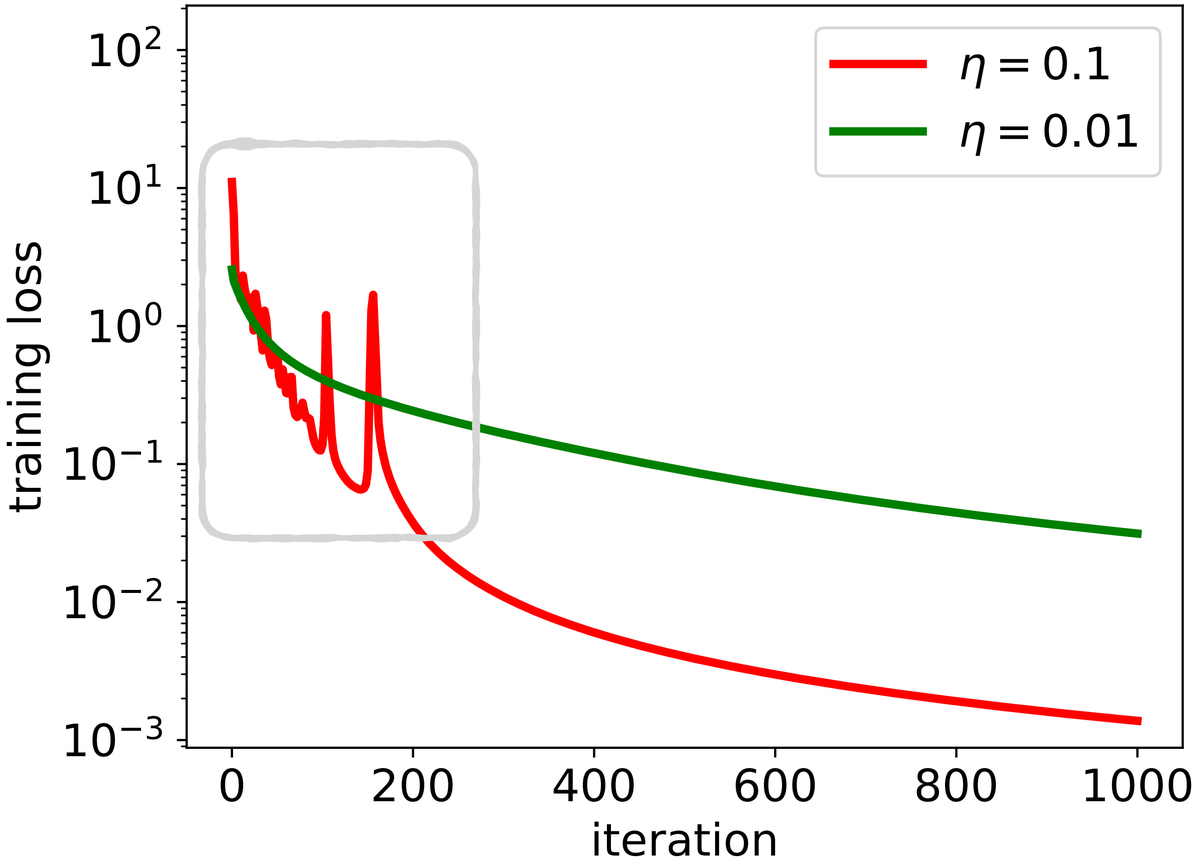
Implicit Bias of Gradient Descent

for Logistic Regression at the *Edge of Stability*

Jingfeng Wu (JHU -> Berkeley), 05/2023
Joint work with Vladimir Braverman (Rice) and Jason D. Lee (Princeton)

