Momentum *Further Constrains Sharpness at the* **Edge of** *Stochastic* **Stability**

Arseniy Andreyev*, Advikar Ananthkumar*, Marc Walden*, Tomaso Poggio, Pierfrancesco Beneventano

Preliminaries about optimization

- Most arguments in optimization rely on the fact that training ***is stable***.

Descent Lemma:

$$L_{t+1} - L_t \leq -\eta(1 - \eta\lambda_{\max}) \cdot \|\nabla L\|^2 \quad \text{if the top Hessian eigenvalue } \lambda_{\max} < 2/\eta$$

Not for neural networks!

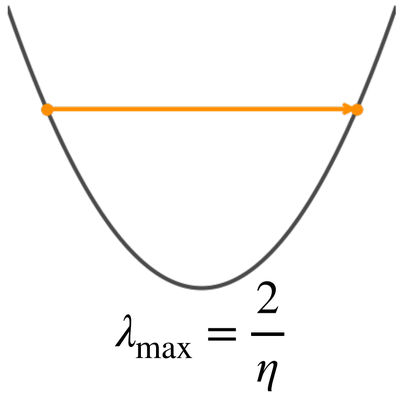
η the step size or learning rate

- Examples are: *ending up in stationary points, speed of convergence, ...*
- ***The stability assumption*** is not an issue in general: *you can make the step size smaller* and then everything works.

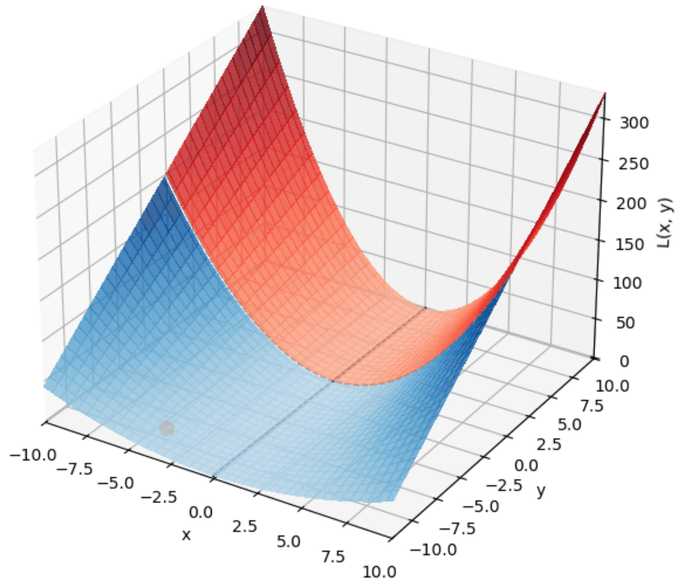
What is Edge of Stability

- *Cohen et al. (2021)*++ observed that for neural networks:
    1. Curvature grows.
    2. Stabilizes at the instability threshold.

