# Reimagining Gradient Descent

## Large Stepsize, Oscillation, Acceleration

Jingfeng Wu

Simons Institute for the Theory of Computing

Berkeley University of California

# Gradient descent

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

Cauchy, 1847

"GD ≈ discrete time gradient flow"

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

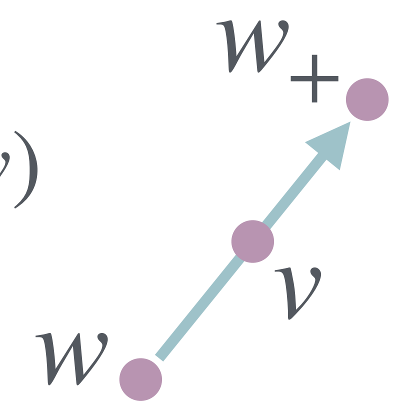
$$= -\|\nabla L(w)\|^2 \mathrm{d}t$$

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

*how to select stepsize?*

# Small stepsize for stability

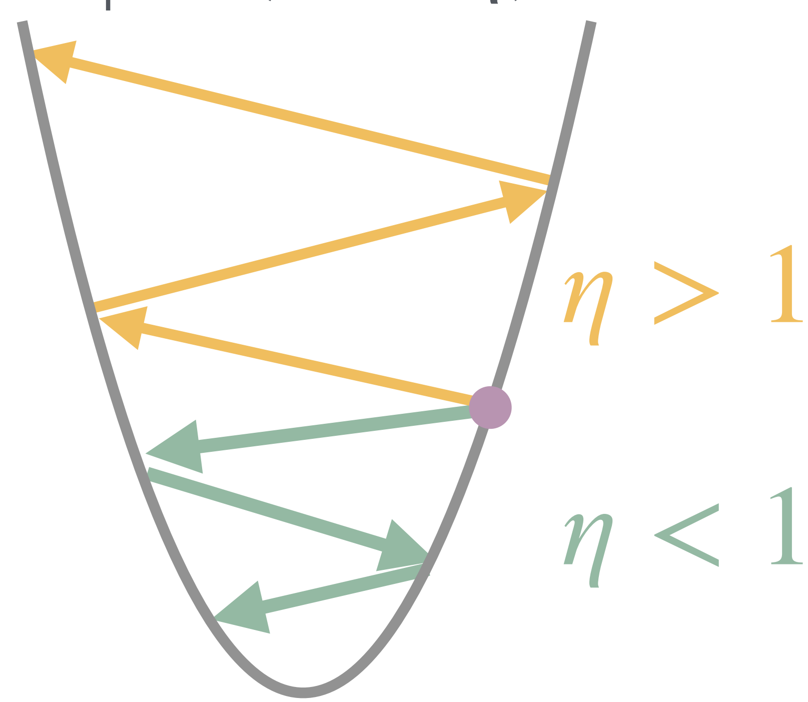
$$
\begin{aligned}
L(w_+) &= L(w - \eta \nabla L(w)) \\
&= L(w) - \eta \|\nabla L(w)\|^2 + \frac{\eta^2}{2} \nabla L(w)^\top \nabla^2 L(v) \nabla L(w) \\
&\leq L(w) - \eta \|\nabla L(w)\|^2 \left( \boxed{1 - \frac{\eta}{2} \|\nabla^2 L(v)\|_2} \right)
\end{aligned}
$$




