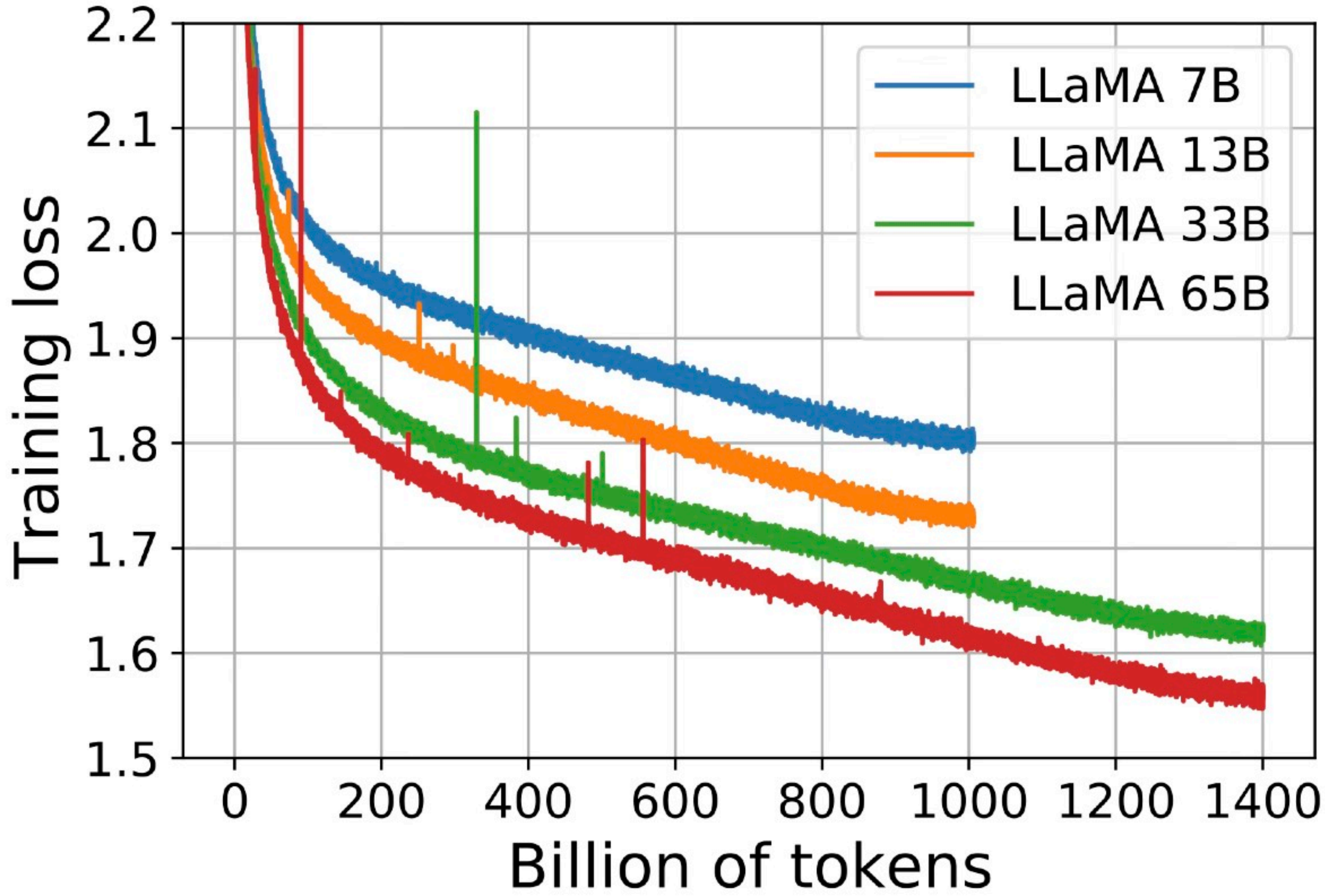
Theoretical Insights on Training Instability in Deep Learning

NeurIPS 2025 Tutorial

Jingfeng Wu — UC Berkeley
Yu-Xiang Wang — UC San Diego
Maryam Fazel — University of Washington

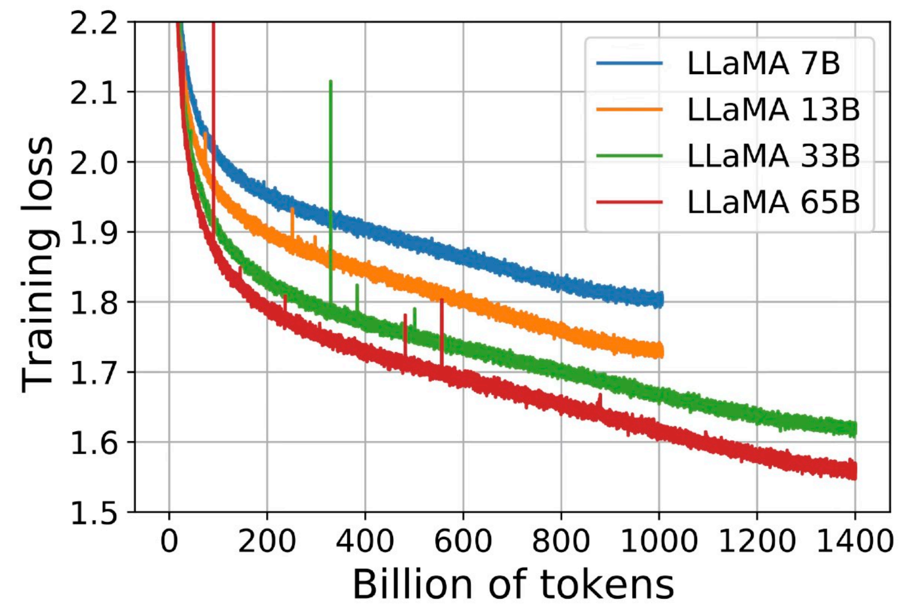
NEURAL INFORMATION PROCESSING SYSTEMS

An LLM pretraining curve



“online” AdamW, batch size = 4M, internet data, transformer

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



r/MachineLearning • 12d ago
Previous-Raisin1434

[R] Why loss spikes?

