Reimagining Gradient Descent

Large Stepsize, Oscillation, Acceleration

Jingfeng Wu

Berkeley
UNIVERSITY OF CALIFORNIA

Gradient descent

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

Cauchy, 1847

"GD ≈ discrete time gradient flow"

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

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

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

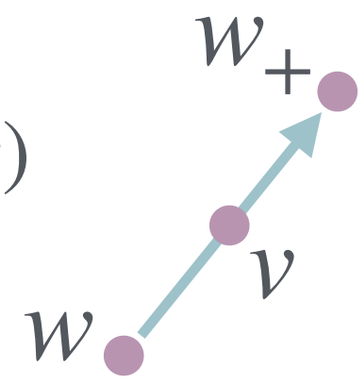
because of stepsize, GD ≠ discrete time flow

Small stepsize for stability

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

$$= L(w) - \eta \|\nabla L(w)\|^2 + \frac{\eta^2}{2} \nabla L(w)^\top \nabla^2 L(v) \nabla L(w)$$

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



$$\eta < \frac{2}{\sup \|\nabla^2 L(\,\cdot\,)\|}$$

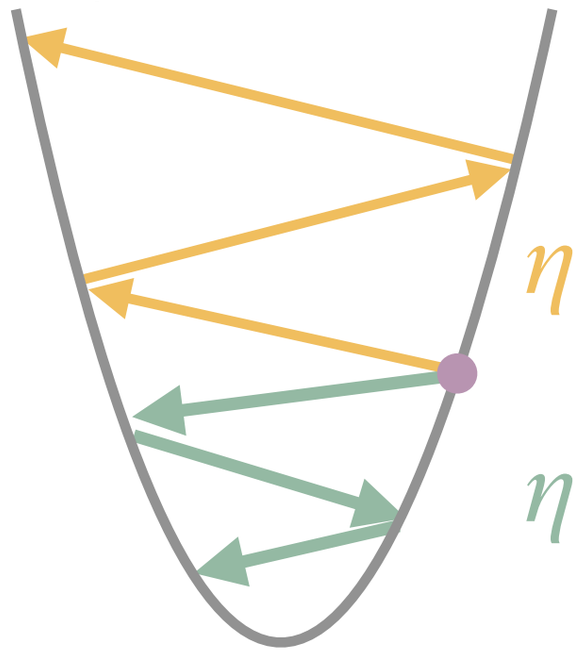
Descent lemma:

for small η, $L(w_t)$ decreases monotonically

for large η, $L(w_t)$ diverges in "bad" cases

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

$$w_+ = (1 - 2\eta) w$$



Classical theory

Let L be 1-smooth with a finite minimizer w^*. For GD with $\eta = 1$,

descent lemma $\quad L(w_t) \downarrow$

convexity $\quad L(w_t) - \min L \leq \dfrac{\|w_0 - w^*\|^2}{2t}$

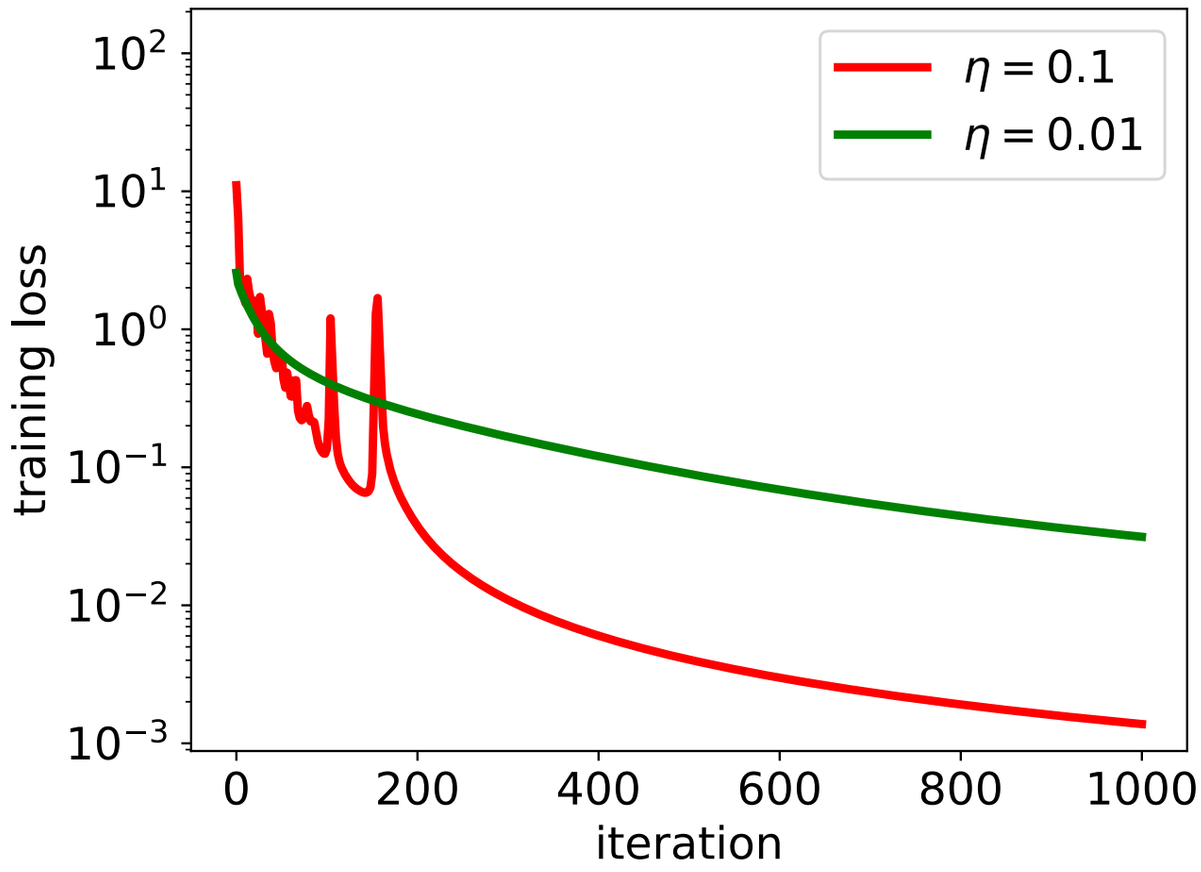
α-strong convexity $\quad L(w_t) - \min L \leq e^{-\alpha t}(L(w_0) - \min L)$

Nesterov's momentum accelerates GD to

$$O\left(1/t^2\right) \text{ and } O\left(e^{-\sqrt{\alpha}t}\right)$$

these are minimax optimal among first-order methods

Experiment (3-layer net, MNIST)