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

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

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




**Descent lemma**:

for small $\eta$, $L(w_t)$ decreases monotonically

for large $\eta$, $L(w_t)$ diverges in “bad” cases

# Classical theory

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

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

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

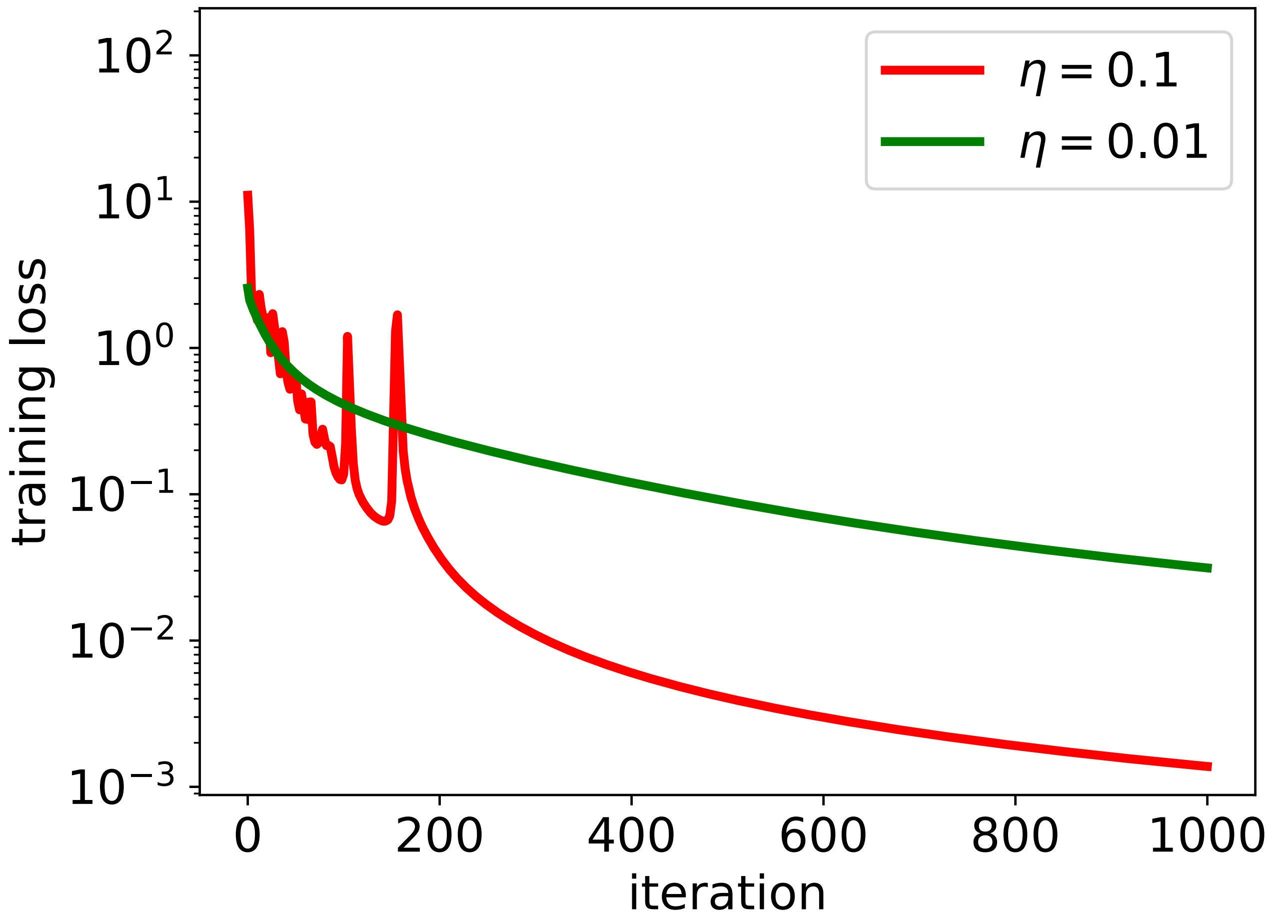
$\alpha$**-strong convexity** $L(w_t) - \min L \leq e^{-\alpha t}(L(w_0) - \min L)$

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)



