Towards a Less Conservative Theory of Machine Learning

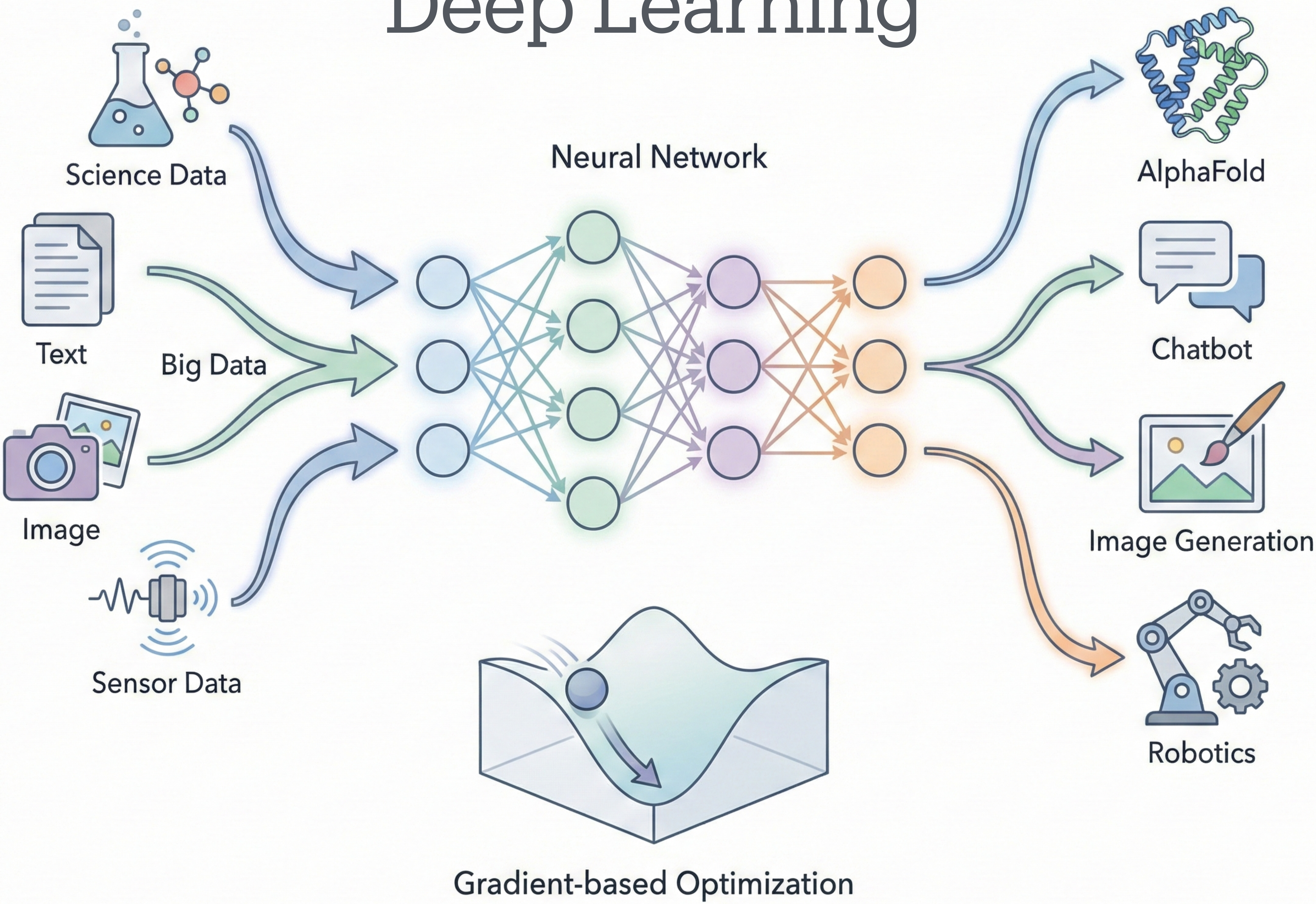
Unstable Optimization & Implicit Regularization

Jingfeng Wu

Simons Institute for the Theory of Computing

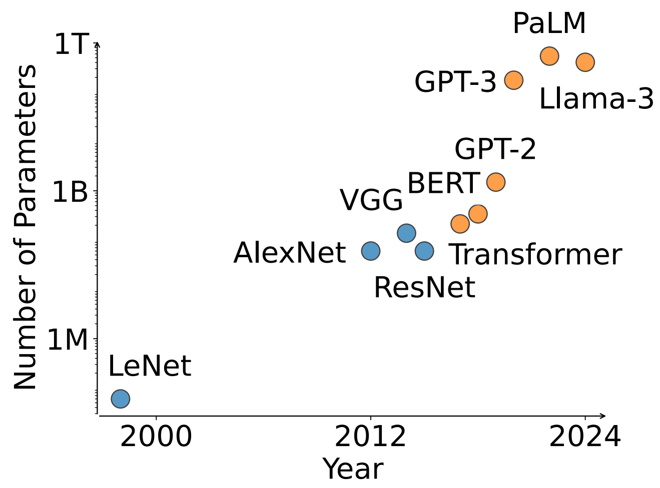
Berkeley University of California

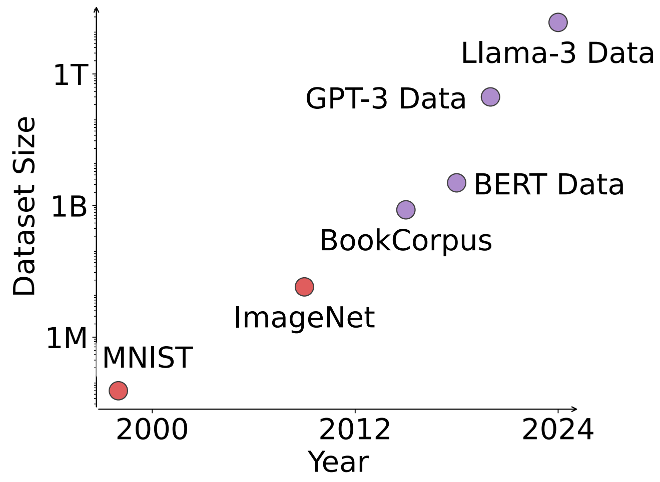
Deep Learning



What makes deep learning thrive?

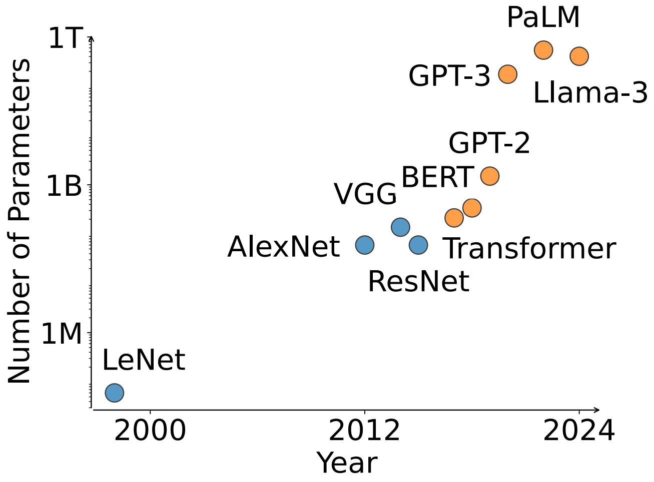
scaling

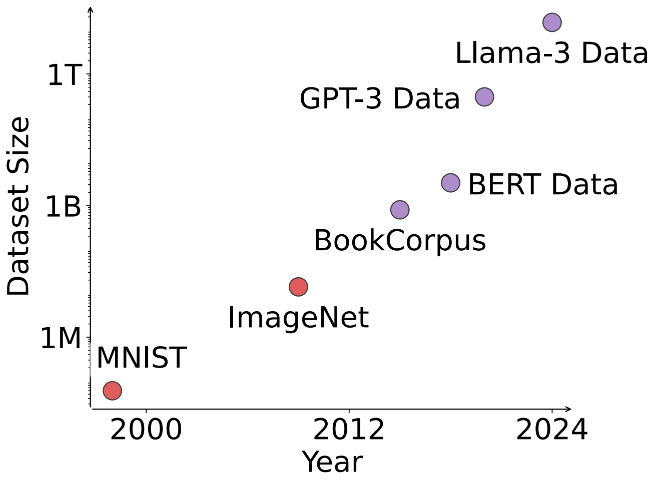




What makes deep learning thrive?

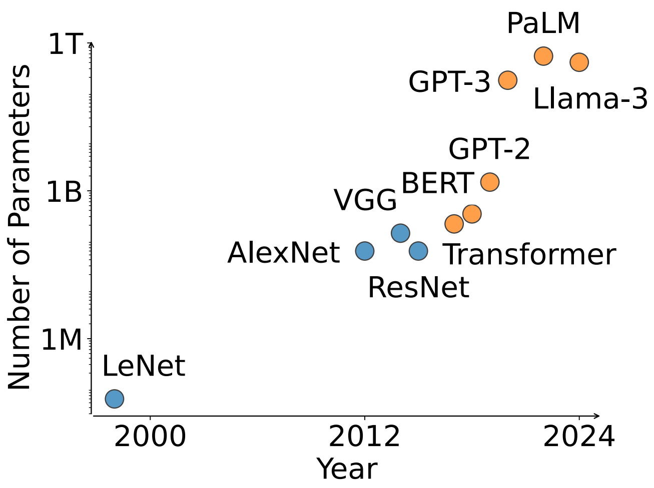
scaling	new mechanisms

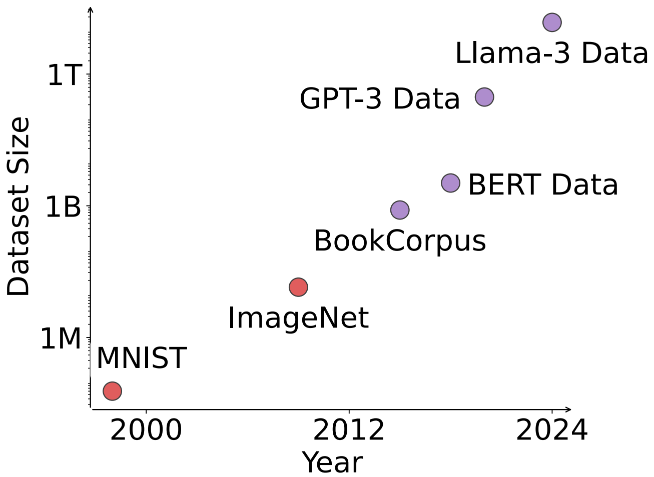




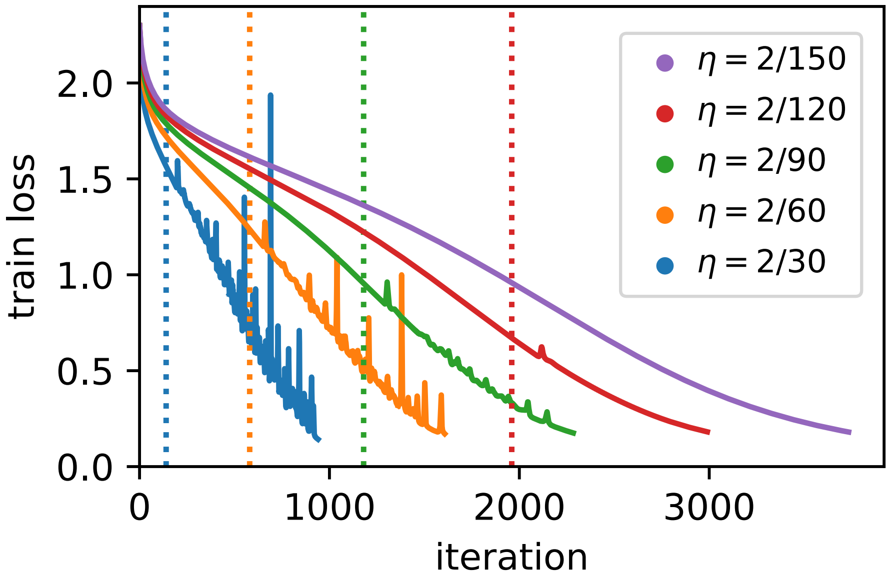
What makes deep learning thrive?

scaling	new mechanisms



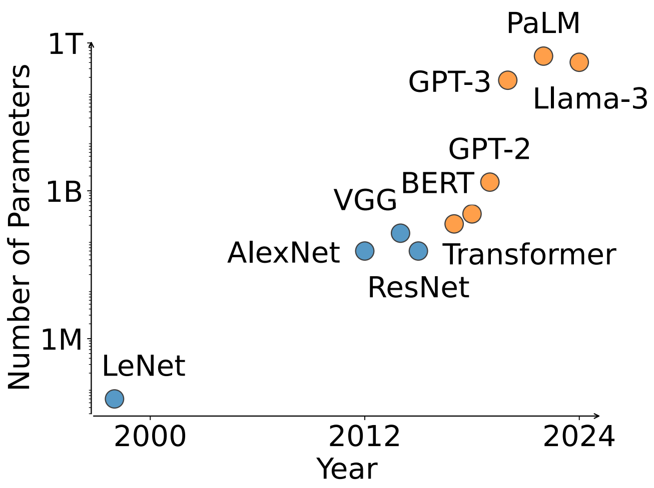


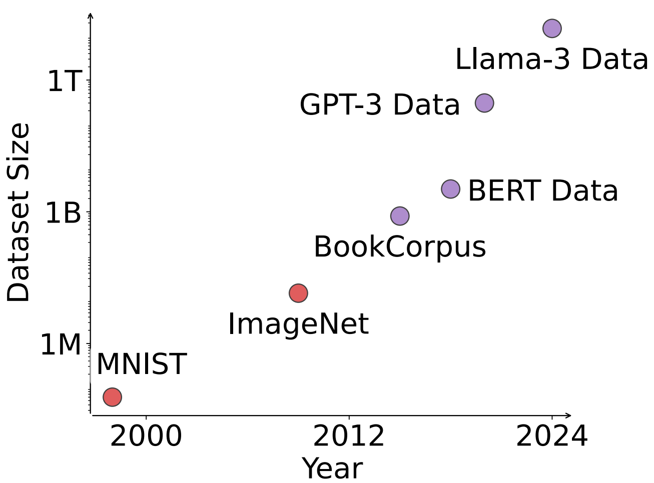
optimization



What makes deep learning thrive?

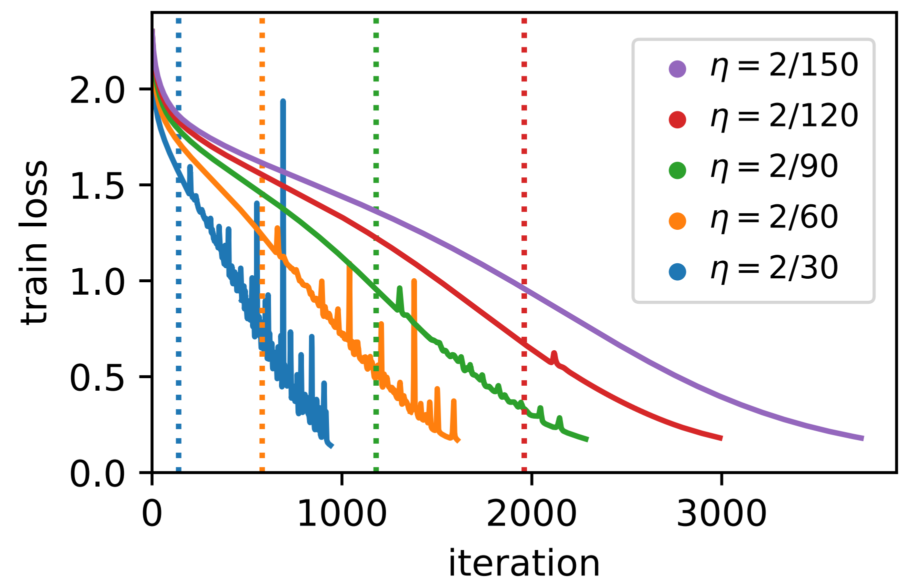
scaling	new mechanisms





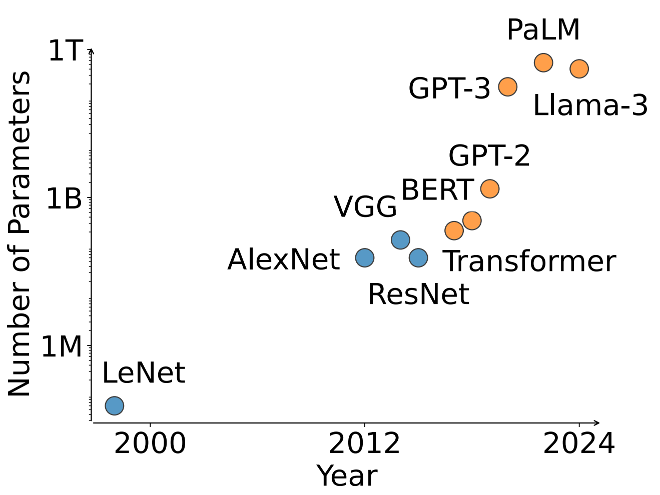
optimization

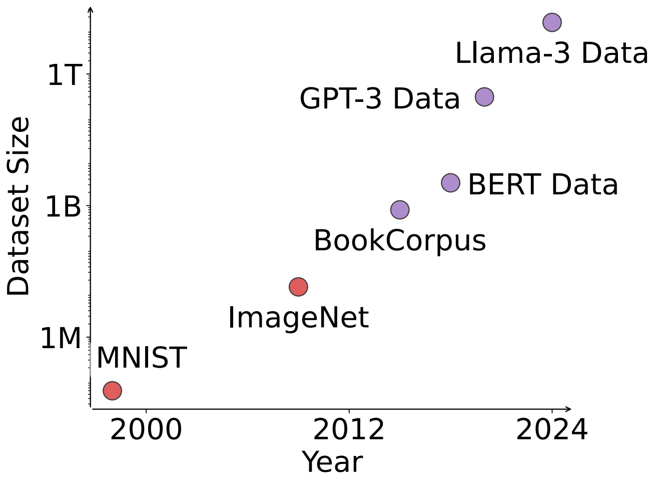
training instability



What makes deep learning thrive?

scaling	new mechanisms



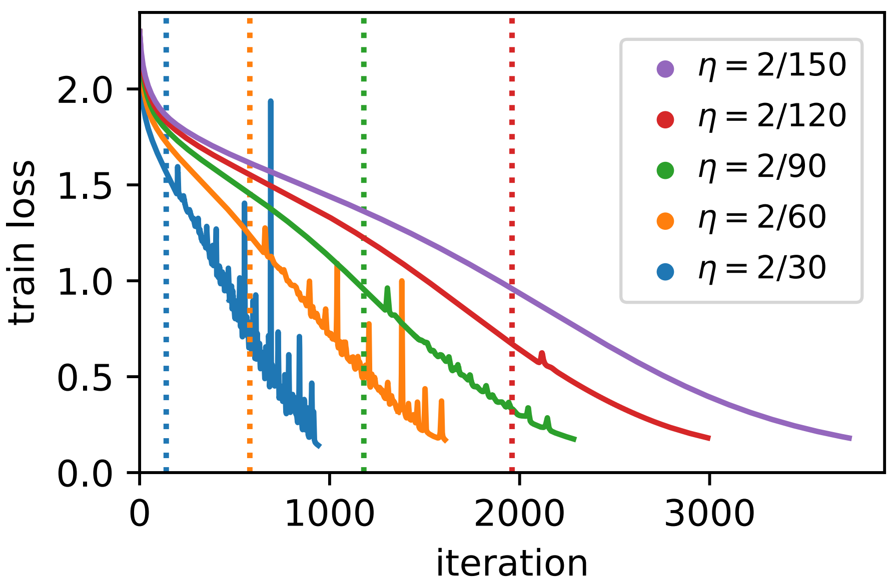


optimization

training instability

"edge of stability"

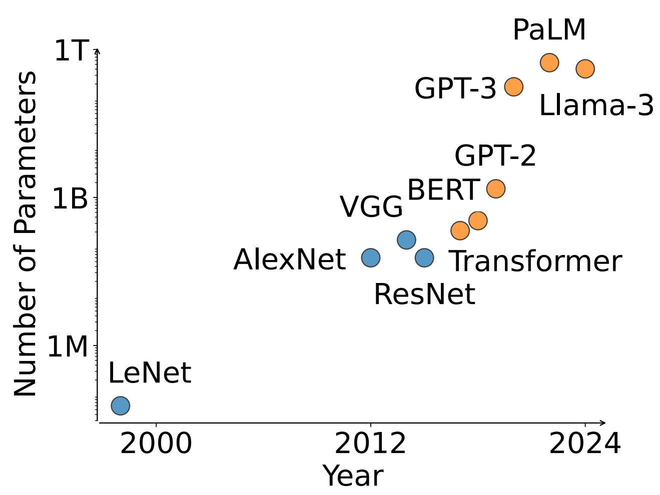
...

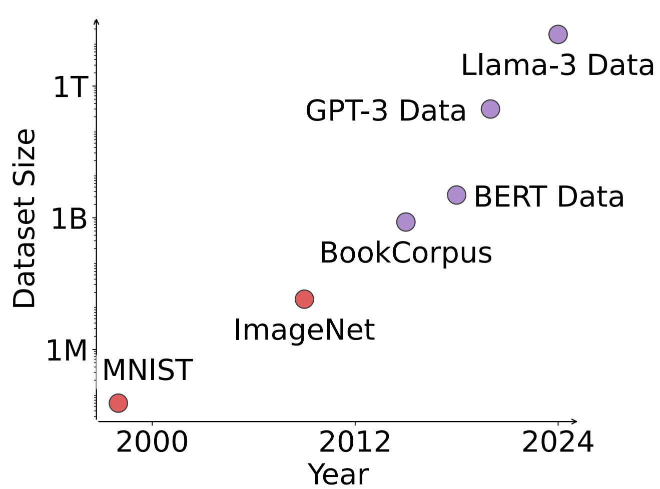


Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021

What makes deep learning thrive?

scaling	new mechanisms



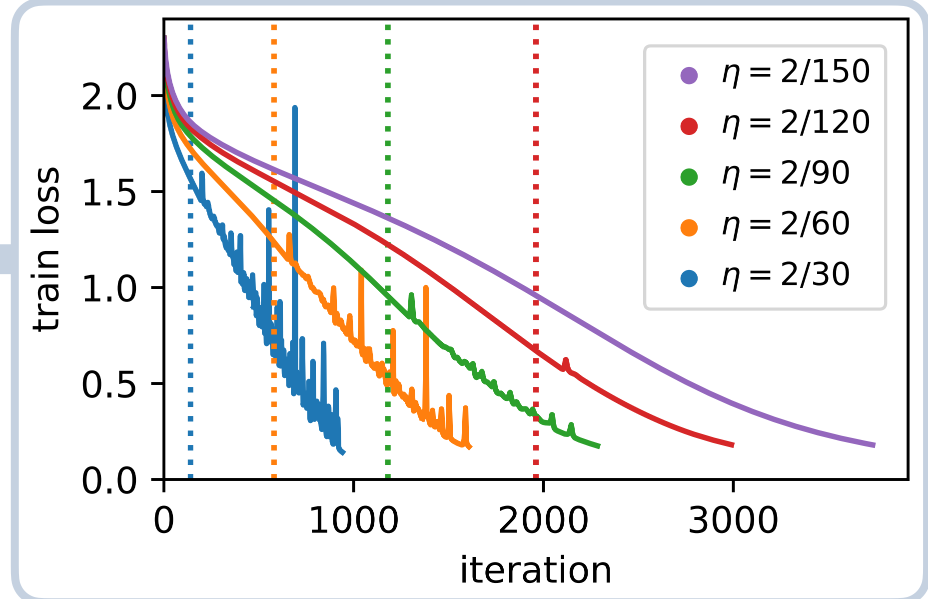


optimization

training instability

"edge of stability"

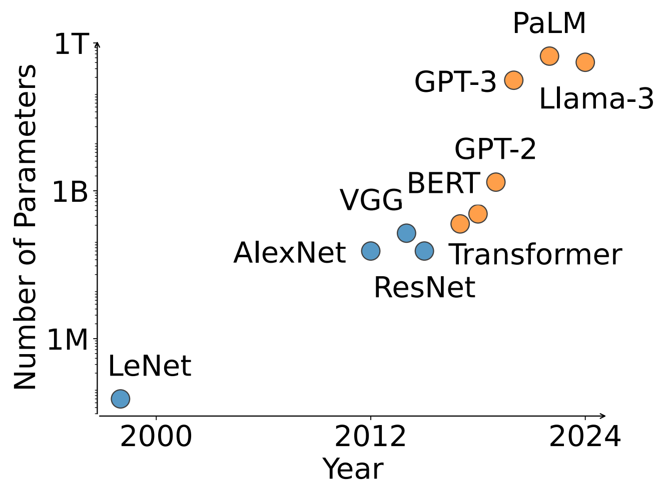
...

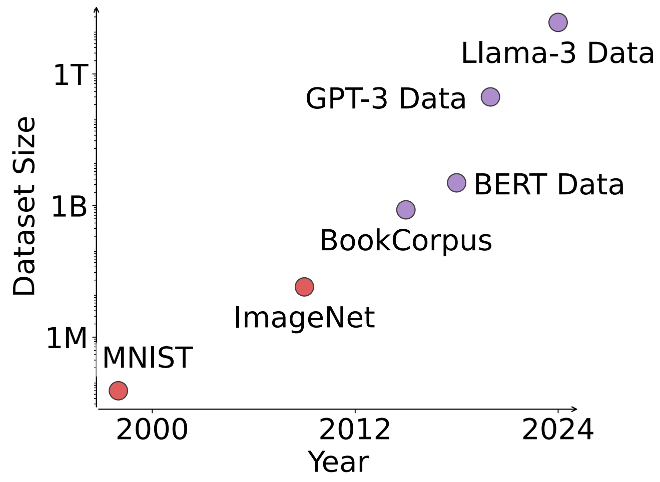


Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021

What makes deep learning thrive?

scaling

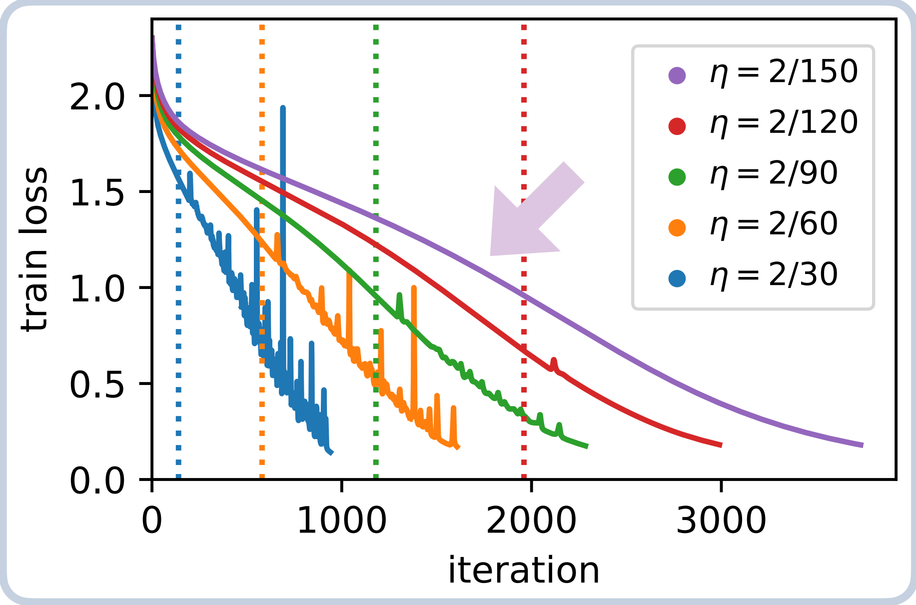




new mechanisms

optimization

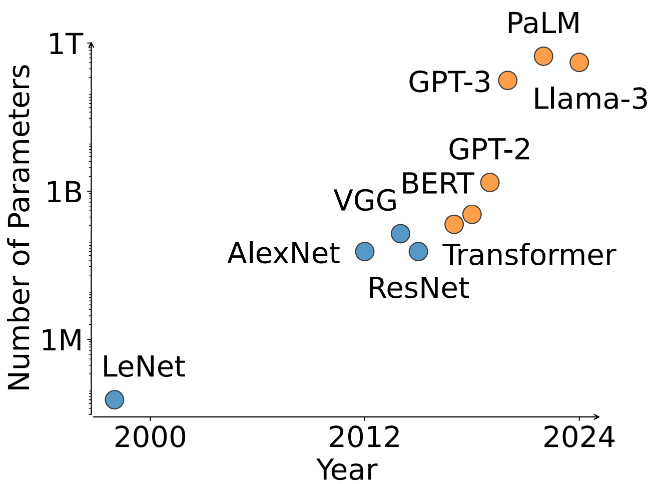
- training instability
- "edge of stability"
- ...

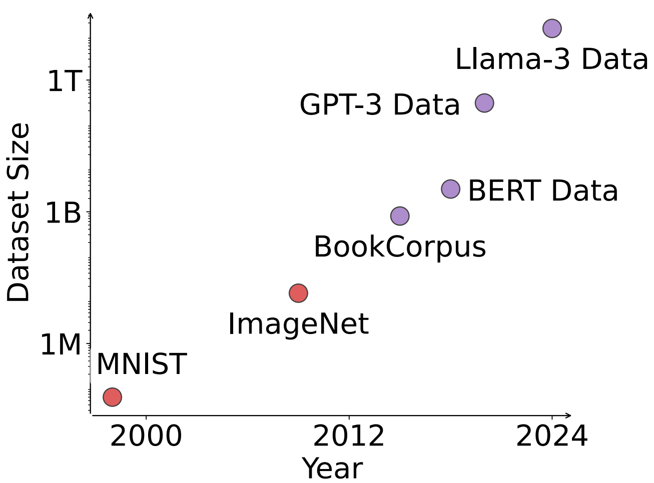


Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021

What makes deep learning thrive?

scaling





new mechanisms

optimization

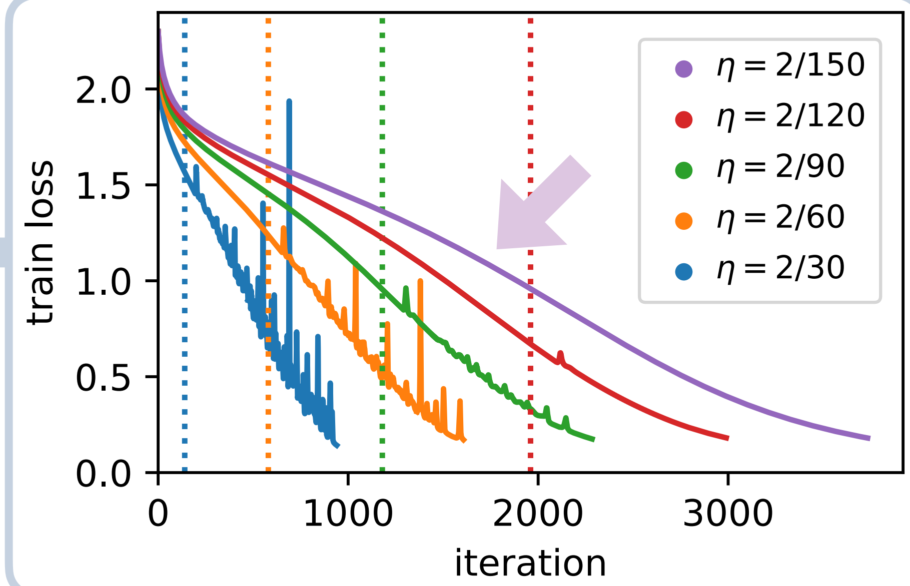
training instability

“edge of stability”

...

large stepsize

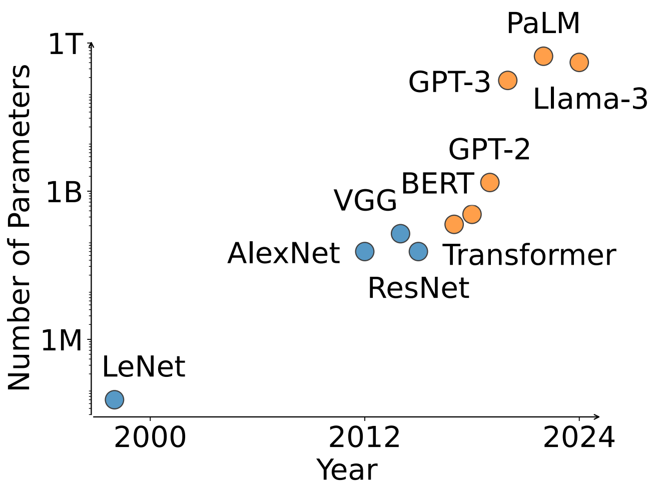
unstable optimization

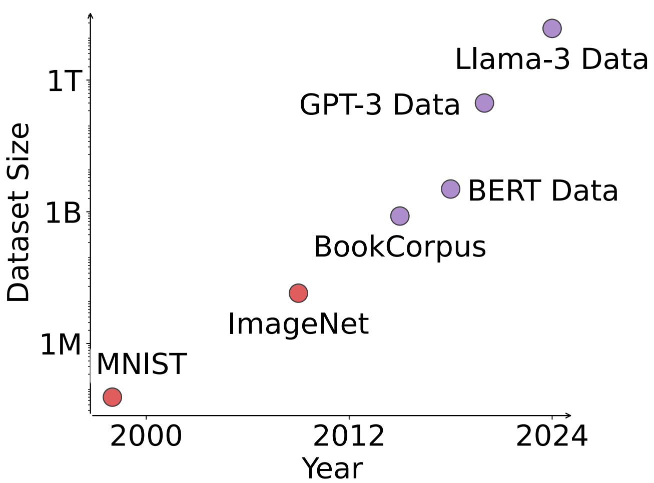


Cohen, Kaur, Li, Kolter, Talwalkar. “Gradient descent on neural networks typically occurs at the edge of stability.” ICLR 2021

What makes deep learning thrive?

scaling





new mechanisms

optimization

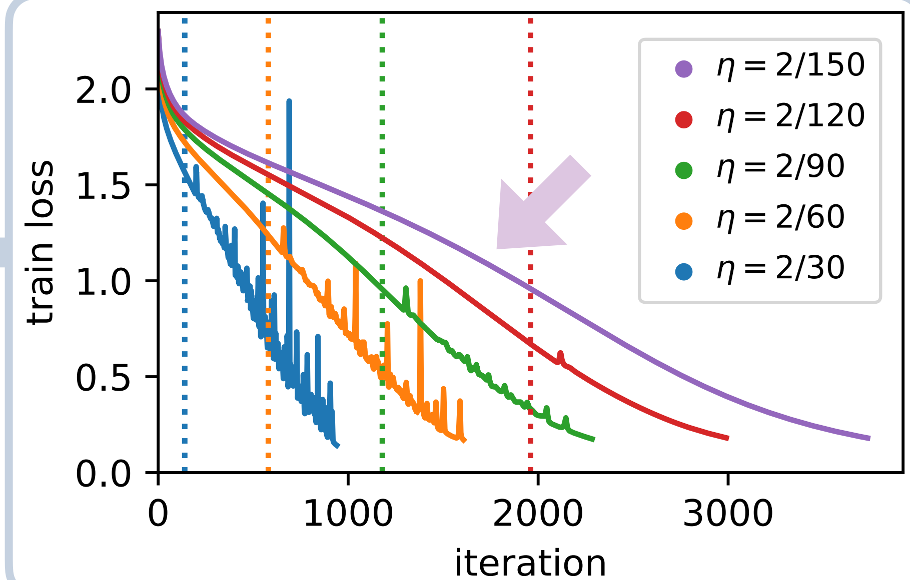
training instability

“edge of stability”

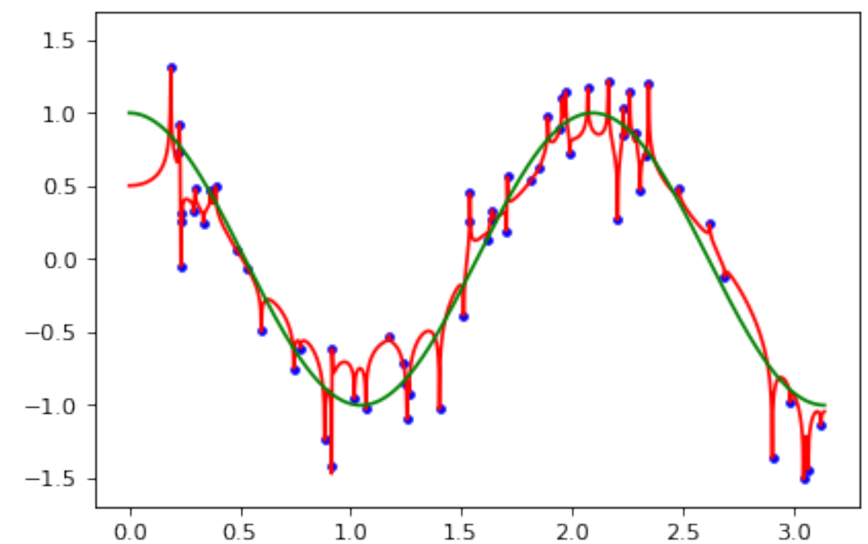
...

large stepsize

unstable optimization



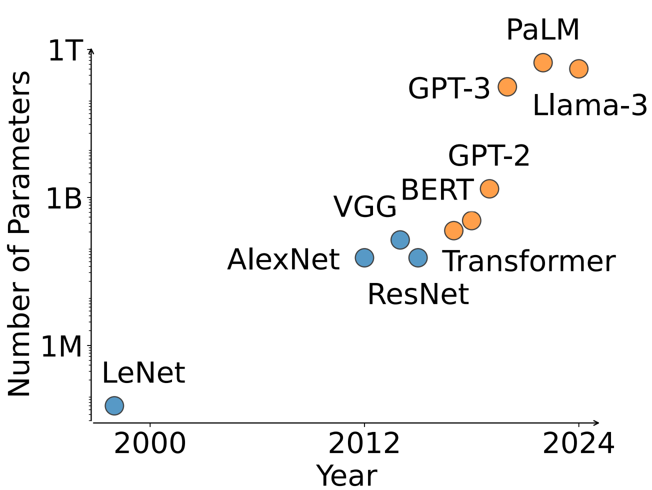
generalization

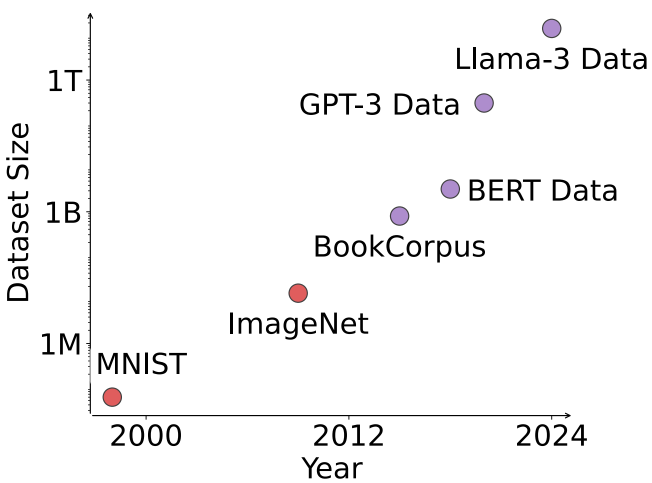


Cohen, Kaur, Li, Kolter, Talwalkar. “Gradient descent on neural networks typically occurs at the edge of stability.” ICLR 2021

What makes deep learning thrive?

scaling





new mechanisms

optimization

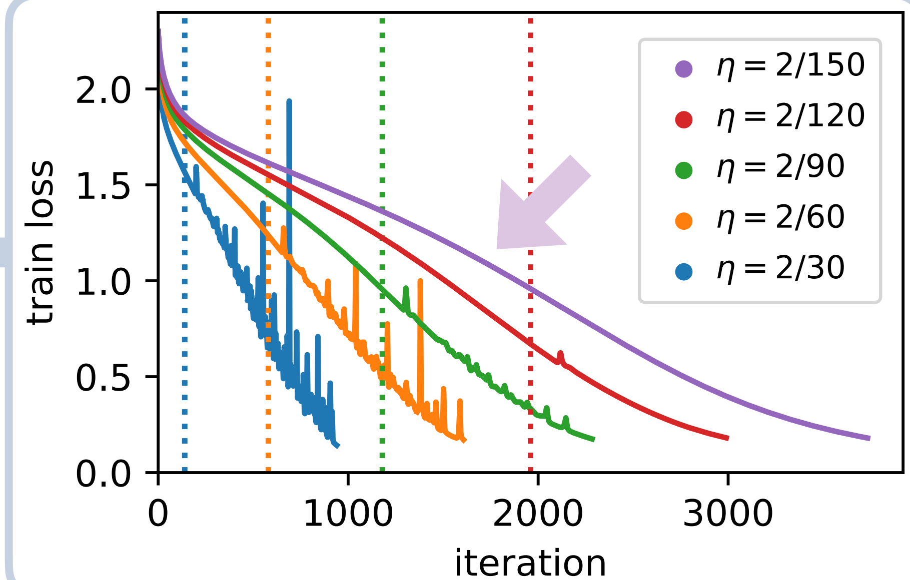
training instability

"edge of stability"

...