Research

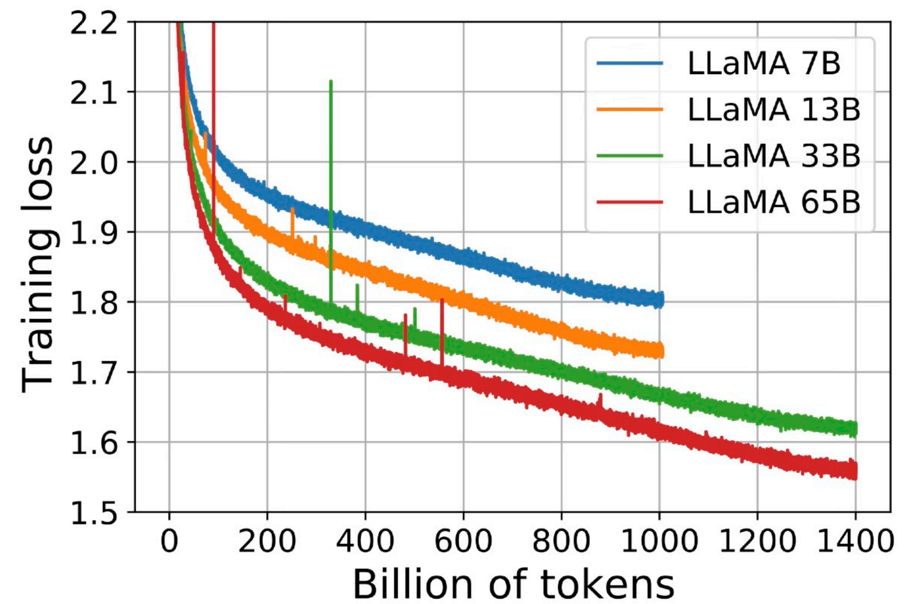
During the training of a neural network, a very common phenomenon is that of loss spikes, which can cause large gradient and destabilize training. Using a learning rate schedule with warmup, or clipping gradients can reduce the loss spikes or reduce their impact on training.

However, I realised that I don't really understand why there are loss spikes in the first place. Is it due to the input data distribution? To what extent can we reduce the amplitude of these spikes? Intuitively, if the model has already seen a representative part of the dataset, it shouldn't be too surprised by anything, hence the gradients shouldn't be that large.

Do you have any insight or references to better understand this phenomenon?

62 | 20 | Share

https://www.reddit.com/r/MachineLearning/comments/1odfuwe/r_why_loss_spikes/



r/MachineLearning • 12d ago
Previous-Raisin1434

[R] Why loss spikes?

Research

During the training of a neural network, a very common phenomenon is that of loss spikes, which can cause large gradient and destabilize training. Using a learning rate schedule with warmup, or clipping gradients can reduce the loss spikes or reduce their impact on training.

However, I realised that I don't really understand why there are loss spikes in the first place. Is it due to the input data distribution? To what extent can we reduce the amplitude of these spikes? Intuitively, if the model has already seen a representative part of the dataset, it shouldn't be too surprised by anything, hence the gradients shouldn't be that large.

Do you have any insight or references to better understand this phenomenon?

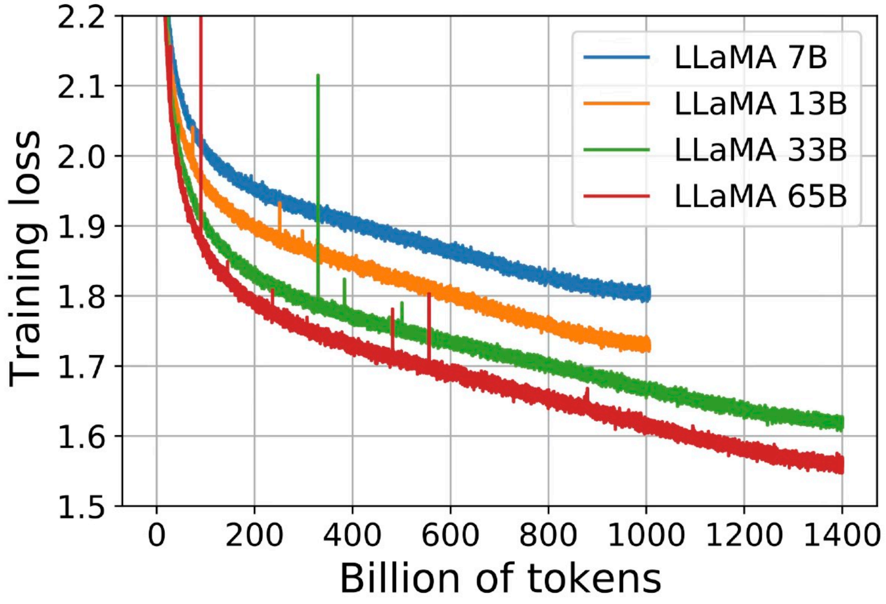
62 20 Share

yes, we do!