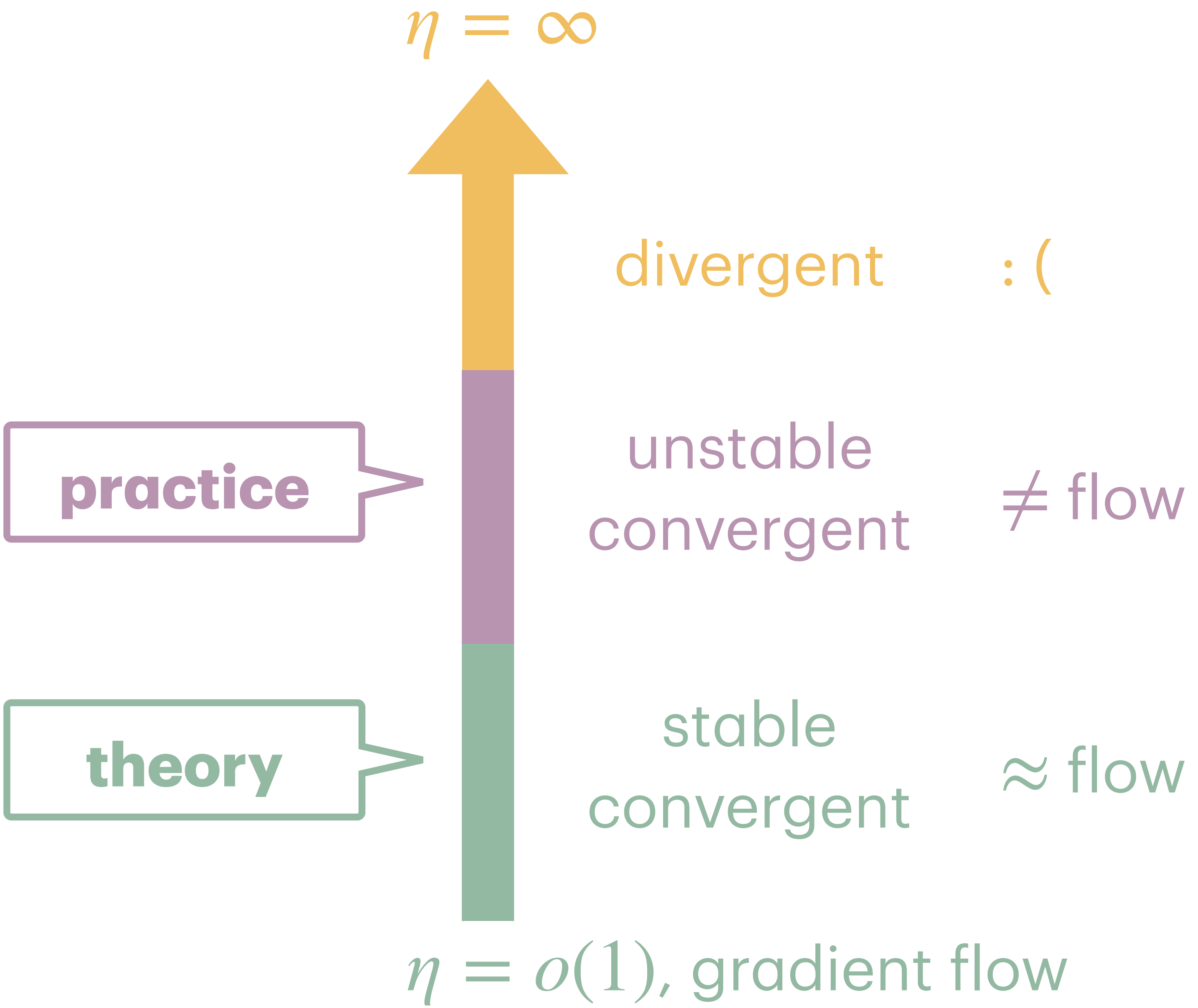

large stepsize is

- unstable
- but faster

"edge of stability"

Cohen, Kaur, Li, Kolter, 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*, Yu*. "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency." COLT 2024

