Implicit Bias of Gradient Descent

for Logistic Regression at the Edge of Stability

NeurIPS 2023 Spotlight

Jingfeng Wu (JHU -> Berkeley)
with Vladimir Braverman (Rice) and Jason D. Lee (Princeton)

Gradient Descent

$$w_+ = w - \boxed{\eta} \cdot \nabla L(w)$$

making how much update?
AKA., stepsize / learning rate?

Cauchy, 1847

Optimization theory, oversimplified

[descent lemma]

For **small** η, $L(w_t)$ decreases **monotonically**, GD works

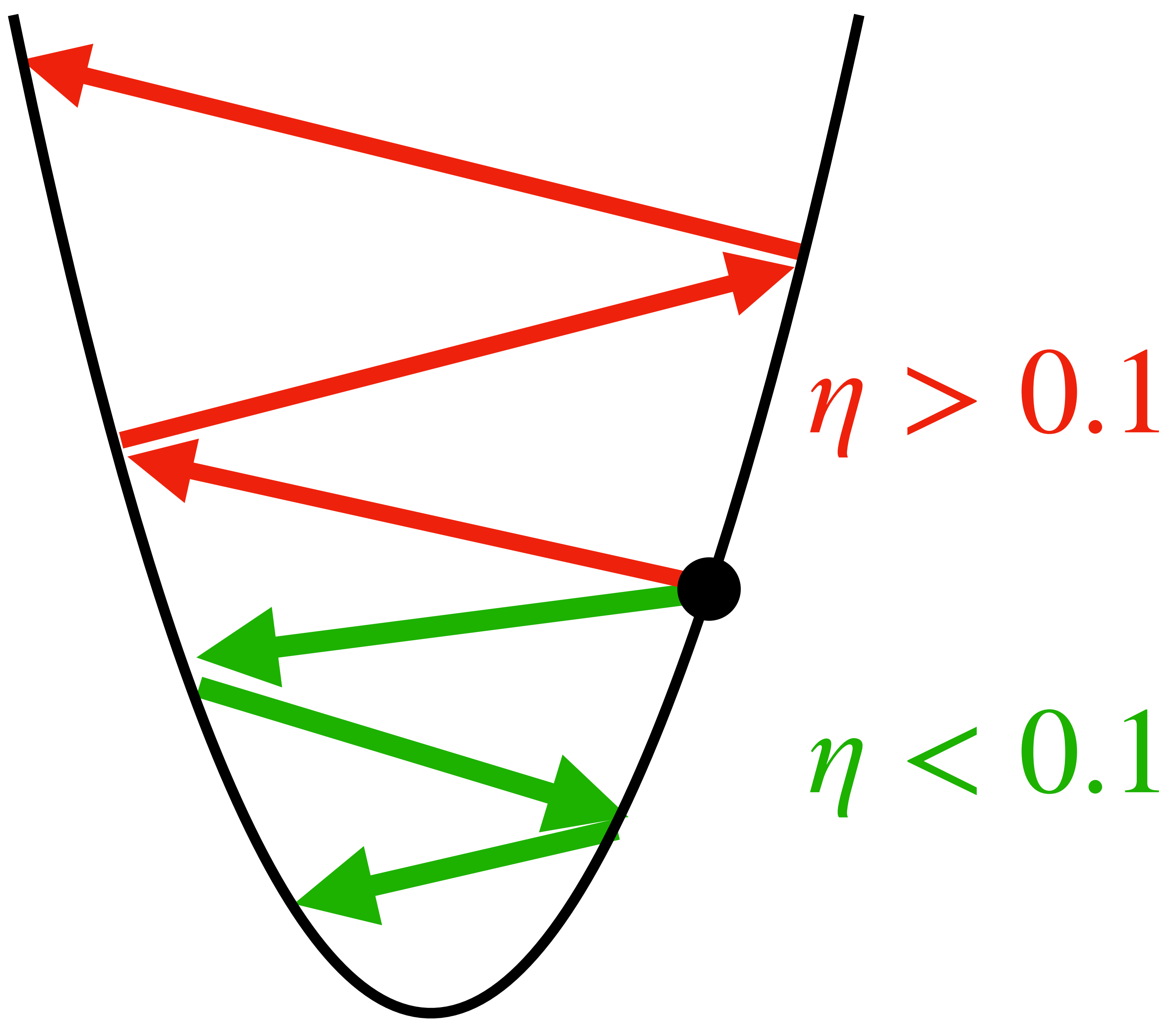
For large η, GD does not work for quadratics