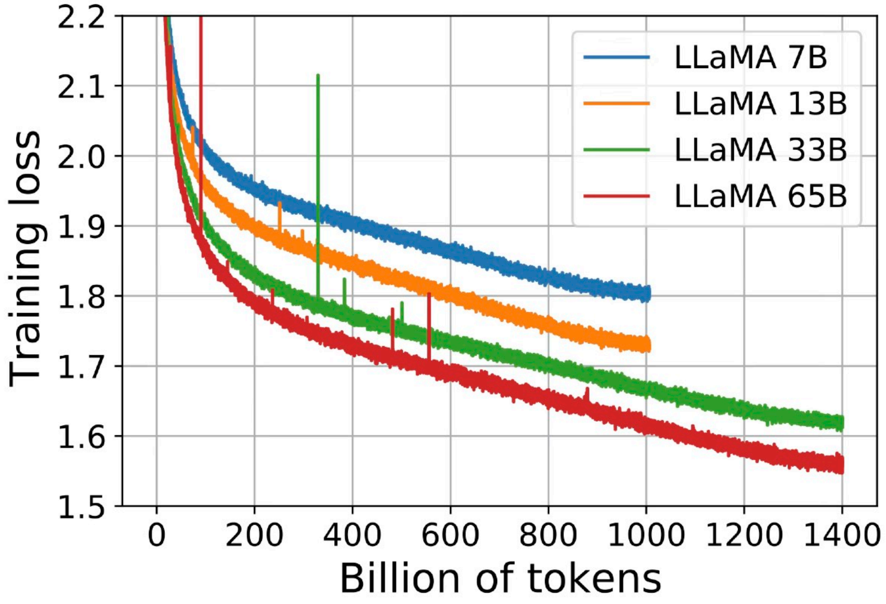
https://www.reddit.com/r/MachineLearning/comments/1odfuwe/r_why_loss_spikes/

Why loss spikes

$$\theta_+ = \theta - \text{stepsize} \times \text{"gradient"}$$



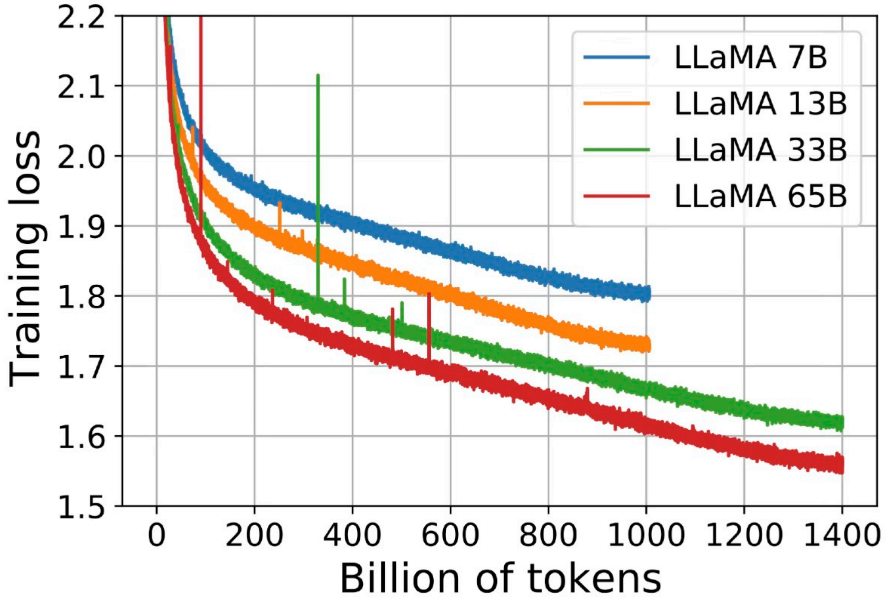
Why loss spikes



$\theta_+ = \theta -$ stepsize $\times$ "gradient"

data randomness $\leftarrow$ unlucky mini-batch

Why loss spikes

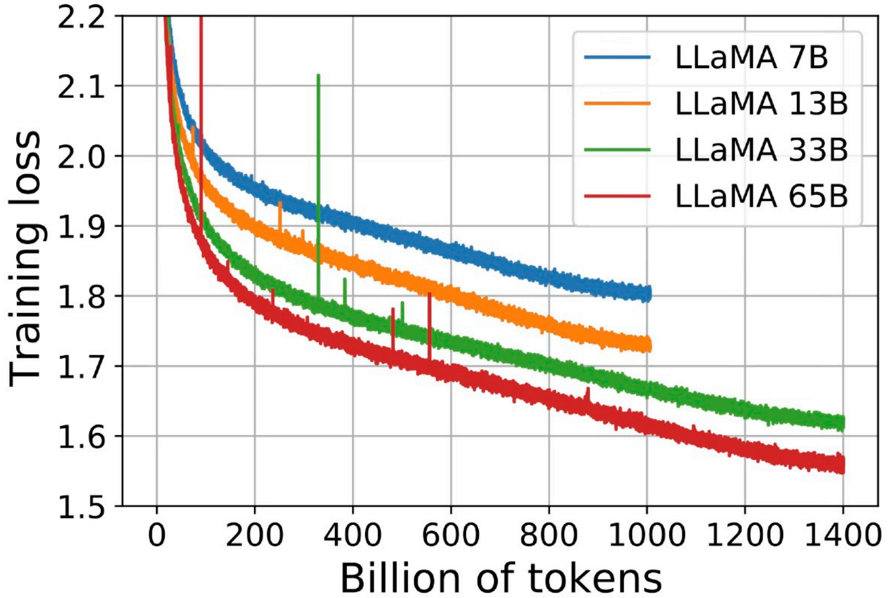


$\theta_+ = \theta -$ stepsize $\times$ "gradient"