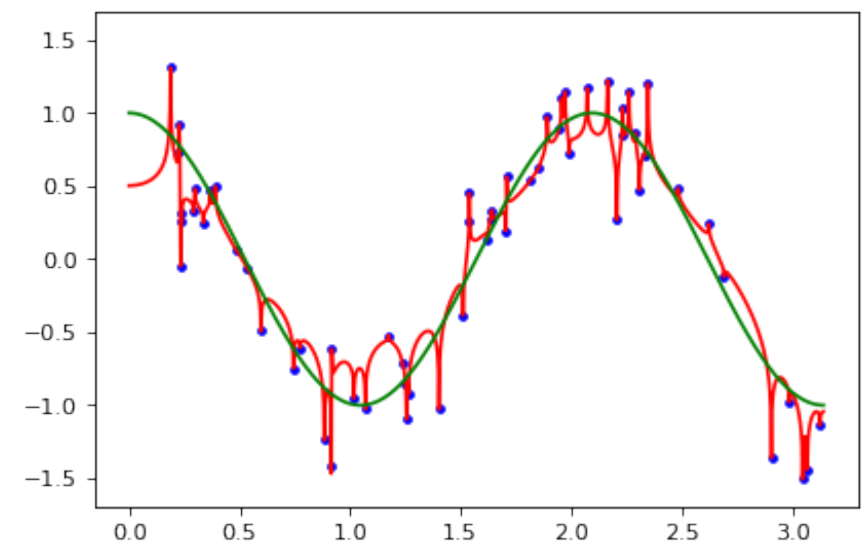
large stepsize

unstable optimization



generalization

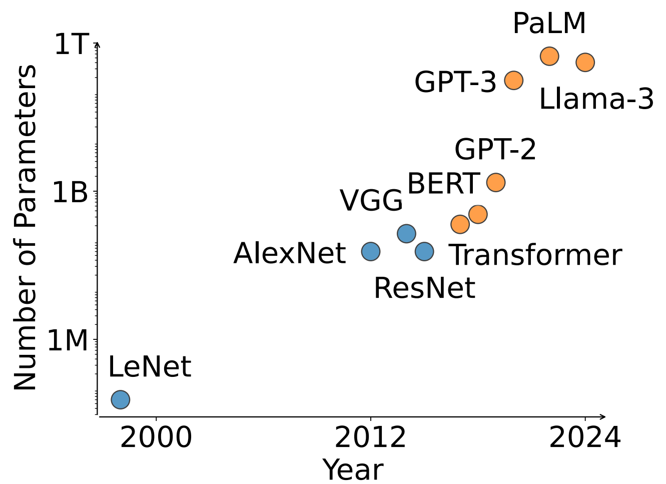
overparameterization

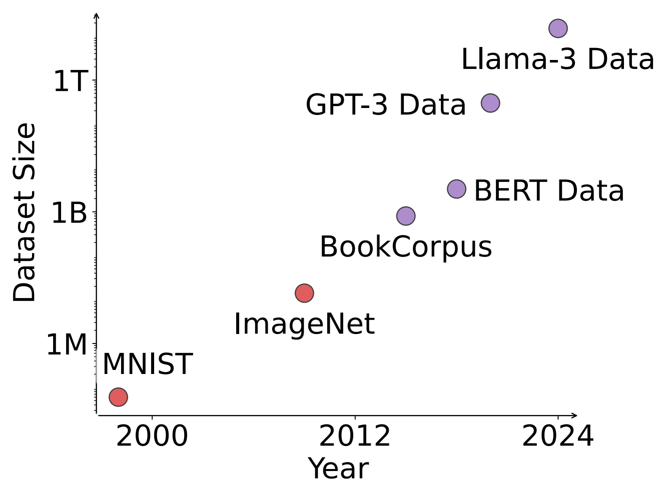


Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021

What makes deep learning thrive?

scaling





new mechanisms

optimization

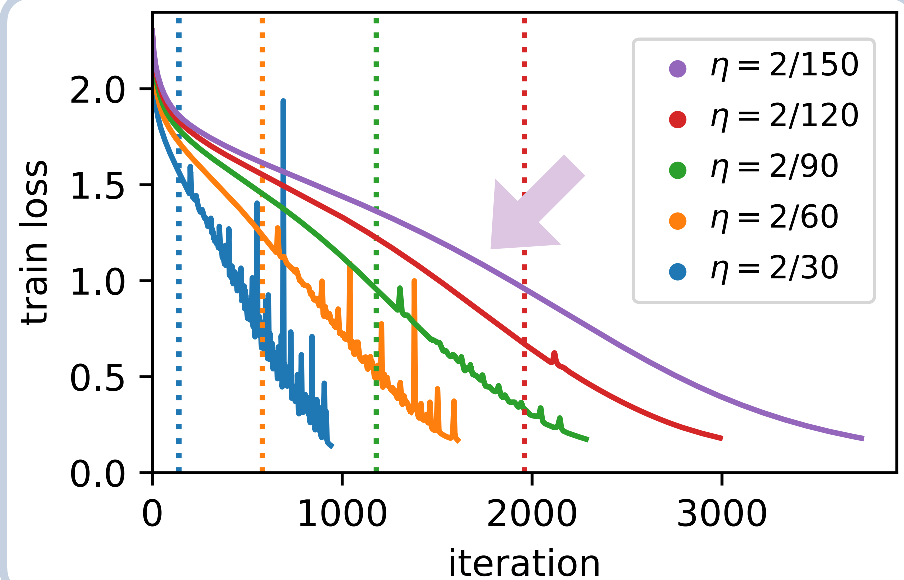
training instability

"edge of stability"

...

large stepsize

unstable optimization

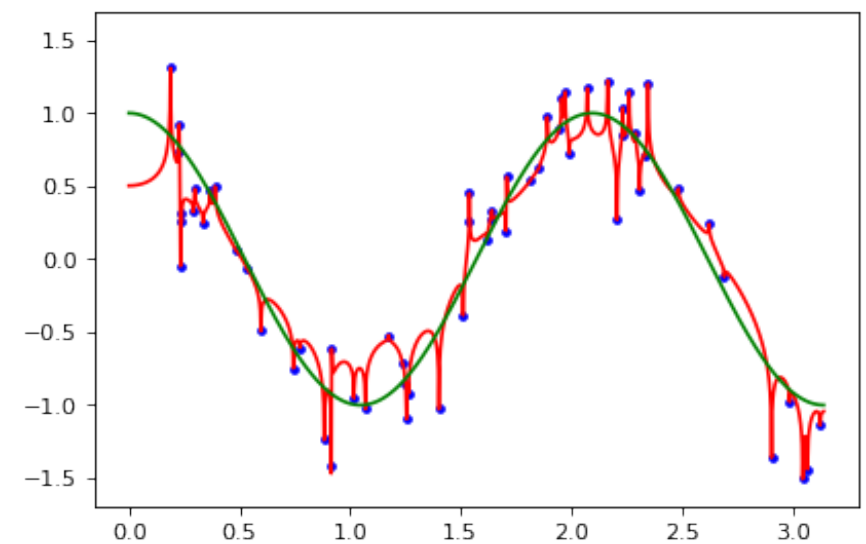


generalization

overparameterization

"benign overfitting"

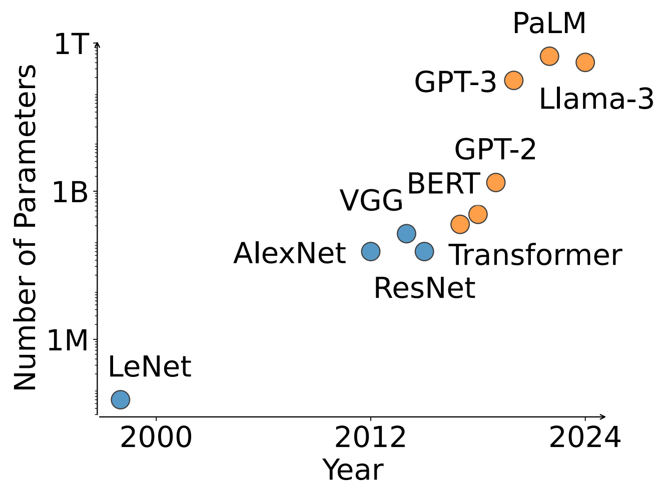
...

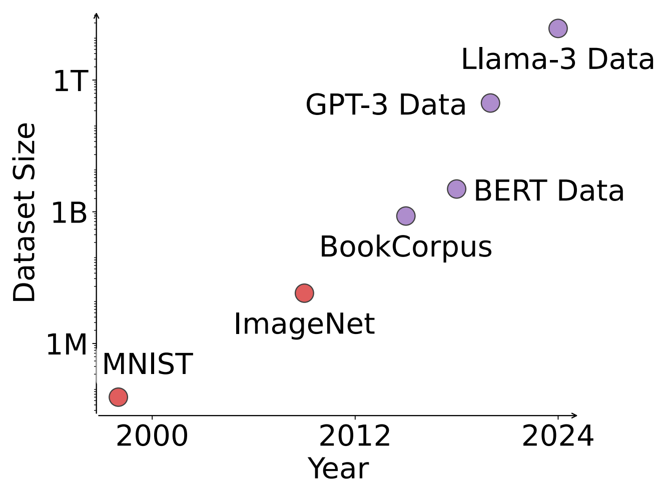


Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021
Bartlett, Long, Lugosi, Tsigler. "Benign overfitting in linear regression." PNAS 2020
Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

What makes deep learning thrive?

scaling





new mechanisms

optimization

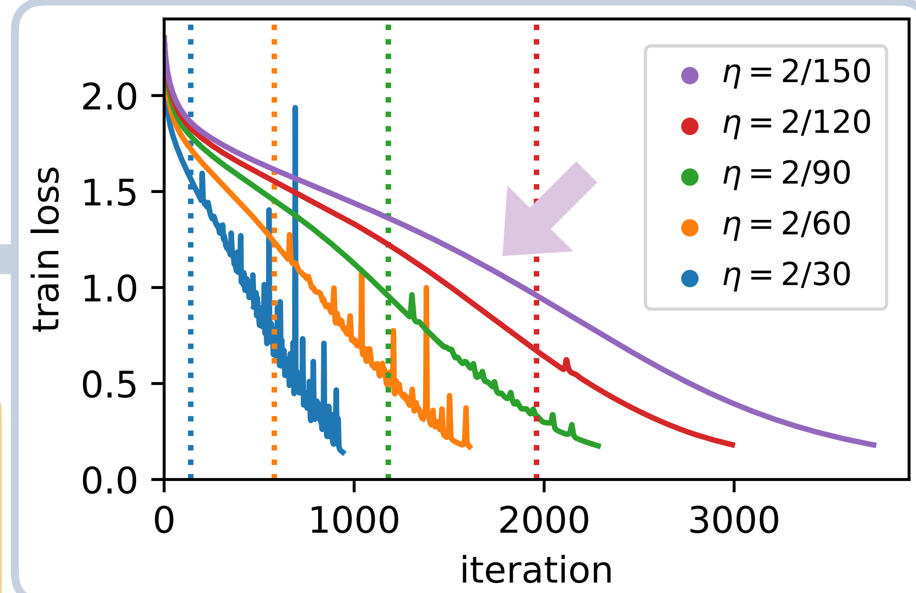
training instability

"edge of stability"

...

large stepsize

unstable optimization

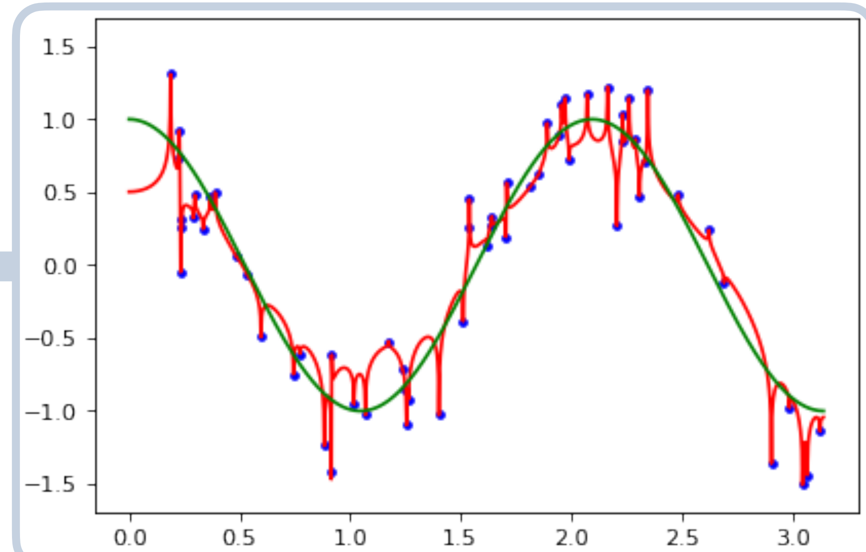


generalization

overparameterization

"benign overfitting"

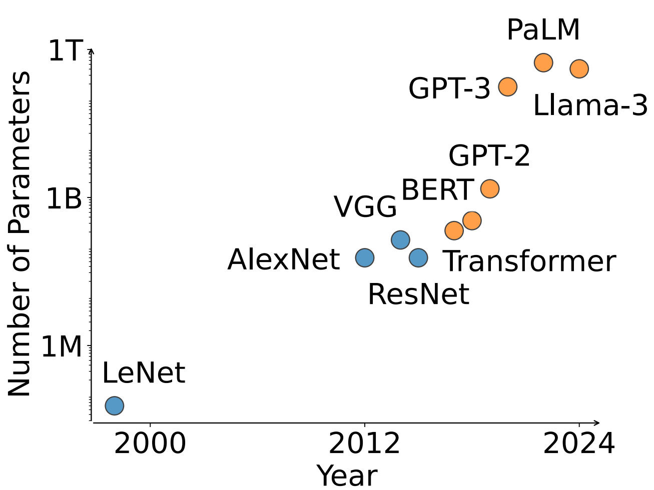
...

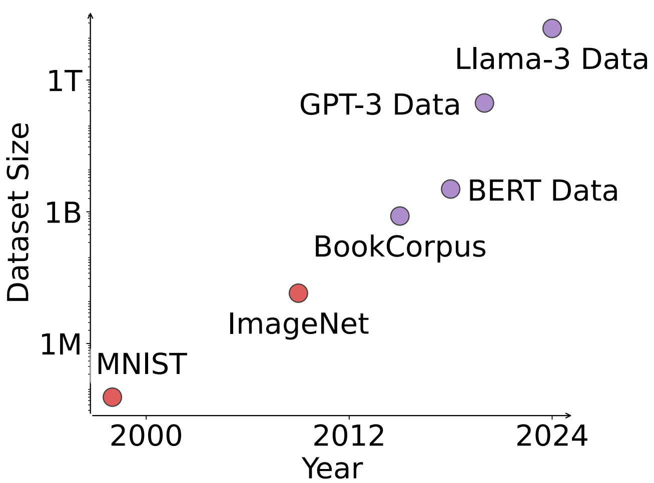


Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021
Bartlett, Long, Lugosi, Tsigler. "Benign overfitting in linear regression." PNAS 2020
Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

What makes deep learning thrive?

scaling





new mechanisms

optimization

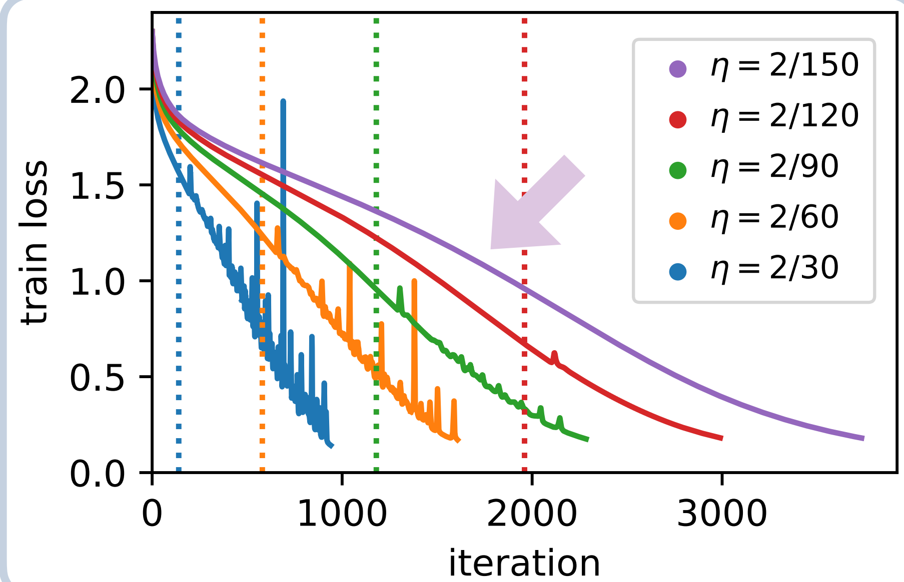
training instability

"edge of stability"

...

large stepsize

unstable optimization



generalization

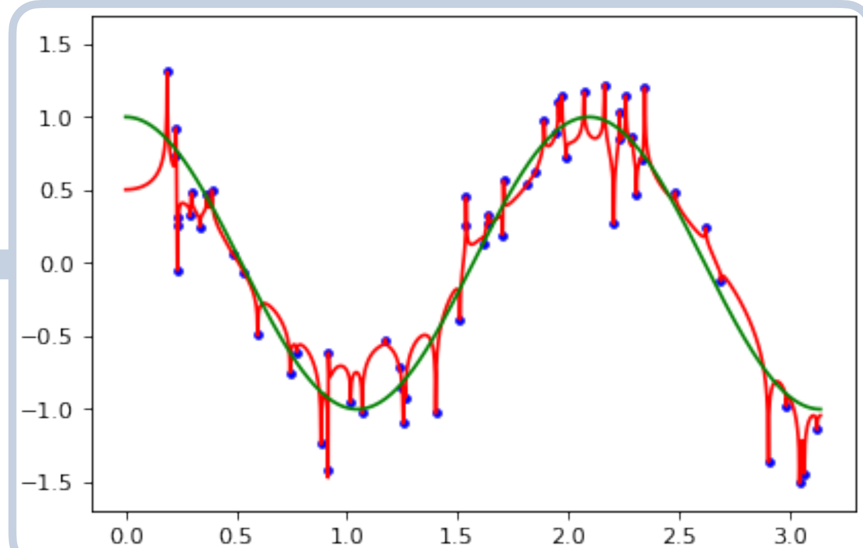
overparameterization

"benign overfitting"

...

optimization path

implicit regularization



Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021
Bartlett, Long, Lugosi, Tsigler. "Benign overfitting in linear regression." PNAS 2020
Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

My research

deep learning = scaling + new mechanisms

My research

deep learning = scaling + new mechanisms

My research

not covered by classical theory

deep learning = scaling + new mechanisms

My research

💡💡💡

not covered by classical theory

deep learning = scaling + new mechanisms

My research

💡 💡 💡

not covered by classical theory

deep learning = scaling + **new mechanisms**

not specific to deep learning

My research

💡💡💡

not covered by
classical theory

deep learning = scaling + new mechanisms

ML fields
(beyond DL)

not specific to
deep learning

✨✨✨

My research

💡 💡 💡

not covered by classical theory

deep learning = scaling + **new mechanisms**

ML fields (beyond DL)

not specific to deep learning

✨✨✨

Approach. Demystify new mechanisms in **sandboxes**

My research

💡 💡 💡

not covered by classical theory

deep learning = scaling + new mechanisms

ML fields (beyond DL)

not specific to deep learning

✨✨✨

Approach. Demystify new mechanisms in **sandboxes**

simple

My research

💡 💡 💡

not covered by classical theory

deep learning = scaling + new mechanisms

ML fields (beyond DL)

not specific to deep learning

✨✨✨

Approach. Demystify new mechanisms in **sandboxes**

simple

meaningful

Contribution 1: unstable optimization

large stepsize accelerates gradient descent in logistic regression

Contribution 2: implicit regularization

gradient descent dominates ridge regression in linear regression

Contribution 3: from theory to practice

principled parallelization method for training language models

Contribution 1: unstable optimization

large stepsize accelerates gradient descent in logistic regression

- "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency"
  W, Peter Bartlett, Matus Telgarsky, Bin Yu
  COLT 2024

- "Large stepsizes accelerate gradient descent for regularized logistic regression"
  W*, Pierre Marion*, Peter Bartlett
  NeurIPS 2025

Unstable optimization

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Unstable optimization

how to choose η?

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Unstable optimization

how to choose η?

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

classical theory:

small stepsize for stability

$L(\theta_t) \downarrow$

Unstable optimization

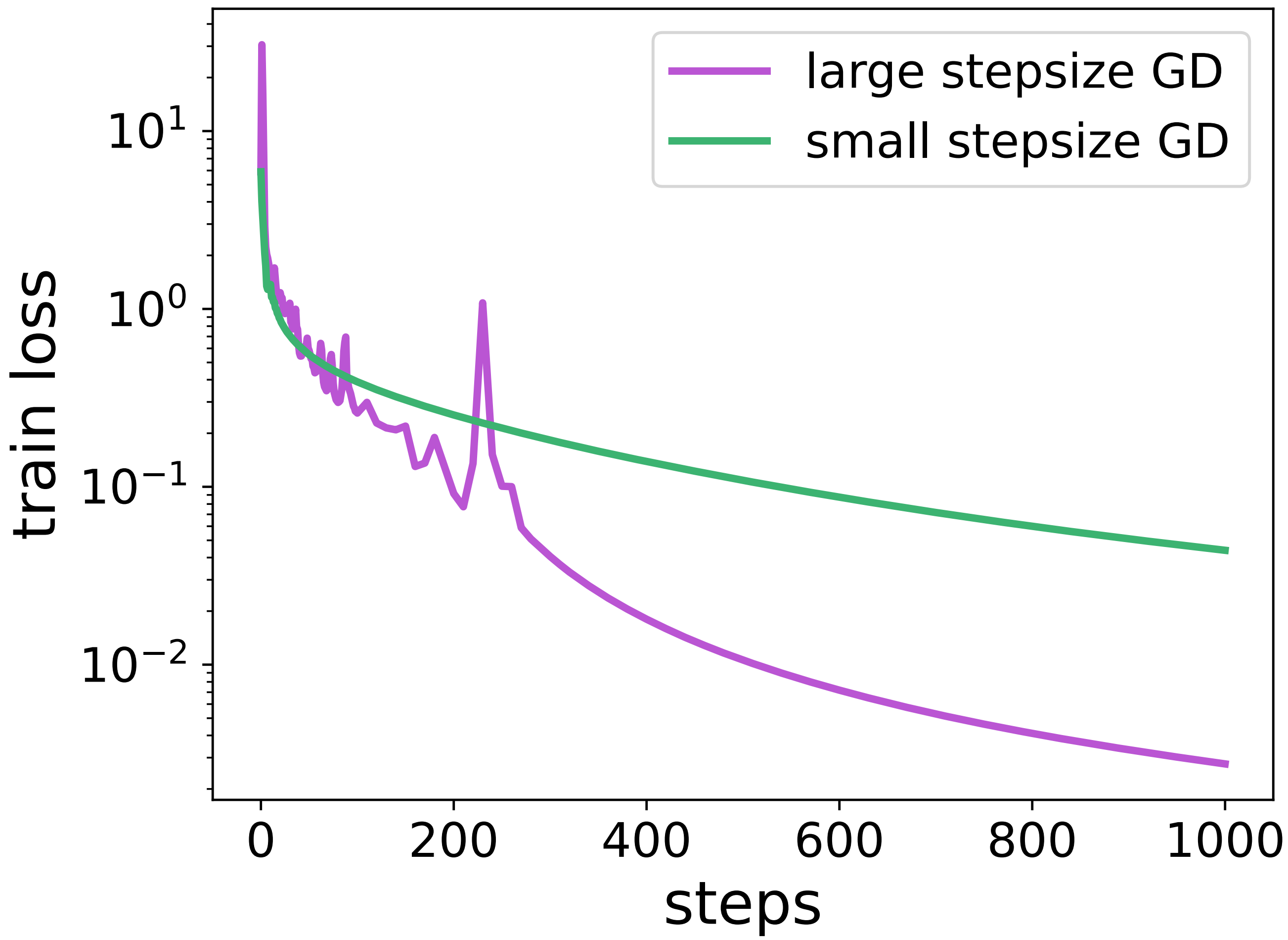
how to choose η?

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

classical theory:

small stepsize for stability

$L(\theta_t) \downarrow$



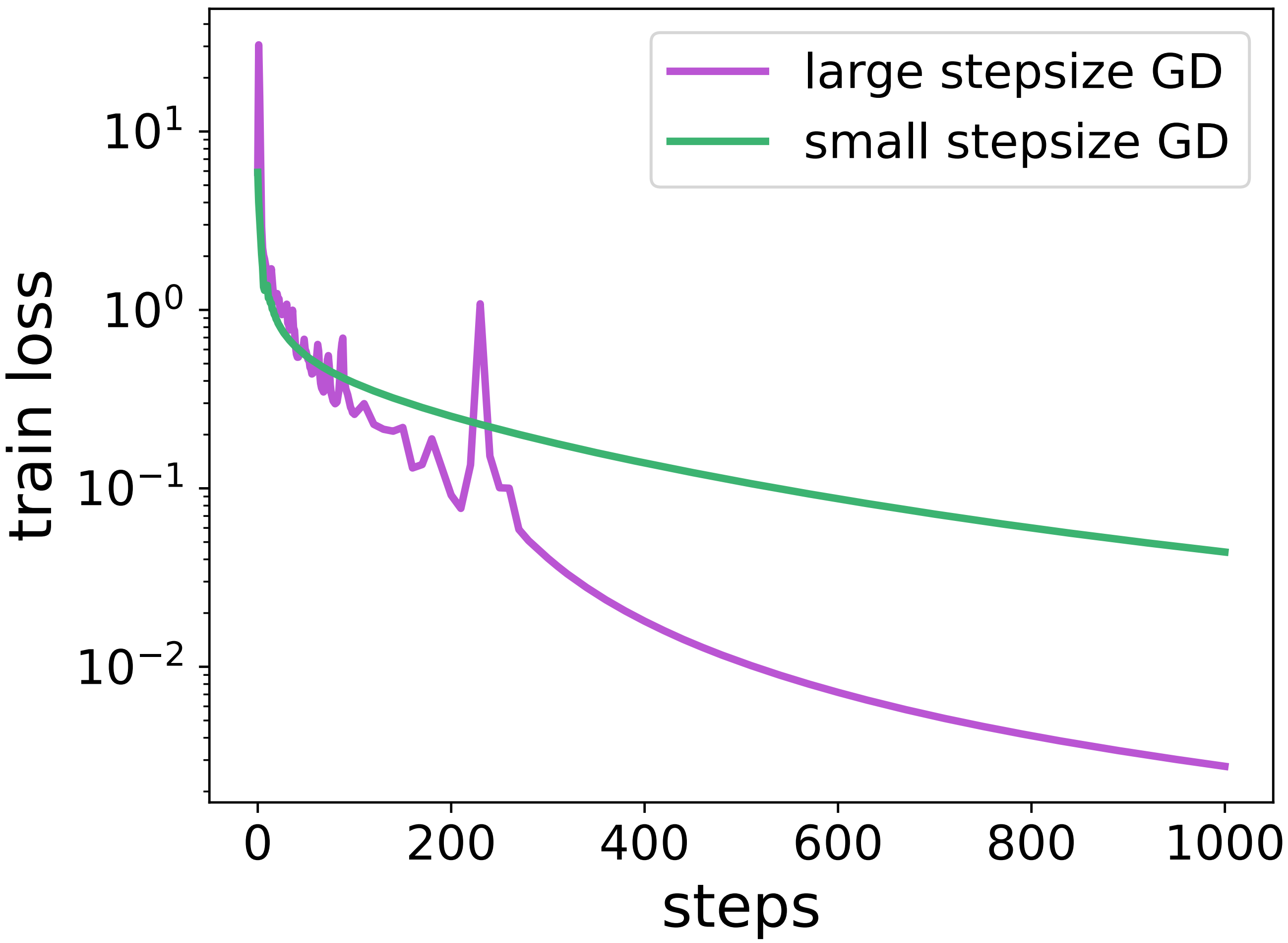
MLP, GD, classification task

Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021
Wu, Wang, Fazel. "Theoretical insights on training instability in deep learning." Tutorial@NeurIPS 2025

Unstable optimization

how to choose η?

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$



classical theory:
small stepsize for stability
$L(\theta_t) \downarrow$

practice:
best stepsize is large
oscillatory loss

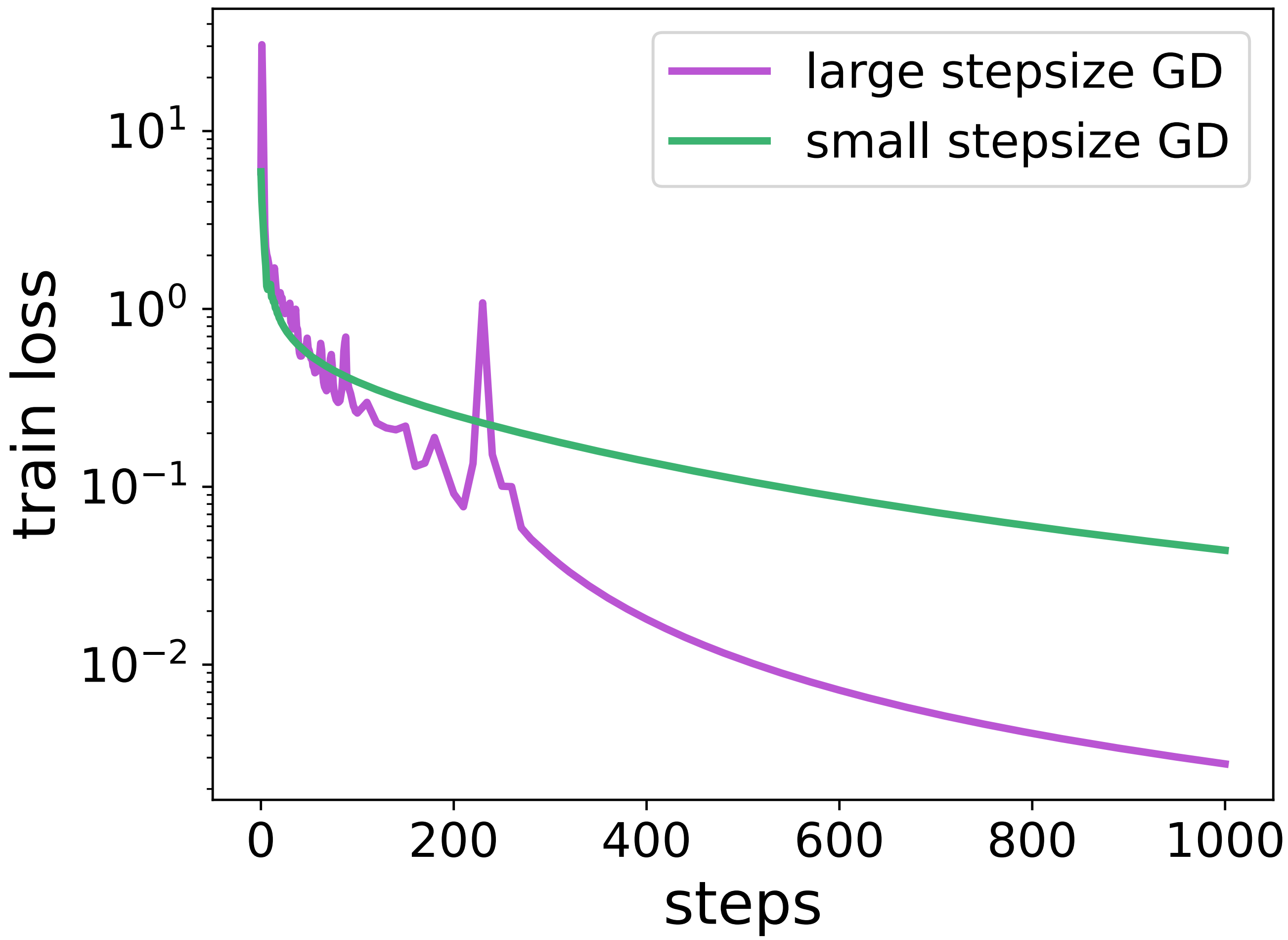
MLP, GD, classification task

Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021
Wu, Wang, Fazel. "Theoretical insights on training instability in deep learning." Tutorial@NeurIPS 2025

Unstable optimization

how to choose η?

