_t^\top\mathbf{P}\mathbf{z}_t, \mathbf{J}_t^\top\mathbf{P}\mathbf{z}_t) \tag{22}$$

$$\mathbf{J}_{t+1} - \mathbf{J}_t = -\eta\mathbf{Q}(\mathbf{J}_t^\top\mathbf{P}_t\mathbf{z}_t, \cdot)\,. \tag{23}$$

where $\mathbf{P}_t$ is a projection matrix which chooses the batch. It is a $D \times D$ diagonal matrix with exactly B random 1s on the diagonal, with all other entries 0. This corresponds to choosing B random elements, without replacement, at each timestep.

For SAM with SGD, the SAM step is replaced with $\rho\mathbf{J}_t\mathbf{P}_t\mathbf{z}_t$ as well. Expanding to lowest order, we have:

$$\mathbf{z}_{t+1} - \mathbf{z}_t = -\eta(\mathbf{J}_t\mathbf{J}_t^\top + \rho(\mathbf{J}_t\mathbf{J}_t^\top)\mathbf{P}_t(\mathbf{J}_t\mathbf{J}_t^\top))\mathbf{P}_t\mathbf{z}_t + O(||\mathbf{z}||^2) \tag{24}$$

$$\begin{aligned}\mathbf{J}_{t+1} - \mathbf{J}_t = &-\eta\mathbf{Q}(\mathbf{J}_t^\top(1 + \rho\mathbf{P}_t\mathbf{J}_t\mathbf{J}_t^\top)\mathbf{P}_t\mathbf{z}_t, \cdot) - \rho\eta\mathbf{Q}([\mathbf{P}_t\mathbf{z}_t]^\top\mathbf{Q}(\mathbf{J}_t^\top\mathbf{P}_t\mathbf{z}_t, \cdot), \cdot) \\ &+ O(\rho^2\eta^2||\mathbf{z}||^2) + O(\eta^3||\mathbf{z}||^3)\end{aligned} \tag{25}$$

Consider the dynamics of $\mathbf{z}$. Taking the average over $\mathbf{P}_t$, we note that $\mathrm{E}[\mathbf{P}] = \beta\mathbf{I}$. For any fixed $D \times D$ matrix $\mathbf{M}$, we also have:

$$\mathrm{E}[\mathbf{P}_t\mathbf{M}\mathbf{P}_t] = \beta^2\mathbf{M} + \beta(1-\beta)\mathrm{diag}(\mathbf{M}) + O(D^{-1}) \tag{26}$$

Substituting, we have:

$$\mathrm{E}_{\mathbf{P}_t}[\mathbf{z}_{t+1} - \mathbf{z}_t] = -\eta\beta\mathbf{J}_t\mathbf{J}_t^\top(1 + \rho[\beta(\mathbf{J}_t\mathbf{J}_t^\top) + (1-\beta)\mathrm{diag}(\mathbf{J}_t\mathbf{J}_t^\top)])\mathbf{z}_t + O(||\mathbf{z}||^2) + O(D^{-1}) \tag{27}$$

Now consider the dynamics of $\mathbf{J}$. First we averaging over random initial $\mathbf{Q}$. At the first step we have:

$$\mathrm{E}_{\mathbf{Q}}[\mathbf{J}_1 - \mathbf{J}_0]_{\alpha i} = -\rho\eta\mathrm{E}[\mathbf{Q}_{\alpha ij}(\mathbf{P}\mathbf{z})_\beta\mathbf{Q}_{\beta jk}\mathbf{J}_{\gamma k}(\mathbf{P}\mathbf{z})_\gamma] + O(\rho^2\eta^2||\mathbf{z}||^2) + O(\eta^3||\mathbf{z}||^3) \tag{28}$$

$$\mathrm{E}_{\mathbf{Q}}[\mathbf{J}_1 - \mathbf{J}_0]_{\alpha i} = -\rho\eta P(\mathbf{P}\mathbf{z})_\alpha\mathbf{J}_{\gamma i}(\mathbf{P}\mathbf{z})_\gamma + O(\rho^2\eta^2||\mathbf{z}||^2) + O(\eta^3||\mathbf{z}||^3) \tag{29}$$