data randomness $\leftarrow$ unlucky mini-batch

numerical overflow $\leftarrow$ insufficient precision

Why loss spikes



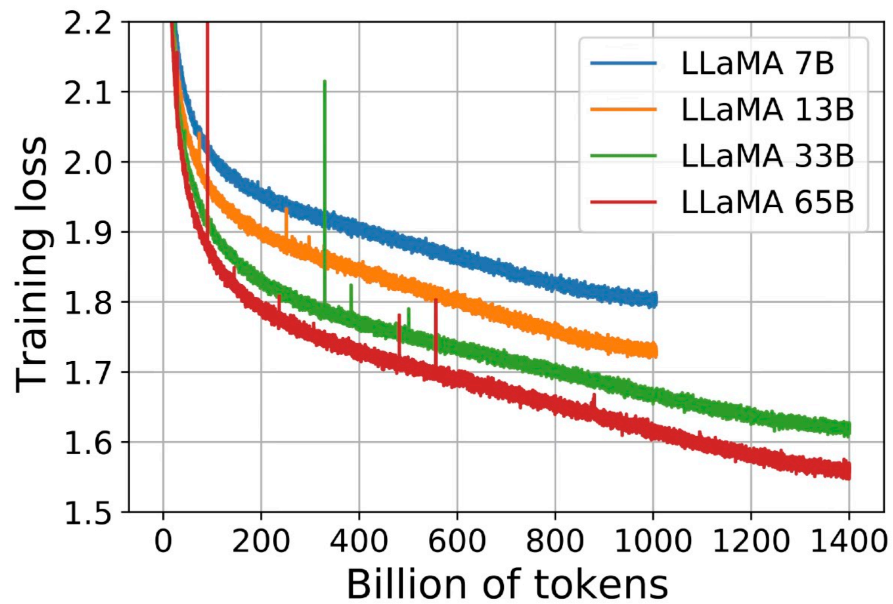
$\theta_+ = \theta -$ stepsize $\times$ "gradient"

data randomness ← unlucky mini-batch

numerical overflow ← insufficient precision

loss landscape ← varying layer-wise curvature

Why loss spikes



$\theta_+ = \theta -$ stepsize $\times$ "gradient"

data randomness ← unlucky mini-batch

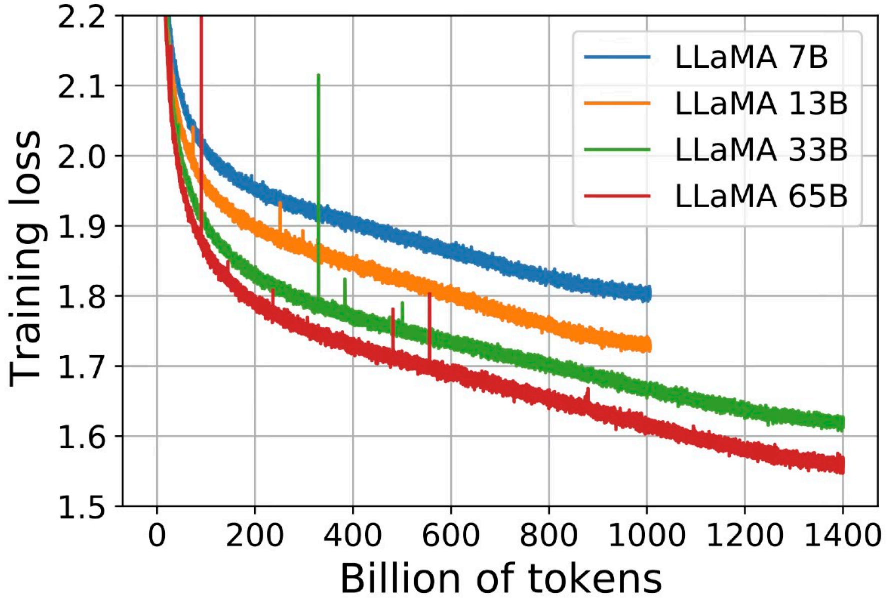
numerical overflow ← insufficient precision

loss landscape ← varying layer-wise curvature

....

