Coffee Talk #5

February 3, 2022

Ozgur Taylan Turan

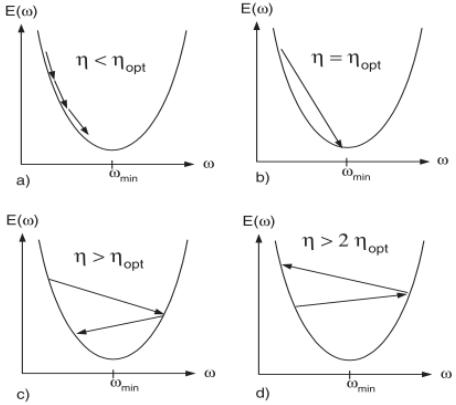
Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability[1]

[1] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar (2021). “Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability”. In: pp. 1–80. arXiv: 2103.00065

Why This Paper?

Interesting empirical result on gradient descent...

Stability of Gradient Descent

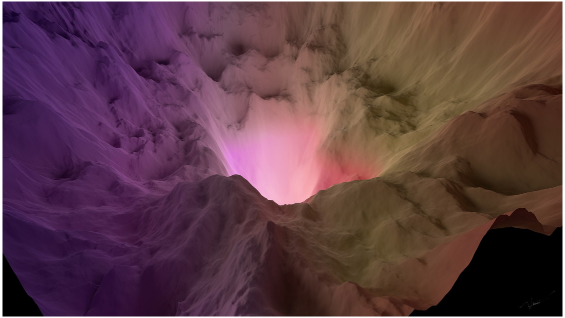


[1]

- **Quadratic** Objective: E(ω)
- $\omega_{t+1} = \omega_t - \eta E'(\omega)$
- Learning Rate: η
- $\eta_{opt} = (E''(\omega))^{-1}$ inverse of Hessian
- If $\eta > 2\eta_{opt}$ $\quad\rightarrow$ Divergence

[1] G. B. Orr and K.-R. Müller (1998). *Neural Networks: Tricks of the Trade, this book is an outgrowth of a 1996 NIPS workshop.* ISBN: 3-540-65311-2. arXiv: 9780201398298

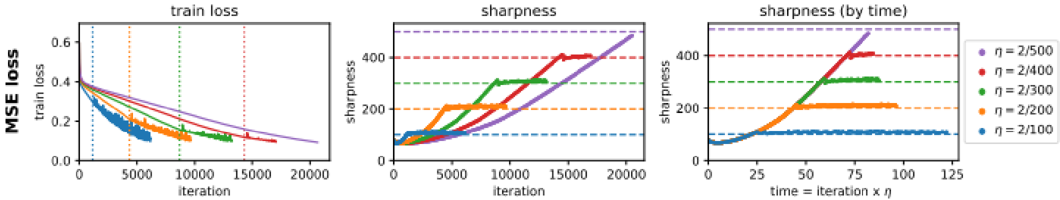
Gradient Descent on Neural Networks



https://losslandscape.com

- Losses ($\mathcal{L}(\omega)$) are not globally quadratic!
- But, second order Taylor expansion around any point in parameter space is Quadratic!
- Then, if $\mathbf{H} > \frac{2}{\eta} \quad \rightarrow$ Divergence
- Hessian largest eigenvalue = Sharpness

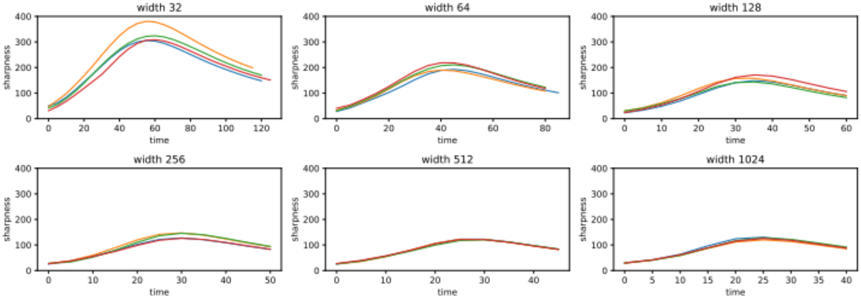
Progressive Sharpening



- Full-batch, vanilla-GD
- CIFAR-10/subset of 5000 examples
- Fully-connected/two-layer/200-width/tanh/stop-99% acc.

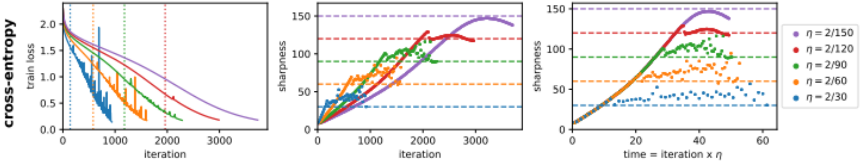
Furthermore

- Effect of width? Lesser degree
- Other Losses? Different behaviour
- Other experiments? Changing arch.+tasks



Furthermore

- Effect of width? Lesser degree
- Other Losses? Different behaviour
- Other experiments? Changing arch.+tasks



Furthermore

- Effect of width? Lesser degree
- Other Losses? Different behaviour
- Other experiments? Changing arch.+tasks

Conclusions

- Training loss decrease is non-monotonic
- *L-smoothness* assumption might be in jeopardy... (convergence analysis)
- Edge of Stability is inherently non-quadratic