Averaging over $\mathbf{P}$ as well, we have:

$$\mathrm{E}_{\mathbf{Q},\mathbf{P}}[\mathbf{J}_1 - \mathbf{J}_0] = -\rho\eta P(\beta^2\mathbf{z}\mathbf{z}^\top + \beta(1-\beta)\mathrm{diag}(\mathbf{z}\mathbf{z}^\top))\mathbf{J} + O(\rho^2\eta^2||\mathbf{z}||^2) + O(\eta^3||\mathbf{z}||^3) + O(D^{-1}) \tag{30}$$

□

Theorem 2.1 can be derived by first setting $\beta = 1$. Given a singular value σ_α corresponding to singular vectors $\mathbf{w}_{al}$ and $\mathbf{v}_\alpha$ we have $\sigma_\alpha = \mathbf{w}_\alpha^\top\mathbf{J}\mathbf{v}_\alpha$. For small learning rates, the singular value of $\mathbf{J}_1$ can be written in terms of the SVD of $\mathbf{J}_0$ as

$$\sigma_\alpha(\mathbf{J}_1) = \mathbf{w}_\alpha(\mathbf{J}_0)^\top\mathbf{J}_1\mathbf{v}_\alpha(\mathbf{J}_0) + O(\eta^2) \tag{31}$$

Therefore we can write

$$\sigma_\alpha(\mathbf{J}_1) - \sigma_\alpha(\mathbf{J}_0) = \mathbf{w}_\alpha(\mathbf{J}_0)^\top(\mathbf{J}_1 - \mathbf{J}_0)\mathbf{v}_\alpha(\mathbf{J}_0) + O(\eta^2) \tag{32}$$

Averaging over $\mathbf{Q}$ and $\mathbf{P}$ completes the theorem.

A.3. Numerical results

The numerical results in Figure 2 were obtained by training the model defined by the update Equation 6 in $\mathbf{z}$ and $\mathbf{J}$ space directly. The tensor $\mathbf{Q}$ was randomly initialized with i.i.d. Gaussian elements at initialization, and $\mathbf{z}$ and $\mathbf{J}$ were randomly initialized as well following the approach in (Agarwala et al., 2022). The figures correspond to 5 independent initializations with the same statistics for $\mathbf{Q}$, $\mathbf{z}$, and $\mathbf{J}$. All plots used $D = 200$ datapoints with $P = 400$ parameters.

For small η, the loss converges exponentially to 0. In particular, the projection onto the largest eigenmode decreases quickly , which by the analysis of Theorem 2.1 suggests that SAM has only a small effect on the largest eigenvalues.

For larger η, by increasing ρ the SAM dynamics seems to access the edge of stability regime, where non-linear effects can stabilize the large eigenvalues of the curvature. One way the original edge of stability dynamics was explored was to examine trajectories at different learning rates (Cohen et al., 2022a). At small learning rate, training loss decreases monotonically; at intermediate learning rates, the edge of stability behavior causes non-monotonic but still stable learning trajectories, and finally, at large learning rate the training is unstbale.

We can similarly increase the SAM radius ρ for fixed learning rate, and see an analogous transition. If we pick η such that the trajectory doesn't reach the non-linear edge of stability regime, and increase ρ, we see that SAM eventually leads to a decrease in the largest eigenvalues, before leading to unstable behavior (Figure 7, left). If we plot $\eta(\lambda_{max} + \rho\lambda_{max}^2)$, we see that this normalized, effective eigenvalue stabilizes very close to 2 for a range of ρ, and for larger ρ stabilizes near but not exactly at 2 (Figure 7, right). We leave a more detailed understanding of this stabilization for future work.

B. SAM edge of stability

B.1. Proof of Theorem 2.3

We prove the following theorem, which gives us an alternate edge of stability for SAM:

Theorem 2.3. Consider a $\mathcal{C}^\infty$ model $\mathbf{f}(\boldsymbol{\theta})$ trained using Equation 6 with MSE loss. Suppose that there exists a point $\boldsymbol{\theta}^*$ where $\mathbf{z}(\boldsymbol{\theta}^*) = 0$. Suppose that for some $\epsilon > 0$, we have the lower bound $\epsilon < \eta\lambda_i(1 + \rho\lambda_i)$ for the eigenvalues of the positive-definite symmetric matrix $\mathbf{J}(\boldsymbol{\theta}^*)\mathbf{J}(\boldsymbol{\theta}^*)^\top$. Given bounds on the largest eigenvalues, there are two regimes:

Convergent regime. If $\eta\lambda_i(1 + \rho\lambda_i) < 2 - \epsilon$ for all for all eigenvalues λ_i of $\mathbf{J}(\boldsymbol{\theta}^*)\mathbf{J}(\boldsymbol{\theta}^*)^\top$, there exists a neighborhood U of $\boldsymbol{\theta}^*$ such that $\lim_{t\to\infty}\mathbf{z}_t = 0$ with exponential convergence for any trajectory initialized at $\boldsymbol{\theta}_0 \in U$.