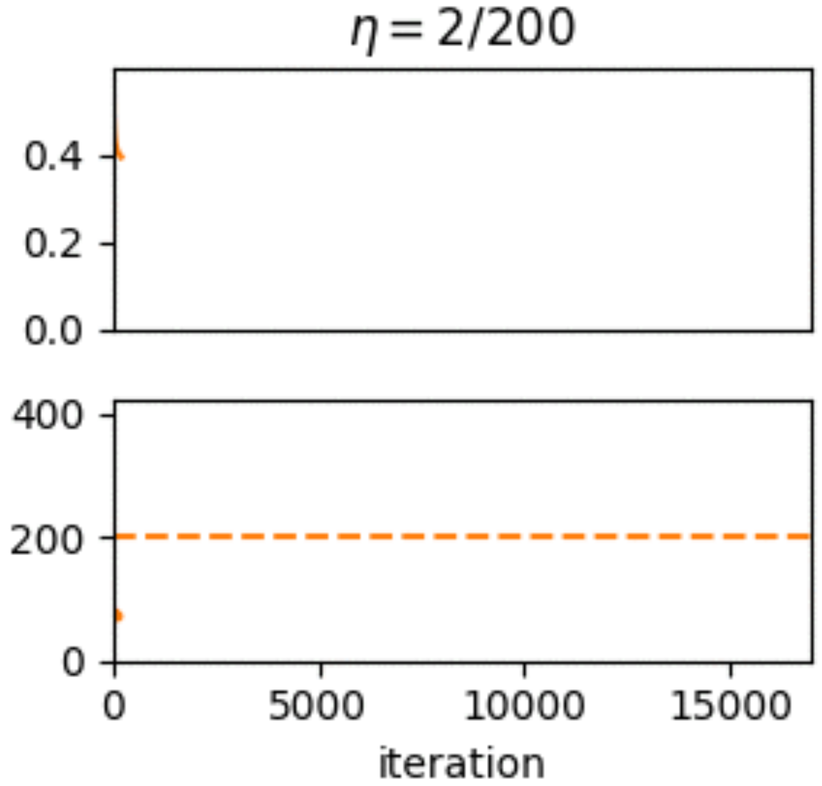
IN-BETWEEN: $|1 - \eta \cdot \lambda_{\max}| = 1$



$$\lambda_{\max} = \frac{2}{\eta}$$



DAMIAN ET AL. (2022), ARXIV: HTTPS://ARXIV.ORG/PDF/2209.15594.



COHEN ET AL. (2021), ARXIV: HTTPS://ARXIV.ORG/PDF/2103.00065.

Open problems

- All of this was for full-batch!
- Real training is:
  1. mini-batch
  2. with momentum, β
  3. adaptive

	full-batch	**mini-batch**
vanilla	Cohen et al., 2021: $\lambda_{\max} \sim 2/\eta$	A&B (2024) Batch Sharpness $\sim 2/\eta$
momentum	Cohen et al., 2021: *momentum allows higher sharpness* $\lambda_{\max} \sim 2/\eta \cdot (1+\beta)$	**this work**
