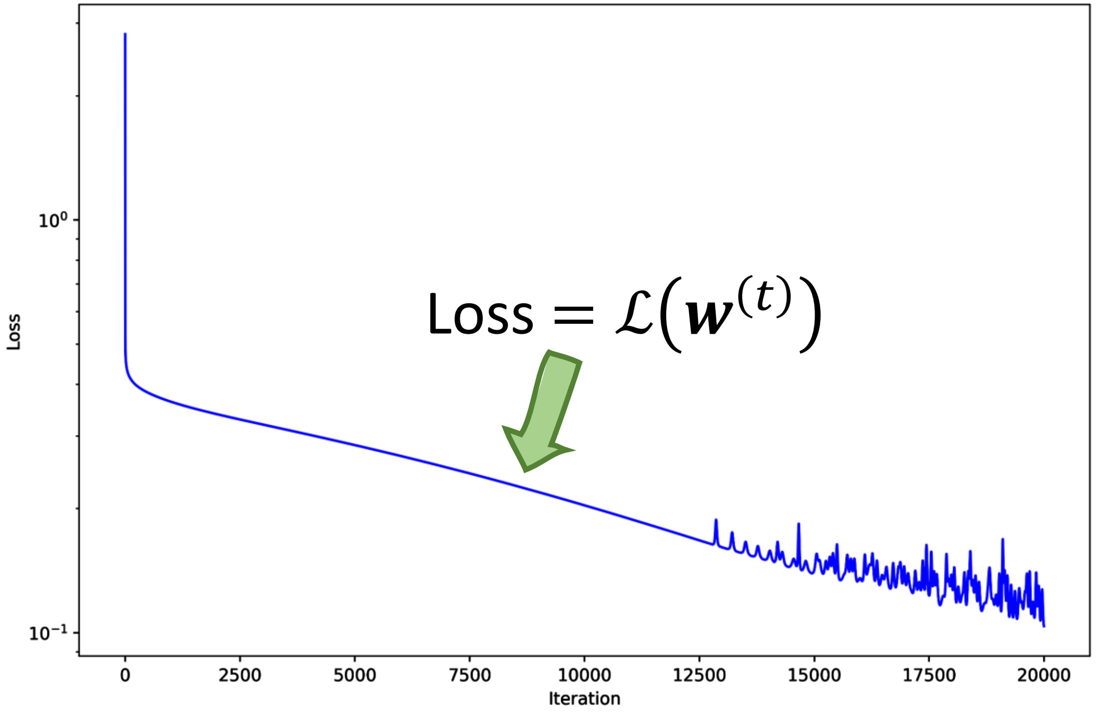
Gradient Descent Monotonically Decreases the Sharpness of Gradient Flow Solutions in Scalar Networks and Beyond

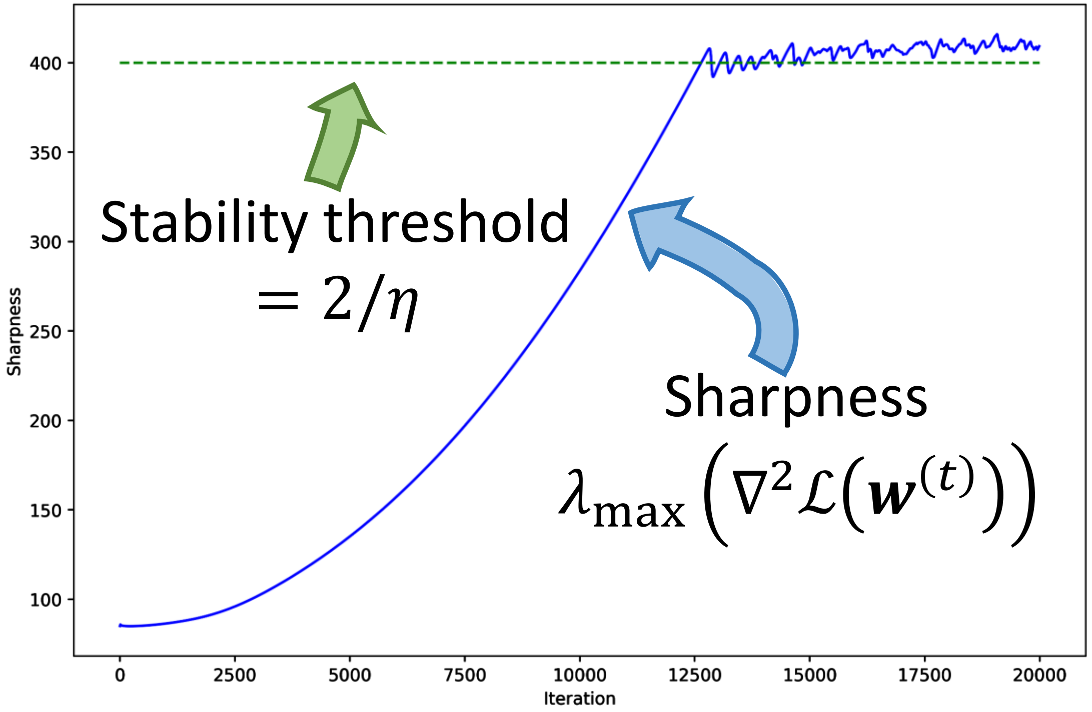
Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, Yair Carmon

