Second-order Regression Models Exhibit Progressive Sharpening to the Edge of Stability

Atish Agarwala, Fabian Pedregosa, Jeffrey Pennington

Google Research

Progressive Sharpening

Prior work (Cohen et al. 2021) has shown tendency of many deep architectures to

1. Sharpness (largest Hessian eigenvalue) increases throughout optimisation
2. Eventually sharpness hovers around 2 / step-size

Fully-connected net on CIFAR-10 5k subset

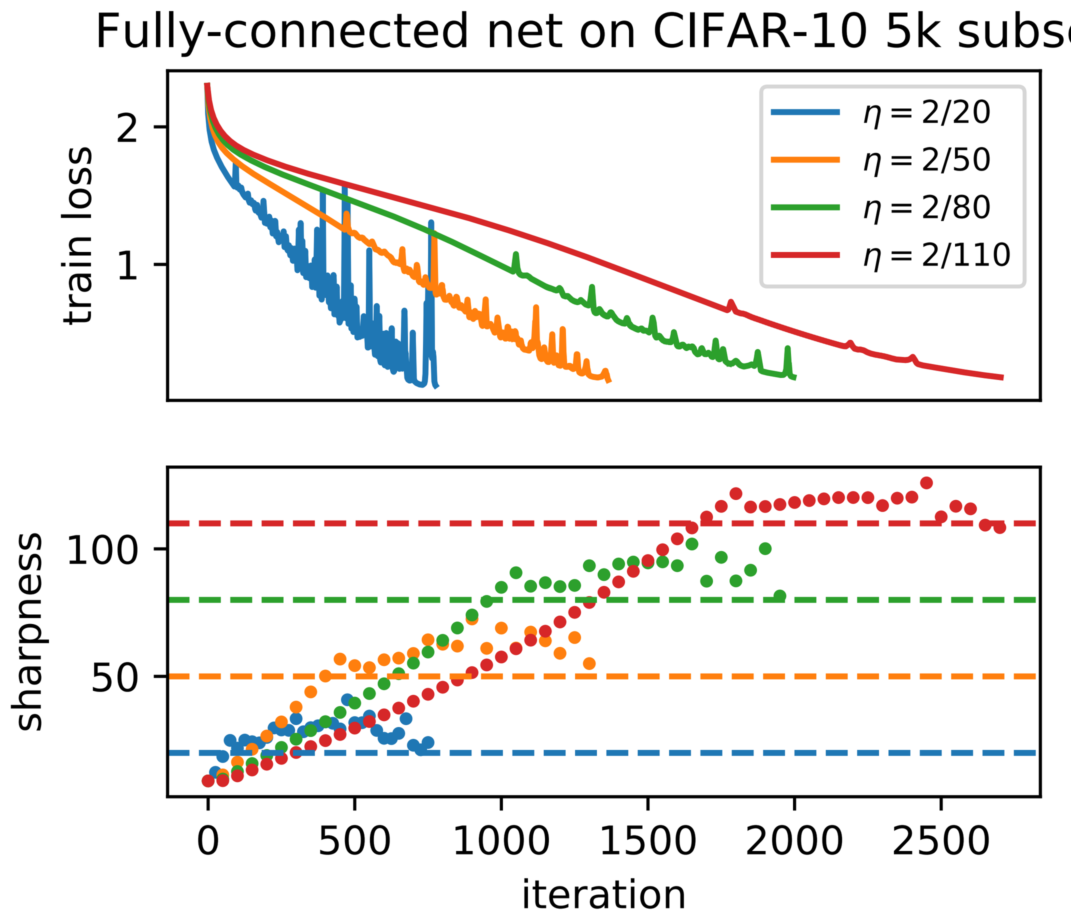


Figure Source: (Cohen et al. 2021)

Implications for Optimization

1. No global bound on *L*-smoothness (aka sharpness), depends on step-size.
2. Quadratic objectives don't exhibit these dynamic = not a good model
3. What is driving the optimization to not diverge?