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$



MLP, GD, classification task

classical theory:
small stepsize for stability
$L(\theta_t) \downarrow$

practice:
best stepsize is large
oscillatory loss

classical theory fails to predict best stepsize

Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021
Wu, Wang, Fazel. "Theoretical insights on training instability in deep learning." Tutorial@NeurIPS 2025

Classical theory

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer

Classical theory

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer

Descent lemma.

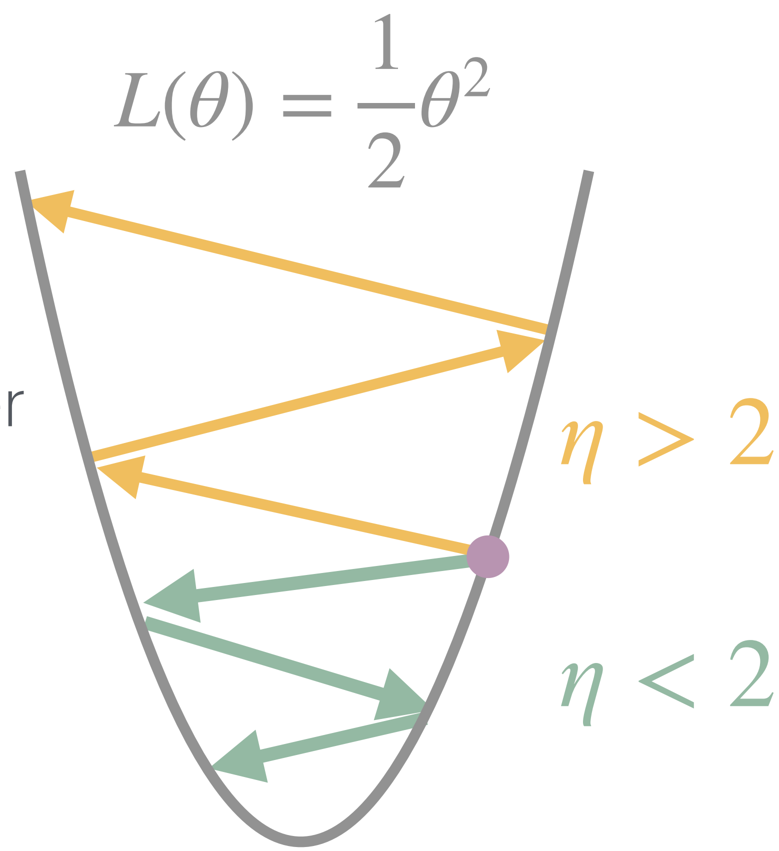
- **small stepsize** $\eta < 2 \Rightarrow L(\theta_t) \downarrow$
- **large stepsize** $\eta > 2 \Rightarrow L(\theta_t) \uparrow \infty$ for quadratics

Classical theory

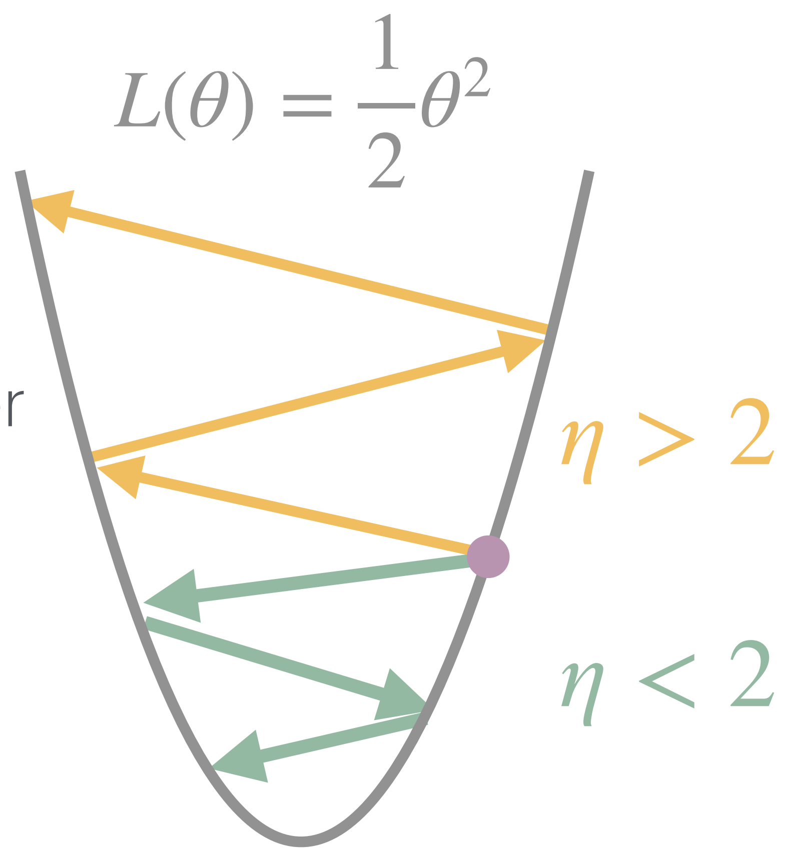
Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer

Descent lemma.

- **small stepsize** $\eta < 2 \Rightarrow L(\theta_t) \downarrow$
- **large stepsize** $\eta > 2 \Rightarrow L(\theta_t) \uparrow \infty$ for quadratics



Classical theory



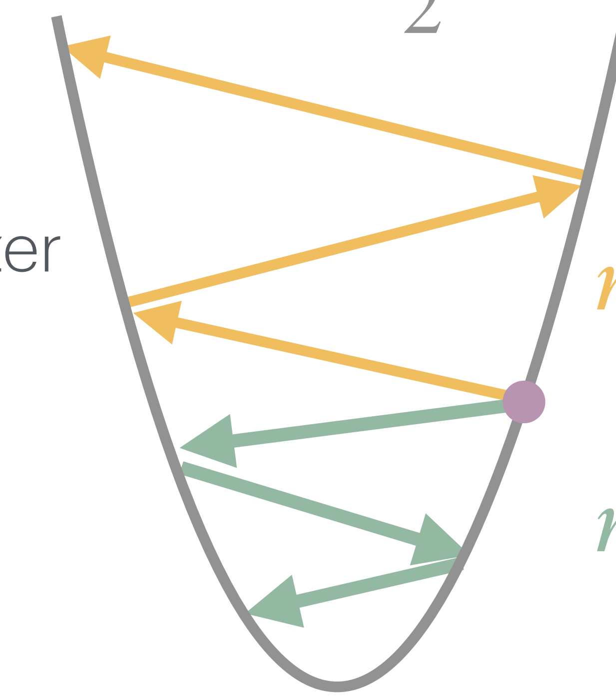
Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer

Descent lemma.

- **small stepsize** $\eta < 2 \Rightarrow L(\theta_t) \downarrow$
- **large stepsize** $\eta > 2 \Rightarrow L(\theta_t) \uparrow \infty$ for quadratics

Classical theory

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer



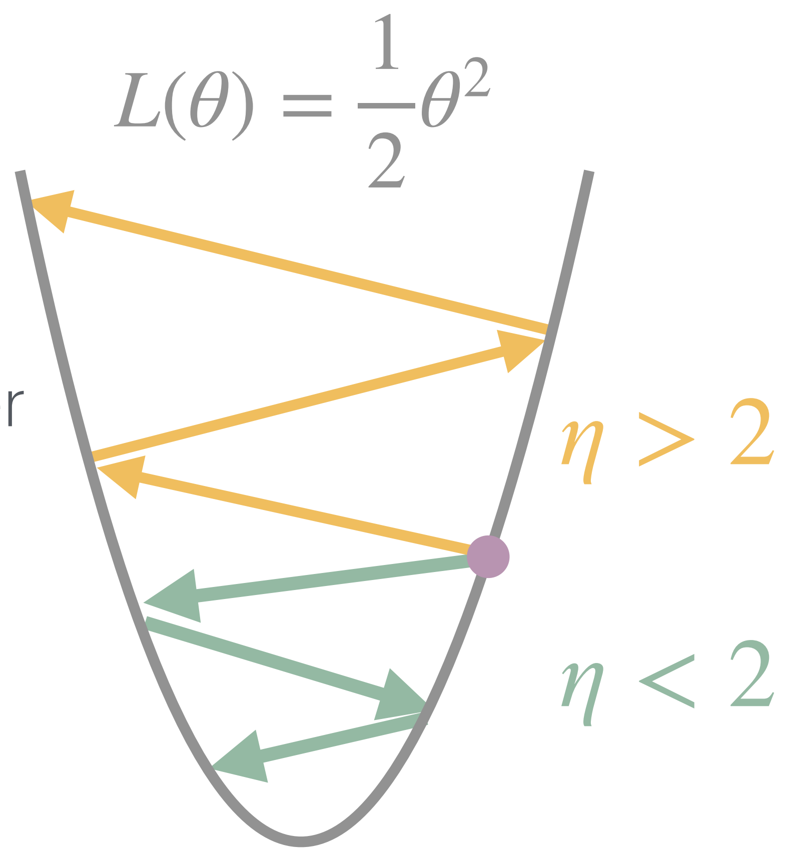
Descent lemma.

- **small stepsize** $\eta < 2 \Rightarrow L(\theta_t) \downarrow$
- **large stepsize** $\eta > 2 \Rightarrow L(\theta_t) \uparrow \infty$ for quadratics

Rates. GD with $\eta = 1$ achieves

- **convexity** $L(\theta_t) - \min L \leq O(1/t)$
- λ**-strong convexity** $L(\theta_t) - \min L \leq \epsilon$ for $t = O\big(\kappa \ln(1/\epsilon)\big)$

Classical theory



Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer

Descent lemma.

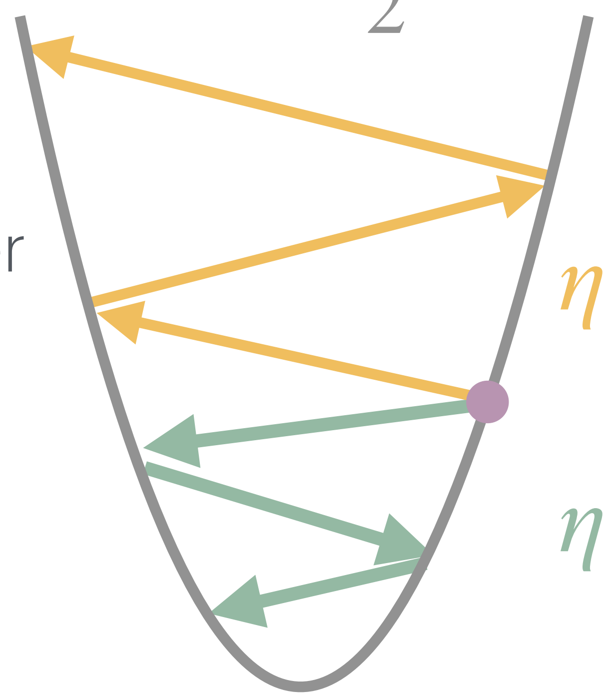
- **small stepsize** $\eta < 2 \Rightarrow L(\theta_t) \downarrow$
- **large stepsize** $\eta > 2 \Rightarrow L(\theta_t) \uparrow \infty$ for quadratics

Rates. GD with $\eta = 1$ achieves

- **convexity** $L(\theta_t) - \min L \leq O(1/t)$
- **λ-strong convexity** $L(\theta_t) - \min L \leq \epsilon$ for $t = O\big(\kappa \ln(1/\epsilon)\big)$

  condition number $\kappa = 1/\lambda \gg 1$

Classical theory



Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer

Descent lemma.

- **small stepsize** $\eta < 2 \Rightarrow L(\theta_t) \downarrow$
- **large stepsize** $\eta > 2 \Rightarrow L(\theta_t) \uparrow \infty$ for quadratics

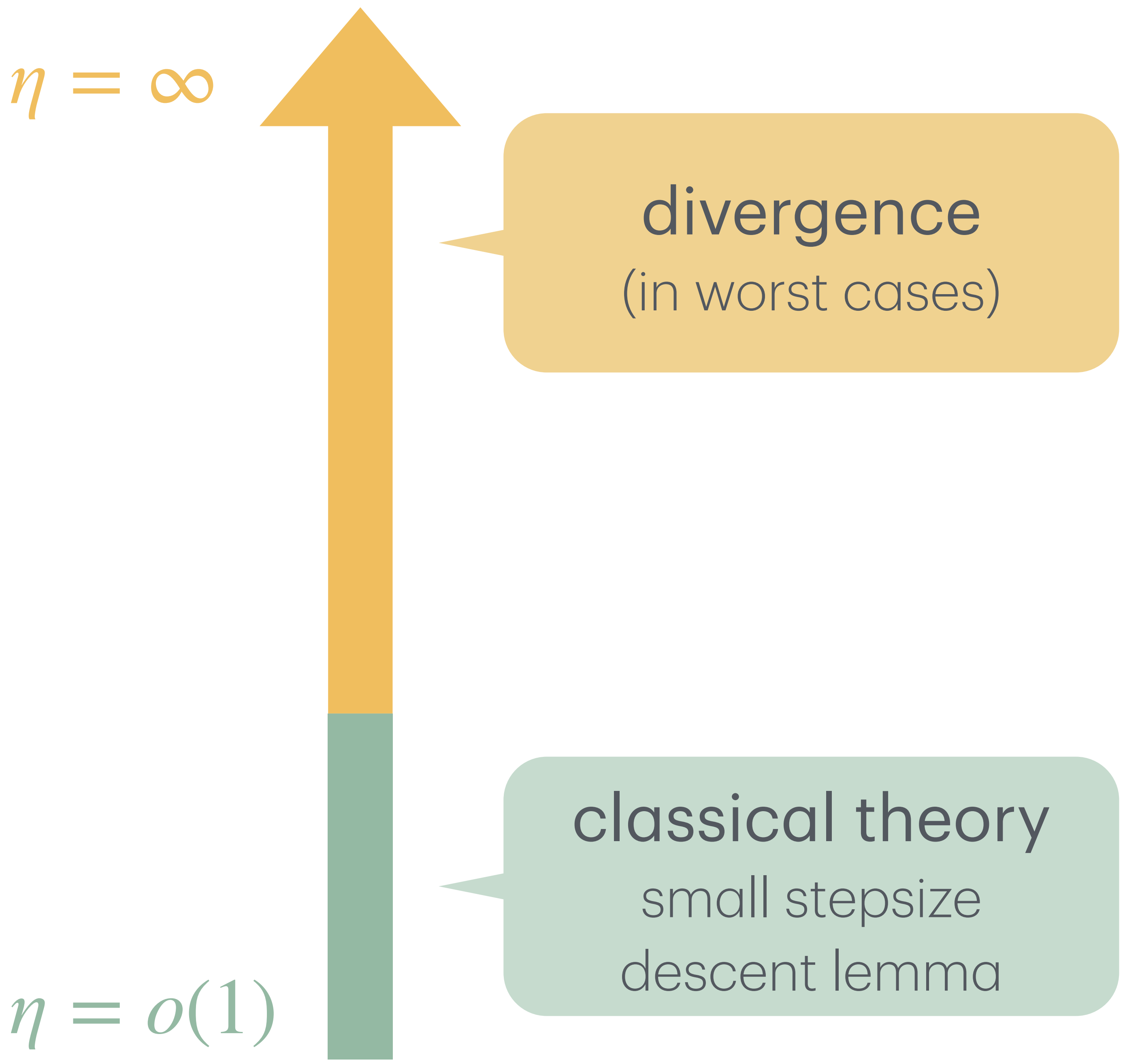
Rates. GD with $\eta = 1$ achieves

- **convexity** $L(\theta_t) - \min L \leq O(1/t)$
- **λ-strong convexity** $L(\theta_t) - \min L \leq \epsilon$ for $t = O\big(\kappa \ln(1/\epsilon)\big)$

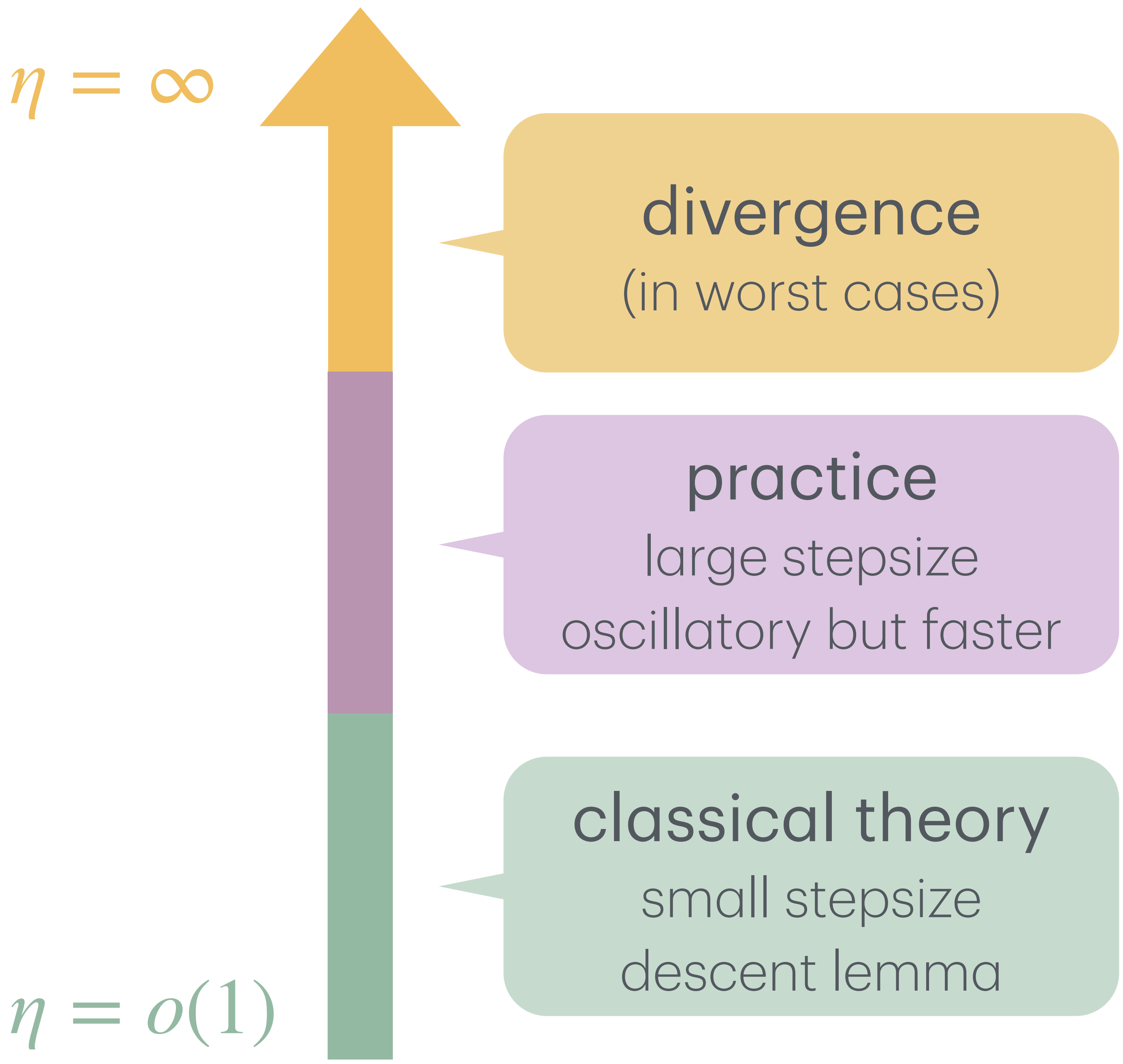
condition number
$\kappa = 1/\lambda \gg 1$

acceleration by Nesterov's momentum:
$O\big(1/t^2\big)$ & $O\big(\sqrt{\kappa}\ln(1/\epsilon)\big)$

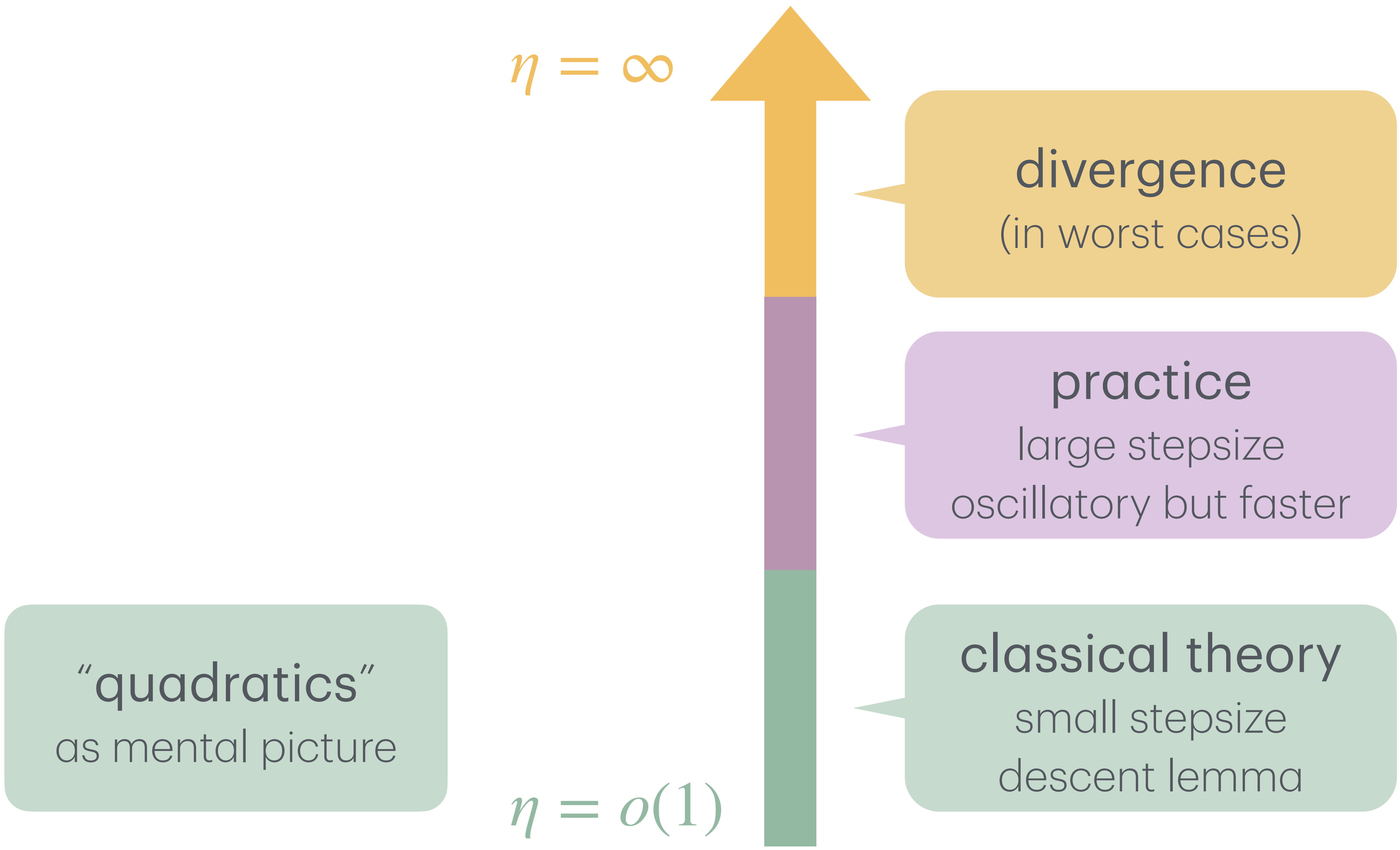
Stepsize


$\eta = \infty$
divergence
(in worst cases)
classical theory
small stepsize
descent lemma
$\eta = o(1)$

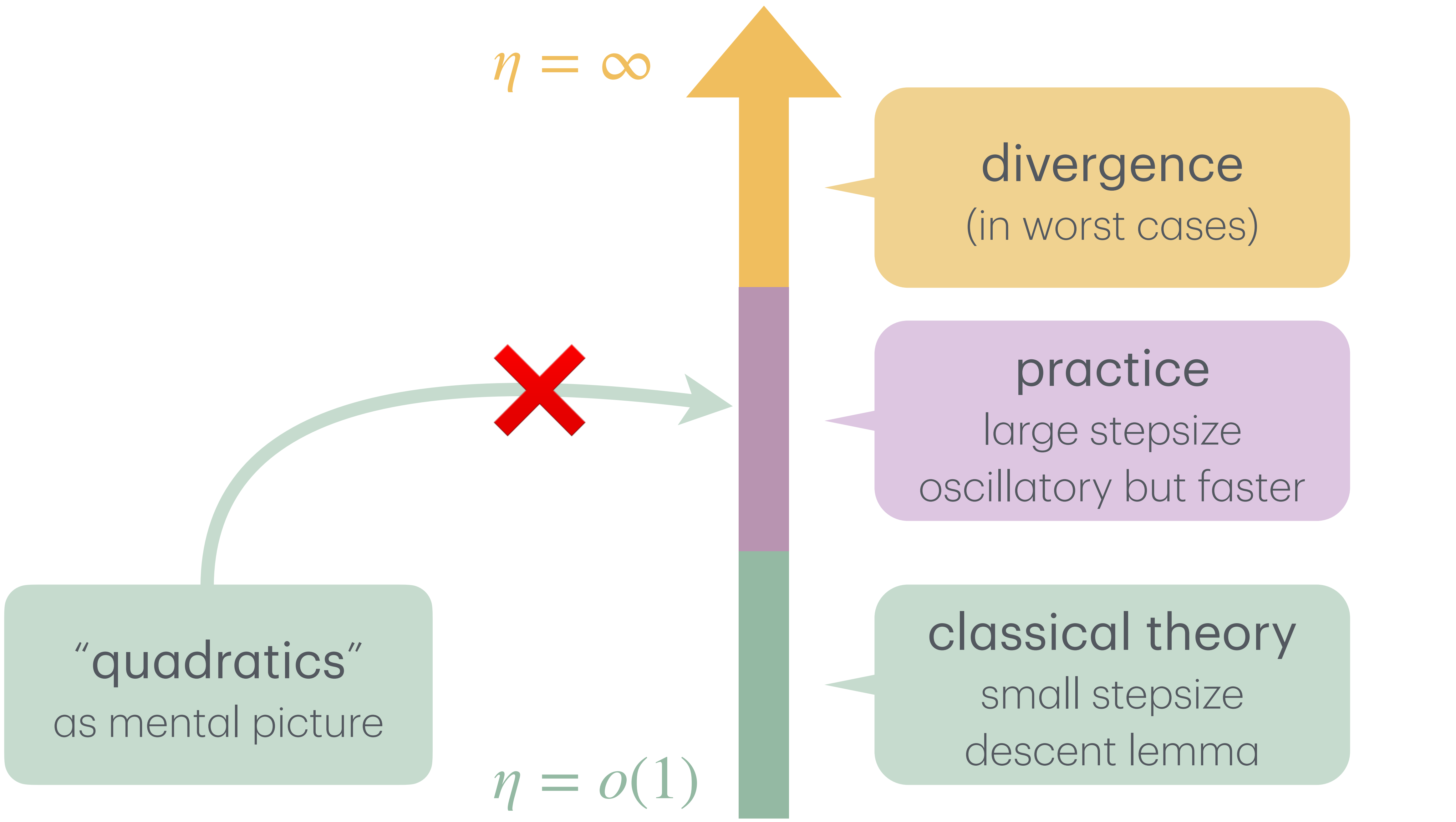
Stepsize


$\eta = \infty$
divergence
(in worst cases)
practice
large stepsize
oscillatory but faster
classical theory
small stepsize
descent lemma
$\eta = o(1)$

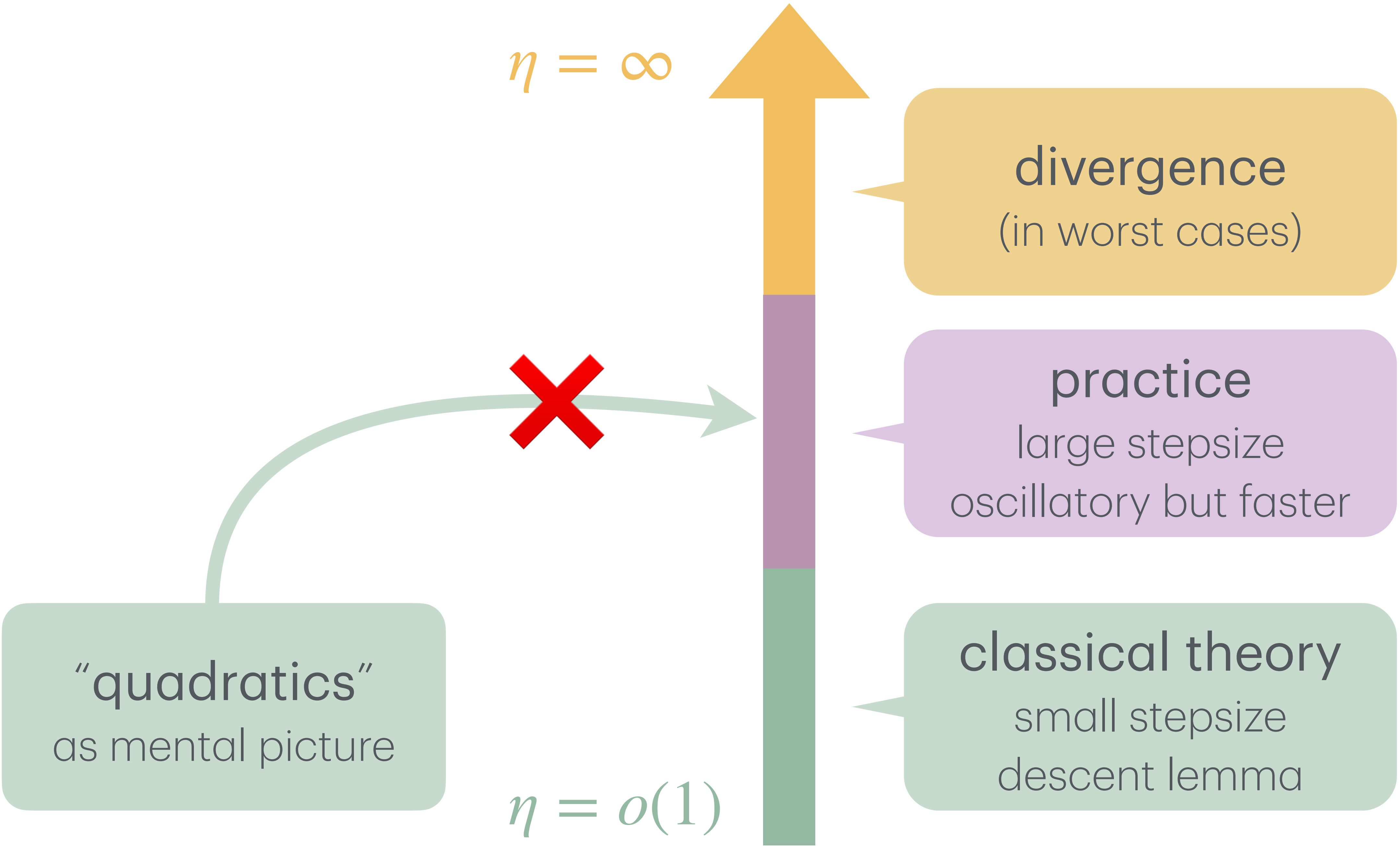
Stepsize


$\eta = \infty$
divergence
(in worst cases)
practice
large stepsize
oscillatory but faster
classical theory
small stepsize
descent lemma
"quadratics"
as mental picture
$\eta = o(1)$

Stepsize


$\eta = \infty$
divergence
(in worst cases)
practice
large stepsize
oscillatory but faster
"quadratics"
as mental picture
classical theory
small stepsize
descent lemma
$\eta = o(1)$

Stepsize



Related works

- Altschuler, Parrilo. "Acceleration by stepsize hedging I: multi-step descent and the silver stepsize schedule." Journal of the ACM 2024
- Grimmer, Shu, Wang. "Composing optimized stepsize schedules for gradient descent." Mathematics of Operations Research 2025

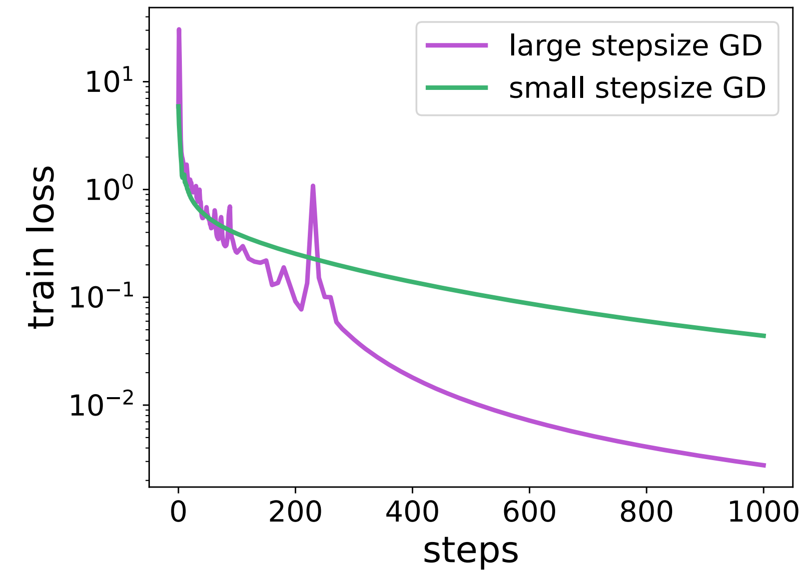
…

Seeking simplest sandbox

......

deep learning

unstable optimization observed

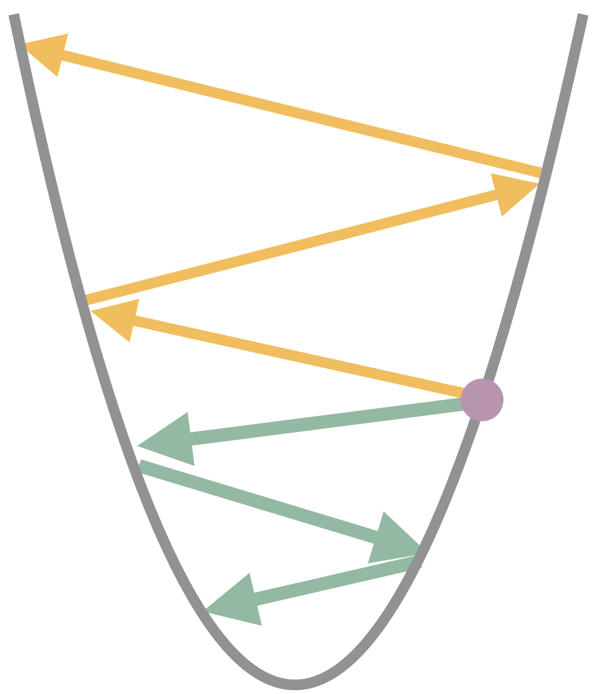


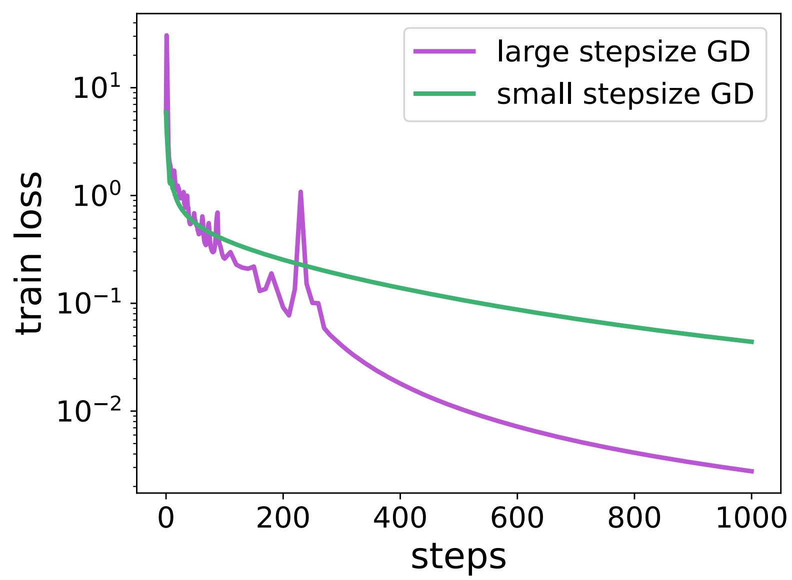
Seeking simplest sandbox

linear regression ...... deep learning

unstable optimization impossible

unstable optimization observed





Seeking simplest sandbox

linear regression

logistic regression

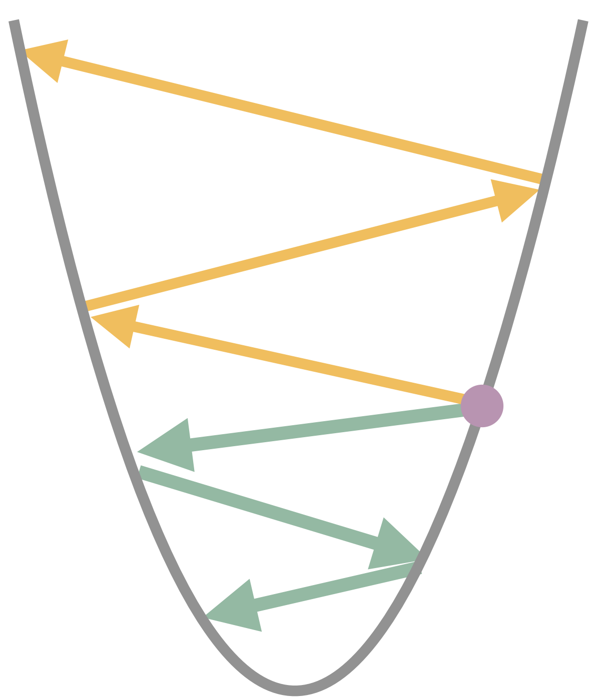
......

