Any-stepsize Gradient Descent for Separable Data under Fenchel–Young Losses

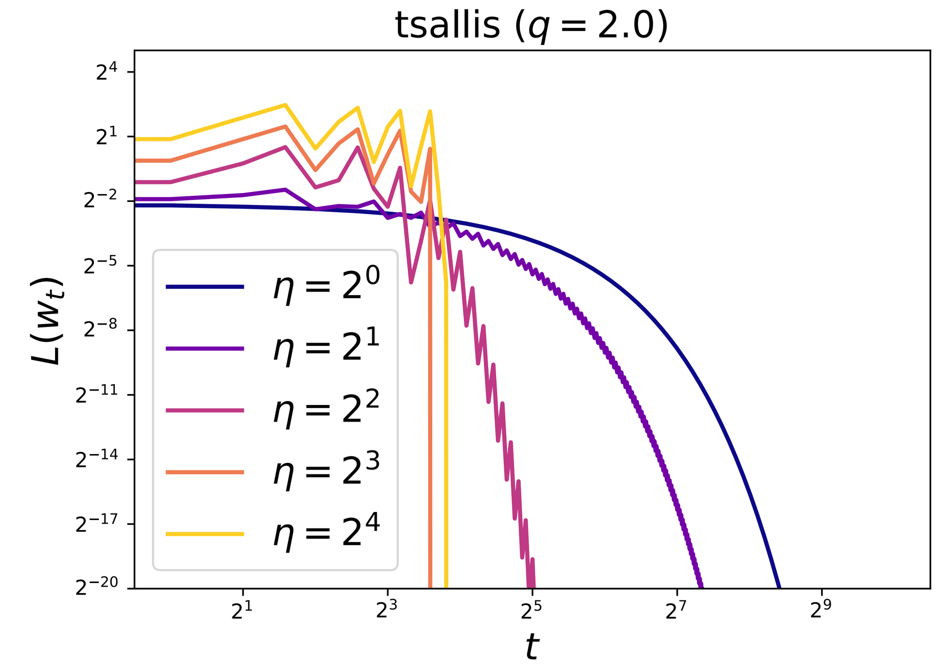
Han Bao (The Institute of Statistical Mathematics) Shinsaku Sakaue (CyberAgent) Yuki Takezawa (Kyoto University / OIST)

Unlike the classical GD analysis ...

$$\underbrace{L(\mathbf{w}_{t+1}) \le L(\mathbf{w}_t) + \left(\frac{\beta}{2}\eta - 1\right)\eta\|\nabla L(\mathbf{w}_t)\|_2^2}_{\beta\text{-smoothness}} \underbrace{< L(\mathbf{w}_t)}_{\text{by } \eta < 2/\beta} \quad \text{(descent lemma)}$$

We show GD convergence for <u>any stepsize η</u> under linear, binary classification.

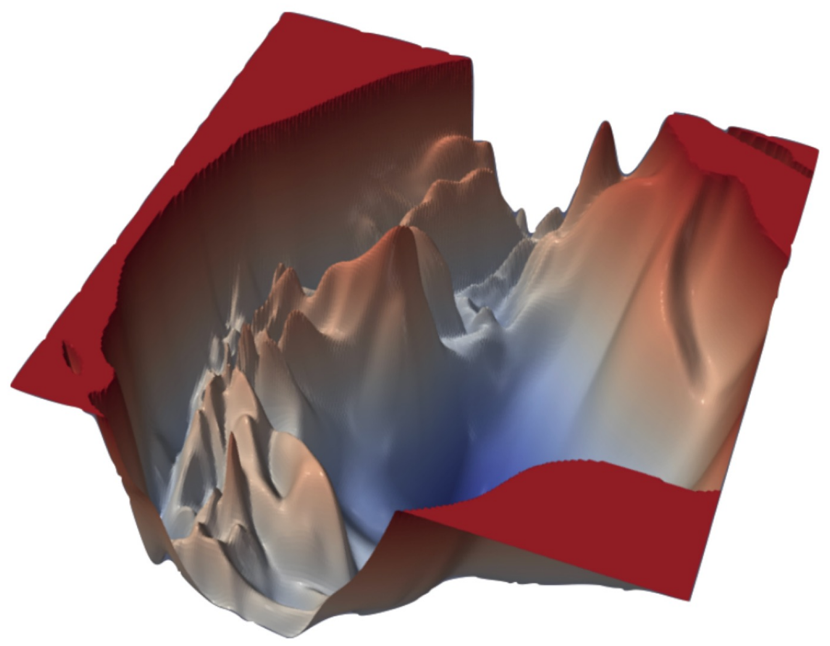
For long enough iterations $T \gtrsim n\gamma^{-2}(1+\eta^{-1})\varepsilon^{-\alpha}$ linear classification risk is ε-optimal $\min_{t\in[T]} L(\mathbf{w}_t) \le \varepsilon$



