Second-order regression models exhibit progressive sharpening to the edge of stability

Atish Agarwala [1] **Fabian Pedregosa** [1] **Jeffrey Pennington** [1]

Abstract

Recent studies of gradient descent with large step sizes have shown that there is often a regime with an initial increase in the largest eigenvalue of the loss Hessian (progressive sharpening), followed by a stabilization of the eigenvalue near the maximum value which allows convergence (edge of stability). These phenomena are intrinsically non-linear and do not happen for models in the constant Neural Tangent Kernel (NTK) regime, for which the predictive function is approximately linear in the parameters. As such, we consider the next simplest class of predictive models, namely those that are quadratic in the parameters, which we call second-order regression models. For quadratic objectives in two dimensions, we prove that this second-order regression model exhibits progressive sharpening of the NTK eigenvalue towards a value that differs slightly from the edge of stability, which we explicitly compute. In higher dimensions, the model generically shows similar behavior, even without the specific structure of a neural network, suggesting that progressive sharpening and edge-of-stability behavior aren't unique features of neural networks, and could be a more general property of discrete learning algorithms in high-dimensional non-linear models.

1. Introduction

A recent trend in the theoretical understanding of deep learning has focused on the *linearized* regime, where the Neural Tangent Kernel (NTK) controls the learning dynamics (Jacot et al., 2018; Lee et al., 2019). The NTK describes learning dynamics of all networks over short enough time horizons, and can describe the dynamics of wide networks over large time horizons. In the NTK regime, there is a function-space ODE which allows for explicit characterization of the network outputs (Jacot et al., 2018; Lee et al., 2019; Yang, 2021). This approach has been used across the board to gain insights into wide neural networks, but it suffers a major limitation: the model is linear in the parameters, so it describes a regime with relatively trivial dynamics that cannot capture feature learning and cannot accurately represent the types of complex training phenomena often observed in practice.

While other large-width scaling regimes can preserve some non-linearity and allow for certain types of feature learning (Bordelon & Pehlevan, 2022; Yang et al., 2022), such approaches tend to focus on the small learning-rate or continuous-time dynamics. In contrast, recent empirical work has highlighted a number of important phenomena arising from the non-linear discrete dynamics in training practical networks with large learning rates (Neyshabur et al., 2017; Gilmer et al., 2022; Ghorbani et al., 2019; Foret et al., 2022). In particular, many experiments have shown the tendency for networks to display *progressive sharpening* of the curvature towards the *edge of stability*, in which the maximum eigenvalue of the loss Hessian increases over the course of training until it stabilizes at a value equal to roughly two divided by the learning rate, corresponding to the largest eigenvalue for which gradient descent would converge in a quadratic potential (Wu et al., 2018; Giladi et al., 2020; Cohen et al., 2022a;b).

In order to build a better understanding of this behavior, we introduce a class of *quadratic regression models* which display all the relevant phenomenology yet are simple enough to admit numerical and analytic understanding. Models of this type have been investigated previously in order to understand other phemomena that occur beyond the NTK regime (Zhu et al., 2022a; Roberts et al., 2022). In a particular low-dimensional setting, we prove that the maximum NTK eigenvalue converges to (near) the edge of stability, and we demonstrate empirically that progressive sharpening to the edge of stability occurs *generically* in the large datapoint, large model limit. Finally, we conduct a numerical analysis on the properties of a real neural network and use tools from our theoretical analysis to show that edge-of-stability behavior "in the wild" shows some of the same patterns as the theoretical models.

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

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1.1. Concurrent work

Several concurrent works have investigated similar questions about edge of stability and have developed insights that nicely complement the conclusions we develop here.

Zhu et al. (2022b) propose a minimalist model in the form of a degree-4 function (or degree-8 objective) and prove that sharpness converges to a value close to, but not equal to, the edge of stability. While these results are similar to our results in the low-dimensional setting, we find that edge-of-stability behavior can occur for degree-2 functions, and that the final sharpness differs from the convergence threshold by a precise problem-dependent amount which we calculate.

Damian et al. (2022) develop an explanatory model for stabilization of sharpness near the edge of stability based on the nonlinear self-interaction of the top eigenmode of the Hessian as well as its effective interaction with the other eigenmodes. The low-dimensional dynamics capture the full behavior over short timescales when the sharpness is near the stability threshold, and closely resemble the differential equations we develop in the low-dimensional limit of our model. Our setup is somewhat less general, as it is restricted to quadratic models, but it affords us some additional insights beyond what was possible in Damian et al. (2022). In particular, we are able to prove that progressive sharpening occurs generically in high-dimensional models, at least at early times, and we find that the final curvature deviates by a predictable amount from the convergence threshold.

2. Quadratic regression model

We begin by defining our basic *quadratic regression model*. Given a P-dimensional parameter vector θ, the D-dimensional output $\mathsf{f}(\theta)$ is given by

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

Here y is a D-dimensional vector, 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, $\mathbf{Q}(\theta,\theta)$ takes two P-dimensional vectors as input, and outputs a D-dimensional vector verifying $\mathbf{Q}(\theta,\theta') = \mathbf{Q}(\theta',\theta)$. If $\mathbf{Q} = 0$, the model corresponds to linear regression. We can recover the $D \times P$-dimensional Jacobian $\mathsf{J} = \partial\mathsf{f}/\partial\theta$ using

$$\mathsf{G}_{i} = \left.\frac{\partial \mathsf{f}}{\partial \theta_i}\right|_{\theta=0}; \quad \mathbf{Q}_{ij} = \left.\frac{\partial^2 \mathsf{f}}{\partial\theta_i \partial\theta_j}\right|_{\theta=0} \implies \mathsf{J} = \mathsf{G} + \mathbf{Q}(\theta,\theta). \tag{2}$$

Any model can be converted into a quadratic regression model by a second-order Taylor expansion around any differentiable point. Quadratic expansions of shallow MLPs have been previously studied (Bai & Lee, 2020; Zhu et al., 2022a), and the perturbation theory for small $\mathbf{Q}$ is studied in Roberts et al. (2022). Other related models are detailed in Appendix A. We will provide evidence that even random, unstructured quadratic regression models lead to *edge of stability* (EOS) behavior.

In this work we focus on the MSE loss setting. More explicitly, given a D-dimensional target vector $\hat{\mathsf{y}}$, the loss $L(\theta)$ can be written in terms of the residuals $\mathsf{z} = \mathsf{f}(\theta) - \hat{\mathsf{y}}$:

$$L(\theta) = \frac{1}{2}\sum \mathsf{z}^2 = \frac{1}{2}\sum \left(\mathsf{y} - \hat{\mathsf{y}} + \mathsf{G}^{\top}\theta + \frac{1}{2}\mathbf{Q}(\theta,\theta)\right)^2. \tag{3}$$

As we will show, in this setting the dynamics can be written in function space in terms of z and J alone.

