# Large Stepsizes Accelerate GD

## for Regularized Logistic Regression


Jingfeng Wu

with Pierre Marion and Peter Bartlett


NEURAL INFORMATION PROCESSING SYSTEMS

# Gradient descent

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

"GD ≈ discrete time gradient flow"

$$\mathrm{d}w = -\nabla L(w)\mathrm{d}t \quad \Rightarrow \quad \mathrm{d}L(w) = \nabla L(w)^\top \mathrm{d}w$$

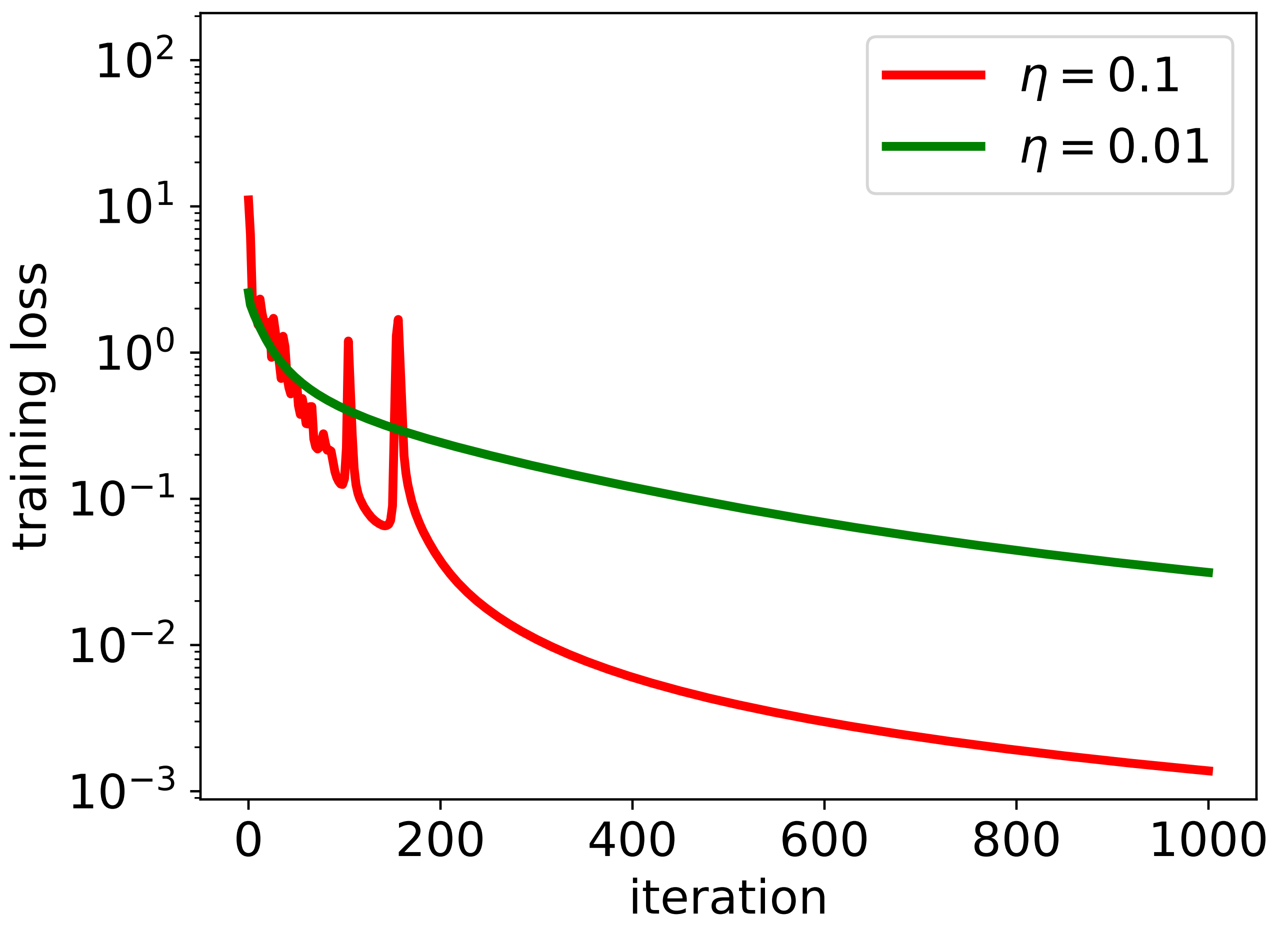
$$= -\|\nabla L(w)\|^2 \mathrm{d}t$$

$$\Rightarrow \quad L(w) \downarrow$$

small $\eta$ $\Rightarrow$ convex optimization theory

Cauchy, 1847

# Experiment (3-layer net, MNIST)




large stepsize is

- unstable
- but faster

this work analyzes benefits of large stepsize

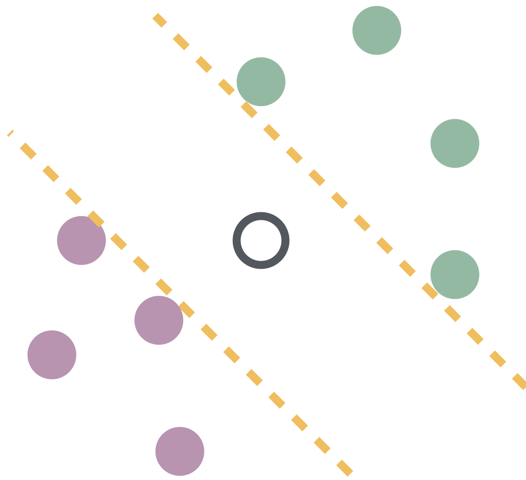
# Regularized logistic regression

$$\tilde{L}(w) = L(w) + \frac{\lambda}{2}\|w\|^2 \qquad L(w) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i x_i^\top w)\right)$$

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

**Assumption** (bounded + separable)

- $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1,\ldots,n$
- $\exists$ unit vector $w^*$, $\min_i y_i x_i^\top w^* \ge \Theta(1)$



"almost surely" separable when overparameterized

# Classical theory prediction

- $\Theta(1)$-smooth
- $\lambda$-strongly convex
- condition number $\kappa = \Theta(1/\lambda)$
- finite minimizer $w_\lambda$, $\|w_\lambda\| = O(\ln(1/\lambda))$

**Classical theory**

For $\eta = \Theta(1)$, $\tilde{L}(w_t) \downarrow$ and

$$\tilde{L}(w_t) - \min \tilde{L} \leq \epsilon \text{ for } t = \tilde{O}(1/\lambda)$$

improved to $\tilde{O}(1/\lambda^{1/2})$ by Nesterov

# Theorem (small $\lambda$)

$\eta_{\max} = \Theta(1/\lambda^{1/2})$