inherent instability ← **stepsize / learning rate**

Why loss spikes



$\theta_+ = \theta - \text{stepsize} \times \text{"gradient"}$

data randomness ← unlucky mini-batch

numerical overflow ← insufficient precision

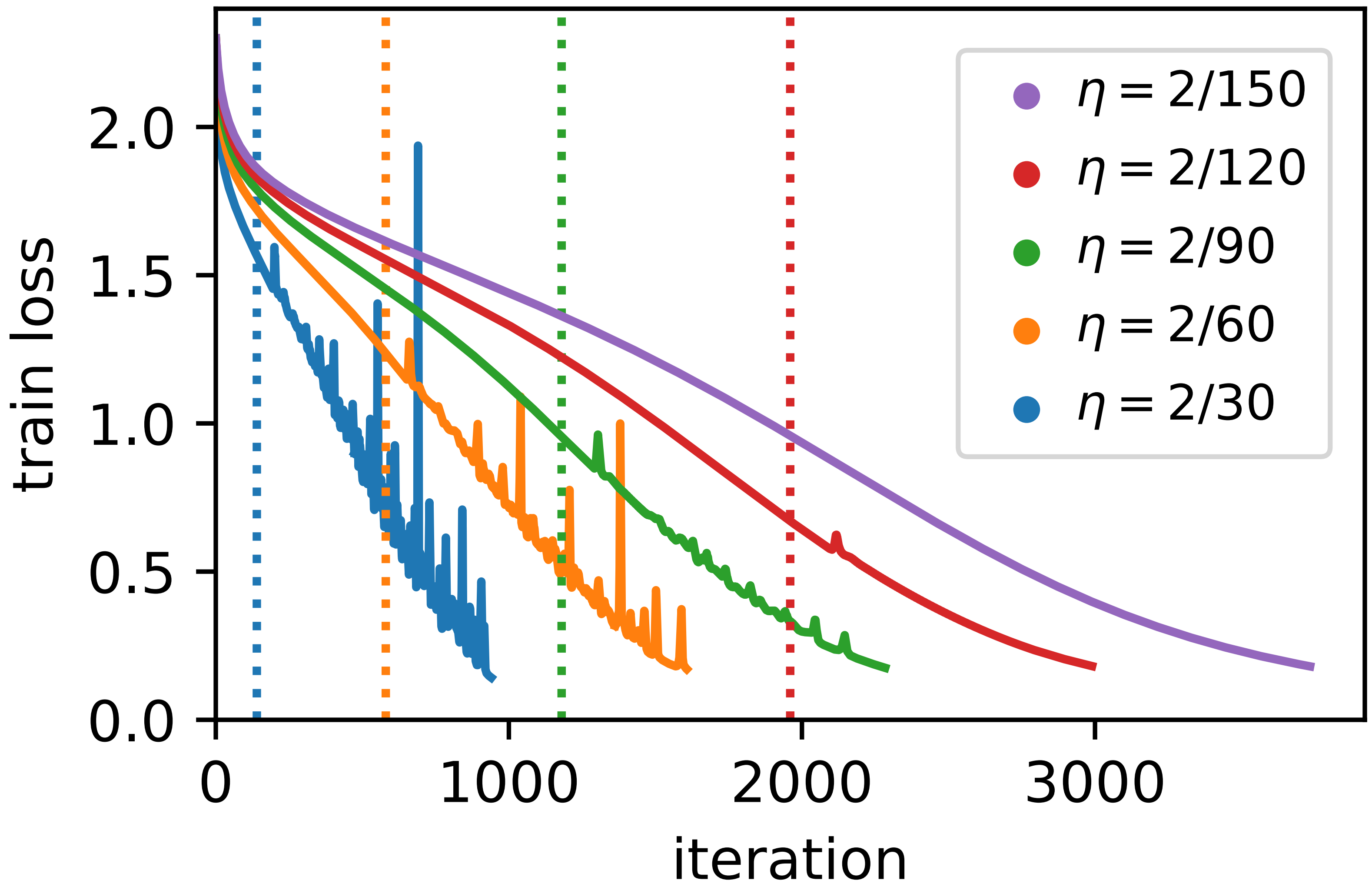
loss landscape ← varying layer-wise curvature

....