What happens in OPT for DL



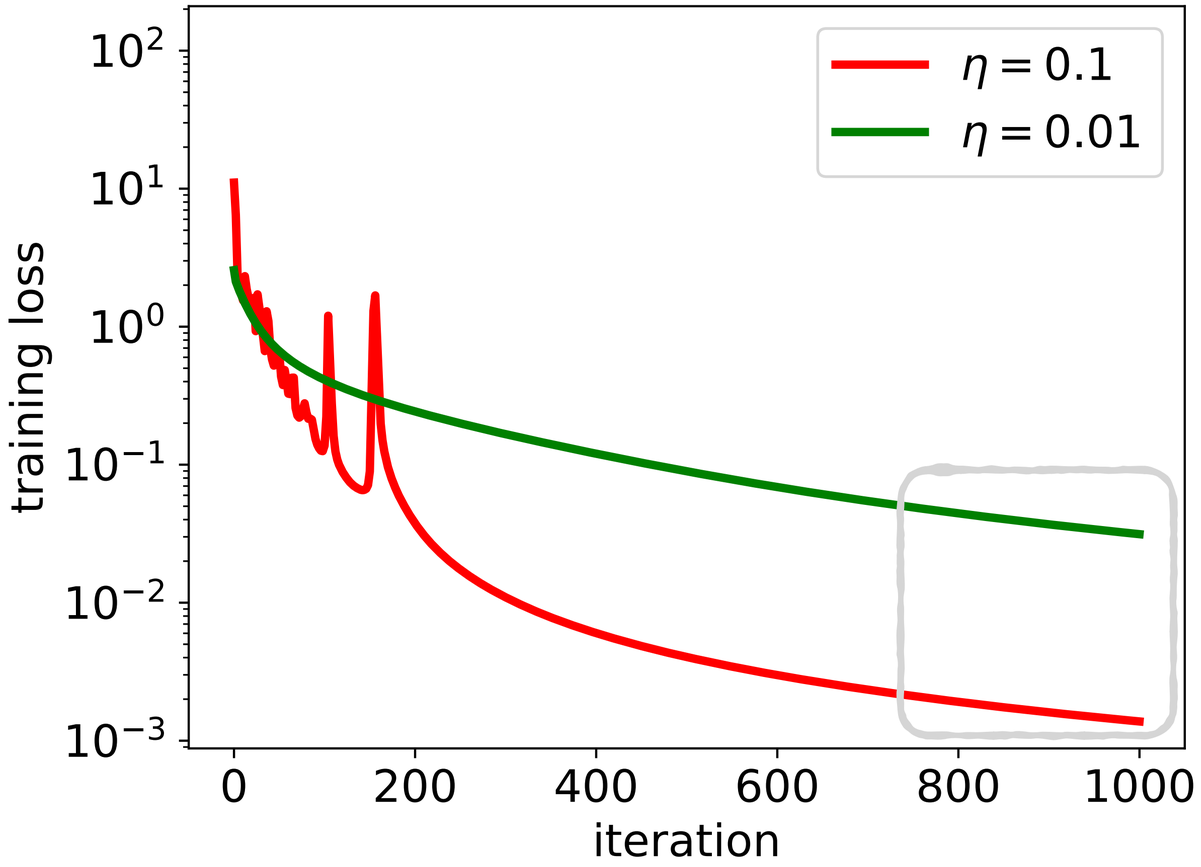
4-layer fully connected net

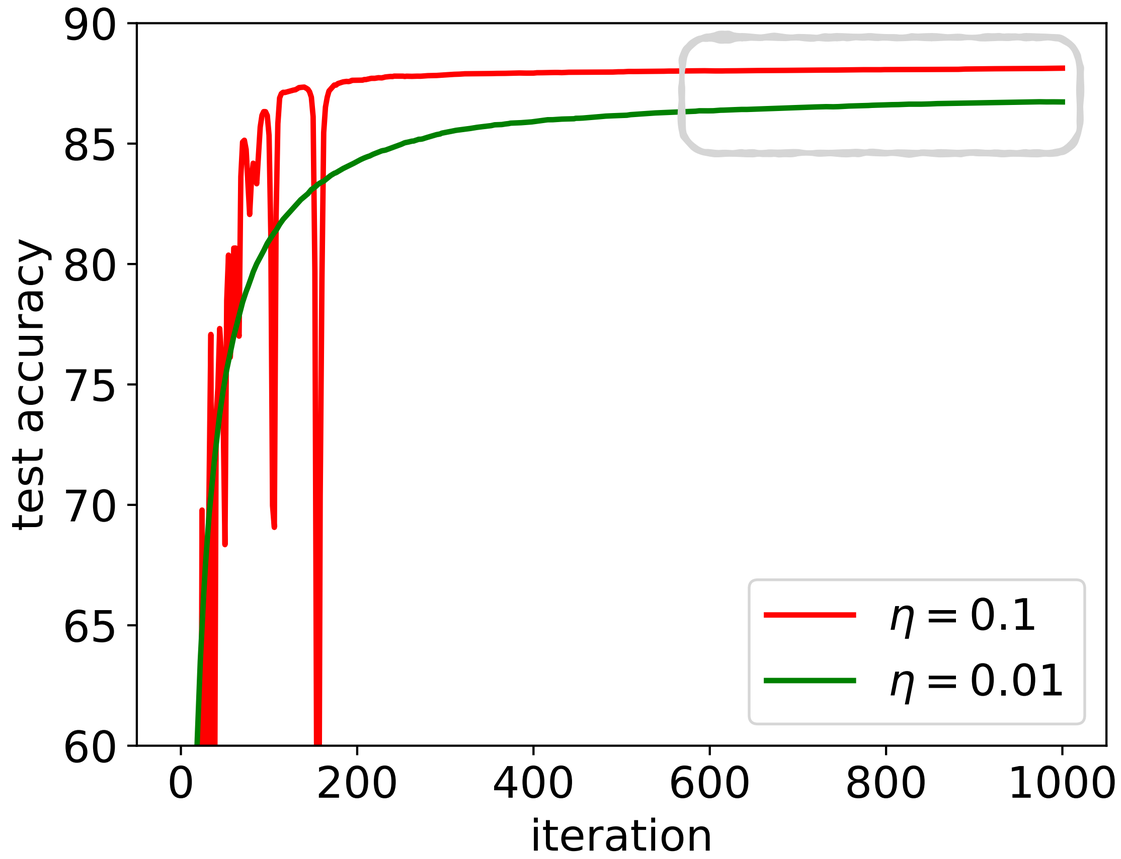
1,000 samples from MNIST

gradient descent (GD)