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

$$= L(w) - \eta \cdot \|\nabla \ell(w)\|^2 + \frac{\eta^2}{2} \cdot \nabla L(w)^\top \cdot \nabla^2 L(w) \cdot \nabla L(w) - O(\eta^3)$$

$$\leq L(w) - \eta \cdot \boxed{\left(1 - \frac{\eta}{2} \cdot \|\nabla^2 L(w)\|_2\right)} \cdot \|\nabla L(w)\|^2 - O(\eta^3)$$

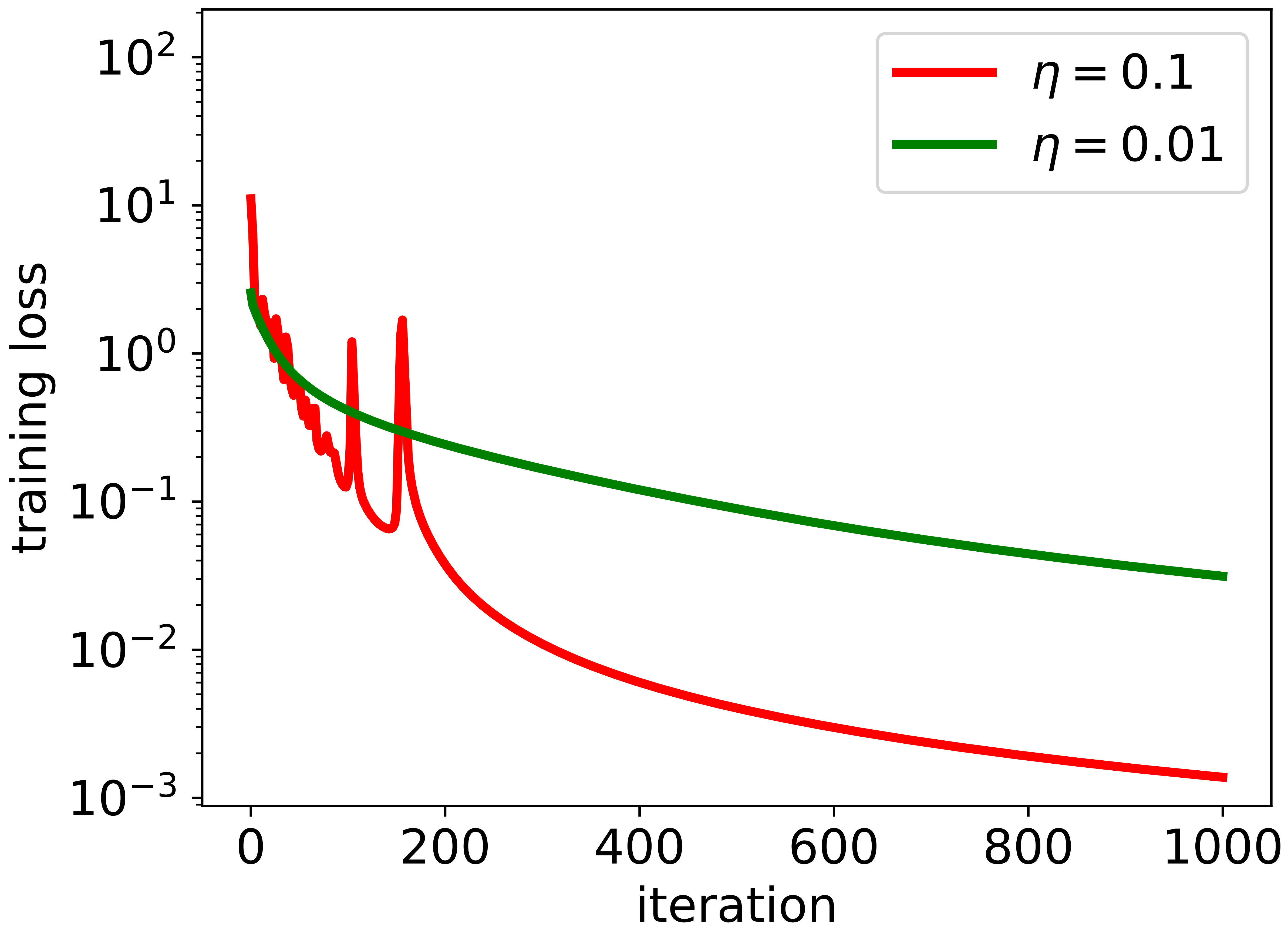
quadratic landscape



$$L(w) = 10w^2$$

$$w \leftarrow w - \eta \cdot 20w$$

Numbers in DL classification



small stepsize works;