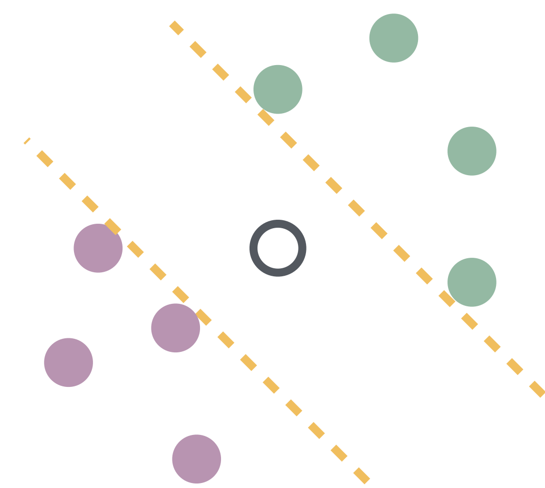
# Logistic regression

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

smooth, convex
non-strongly convex

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

**Assumption** (bounded + separable)



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

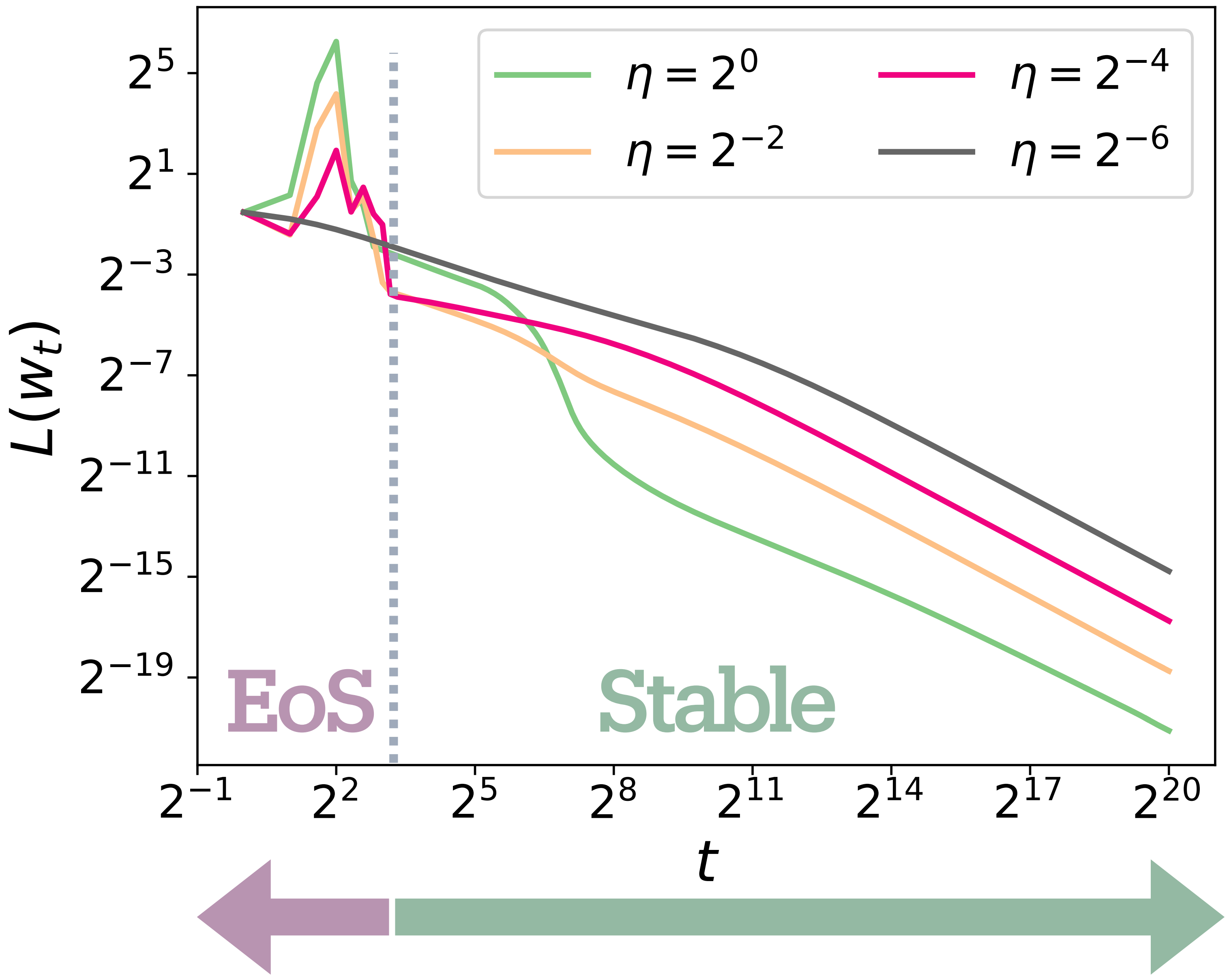
"almost surely" when overparameterized

**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 $\tau$ steps for

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

$\tau = \Theta(\eta)$

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

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

"flow rate"

1. Convergence for **every** $\eta$
2. Large $\eta$: 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 $w^*$, $\min_i y_i x_i^\top w^* > \gamma > 0$