Risk oscillation caused by large stepsize

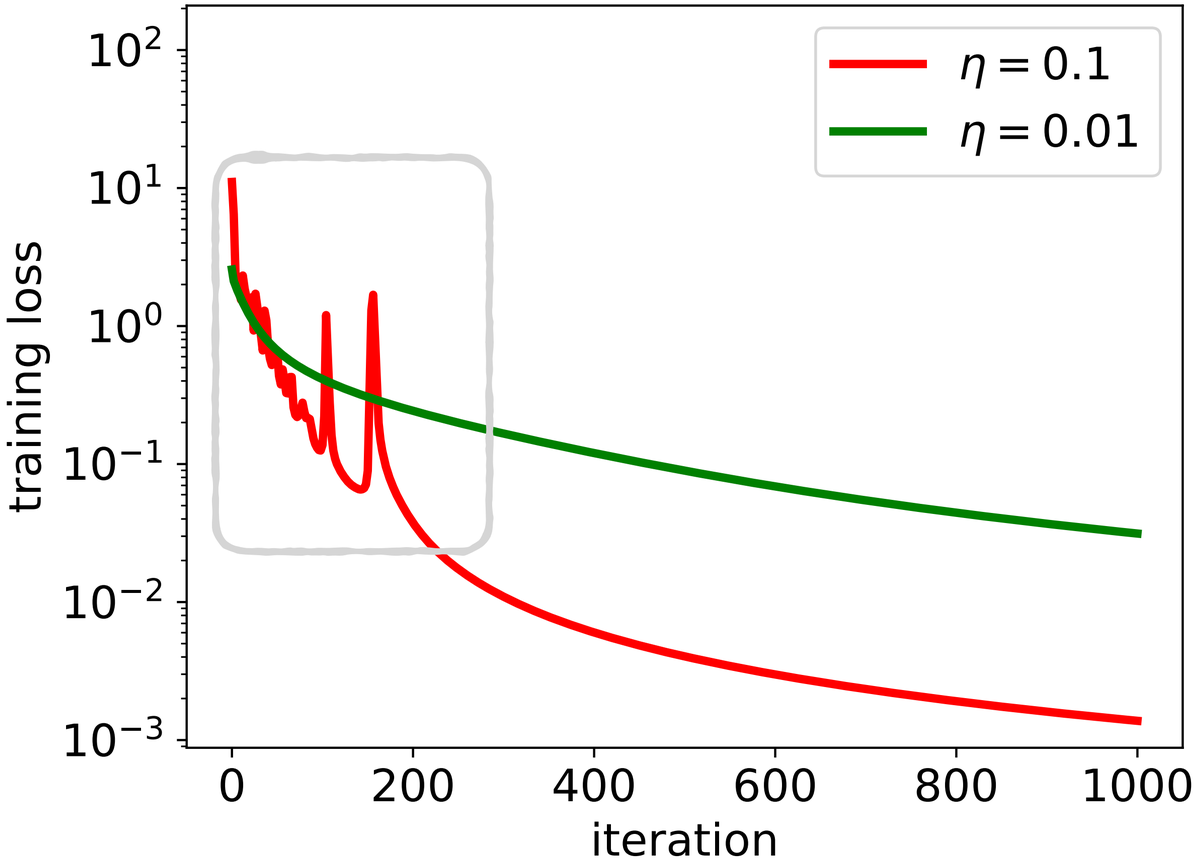
Good hyperparameter is often large

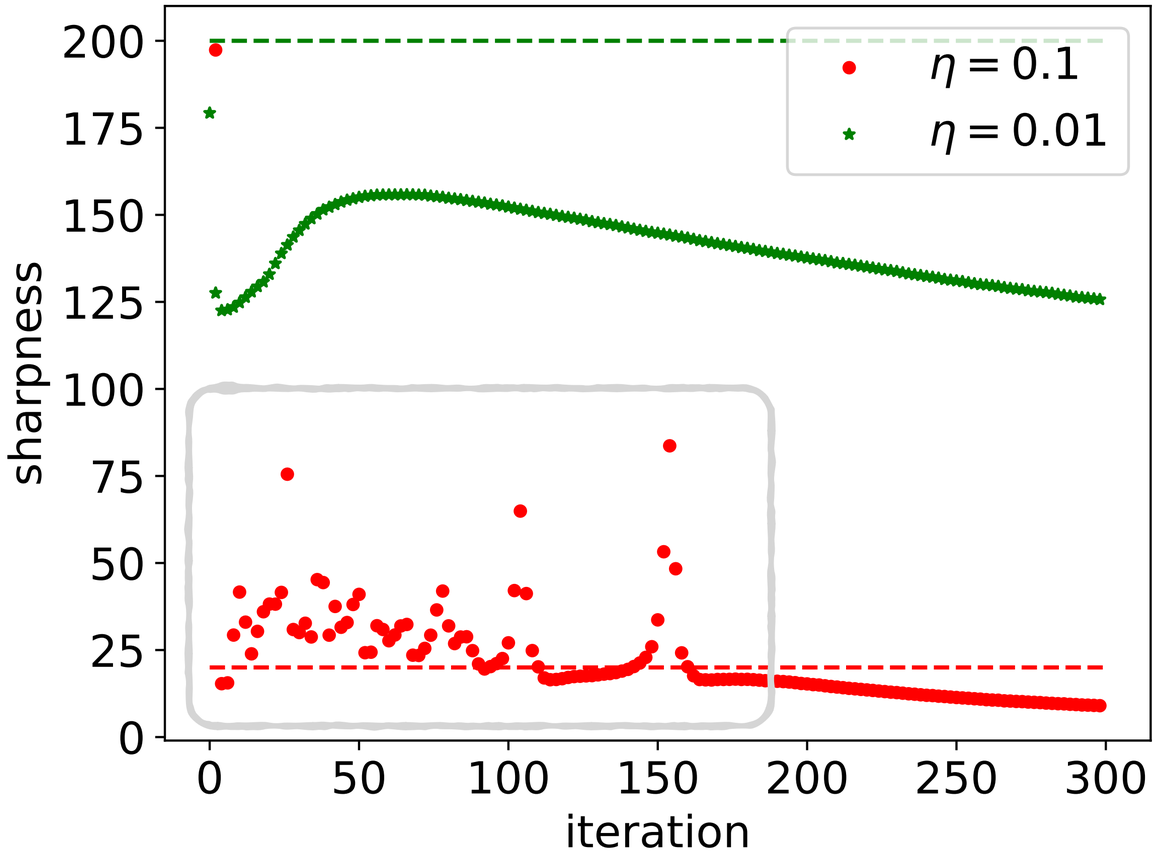




Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y. and He, K., 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*.

Edge of stability (EoS) $\eta > 2 \,/\, \|\nabla^2 L(w)\|_2$





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. *arXiv preprint arXiv:2103.00065*.