large stepsize also works

non-monotonically

edge of stability

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

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

Problem simplification

10 classes -> 2 classes   NN -> linear model (w/o bias)   => logistic regression

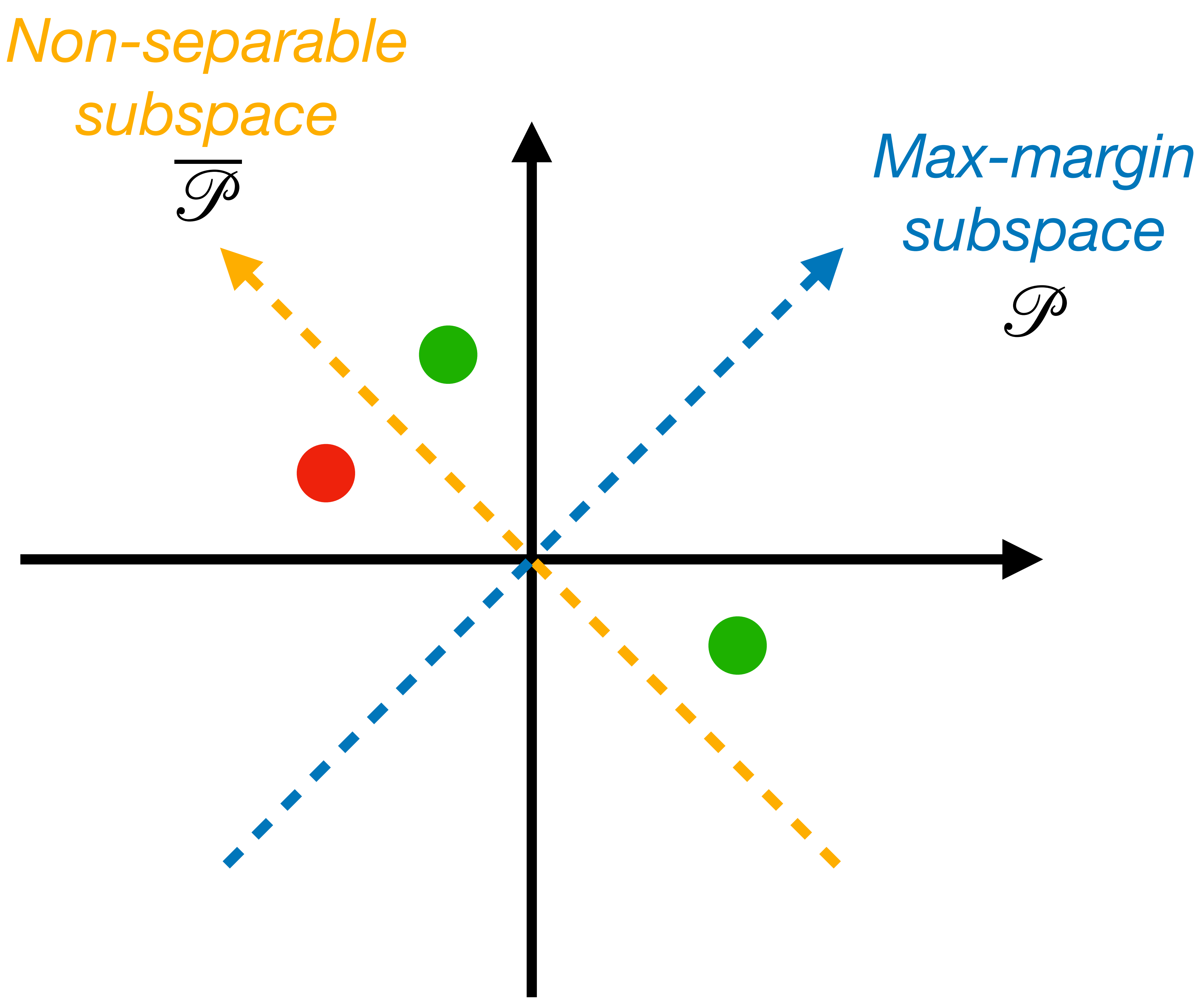
- training data $\big(x_i,\, y_i = 1\big)_{i=1}^{n}$
- **Assumption** 1: $\exists w,\ \langle w, x_i \rangle > 0,\ i = 1,\ldots,n$
- logistic loss

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

- constant-stepsize GD

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

Key idea: space decomposition



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: supp. vectors span the space