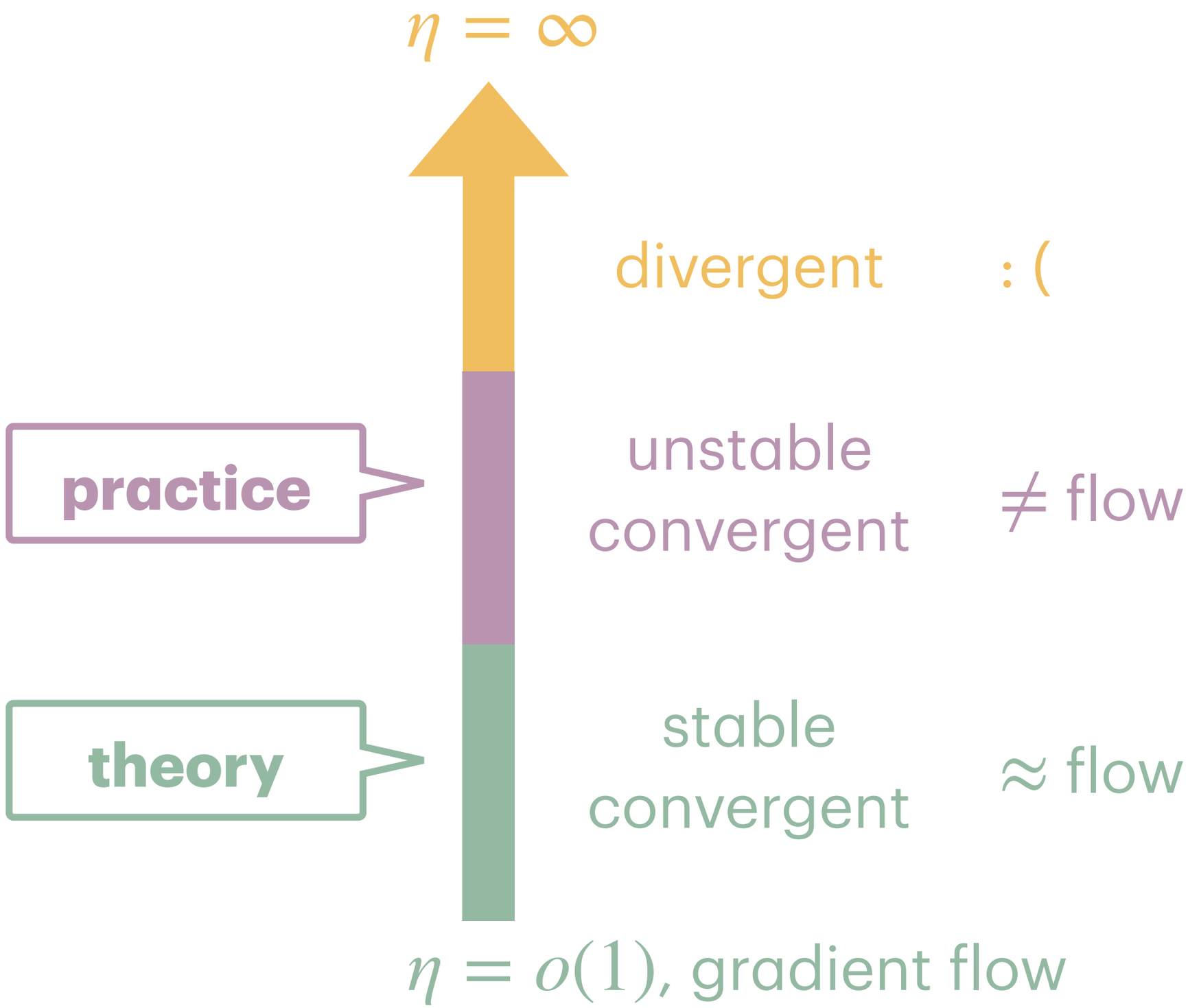
large stepsize is

- unstable
- but faster

"edge of stability"

Cohen, Kaur, Li, Kolter, and Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021

Stepsize?


$\eta = \infty$
divergent : (
practice
unstable
convergent
$\neq$ flow
theory
stable
convergent
$\approx$ flow
$\eta = o(1)$, gradient flow

(1/3) Seeking “simplest” answer

linear regression

logistic regression

.....

deep learning

unstable convergence impossible

observable & provable

unstable convergence observed

Peter Bartlett

Matus Telgarsky

Bin Yu

Wu, Bartlett*, Telgarsky*, and Yu*. “Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency.” COLT 2024

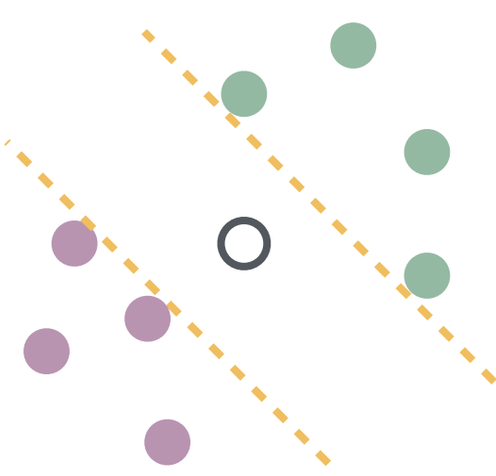
Logistic regression

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

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

Assumption (bounded + separable)

- $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1, \ldots, n$
- $\exists$ unit vector $w^*,\ \min_i y_i x_i^\top w^* \ge \gamma > 0$