A Second-order Model

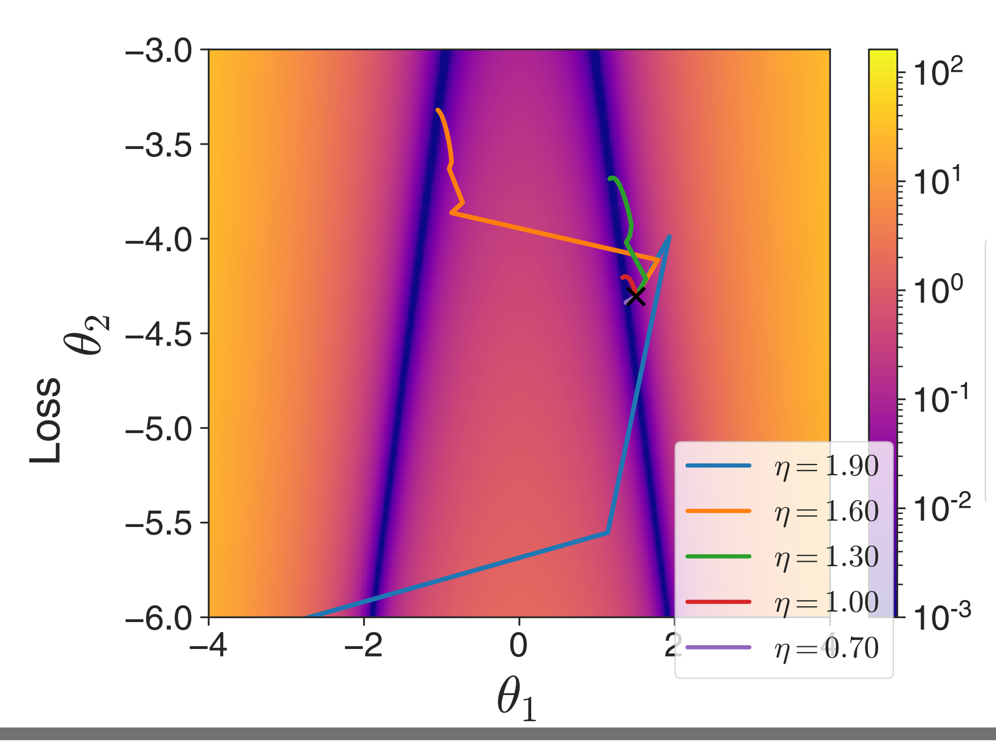
We propose a tractable model that exhibits progressive sharpening. In its simplest form

$$\mathcal{L}(\theta) = (f(\theta) - E)^2$$

where $f(\theta) = \theta^\top Q \theta$, $\theta \in \mathbb{R}^2$

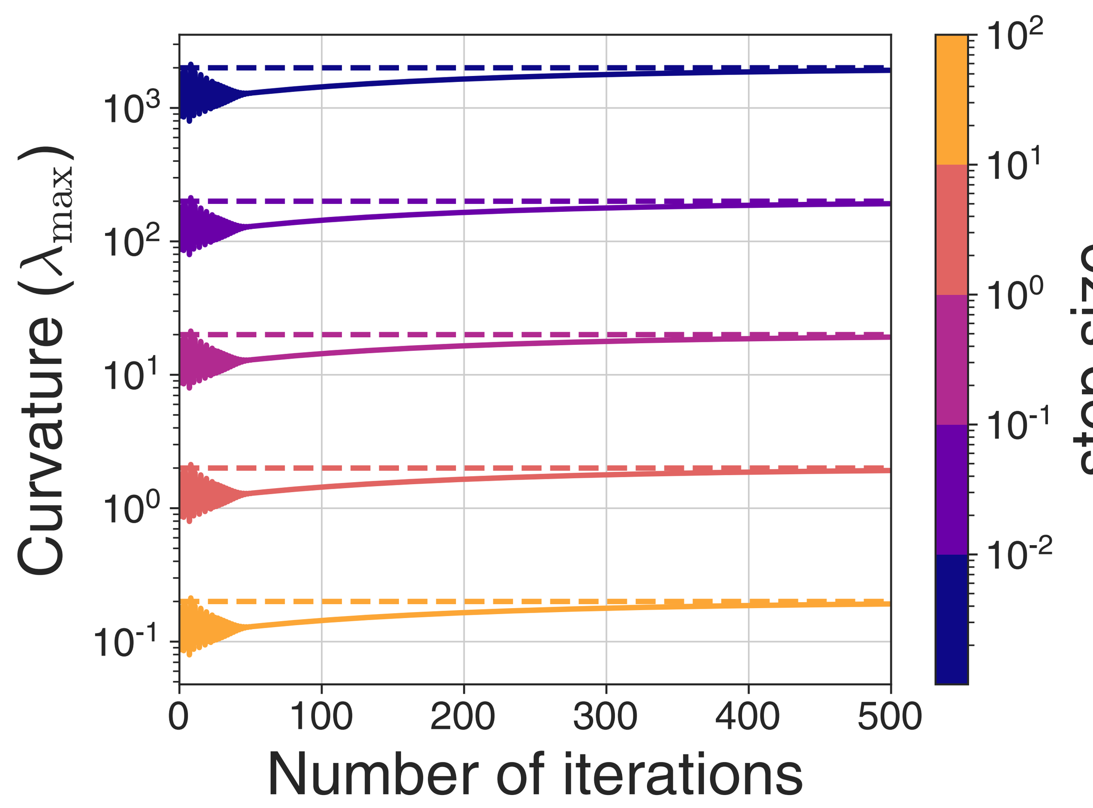
- Can be seen as quadratic regression with a quadratic predictive model, one datapoint
- Unlike NTK, f is quadratic in θ