3. Low-dimensional dynamics

We begin by focusing on the dynamics for $D = 1$ - a single datapoint. Without loss of generality, we can write the loss function as:

$$L(\theta) = \frac{1}{2}\left(\frac{1}{2}\theta^{\top}\mathbf{Q}\theta - E\right)^2 \tag{4}$$

for a $P \times P$ matrix $\mathbf{Q}$ and scalar E which define the problem. We will analyze the dynamics in terms of the (scalar) residual $z = \frac{1}{2}\theta^{\top}\mathbf{Q}\theta - E$ and the $1 \times P$ dimensional Jacobian $\mathsf{J} = \partial z/\partial\theta$. In particular, we will be interested in the dynamics of the scalar curvature $\mathsf{J}\mathsf{J}^{\top}$ (the neural tangent kernel, or NTK).

3.1. Gradient flow

We first consider gradient flow (GF) dynamics of the loss L with respect to the parameters θ. Given a scaling factor η, the GF dynamics are given by

$$\dot{\theta} = -\eta \nabla L = -\eta z\frac{\partial z}{\partial\theta} = -\frac{\eta}{2}\theta^{\top}\mathbf{Q}\theta\,\mathbf{Q}\theta - \eta E\,\mathbf{Q}\theta. \tag{5}$$

The dynamics of z and J close:

$$\dot{z} = -\eta(\mathsf{J}\mathsf{J}^{\top})z, \quad \dot{\mathsf{J}} = -\eta z\mathsf{J}\mathbf{Q}. \tag{6}$$

The curvature is a scalar, described by the NTK $\mathsf{J}\mathsf{J}^{\top}$. In these coordinates, we have $E = \mathsf{J}\mathbf{Q}^{+}\mathsf{J}^{\top}z$, where $\mathbf{Q}^{+}$ denotes the Moore-Penrose pseudoinverse.

We can make some coordinate transformations to simplify the dynamics. We define:

$$\bar{z} = \eta z, \; T(k) = \mathsf{J}\mathbf{Q}^{k}\mathsf{J}^{\top}. \tag{7}$$

We note that $T(0) = \eta\mathsf{J}\mathsf{J}^{\top}$ - that is, the curvature normalized by the scaling factor. The dynamical equations are:

$$\frac{d\bar{z}}{dt} = -\bar{z}T(0), \quad \frac{dT(k)}{dt} = -2\bar{z}T(k+1). \tag{8}$$

Figure 1. $D = 1$ loss landscape $L(\theta)$ as a function of the parameters, where $P = 2$, $E = 0$ and Q has eigenvalues 1 and -0.1. The GD trajectories (initialized at $(1.5, -4.32)$, marked with an x) converge to minima with larger curvature than at initialization and therefore show progressive sharpening (left). The two-step dynamics, in which we consider only even iteration numbers, exhibit fewer oscillations near the edge of stability (right).

In order to study edge of stability behavior, we need initializations which allow the curvature $\tilde{T}(0)$ to increase over time - a phenomenon known as *progressive sharpening*. Progressive sharpening has been shown to be ubiquitous in machine learning models (Cohen et al., 2022a), so any useful phenomenological model should show it as well. In Appendix B.1, we confirm that the model shows progressive sharpening over a range of initial conditions.

3.2. Gradient descent

We will now show that for gradient descent (GD) dynamics with learning rate η, there is a family of initializations where the $D = 1$ quadratic regression model shows edge of stability (EOS) behavior. For this model, we define EOS behavior as a setting where the dynamics causes λ_{max}, the maximum eigenvalue value of the NTK $JJ^\top$, to remain close to the critical value $2/\eta$. This critical value corresponds to the largest learning rate where the dynamics converges exponentially near a minimum of the loss. This boundary corresponds to $\tilde{T}(0) = 2$ - hence the interpretation of $\tilde{T}(0)$ as the rescaled curvature.

We will show that there is a family of models such that, for any distance, there is a model which leads to EOS convergence to that distance away from the edge of stability - for any initialization in a non-trivial volume of initializations. Additionally we will show that this initialization volume is uniform across models in a natural set of coordinates.

We will prove this form of EOS in $D = 1$ model, and find empirical evidence that this holds for for large D as well. Note that we define the EOS with respect to the maximum NTK eigenvalue instead of the maximum loss Hessian eigenvalue from (Cohen et al., 2022a); see Appendix A.1 for discussion of why this is appropriate for MSE loss.

When Q has both positive and negative eigenvalues, the loss landscape is the square of a hyperbolic parabaloid (Figure 1, left). As suggested by the gradient flow analysis, this causes some trajectories to increase their curvature before convergence. The final curvature naturally depends on both the initialization and learning rate. One of the challenges in analyzing the gradient descent dynamics is that they rapidly and heavily oscillate around minima for large learning rates. One way to mitigate this issue is to consider only every other step (Figure 1, right). We will use this observation to analyze the gradient descent (GD) dynamics directly to find configurations where these trajectories show edge-of-stability behavior.

Let θ_t be the parameters at step t. The gradient descent equations are given by:

$$\theta_{t+1} - \theta_t = -\nabla L = -\frac{\eta}{2} r_t^\top Q r_t - E^\top Q r_t. \quad (9)$$

In the $z - T$ coordinates, the gradient descent equations become (Appendix B.2):

$$z_{t+1} - z_t = -z_t T_t(0) + \frac{1}{2}(z_t^2) T_t(1) \quad (10)$$

$$T_{t+1}(k) - T_t(k) = -z_t(2T_t(k+1) - z_t T_t(k+2)). \quad (11)$$

If Q is invertible, $\tilde{E} = E = T(-1) \to z = 0$. By definition $T_t(0) = J_t J_t^\top = \lambda_{max;t}$ is the (rescaled) NTK eigenvalue. In these coordinates, EOS behavior corresponds to dynamics which keep $T_t(0)$ near the value 2 as z_t goes to 0.

In the remainder of this section, we will focus on the $P = 2$ case. As we will see, this lets us write the dynamics in terms of z and $T(0)$ (the residuals and curvature) alone.

3.2.1. Reduction to catapult dynamics

If the eigenvalues of Q are $\{-\omega, \omega\}$, and $\tilde{E} = 0$, the model becomes equivalent to a single hidden layer linear network

Figure 2. Left: plotting curvature for $D = 1, P = 2$ model shows EOS behavior for various step sizes ($\eta = 5 \cdot 10^{-3}$). Middle: plotting every other iterate, we see that a variety of initializations (black x's), trajectories in $(z, T(0))$ space stay near the nullcline $(z, f_z(z))$ - the curve where $z_{t+2} - z_t = 0$. Arrows represent direction of dynamical flows. Right: Changing variables to $(z, T(0) - f_z(z))$ shows quick concentration to a curve of near-constant, small, negative γ

with one training datapoint (Appendix A.2) - also known as the catapult phase dynamics. This model doesn't exhibit sharpening or edge-of-stability behavior (Lewkowycz et al., 2020). We can prove this in the z, $T(0)$ coordinates.