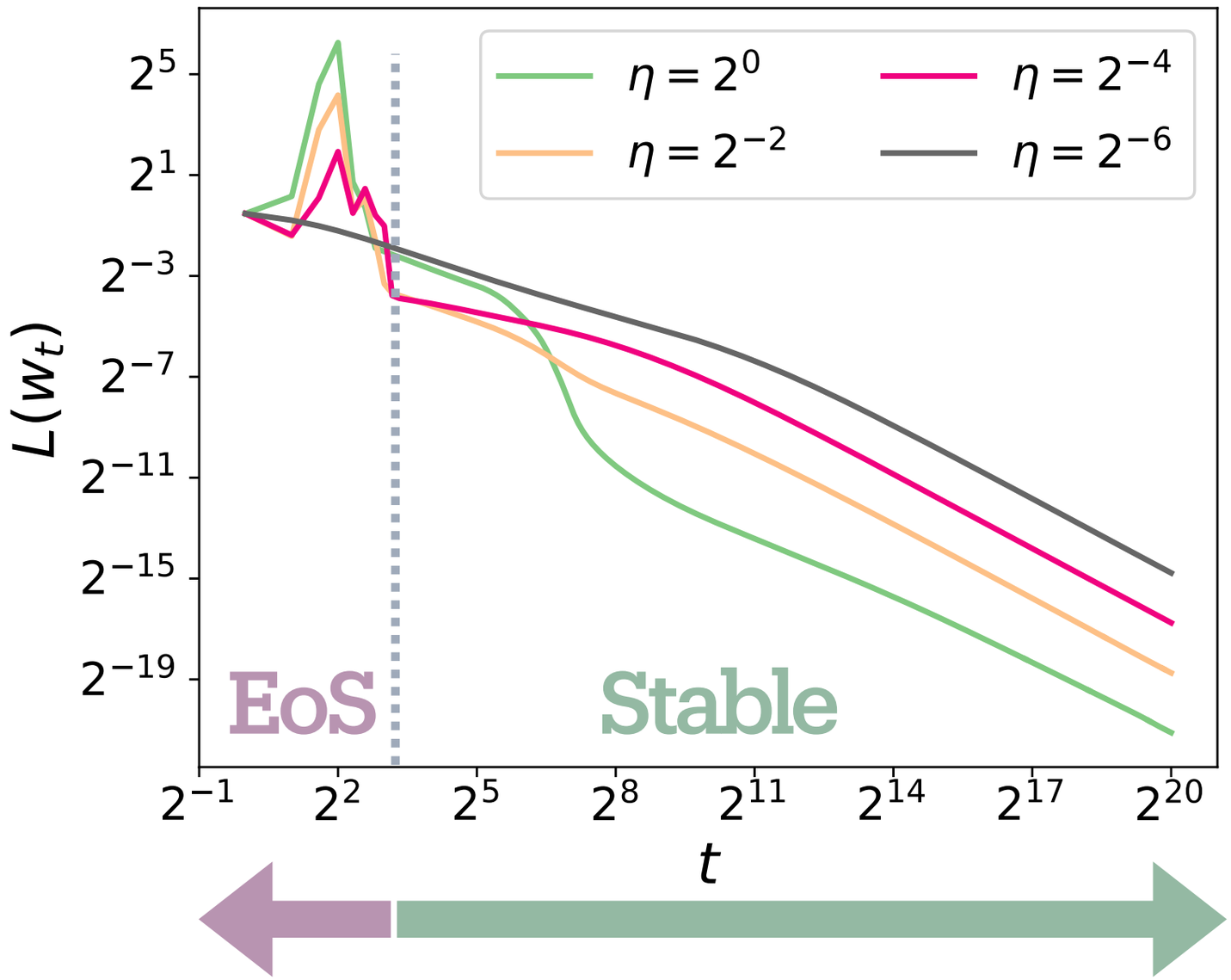


Classical theory

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

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

MNIST “0” vs “8”



Stable phase: $L(w_t) \downarrow$ from t and onwards
EoS phase: otherwise

Theorem

Phase transition. GD exists EoS in τ steps for

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

Stable phase. From τ and onwards

$$L(w_{\tau+t}) = \tilde{O}\left(\frac{1}{\eta t}\right)$$

1. Convergence for **every** η
2. Large η: faster in stable phase but stays longer in EoS
3. Given #steps $T \geq \Theta(n)$, if choose $\eta = \Theta(T)$, then