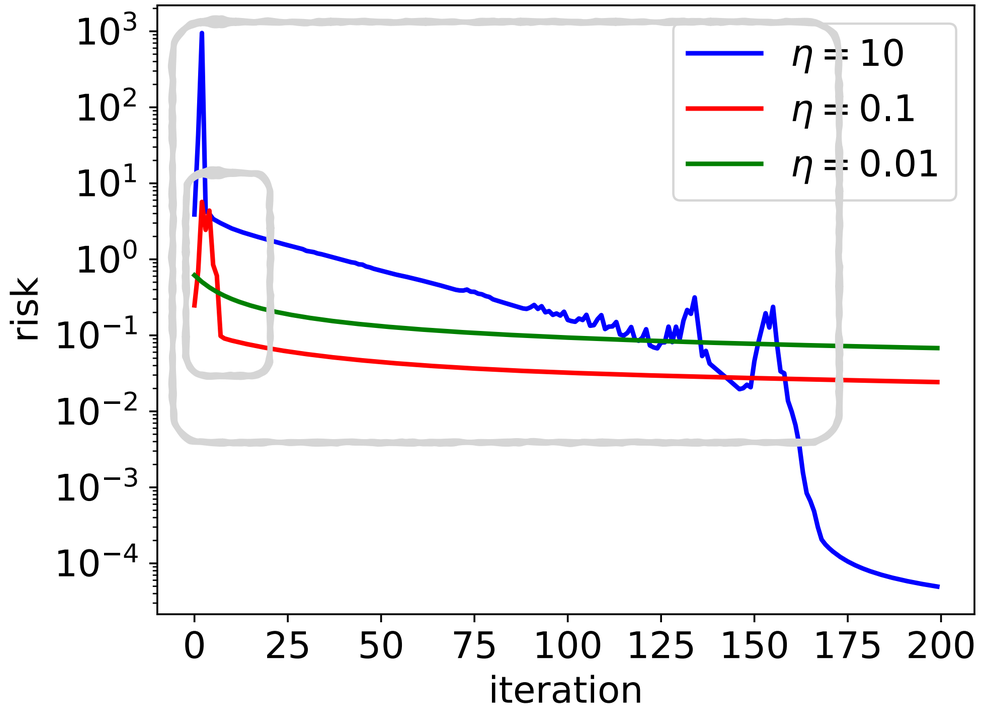
Risks converge while oscillating

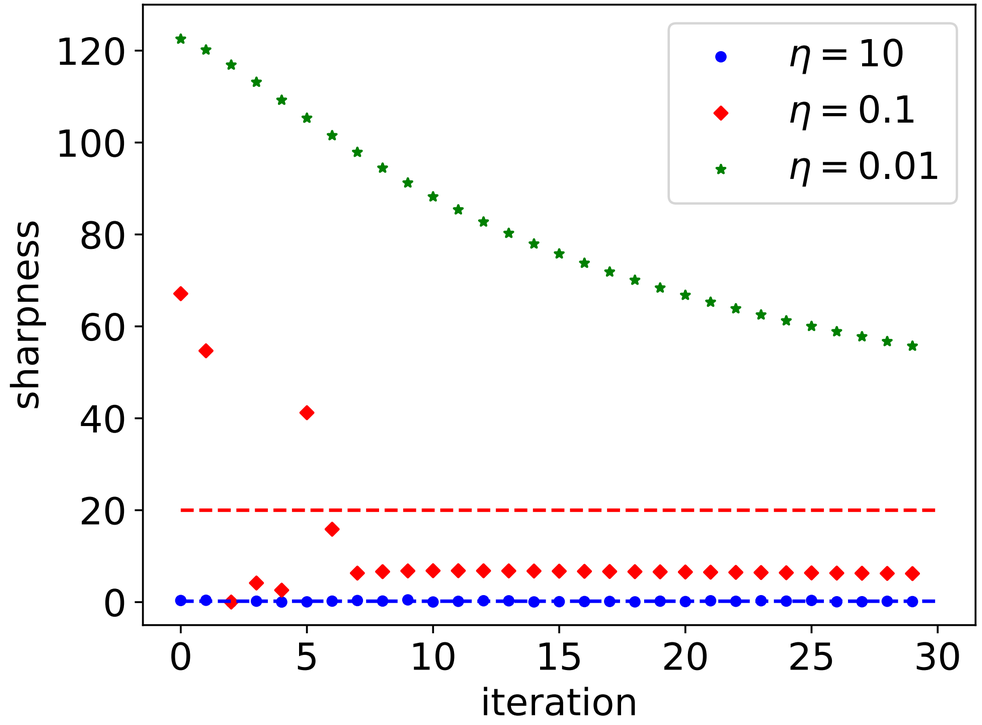
by chance or by math?

Remove “distractions”

10 classes -> 2 classes (y = “0” or “8”)

NN -> linear model (w/o bias)