We assume without loss of generality that the eigenvalues are $\{-1, 1\}$ - which can be accomplished by rescaling x. We can rewrite the dynamics in terms of z and the curvature $T(0)$ only (Appendix B.3):

$$z_{t+1} - z_t = -\eta z_t T_t(0) + \frac{1}{2}\eta^2 z_t^2 (2\eta z_t + \eta^2) \tag{12}$$

$$T_{t+1}(0) - T_t(0) = -2\eta z_t (2\eta z_t + \eta^2) + \eta^2 z_t^2 T_t(0) + \cdots \tag{13}$$

For $\eta^2 = 0$, we can see that $\text{sign}(\Delta T(0)) = \text{sign}(2 - T_t(0))$, as in (Lewkowycz et al., 2020) - so convergence implies strictly decreasing curvature. For $\eta^2 > 0$, there is a region where the curvature can increase (Appendix B.3). However, there is still no EOS behavior - if $T(0)$ starts away from $2/\eta$, there is no mechanism to stabilize it near the edge.

3.2.2. Edge of stability regime

In this section, we consider the case in which Q has two eigenvalues - one of which is large and positive, and the other one small and negative. Without loss of generality, we assume that the largest eigenvalue of Q is 1. We denote the second eigenvalue by $-\gamma$, for $0 < \gamma \leq 1$. With this notation we can write the dynamical equations (Appendix B.3) as

$$z_{t+1} - z_t = -\eta z_t T_t(0) + \frac{1}{2}\eta^2 z_t^2 ((1+\gamma) T_t(0) + \gamma(2\eta z_t + \eta^2)) \tag{14}$$

$$T_{t+1}(0) - T_t(0) = -2\eta z_t \left(\gamma(2\eta z_t + \eta^2) + (1+\gamma) T_t(0) \right) + \eta^2 z_t^2 \left[T_t(0) + \gamma(1+\gamma)(T_t(0) - \eta^2 - 2\eta z_t) \right] + \cdots \tag{15}$$

For small η, there are trajectories where λ_{max} is initially away from $2/\eta$ but converges towards it (Figure 2, left) - in other words, EOS behavior. We used a variety of step sizes but initialized at pairs initialized at pairs $(z_0, T_0(0))$ to show the universality of the z-$T(0)$ coordinates.

In order to quantitatively understand the progressive sharpening and edge of stability, it is useful to look at the two-step dynamics. One additional motivation for studying the two-step dynamics follows from the analysis of gradient descent on linear least squares (i.e., linear model) with a large step size η. For every coordinate $\tilde{z}$, the one-step and two-step dynamics are

$$\tilde{z}_{t+1} - \tilde{z}_t = -\eta \tilde{z}_t \text{ and } \tilde{z}_{t+2} - \tilde{z}_t = (1-\eta)^2 \tilde{z}_t \tag{16}$$

While the dynamics converge for $\eta < 2$, if $\eta > 1$ the one-step dynamics alternate sign when approaching minimum, whereas the the two-step dynamics maintain the sign of $\tilde{z}$ and the trajectories exhibit no oscillations.

Likewise, plotting every other iterate in the two parameter model more clearly demonstrates the phenomenology. For small η, the dynamics shows the distinct phases described in (Li et al., 2022): an initial increase in $T(0)$, a slow increase in z, then a decrease in $T(0)$, and finally a slow decrease of z while $T(0)$ remains near $2/\eta$ (Figure 2, middle).

Unfortunately, the two-step version of the dynamics defined by Equations 14 and 15 are more complicated – they are 3rd order in $T(0)$ and 9th order in z; see Appendix B.4 for a more detailed discussion. However, we can analyze the dynamics as z goes to 0. In order to understand the mechanisms of the EOS behavior, it is useful to understand the nullclines of the two step dynamics. The nullclines are the curves $(z, f_z(z))$ and $(z, f_T(z))$ consisting of initializations such that z and $T(0)$, respectively, are unchanged after two steps.

We define the multinomials $p_z(z, T)$ and $p_T(z, T)$ such that

Figure 3. GD dynamics in the quadratic regression model. Loss functions are monotonic for small η_z (left). For larger η_z, final λ_{max} is higher (right). As sharpening drives λ_{max} near 2, non-linear effects induce EOS behavior.

$z_{t+1} - z_t = p_z(z_t, T_t)$ and $T_{t+1}(0) - T_t(0) = p_T(z_t, T_t)$. Then $f_z(z)$ and $f_T(z)$ obey the implicit equations:

$$\begin{aligned} p_z(p_z(z, f_z(z)), p_T(z, f_z(z))) &= 0 \\ p_T(p_z(z, f_T(z)), p_T(z, f_T(z))) &= 0 \end{aligned} \tag{17}$$

These expressions are polynomials cubic in $f_z(z)$ and $f_T(z)$ respectively - so there are three possible solutions as z goes to 0. We are particularly interested in the solution that goes through $z = 0, T(0) = 2$ - that is, the critical point corresponding to EOS.

Calculations detailed in Appendix B.4 show that the distance between the two nullclines is linear in η, so they become close as η goes to 0. (Figure 2, middle). In addition, the trajectories stay near f_z - which gives rise to EOS behavior. This suggests that the dynamics are slow near the nullclines, and trajectories appear to be approaching an attractor. We can find the structure of the attractor by changing variables to $y_t - T_t(0) - f_z(z_t)$ - the distance from the z nullcline.

We can build intuition by expanding the dynamics to lowest order in z and y. A direct calculation in Appendix B.5 gives us the approximation:

$$z_{t+2} - z_t = 2y_t z_t + O(y_t^2 z_t) + O(y_t z_t^2) \tag{18}$$

$$y_{t+2} - y_t = -2(4 - 3\eta + 4\eta^2) y_t z_t^2 - 4\eta z_t^2 + O(\eta z_t^3 + y^2 z_t^2) \tag{19}$$

We immediately see that z changes slowly for small y - since we chose coordinates where $z_{t+2} - z_t = 0$ when $y = 0$. We can also see that $y_{t+2} - y_t$ is $O(\eta)$ for $y_t = 0$ - so for small η, the y dynamics is slow too. Moreover, we see that the coefficient of the z_t^2 term is negative - the changes in z tend to drive y (and therefore $T(0)$) to decrease. The coefficient of the y_t term is negative as well; the dynamics of y tends to be contractive. The key is that the contractive behavior takes y to an $O(\eta)$ fixed point at a rate proportional to η^2, while the dynamics of z are proportional to η. This suggests a separation of timescales $\eta^2 \ll \eta$, where y first equilibrates to a fixed value, and then z converges to 0 (Figure 2, right).