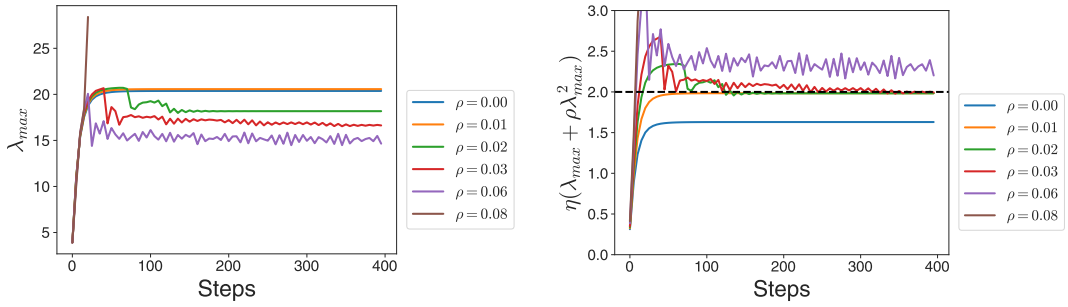

Figure 7. For fixed η, as ρ increases the largest eigenvalue of $\mathbf{J}\mathbf{J}^\top$ decreases, until training is no longer stable (left). For intermediate ρ, the eigenvalue is very well predicted by $\eta(\lambda_{max} + \rho\lambda_{max}^2) = 2$ (right); however there is also a range of ρ where training is still stable but $\eta(\lambda_{max} + \rho\lambda_{max}^2) > 2$ (purple curve).

Divergent regime. If $\eta\lambda_i(1 + \rho\lambda_i) > 2 + \epsilon$ for some eigenvector $\mathbf{v}_i$ of $\mathbf{J}(\boldsymbol{\theta}^*)\mathbf{J}(\boldsymbol{\theta}^*)^\top$, then there exists some q_{min} such that for any $q < q_{min}$, given $B_q(\boldsymbol{\theta}^*)$, the ball of radius q around $\boldsymbol{\theta}^*$, there exists some initialization $\boldsymbol{\theta}_0 \in B_q(\boldsymbol{\theta}^*)$ such that the trajectory $\{\boldsymbol{\theta}_t\}$ leaves $B_q(\boldsymbol{\theta}^*)$ at some time t.

Proof. The SAM update for MSE loss can be written as:

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = -\eta\mathbf{J}^\top(\boldsymbol{\theta}_t + \rho\mathbf{J}_t^\top\mathbf{z}_t)\mathbf{z}(\boldsymbol{\theta}_t + \rho\mathbf{J}_t^\top\mathbf{z}_t) \tag{33}$$

We will use the differentiability of $f(\boldsymbol{\theta})$ to Taylor expand the update step, and divide it into a dominant linear piece which leads to convergence, and an higher order term which can be safely ignored in the long term dynamics.

Since the model $f(\boldsymbol{\theta})$ is analytic at $\boldsymbol{\theta}^*$, there is a neighborhood U_r of $\boldsymbol{\theta}^*$ with the following properties: for $\boldsymbol{\theta} \in U_r$, $\mathbf{z}$ and $\mathbf{J}$ there exists a radius r such that

$$\mathbf{z}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) - \mathbf{z}(\boldsymbol{\theta}) = \mathbf{J}\Delta\boldsymbol{\theta} + \frac{1}{2}\frac{\partial^2\mathbf{z}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\Delta\boldsymbol{\theta}, \Delta\boldsymbol{\theta}) + \dots \tag{34}$$

$$\mathbf{J}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta}) = \frac{\partial^2\mathbf{z}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\Delta\boldsymbol{\theta}, \cdot) + \frac{1}{2}\frac{\partial^3\mathbf{z}}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_3}(\Delta\boldsymbol{\theta}, \Delta\boldsymbol{\theta}, \cdot) + \dots \tag{35}$$

for all $||\Delta\boldsymbol{\theta}|| < r$. The derivatives which define the power series are taken at $\boldsymbol{\theta}$. By Taylor's theorem, there exists some $b > 0$ such that we have the bounds