$$\tau \leq T/2 \ \text{ and } \ L(w_T) = \tilde{O}(1/T^2)$$

acceleration by large stepsize

A “non-quadratic” picture

$\exists$ unit vector w^*, $\min_i y_i x_i^\top w^* > \gamma > 0$

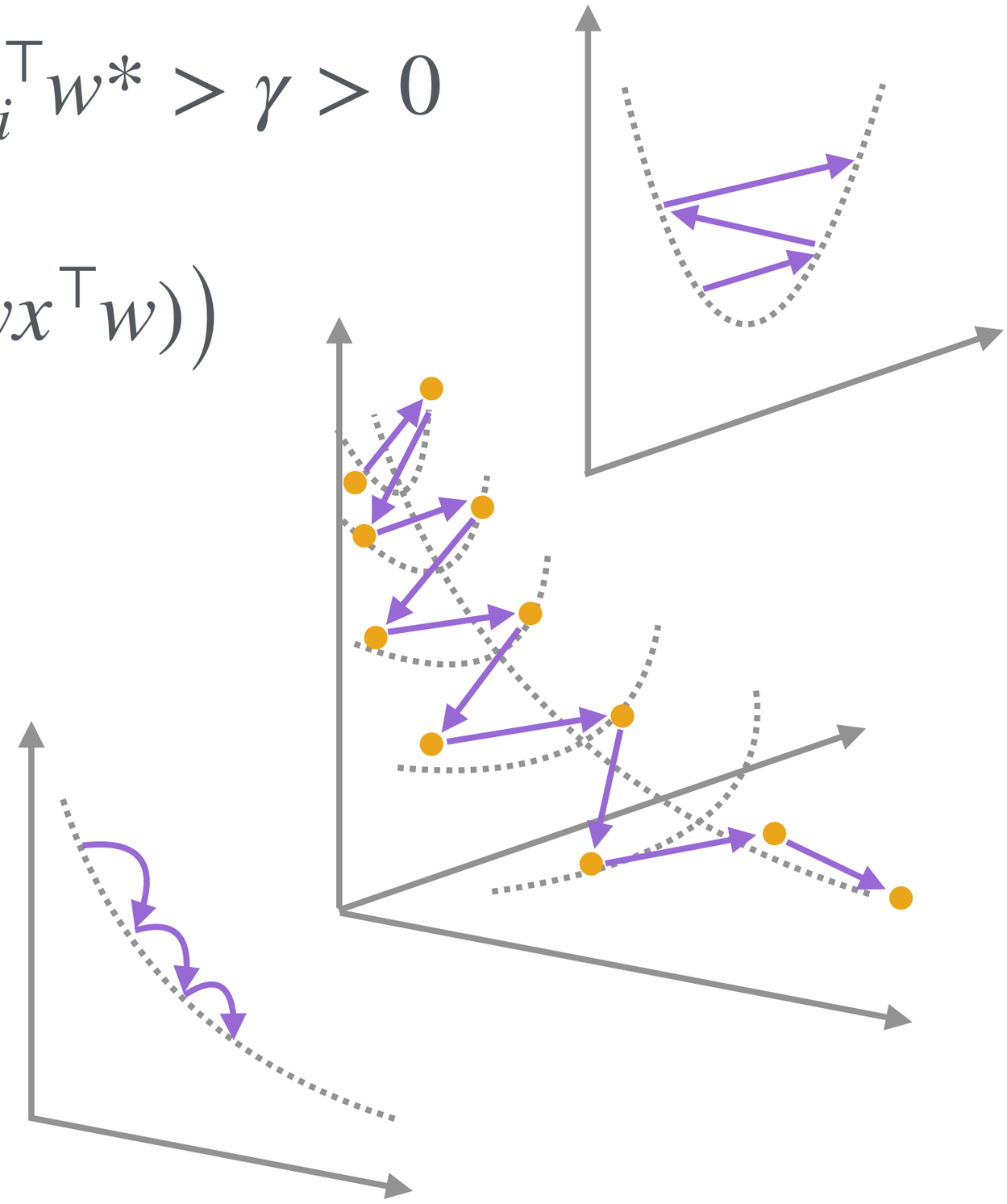
$$L(w) = \hat{\mathbb{E}} \ln\left(1 + \exp(-yx^\top w)\right)$$

minimizer at ∞

$$\lim_{\lambda \to \infty} L(\lambda w^*) = 0$$

self-bounded

$$\|\nabla^2 L\| \le L$$



Proof

$$\|w_{t+1} - u\|^2 = \|w_t - u\|^2 + 2\eta\langle \nabla L(w_t), u - w_t\rangle + \eta^2\|\nabla L(w_t)\|^2$$

$$= \|w_t - u\|^2 + 2\eta\langle \nabla L(w_t), u_1 - w_t\rangle$$

$$+\eta^2\Big(\langle \nabla L(w_t), 2u_2/\eta\rangle + \|\nabla L(w_t)\|^2\Big)$$

$\langle \nabla L(w), w^*\rangle < 0$ => ≤ 0 if $u_2 = w^* \cdot \Theta(\eta)$

$\|\nabla L(w)\| \leq 1$

$$\leq \|w_t - u\|^2 + 2\eta\langle \nabla L(w_t), u_1 - w_t\rangle$$

$$\leq \|w_t - u\|^2 + 2\eta\Big(L(u_1) - L(w_t)\Big)$$

Telescoping the sum...

Two extensions

minimizer at ∞

$$\lim_{\lambda \to \infty} L(\lambda w^*) = 0$$

finite minimizer

e.g. regularization

unstable convergence under finite minimizer

self-bounded

$$\|\nabla^2 L\| \leq L$$

enabling “tricks”

e.g. adaptive GD
[Ji & Telgarsky 2021]

large stepsizes for GD variants

(2/3) Large stepsize for adaptive GD

self-bounded

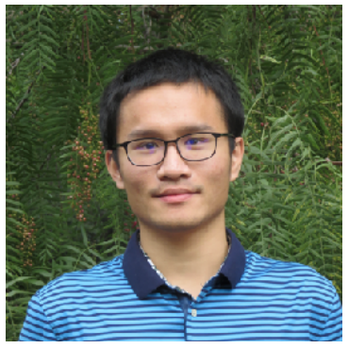
$$\|\nabla^2 L\| \leq L$$

enabling "tricks"

e.g. adaptive GD
[Ji & Telgarsky 2021]

large stepsizes for GD variants

Ruiqi Zhang



Licong Lin

Peter Bartlett

Zhang, **Wu**, Lin, Bartlett. "Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes." ICML 2025