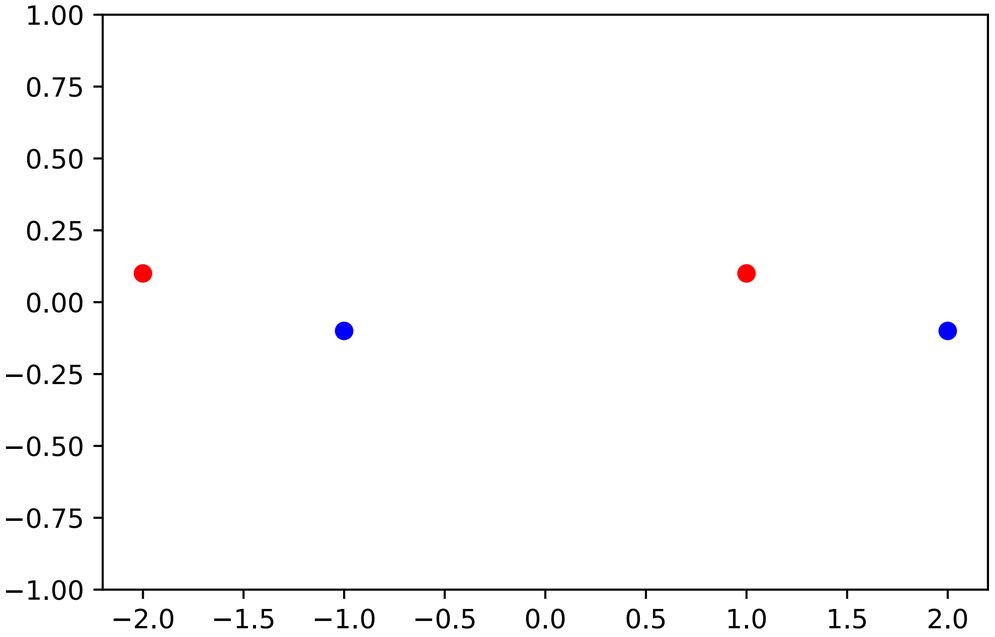


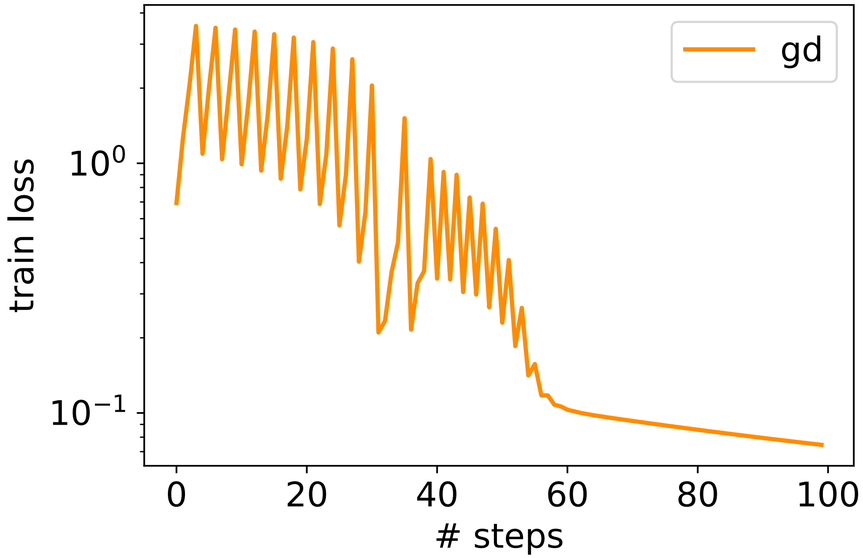


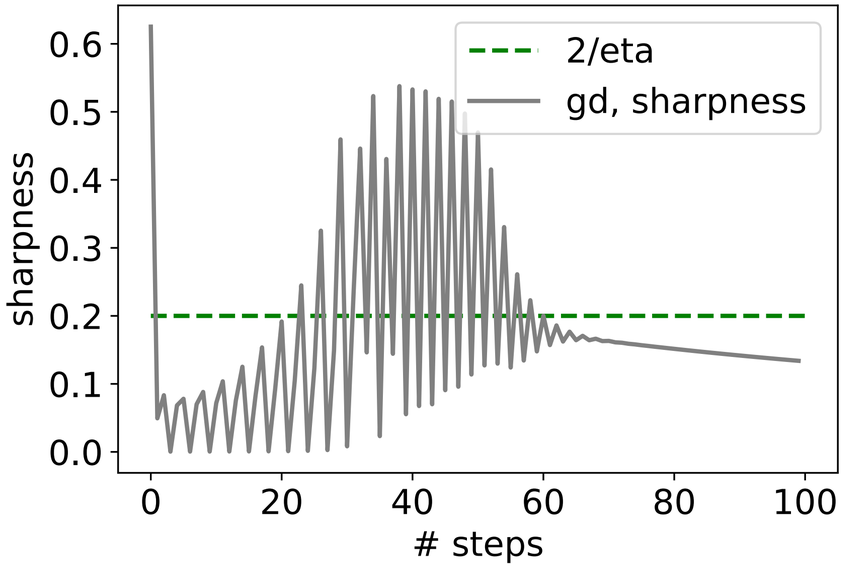
Logistic regression on “linearly separable” data