$$||\mathbf{z}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) - \mathbf{z}(\boldsymbol{\theta}) - \mathbf{J}\Delta\boldsymbol{\theta}|| \leq b||\Delta\boldsymbol{\theta}||^2 \tag{36}$$

$$||\mathbf{J}(\boldsymbol{\theta} + \Delta\boldsymbol{\theta}) - \mathbf{J}(\boldsymbol{\theta}) - \frac{\partial^2\mathbf{z}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\Delta\boldsymbol{\theta}, \cdot)|| \leq b||\Delta\boldsymbol{\theta}||^2 \tag{37}$$

for $||\Delta\boldsymbol{\theta}|| < r$ uniformly over U_r.

In addition, since $\mathbf{J}(\boldsymbol{\theta}^*)\mathbf{J}(\boldsymbol{\theta}^*)^\top$ has $\epsilon < \eta\lambda_i(1 + \rho\lambda_i)$ for all eigenvalues λ_i, there exists a neighborhood $V_{\epsilon,1/2}$ of $\boldsymbol{\theta}^*$ such that $\epsilon/2 < \eta\lambda_i(1 + \rho\lambda_i)$ for all eigenvalues λ_i of $\mathbf{J}\mathbf{J}^\top$, as well as λ_{max} of $\mathbf{J}\mathbf{J}^\top$ is bounded by $\eta\lambda_i(1 + \rho\lambda_i) < 2 - \epsilon/2$ in the *convergent* case, and $2\lambda_{max}(\boldsymbol{\theta}^*)$ in the *divergent* case for any $\boldsymbol{\theta} \in V_{\epsilon,1/2}$. Finally, for any $d > 0$, there exists a connected neighborhood T_d of $\boldsymbol{\theta}$ given by the set of points where $||\mathbf{z}|| < d$.

Consider the (non-empty) neighborhood $X_{r,a,d} = T_d \cap U_r \cap V_{\epsilon,1/2}$ given by the intersection of these sets. To recap, points $\boldsymbol{\theta}$ in $X_{r,d}$ have the following properties:

- $\mathbf{z}$ and $\mathbf{J}$ have power series representations around $\boldsymbol{\theta}$ with radius $r > 0$.
- The second-order and higher terms are bounded by $b||\Delta\boldsymbol{\theta}||^2$ uniformly on $X_{r,d}$, independently of d.
- $||\mathbf{z}(\boldsymbol{\theta})|| < d$.

- The eigenvalues of $\mathbf{J}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})^\top$ are bounded from below by $\epsilon/2 < \eta\lambda_i(1+\rho\lambda_i)$ and above by $\eta\lambda_i(1+\rho\lambda_i) < 2-\epsilon/2$ (convergent case) or $2\lambda_{max}(\boldsymbol{\theta}^*)$ (divergent case).

We now proceed with analyzing the dynamics. If $||\rho\mathbf{J}_t\mathbf{z}_t|| < r$, then we have:

$$\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = -\eta(\mathbf{J}_t^\top + \rho\mathbf{J}_t^\top\mathbf{J}_t\mathbf{J}_t^\top)\mathbf{z}_t + O(b||\rho\mathbf{J}_t^\top\mathbf{z}_t||^2) \tag{38}$$

We note that $||\rho\mathbf{J}_t\mathbf{z}_t|| < A||\mathbf{z}_t||$ on $X_{r,d}$ for some constant A independent of d, since the singular values of $\mathbf{J}_t$ are bounded uniformly from above. Therefore, if we choose $d < r/A$, the power series representation exists for all $\boldsymbol{\theta} \in X_{r,d}$.

If $||\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t|| < r$, then both $\mathbf{z}(\boldsymbol{\theta}_{t+1}) - \mathbf{z}(\boldsymbol{\theta}_t)$ as well as $\mathbf{J}(\boldsymbol{\theta}_{t+1}) - \mathbf{J}(\boldsymbol{\theta}_t)$ can be represented as power series centered around $\boldsymbol{\theta}_t$. We can again use the uniform bound on the singular values of $\mathbf{J}$, as well as the uniform bound on the error terms to choose d small enough such that $||\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t|| < r$ always for $\boldsymbol{\theta}_t \in X_{r,d}$.