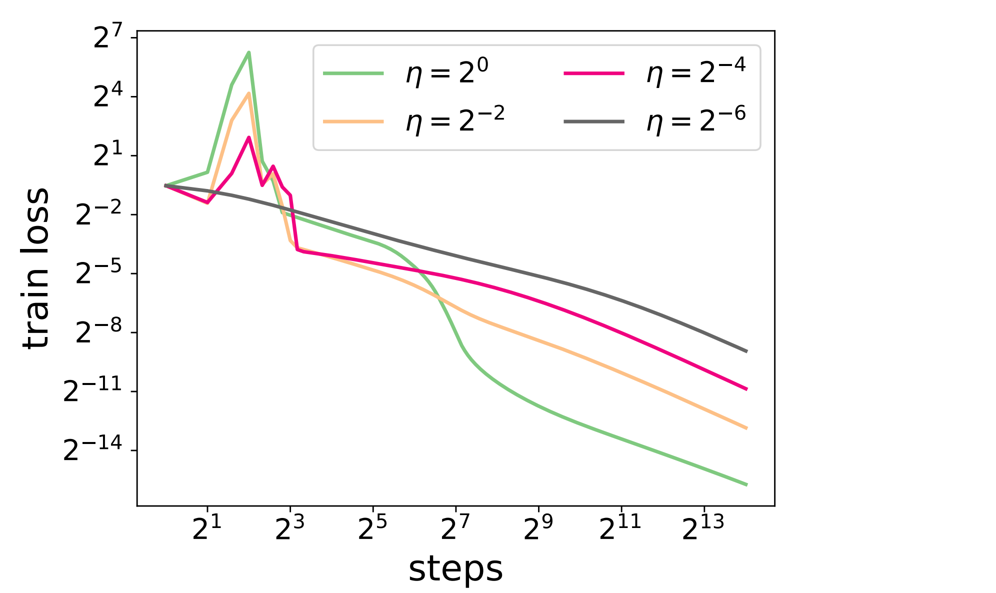
deep learning

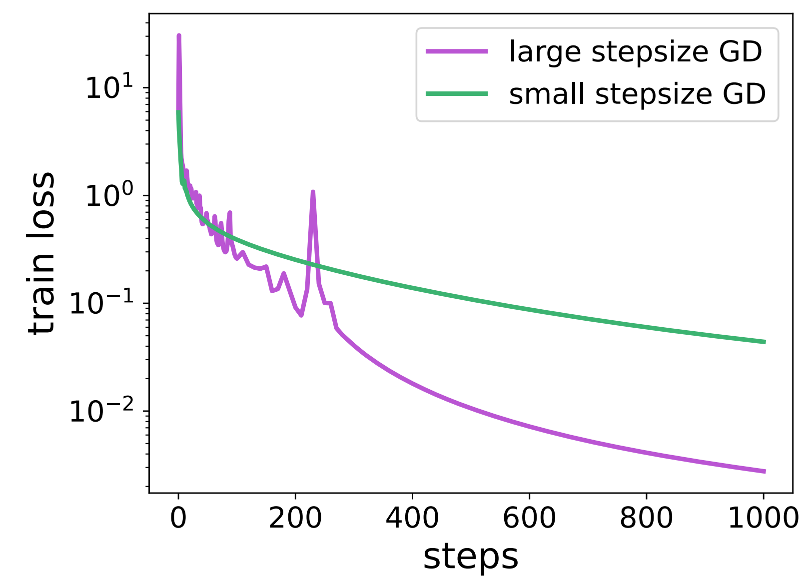
unstable optimization impossible

observable & provable

unstable optimization observed







Logistic regression

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

Gradient Descent $$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

Logistic regression

empirical risk $\quad L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln\big(1 + \exp(-y_i x_i^\top \theta)\big)$

Gradient Descent $\quad \theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Assumption. $\|x_i\| \leq 1,\ y_i \in \{\pm 1\},\ i = 1,\ldots,n$

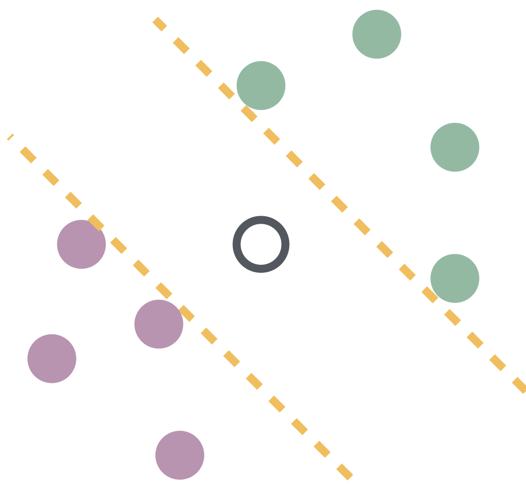
Logistic regression

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

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Assumption. $\|x_i\| \leq 1$, $y_i \in \{\pm 1\}$, $i = 1, \ldots, n$

- $\exists$ unit vector θ^*, $\min_i y_i x_i^\top \theta^* \geq \gamma > 0$



Logistic regression

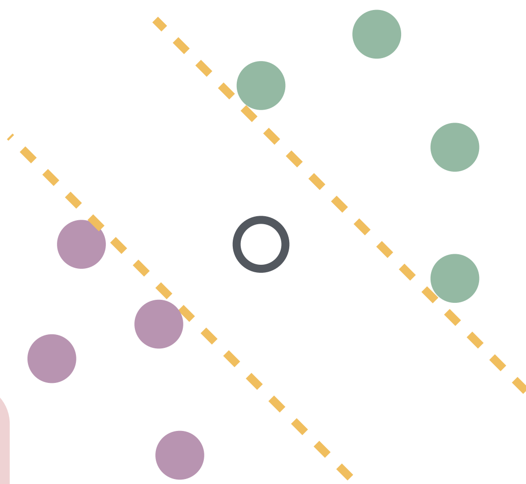
empirical risk $\quad L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i x_i^\top \theta)\right)$

Gradient Descent $\quad \theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Assumption. $\|x_i\| \leq 1,\ y_i \in \{\pm 1\},\ i = 1, \ldots, n$

- $\exists$ unit vector θ^*, $\min_i y_i x_i^\top \theta^* \geq \gamma > 0$

implied by overparameterization



Logistic regression

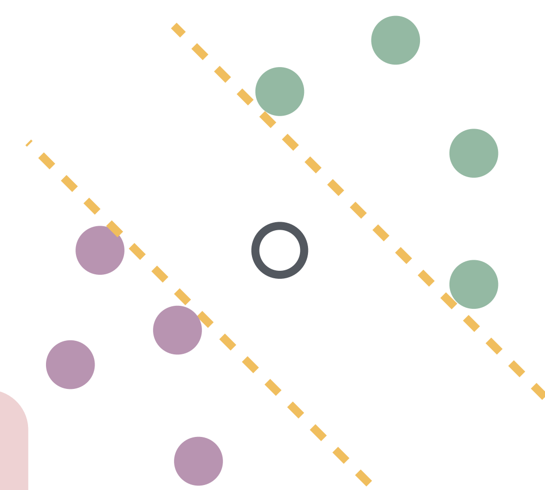
empirical risk $\quad L(\theta) = \dfrac{1}{n}\sum_{i=1}^{n} \ln\big(1 + \exp(-y_i x_i^\top \theta)\big)$

Gradient Descent $\quad \theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Assumption. $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1,\ldots,n$

- $\exists$ unit vector θ^*, $\min_i y_i x_i^\top \theta^* \ge \gamma > 0$
- η, t grow while $n, \gamma = \Theta(1)$

implied by overparameterization



Logistic regression

smooth, convex
non-strongly convex

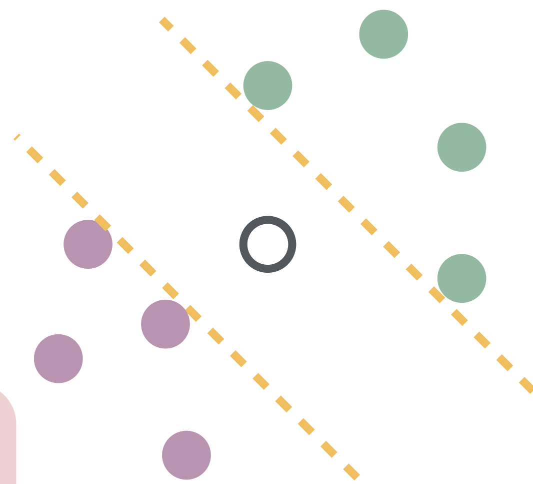
empirical risk $L(\theta) = \dfrac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i x_i^\top \theta)\right)$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Assumption. $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1,\ldots,n$

- $\exists$ unit vector θ^*, $\min_i y_i x_i^\top \theta^* \ge \gamma > 0$
- η, t grow while $n, \gamma = \Theta(1)$

implied by
overparameterization

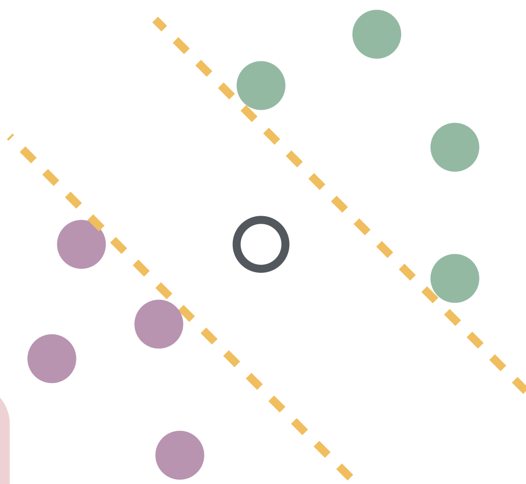


Logistic regression

smooth, convex
non-strongly convex

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

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$



Assumption. $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1, \ldots, n$

- $\exists$ unit vector θ^*, $\min_i y_i x_i^\top \theta^* \ge \gamma > 0$
- η, t grow while $n, \gamma = \Theta(1)$

implied by
overparameterization

Classical theory. For $\eta = \Theta(1)$, $L(\theta_t) \downarrow$ and $L(\theta_t) = \tilde{O}(1/t)$

Logistic regression

smooth, convex
non-strongly convex

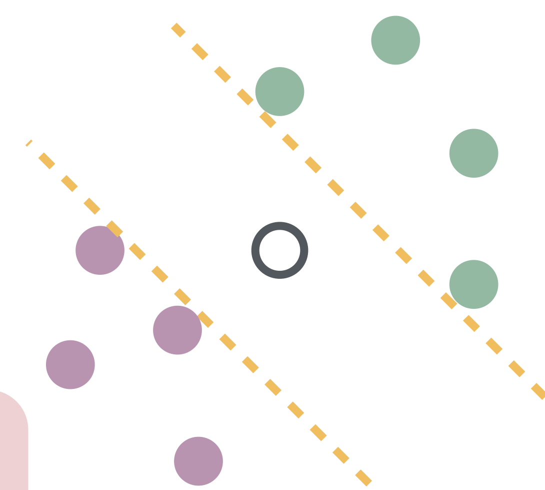
empirical risk $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \exp(-y_i x_i^\top \theta)\right)$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$

Assumption. $\|x_i\| \le 1,\ y_i \in \{\pm 1\},\ i = 1,\ldots,n$

- $\exists$ unit vector θ^*, $\min_i y_i x_i^\top \theta^* \ge \gamma > 0$
- η, t grow while $n, \gamma = \Theta(1)$

implied by
overparameterization



Classical theory. For $\eta = \Theta(1)$, $L(\theta_t) \downarrow$ and $L(\theta_t) = \tilde{O}(1/t)$

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

Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \tilde{O}(1/T^2)$$

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

Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \tilde{O}(1/T^2)$$

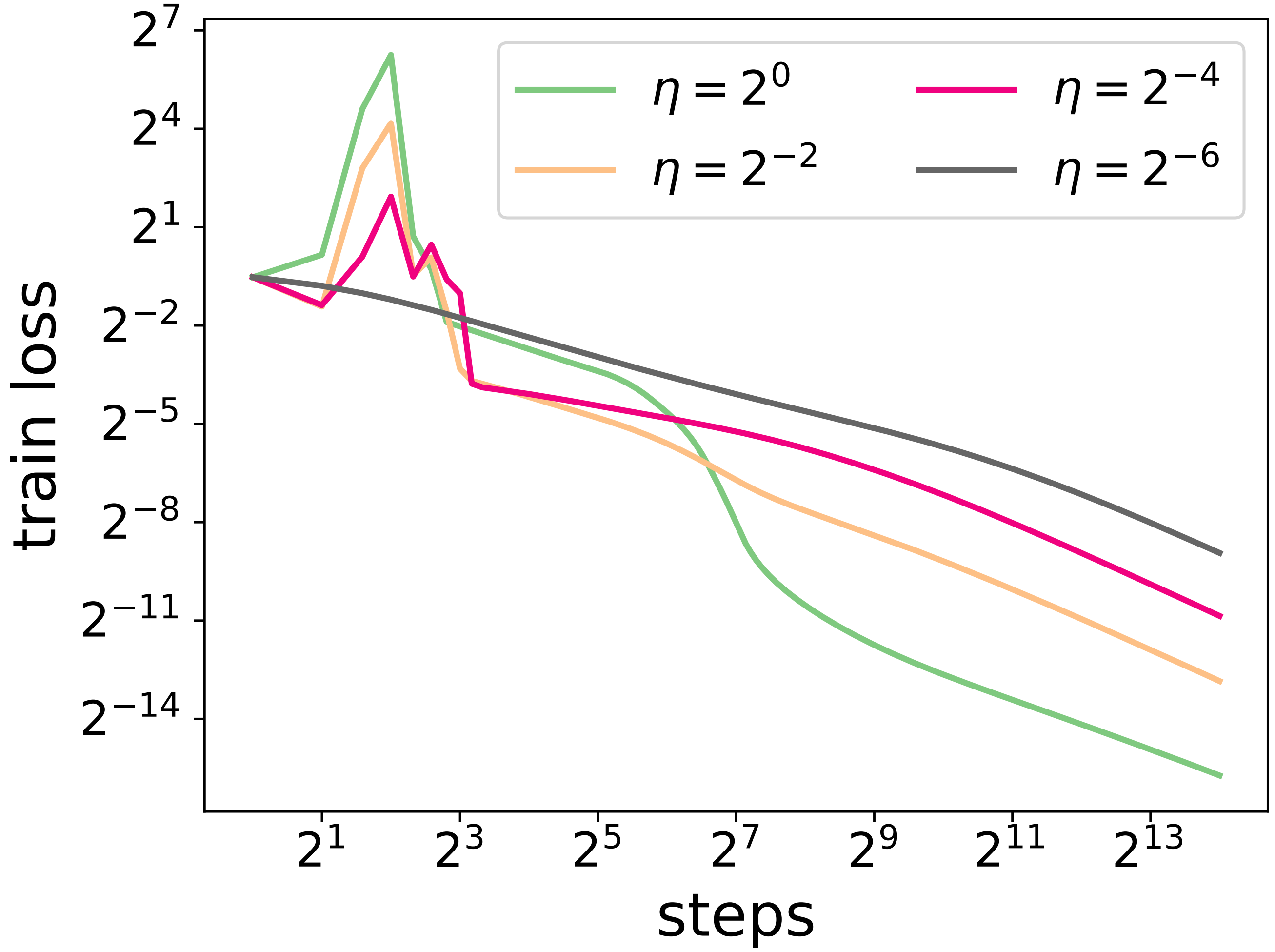
match Nesterov

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

Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \tilde{O}(1/T^2) \quad \text{match Nesterov}$$



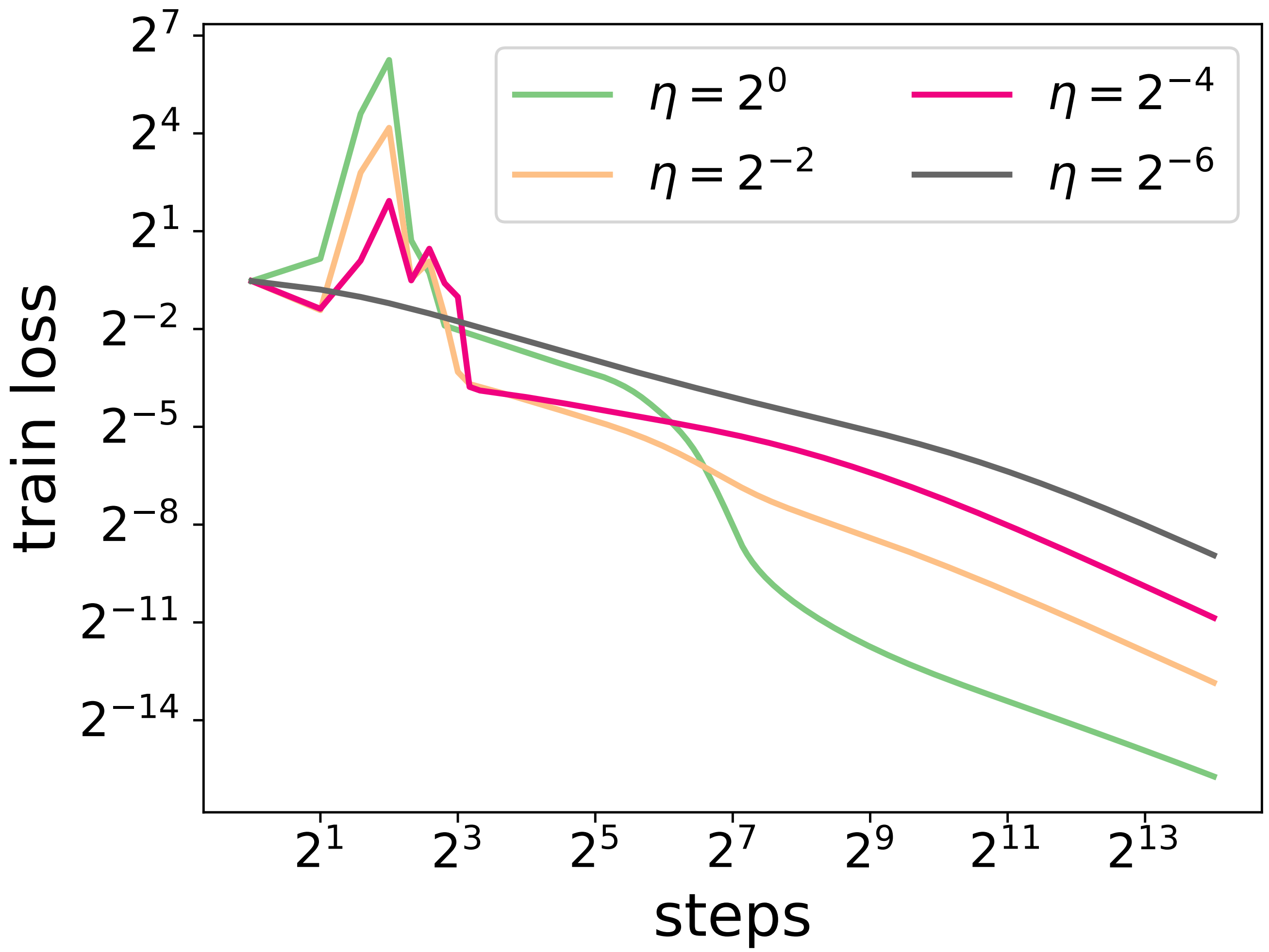
Wu, Bartlett, Telgarsky, Yu. "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency." COLT 2024

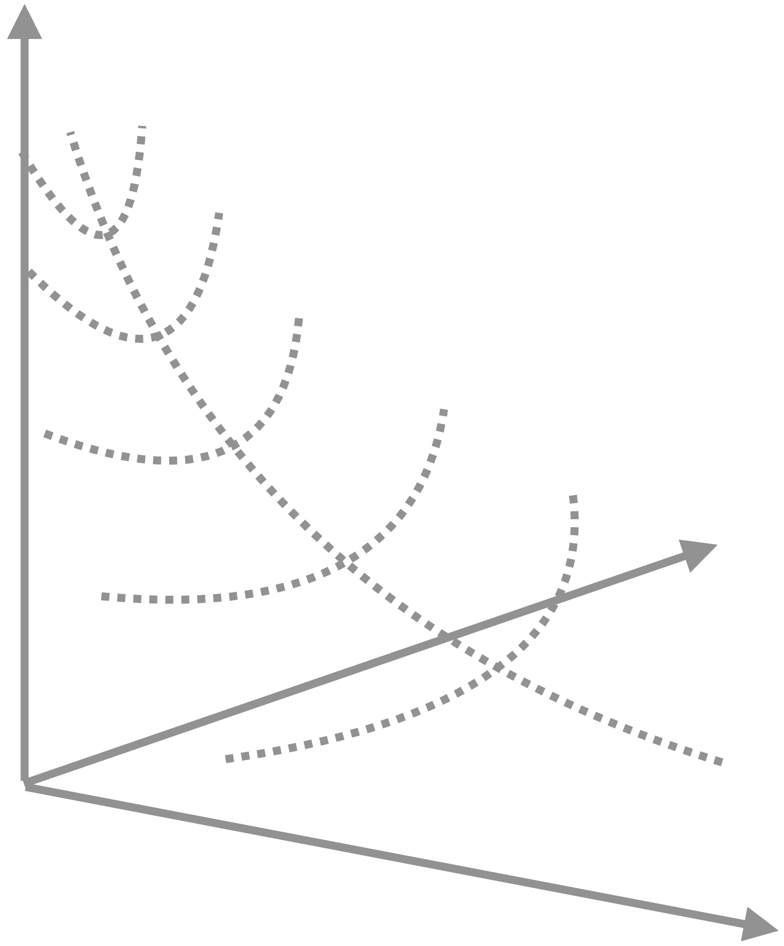
Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \boxed{\tilde{O}(1/T^2)}$$

match Nesterov



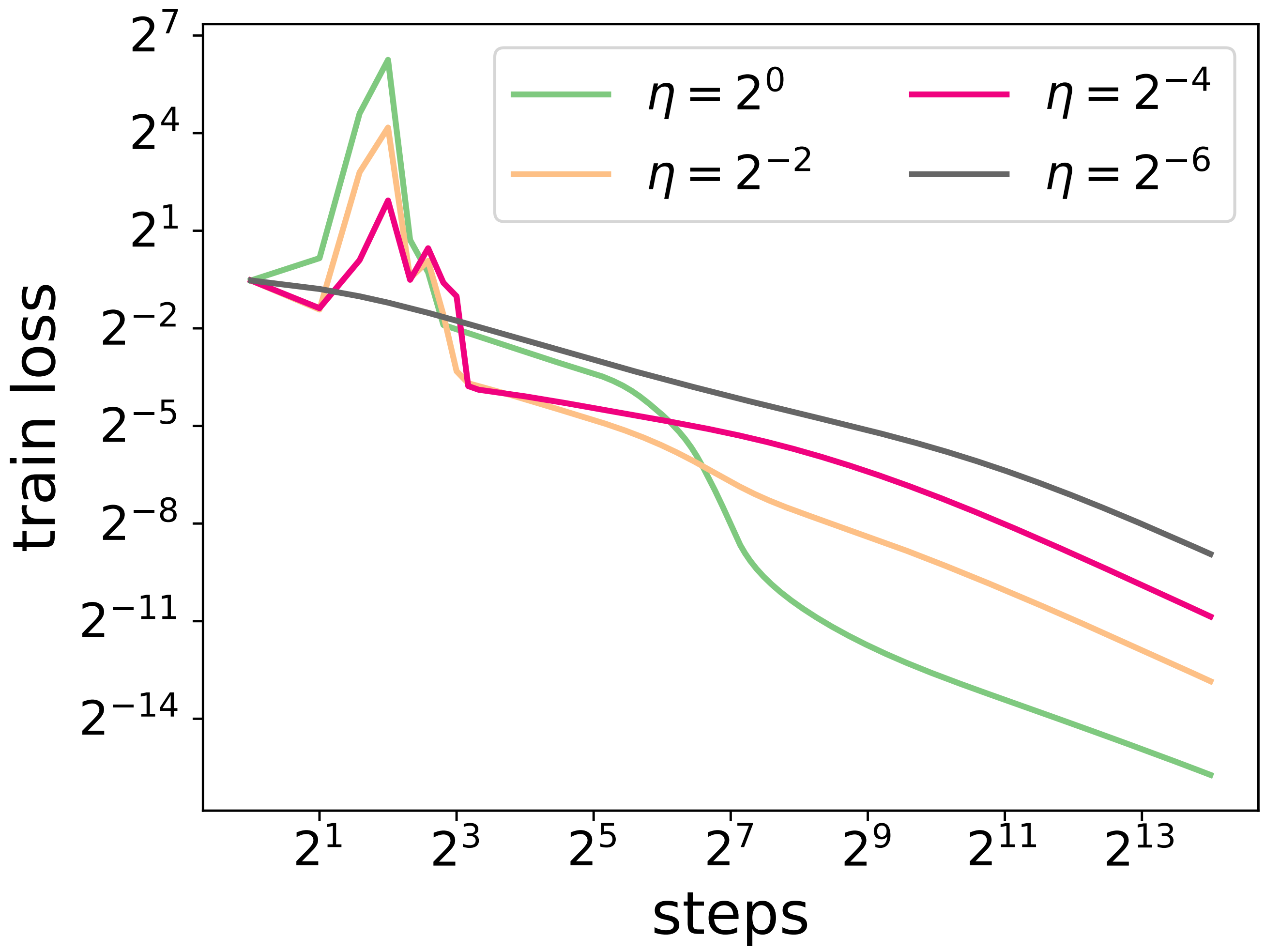


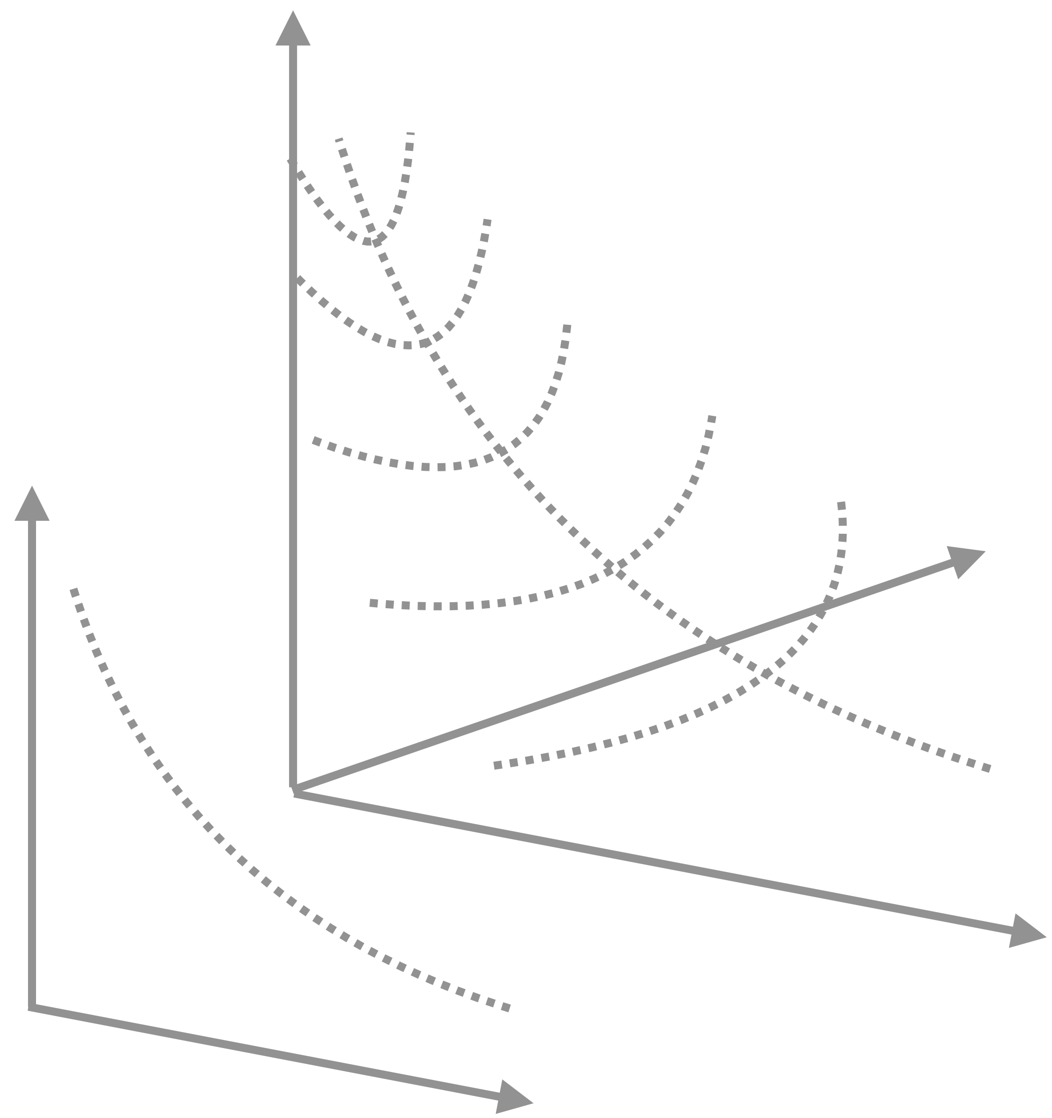
Wu, Bartlett, Telgarsky, Yu. "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency." COLT 2024

Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \tilde{O}(1/T^2)$$ match Nesterov



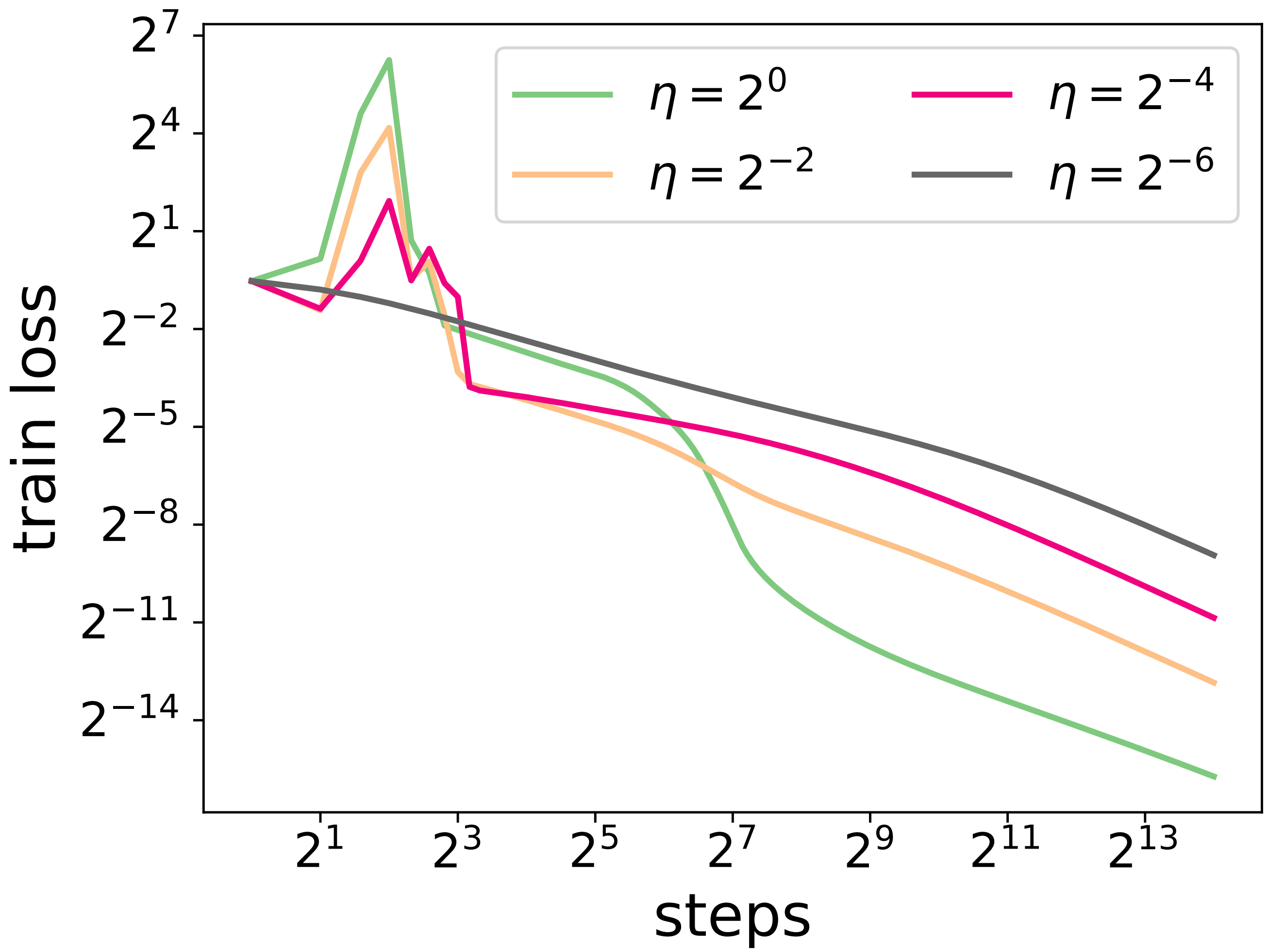


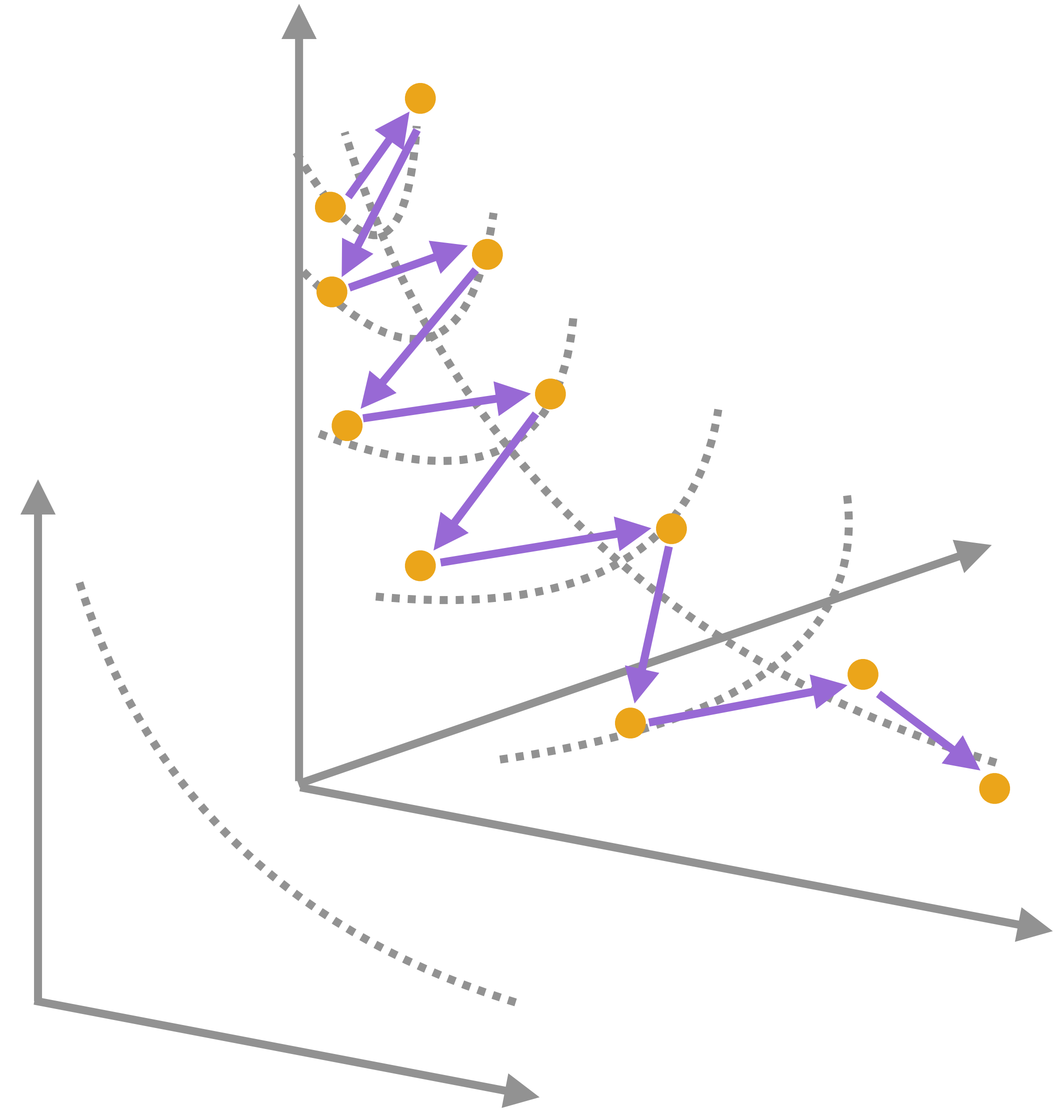
Wu, Bartlett, Telgarsky, Yu. "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency." COLT 2024

Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \tilde{O}(1/T^2) \quad \text{match Nesterov}$$



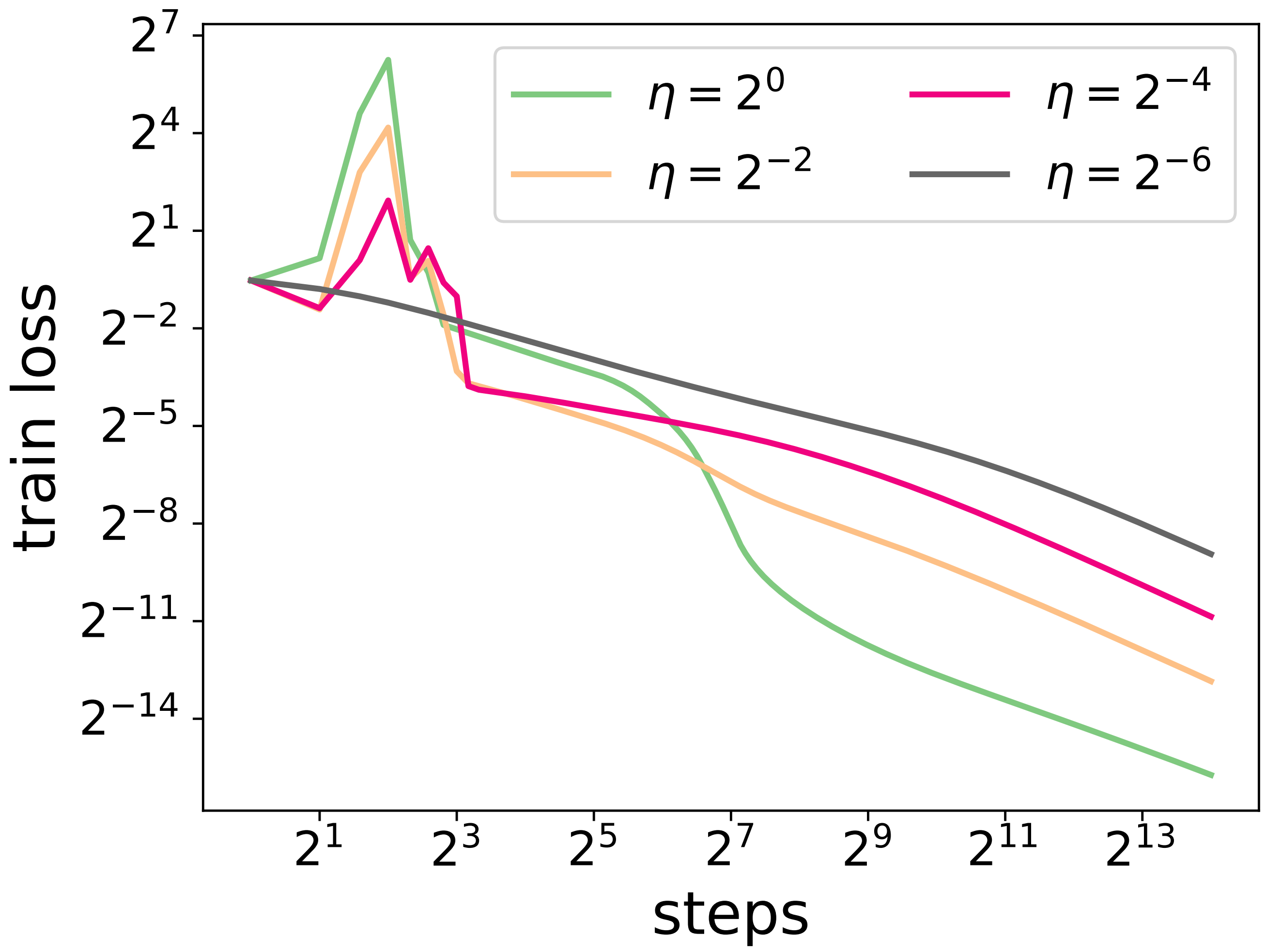


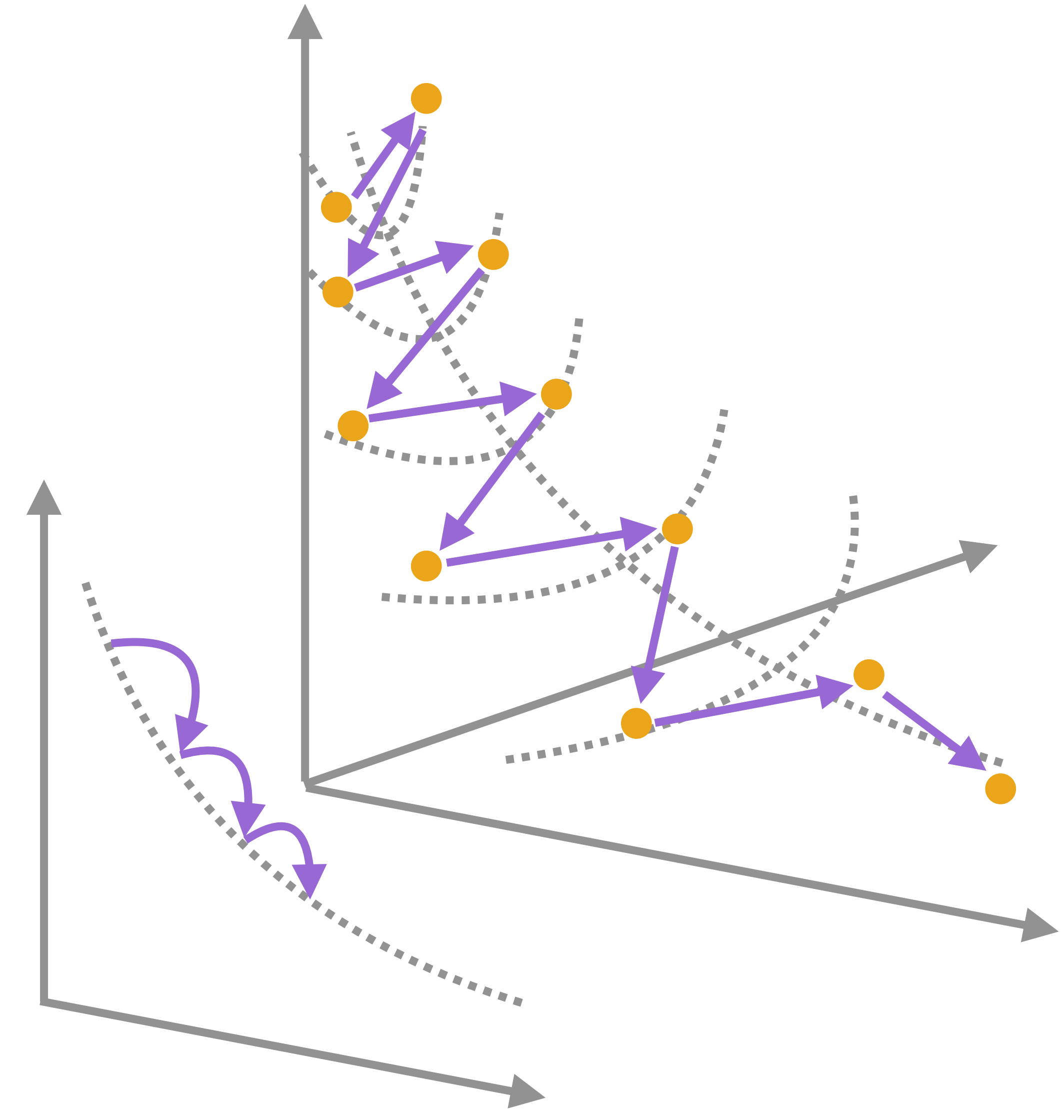
Wu, Bartlett, Telgarsky, Yu. "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency." COLT 2024

Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \tilde{O}(1/T^2)$$ match Nesterov



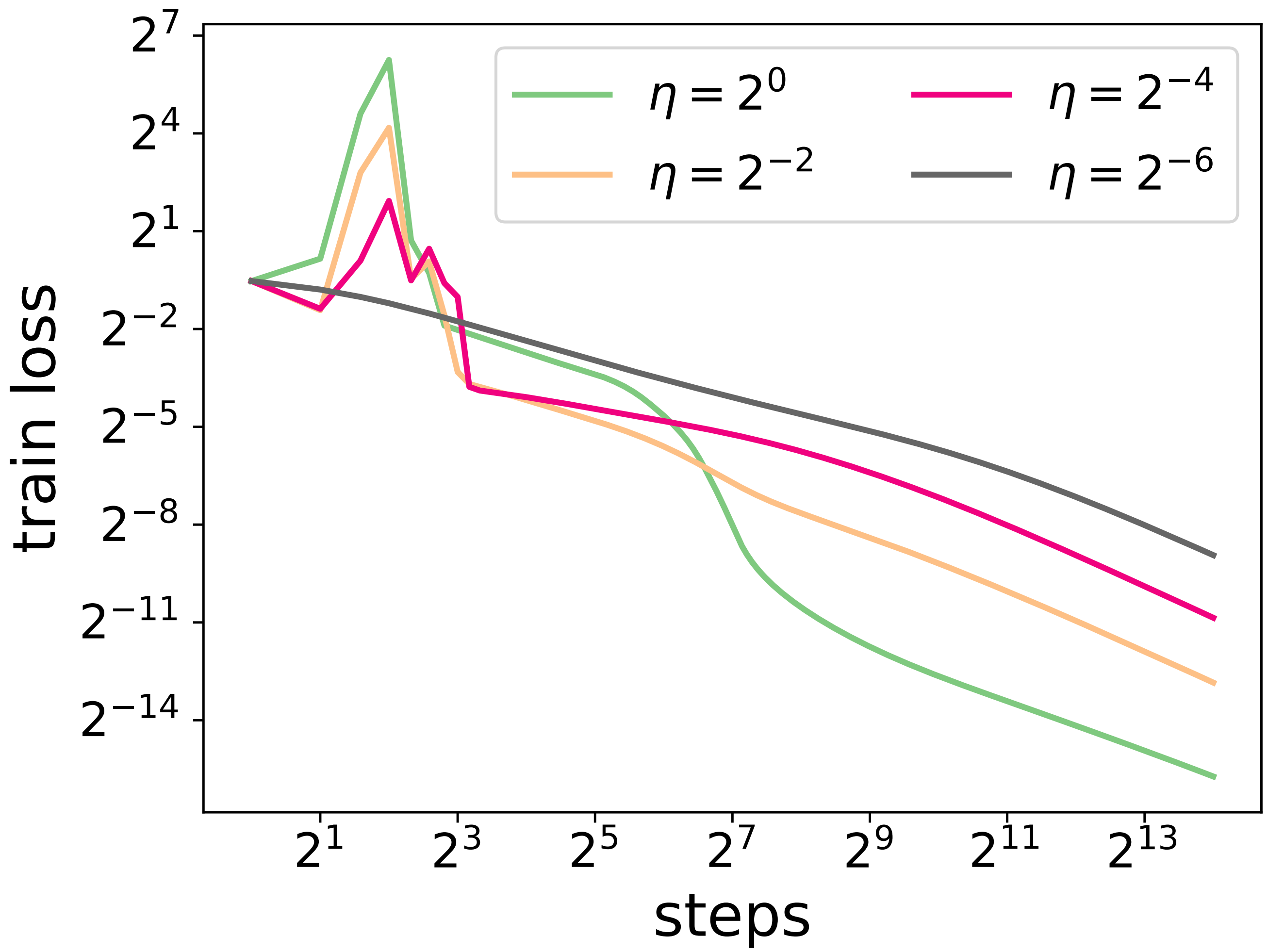


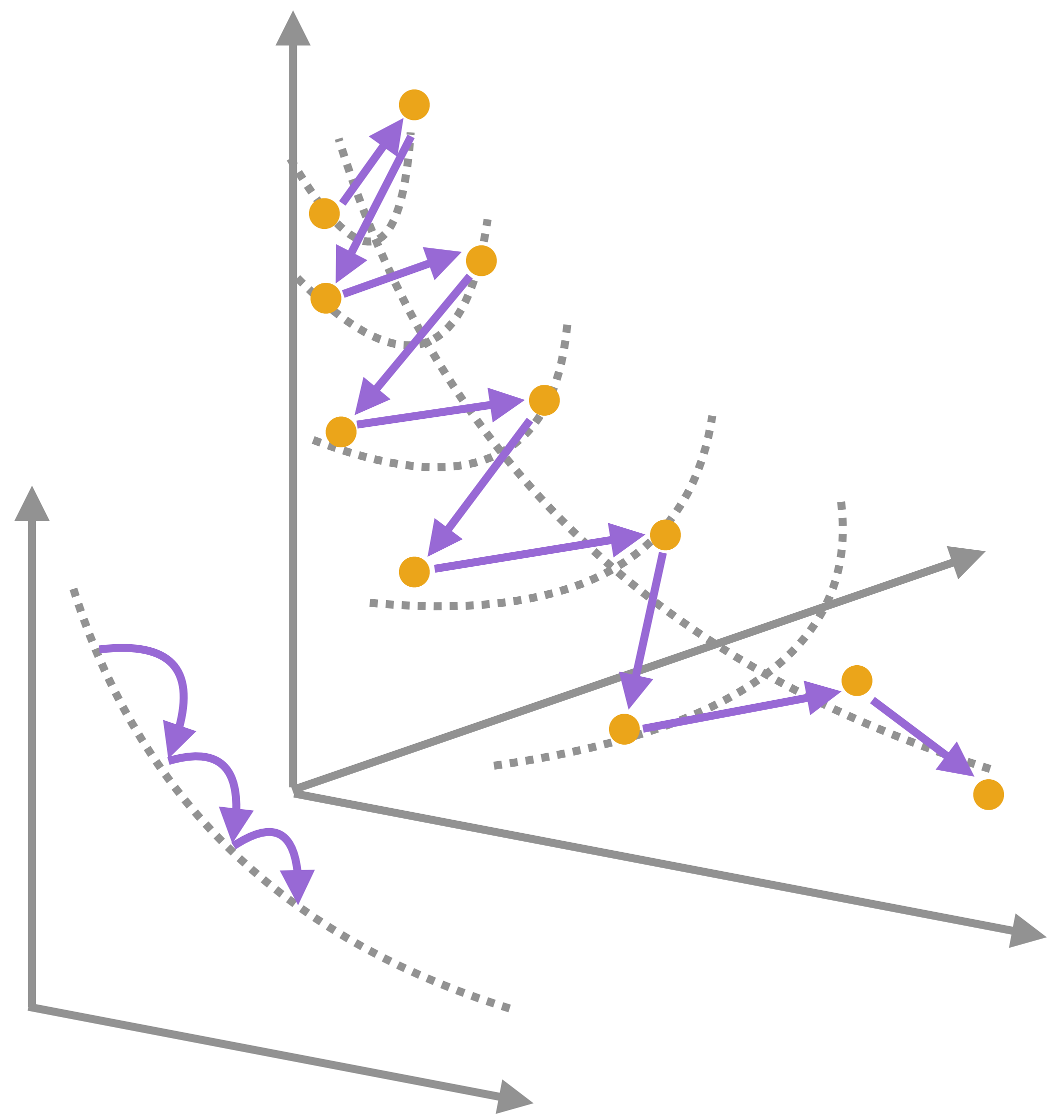
Wu, Bartlett, Telgarsky, Yu. "Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency." COLT 2024

Large stepsize accelerates GD

Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

$$L(\theta_T) = \tilde{O}(1/T^2)$$ match Nesterov

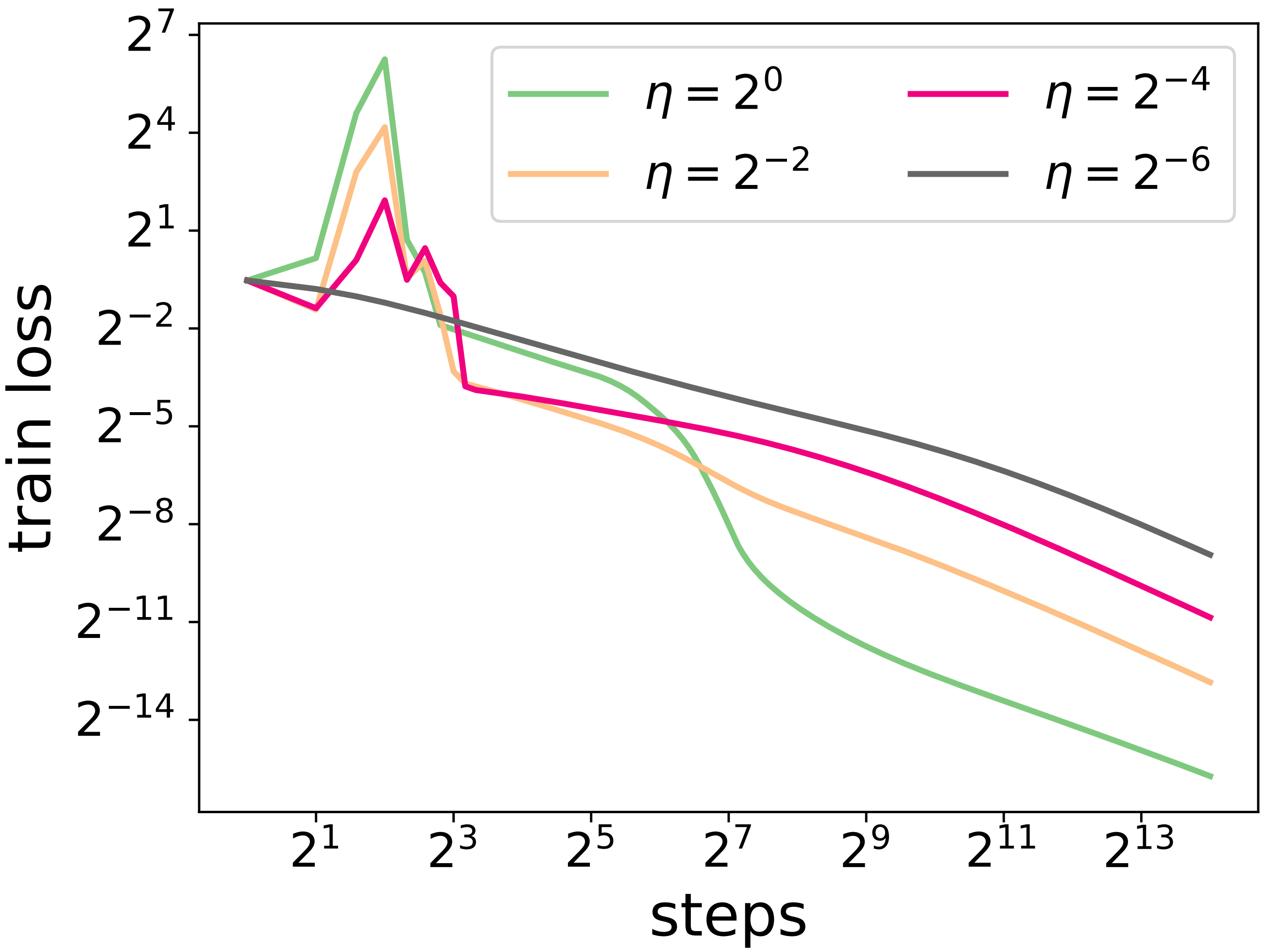




"open valley" as mental picture

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

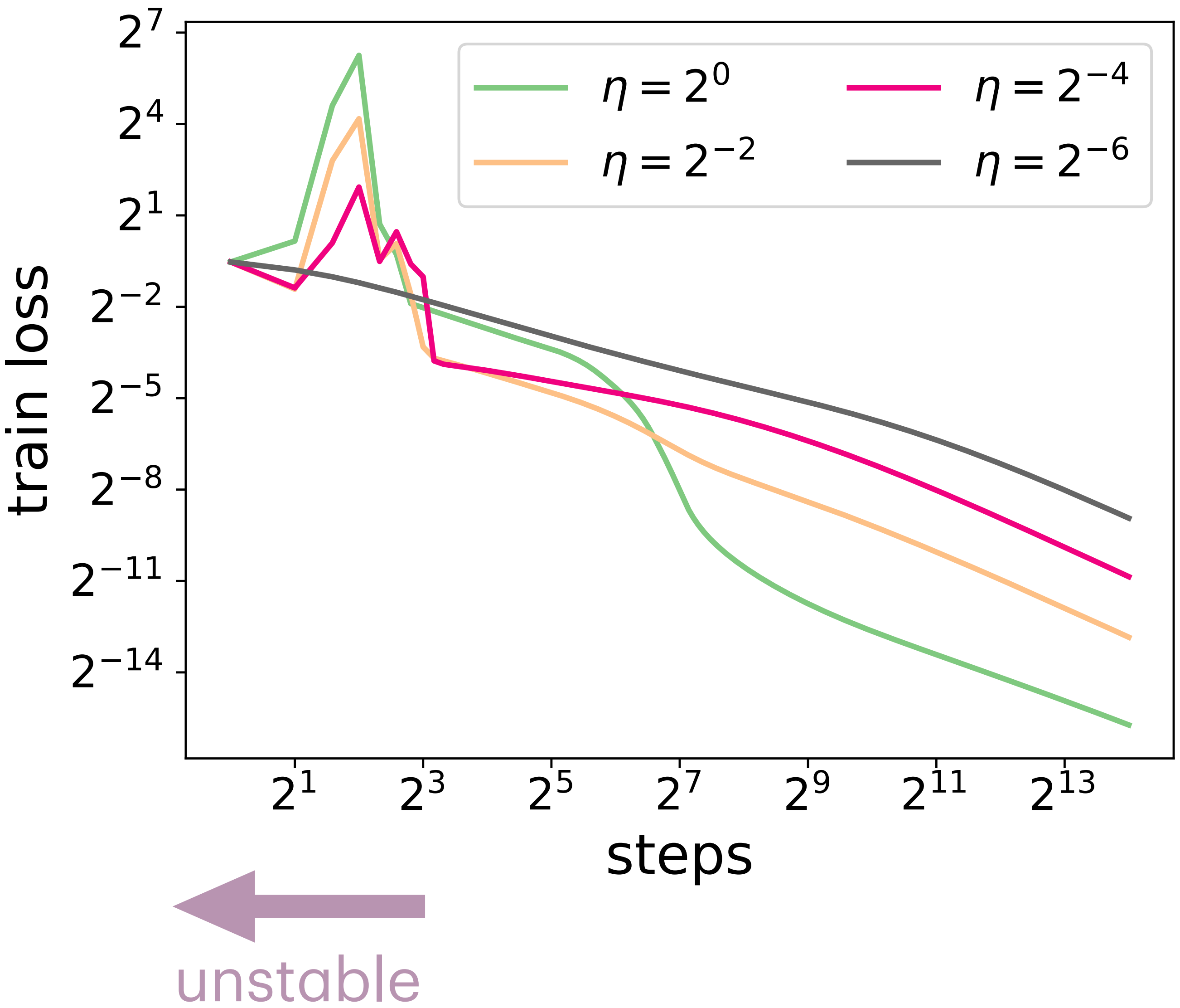
More details



Theorem. For all η, we have

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

More details



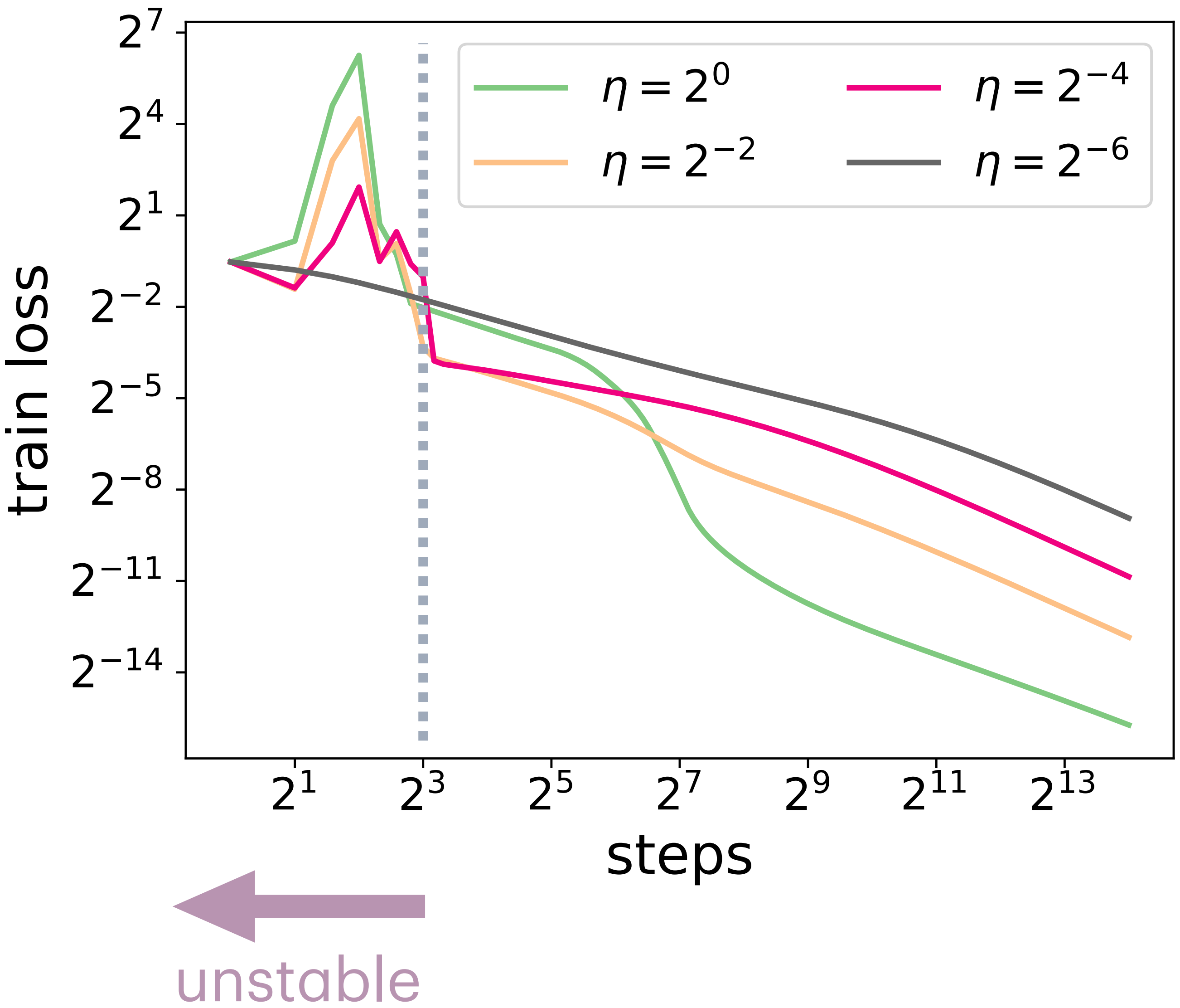
Theorem. For all η, we have

- **Unstable phase:**

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\theta_k) = \tilde{O}\left(\frac{1+\eta^2}{\eta t}\right)$$

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

More details



Theorem. For all η, we have

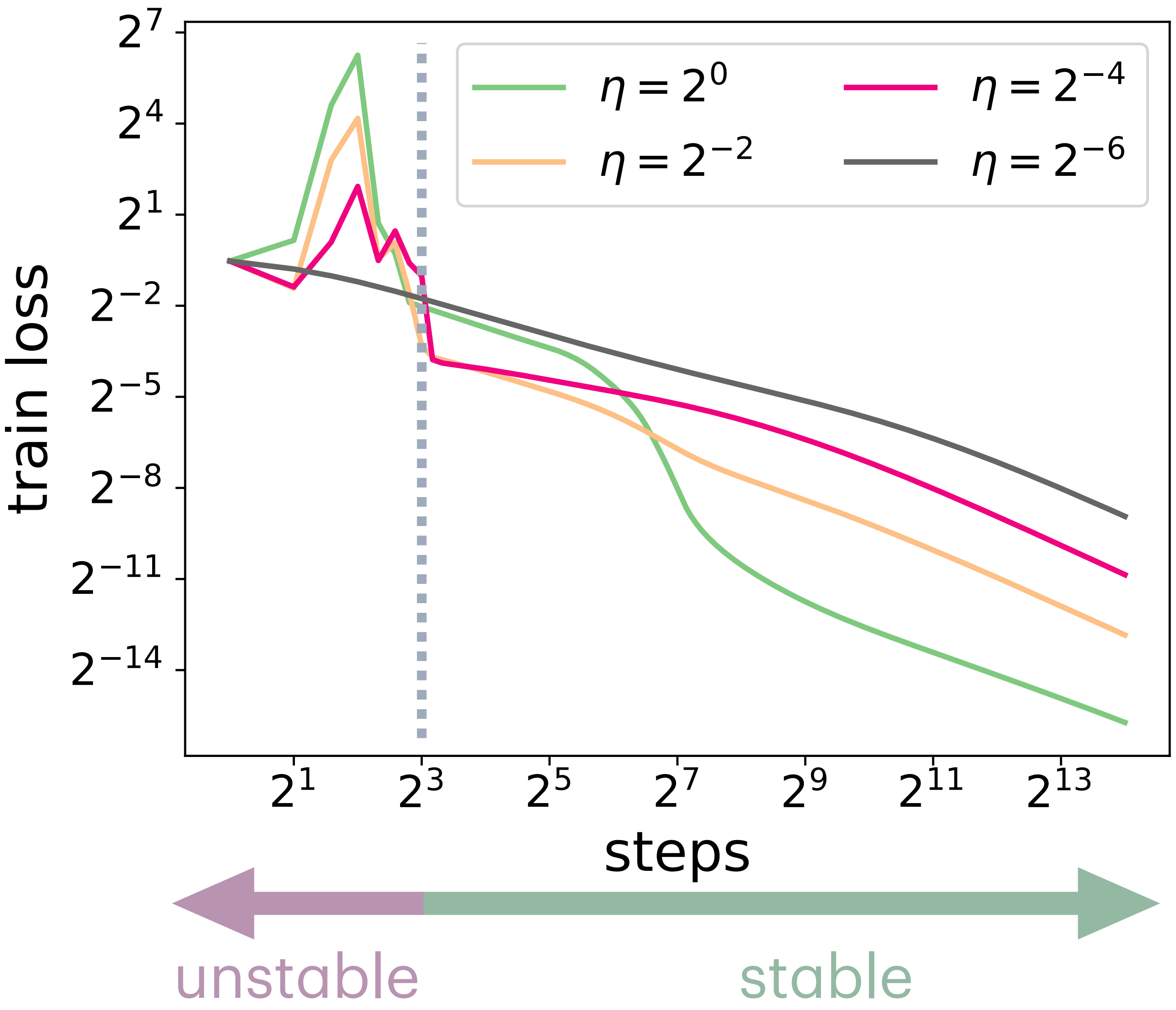
- **Unstable phase:**

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\theta_k) = \tilde{O}\left(\frac{1+\eta^2}{\eta t}\right)$$

- **Phase transition:** Unstable phase ends in $\tau = O(\eta)$ steps

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

More details



Theorem. For all η, we have

- **Unstable phase:**

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\theta_k) = \tilde{O}\left(\frac{1+\eta^2}{\eta t}\right)$$

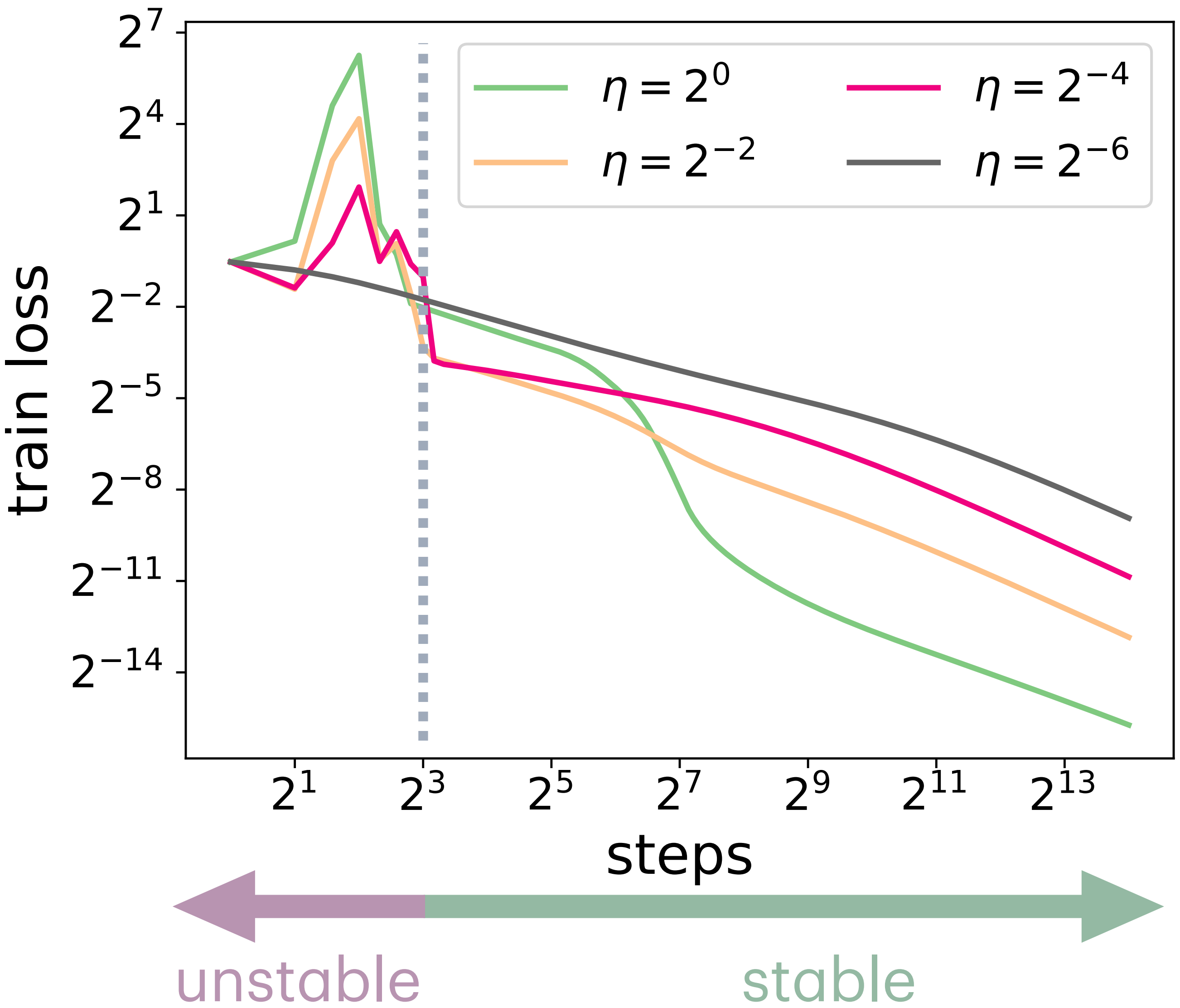
- **Phase transition:** Unstable phase ends in $\tau = O(\eta)$ steps