A even more minimal example

Logistic regression on four 2D samples







Problem setup

Logistic regression on separable data

- Samples $\left(x_i,\ y_i = 1\right)_{i=1}^{n}$

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

- Risk

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

- Constant-stepsize GD

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

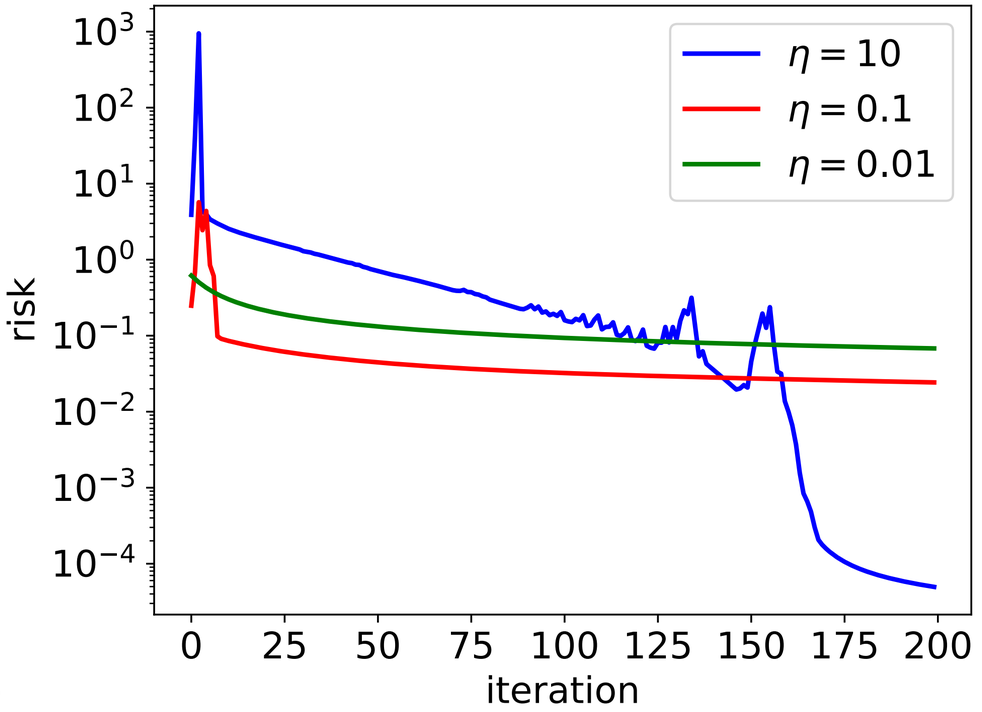
What we know so far

Logistic regression on separable data

If the stepsize is small (e.g., $\eta = 0.01$)

- Risk minimization

$$L(w_t) \leq \tilde{\mathcal{O}}\left(\frac{1}{t}\right)$$ by descent lemma



- Margin maximization

$$\left| \frac{w_t}{\|w_t\|_2} - \frac{\hat{w}}{\|\hat{w}\|_2} \right| \leq \tilde{\mathcal{O}}\left(\frac{1}{\log(t)}\right)$$

SVM solution

$$\hat{w} = \arg\min \|w\|_2, \text{ s.t. } \forall i, \langle w, x_i \rangle \geq 1$$

- Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., & Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, *19*(1), 2822-2878.
- Ji, Z., & Telgarsky, M. (2018). Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*.

A gap between theory and practice

Small stepsize ✅

guaranteed risk minimization,
margin maximization,
…

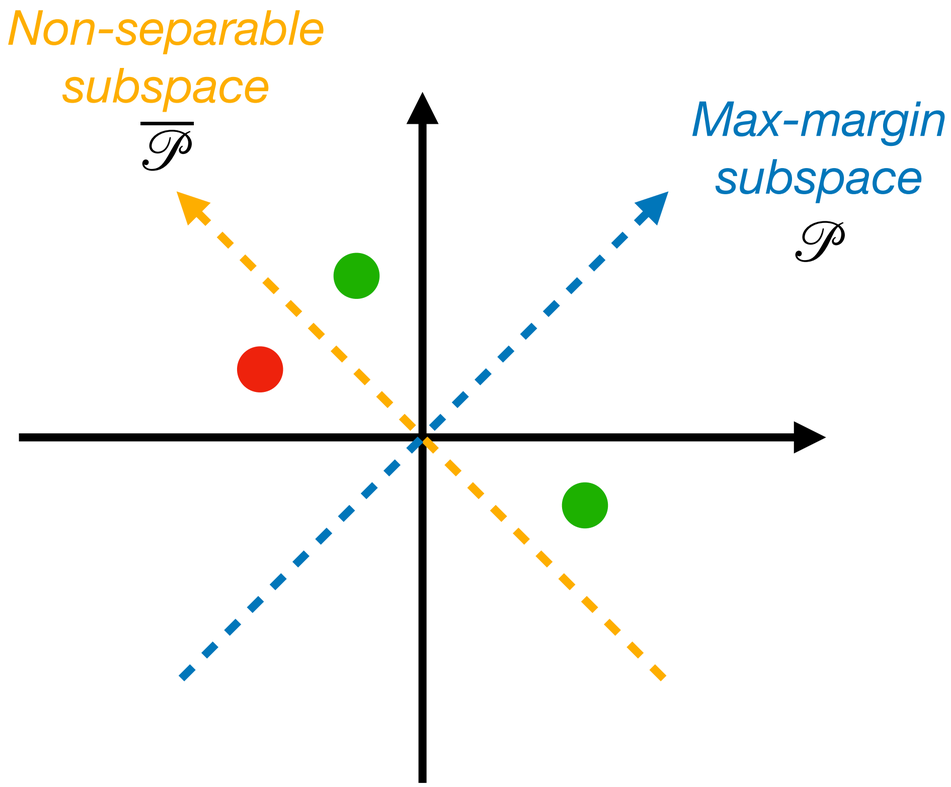
Large stepsize works well

trust me bro!
tried 3 different seeds

Our Results

(any) const-stepsize
+
logistic regression
+
separable data

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 $\langle v, x_i \rangle > 0$, $\langle v, x_j \rangle < 0$