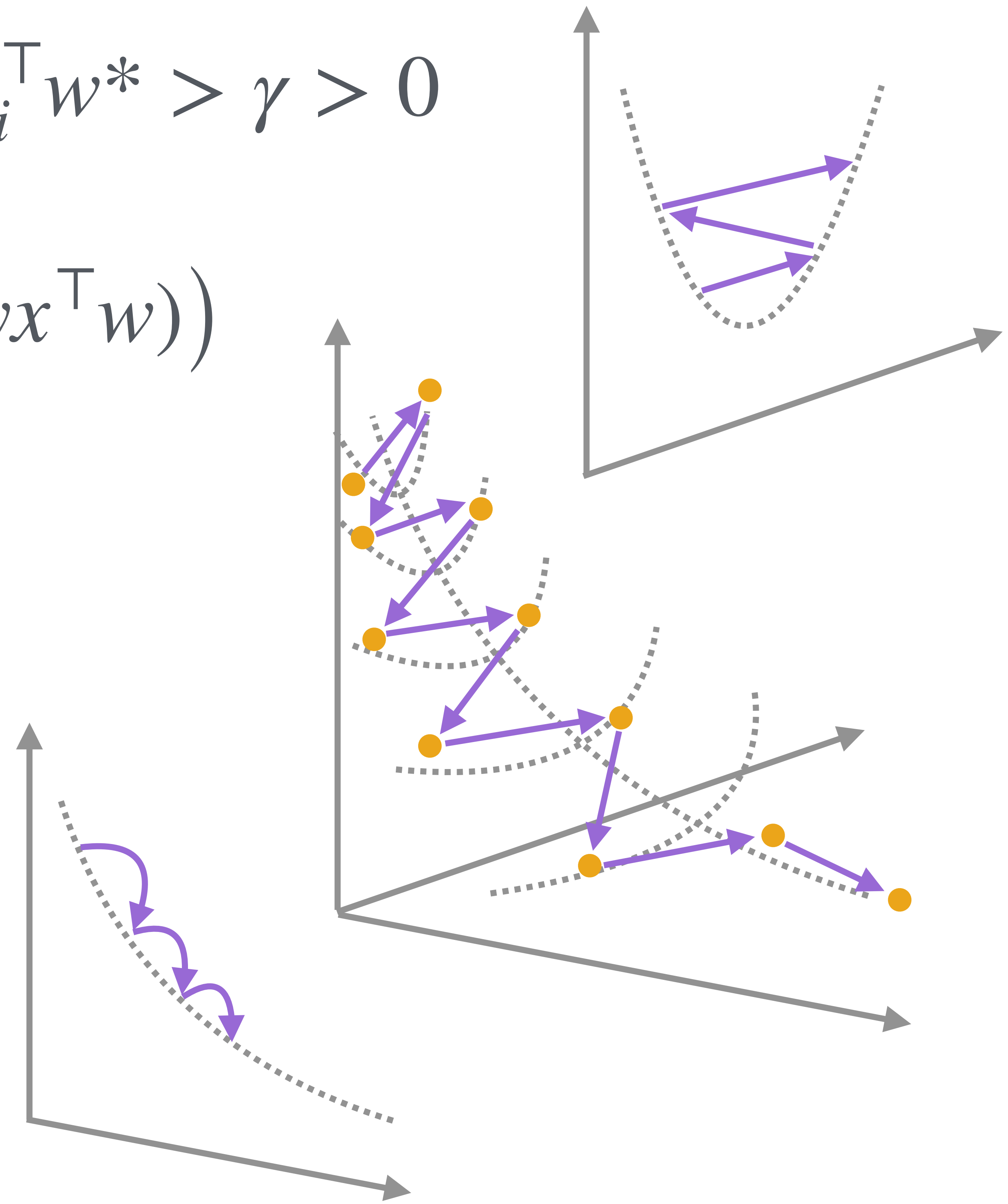
$$L(w) = \hat{\mathbb{E}} \ln\big(1 + \exp(-y x^\top w)\big)$$

***minimizer at*** $\infty$

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

***self-bounded***

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



# Proof

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

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

local tells a bit about global

$$+\eta^2\left(\boxed{\langle \nabla L(w_t), 2u_2/\eta\rangle + \|\nabla L(w_t)\|^2}\right)$$

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

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

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

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

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***

Ji & Telgarsky. "Characterizing the implicit bias via a primal-dual analysis.” ALT 2021.