- **Stable phase:** $L(\theta_{\tau+t}) \downarrow$ and

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

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

More details



Theorem. For all η, we have

- **Unstable phase:**

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\theta_k) = \tilde{O}\left(\frac{1+\eta^2}{\eta t}\right)$$

- **Phase transition:** Unstable phase ends in $\tau = O(\eta)$ steps

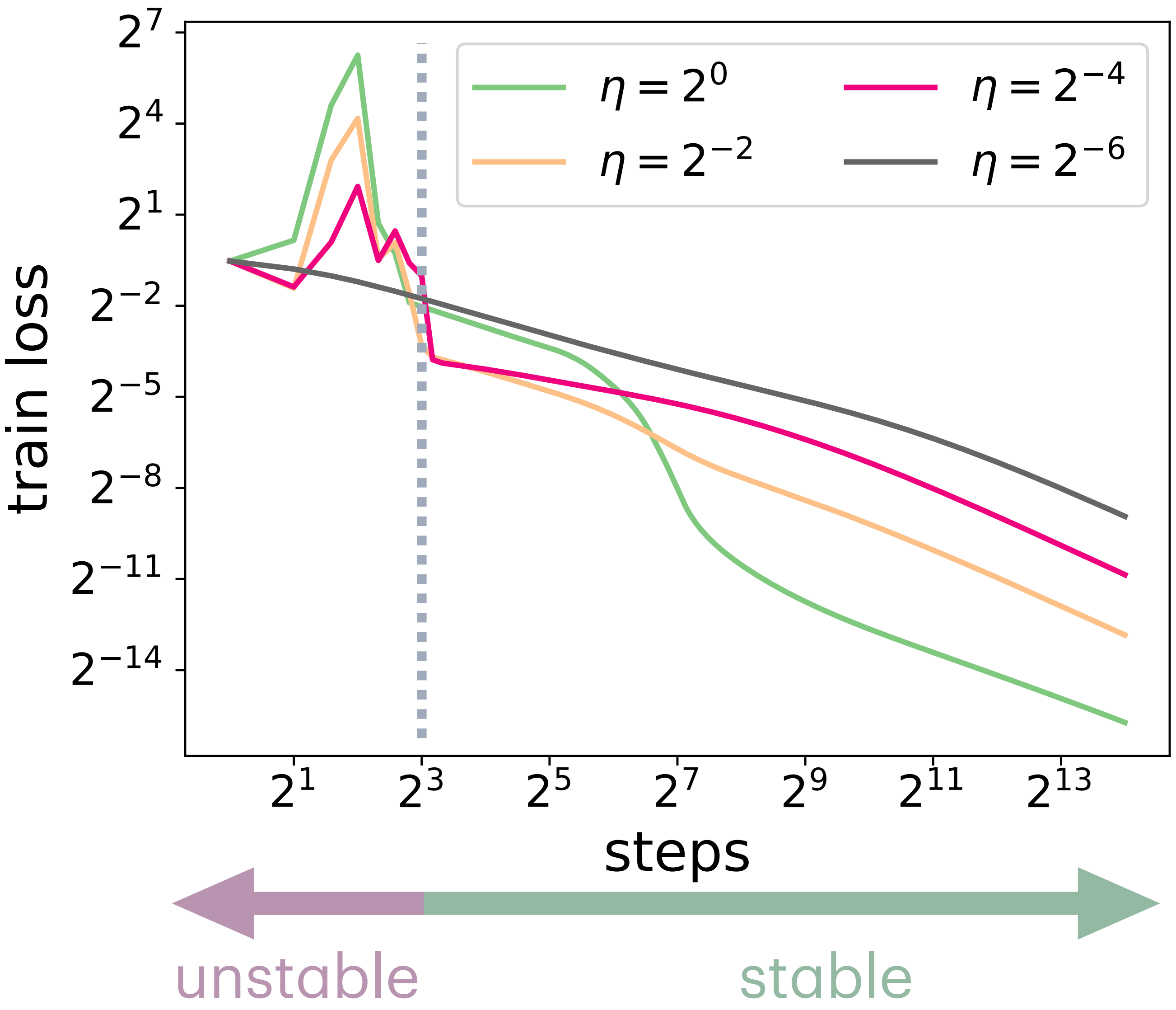
- **Stable phase:** $L(\theta_{\tau+t}) \downarrow$ and

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

tradeoff

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

More details



Theorem. For all η, we have

- **Unstable phase:**

$$\frac{1}{t}\sum_{k=0}^{t-1} L(\theta_k) = \tilde{O}\left(\frac{1+\eta^2}{\eta t}\right)$$

- **Phase transition:** Unstable phase ends in $\tau = O(\eta)$ steps

- **Stable phase:** $L(\theta_{\tau+t}) \downarrow$ and

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

tradeoff

"instability" is needed for acceleration

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

Regularized logistic regression

regularized empirical risk $\tilde{L}(\theta) = L(\theta) + \frac{\lambda}{2}||\theta||^2$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla \tilde{L}(\theta_t)$

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

Regularized logistic regression

regularized empirical risk $\tilde{L}(\theta) = L(\theta) + \boxed{\frac{\lambda}{2}||\theta||^2}$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla \tilde{L}(\theta_t)$

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

Regularized logistic regression

regularized empirical risk $\tilde{L}(\theta) = L(\theta) + \boxed{\frac{\lambda}{2}\|\theta\|^2}$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla \tilde{L}(\theta_t)$

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

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

Regularized logistic regression

regularized empirical risk $\tilde{L}(\theta) = L(\theta) + \boxed{\frac{\lambda}{2}\|\theta\|^2}$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla \tilde{L}(\theta_t)$

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

Classical theory. For $\eta = \Theta(1)$, we have $\tilde{L}(\theta_t) \downarrow$ and

$$\tilde{L}(\theta_t) - \min \tilde{L} \leq \epsilon \text{ for } t = O\left(\kappa \ln(1/\epsilon)\right)$$

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

Regularized logistic regression

regularized empirical risk $\tilde{L}(\theta) = L(\theta) + \boxed{\frac{\lambda}{2}\|\theta\|^2}$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla \tilde{L}(\theta_t)$

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

Classical theory. For $\eta = \Theta(1)$, we have $\tilde{L}(\theta_t) \downarrow$ and

$\tilde{L}(\theta_t) - \min \tilde{L} \le \epsilon$ for $t = O\big(\kappa \ln(1/\epsilon)\big)$ Nesterov: $\tilde{O}\big(\sqrt{\kappa}\big)$

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

Regularized logistic regression

regularized empirical risk $\tilde{L}(\theta) = L(\theta) + \boxed{\frac{\lambda}{2}\|\theta\|^2}$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla \tilde{L}(\theta_t)$

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

Classical theory. For $\eta = \Theta(1)$, we have $\tilde{L}(\theta_t) \downarrow$ and

$\tilde{L}(\theta_t) - \min \tilde{L} \le \epsilon$ for $t = O\big(\kappa \ln(1/\epsilon)\big)$ Nesterov: $\tilde{O}\big(\sqrt{\kappa}\big)$

Our theory. Assume $\lambda \le \Theta(1)$. For $\eta = \Theta\big(\sqrt{\kappa}\big)$, we have

$\tilde{L}(\theta_t) - \min \tilde{L} \le \epsilon$ for $t = O\big(\sqrt{\kappa}\ln(1/\epsilon)\big)$

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

Regularized logistic regression

regularized empirical risk $\tilde{L}(\theta) = L(\theta) + \boxed{\frac{\lambda}{2}\|\theta\|^2}$

Gradient Descent $\theta_{t+1} = \theta_t - \eta \nabla \tilde{L}(\theta_t)$

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

Classical theory. For $\eta = \Theta(1)$, we have $\tilde{L}(\theta_t) \downarrow$ and

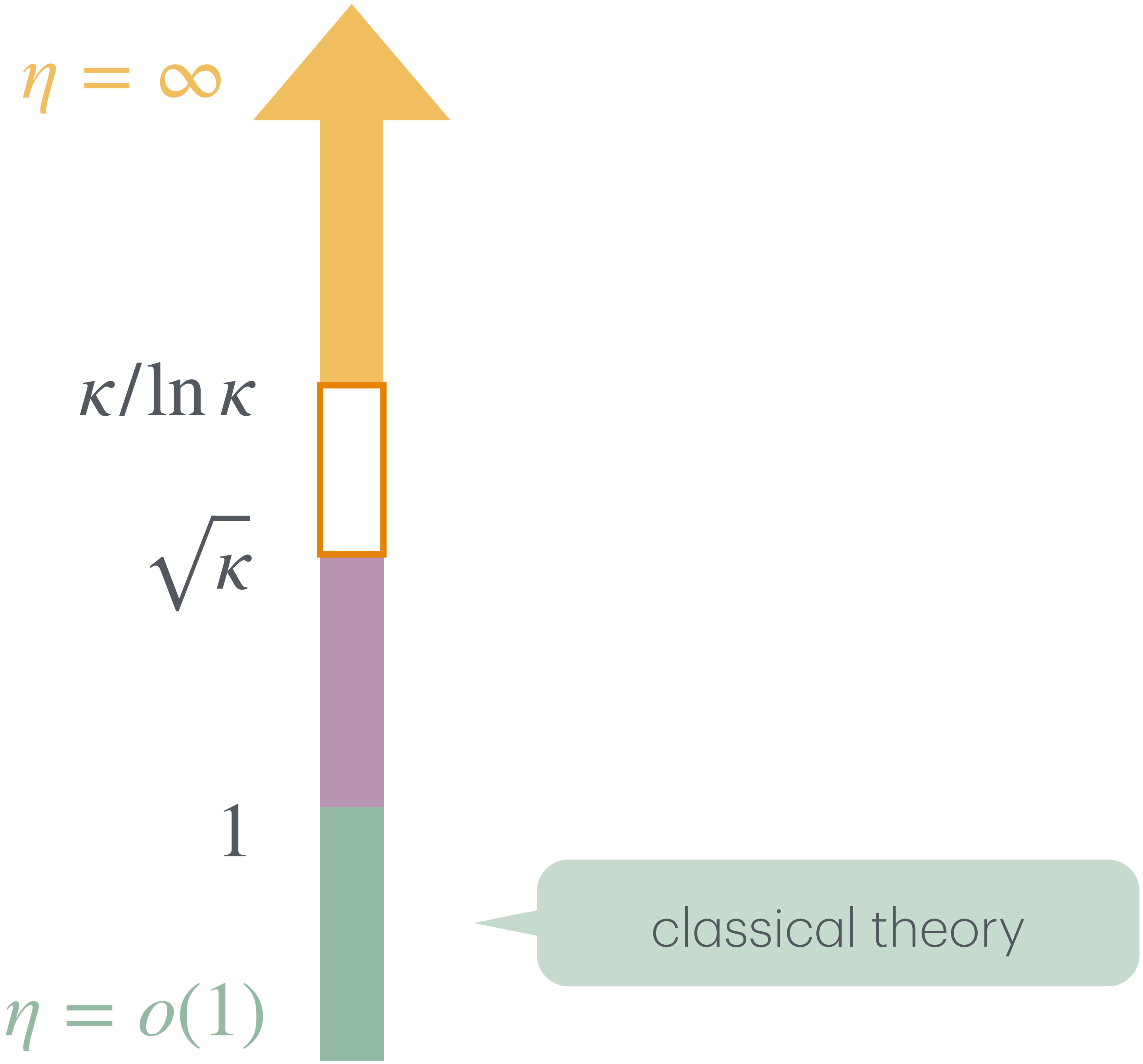
$\tilde{L}(\theta_t) - \min \tilde{L} \le \epsilon$ for $t = O\big(\kappa \ln(1/\epsilon)\big)$ Nesterov: $\tilde{O}\big(\sqrt{\kappa}\big)$

Our theory. Assume $\lambda \le \Theta(1)$. For $\eta = \Theta\big(\sqrt{\kappa}\big)$, we have

$\tilde{L}(\theta_t) - \min \tilde{L} \le \epsilon$ for $t = O\big(\sqrt{\kappa} \ln(1/\epsilon)\big)$ match Nesterov via large stepsize

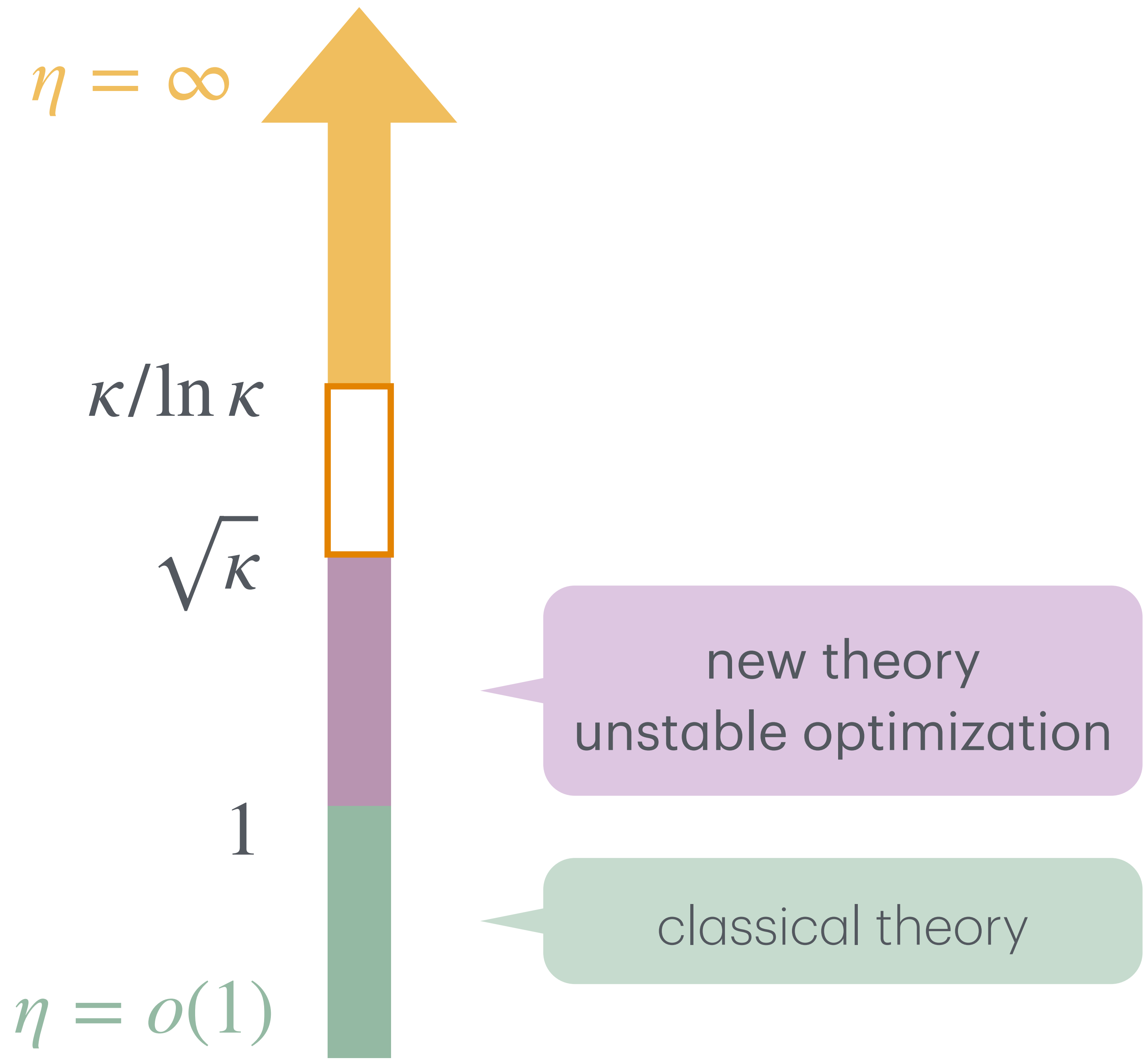
Wu, Marion, Bartlett. "Large stepsizes accelerate gradient descent for regularized logistic regression." NeurIPS 2025

Stepsize diagram revisited



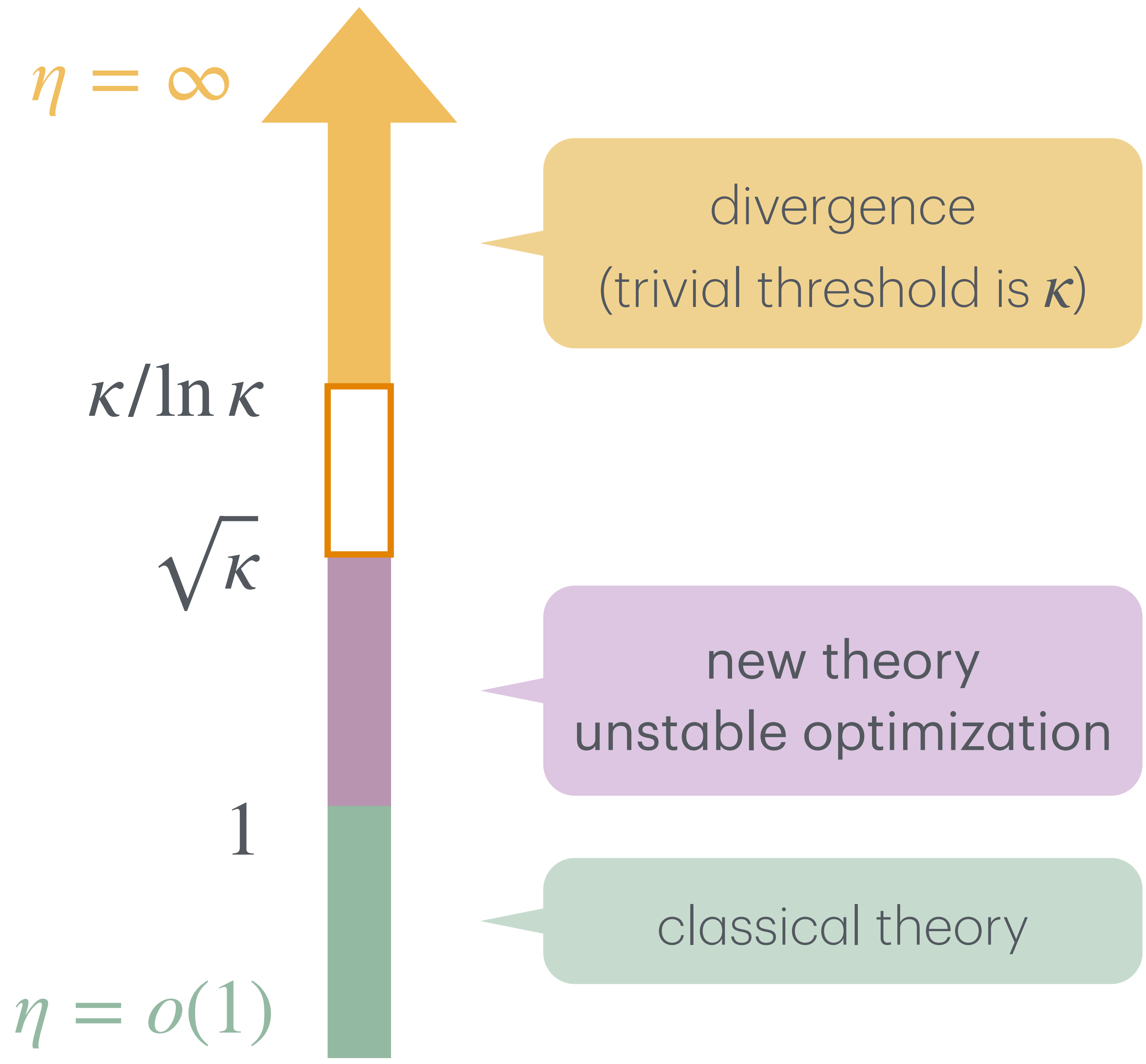
Wu, Marion, Bartlett. "Large stepsizes accelerate gradient descent for regularized logistic regression." NeurIPS 2025

Stepsize diagram revisited



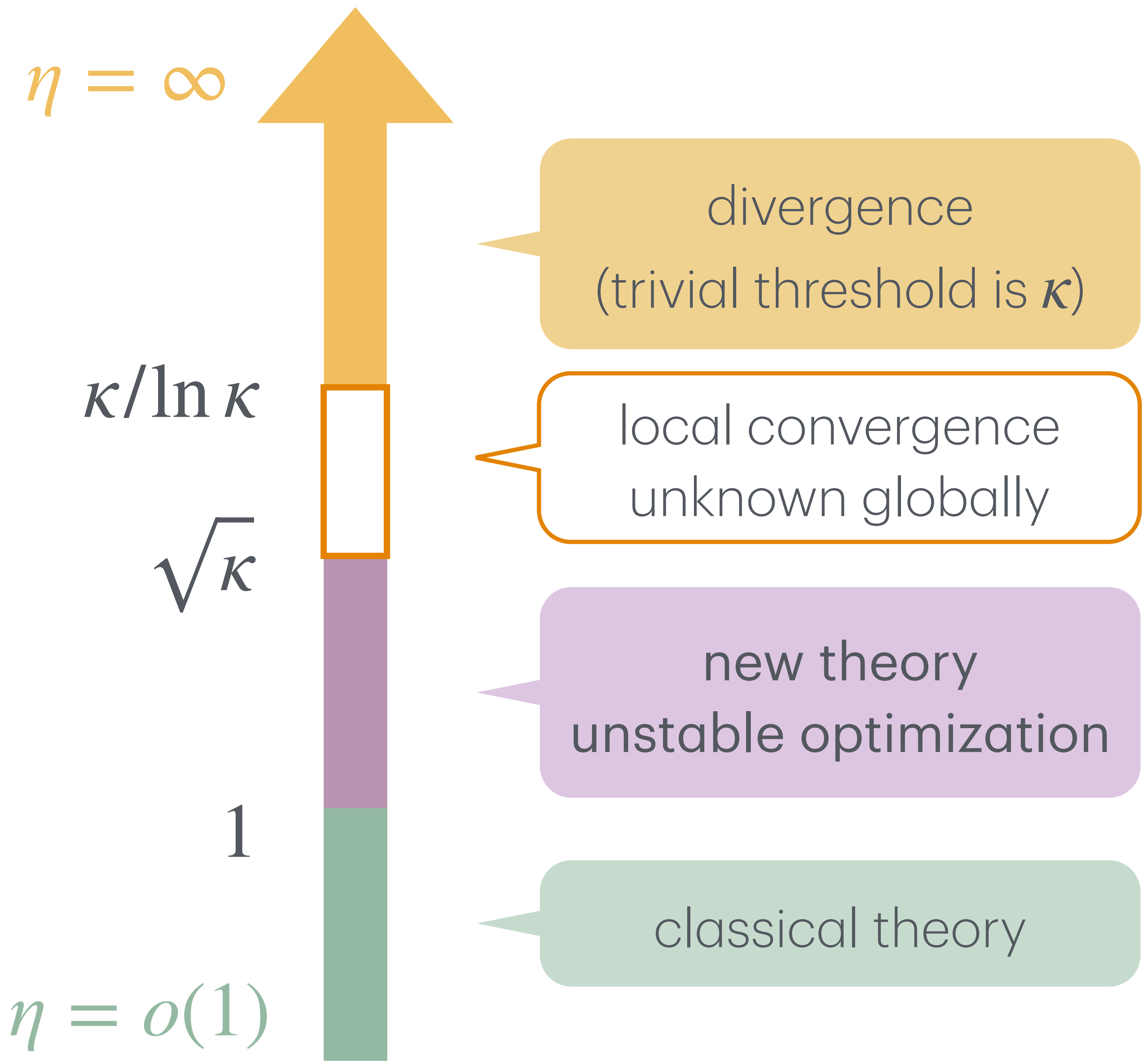
Wu, Marion, Bartlett. "Large stepsizes accelerate gradient descent for regularized logistic regression." NeurIPS 2025

Stepsize diagram revisited



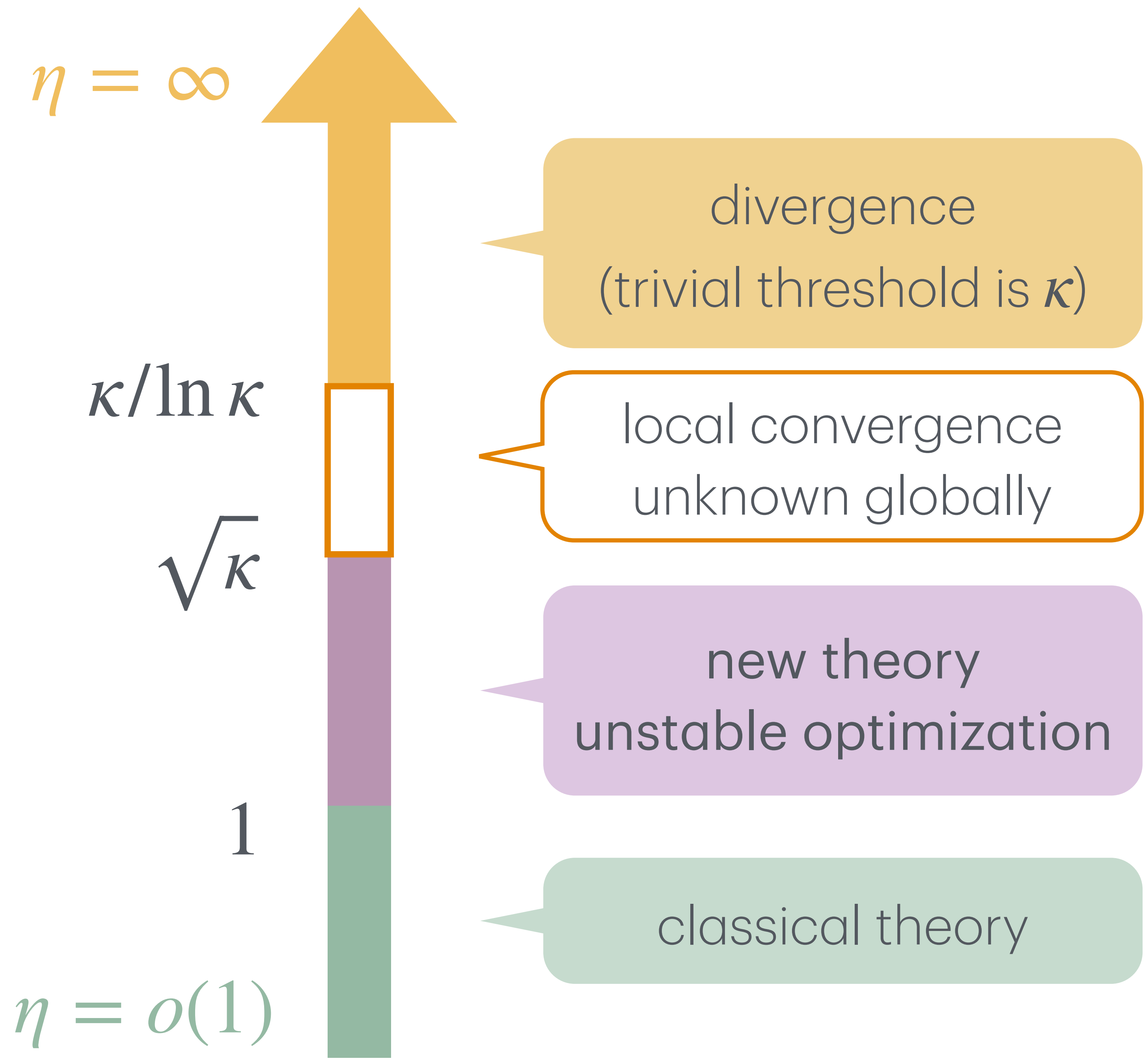
Wu, Marion, Bartlett. "Large stepsizes accelerate gradient descent for regularized logistic regression." NeurIPS 2025

Stepsize diagram revisited



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

Stepsize diagram revisited



three extensions

- general loss functions (WBTY'24)
- *two-layer networks (WBTY'24, CWMLB'24)
- large adaptive stepsizes (ZWLB'25)

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

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

Contribution 2: implicit regularization

gradient descent dominates ridge regression in linear regression

- “Risk comparisons in linear regression: implicit regularization dominates explicit regularization”
  W, Peter Bartlett, Jason Lee, Sham Kakade, Bin Yu
  arXiv 2025.09

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**

e.g., norm regularization

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent
$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**
e.g., norm regularization

"new" mechanism: **implicit control**
e.g., GD path has low-complexity

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent
$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**
e.g., norm regularization

"new" mechanism: **implicit control**
e.g., GD path has low-complexity

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**
e.g., norm regularization

"new" mechanism: **implicit control**
e.g., GD path has low-complexity

Bartlett. "For valid generalization the size of the weights is more important than the size of the network." NeurIPS 1996
Bühlmann, Yu. "Boosting with the L_2 loss: regression and classification." JASA 2003

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**
e.g., norm regularization

"new" mechanism: **implicit control**
e.g., GD path has low-complexity

Bartlett. "For valid generalization the size of the weights is more important than the size of the network." NeurIPS 1996
Bühlmann, Yu. "Boosting with the L_2 loss: regression and classification." JASA 2003
Neyshabur, Tomioka, Srebro. "In search of the real inductive bias: on the role of implicit regularization in deep learning." ICLR 2015 (workshop)
Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

Implicit regularization

$$\text{test error} \lesssim \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**
e.g., norm regularization

"new" mechanism: **implicit control**
e.g., GD path has low-complexity

Bartlett. "For valid generalization the size of the weights is more important than the size of the network." NeurIPS 1996

Bühlmann, Yu. "Boosting with the L_2 loss: regression and classification." JASA 2003

Neyshabur, Tomioka, Srebro. "In search of the real inductive bias: on the role of implicit regularization in deep learning." ICLR 2015 (workshop)

Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

#citations ≥ 6026 (conf. ver.) + 3633 (journal ver.)

Implicit regularization

$$\text{test error} \leq \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**

e.g., norm regularization

"new" mechanism: **implicit control**

e.g., GD path has low-complexity

how do they compare?

(hint: they're not equivalent)

Bartlett. "For valid generalization the size of the weights is more important than the size of the network." NeurIPS 1996

Bühlmann, Yu. "Boosting with the L_2 loss: regression and classification." JASA 2003

Neyshabur, Tomioka, Srebro. "In search of the real inductive bias: on the role of implicit regularization in deep learning." ICLR 2015 (workshop)

Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

#citations ≥ 6026 (conf. ver.) + 3633 (journal ver.)

Implicit regularization

$$\text{test error} \lesssim \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**
e.g., norm regularization

"new" mechanism: **implicit control**
e.g., GD path has low-complexity

Challenges:

non-convexity
confounders
...

how do they compare?
(hint: they're not equivalent)

Bartlett. "For valid generalization the size of the weights is more important than the size of the network." NeurIPS 1996
Bühlmann, Yu. "Boosting with the L_2 loss: regression and classification." JASA 2003
Neyshabur, Tomioka, Srebro. "In search of the real inductive bias: on the role of implicit regularization in deep learning." ICLR 2015 (workshop)
Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

#citations ≥ 6026 (conf. ver.) + 3633 (journal ver.)

Implicit regularization

$$\text{test error} \lesssim \text{training error} + \sqrt{\frac{\text{complexity}}{n}}$$

Gradient Descent

$$\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$$

classical mechanism: **explicit control**
e.g., norm regularization

"new" mechanism: **implicit control**
e.g., GD path has low-complexity

how do they compare?
(hint: they're not equivalent)

Challenges:

non-convexity

confounders

...

Sandbox:

linear regression

Bartlett. "For valid generalization the size of the weights is more important than the size of the network." NeurIPS 1996

Bühlmann, Yu. "Boosting with the L_2 loss: regression and classification." JASA 2003

Neyshabur, Tomioka, Srebro. "In search of the real inductive bias: on the role of implicit regularization in deep learning." ICLR 2015 (workshop)

Zhang, Bengio, Hardt, Recht, Vinyals. "Understanding deep learning requires rethinking generalization." ICLR 2017

#citations ≥ 6026 (conf. ver.) + 3633 (journal ver.)

Implicit dominates explicit regularization

For **all** linear regression problems (Gaussian-design):

early-stopped GD is

- always no worse
- sometimes much better

than ℓ_2-regularization (ridge regression)

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Implicit dominates explicit regularization

For **all** linear regression problems (Gaussian-design):

early-stopped GD is

- always no worse
- sometimes much better

than ℓ_2-regularization (ridge regression)

"James-Stein dominates OLS" $\Rightarrow$ statistical learning

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Implicit dominates explicit regularization

For **all** linear regression problems (Gaussian-design):

early-stopped GD is

- always no worse
- sometimes much better

than ℓ_2-regularization (ridge regression)

"James-Stein dominates OLS" $\Rightarrow$ statistical learning

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Linear regression

$$x \sim \mathsf{N}(0, \Sigma), \quad y = x^\top \theta^* + \mathsf{N}(0, 1) \;\; \text{for} \;\; \|\theta^*\|_\Sigma \lesssim 1$$

Linear regression

$$x \sim \mathrm{N}(0, \Sigma), \quad y = x^\top \theta^* + \mathrm{N}(0, 1) \text{ for } \|\theta^*\|_\Sigma \lesssim 1$$

nontrivial label noise

Linear regression

$$x \sim \mathsf{N}(0, \Sigma), \quad y = x^\top \theta^* + \mathsf{N}(0, 1) \ \text{ for } \ \|\theta^*\|_\Sigma \lesssim 1$$

problem instance given by (Σ, θ^*)

nontrivial label noise

Linear regression

$$x \sim \mathrm{N}(0, \Sigma), \quad y = x^\top \theta^* + \mathrm{N}(0, 1) \ \text{ for } \ \|\theta^*\|_\Sigma \lesssim 1$$

problem instance given by (Σ, θ^*)

nontrivial label noise

(excess) test error / prediction risk

$$\begin{aligned} \mathrm{risk}(\theta) &= \mathbb{E}(y - x^\top \theta)^2 - \mathbb{E}(y - x^\top \theta^*)^2 \\ &= \|\theta - \theta^*\|_\Sigma^2 \end{aligned}$$

Linear regression

$x \sim \mathrm{N}(0, \Sigma), \quad y = x^\top \theta^* + \mathrm{N}(0, 1)$ for $\|\theta^*\|_\Sigma \lesssim 1$

problem instance given by (Σ, θ^*)

nontrivial label noise

(excess) test error / prediction risk

$$\begin{aligned} \mathrm{risk}(\theta) &= \mathbb{E}(y - x^\top \theta)^2 - \mathbb{E}(y - x^\top \theta^*)^2 \\ &= \|\theta - \theta^*\|_\Sigma^2 \end{aligned}$$

n iid samples $(x_1, y_1), \ldots, (x_n, y_n)$

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Explicit / implicit regularization

empirical risk $\quad L(\theta) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i^\top\theta - y_i)^2$

Explicit / implicit regularization

empirical risk $L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i^\top\theta - y_i)^2$

ridge regression

$$\hat{\theta}_\lambda^{\text{ridge}} = \arg\min L(\theta) + \lambda\|\theta\|^2$$
$$= (X^\top X + n\lambda I)^{-1}X^\top Y$$

Explicit / implicit regularization

empirical risk $L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i^\top\theta - y_i)^2$

hyperparameter: $\lambda \geq 0$

ridge regression

$$\hat{\theta}_\lambda^{\text{ridge}} = \arg\min L(\theta) + \lambda||\theta||^2$$
$$= (X^\top X + n\lambda I)^{-1}X^\top Y$$

Explicit / implicit regularization

empirical risk $L(\theta) = \dfrac{1}{n} \sum_{i=1}^{n} (x_i^\top \theta - y_i)^2$

ridge regression

hyperparameter: $\lambda \geq 0$

$$\hat{\theta}_\lambda^{\text{ridge}} = \arg\min L(\theta) + \lambda ||\theta||^2$$
$$= (X^\top X + n\lambda I)^{-1} X^\top Y$$

gradient descent

- $\theta_0 = 0$
- for $s = 1, \ldots, t$

$$\theta_s = \theta_{s-1} - \eta \nabla L(\theta_s)$$

- $\hat{\theta}_t^{\text{gd}} = \theta_t$

Explicit / implicit regularization

empirical risk $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} (x_i^\top \theta - y_i)^2$

ridge regression

hyperparameter: $\lambda \geq 0$

$$\hat{\theta}_\lambda^{\text{ridge}} = \arg\min L(\theta) + \lambda \|\theta\|^2$$
$$= (X^\top X + n\lambda I)^{-1} X^\top Y$$

gradient descent

hyperparameter: $t \geq 0$

- $\theta_0 = 0$
- for $s = 1, \ldots, t$

$$\theta_s = \theta_{s-1} - \eta \nabla L(\theta_s)$$

- $\hat{\theta}_t^{\text{gd}} = \theta_t$

Explicit / implicit regularization

empirical risk $\quad L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i^\top\theta - y_i)^2$

ridge regression

hyperparameter: $\lambda \geq 0$

$$\hat{\theta}_\lambda^{\text{ridge}} = \arg\min L(\theta) + \lambda\|\theta\|^2$$
$$= (X^\top X + n\lambda I)^{-1} X^\top Y$$

gradient descent

- $\theta_0 = 0$
- for $s = 1,\ldots,t$

hyperparameter: $t \geq 0$

$$\theta_s = \theta_{s-1} - \eta \nabla L(\theta_s)$$

- $\hat{\theta}_t^{\text{gd}} = \theta_t$

fix $0 < \eta \leq 1/\|\nabla^2 L\|$; otherwise, rescale time

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

$$\mathsf{risk}\big(\hat{\theta}_t^{\mathsf{gd}}\big) \lesssim \mathsf{risk}\big(\hat{\theta}_\lambda^{\mathsf{ridge}}\big)$$

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\big(\hat{\theta}_t^{\mathsf{gd}}\big) \lesssim \mathsf{risk}\big(\hat{\theta}_\lambda^{\mathsf{ridge}}\big)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

$$\mathsf{risk}\big(\hat{\theta}_t^{\mathsf{gd}}\big) \lesssim \frac{1}{n^{0.33}} \mathsf{risk}\big(\hat{\theta}_\lambda^{\mathsf{ridge}}\big)$$

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

natural problems
e.g., power-law

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \frac{1}{n^{0.33}} \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

natural problems
e.g., power-law

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \frac{1}{n^{0.33}} \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Prior work. Assume $\mathbb{E}\theta^* = 0$, $\mathbb{E}\theta^{*\otimes 2} \propto I$. For $t \propto 1/\lambda$, we have

$$\mathbb{E}\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \leq 1.69 \mathbb{E}\mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09
Ali, Kolter, Tibshirani. "A continuous-time view of early stopping for least squares regression." AISTATS 2019

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1, \lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

natural problems
e.g., power-law

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \frac{1}{n^{0.33}} \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Prior work. Assume $\mathbb{E}\theta^* = 0$, $\mathbb{E}\theta^{*\otimes 2} \propto I$. For $t \propto 1/\lambda$, we have

$$\inf_\lambda \mathbb{E}\mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right) \leq \mathbb{E}\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \leq 1.69\,\mathbb{E}\mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Wu, Bartlett, Lee, Kakade, Yu. “Risk comparisons in linear regression: implicit regularization dominates explicit regularization.” arXiv 2025.09
Ali, Kolter, Tibshirani. “A continuous-time view of early stopping for least squares regression.” AISTATS 2019

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

natural problems
e.g., power-law

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \frac{1}{n^{0.33}} \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Prior work. Assume $\mathbb{E}\theta^* = 0$, $\mathbb{E}\theta^{*\otimes 2} \propto I$. For $t \propto 1/\lambda$, we have

$$\inf_\lambda \mathbb{E}\mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right) \leq \mathbb{E}\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \leq 1.69\,\mathbb{E}\mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

"GD = ridge" under isotropic prior

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Ali, Kolter, Tibshirani. "A continuous-time view of early stopping for least squares regression." AISTATS 2019

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

natural problems
e.g., power-law

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \frac{1}{n^{0.33}} \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Prior work. Assume $\mathbb{E}\theta^* = 0$, $\mathbb{E}\theta^{*\otimes 2} \propto I$. For $t \propto 1/\lambda$, we have

$$\inf_\lambda \mathbb{E}\mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right) \leq \mathbb{E}\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \leq 1.69\,\mathbb{E}\mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

"GD = ridge" under isotropic prior

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Ali, Kolter, Tibshirani. "A continuous-time view of early stopping for least squares regression." AISTATS 2019

GD dominates ridge

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

natural problems
e.g., power-law

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim \frac{1}{n^{0.33}} \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Prior work. Assume $\mathbb{E}\theta^* = 0$, $\mathbb{E}\theta^{*\otimes 2} \propto I$. For $t \propto 1/\lambda$, we have

Remark 8. The Bayes perspective here is critical; the proof breaks down for prediction risk, at an arbitrary fixed β_0, and it is not clear to us whether the result is true for prediction risk in general.