The intuition from the lowest order terms can be formalized, and gives us a prediction of $\lim_{t\to\infty} y_t = -\eta/2$, confirmed by the following theorem (proof in Appendix B.6):

Theorem 3.1. Consider a quadratic regression model with $D = 1, P = 2, E = 0$, and Q with eigenvalues $\{\alpha, 1\}$. There exists an $\eta_c > 0$ and an η-independent neighborhood $U \subset \mathbb{R}^3$ such that for $0 < \eta < \eta_c$ and all initializations such that $(z_0, \lambda_{max}(0), \beta) \in U$, sharpness converges to a value near the edge of stability:

$$\lim_{t\to\infty} \lambda_{max}(t) = (2 - \eta/2)/\eta + O(\eta^2). \tag{20}$$

Therefore, unlike the catapult phase model, the small η provably has EOS behavior - whose mechanism is well-understood by the z-y coordinate transformation.

4. High dimensional dynamics

In this section we analyze the dynamics for large D. We focus on randomly initialized Q, and prove that progressive sharpening is ubiquitous. We then demonstrate that the quadratic regression model shows edge of stability behavior over a range of initializations.

4.1. Gradient flow dynamics

The gradient flow dynamics on an MSE loss function L can be written generally as

$$\dot{\theta} = -\frac{\partial L(\theta)}{\partial \theta} = -J^\top \dot{z}. \tag{21}$$

where J is the $D \times P$-dimensional Jacobian, and $\dot{z}$ is the D-dimensional vector of residuals $f(x) - y$. For the quadratic regression model, the dynamics of z and J once again close:

$$\dot{z} = -JJ^\top z, \quad \dot{J} = -Q(J^\top z, \cdot). \tag{22}$$

When $Q = 0$ (linearized regime), J is constant, the dynamics are then linear in z, and are controlled by the eigenstructure of the $D \times D$ matrix $JJ^\top$, the empirical NTK.

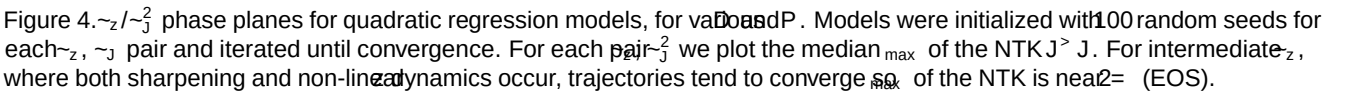

Figure 4. σ_z/σ_J^2 phase planes for quadratic regression models, for various $D = P$. Models were initialized with 100 random seeds for each σ_z, σ_J pair and iterated until convergence. For each pair σ_z, σ_J^2 we plot the median λ_{max} of the NTK $JJ^\top$. For intermediate σ_z, where both sharpening and non-linear dynamics occur, trajectories tend to converge so λ_{max} of the NTK is near $2/\eta$ (EOS).

We are interested in settings where progressive sharpening occurs under GF. We can study the dynamics of the maximum eigenvalue λ_{max} of $JJ^\top$ at early times for random initializations. In Appendix C.1, we prove the following theorem:

Theorem 4.1. Let z, J, and Q be initialized with i.i.d. elements with zero mean and variances σ_z^2, σ_J^2, and 1 respectively, with distributions invariant to rotation in data and parameter space, and have finite fourth moments. Let λ_{max} be the largest eigenvalue of $JJ^\top$. In the limit of large D and P, with fixed ratio D/P, at initialization we have

$$E[\dot{\lambda}_{max}(0)] = 0\,; \quad E[\ddot{\lambda}_{max}(0)]/E[\lambda_{max}(0)] = \sigma_z^2 \tag{23}$$

where E denotes the expectation over z, J, and Q at initialization.

Much like in the $D = 1$ case, Theorem 4.1 suggests that it is easy to find initializations that show progressive sharpening - and increasing σ_z makes sharpening more prominent.

4.2. Gradient descent dynamics

We now consider finite-step size gradient descent (GD) dynamics. The dynamics for are given by:

$$\theta_{t+1} = \theta_t - \eta J_t^\top z_t\,. \tag{24}$$

In this setting, the dynamical equations can be written as

$$z_{t+1} - z_t = -\eta J_t J_t^\top z_t + \frac{1}{2}\eta^2 Q(J_t^\top z_t, J_t^\top z_t) \tag{25}$$

$$J_{t+1} - J_t = -\eta Q(J_t^\top z_t, \cdot)\,. \tag{26}$$

If $Q = 0$, the dynamics reduce to discrete gradient descent in a quadratic potential - which converges iff $\lambda_{max} < 2/\eta$.

One immediate question is: when does the η^2 in Equation 25 affect the dynamics? Given that it scales with higher powers of η and z than the first term, we can conjecture that the ratio of the magnitudes of the terms, r_{NL}, is proportional to $\eta\|z\|_2$ and σ_J. A calculation in Appendix C.2 shows that, for random rotationally invariant initializations, we have:

$$r_{NL} \sim \frac{E[\|\frac{1}{2}\eta^2 Q(J_0^\top z_0, J_0^\top z_0)\|_2^2]}{E[\|\eta J_0 J_0^\top z_0\|_2^2]}^{1/2} = \frac{1}{2}\eta\sigma_z\sqrt{D}\,; \tag{27}$$

where the initialization statistics are defined as in Theorem 4.1. This confirms that increasing the learning rate and the residual magnitude $\|z\|$ increases the deviation of the dynamics from GF, and reveals that the amount of non-linearity is insensitive to σ_J.

We can see this phenomenology in the dynamics of the GD equations (Figure 3). Here we plot different trajectories for random initializations of the type in Theorem 4.1 with $D = 60$, $P = 120$, and $\eta = 1$. As σ_z increases, so does the curvature λ_{max} (as suggested by Theorem 4.1), and when σ_z is $O(1)$, the dynamics is non-linear (as predicted by r_{NL}) and EOS behavior emerges. This suggests that the second term in Equation 25 is crucial for the stabilization of λ_{max}.

We can confirm this more generally by initializing over various η, D, P, σ_z, and σ_J over multiple seeds, and plotting the resulting phase diagram of the final λ_{max} reached. We can simplify the plotting with some rescaling of parameters and initializations. In the rescaled variables

$$\tilde{z} = \eta z\,; \quad \tilde{J} = \eta^{1/2} J\,; \tag{28}$$

the dynamics are equivalent to Equations 25 and 26 with $\eta = 1$. As in the $\tilde{z} - T(0)$ model of Equations 56–57, λ_{max} in the rescaled coordinates is equivalent to λ_{max}/η in the unscaled coordinates. We can also define rescaled initializations for z and J. If we set

$$\tilde{\sigma}_z = \eta\sigma_z/\sqrt{D}\,; \quad \tilde{\sigma}_J = \eta\sigma_J/(DP)^{1/4}\,; \tag{29}$$

then we have $r_{NL} = \tilde{\sigma}_z$ which allows for easier comparison across $(D; P)$ pairs.