# (2/3) Large stepsize for adaptive GD

*self-bounded*

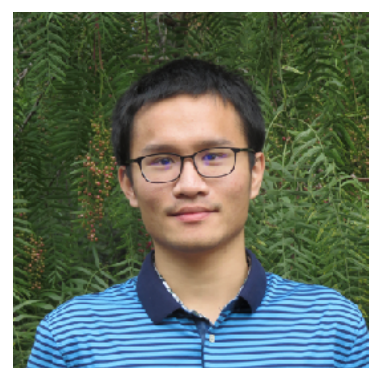
$\|\nabla^2 L\| \leq L$

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

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

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

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

adapt to curvature

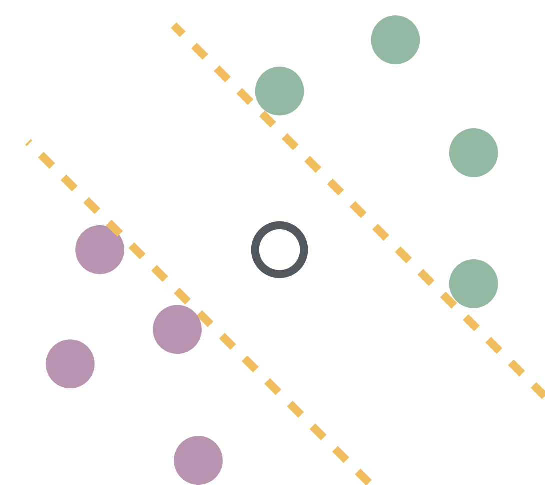
$$w_{t+1} = w_t - \eta\nabla\phi(w_t) \qquad \phi(w) = -\ell^{-1}(L(w))$$

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

**[Ji & Telgarsky, 2021]**

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

large stepsize makes adaptive GD even faster



# Theorem

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

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

$\leq \exp(-\Theta(\eta))$

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"

no more "flat" region

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 $\gamma$ such that: for any first-order batch method

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

matching "Perceptron"
[Novikoff, 1962, or earlier]

first-order batch method:

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

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

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

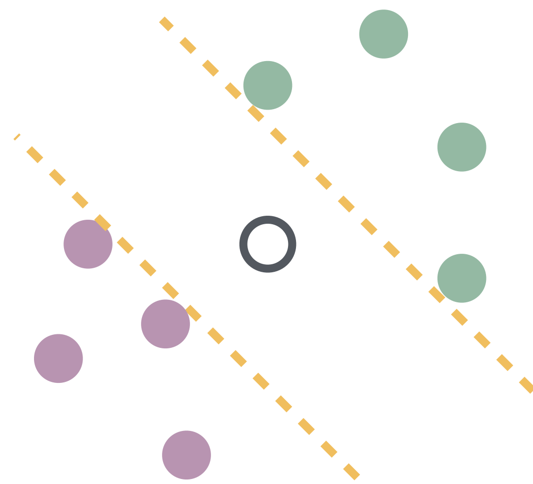
Peter Bartlett

**Wu***, Marion*, Bartlett. "Large stepsizes accelerate gradient descent for regularized logistic regression." NeurIPS 2025

# Regularized logistic regression

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

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



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

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

**Classical theory**

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

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

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 EoS 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 EoS in $\tau$ steps for