Assume separability and

$$\lambda \le \Theta\left(\frac{1}{n \ln n}\right) \quad \eta \le \Theta\left( \min\left\{\frac{1}{\lambda^{1/2}}, \frac{1}{n\lambda}\right\}\right)$$

**Phase transition.** GD exists unstable phase in $\tau$ steps for

$$\tau := \max\{\eta, n, n/\eta \ln(n/\eta)\}$$

$\tau = \Theta(1/\lambda^{1/2})$

**Stable phase.** From $\tau$ and onward

$$\tilde{L}(w_{\tau+t}) - \min \tilde{L} \lesssim \exp(-\lambda\eta t)$$

$t = \Theta(\ln(1/\epsilon)/\lambda^{1/2})$

for small $\lambda$, large stepsize GD matches Nesterov

# Theorem (general $\lambda$)

Assume separability and

$\eta_{\max} = \Theta(1/\lambda^{1/3})$

$$\lambda \le \Theta(1), \quad \eta \le \Theta(1/\lambda^{1/3})$$

**Phase transition.** GD exists unstable phase in $\tau$ steps for

$$\tau := \Theta(\eta^2)$$

$\tau = \Theta(1/\lambda^{2/3})$

**Stable phase.** From $\tau$ and onward

$$\tilde{L}(w_{\tau+t}) - \min \tilde{L} \lesssim \exp(-\lambda\eta t)$$

$t = \Theta(\ln(1/\epsilon)/\lambda^{2/3})$

for general $\lambda$, large stepsize is faster than small stepsize

$\tilde{O}(1/\lambda^{2/3})$ $\tilde{O}(1/\lambda)$

# Margin-based generalization

Assume $(x_i, y_i)_{i=1}^n$ are iid copies of $(x, y)$, where a.s.

- $\|x\| \leq 1,\ y \in \{\pm 1\}$
- $\exists$ unit vector $w^*$, $yx^\top w^* \geq \Theta(1)$

**Corollary.** The best known test error upper bound is $\tilde{O}(1/n)$. To get $\tilde{O}(1/n)$ rate, GD takes

- $O(n)$ steps with $\lambda = 0$ and $\eta = \Theta(1)$
- $O(n)$ steps with $\lambda = 1/n$ and $\eta = 1$
- $\tilde{O}(n^{2/3})$ steps with $\lambda = 1/n$ and $\eta = \Theta(n^{1/3})$

large stepsize accelerates GD without overfitting

# Stepsize diagram