Figure 5. A FCN trained on CIFAR shows multiple cycles of sharpening and edge-of-stability behavior. The projection of the training set residuals $f(X;\theta) - Y$ onto the top NTK eigenmode v_1, increases in magnitude and oscillates around 0 (left). Plotting dynamics every two steps removes high frequency oscillations (middle). The largest eigenvalue crosses the edge of stability multiple times, but the second largest eigenvalue λ_2 remains below the edge of stability (right). Dynamics is similar for Hessian eigenvalues (dashed).

Using this initialization scheme, we can plot the final value of λ_{max} reached as a function of $\tilde{\alpha}_z$ and $\tilde{\alpha}_J$ for 100 independent random initializations for each $\tilde{\alpha}_z, \tilde{\alpha}_J$ pair (Figure 4). We see that the key is for $\alpha_{NL} = \tilde{\alpha}_z$ to be $O(1)$ - corresponding to both progressive sharpening and non-linear dynamics near initialization. In particular, initializations with small $\tilde{\alpha}_J$ values which converge at the EOS correspond to trajectories which first sharpen, and then settle near $\lambda_{max} = 2/\eta$. Large $\tilde{\alpha}_z$ and large $\tilde{\alpha}_J$ dynamics diverge. There is a small band of initial $\tilde{\alpha}_J$ over a wide range of $\tilde{\alpha}_z$ which have final $\lambda_{max} \approx 2/\eta$; these correspond to models initialized near the EOS, which stay near it.

This suggests that progressive sharpening and edge of stability aren't uniquely features of neural network models, and could be a more general property of learning in high-dimensional, non-linear models.

5. Connection to real world models

In this section we conduct numerical experiments in "real world" models, and compare the behavior to our theory on simplified models. Following (Cohen et al., 2022a), we trained a 2-hidden layer tanh network using the squared loss on 5000 examples from CIFAR10 with learning rate 10^{-2} - a setting which shows edge of stability behavior. Close to the onset of EOS, we approximately computed λ_1, the largest eigenvalue of $JJ^\top$, and its corresponding eigenvector v_1 using a Lanczos method (Ghorbani et al., 2019; Novak et al., 2019). We use v_1 to compute $z_1 = v_1^\top z$, where z is the vector of residuals $f(X;\theta) - Y$ for neural network function f, training inputs X, labels Y, and parameters θ. The EOS behavior in the NTK is similar to the EOS behavior defined with respect to the full Hessian in (Cohen et al., 2022a) (Figure 5, left and right). Once again, plotting the trajectories at every other step gets rid of the high frequency oscillations (Figure 5, middle). Unlike the $D = 1, P = 2$ model, there are multiple crossings of the critical line $\lambda_{max} = 2/\eta$ line.

There is evidence that low-dimensional features of a quadratic regression model could be used to explain some aspects of EOS behavior. We empirically compute the the second derivative of the output $f(x;\theta)$ by automatic differentiation. We denote by $Q(\theta;\theta)$ the resulting tensor. We can use matrix-vector products to compute the spectrum of the matrix $Q_1 \equiv v_1 \cdot Q(\theta;\theta)$, which is projection of the output of Q in the v_1 direction, without instantiating Q in memory (Figure 6, left). This figure reveals that the spectrum does not shift much from step 3200 to 3900 (the range of our plots). This suggests that Q doesn't change much as these EOS dynamics are displayed. We can also see that Q is much larger in the v_1 direction than a random direction.

Let y be defined as $y = \lambda_1 - 2/\eta$. Plotting the two-step dynamics of z_1 versus $2yz$ we see a remarkable agreement (Figure 6, middle). This is the same form that the dynamics of z takes in our simplified model. It can also be found by iterating Equation 25 twice with fixed Jacobian for $y = \lambda_1 - 2/\eta$ and discarding terms higher order in η. This suggests that during this particular EOS behavior, much like in our simplified model the dynamics of the eigenvalue is more important than any rotation in the eigenbasis.

The dynamics of y is more complicated. $y_{t+2} - y_t$ is anti-correlated with z_1^2 but there is no low-order functional form in terms of y and z_1 (Appendix D.2). We can get some insight into the stabilization by plotting the ratio of $\eta^2 Q_1(Jz_1v_1; Jz_1v_1)$ (the non-linear contribution to the z_1 dynamics from the v_1 direction) and $\eta\lambda_1 z_1$ (the linearized contribution), and compare it to the dynamics of y (Figure 6, right). The ratio is small during the initial sharpening, but becomes $O(1)$ shortly before the curvature decreases for the first time. It remains $O(1)$ through the rest of the dynamics. This suggests that the non-linear feedback from the dynamics of the top eigenmode onto itself is crucial to understanding the EOS dynamics.

Figure 6.Q is approximately constant during edge-of-stability dynamics for FCN trained on CIFAR10 (left). Projection onto largest eigendirectionv_1 (blue and orange) is larger than projection onto random direction (green). Two step difference $(z_1)_{t+2} - (z_1)_t$ is well approximated by$2z_1$y (middle), leading order term of models with xed eigenbasis. Non-linear dynamical contribution $^2Q_1(Jz_1v_1; Jz_1v_1)$ is small during sharpening, but becomes large immediately preceding decrease in top eigenvalue (right) - as is the case in the simple model.

6. Discussion

6.1. Lessons learned from quadratic regression models

The main lesson to be learned from the quadratic regression models is that behavior like progressive sharpening (for both GF and GD) and edge-of-stability behavior (for GD) may be common features of high-dimensional gradient-based training of non-linear models. Indeed, these phenomena can be revealed in simple settings without any connection to deep learning models: with mild tuning our simpli ed model, which corresponds to1datapoint and2 parameters can provably show EOS behavior. This combined with the analysis of the CIFAR model suggest that the general mechanism may have a low-dimensional description.

Quadratic approximations of real models quantitatively can capture the early features of EOS behavior (the initial return to $_{max} < 2=$), but do not necessarily capture the magnitude and period of subsequent oscillations – these require higher order terms (Appendix D.3). Nevertheless, the quadratic approximation does correctly describe much of the qualitative behavior, including the convergence of$_{max}$ to a limiting two-cycle that oscillates around$2=$, with an average valuebelow$2=$. In the simpli ed two-parameter model, it is possible to analytically predict the nal value at convergence, and indeed we nd that it deviates slightly from the value$2=$. In addition, both the theoretical and realistic models suggest that the curvature is controlled by a low-dimensional feedback mechanism mediated byz_1^2 mediated viaQ_1.

A key feature of all the models studied in this work is that looking at every-other iterate (the two-step dynamics) greatly aids in understanding the models theoretically and empirically. Near the edge of stability, this makes the changes in the top eigenmode small. In the simpli ed model, the slow$z$ dynamics (and related slow$V(0)$ dynamics) allowed for the detailed theoretical analysis; in the CIFAR model, the two-step dynamics is slowly varying in bothz_1 and $_{max}$. The quantitative comparisons of these small changes may help uncover any universal mechanisms/canonical forms that explain EOS behavior in other systems and scenarios.