Assumption 3: $\alpha_i > 0$ for supp. vectors

Then there must exist $\langle v, x_i \rangle > 0$, $\langle v, x_j \rangle < 0$

Provably convergence under logistic loss

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

A. in the max-margin subspace,

margin

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

B. in the non-separable subspace,

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

C. moreover,

strongly convex

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

D. risk is bounded by

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

Negative example under exp loss

Consider exp loss on two 2D samples

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

Assume that

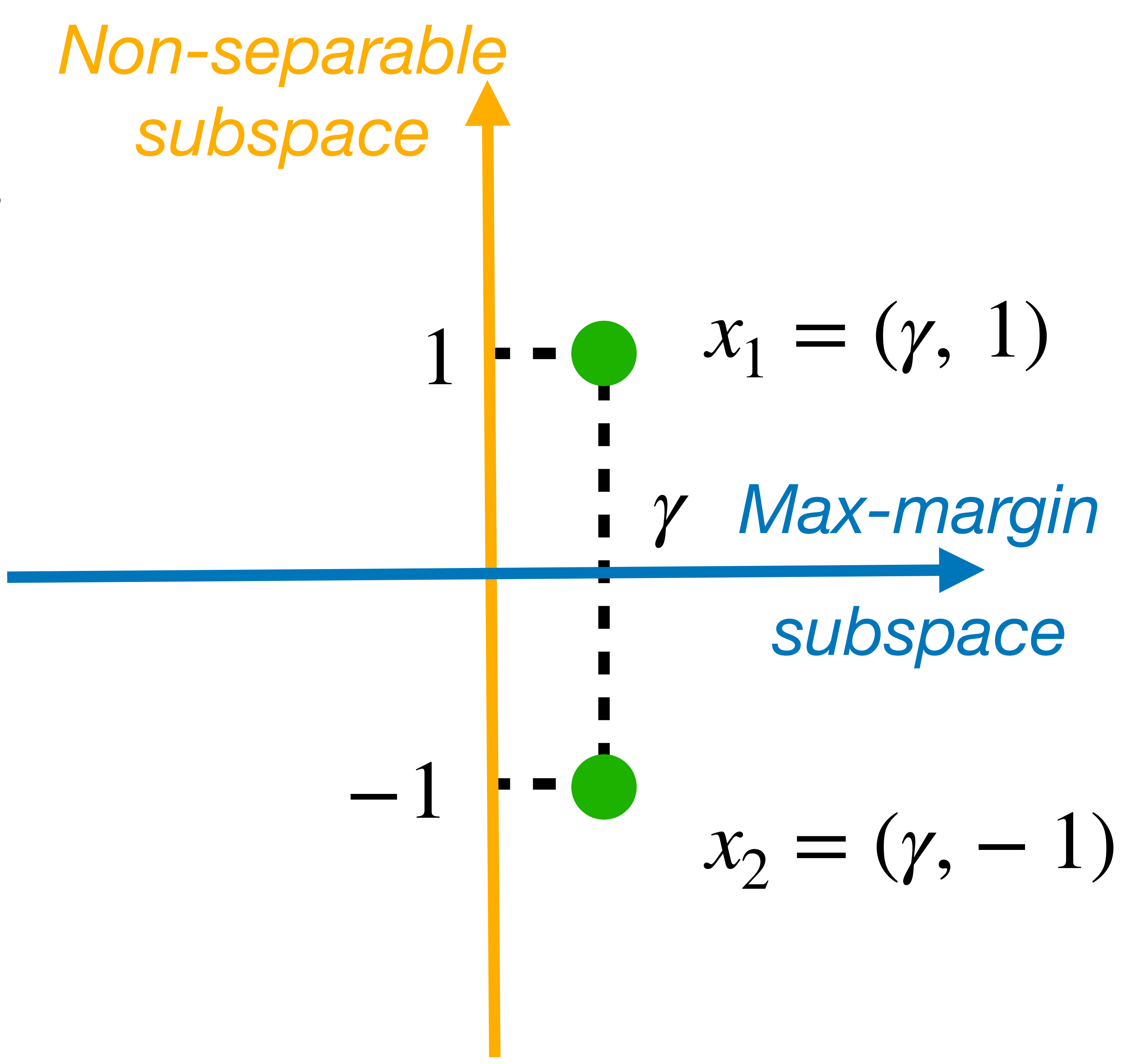
$$0 \leq v_0 \leq 2, \ \ |\bar{v}_0| \geq 1, \ 0 < \gamma < 1/4, \ \eta \geq 4.$$

Then

A. $v_t \to \infty$

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

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



Feel free to use large stepsizes

under logistic loss!