adaptive + momentum	Cohen et al., 2022 $\lambda_{\max,P} \sim 2/\eta \cdot \frac{1+\beta}{1-\beta}$	**preliminary observations**

Batch Sharpness = “directional curvature of mini-batch landscape”

Open problems

Do optimizers with momentum train in a regime of instability?

How does the threshold shift?

Why do we care

Do they also for practical training procedures?

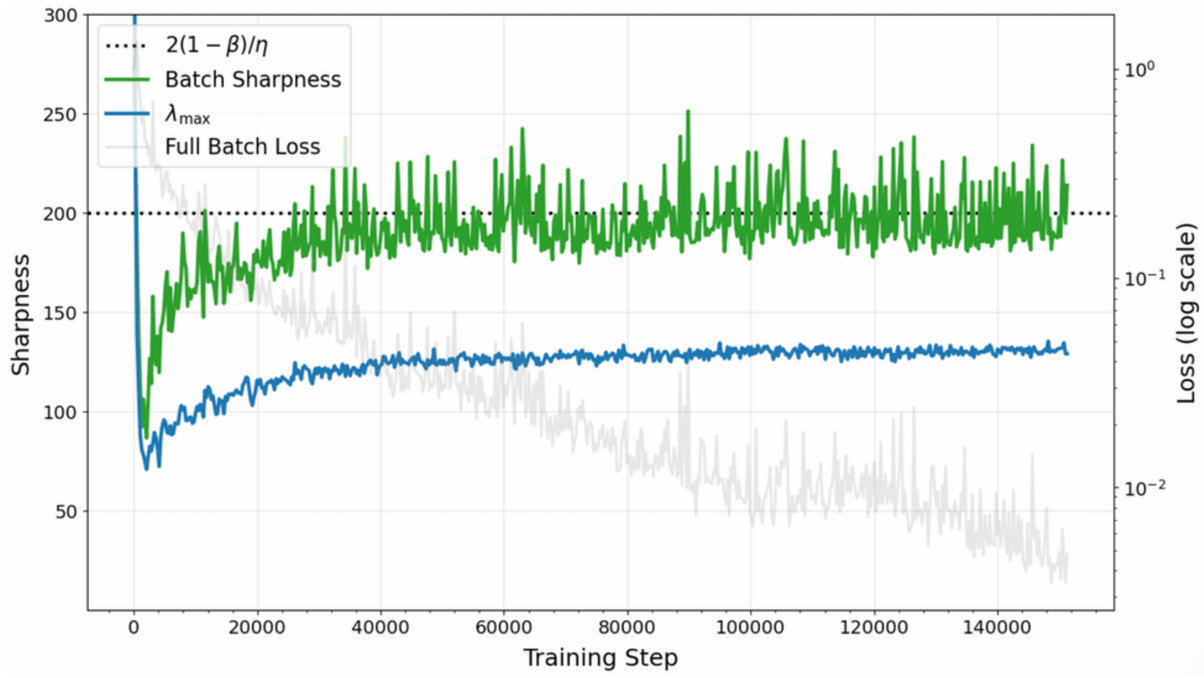
- **Classical arguments break.**
- At the EoS training **happens differently** (*Kwag et al. 2026*).
- This is a **mechanism for** SGD inducing flatness.