GD dominates ridge instead of "GD = ridge"

Theorem. For **every** (Σ, θ^*), $n \geq 1$, $\lambda \geq 0$, there exists $t \geq 0$ such that, w.p. ≥ 0.99

e.g., $\eta t = 1/\lambda$

$$\text{risk}\left(\hat{\theta}_t^{\text{gd}}\right) \lesssim \text{risk}\left(\hat{\theta}_\lambda^{\text{ridge}}\right)$$

There exists (Σ, θ^*) and t_n such that for **any** λ, w.p. ≥ 0.99

natural problems
e.g., power-law

$$\text{risk}\left(\hat{\theta}_t^{\text{gd}}\right) \lesssim \frac{1}{n^{0.33}}\text{risk}\left(\hat{\theta}_\lambda^{\text{ridge}}\right)$$

Prior work. Assume $\mathbb{E}\theta^* = 0$, $\mathbb{E}\theta^{*\otimes 2} \propto I$. For $t \propto 1/\lambda$, we have

Remark 8. The Bayes perspective here is critical; the proof breaks down for prediction risk, at an arbitrary fixed β_0, and it is not clear to us whether the result is true for prediction risk in general.

Approach

“dominance”: always no worse, sometimes much better

Approach

“dominance”: always no worse, sometimes much better

(TB’23) There is an analytic function f such that, w.p. ≥ 0.99

$$f(\Sigma, \theta^*, n, \lambda) \approx \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

Tsigler & Bartlett. “Benign overfitting in ridge regression.” JMLR 2023

Approach

"dominance": always no worse, sometimes much better

(TB'23) There is an analytic function f such that, w.p. ≥ 0.99

$$\boxed{\text{risk}\left(\hat{\theta}_t^{\text{gd}}\right) \lesssim f(\Sigma, \theta^*, n, \lambda) \quad \text{for } \eta t = 1/\lambda} \approx \text{risk}\left(\hat{\theta}_\lambda^{\text{ridge}}\right)$$

Tsigler & Bartlett. "Benign overfitting in ridge regression." JMLR 2023

Approach

"dominance": always no worse, sometimes much better

(TB'23) There is an analytic function f such that, w.p. ≥ 0.99

$$\text{risk}\left(\hat{\theta}_t^{\text{gd}}\right) \lesssim f(\Sigma, \theta^*, n, \lambda) \approx \text{risk}\left(\hat{\theta}_\lambda^{\text{ridge}}\right)$$

for $\eta t = 1/\lambda$

Tsigler & Bartlett. "Benign overfitting in ridge regression." JMLR 2023

Approach

“dominance”: always no worse, sometimes much better

(TB’23) There is an analytic function f such that, w.p. ≥ 0.99

$$\mathrm{risk}\left(\hat{\theta}_t^{\mathrm{gd}}\right) \lesssim f(\Sigma, \theta^*, n, \lambda) \approx \mathrm{risk}\left(\hat{\theta}_\lambda^{\mathrm{ridge}}\right)$$

for $\eta t = 1/\lambda$

$$\hat{\theta}_\lambda^{\mathrm{ridge}} = g(X^\top X) X^\top Y \qquad \hat{\theta}_t^{\mathrm{gd}} = h(X^\top X) X^\top Y$$

Tsigler & Bartlett. “Benign overfitting in ridge regression.” JMLR 2023

Approach

"dominance": always no worse, sometimes much better

(TB'23) There is an analytic function f such that, w.p. ≥ 0.99

$$\mathrm{risk}\left(\hat{\theta}_t^{\mathrm{gd}}\right) \lesssim f(\Sigma, \theta^*, n, \lambda) \approx \mathrm{risk}\left(\hat{\theta}_\lambda^{\mathrm{ridge}}\right)$$

for $\eta t = 1/\lambda$

$$\hat{\theta}_\lambda^{\mathrm{ridge}} = g(X^\top X)X^\top Y$$

$$\hat{\theta}_t^{\mathrm{gd}} = h(X^\top X)X^\top Y$$

polynomial filter

$$g : z \mapsto \frac{1}{z + n\lambda}$$

Tsigler & Bartlett. "Benign overfitting in ridge regression." JMLR 2023

Approach

"dominance": always no worse, sometimes much better

(TB'23) There is an analytic function f such that, w.p. ≥ 0.99

$$\mathsf{risk}\left(\hat{\theta}_t^{\mathsf{gd}}\right) \lesssim f(\Sigma, \theta^*, n, \lambda) \approx \mathsf{risk}\left(\hat{\theta}_\lambda^{\mathsf{ridge}}\right)$$

for $\eta t = 1/\lambda$

$$\hat{\theta}_\lambda^{\mathsf{ridge}} = g(X^\top X)X^\top Y$$

polynomial filter

$$g : z \mapsto \frac{1}{z + n\lambda}$$

$$\hat{\theta}_t^{\mathsf{gd}} = h(X^\top X)X^\top Y$$

exponential filter

$$h : z \mapsto \frac{1 - (1 - z\eta/n)^t}{z}$$

Tsigler & Bartlett. "Benign overfitting in ridge regression." JMLR 2023

Approach

“dominance”: always no worse, sometimes much better

(TB’23) There is an analytic function f such that, w.p. ≥ 0.99

$$\text{risk}\left(\hat{\theta}_t^{\text{gd}}\right) \lesssim f(\Sigma, \theta^*, n, \lambda) \eqsim \text{risk}\left(\hat{\theta}_\lambda^{\text{ridge}}\right)$$

for $\eta t = 1/\lambda$

$$\hat{\theta}_\lambda^{\text{ridge}} = g(X^\top X)X^\top Y \qquad \hat{\theta}_t^{\text{gd}} = h(X^\top X)X^\top Y$$

polynomial filter

$$g : z \mapsto \frac{1}{z + n\lambda}$$

more effective when θ^* decays fast

exponential filter

$$h : z \mapsto \frac{1 - (1 - z\eta/n)^t}{z}$$

Tsigler & Bartlett. “Benign overfitting in ridge regression.” JMLR 2023

GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD is inadmissible

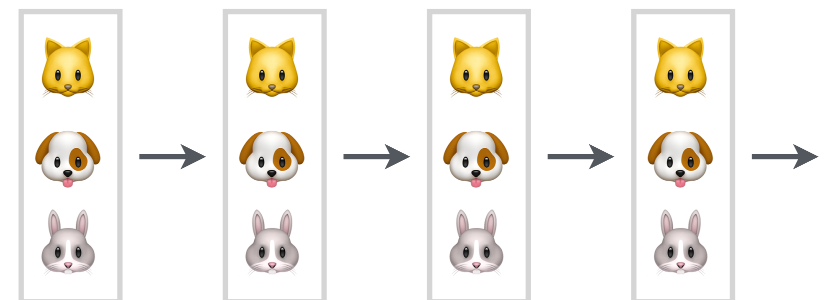
Theorem. In linear regression, GD dominates ridge;
multi-pass SGD (with replacement) = GD

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD



Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD

online SGD $\lessgtr$ GD

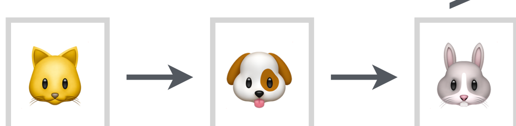
Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD

online SGD $\lessgtr$ GD



Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

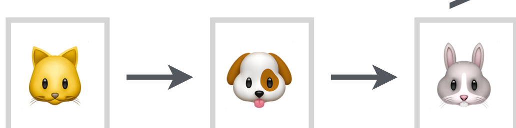
GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD

high-dim; related to benign overfitting

online SGD $\lesseqgtr$ GD



Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

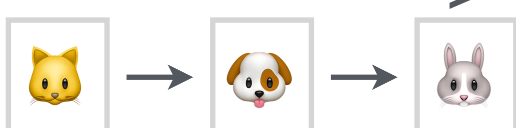
GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD

high-dim; related to benign overfitting

online SGD $\lesssim$ GD



multi-epoch SGD (without replacement) dominates GD

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

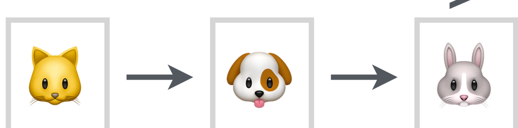
GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

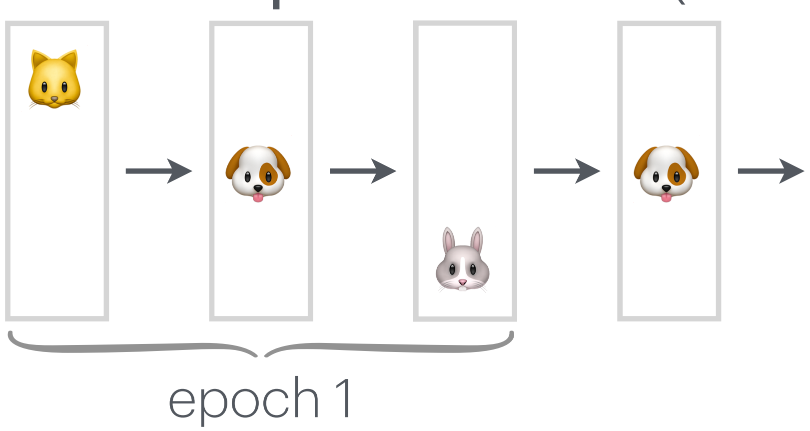
multi-pass SGD (with replacement) = GD

high-dim; related to benign overfitting

online SGD $\lessgtr$ GD



multi-epoch SGD (without replacement) **dominates GD**



Wu, Bartlett, Lee, Kakade, Yu. “Risk comparisons in linear regression: implicit regularization dominates explicit regularization.” arXiv 2025.09

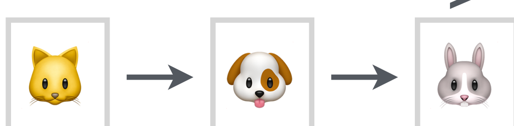
GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD

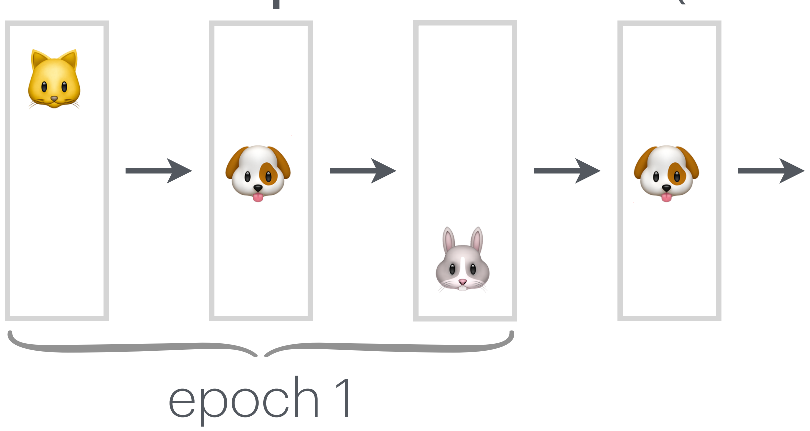
high-dim; related to benign overfitting

online SGD $\lessgtr$ GD



$\geq \max\{\text{online SGD, GD}\}$

multi-epoch SGD (without replacement) **dominates GD**



Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

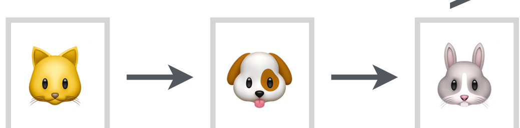
GD is inadmissible

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD

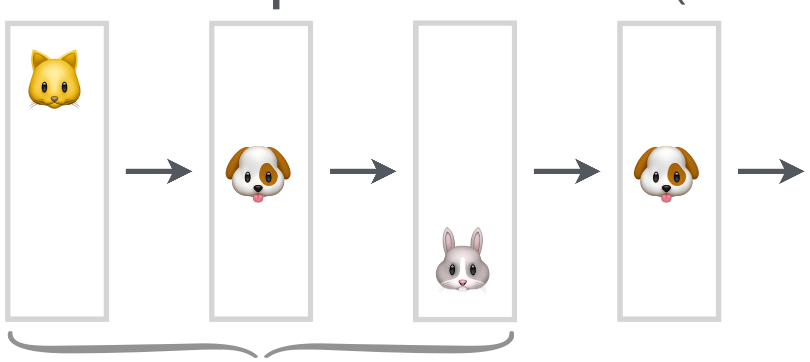
high-dim; related to benign overfitting

online SGD $\lessgtr$ GD



$\geq$ max{online SGD, GD}

multi-epoch SGD (without replacement) **dominates GD**



the SGD variant used in deep learning

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Contribution 3: from theory to practice

principled parallelization method for training language models

- "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling"
  Alexandru Meterez*, Depen Morwani*, W, Costin-Andrei Oncescu, Cengiz Pehlevan, Sham Kakade
  ICLR 2026

Language Model (LM) training

Practice. LM training is “online”: #data ∝ #flops

Language Model (LM) training

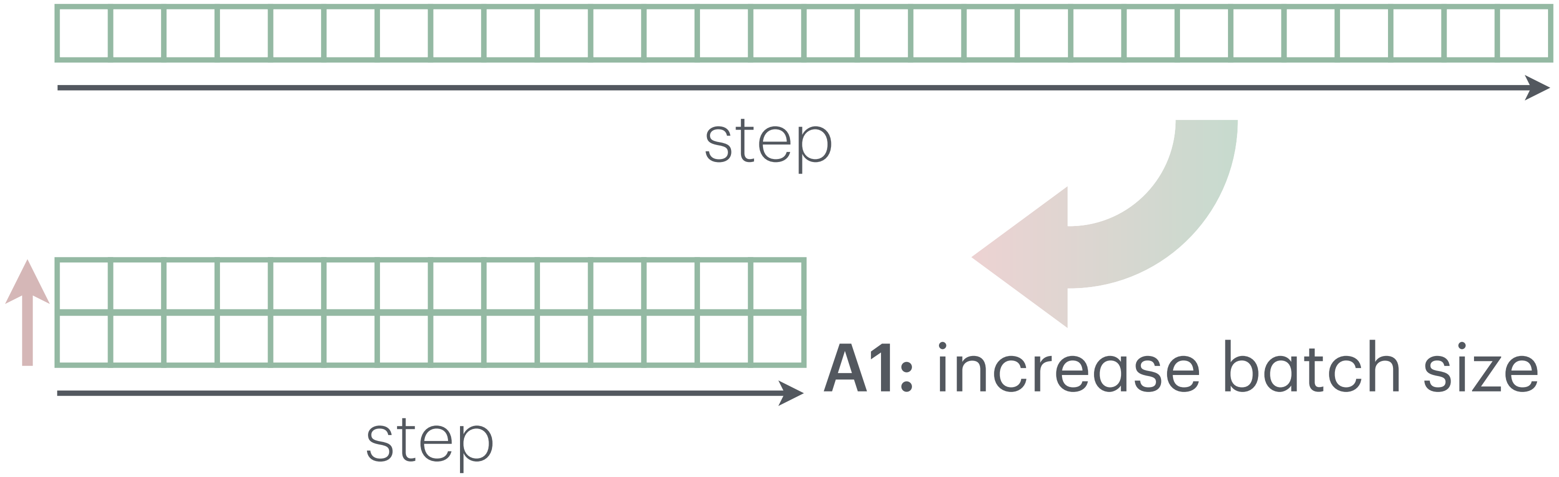
Practice. LM training is “online”: #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?

Language Model (LM) training

Practice. LM training is "online": #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?

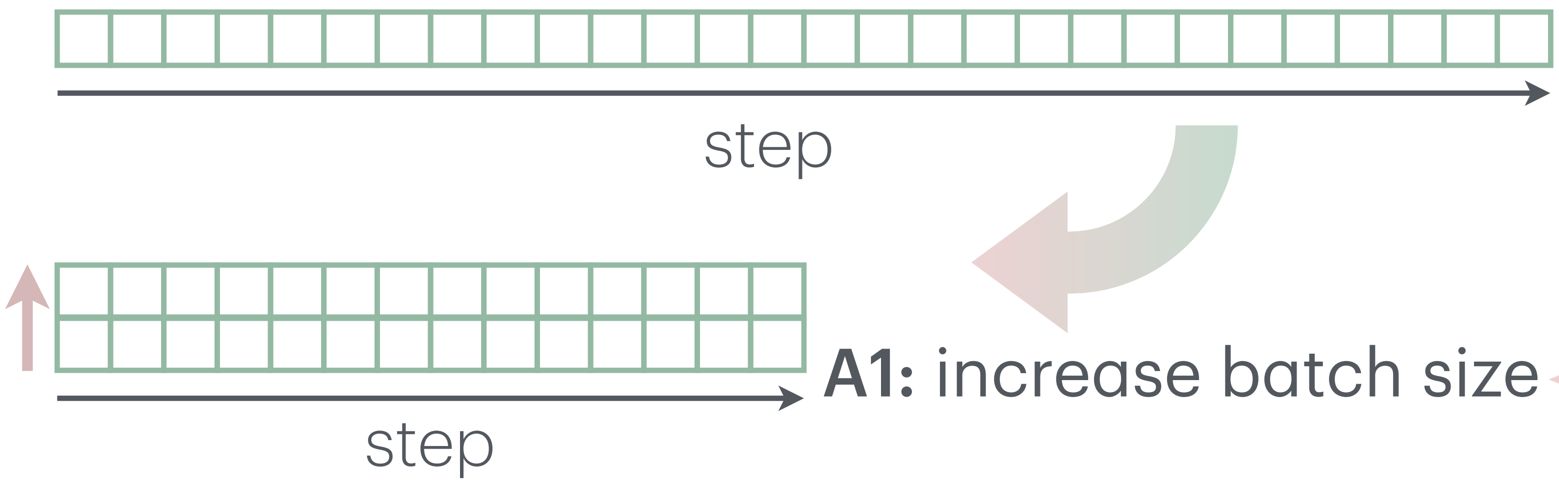


A1: increase batch size

Language Model (LM) training

Practice. LM training is “online”: #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?

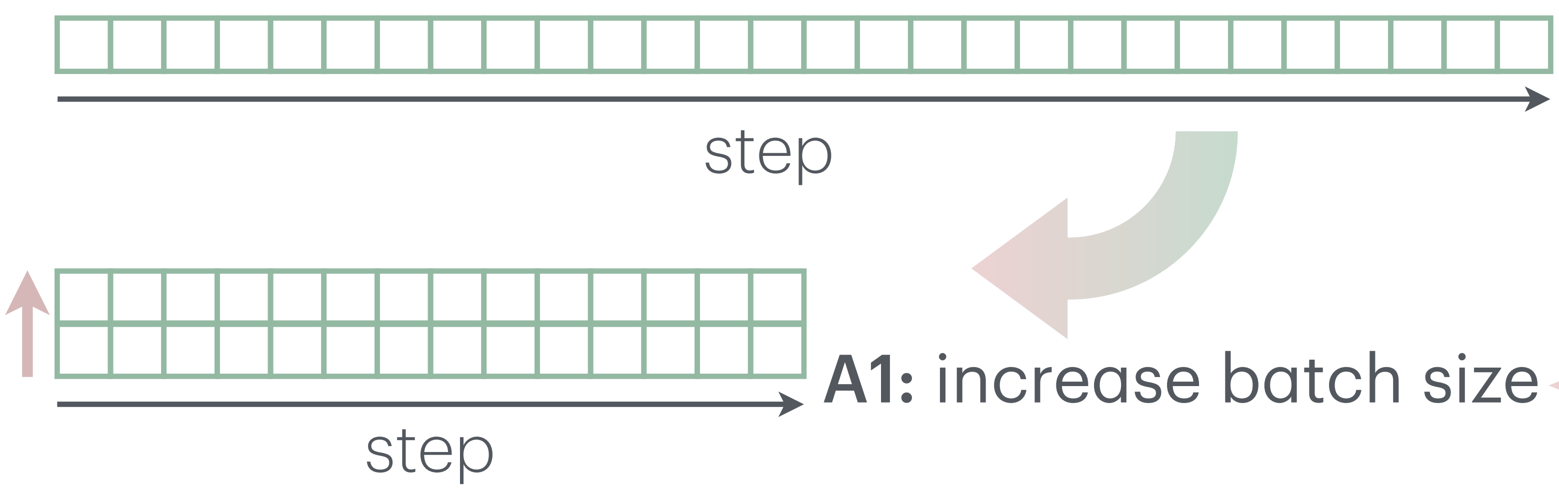


cannot exceed
“critical batch size”

Language Model (LM) training

Practice. LM training is "online": #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?



A1: increase batch size

cannot exceed
"critical batch size"

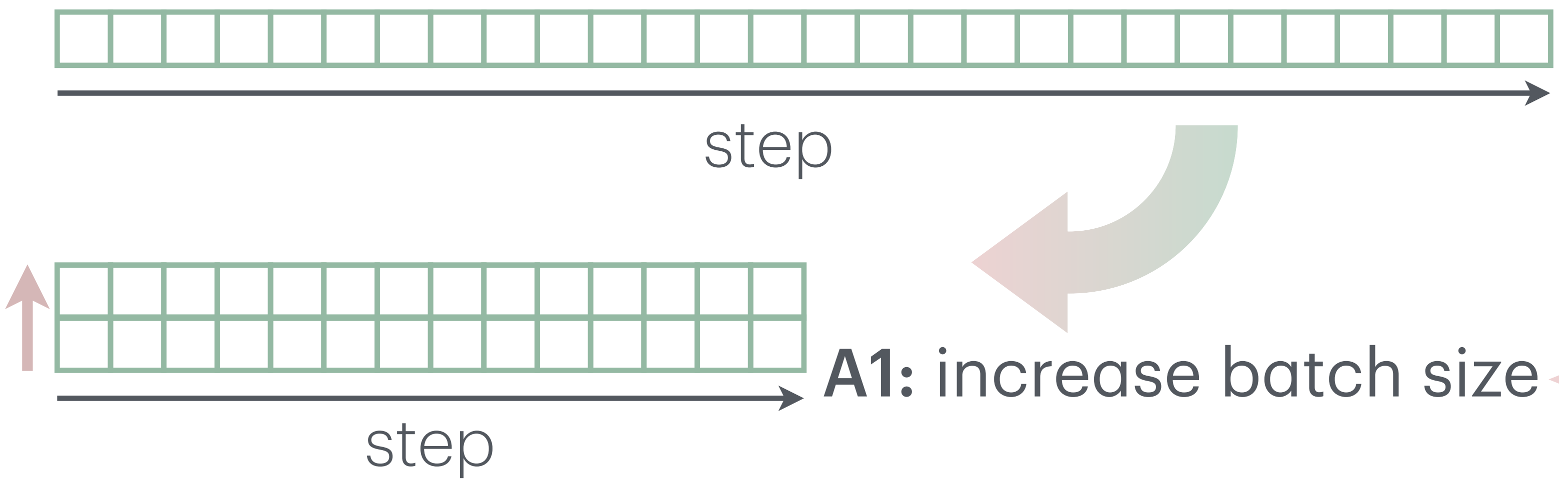
1. known in literature pre-LM

Shallue, Lee, Antognini, Sohl-Dickstein, Frostig, Dahl. "Measuring the effects of data parallelism on neural network training." JMLR 2019

Language Model (LM) training

Practice. LM training is “online”: #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?



A1: increase batch size

cannot exceed
“critical batch size”

1. known in literature pre-LM
2. provable in linear regression

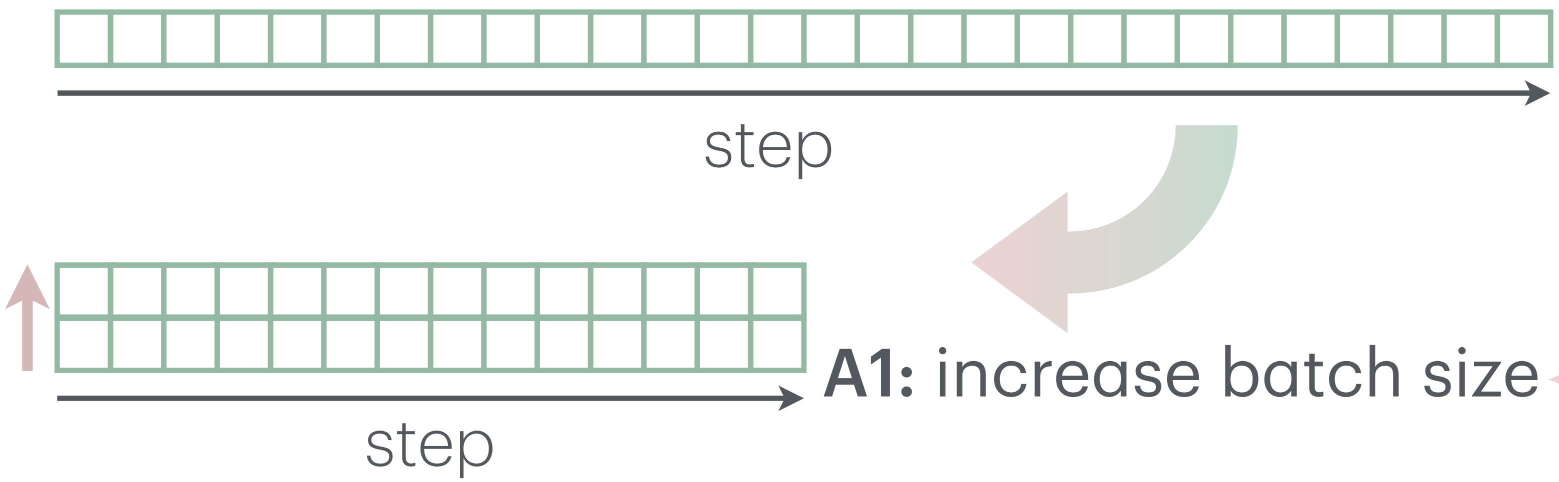
Shallue, Lee, Antognini, Sohl-Dickstein, Frostig, Dahl. “Measuring the effects of data parallelism on neural network training.” JMLR 2019

Wu, Zou, Braverman, Gu, Kakade. “Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression.” ICML 2022

Language Model (LM) training

Practice. LM training is "online": #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?



cannot exceed
"critical batch size"

1. known in literature pre-LM
2. provable in linear regression
3. quantified by us in LM

Shallue, Lee, Antognini, Sohl-Dickstein, Frostig, Dahl. "Measuring the effects of data parallelism on neural network training." JMLR 2019

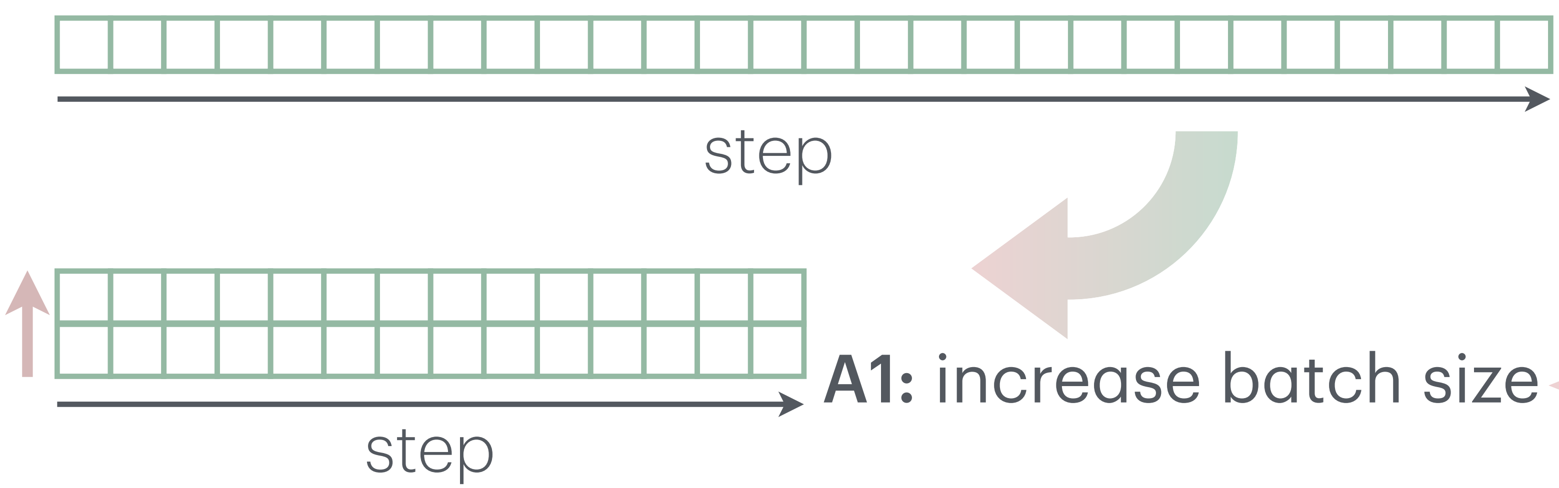
Wu, Zou, Braverman, Gu, Kakade. "Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression." ICML 2022

Zhang, Morwani, Vyas, **Wu**, Zou, Ghai, Foster, Kakade. "How does critical batch size scale in pre-training?" ICLR 2025

Language Model (LM) training

Practice. LM training is “online”: #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?



cannot exceed
“critical batch size”

1. known in literature pre-LM
2. provable in linear regression
3. quantified by us in LM

more parallelization?

Shallue, Lee, Antognini, Sohl-Dickstein, Frostig, Dahl. “Measuring the effects of data parallelism on neural network training.” JMLR 2019

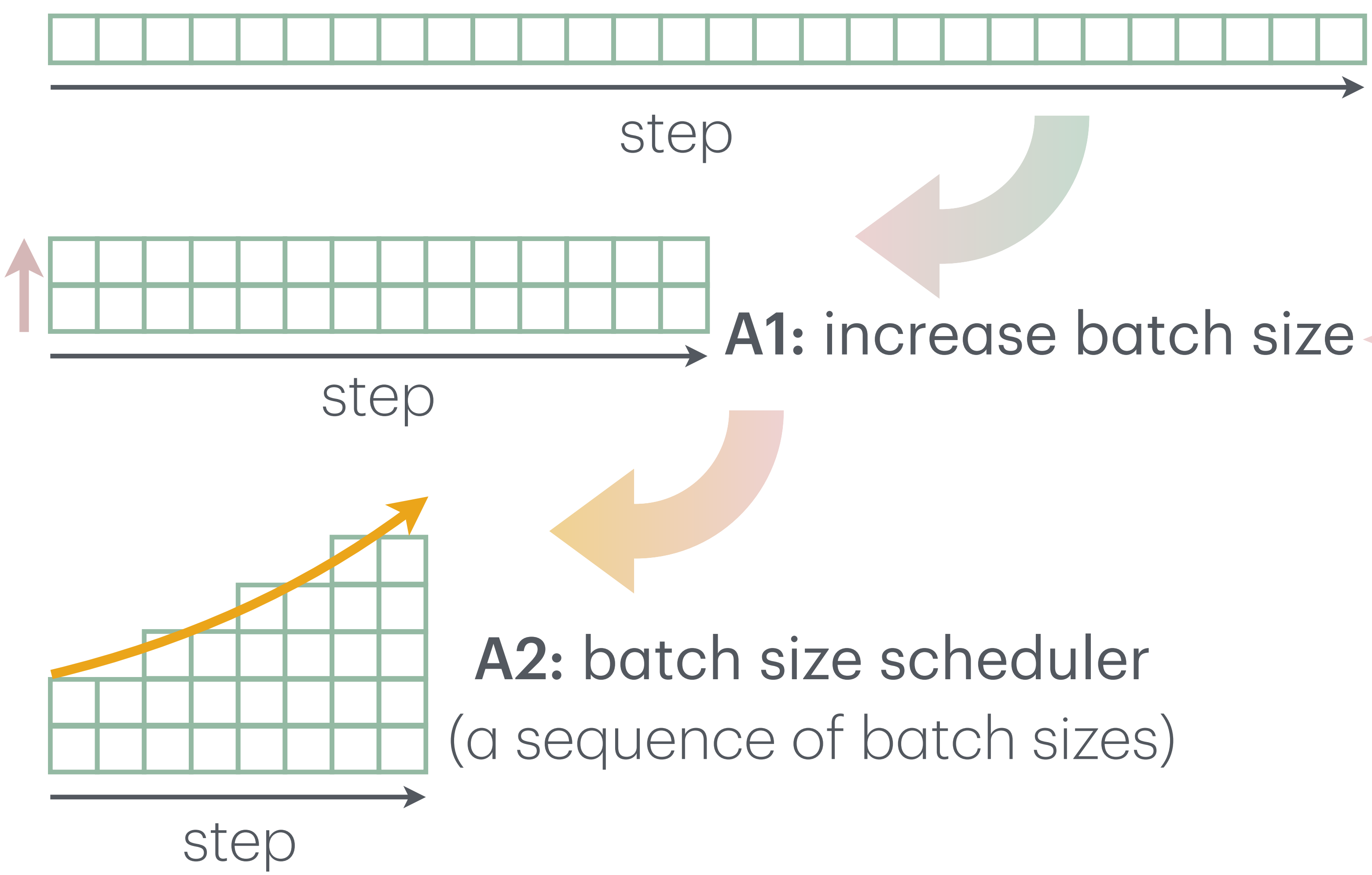
Wu, Zou, Braverman, Gu, Kakade. “Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression.” ICML 2022

Zhang, Morwani, Vyas, **Wu**, Zou, Ghai, Foster, Kakade. “How does critical batch size scale in pre-training?” ICLR 2025

Language Model (LM) training

Practice. LM training is "online": #data ∝ #flops

Question. Fixing #flops, same test error with fewer steps?



cannot exceed
"critical batch size"

1. known in literature pre-LM
2. provable in linear regression
3. quantified by us in LM

more parallelization?