Implicit Bias

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

A. In the max-margin subspace,

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

B. In the non-separable subspace,

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

C. In the non-separable subspace,

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

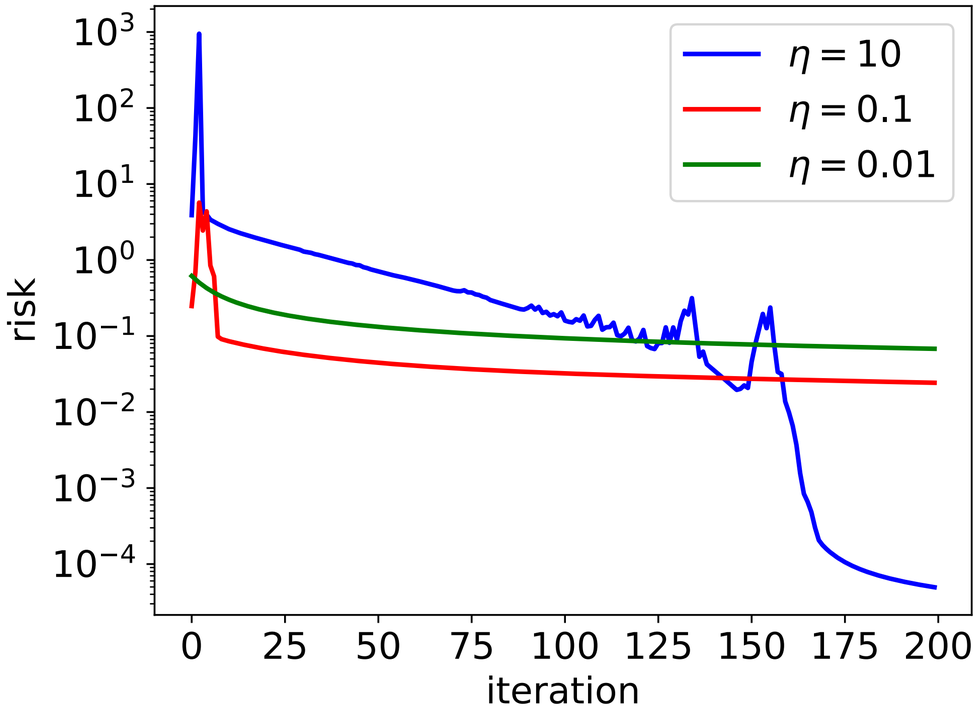
strongly convex

Risk minimization

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

D. Risk is bounded by

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



theory based!

large stepsize => risk still converges

possibly non-monotonically

Feel free to use large stepsizes?

GD can diverge 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

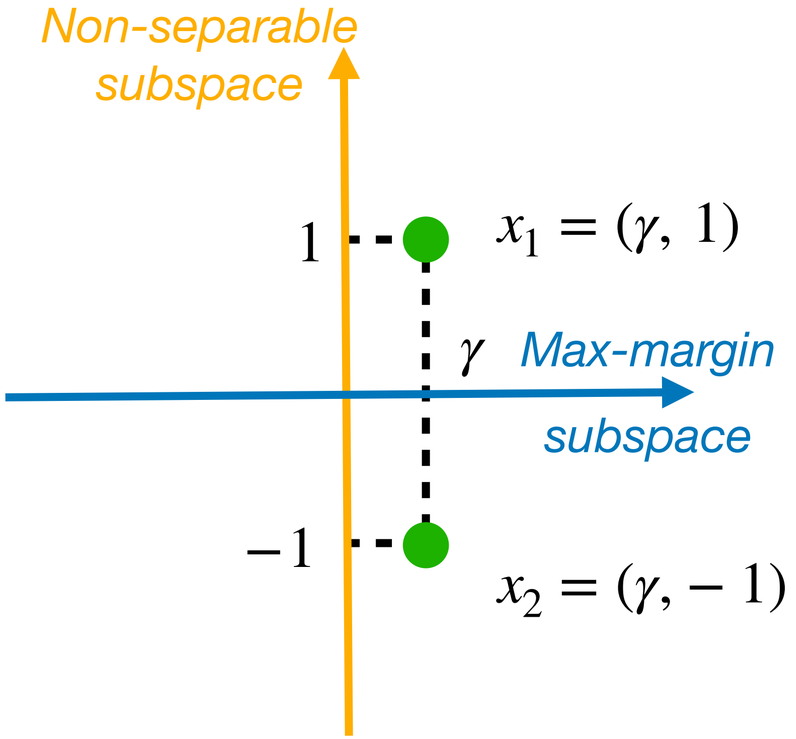
$$0 \le v_0 \le 2, \ \ |\bar{v}_0| \ge 1, \ 0 < \gamma < 1/4, \ \eta \ge 4.$$

Then

A. $v_t \to \infty$

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

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



EoS ≠ small-stepsize

logistic loss > exp loss

Feel free to use large stepsizes under logistic loss!

Techniques Overview

A new approach for handling EoS

Existing approaches

1. Show risk convergence

   (by descent lemma)

2. Show iterate limiting behaviors

descent lemma is broken in EoS

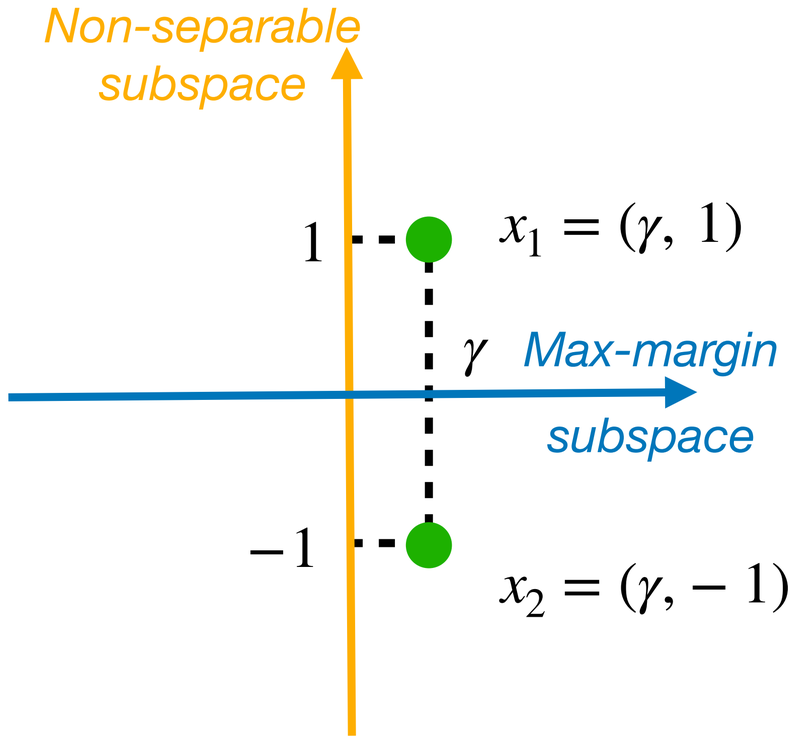
Our approach

1. Study iterate limiting behaviors
2. Show risk convergence

   (by iterate limits)

- Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., & Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, *19*(1), 2822-2878.
- Ji, Z., & Telgarsky, M. (2018). Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*.