GD (w/ Tsallis 2-loss) converges quickly

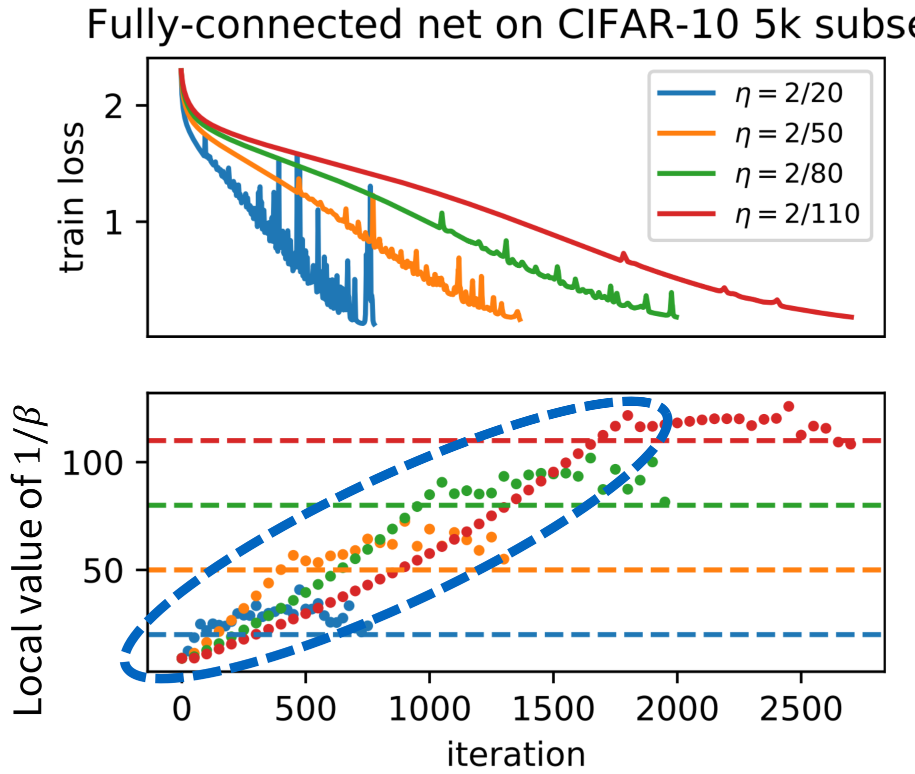
Background: Edge of stability (EoS)

Loss landscapes of NNs tend to be highly nonsmooth ...



Li et al. (NeurIPS2018) "Visualizing the loss landscape of neural nets"

GD keeps working with excessively large stepsize!



Initially $\eta < 2/\beta$ is not satisfied

Q. Why GD converging beyond $\eta < 2/\beta$?

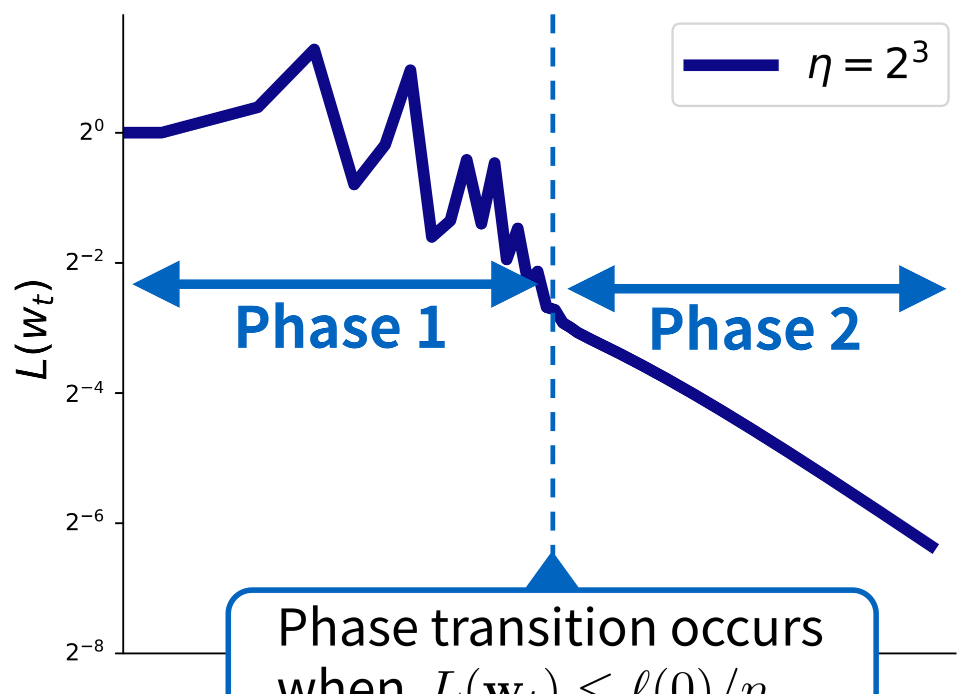
Cohen et al. (ICLR2021) "Gradient descent on neural networks typically occurs at the edge of stability"