Tel-Aviv University ,Technion

ICML
International Conference On Machine Learning

Motivation: the edge of stability phenomenon





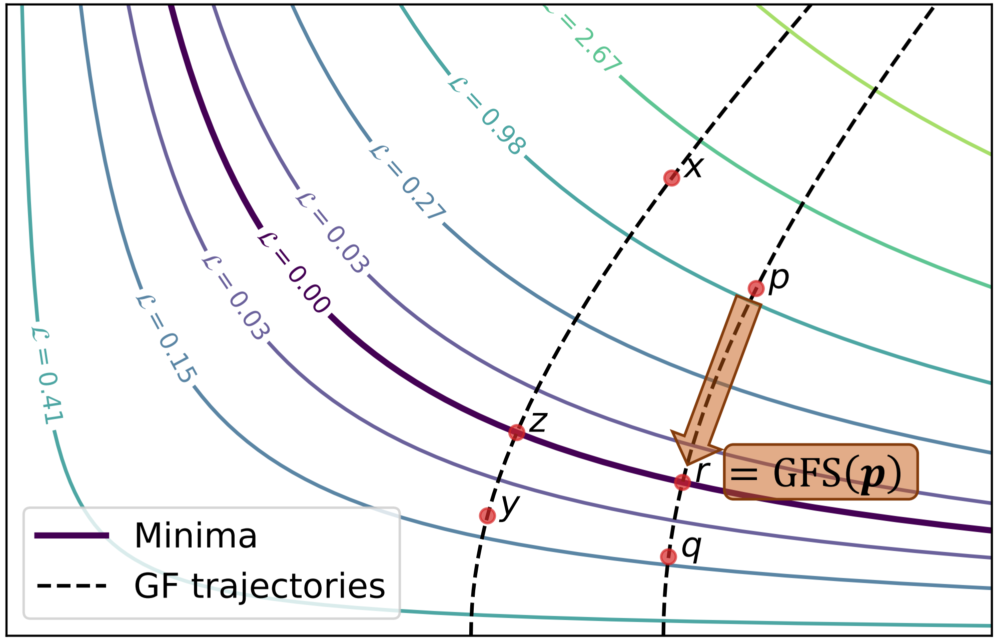
Cohen et al. (2021)

Gradient descent on neural networks typically occurs at the edge of stability

Explaining EoS convergence via gradient flow solution sharpness

$\mathrm{GFS}(\boldsymbol{w})$ = the end of the gradient flow trajectory starting from $\boldsymbol{w}$