Work out the two samples



$$L(v, \bar{v}) = \log\left(1 + e^{-\gamma v - \bar{v}}\right) + \log\left(1 + e^{-\gamma v + \bar{v}}\right)$$

$$v_{t+1} = v_t - \eta \cdot g_t \qquad \bar{v}_{t+1} = \bar{v}_t - \eta \cdot \bar{g}_t$$

$$g_t := -\gamma \cdot \left(\frac{1}{1 + e^{\gamma v_t + \bar{v}_t}} + \frac{1}{1 + e^{\gamma v_t - \bar{v}_t}}\right) \qquad \bar{g}_t := -\left(\frac{1}{1 + e^{\gamma v_t + \bar{v}_t}} - \frac{1}{1 + e^{\gamma v_t - \bar{v}_t}}\right)$$

For simplicity, assume $v_0 = 0,\ |\bar{v}_0| > 0.$

Step 1: $(\bar{v}_t)_{t\geq 0}$ are uniformly bounded

$$\bar{g}_t := -\left(\frac{1}{1+e^{\gamma v_t+\bar{v}_t}} - \frac{1}{1+e^{\gamma v_t-\bar{v}_t}}\right)$$

$$\begin{aligned}
|\bar{v}_{t+1}| &= ||\bar{v}_t| - \eta \cdot |\bar{g}_t|| \\
&\leq \max\{\,|\bar{v}_t|,\, \eta \cdot |\bar{g}_t|\,\} && \bar{v}_t \text{ and } \bar{g}_t \text{ share the same sign} \\
&\leq \max\{\,|\bar{v}_t|,\, \eta\} && \bar{g}_t \text{ is bounded} \\
&\leq \dots \\
&\leq \max\{\,|\bar{v}_0|,\, \eta\} && \text{induction}
\end{aligned}$$

Step 2: $g_t \approx -\gamma \cdot e^{-\gamma v_t} \cdot \Theta(1)$**, so** $v_t \approx \log(t)/\gamma$ *=> margin gets max-ed*

Step 3: $\bar{g}_t \approx e^{-\gamma v_t} \cdot \nabla G(\bar{v}_t)$**, so**

$$\bar{v}_{t+1} \approx \bar{v}_t - \eta_t \cdot \nabla G(\bar{v}_t)\text{, where } \eta_t = \eta \cdot e^{-\gamma v_t} \approx \Theta(1)/t$$

Step 4: a modified descent lemma

$$
\begin{aligned}
G(\bar{v}_{t+1}) &\le G(\bar{v}_t) + \langle \nabla G(\bar{v}_t), \bar{v}_{t+1} - \bar{v}_t \rangle + \frac{\beta}{2} \cdot \|\bar{v}_{t+1} - \bar{v}_t\|_2^2 \\
&\approx G(\bar{v}_t) - \eta_t \cdot \|\nabla G(\bar{v}_t)\|_2^2 + \frac{\beta}{2} \cdot \eta_t^2 \cdot \|\nabla G(\bar{v}_t)\|_2^2 \quad (+\text{higher orders}) \\
&\le G(\bar{v}_t) + \Theta(1) \cdot \eta_t^2 \cdot \|\nabla G(\bar{v}_t)\|_2^2 \\
&\le G(\bar{v}_t) + \frac{1}{t^2} \cdot \Theta(1)
\end{aligned}
$$

The "increase" of risk must decrease

So for $T \ge t \ge 1$, $G(\bar{v}_T) \le G(\bar{v}_t) + \frac{1}{t} \cdot \Theta(1)$ *For small-stepsize, $G(\bar{v}_T) < G(\bar{v}_t)$*

Step 5: convergence of $\bar{v}_t$

$$\begin{aligned}\|\bar{v}_{t+1} - \bar{v}_*\|_2^2 &= \|\bar{v}_t - \bar{v}_*\|_2^2 + 2 \cdot \langle \bar{v}_t - \bar{v}_*, \bar{v}_{t+1} - \bar{v}_t \rangle + \|\bar{v}_{t+1} - \bar{v}_t\|_2^2 \\ &\approx \|\bar{v}_t - \bar{v}_*\|_2^2 - 2\eta_t \cdot \langle \bar{v}_t - \bar{v}_*, \nabla G(\bar{v}_t) \rangle + \eta_t^2 \cdot \|\nabla G(\bar{v}_t)\|_2^2 \\ &\leq \|\bar{v}_t - \bar{v}_*\|_2^2 - 2\eta_t \cdot \Big(G(\bar{v}_t) - G(\bar{v}_*)\Big) + \frac{1}{t^2} \cdot \Theta(1)\end{aligned}$$

$$\sum_{t=t_0}^{T} \eta_t \cdot \Big(G(\bar{v}_t) - G(\bar{v}_*)\Big) \leq \Theta(1) + \sum_{t=t_0}^{T} \frac{1}{t^2} \cdot \Theta(1) \leq \Theta(1)$$

$$G(\bar{v}_T) - G(\bar{v}_*) \leq \frac{\sum_{t=t_0}^{T} \eta_t \cdot \Big(G(\bar{v}_t) - G(\bar{v}_*)\Big) + \sum_{t=t_0}^{T} \eta_t \cdot \frac{\Theta(1)}{t}}{\sum_{t=t_0}^{T} \eta_t} \leq \frac{\Theta(1)}{\sum_{t=t_0}^{T} \frac{1}{t}} \approx \frac{\Theta(1)}{\log(T)}$$

Take away

- Convergence/implicit bias theory still holds in the EoS regime
- However, only for logistic loss not for exponential loss
- New analysis ideas

What's next?

- What happens in poly time?
- Population risk?
- Non-linear model? Other Algos?…