inherent instability ← stepsize / learning rate

in DL, all efficient stepsizes are "large", causing training instability

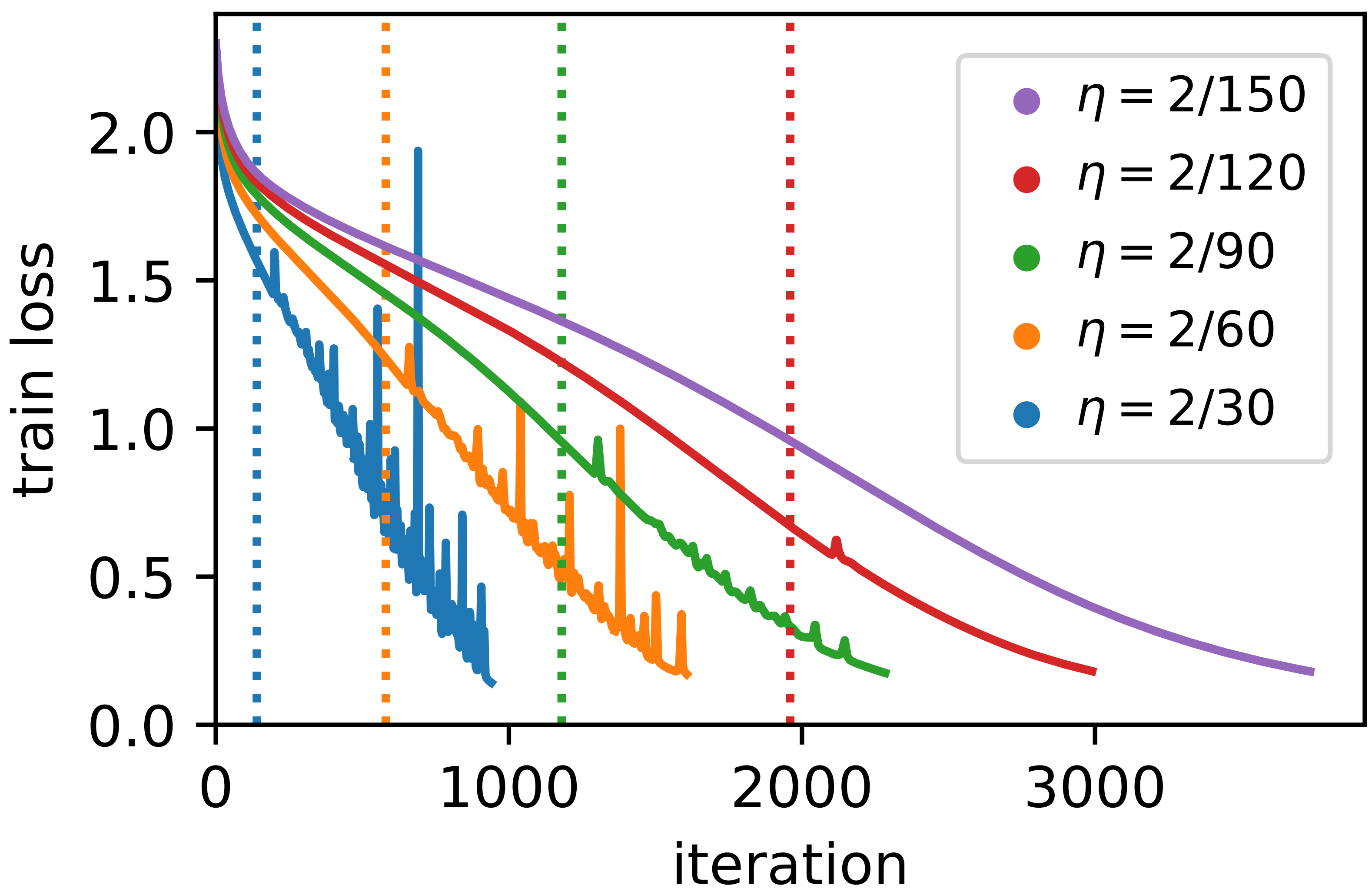
Sandbox: GD + MLP



gradient descent, full batch, 5k subset of CIFAR-10, MLP

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

Sandbox: GD + MLP

