Implicit Bias of Gradient Descent for Logistic Regression at the Edge of Stability

Jingfeng Wu[1], Vladimir Braverman[2], Jason Lee[3]
[1]Johns Hopkins University (=> UC Berkeley),
[2]Rice University, [3]Princeton University

NEURAL INFORMATION PROCESSING SYSTEMS

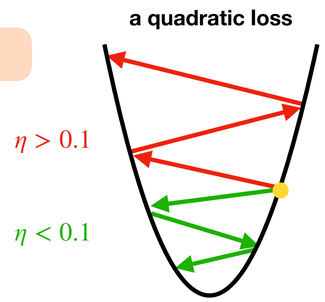
Background

$$w_+ = w - \eta \nabla L(w)$$

How to choose stepsize / learning rate?

Descent Lemma

- For **small** η, $L(w_t)$ decreases **monotonically**, GD works
- For large η, GD does not work for quadratics

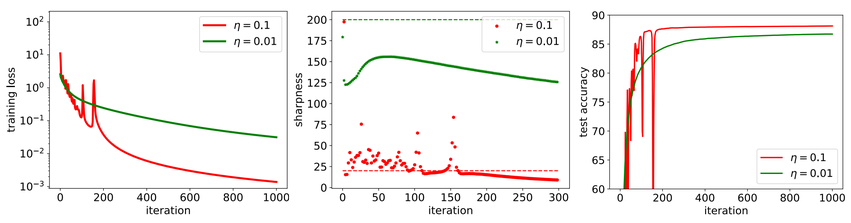


$$\begin{aligned} L(w_+) &= L(w - \eta \cdot \nabla \ell(w)) \\ &= L(w) - \eta \cdot \|\nabla \ell(w)\|^2 + \frac{\eta^2}{2} \cdot \nabla L(w)^\top \cdot \nabla^2 L(w) \cdot \nabla L(w) - O(\eta^3) \\ &\le L(w) - \eta \cdot \left(1 - \frac{\eta}{2} \cdot \|\nabla^2 L(w)\|_2\right) \cdot \|\nabla L(w)\|^2 - O(\eta^3) \end{aligned}$$

Edge of Stability

small stepsize works; large stepsize also works

non-monotonicity against descent lemma



3-layer net + 1,000 samples from MNIST+ GD with const-stepsize

Theory

- binary classification data $(x_i, y_i = 1)_{i=1}^n$
- logistic loss + linear model

$$L(w) := \sum_i \log\Big(1 + \exp\big(-\langle w, x_i\rangle\big)\Big)$$

- constant-stepsize GD

$$w_t = w_{t-1} - \eta \cdot \nabla L(w_{t-1})$$

- **Assumption 1.** $\exists w,\ \langle w, x_i\rangle > 0,\ i = 1, \ldots, n$

Convergence at EoS

For **every** constant stepsize $\eta > 0$:

A. in the max-margin subspace,

$$\mathscr{P} \circ w_t \ge \frac{1}{\gamma} \cdot \log(t) + \Theta(1)$$

B. in the non-separable subspace,

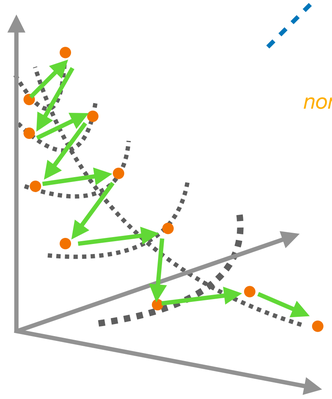
$$\|\overline{\mathscr{P}} \circ w_t\|_2 \le \Theta(1)$$

C. moreover,

$$G\big(\overline{\mathscr{P}} \circ w_t\big) - \min G(\cdot) \le \frac{\Theta(1)}{\log(t)}, \text{ where } G(v) := \sum_{x \in \text{supp.}} \exp\big(-\big\langle \overline{\mathscr{P}} \circ x,\ v\big\rangle\big)$$

D. risk is bounded by

$$L(w_t) \le \frac{\Theta(1)}{t}$$



loss landscape

Negative Results for Exp Loss

Consider exp loss on two 2D samples

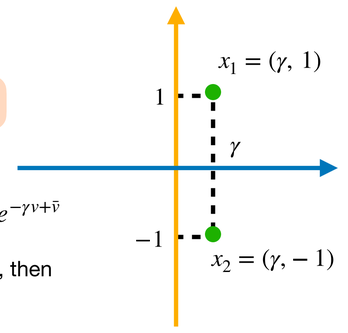
$$L(w) = \textstyle\sum_i e^{-\langle w, x_i\rangle} \quad <=> \quad L(v, \bar{v}) = e^{-\gamma v - \bar{v}} + e^{-\gamma v + \bar{v}}$$

If $0 \le v_0 \le 2,\ |\bar{v}_0| \ge 1,\ 0 < \gamma < 1/4,\ \eta \ge 4$, then

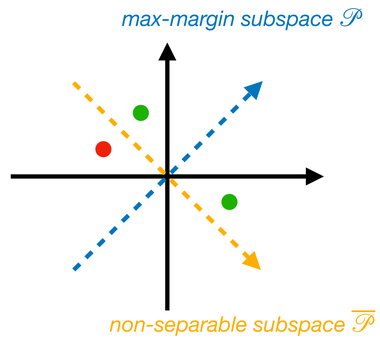
A. $v_t \to \infty$

B. $|\bar{v}_t| > 2\gamma v_t$ and $\bar{v}_t$ flips sign every iteration

C. $L(v_t, \bar{v}_t) \to \infty$



Key Idea



Dual form: $\hat{w} = \alpha_1 \cdot x_1 + \ldots + \alpha_s \cdot x_s$

Orthogonal: $0 = \langle v, \hat{w}\rangle$

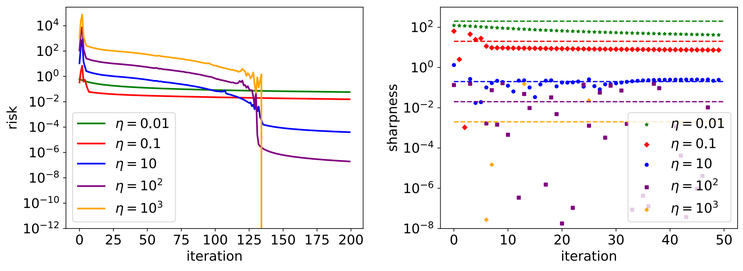
$0 = \alpha_1 \cdot \langle v, x_1\rangle + \ldots + \alpha_s \cdot \langle v, x_s\rangle$

Assumption 2. $\alpha_i > 0$ for supp. vectors

Assumption 3. supp. vectors span the space

Lemma. There must exist $\langle v, x_i\rangle > 0,\ \langle v, x_j\rangle < 0$

Simulation



logistic regression on 1,000 samples of "0" and "8" from MNIST

Conclusion & References

- *convergence of GD with large stepsize*
- *"valley" instead of "quadratic basin"*
- *benefits of logisitic loss*

1. Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., & Srebro, N. (2018). The implicit bias of gradient descent on separable data. JMLR 2019.
2. Ji, Z., & Telgarsky, M. (2018). Risk and parameter convergence of logistic regression. COLT 2019.
3. Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., & Talwalkar, A. (2021). Gradient descent on neural networks typically occurs at the edge of stability. ICLR 2021