This objective has multiple solutions with different degrees of sharpness



Quadratic Model Exhibits Progressive Sharpening

When $\text{eigenvalues}(Q) = \{1, -\varepsilon\}$, with $0 < \varepsilon \ll 1$ then we observe progressive sharpening



Main Result

Theorem 1 *There exists an $\epsilon_c > 0$ such that for a quadratic regression model with $E = 0$ and eigenvalues $\{-\epsilon, 1\}$, $\epsilon \leq \epsilon_c$. there exists a neighborhood $U \subset \mathbb{R}^2$ and interval $[\eta_1, \eta_2]$ such that for initial $\boldsymbol{\theta} \in U$ and learning rate $\eta \in [\eta_1, \eta_2]$, the model displays edge-of-stability behavior:*

$$2/\eta - \delta_\lambda \leq \lim_{t\to\infty} \lambda_{\max} \leq 2/\eta\,,$$

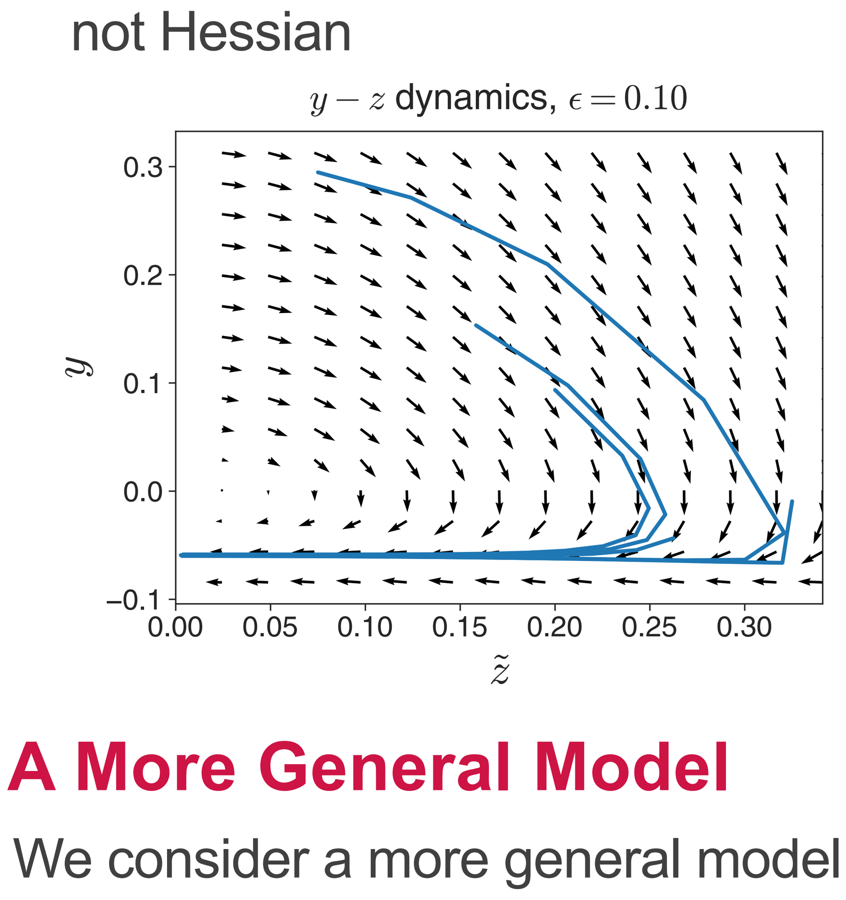
for δ_λ of $O(\epsilon)$.