Therefore, for sufficiently small d, we have:

$$\mathbf{z}(\boldsymbol{\theta}_{t+1}) - \mathbf{z}(\boldsymbol{\theta}_t) = \mathbf{z}_{t+1} - \mathbf{z}_t = -\eta\mathbf{J}_t\mathbf{J}_t^\top[(1+\rho\mathbf{J}_t\mathbf{J}_t^\top)\mathbf{z}_t] + O(h||\mathbf{z}_t||^2) \tag{39}$$

$$\mathbf{J}(\boldsymbol{\theta}_{t+1}) - \mathbf{J}(\boldsymbol{\theta}_t) = -\eta\frac{\partial^2\mathbf{z}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\mathbf{J}_t^\top\mathbf{z}_t, \cdot) + O(h||\mathbf{z}_t||^2) \tag{40}$$

for some constant h independent of d.

We first analyze the dynamics in the convergent case. We first establish that $||\mathbf{z}||^2$ decreases exponentially at each step, and then confirm that the trajectory remains in $X_{\epsilon,1/2}$. Consider a single step in the eigenbasis of $\mathbf{J}_t\mathbf{J}_t^\top$. Let $z(i)$ be the projection $\mathbf{v}_i \cdot \mathbf{z}$ for eigenvector $\mathbf{v}_i$ corresponding to eigenvalue λ_i. Then we have:

$$z(i)_{t+1}^2 - z(i)_t^2 = (-\eta\lambda_i(1+\rho\lambda_i)z(i)_t + O(||\mathbf{z}_t||^2))([2-\eta\lambda_i(1+\rho\lambda_i)]z(i)_t + O(||\mathbf{z}_t||^2)) \tag{41}$$

From our bounds, we have

$$z(i)_{t+1}^2 - z(i)_t^2 \leq -\frac{1}{2}\epsilon z(i)_t^2 + c||\mathbf{z}_t||^3 \tag{42}$$

By uniformity of the Taylor approximation we again have that c is uniform, independent of a and d. Summing over the eigenmodes, we have:

$$||\mathbf{z}_{t+1}||^2 - ||\mathbf{z}_t||^2 \leq -\frac{1}{2}\epsilon||\mathbf{z}_t||^2 + Dc||\mathbf{z}_t||^3 \tag{43}$$

If we choose $d < \frac{\epsilon}{4cD}$, then we have

$$||\mathbf{z}_{t+1}||^2 - ||\mathbf{z}_t||^2 \leq -\frac{1}{4}\epsilon||\mathbf{z}_t||^2 \tag{44}$$

Therefore $||\mathbf{z}_{t+1}||^2$ decreases by a factor of at least $1-\epsilon/4$ each step.

In order to complete the proof over the convergent regime, we need to show that $\mathbf{J}_t$ changes by a small enough amount that the upper and lower bounds on the eigenvalues are still valid - that is, the trajectory remains in $X_{\epsilon,1/2}$. Suppose the dynamics begins with initial residuals $\mathbf{z}_0$, and remains within $X_{\epsilon,1/2}$ for t steps. Consider the $t+1$th step. The total change in $\mathbf{J}$ can be bounded by:

$$||\mathbf{J}_{t+1} - \mathbf{J}_0|| \leq B\sum_t ||\mathbf{z}_t|| + C\sum_t ||\mathbf{z}_t||^2 \tag{45}$$

for some constants B and C independent of d. The first term comes from a uniform upper bound on $-\eta\frac{\partial^2\mathbf{z}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\mathbf{J}_t^\top\mathbf{z}_t, \cdot)$, and the second term comes from the uniform upper bound on the higher order corrections to changes in $\mathbf{J}$ for each step. Using the bound on $||\mathbf{z}_t||$, we have:

$$||\mathbf{J}_{t+1} - \mathbf{J}_0|| \leq \frac{4(B+C)}{\epsilon}||\mathbf{z}_0|| \tag{46}$$

If the right hand side of the inequality is less than $\epsilon^{(1+\delta)/2}$, for any $\delta > 0$, then the change in the singular values is $o(\epsilon^{1/2})$, the change in the eigenvalues of $\mathbf{J}\mathbf{J}^\top$ is $o(\epsilon)$, and the trajectory remains in $V_{\epsilon,1/2}$ at time $t+1$. Let $d \leq \frac{1}{4(B+C)}\epsilon^{(3+\delta)/2}$. Then, $||\mathbf{J}_{t+1} - \mathbf{J}_0|| \leq \epsilon^{(1+\delta)/2}$ for all t.