Adaptive GD

$$L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i x_i^\top w) \qquad \ell(t) = \ln\Big(1 + \exp(-t)\Big)$$

$$w_{t+1} = w_t - \eta\Big((-\ell^{-1})' \circ L(w_t)\Big)\nabla L(w_t)$$

$$\approx w_t - \frac{\eta}{L(w_t)}\nabla L(w_t)$$

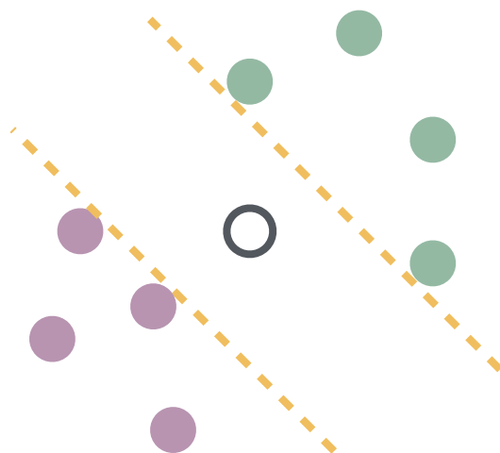
$$w_{t+1} = w_t - \eta \nabla \phi(w_t) \qquad \phi(w) = -\ell^{-1}(L(w))$$

$$\approx \ln \sum \exp(-y_i x_i^\top w)$$

[Ji & Telgarsky, 2021]

For $\eta = \Theta(1)$, $L(w_t) \downarrow$ and $L(w_t) \le \exp(-\Theta(t))$

large stepsize makes adaptive GD even faster



Theorem

Assume separability with margin γ. For $t \geq 1/\gamma^2$, we have

$$L(\bar{w}_t) \leq \exp\Big(-\Theta(\gamma^2 \eta t)\Big), \quad \text{where } \bar{w}_t = \frac{1}{t}\sum_{k=1}^{t} w_k$$

1. Arbitrarily small error in $1/\gamma^2$ steps

$$\lim_{\eta \to \infty} L(\bar{w}_t) = 0 \quad \text{for} \quad t = 1/\gamma^2$$

2. Averaged iterate, no “stable phase”

3. small < large < small adaptive << large adaptive

$\tilde{O}(1/\epsilon)$ $\tilde{O}(1/\epsilon^{1/2})$ $O(\ln(1/\epsilon))$ $O(1)$

Theorem (lower bound)

$\forall w_0$, $\exists\ (x_i, y_i)_{i=1}^n$ with margin γ such that: for any first-order batch method

$$\min_i y_i x_i^\top w_t > 0 \quad \Rightarrow \quad t \geq \Omega(1/\gamma^2)$$

First-order batch method:

$$w_t \in w_0 + \text{span}\{\nabla L(w_0), \ldots, \nabla L(w_{t-1})\}$$

where $L(w) = \hat{\mathbb{E}}\ell(yx^\top w)$ for any ℓ

adaptive GD + large stepsize = minimax optimal

(3/3) Large stepsize under finite minimizer

minimizer at ∞

$$\lim_{\lambda \to \infty} L(\lambda w^*) = 0$$

finite minimizer

e.g. regularization

unstable convergence under finite minimizer

Pierre Marion

Peter Bartlett

Wu*, Marion*, and Bartlett. “Large stepsizes accelerate gradient descent for regularized logistic regression.” arXiv:2506.02336

Regularized logistic regression

$$\tilde{L}(w) = L(w) + \frac{\lambda}{2}\|w\|^2 \qquad L(w) = \frac{1}{n}\sum_i \ell(y_i x_i^\top w)$$

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

λ-strongly convex, $\Theta(1)$-smooth, $\kappa = \Theta(1/\lambda)$

finite minimizer w_λ, $\|w_\lambda\| = O(\ln(1/\lambda))$

Classical theory

$\tilde{O}(1/\lambda)$

For $\eta = \Theta(1)$, $\tilde{L}(w_t) \downarrow$ and $\tilde{L}(w_t) - \min \tilde{L} \le \epsilon$ for $t = O(\kappa \ln(1/\epsilon))$

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

Theorem (small λ)