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

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

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

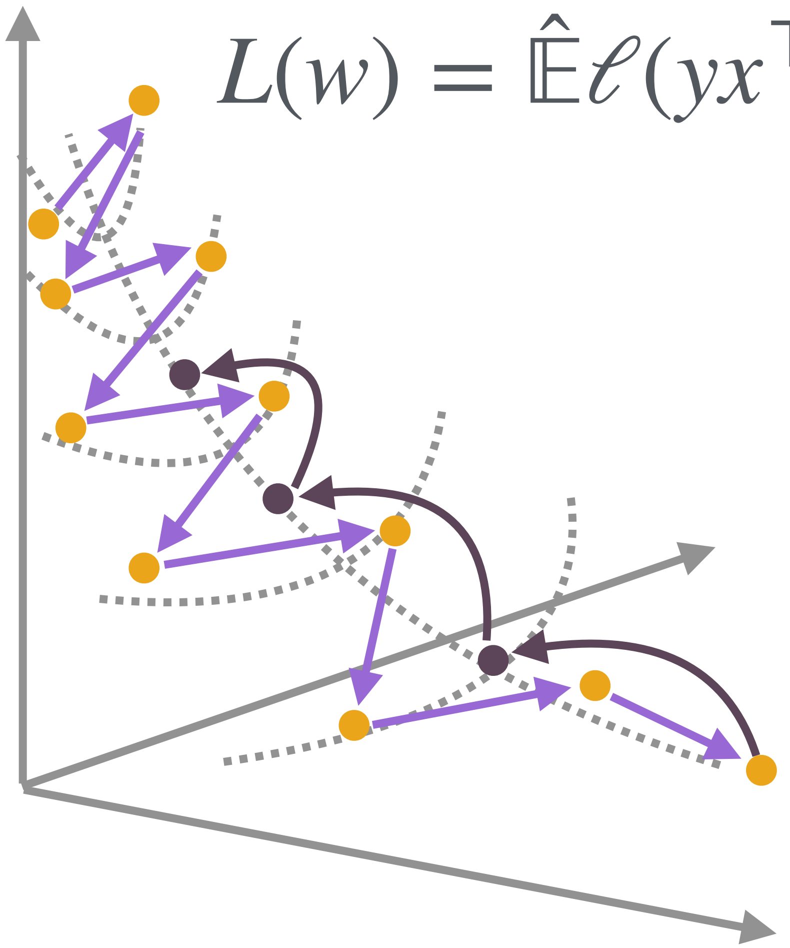
$$\tilde{L}(w_{\tau+t}) - \min \tilde{L} \lesssim \exp(-\lambda \eta t)$$

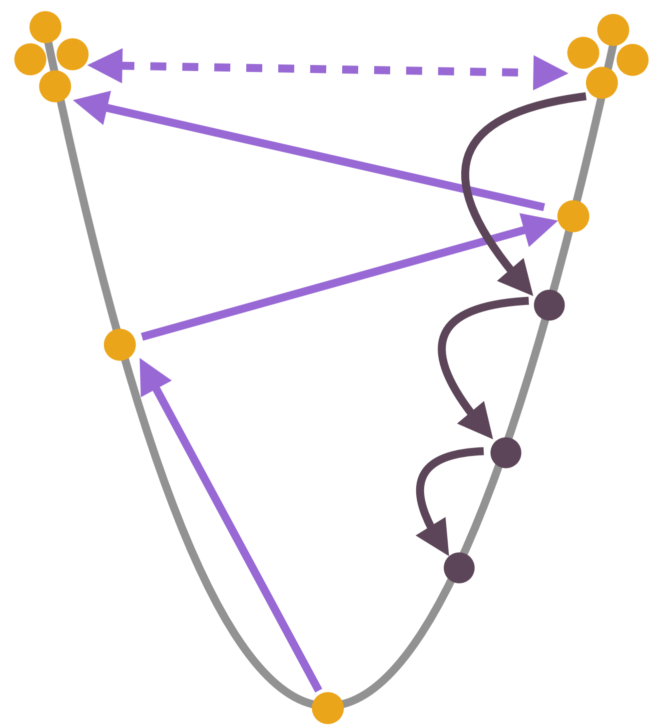
$t = \Theta(\ln(1/\epsilon)/\lambda^{2/3})$

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

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

# A new picture




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

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

# 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 $w^*$, $yx^\top w^* \geq \gamma > 0$

**[Classical fast rate]** For the test error, w.h.p.

$$L_{\text{test}}(\hat{w}) := \mathbb{E} \ln(1 + e^{-yx^\top \hat{w}}) \lesssim L(\hat{w}) + \tilde{O}(1)\frac{\max\{1, \|\hat{w}\|^2\}}{n}$$