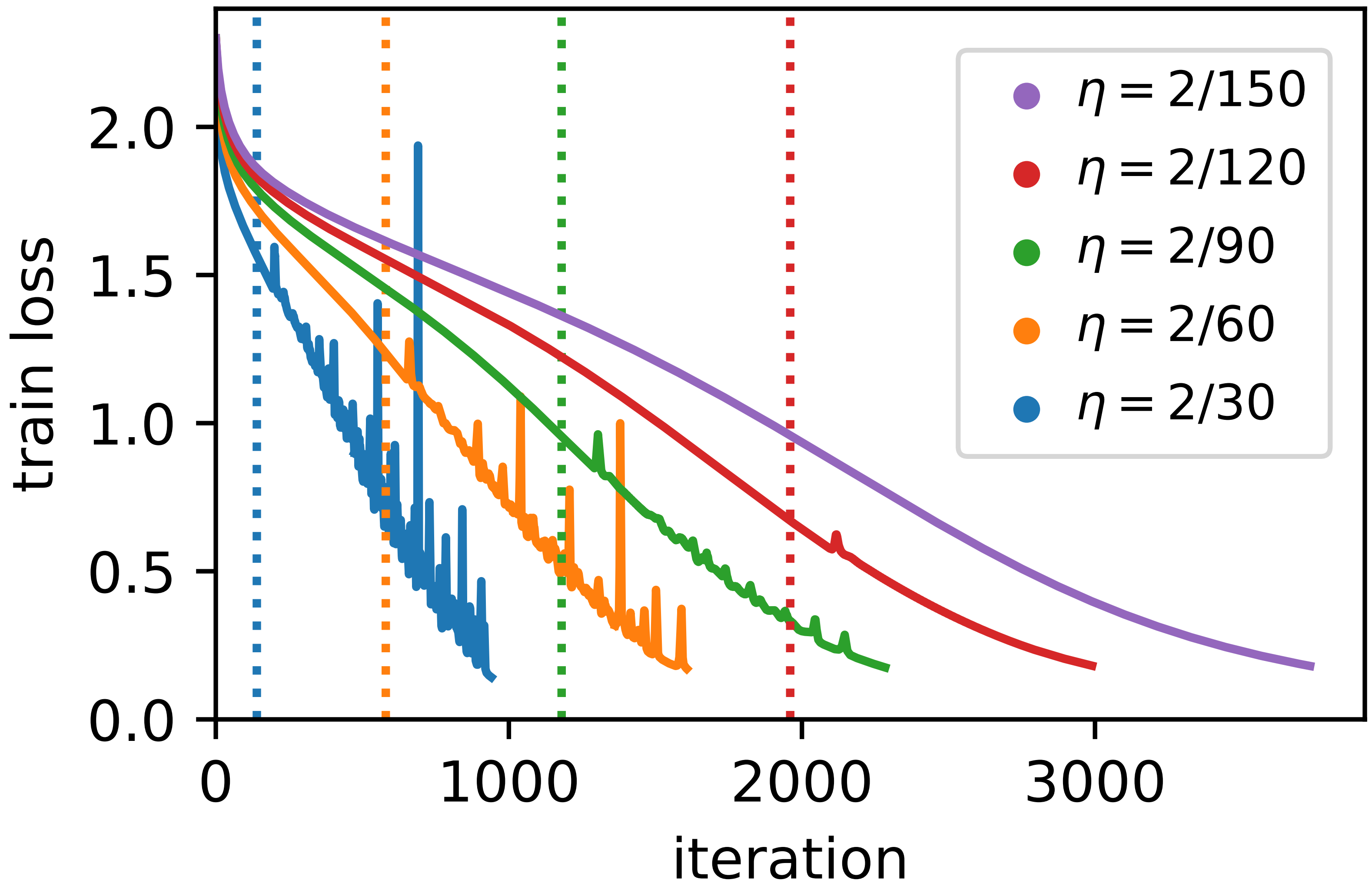


- no randomness
- mild overflow
- OK landscape

gradient descent, full batch, 5k subset of CIFAR-10, MLP

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

Sandbox: GD + MLP



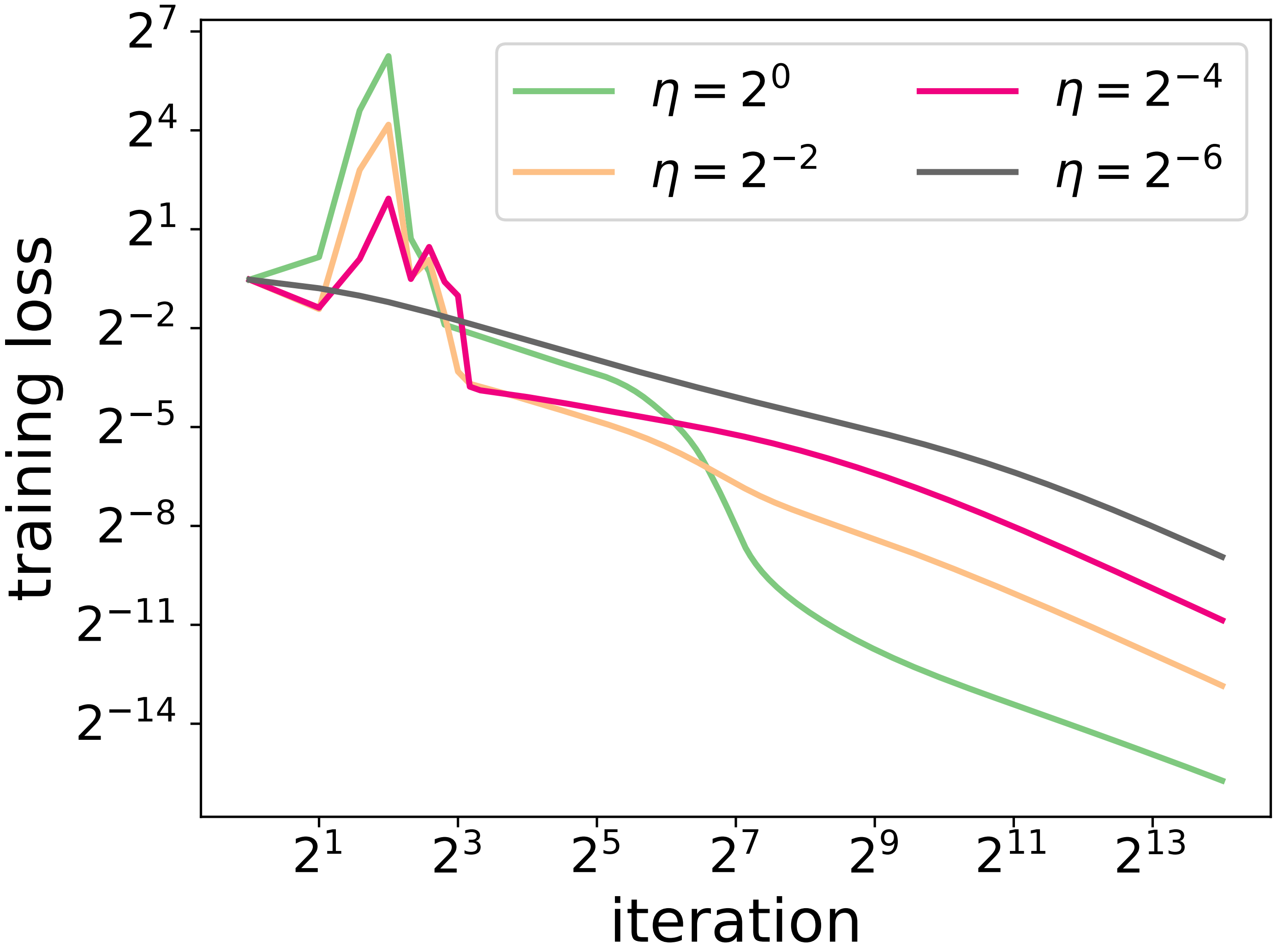
- no randomness
- mild overflow
- OK landscape

but still unstable
(in efficient runs)