Analysis — key insights

1. Derive recurrence for output $z_t \overset{\text{def}}{=} f(\theta_t)$ Instead of parameters.
2. Write recurrence for every other iteration — removes oscillations.

$$\tilde{z}_{t+2} - \tilde{z}_t = 2y_t\tilde{z}_t + O(y_t^2\tilde{z}_t) + O(y_t\tilde{z}_t^2)$$
$$y_{t+2} - y_t = -2(4 - 3\epsilon + 4\epsilon^2)y_t\tilde{z}_t^2 - 4\epsilon\tilde{z}_t^2 + \epsilon O(\tilde{z}_t^3)\,.$$

3. Study dynamical system as $z_t \to 0$, i.e., as model converges to solution.
4. Analysis is performed on empirical NTK, not Hessian



A More General Model

We consider a more general model, where.

$f : \mathbb{R}^d \to \mathbb{R}^n$ and

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

General Model Exhibits Progressive Sharpening

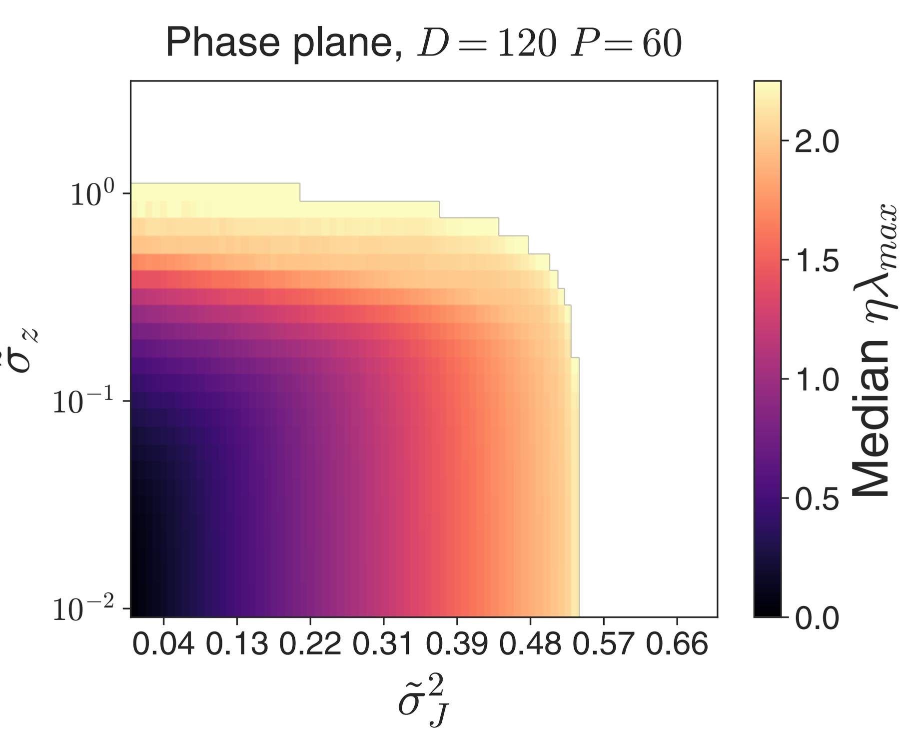
Provably at initialization:

Theorem 2 *Let $\mathbf{z}$, $\mathbf{J}$, and $\mathbf{Q}$ 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 $\lambda_{\max}$ be the largest eigenvalue of $\mathbf{J}\mathbf{J}^\top$. In the limit of large D and P, with fixed ratio D/P, at initialization we have*

$$\mathrm{E}[\dot{\lambda}_{\max}(0)] = 0,\ \mathrm{E}[\ddot{\lambda}_{\max}(0)]/\mathrm{E}[\lambda_{\max}(0)] = \sigma_z^2 \quad (22)$$

where E denotes the expectation over $\mathbf{z}$, $\mathbf{J}$, and $\mathbf{Q}$ at initialization.