6.2. Future work

One avenue for future work is to quantitatively understand progressive sharpening and EOS behavior in the quadratic regression model for large and$P$. In particular, computing the nal deviation$2 - _{max}$ in the edge-of-stability regime as a function of$_z$, $_J$, and$D=P$ remains an interesting open question. It would also be useful to understand how higher order terms affect the training dynamics - particularly the details of the oscillations around$= 2$.

Finally, our analysis has not touched on the feature learning aspects of the model. In the quadratic regression model, feature learning is encoded in the relationship between$J$ and$z$, and in particular the relationship between$z$ and the eigenstructure of$JJ^>$. Understanding how$Q$ mediates the dynamics of these two quantities may provide a quantitative basis for understanding feature learning which is complementary to existing theoretical approaches (Roberts et al., 2022; Bordelon & Pehlevan, 2022; Yang et al., 2022).

References

Bai, Y. and Lee, J. D. Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks. In*International Conference on Learning Representations*, March 2020.

Bordelon, B. and Pehlevan, C. Self-Consistent Dynamical Field Theory of Kernel Evolution in Wide Neural Networks, May 2022.

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.

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.

Giladi, N., Nacson, M. S., Hoffer, E., and Soudry, D. At Stability's Edge: How to Adjust Hyperparameters to Preserve Minima Selection in Asynchronous Training of Neural Networks? In *Eighth International Conference on Learning Representations*, April 2020.

Gilmer, J., Ghorbani, B., Garg, A., Kudugunta, S., Neyshabur, B., Cardoze, D., Dahl, G. E., Nado, Z., and Firat, O. A Loss Curvature Perspective on Training Instabilities of Deep Learning Models. In *International Conference on Learning Representations*, March 2022.

Huang, J. and Yau, H.-T. Dynamics of Deep Neural Networks and Neural Tangent Hierarchy. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 4542–4551. PMLR, November 2020.

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.

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.

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.

Li, Z., Wang, Z., and Li, J. Analyzing Sharpness along GD Trajectory: Progressive Sharpening and Edge of Stability, July 2022.

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.

Novak, R., Xiao, L., Hron, J., Lee, J., Alemi, A. A., Sohl-Dickstein, J., and Schoenholz, S. S. Neural Tangents: Fast and Easy Infinite Neural Networks in Python. *arXiv:1912.02803 [cs, stat]*, December 2019.

Roberts, D. A., Yaida, S., and Hanin, B. *The Principles of Deep Learning Theory*, May 2022. doi: 10.1017/9781009023405.

Wu, L., Ma, C., and E, W. How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Yang, G. Tensor Programs I: Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes. *arXiv:1910.12478 [cond-mat, physics:math-ph]*, May 2021.

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer, March 2022.

Zhu, L., Liu, C., Radhakrishnan, A., and Belkin, M. Quadratic models for understanding neural network dynamics, May 2022a.

Zhu, X., Wang, Z., Wang, X., Zhou, M., and Ge, R. Understanding Edge-of-Stability Training Dynamics with a Minimalist Example, October 2022b.

A. Connection to other models

A.1. Hessian versus NTK maximum eigenvalue

In this work we focus on EOS dynamics of the largest eigenvalue of the NTK, rather than the Hessian as in (Cohen et al., 2022a). We note that a version of Theorem 3.1 is true for the maximum Hessian eigenvalue as well. In general, the Hessian can be written as

$$\frac{\partial^2 L}{\partial \theta^2} = \nabla L \cdot \frac{\partial^2 z}{\partial \theta^2} + J^T \frac{\partial^2 L}{\partial z^2} J \tag{30}$$

For squared loss in particular, we have

$$\frac{\partial^2 L}{\partial \theta^2} = \nabla L \cdot \frac{\partial^2 z}{\partial \theta^2} + J^T J \tag{31}$$

As the loss gradient goes to 0, the Hessian eigenvalues approach the eigenvalues of JJ^T, whose non-zero eigenvalues are the same as those of the empirical NTK JJ^T. Since the theorem involves behavior as t goes to convergence, the maximum NTK and maximum Hessian eigenvalues are equal in the limit, and the same EOS behavior applied in both cases.

For the higher dimensional models (quadratic regression model and fully connected network on CIFAR10), our experiments show that the maximum NTK eigenvalue shows edge of stability behavior. The CIFAR model is the same as the one in (Cohen et al., 2022a) which was used to illustrate the edge of stability in terms of the maximum Hessian eigenvalues. Therefore we focused on the NTK version of EOS in our paper, as we found it more amenable to theoretical analysis and explanation.

There are almost certainly cases where EOS behavior is displayed in the Hessian eigenvalues but not the NTK eigenvalues, particularly in cases where the loss is highly non-isotropic in the outputs (that is, $\frac{\partial^2 L}{\partial z^2}$ is far from a multiple of the identity matrix). As pointed out in previous works in these cases even the Hessian-based EOS is more difficult to analyze (Cohen et al., 2022a). We leave understanding of EOS with more complicated loss functions for future work.

A.2. One-hidden layer linear network

Consider a one hidden layer network with a scalar output:

$$f(x) = v^\top U x \tag{32}$$

where x is an input vector of length N, U is a $K \times N$ dimensional matrix, and v is a K dimensional vector. We note that

$$\frac{\partial^2 f(x)}{\partial v_i \partial v_j} = \frac{\partial^2 f(x)}{\partial U_{ij} \partial U_{kl}} = 0; \quad \frac{\partial^2 f(x)}{\partial v_i \partial U_{jk}} = \delta_{ij} x_k \tag{33}$$

where δ_{ij} is the Kroenecker delta. For a fixed training set, this second derivative is constant; therefore, the one-hidden layer linear network is a quadratic regression model of the type studied in Section 4.

In the particular case of a single datapoint x, we can compute the eigenvectors of the Q matrix. Let $(w; W)$ be an eigenvector of Q, representing the v and U components respectively. The eigenvector equations are

$$\lambda w_i = x_m \delta_{ij} W_{jm} \tag{34}$$

$$\lambda W_{jm} = x_m \delta_{ij} w_i \tag{35}$$

Simplifying, we have:

$$\lambda w = Wx \tag{36}$$

$$\lambda W = wx^\top \tag{37}$$

We have two scenarios. The first is that $\lambda = 0$. In this case, we have $w = 0$, and W is a matrix with x in its nullspace. The latter condition gives us M constraints on MN equations - for a total of $M(N-1)$ of our $M(N+1)$ total eigenmodes.

If $\lambda \neq 0$, then combining the equations we have the conditions:

$$\lambda^2 w = (x \cdot x) w \tag{38}$$

$$! \ ^2 W = W x x^{>} \tag{39}$$

This gives us $! = \sqrt{x \ x}$. We know from Equation 37 that W is low rank. Therefore, we can guess a solution of the form

$$W_{;i} = e_i x^{>} \tag{40}$$

where the e_i are the M coordinate vectors. This suggests that we have

$$w_{;i} = (\sqrt{x \ x}) e_i \tag{41}$$

This gives us our $nal \ 2M$ eigenmodes.