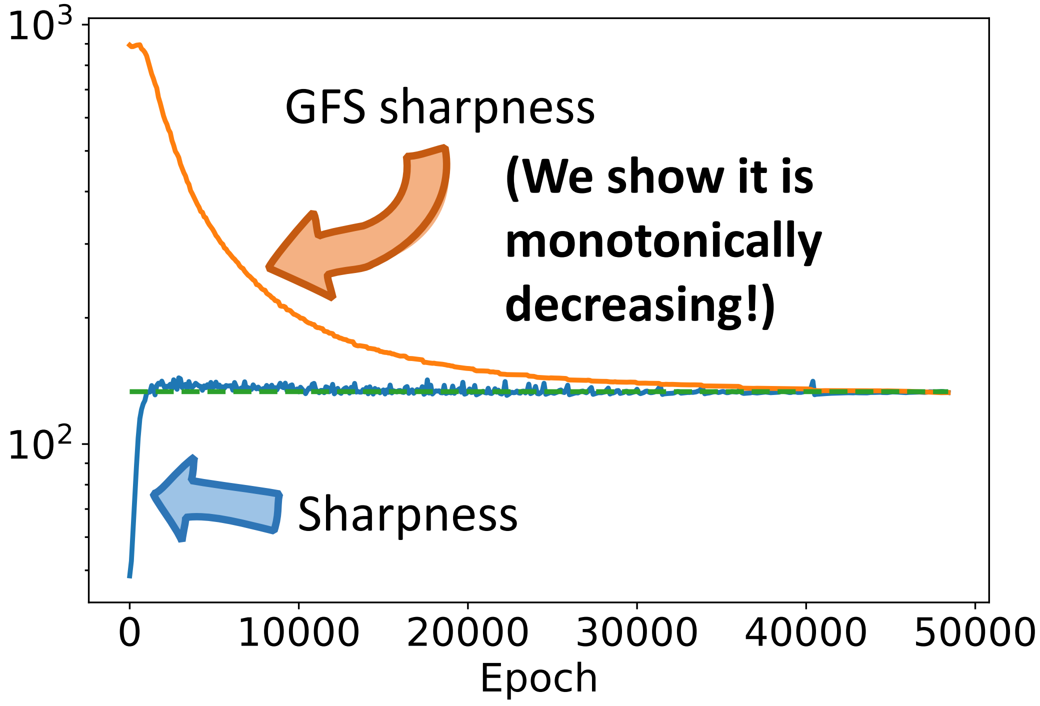
$\phi(\boldsymbol{w})$ = the GFS sharpness of $\boldsymbol{w}$ = $\lambda_{\max}\left(\nabla^2 \mathcal{L}(\mathrm{GFS}(\boldsymbol{w}))\right)$



Explaining EoS convergence via gradient flow solution sharpness

$$\mathcal{L}(\boldsymbol{w}) \approx 0 \implies \phi(\boldsymbol{w}) \approx \lambda_{\max}\left(\nabla^2 \mathcal{L}(\boldsymbol{w})\right)$$

Understanding $\lim_{t\to\infty} \phi\left(w^{(t)}\right)$ **allow us to understand EoS convergence**



Theory for scalar networks

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}(w_1 w_2 w_3 \cdots w_D - 1)^2$$

Theorem (*Informal*)**:** Under a weak assumption on the initialization $\boldsymbol{w}^{(0)}$ then for all t≥ 0:

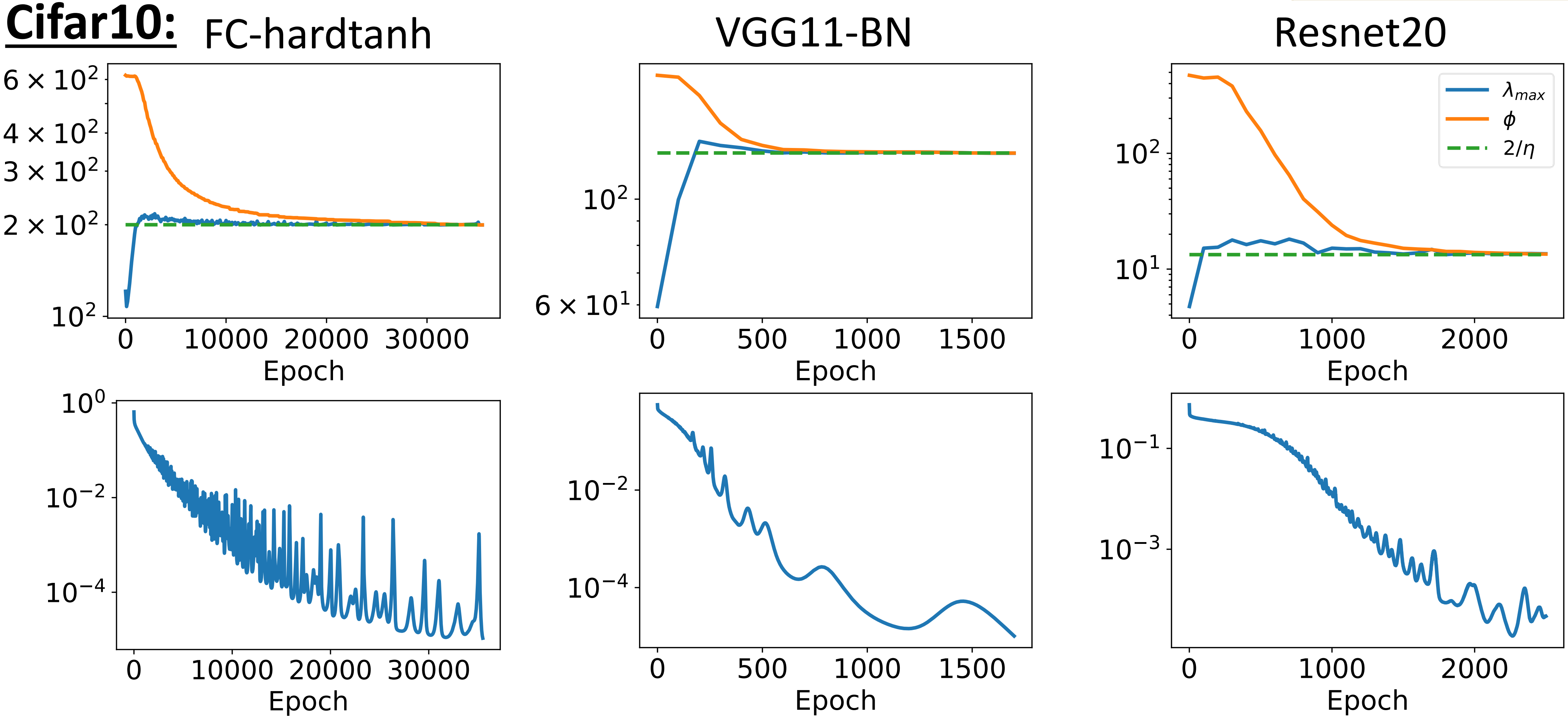
- The assumption hold for $\boldsymbol{w}^{(t)}$
- $\phi\left(\boldsymbol{w}^{(t+1)}\right) \leq \phi\left(\boldsymbol{w}^{(t)}\right)$

Theorem (*Informal*)**:** If for some $t \geq 0$ and $\delta \in (0,0.4)$, the assumption hold for $\boldsymbol{w}^{(t)}$, $\phi\left(\boldsymbol{w}^{(t)}\right) = \frac{2-\delta}{\eta}$ and $\mathcal{L}\left(\boldsymbol{w}^{(t)}\right) = \mathcal{O}(\delta^2)$ then:

- $\lim_{k\to\infty} \phi\left(\boldsymbol{w}^{(k)}\right) \geq \frac{2(1-\delta)}{\eta}$
- The loss converges exponentially to 0