Therefore the trajectory remains within $X_{r,d}$, and $||\mathbf{z}_t||$ converges exponentially to 0, for any d sufficiently small. Therefore there is a neighborhood of $\boldsymbol{\theta}^*$ where $||\mathbf{z}||$ converges exponentially to 0.

Now we consider the divergent regime. We will show that we can find initializations with arbitrarily small $||\mathbf{z}||$ and $||\boldsymbol{\theta} - \boldsymbol{\theta}^*||$ which eventually have increasing $||\mathbf{z}||$.

Since $\mathbf{J}\mathbf{J}^\top$ is full rank, there exists some $\boldsymbol{\theta}_0$ in any neighborhood of $\boldsymbol{\theta}^*$ such that $|\mathbf{v}_m \cdot \mathbf{z}(\boldsymbol{\theta}_0)| > 0$ where $\mathbf{v}_m$ is the direction of the largest eigenvalue of $\mathbf{J}\mathbf{J}^\top$. Consider such a $\boldsymbol{\theta}_0$ in $X_{r,d}$ (and therefore in T_d as well. The change in the magnitude of this component m of $\mathbf{z}$ is bounded from below by

$$z(m)_1^2 - z(m)_0^2 \geq \frac{1}{2}\epsilon z(m)_1^2 - c||\mathbf{z}_0||^3 \tag{47}$$

Again the correction is uniformly bounded independent of d. Therefore the bound becomes

$$z(m)_1^2 - z(m)_0^2 \geq \frac{1}{4}\epsilon z(m)_0^2 \tag{48}$$

Choose d_{min} such that the above bound holds for $d < d_{min}$. Furthermore, choose q_{min} so that the ball $B_{q_{min}}(\boldsymbol{\theta}^*) \subset X_{r,d_{min}}$. Pick an initialization $\boldsymbol{\theta}_0 \in B_q(\boldsymbol{\theta}^*)$ for $q < q_{min}$.

After a single step, there are two possibilities. The first is that $\boldsymbol{\theta}_1$ is no longer in $B_{q_{min}}(\boldsymbol{\theta}^*)$. In this case the trajectory has left $B_q(\boldsymbol{\theta}^*)$ and the proof is complete.

The second is that $\boldsymbol{\theta}_1$ remains in $B_{q_{min}}(\boldsymbol{\theta}^*)$. In this case, $z(m)_1^2$ is bounded from below by $(1 + 1/4\epsilon)z(m)_0^2$. If the trajectory remains in $B_{q_{min}}(\boldsymbol{\theta}^*)$ for t steps, we have the bound:

$$||\mathbf{z}_t||^2 \geq (1 + 1/4\epsilon)^t z(m)_0^2 \tag{49}$$

Therefore, at some finite time t, $||\mathbf{z}_t||^2 \geq d$, and $\boldsymbol{\theta}$ leaves $X_{r,d_{min}}$. Therefore it leaves $B_q(\boldsymbol{\theta}^*)$. This is true for any $q < q_{min}$. This completes the proof for the divergent case.

□

C. CIFAR-10 experiment details

C.1. Effect of learning rate

The analysis in Section 4 focused on the optimal learning rate $\eta = 0.3$, which is well in the edge of stability regime. By performing a learning rate sweep, we can see that the SAM-EOS has similar phenomenology to traditional EOS (Figure 8). For small learning rates, the edge of stability is not reached, and training proceeds without issue. For intermediate learning rates, the regular EOS is not reached but the SAM-EOS is reached. Finally for larger learning rates, the training dynamics becomes more unstable and eventually diverges (for $\eta \approx 0.5$, not plotted here).

We also note that the relative effect of SAM becomes weaker as learning rate increases for fixed ρ. This is consistent with the theory which suggests that the ratio $r = \rho/\eta$ controls the SAM correction to the learning rate normalized eigenvalue used in EOS. This is demonstrated in Figure 8 by the $\eta = 0.4$ curves (red), which are near the GD-EOS, unlike the $\eta = 0.1$ and $\eta = 0.2$ curves which are near the SAM-EOS but not the GD-EOS.

C.2. Cross-entropy loss

Many of the trends observed using MSE loss in Section 4 can also be observed for cross-entropy loss. Eigenvalues generally increase at late times, and there is still a regime where SGD shows EOS behavior in $\eta\lambda_{max}$, while SAM shows EOS behavior in $\eta(\lambda_{max} + \rho\lambda_{max}^2)$ (Figure 9). In addition, the gradient norm is stil stable for much of training, with SGD gradient norm decreasing at the end of training while SAM gradient norms stay relatively constant (Figure 9).