Assume separability and

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

Phase transition. GD exists EoS in τ steps for

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

Stable phase. From τ and onward

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

for small λ, large stepsize GD matches Nesterov

Theorem (general λ)

Assume separability and

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

Phase transition. GD exists EoS in τ steps for

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

Stable phase. From τ and onward

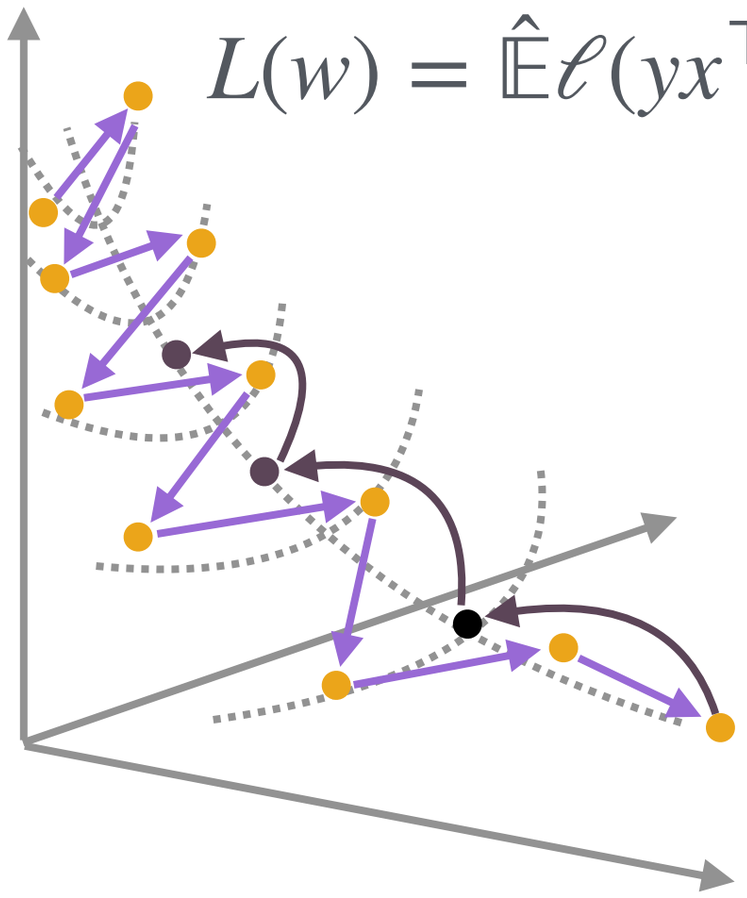
$$\tilde{L}(w_{\tau+t}) - \min \tilde{L} \lesssim \exp(-\lambda\eta t)$$

for general λ, large stepsizes is faster than small stepsizes

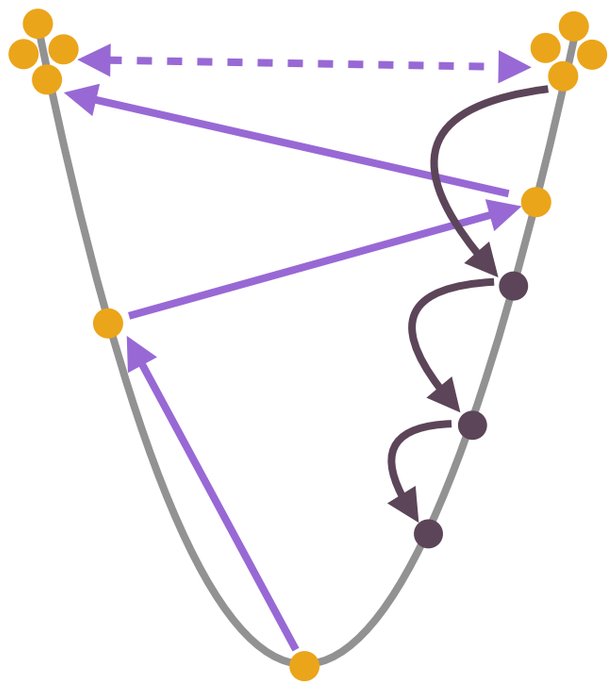
$\tilde{O}(1/\lambda^{2/3})$ $\tilde{O}(1/\lambda)$