tradeoff: fitting data vs estimator norm

Srebro, Sridharan, Tewari. “Smoothness, low noise and fast rates.” NeurIPS 2010

# Acceleration without overfitting

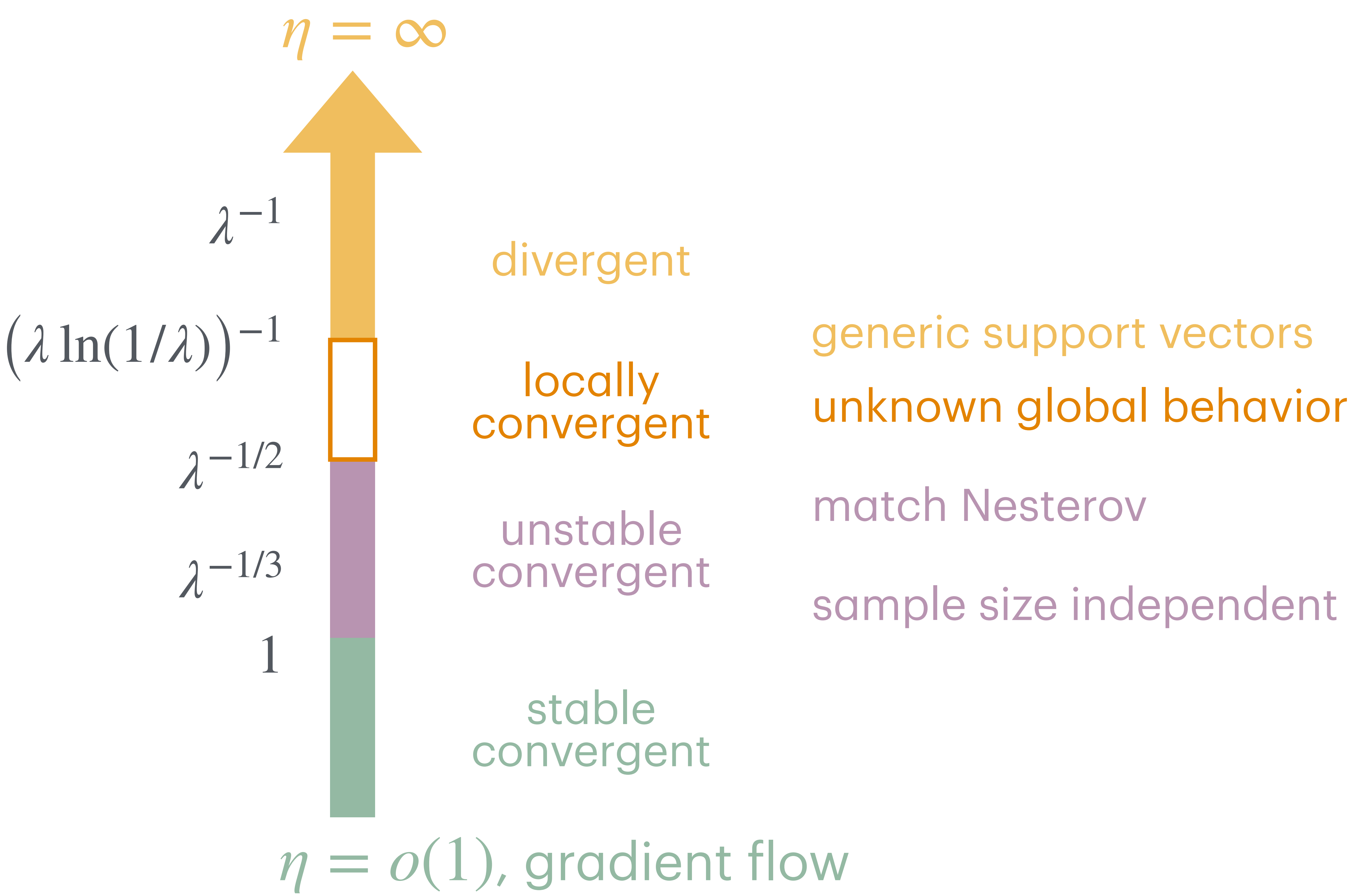
**Corollary.** ERM with $\lambda = 1/n$ gets $\tilde{O}(1/n)$ rate, minimizing the upper bound.

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



# (4/3) More large stepsizes

- other loss functions
- SGD

- networks in kernel regime
- two-layer networks with linear teacher
- implicit bias

**Wu**, Bartlett*, Telgarsky*, 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*, Bartlett. "Large stepsizes accelerate gradient descent for regularized logistic regression." NeurIPS 2025

————————

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

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

# Contribution

provable unstable convergence
in three cases

a general theory?