There are also qualitative differences in the behavior. For example, the eigenvalue decrease starts earlier in training. Decreasing eigenvalues for cross-entropy loss have been previously observed (Cohen et al., 2022a), and there is evidence that the origin of the effect is due to the interaction of the logit magnitude with the softmax function. The gradient magnitudes also have an initial rapid fall-off period. Overall more study is needed to understand how the effects and mechanisms used by SAM depend on the loss used.

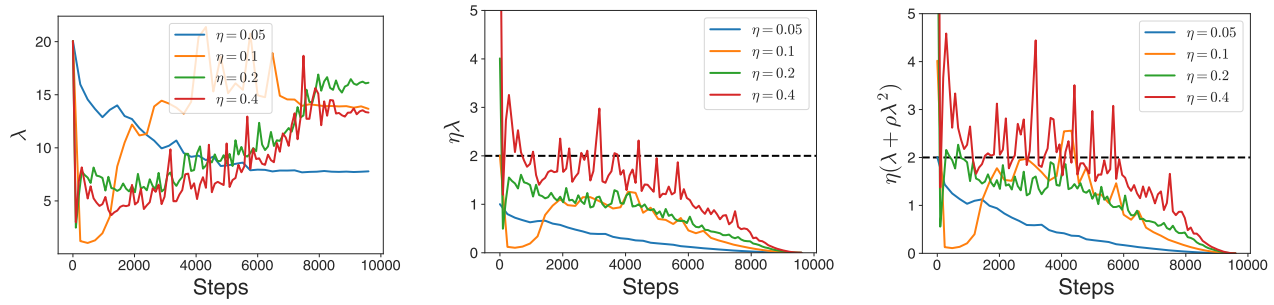

Figure 8. Largest Hessian eigenvalues for CIFAR10 trained with MSE loss and SAM radius $\rho = 0.05$, various learning rates. Behavior is similar to traditional EOS scenarios. For small learning rates, EOS is never reached. For intermediate learning rates, SAM-EOS is reached for at least some time. Finally, for larger learning rates learning becomes more variable and unstable, and model can't train for $\eta = 0.5$.

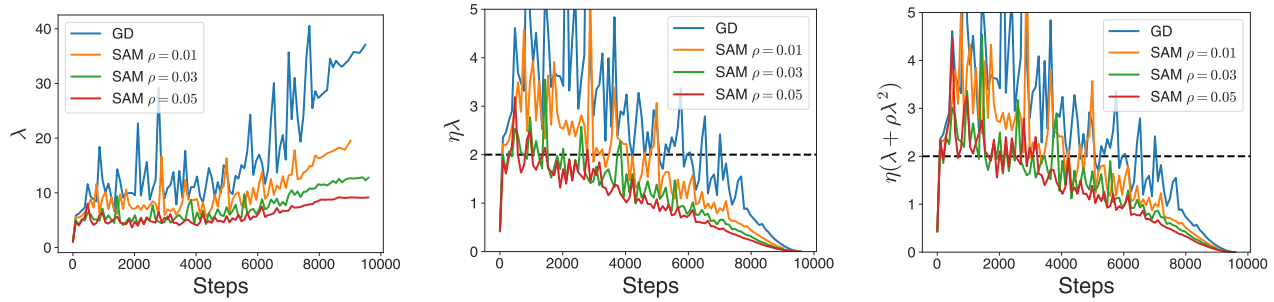

Figure 9. Largest Hessian eigenvalues for CIFAR10 trained with cross-entropy loss. Trends are similar to MSE loss (Figure 4), with the exception that normalized eigenvalues decrease from an earlier time.

C.3. Alternative sharpness measures

While the main focus of the text is focused on the largest eigenvalue of the Hessian, for SGD learning there are a number of alternative sharpness measures. In the recent work of Lee & Jang (2023), the propose the *implicit interaction regularizer* (IIR)

$$\lambda_{IIR} = \mathrm{E}[\mathbf{g}_B^{\mathrm{T}} \mathbf{H} \mathbf{g}_B] / \mathrm{E}[\mathbf{g}_D \cdot \mathbf{g}_D] \tag{50}$$

as an alternative definition of the EOS for SGD. Here $\mathbf{g}_B$ is the minibatch gradient of the loss with respect to parameters (sampled i.i.d. without replacement) and $\mathbf{g}_D$ is the full batch gradient. Averages in this quantity (and section) are taken over i.i.d. sampling of training data for an SGD step. (The notation λ_{IIR} for this quantity is our own and used for consistency with the rest of the text.)