A new picture

$$L(w) = \hat{\mathbb{E}}\ell(yx^\top w)$$



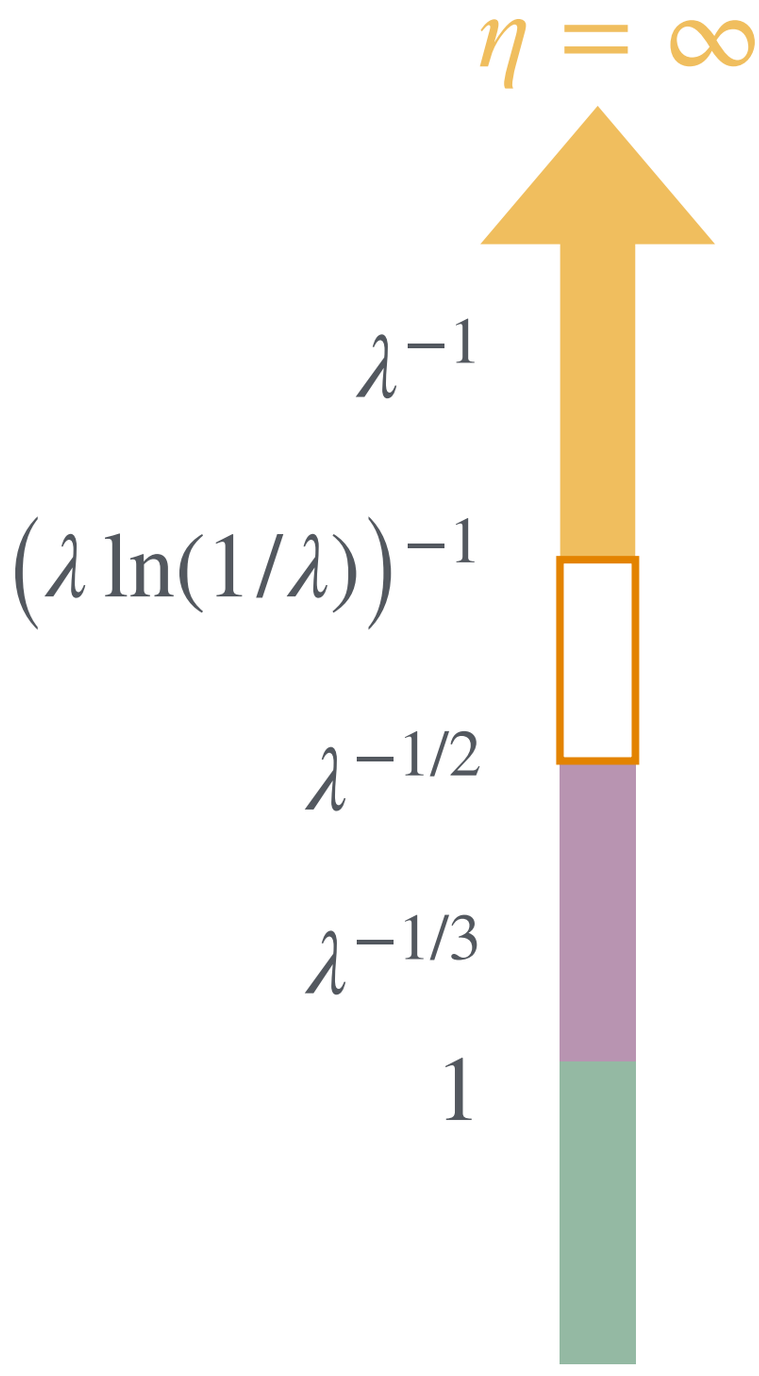
$$R(w) = \frac{\lambda}{2}\|w\|^2$$



EoS. $\tilde{L} \approx L, R \leq \Theta(1)$, "overshoot" $\|w_\lambda\| = O(\ln(1/\lambda))$

Stable. "move back" $\sup \|w_t\| = \Theta(\eta) = \text{poly}(1/\lambda)$

Stepsize diagram



divergent

locally convergent

unstable convergent

stable convergent

generic support vectors

unknown global behavior

match Nesterov

sample size independent

(4/3) More results

- generalization
- SGD
- other loss functions
- networks in kernel regime
- two-layer networks with linear teacher
- implicit bias

Wu, Bartlett*, Telgarsky*, and Yu*. “Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency.” COLT 2024

Zhang, **Wu**, Lin, Bartlett. “Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes.” ICML 2025

Wu*, Marion*, and Bartlett. “Large stepsizes accelerate gradient descent for regularized logistic regression.” arXiv:2506.02336

————————

Cai, **Wu**, Mei, Lindsey, and Bartlett. “Large stepsize GD for non-homogeneous two-layer networks: margin improvement and fast optimization.” NeurIPS 2024

Cai*, Zhou*, **Wu**, Mei, Lindsey, and Bartlett. “Implicit bias of gradient descent for non-homogeneous deep networks.” ICML 2025

Contribution

Provable unstable convergence in three cases

Next: a general theory?