We can analyze the initial values of of the $T(! _i)$ as well. The components of the Jacobian can be written as:

$$(J_v)_i \quad \frac{@f(x)}{@v_i} = U_{im} x_m \tag{42}$$

$$(J_U)_{jm} \quad \frac{@f(x)}{@U_{jm}} = v_j x_m \tag{43}$$

From this form, we can deduce that is orthogonal to the 0 modes. We can also compute the conserved quantity J_+^2 Let the the total weight in the positive eigenmodes, and J^2 be the total weight in the negative eigenmodes. A direct calculation shows that

$$! \ ^1(J_+^2 \quad J^2) = 2 f(x) \tag{44}$$

which implies that $E = 0$.

Therefore, the single-hidden layer linear model on one datapoint is equivalent to the quartic loss model $E = 0$ with and eigenvalues $\sqrt{x \ x}$.

A.3. Connection to (Bordelon & Pehlevan, 2022)

Since the one-hidden layer linear model has constant Q, the models in Section F.1 of (Bordelon & Pehlevan, 2022) fall into the quadratic regression class. In the case of Section F.1.1, Equation 67, we can make the mapping to our model explicit. The dynamics are equivalent to said model with a single eigenvalue $_0$ if we make the identi cations

$$= \tilde{z}; \ H_y = J_0^2; \ _0 = \sqrt{2!}; \ y = E=2 \tag{45}$$

A.4. Connection to NTH

The Neural Tangent Hierarchy (NTH) equations extend the NTK dynamics to account for changes in the tangent kernel by constructing an in nite sequence of higher order tensors which control the non-linear dynamics of learning (Huang & Yau, 2020). Truncation of the NTH equations $at \ 3rd$ order is related to, but not the same as the quadratic regression model, as we will show here.

The $3rd$ order NTH equation describes the change in the tangent kernel $JJ^{>}$. Consider the $D \quad D \quad D$-dimensional kernel K_3 whose elements are given by

$$(K_3) \quad = \frac{@^2 z}{@_i @_j} J_i \ J_j + \frac{@^2 z}{@_i @_j} J_i \ J_j \tag{46}$$

where repeated indices are summed over. In the NTH, for squared loss the change in the NTK is given by

$$\frac{d}{dt} \ JJ^{>} \quad = \quad (K_3) \quad z \tag{47}$$

For $xed \ Q = \frac{@^2 z}{@ @^{0}}$, this equation is identical to the GF equations for the NTK in the quadratic regression model. We note that K_3 is not constant under the quadratic regression model. Conversely, for $K_3 xed \frac{@^2 z}{@ @^{0}}$ is not constant either. Therefore, the two methods can be used to construct different low-order expansions of the dynamics.

B. 2 parameter model

B.1. Progressive sharpening under gradient flow

Given the rescaled variables

$$\tilde{z} = z; \quad T(k) = JQ^k J^\top : \tag{48}$$

we get the dynamical equations

$$\frac{d\tilde{z}}{dt} = \tilde{z}T(0); \qquad \frac{dT(k)}{dt} = 2\tilde{z}T(k+1) : \tag{49}$$

We can show that there are initializations which show progressive sharpening at late times - that is, increase in $T(0)$. Consider an initialization where the $T(k)$ are non-negative for even k, and non-positive for odd k. In this case, given the dynamical equations, the $T(k)$ don't change sign. If $\tilde{z} > 0$, then $T(0)$ is non-decreasing at all times.

One example of such an initialization is the case where the eigenvalues of Q are $\pm$, and the J are initialized with more weight in the negative eigenmodes than the positive ones. There are many other families of initializations with this property.

B.2. Derivation of function space dynamical equations

Consider the gradient descent equations:

$$_{t+1} \quad _t = \quad r \quad L = \frac{}{2} \quad ^\top Q \quad E \quad Q \quad : \tag{50}$$

We can derive the $\tilde{z}$ - T dynamics as follows. We can re-write the update equation as

$$_{t+1} \quad _t = \quad r \quad L = \quad z J^\top : \tag{51}$$

The change in z can be written as

$$z = \quad ^\top Q \quad + \frac{1}{2} \quad ^\top Q \tag{52}$$

Substituting, we have:

$$z = \quad z JJ^\top + \frac{^2}{2} z^2 JQJ^\top \tag{53}$$

We also have

$$J = \quad z JQ \tag{54}$$

We can also write:

$$(\quad JQ^k J^\top) = \quad 2 z JQ^{k+1} J^\top + \quad ^2 z^2 JQ^{k+2} J^\top \tag{55}$$

Converting to the $\tilde{z}$ - T coordinates, the gradient descent equations become:

$$\tilde{z}_{t+1} \quad \tilde{z}_t = \quad \tilde{z}_t T_t(0) + \frac{1}{2}(\tilde{z}_t^2) T_t(1) \tag{56}$$

$$T_{t+1}(k) \quad T_t(k) = \quad \tilde{z}_t (2T_t(k+1) \quad \tilde{z}_t T_t(k+2)) : \tag{57}$$

B.3. Derivation of $\tilde{z}$ - $T(0)$ equations

We can use the conserved quantity E to write the dynamics in terms of $\tilde{z}$ and $T(0)$ only. Without loss of generality, let the eigenvalues are 1 and $$, with 1 1. (We can achieve this by rescaling $\tilde{z}$.) Recall the dynamical equations

$$\tilde{z}_{t+1} \quad \tilde{z}_t = \quad \tilde{z}_t T_t(0) + \frac{1}{2}(\tilde{z}_t^2) T_t(1) \tag{58}$$

$$T_{t+1}(0) \quad T_t(0) = \quad \tilde{z}_t (2T_t(1) \quad \tilde{z}_t T_t(2)) \tag{59}$$

We will find substitutions for $T(1)$ and $T(2)$ in terms of $\tilde{z}$ and $T(0)$. Recall that we have

$$T(\quad 1) = E^- + 2\tilde{z} \tag{60}$$

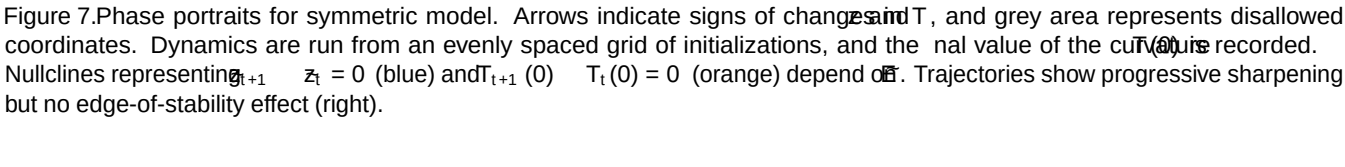

Figure 7. Phase portraits for symmetric model. Arrows indicate signs of changes in $\bar{z}$ and T, and grey area represents disallowed coordinates. Dynamics are run from an evenly spaced grid of initializations, and the final value of the curvature $T(0)$ is recorded. Nullclines representing $\bar{z}_{t+1} - \bar{z}_t = 0$ (blue) and $T_{t+1}(0) - T_t(0) = 0$ (orange) depend on $\bar{E}$. Trajectories show progressive sharpening but no edge-of-stability effect (right).

where $\bar{E}$ is conserved throughout the dynamics (and indeed is a property of the landscape). We will use this definition to solve for $T(1)$ and $T(2)$.