This quantity has the property that the loss is decreasing if $\eta\lambda_{IIR} < 2$, and increasing otherwise. More precisely, $E[\mathcal{L}_{t+1} - \mathcal{L}_t] < 0$ if and only if $\eta\lambda_{IIR} < 2$, up to some correction of $O(\eta^3)$. This quantity has the advantage of explicitly considering SGD dynamics, but the disadvantage of crossing the EOS threshold only when the gradients are sufficiently aligned with the large eigendirections. If the stabilization of the large eigenmodes keeps the overall loss dynamics decreasing, then λ_{IIR} may remain significantly below the edge of stability even if λ_{max} is near it.

By plotting λ_{IIR} for the CIFAR10 models trained on MSE loss ($\eta = 0.3$), we can see that we get similar dynamics to λ_{max} (Figure 11). There is stabilization of $\eta\lambda_{IIR}$, which is significantly below the GD-EOS when SAM is turned on (middle). If we instead compute the SAM-adjusted values $\eta(\lambda_{IIR} + \rho\lambda_{IIR}^2)$, then all curves are closer to the SAM-EOS. Note that $\lambda_{IIR} < \lambda_{max}$, and that particularly for larger ρ the trajectories remain further away (lower) then the edge of stability.

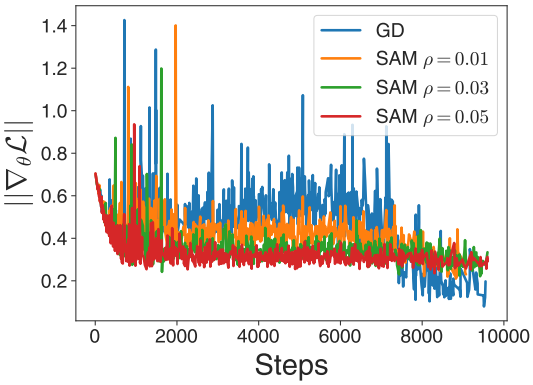

Figure 10. Minibatch gradient magnitudes for CIFAR-10 model trained on cross-entropy loss. Trends are similar to MSE loss (Figure 9), with larger overall variation in gradient values.

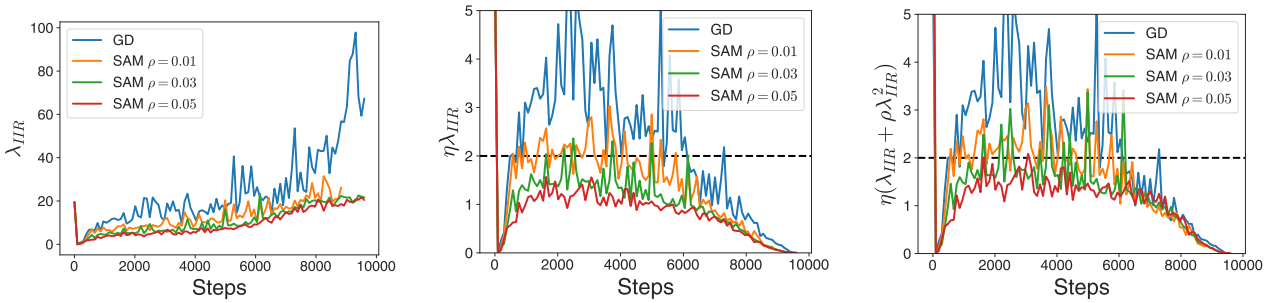

Figure 11. Alternative sharpness measure $\lambda_{IIR} = \mathrm{E}[\mathbf{g}_B^{\mathrm{T}}\mathbf{H}\mathbf{g}_B]/\mathrm{E}[\mathbf{g}_D \cdot \mathbf{g}_D]$ for training on MSE loss. Trajectories are similar to those of λ_{max}, with SAM trajectories generally below the GD-EOS, but $\lambda_{IIR} + \rho\lambda_{IIR}^2$ near the SAM-EOS. λ_{IIR} tends to be smaller than λ_{max}, and accordingly lower than the SAM-EOS threshhold, particularly for larger ρ.