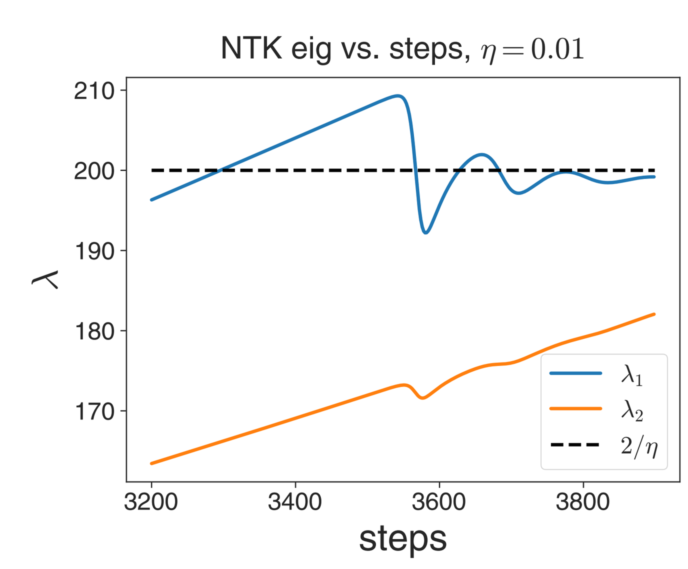
Empirically upon convergence:



Sharpness at convergence is close, but not exactly 2/step-size

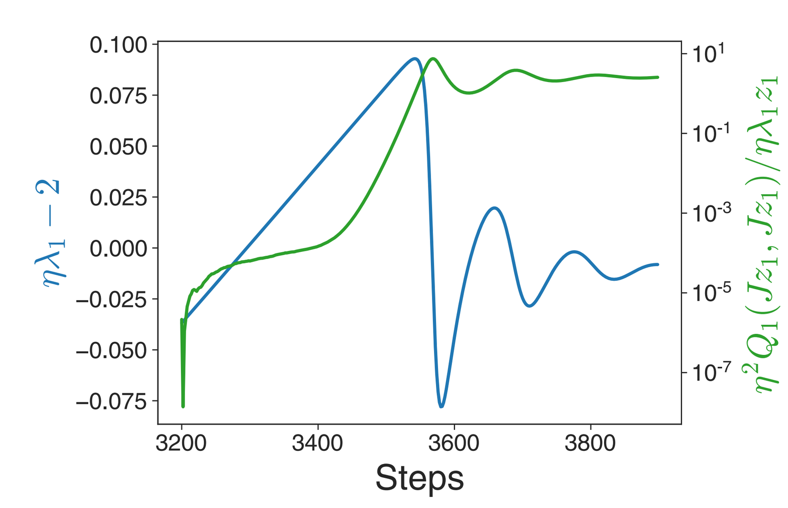
Connection with Deep Models

We trained 2-hidden-layer tank network on CIFAR10



Top: Largest eigenvalue crosses edge of stability multiple times, second largest remains below.

Below: Non-linear dynamical contribution (green) is small during sharpening but becomes large preceding decrease in top eigenvalue



Bibliography

Cohen, Jeremy M., et al. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021.

Lewkowycz, Aitor, et al. "The large learning rate phase of deep learning: the catapult mechanism." arXiv:2003.02218 (2020).

Arora, Sanjeev, et al.. "Understanding Gradient Descent on Edge of Stability in Deep Learning." arXiv:2205.09745 (2022).

Damian, Alex, et al."Self-Stabilization: The Implicit Bias of Gradient Descent at the Edge of Stability." arXiv arXiv:2209.15594 (2022).