Since $P = 2$, we can write $T(-1) = bT(0) + aT(1)$, for coefficients a and b which are valid for all combinations of $\bar{E}$. If $J(-) = 0$, we have $b = 1 - a$. If $J(1) = 0$, we have $1 = (1 - a) + a^2 a$. Solving, we have:

$$T(-1) = (1 - a)T(0) + aT(1) \text{ for } a = \frac{1}{-} \tag{61}$$

The restrictions on $-$ translate to $a \in (-1, 1)$. In terms of the conserved quantity $\bar{E} = T(-1) - 2\bar{z}$, we have:

$$T(-1) = \bar{E} + 2\bar{z} \tag{62}$$

In order to convert the dynamics, we need to solve for $T(1)$ and $T(2)$ in terms of $T(0)$ and $\bar{z}$. We have:

$$T(1) = \frac{1}{a}(T(-1) + (a - 1)T(0)) = \frac{1}{a}\left(\bar{E} + 2\bar{z} + (a - 1)T(0)\right) \tag{63}$$

We also have

$$T(2) = T(0) + \frac{1 - a}{a^2}\left(T(0) - \bar{E} - 2\bar{z}\right) \tag{64}$$

This gives us

$$\bar{z}_{t+1} - \bar{z}_t = -\bar{z}_t T_t(0) + \frac{1}{2a}(\bar{z}_t^2)((a - 1)T_t(0) + 2\bar{z}_t + \bar{E}) \tag{65}$$

$$T_{t+1}(0) - T_t(0) = \frac{2}{a}\bar{z}_t(2\bar{z}_t + \bar{E} + (a - 1)T_t(0)) + \bar{z}_t^2 - T_t(0) + \frac{1 - a}{a^2}\left(T_t(0) - \bar{E} - 2\bar{z}_t\right) \tag{66}$$

If $- = -$ (that is, $a = -1$) we recover the equations from the main text.

The non-negativity of T^{-2} gives us constraints on the values of $\bar{z}$ and T. For $a > 1$ (small negative second eigenvalue), the constraints are:

$$T > 2\bar{z} + \bar{E}; \; T > -(2\bar{z} + \bar{E})/a \tag{67}$$

This is an upward-facing cone with vertex at $-\bar{E}/2$ (Figure 8, left). For $a < -1$, the constraints are

$$-(2\bar{z} + \bar{E})/a < T < 2\bar{z} + \bar{E} \tag{68}$$

This is a sideways facing cone with vertex at $-\bar{E}/2$ (Figure 8, right). We see that in this case, there is a limited set of values of T to converge to. Indeed, for $\bar{E} = 0$, there is no convergence except $T(0) = 0$. This why we focus on the case of one positive and one negative eigenvalue.

We can also solve for the nullclines - the curves where either $\bar{z}_{t+1} - \bar{z}_t = 0$ (blue in Figure 8), or $T_{t+1}(0) - T_t(0) = 0$ (orange in Figure 8). The nullcline $(\bar{z}, f_{\bar{z}}(\bar{z}))$ for $\bar{z}$ is given by

$$f_{\bar{z}}(\bar{z}) = \frac{\bar{z}(2\bar{z} + E^2)}{2a - (a-1)\bar{z}} \tag{69}$$

The nullcline $(\bar{z}, f_T(\bar{z}))$ for $T(0)$ is given by

$$f_T(\bar{z}) = \frac{(a-1)\bar{z} - 2a}{(a^2 - a + 1)\bar{z} - 2a(a-1)}(2\bar{z} + E^2) \tag{70}$$

The line $\bar{z} = 0$ is also a nullcline.

For the symmetric model $= 1$, the structure of the nullclines determines the presence or lack of progressive sharpening. For $E^2 = 0$, there is no sharpening; the phase portrait (Figure 7, left) con rms this as the nullcline $T_{t+1}(0) - T_t(0)$ divides the space into two halves, one which converges, and the other which doesn't. However, even when $E^2 > 0$, the nullclines split, and there is a small region where progressive sharpening can occur (Figure 7, middle). However, there is still no edge-of-stability behavior in this case - there is no region where the trajectories cluster near $= 2/\lambda_{max} =$ (Figure 7, right).

Figure 8. Phase planes of $= 1$, $P = 2$ model. Grey region corresponds to parameters forbidden by positivity constraint $T(1+\eta)^2$. For > 0, allowed region is smaller and intersects $= 0$ at a small range only. Nullclines can be solved for analytically.

B.4. Two-step dynamics

The two-step difference equations can be derived by iterating Equations 14 and 15. We have

$$\bar{z}_{t+2} - \bar{z}_t = p_0(\bar{z}_t,) + p_1(\bar{z}_t,)T_t(0) + p_2(\bar{z}_t,)T_t(0)^2 + p_3(\bar{z}_t,)T_t(0)^3 \tag{71}$$

$$T(0)_{t+2} - T_t(0) = q_0(\bar{z}_t,) + q_1(\bar{z}_t,)T_t(0) + q_2(\bar{z}_t,)T_t(0)^2 + q_3(\bar{z}_t,)T_t(0)^3 \tag{72}$$

Here the p_i and q_i are polynomials in $\bar{z}$, maximum 9th order in $\bar{z}$ and 6th order in . They can be computed explicitly but we choose to omit the exact forms for now.

Numerical simulation of the dynamics for small reveals an edge of stability effect (Figure 9). We see that the distribution of nal values of T for random initializations has a peak near $T(0) = 2$ (right). By plotting the two-step dynamics, we can see that the two-step nullclines which go into $T(0) = 2$ almost coincide (left). By studying these nullclines, we will be able to understand the edge of stability effect.

For xed , we can solve for the two-step nullclines $(\bar{z}_{t+2} - \bar{z}_t = 0)$ and the T nullclines $(T_{t+2}(0) - T_t(0) = 0)$ using Cardano's formula to solve for T as a function of $\bar{z}$. In particular, each nullcline equation has a solution that goes through $\bar{z} = 0$, $T(0) = 2$, independent of . This is the family of solutions that we will focus on.

Let $(\bar{z}, f_{\bar{z},})$ be the nullcline of $\bar{z}$, and let $(\bar{z}, f_{T,})$ be the nullcline of $T(0)$. We will show that the T values of the nullclines, as a function of $\bar{z}$ and , is differentiable around $\bar{z} = 0$, $= 0$.