Shallue, Lee, Antognini, Sohl-Dickstein, Frostig, Dahl. "Measuring the effects of data parallelism on neural network training." JMLR 2019

Wu, Zou, Braverman, Gu, Kakade. "Last iterate risk bounds of SGD with decaying stepsize for overparameterized linear regression." ICML 2022

Zhang, Morwani, Vyas, **Wu**, Zou, Ghai, Foster, Kakade. "How does critical batch size scale in pre-training?" ICLR 2025

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

works with
language models

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

compatible with Adam

works with language models

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

compatible with Adam

works with language models

sandbox: linear regression

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

sandbox:
normalized SGD

compatible
with Adam

works with
language models

sandbox:
linear regression

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

sandbox: normalized SGD

compatible with Adam

works with language models

sandbox: linear regression

Theorem (informal). For normalized SGD, "default" and "*Seesaw*" achieve same test error rate for all linear regression problems

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

sandbox: normalized SGD

compatible with Adam

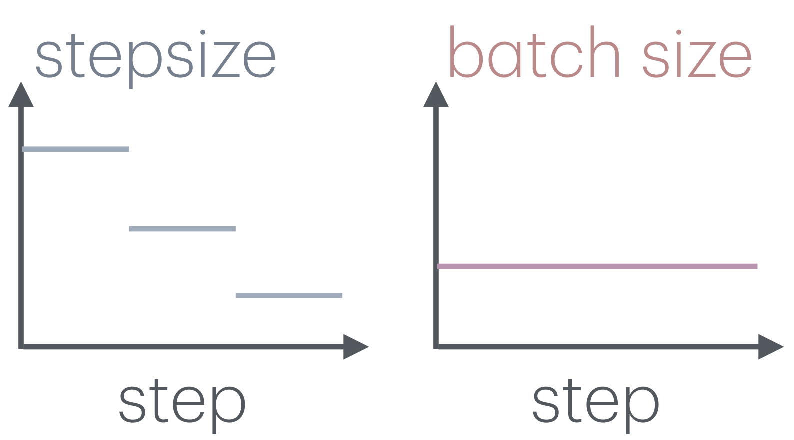
works with language models

sandbox: linear regression

Theorem (informal). For normalized SGD, "default" and "*Seesaw*" achieve same test error rate for all linear regression problems

default: stepsize scheduler

$\eta \to \eta/1.1$, B fixed



Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Seesaw — a principled batch size scheduler

batch size scheduler — same test error with fewer steps?

sandbox: normalized SGD

compatible with Adam

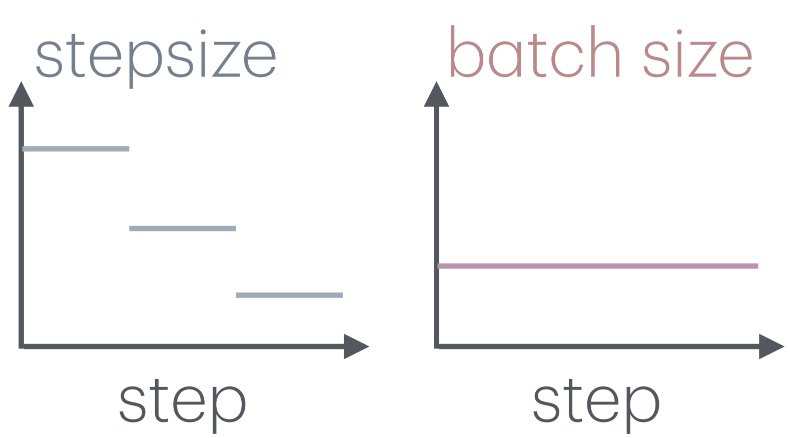
works with language models

sandbox: linear regression

Theorem (informal)**.** For normalized SGD, “default” and “*Seesaw*” achieve same test error rate for all linear regression problems

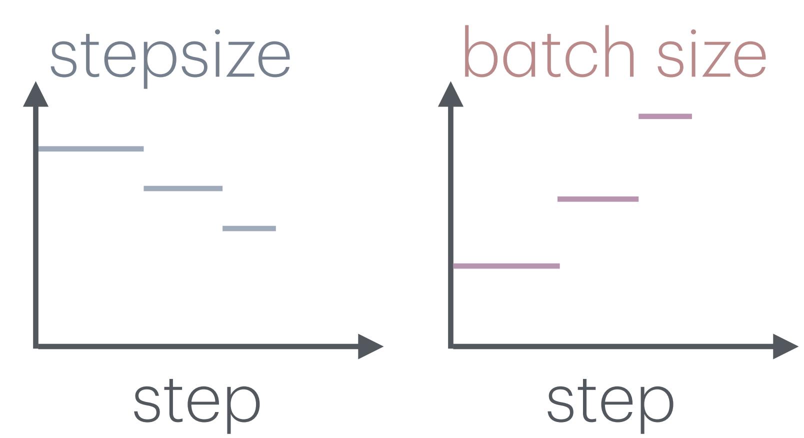
default: stepsize scheduler

$\eta \to \eta/1.1$, B fixed



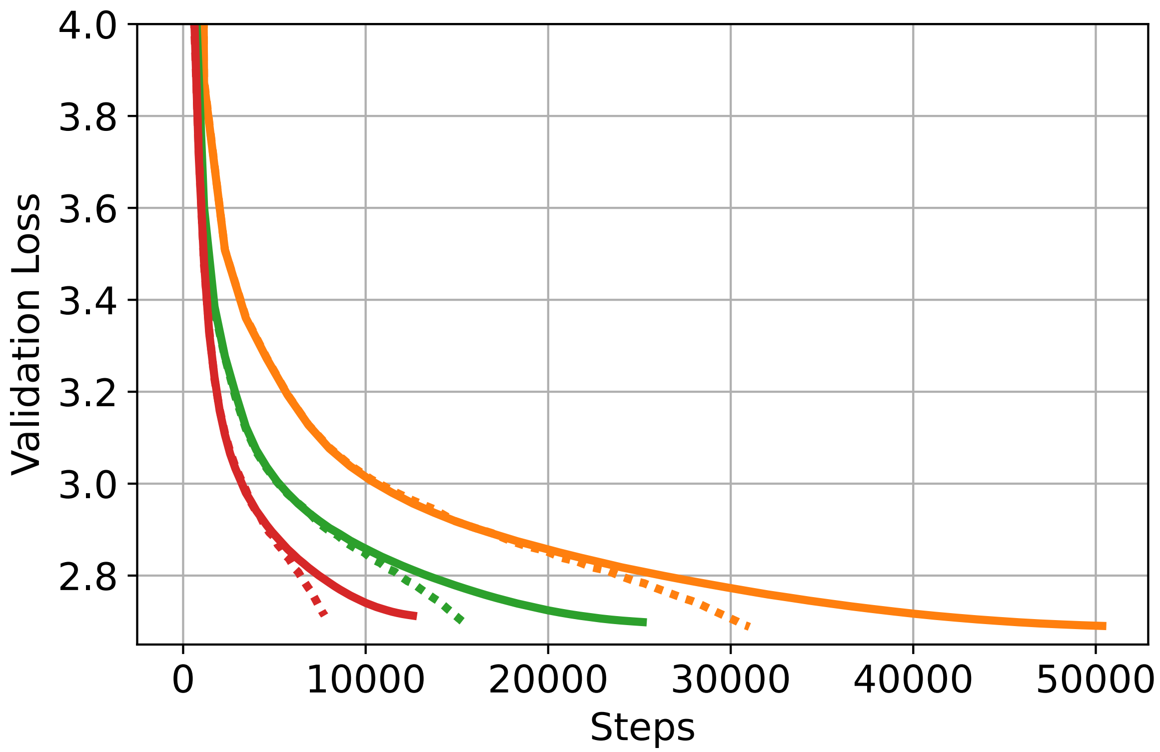
***Seesaw*:** joint scheduler

$\eta \to \eta/\sqrt{1.1}$, $B \to 1.1B$



Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. “*Seesaw*: accelerating training by balancing learning rate and batch size scheduling.” ICLR 2026

Same error (±0.17%), 36% fewer steps

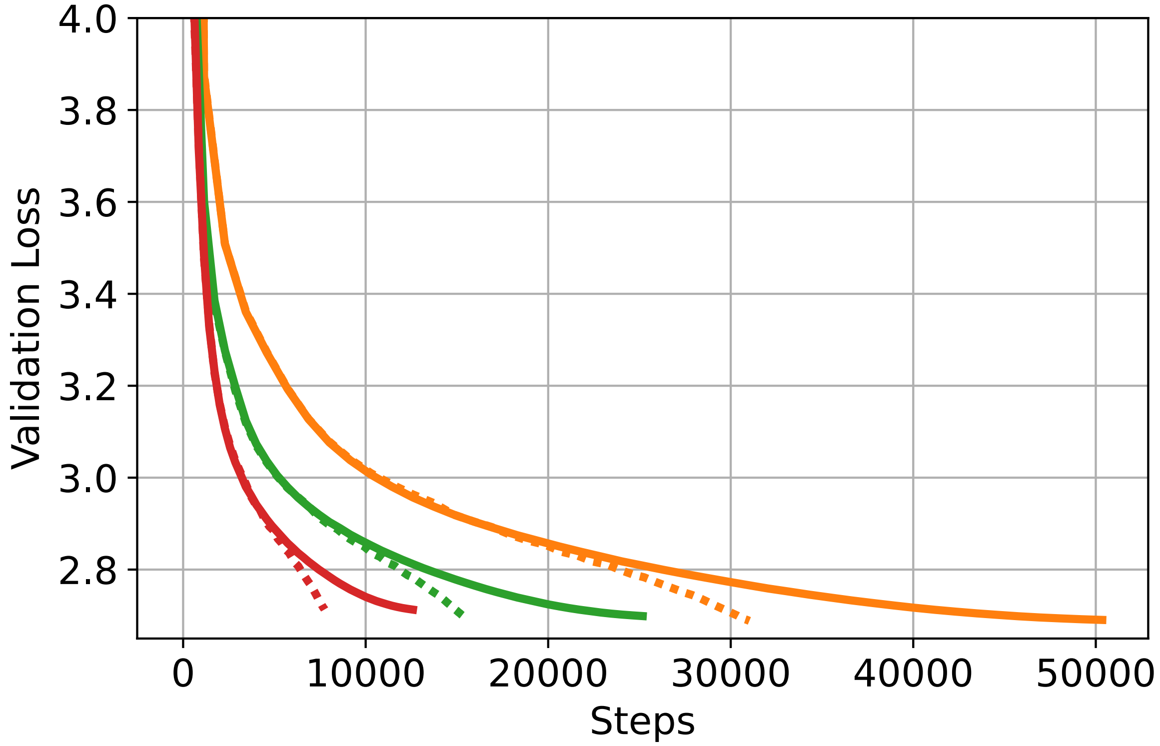


transformer (600M), Adam, C4

initial batch size: 2^8, 2^9, 2^{10} (= CBS)
solid curve: **default** (fixed batch size, cosine stepsize scheduler)
dotted curve: *Seesaw* (ours)

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Same error (±0.17%), 36% fewer steps

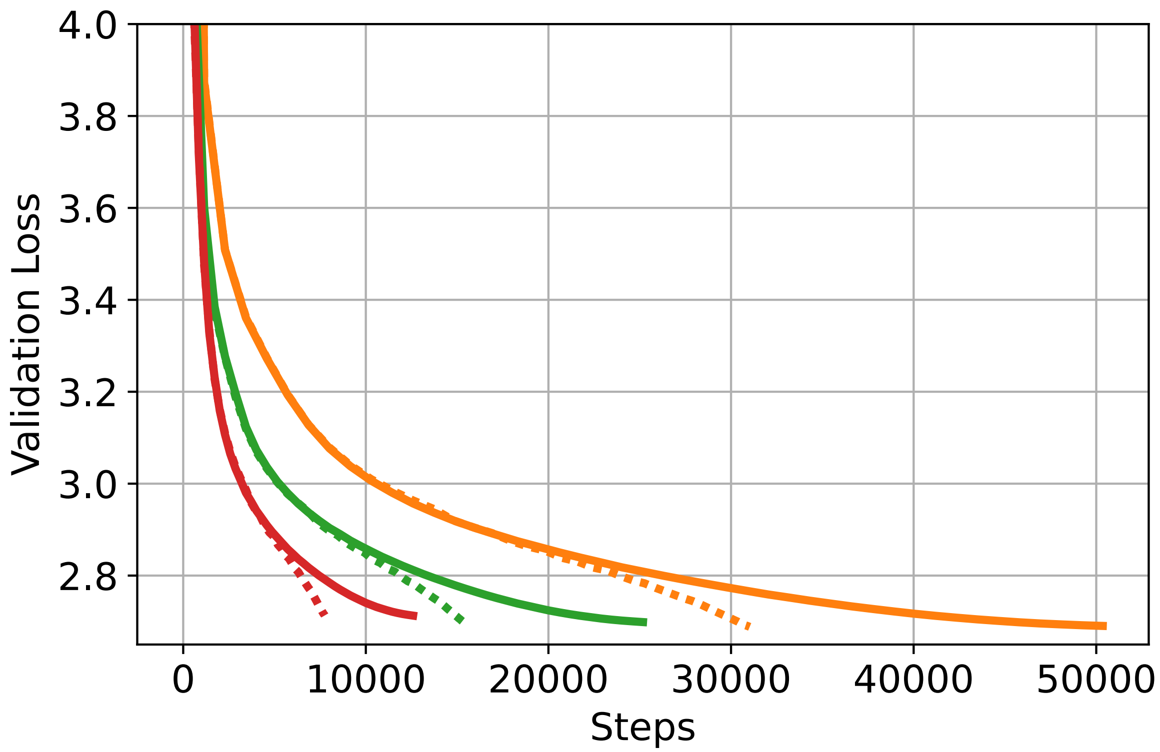


transformer (600M), Adam, C4

initial batch size: 2^8, 2^9, 2^{10} (= CBS)
solid curve: **default** (fixed batch size, cosine stepsize scheduler)
dotted curve: *Seesaw* (ours)

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Same error (±0.17%), 36% fewer steps



transformer (600M), Adam, C4

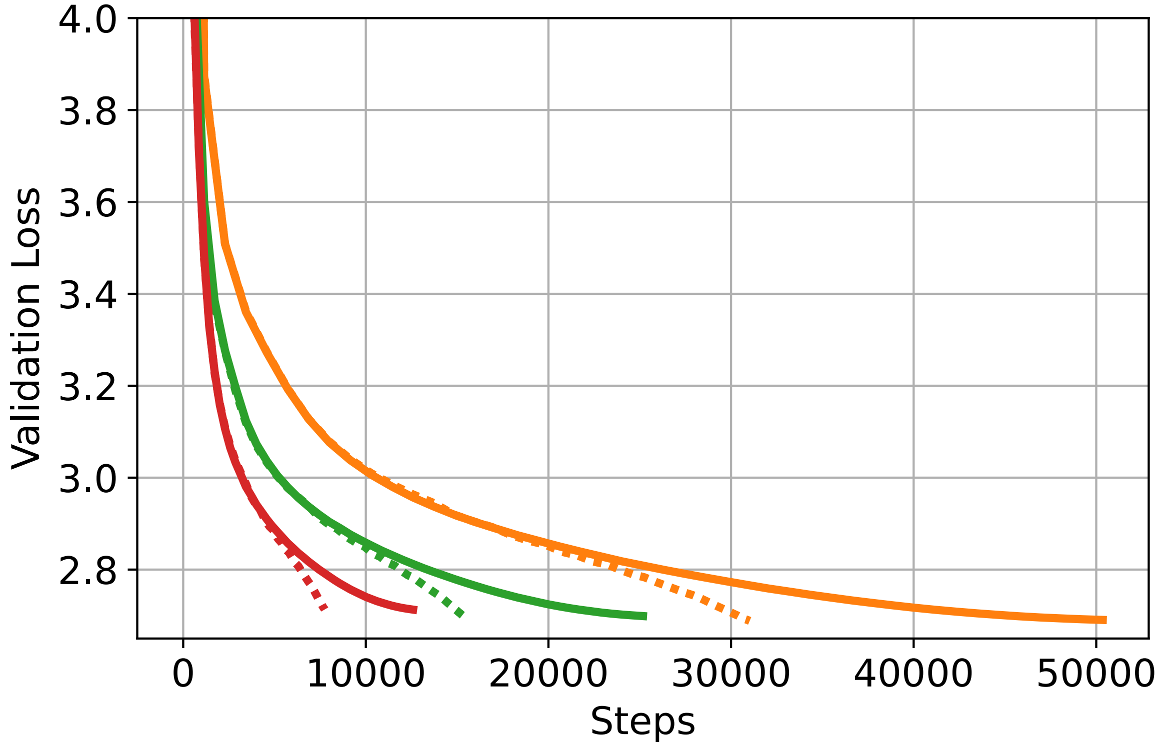
Seesaw

- theory based, practice verified
- blackbox — no extra measures
- #GPU ⬆️ — no free lunch

initial batch size: 2^8, 2^9, 2^{10} (= CBS)
solid curve: **default** (fixed batch size, cosine stepsize scheduler)
dotted curve: *Seesaw* (ours)

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Same error (±0.17%), 36% fewer steps



transformer (600M), Adam, C4

Seesaw

- theory based, practice verified
- blackbox — no extra measures
- #GPU ⬆️ — no free lunch

simple, meaningful sandbox can be predictive!

initial batch size: 2^8, 2^9, 2^{10} (= CBS)
solid curve: default (fixed batch size, cosine stepsize scheduler)
dotted curve: *Seesaw* (ours)

Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026

Summary

Contribution 1: unstable optimization

large stepsize accelerates gradient descent in logistic regression

Contribution 2: implicit regularization

gradient descent dominates ridge regression in linear regression

Contribution 3: from theory to practice

principled parallelization method for training language models

Summary

💡 💡 💡

not covered by classical theory

deep learning = scaling + new mechanisms

ML systems (beyond DL)

not specific to deep learning

✨✨✨

Contribution 1: unstable optimization

large stepsize accelerates gradient descent in logistic regression

Contribution 2: implicit regularization

gradient descent dominates ridge regression in linear regression

Contribution 3: from theory to practice

principled parallelization method for training language models

Summary

classical theory: conservative

“worst-case”, “stable”, ...

optimization | statistics

Summary

my research: less conservative

"instance-wise", "unstable", ...

optimization x statistics

classical theory: conservative

"worst-case", "stable", ...

optimization | statistics

Summary

my aim: theory that is explanatory, predictive, inspirational

my research: less conservative

“instance-wise”, “unstable”, ...

optimization x statistics

classical theory: conservative

“worst-case”, “stable”, ...

optimization | statistics

Going forward

my aim: theory that is explanatory, predictive, inspirational

Going forward

my aim: theory that is explanatory, predictive, inspirational

1. embrace *instability* in optimization theory

Going forward

my aim: theory that is explanatory, predictive, inspirational

1. embrace *instability* in optimization theory
2. explore *synergies* between statistics & optimization

Going forward

my aim: theory that is explanatory, predictive, inspirational

1. embrace *instability* in optimization theory
2. explore *synergies* between statistics & optimization
3. connect theory with *practice*

Going forward

my aim: theory that is explanatory, predictive, inspirational

1. embrace *instability* in optimization theory
2. explore *synergies* between statistics & optimization
3. connect theory with *practice*

instability

new technique?
model, data?
other instabilities?

*check out our tutorial for more discussions

Going forward

my aim: theory that is explanatory, predictive, inspirational

1. embrace *instability* in optimization theory
2. explore *synergies* between statistics & optimization
3. connect theory with *practice*

instability

new technique?
model, data?
other instabilities?

*check out our tutorial for more discussions

stats → opt

new criterion
hyperparameter?
data reuse?

opt → stats

opt algorithm as estimator?

Going forward

my aim: theory that is explanatory, predictive, inspirational

1. embrace *instability* in optimization theory
2. explore *synergies* between statistics & optimization
3. connect theory with *practice*

instability

new technique?
model, data?
other instabilities?

*check out our tutorial for more discussions

stats → opt

new criterion
hyperparameter?
data reuse?

opt → stats

opt algorithm as estimator?

practice

testbed
new question?
new sandbox?
other domain?

Going forward

my aim: theory that is explanatory, predictive, inspirational

1. embrace *instability* in optimization theory
2. explore *synergies* between statistics & optimization
3. connect theory with *practice*

my research: less conservative

"instance-wise", "unstable", ...

optimization x statistics

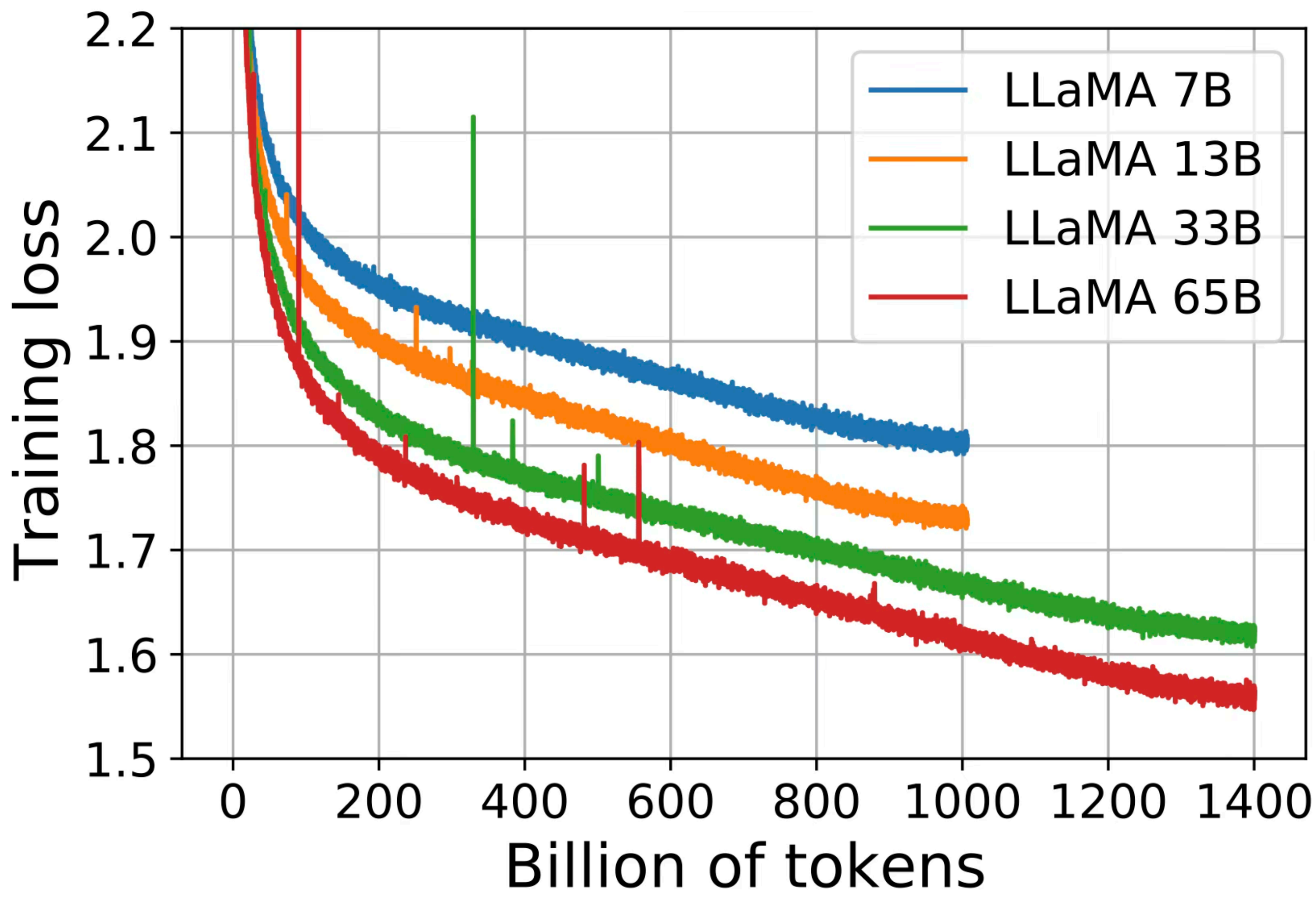
classical theory: conservative

"worst-case", "stable", ...

optimization | statistics

Backup slides

LM training instability



"online" AdamW, batch size = 4M, internet data, transformer

Touvron, Hugo, Izacard, et al. "LLaMA: open and efficient foundation language models." arXiv 2023

Large, adaptive stepsize

Theorem. $\forall \theta_0$, $\exists\ (x_i, y_i)_{i=1}^n$ with margin γ such that: for any first-order batch method

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

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

Large, adaptive stepsize

Theorem. $\forall \theta_0$, $\exists\ (x_i, y_i)_{i=1}^n$ with margin γ such that: for any first-order batch method

$\theta_t \in \theta_0 + \mathrm{span}\{\nabla L(\theta_0), \ldots, \nabla L(\theta_{t-1})\}$
where $L(\theta) = \hat{\mathbb{E}}\ell(yx^\top\theta)$ for any ℓ

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

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

Large, adaptive stepsize

Theorem. $\forall \theta_0$, $\exists\ (x_i, y_i)_{i=1}^n$ with margin γ such that: for any first-order batch method

$\theta_t \in \theta_0 + \mathrm{span}\{\nabla L(\theta_0), \ldots, \nabla L(\theta_{t-1})\}$
where $L(\theta) = \hat{\mathbb{E}} \ell(y x^\top \theta)$ for any ℓ

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

$$\theta_{t+1} = \theta_t - \eta\Big((-\ell^{-1})' \circ L(\theta_t)\Big)\, \nabla L(\theta_t) \approx \theta_t - \frac{\eta}{L(\theta_t)} \nabla L(\theta_t)$$

$$\ell(t) = \ln\big(1 + \exp(-t)\big)$$

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

Large, adaptive stepsize

Theorem. $\forall \theta_0$, $\exists\, (x_i, y_i)_{i=1}^n$ with margin γ such that: for any first-order batch method

$\theta_t \in \theta_0 + \mathrm{span}\{\nabla L(\theta_0), \ldots, \nabla L(\theta_{t-1})\}$
where $L(\theta) = \hat{\mathbb{E}}\ell(yx^\top\theta)$ for any ℓ

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

$$\theta_{t+1} = \theta_t - \eta\Big((-\ell^{-1})' \circ L(\theta_t)\Big)\,\nabla L(\theta_t) \approx \theta_t - \frac{\eta}{L(\theta_t)}\nabla L(\theta_t)$$

$$\ell(t) = \ln\big(1 + \exp(-t)\big)$$

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

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

Therefore, $\lim_{\eta\to\infty} L(\bar{\theta}_t) = 0$ for $t = 1/\gamma^2$

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

Large, adaptive stepsize

Theorem. $\forall \theta_0$, $\exists\ (x_i, y_i)_{i=1}^n$ with margin γ such that: for any first-order batch method

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

$\theta_t \in \theta_0 + \text{span}\{\nabla L(\theta_0), \ldots, \nabla L(\theta_{t-1})\}$
where $L(\theta) = \hat{\mathbb{E}}\ell(yx^\top\theta)$ for any ℓ

$$\theta_{t+1} = \theta_t - \eta\Big((-\ell^{-1})' \circ L(\theta_t)\Big)\,\nabla L(\theta_t) \approx \theta_t - \frac{\eta}{L(\theta_t)}\nabla L(\theta_t)$$

$$\ell(t) = \ln\big(1 + \exp(-t)\big)$$

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

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

Therefore, $\lim_{\eta\to\infty} L(\bar{\theta}_t) = 0$ for $t = 1/\gamma^2$

matching "Perceptron"
(Novikoff'1962, or earlier)

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

Online SGD beats GD

Theorem. Let $n \geq 1$. For a sequence of d-dim problems

$$d \geq n^2 \qquad \theta^* = \begin{bmatrix} n^{0.45} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} n^{-0.9} & & & \\ & 1/d & & \\ & & \ddots & \\ & & & 1/d \end{bmatrix}$$

we have $\|\theta^*\|_\Sigma \leq 1$; moreover, w.p. ≥ 0.99,

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Online SGD beats GD

Theorem. Let $n \geq 1$. For a sequence of d-dim problems

$$d \geq n^2 \qquad \theta^* = \begin{bmatrix} n^{0.45} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} n^{-0.9} & & & \\ & 1/d & & \\ & & \ddots & \\ & & & 1/d \end{bmatrix}$$

we have $\|\theta^*\|_\Sigma \leq 1$; moreover, w.p. ≥ 0.99,

- for all $0 < \eta \lesssim 1$ and $t \geq 0$, $\quad \mathrm{risk}\left(\hat{\theta}_t^{\mathrm{gd}}\right) = \Omega\left(n^{-0.2}\right)$

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Online SGD beats GD

Theorem. Let $n \geq 1$. For a sequence of d-dim problems

$$d \geq n^2 \qquad \theta^* = \begin{bmatrix} n^{0.45} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} n^{-0.9} & & & \\ & 1/d & & \\ & & \ddots & \\ & & & 1/d \end{bmatrix}$$

we have $\|\theta^*\|_\Sigma \leq 1$; moreover, w.p. ≥ 0.99,

- for all $0 < \eta \lesssim 1$ and $t \geq 0$, $\quad \text{risk}\left(\hat{\theta}_t^{\text{gd}}\right) = \Omega\left(n^{-0.2}\right)$
- for $\eta \eqsim 1$, $\quad \text{risk}\left(\hat{\theta}_\eta^{\text{sgd}}\right) = O\left(\ln(n)/n\right)$

Wu, Bartlett, Lee, Kakade, Yu. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025.09

Online SGD beats GD

Theorem. Let $n \geq 1$. For a sequence of d-dim problems

$$d \geq n^2 \qquad \theta^* = \begin{bmatrix} n^{0.45} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} n^{-0.9} & & & \\ & 1/d & & \\ & & \ddots & \\ & & & 1/d \end{bmatrix}$$

we have $\|\theta^*\|_\Sigma \leq 1$; moreover, w.p. ≥ 0.99,

- for all $0 < \eta \lesssim 1$ and $t \geq 0$, $\quad \text{risk}\big(\hat{\theta}_t^{\text{gd}}\big) = \Omega\big(n^{-0.2}\big)$
- for $\eta \eqsim 1$, $\quad \text{risk}\big(\hat{\theta}_\eta^{\text{sgd}}\big) = O\big(\ln(n)/n\big)$

online SGD decays bias exponentially; GD bias ≥ OLS bias

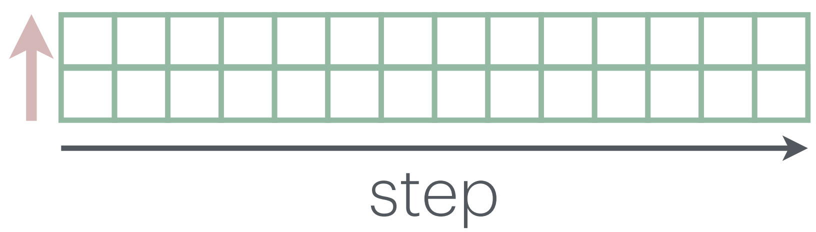
Wu, Bartlett, Lee, Kakade, Yu. “Risk comparisons in linear regression: implicit regularization dominates explicit regularization.” arXiv 2025.09

Comparing with (MLPA’22)

MLPA’22 setup

(constant batch size)

for *multiple* training runs with different constant batch size B, how to set (initial) stepsize η in each run?



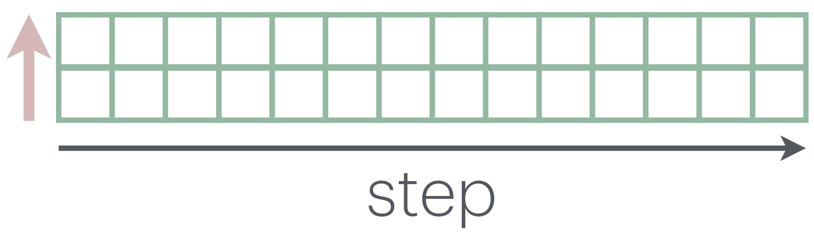
Malladi, Lyu, Panigrahi, Arora. "On the SDEs and scaling rules for adaptive gradient algorithms." NeurIPS 2022

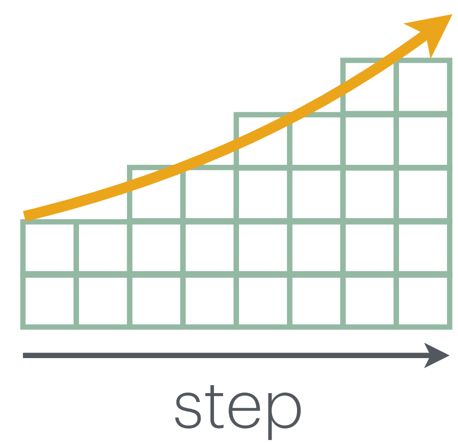
Meterez, Morwani, Wu, Oncescu, Pehlevan, Kakade. “*Seesaw*: accelerating training by balancing learning rate and batch size scheduling.” ICLR 2026

Comparing with (MLPA’22)

MLPA’22 setup
(constant batch size)
for *multiple* training runs with different constant batch size B, how to set (initial) stepsize η in each run?

our setup
(batch size scheduler)
for a *single* training run with given stepsize scheduler $(\eta_t)_{t>0}$, how to design batch size scheduler $(B_t)_{t>0}$?



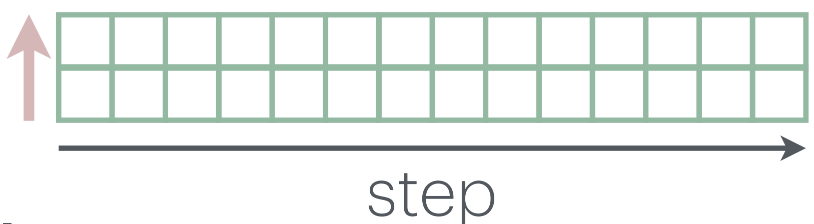


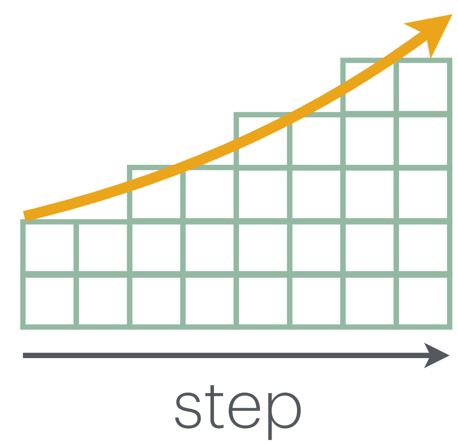
Malladi, Lyu, Panigrahi, Arora. "On the SDEs and scaling rules for adaptive gradient algorithms." NeurIPS 2022
Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. “*Seesaw*: accelerating training by balancing learning rate and batch size scheduling.” ICLR 2026

Comparing with (MLPA'22)

MLPA'22 setup
(constant batch size)
for *multiple* training runs with different constant batch size B, how to set (initial) stepsize η in each run?

our setup
(batch size scheduler)
for a *single* training run with given stepsize scheduler $(\eta_t)_{t>0}$, how to design batch size scheduler $(B_t)_{t>0}$?





Details:

- constant batch size **vs** batch size scheduler
- SDE **vs** linear regression
- infinitesimal stepsize **vs** any stepsize scheduler
- batch **vs** online

Malladi, Lyu, Panigrahi, Arora. "On the SDEs and scaling rules for adaptive gradient algorithms." NeurIPS 2022
Meterez, Morwani, **Wu**, Oncescu, Pehlevan, Kakade. "*Seesaw*: accelerating training by balancing learning rate and batch size scheduling." ICLR 2026