Experiments: neural networks

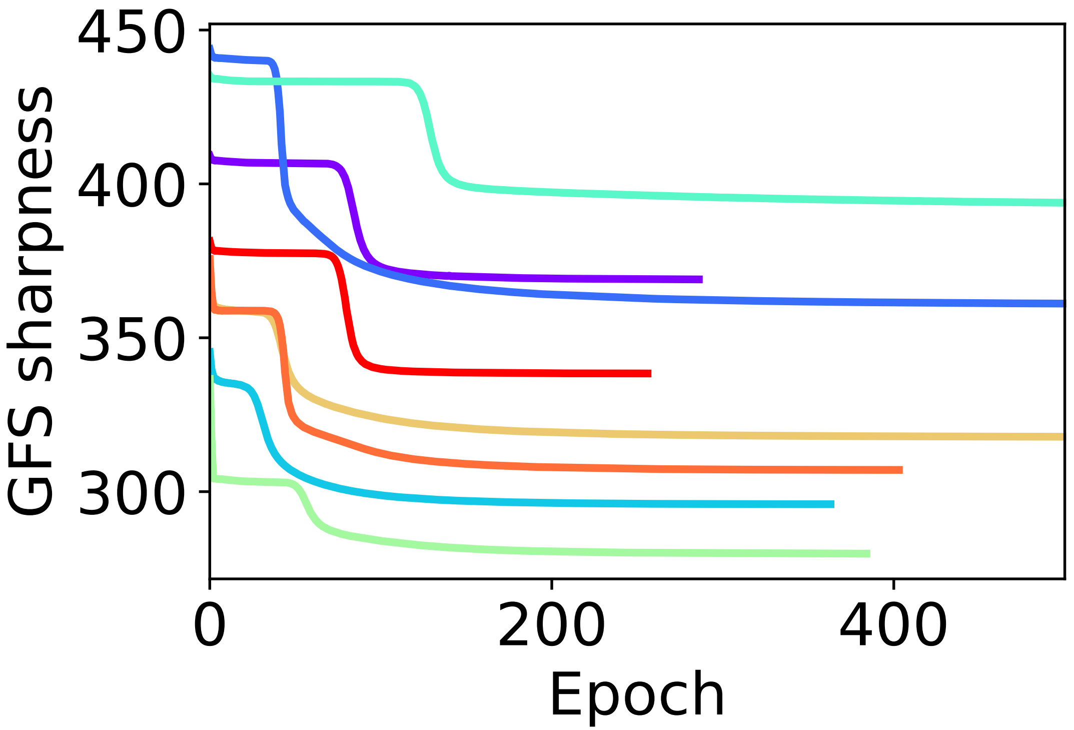
Cifar10:

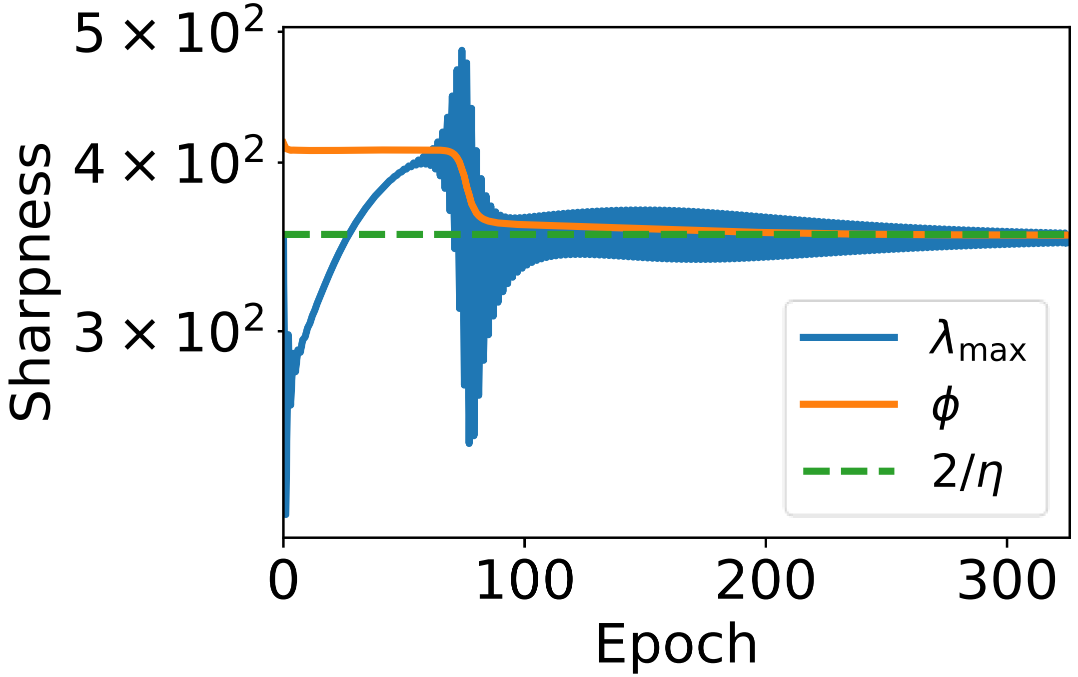


Experiments: squared regression model

$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \boldsymbol{u}_+^2 - \boldsymbol{u}_-^2, \boldsymbol{x} \rangle$

The GFS can be efficiently calculated





Thanks!