Finally a mechanism for *Keskar et al. (2016)*

- Now we know why and we can forecast how the solution depends on *hyperparameters*: **in the stable setting it does not**.

Results

- With addition of momentum:
  - Still have EoS-like regime of instability
  - Batch Sharpness is still the indicative quantity

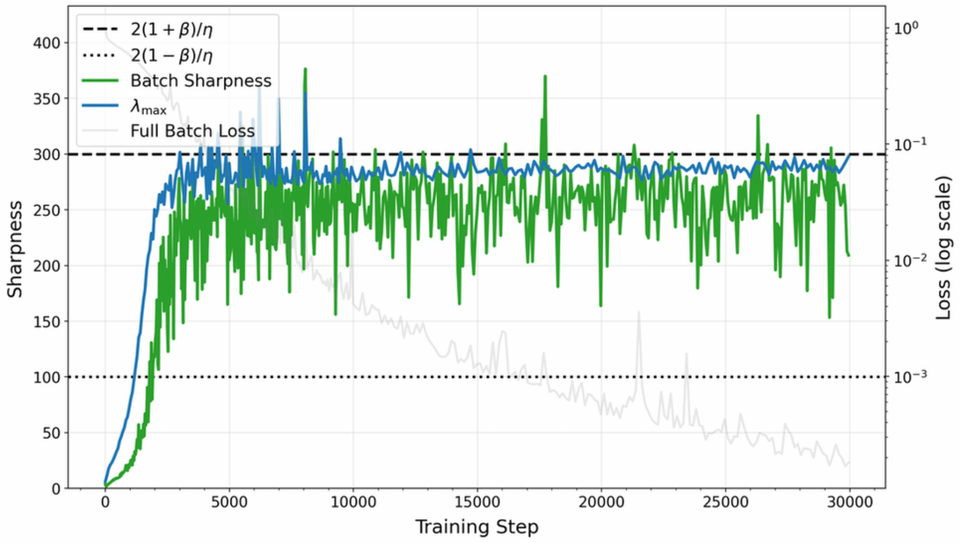


Results

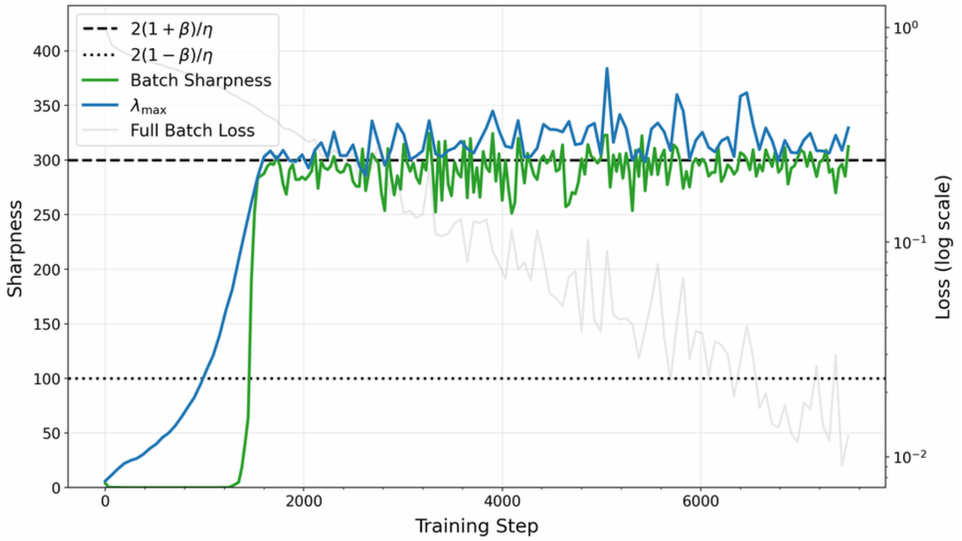
- Effect depends on batch size:
  - Large batch size (as Cohen et al., 2021):
    - As in full-batch regime
    - Momentum induces for higher sharpness:

$\lambda_{\max}$, Batch Sharpness $\sim 2/\eta \cdot (1+\beta)$

batch size = 256



full batch



Results

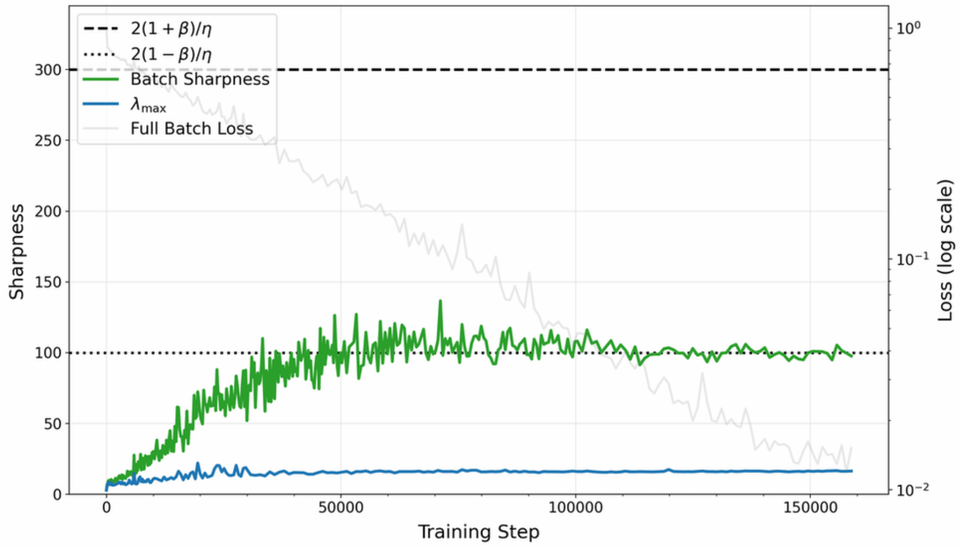
- Effect depends on batch size:
  - Small batch size:

    Sharpness threshold is **lower**:

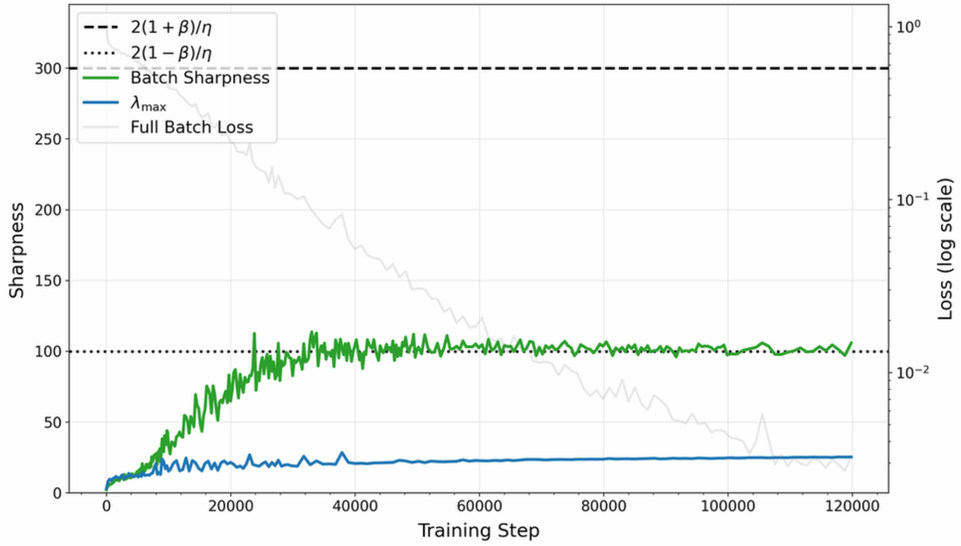
    $$\text{Batch Sharpness} \sim 2/\eta \cdot (1-\beta)$$

- Theoretically, we show this through linear stability analysis

batch size = 4

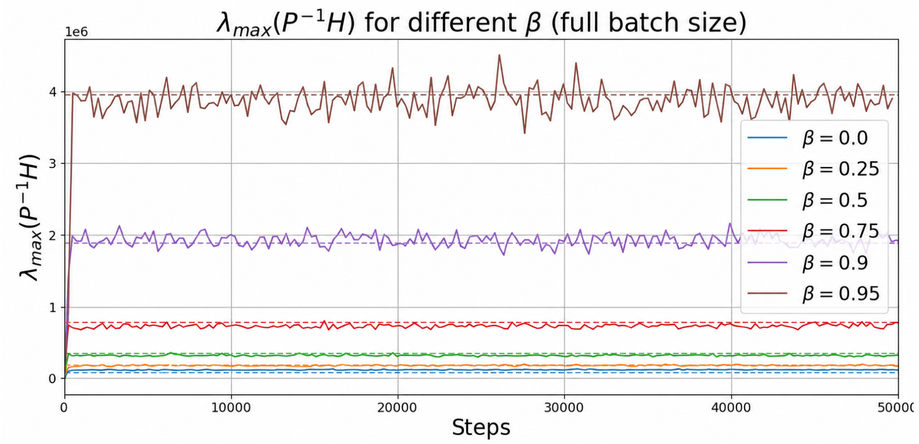


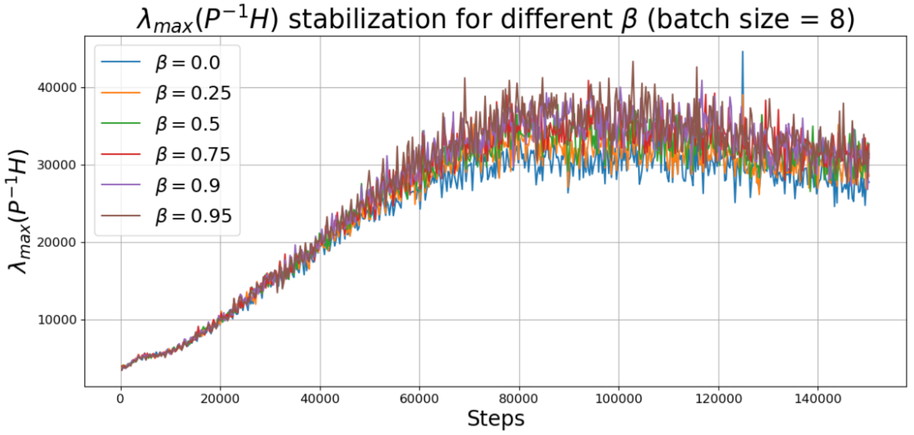
batch size = 8



Future work

- Understanding effect of momentum for other optimizers
  - Preliminary, for Adam:





Thanks