Previous work: 2-phase analysis

Setup

- Binary classification $(\mathbf{x}_i, y_i)_{i\in[n]}$
- Linearly separable data $\langle \mathbf{w}_*, y_i\mathbf{x}_i\rangle \ge \gamma$ for all $i \in [n]$
- Training risk $L(\mathbf{w}) = \frac{1}{n}\sum_{i\in[n]} \ell(\langle \mathbf{w}, y_i\mathbf{x}_i\rangle)$
- Logistic loss $\ell(z) = \ln(1+\exp(-z))$
- Constant stepsize GD $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\nabla L(\mathbf{w}_t)$

Result



Phase 1 (EoS): loss oscillates but its average is controlled

$$\frac{1}{t}\sum_{k=1}^{t-1} L(\mathbf{w}_k) \le \tilde{O}\left(\frac{1}{t}\right)$$

Phase 2 (stable): loss monotonically decreases faster than classical GD rate $O(1/T)$

Thanks to self-bounding property $\ell'' \lesssim \ell$ ▶ $L(\mathbf{w}_T) \le \tilde{O}\left(\frac{1}{T^2}\right)$

☺ Large stepsize η accelerates optimization

☹ But classification is done while EoS phase

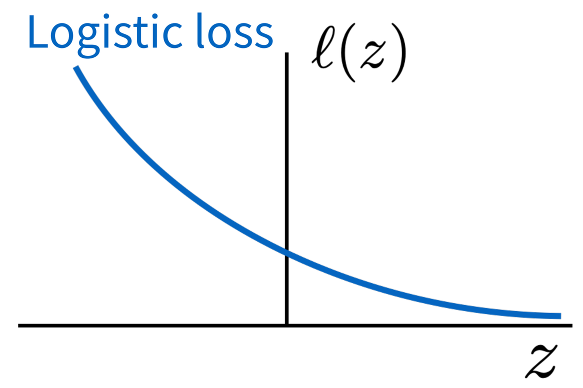
Wu et al. (COLT2024) "Large Stepsize Gradient Descent for Logistic Loss"

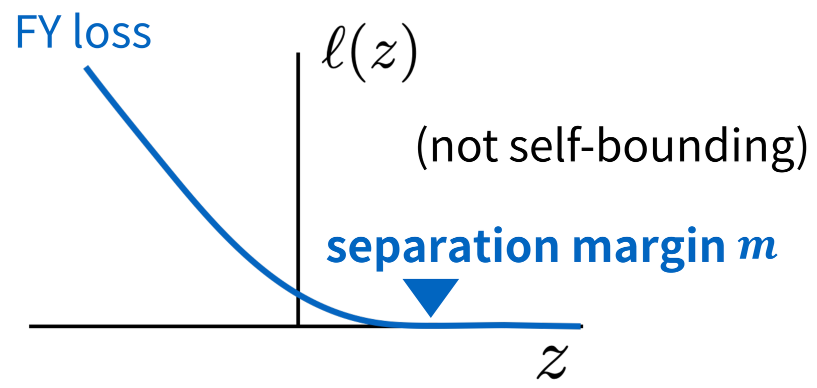
Our result: perceptron analysis

Setup

Use Fenchel-Young loss: $\ell(z) = \phi^*(-z)$

(ϕ: Legendre-type convex function, e.g., Tsallis neg-entropy)





All other setups remain the same as Wu et al. (2024)

Result

Theorem. For twice differentiable & Legendre-type ϕ, GD starting with $\mathbf{w}_0 = \mathbf{0}$ achieves $\min_{t\in[T]} L(\mathbf{w}_t) \le \varepsilon$ after at most $T \gtrsim n\gamma^{-2}(1+\eta^{-1})\varepsilon^{-\alpha}$ steps, where

$$\alpha = \limsup_{\mu\to 0} \frac{\phi'(\mu)}{\mu\phi''(\mu)}\left[1 - \frac{\phi(\mu)}{\mu\phi'(\mu)}\right]$$

Examples.

- Tsallis neg-entropy ($1 < q < 2$): $\alpha = 1/q$
- Tsallis neg-entropy ($q \ge 2$): $\alpha = 1/2$
- Renyi 2-neg-entropy: $\alpha = 1/3$

Proof sketch. By the following perceptron inequality:

$$\underbrace{C_L \cdot t \le \langle \mathbf{w}_t, \mathbf{w}_*\rangle}_{\text{by linear separability}} \le \underbrace{\|\mathbf{w}_t\| \le O(1)}_{\text{by separation margin}}$$

This holds during the risk is ε-suboptimal