gradient descent, full batch, 5k subset of CIFAR-10, MLP

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

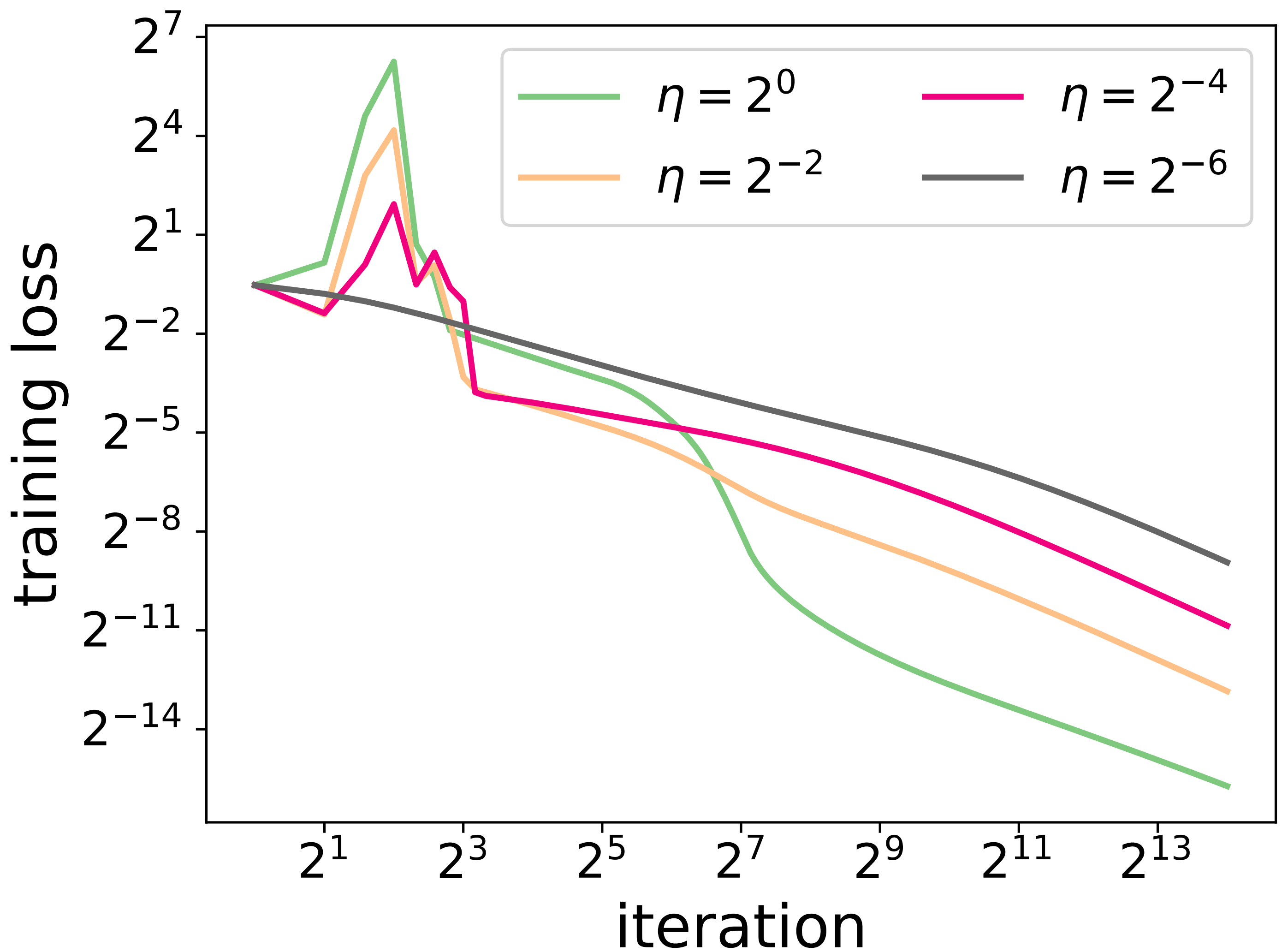
Sandbox²: GD + linear model



GD, 1k subset of MNIST "0" vs "8", logistic regression

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

Sandbox²: GD + linear model

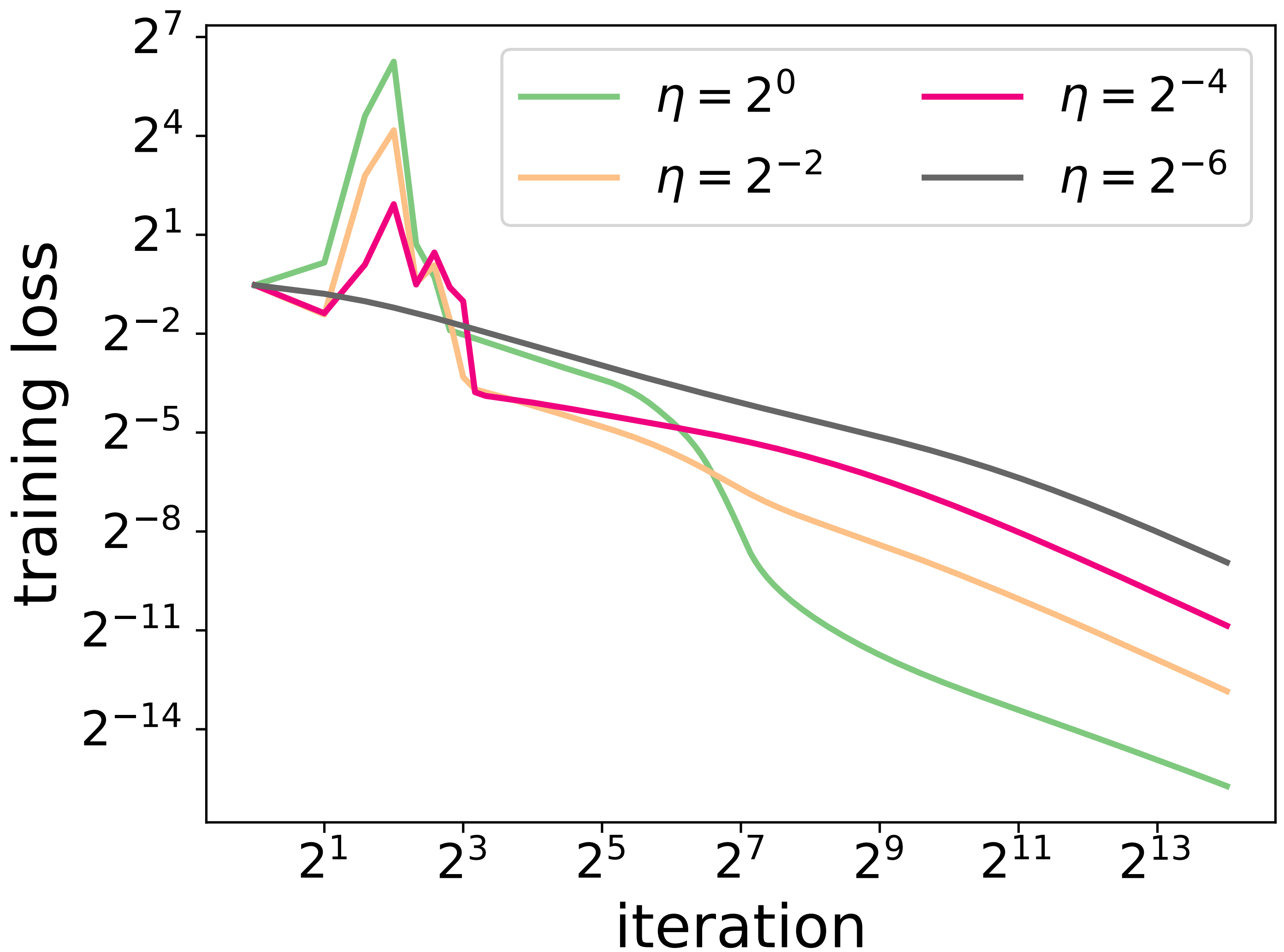


- no randomness
- no overflow
- convex landscape

GD, 1k subset of MNIST "0" vs "8", logistic regression

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

Sandbox²: GD + linear model



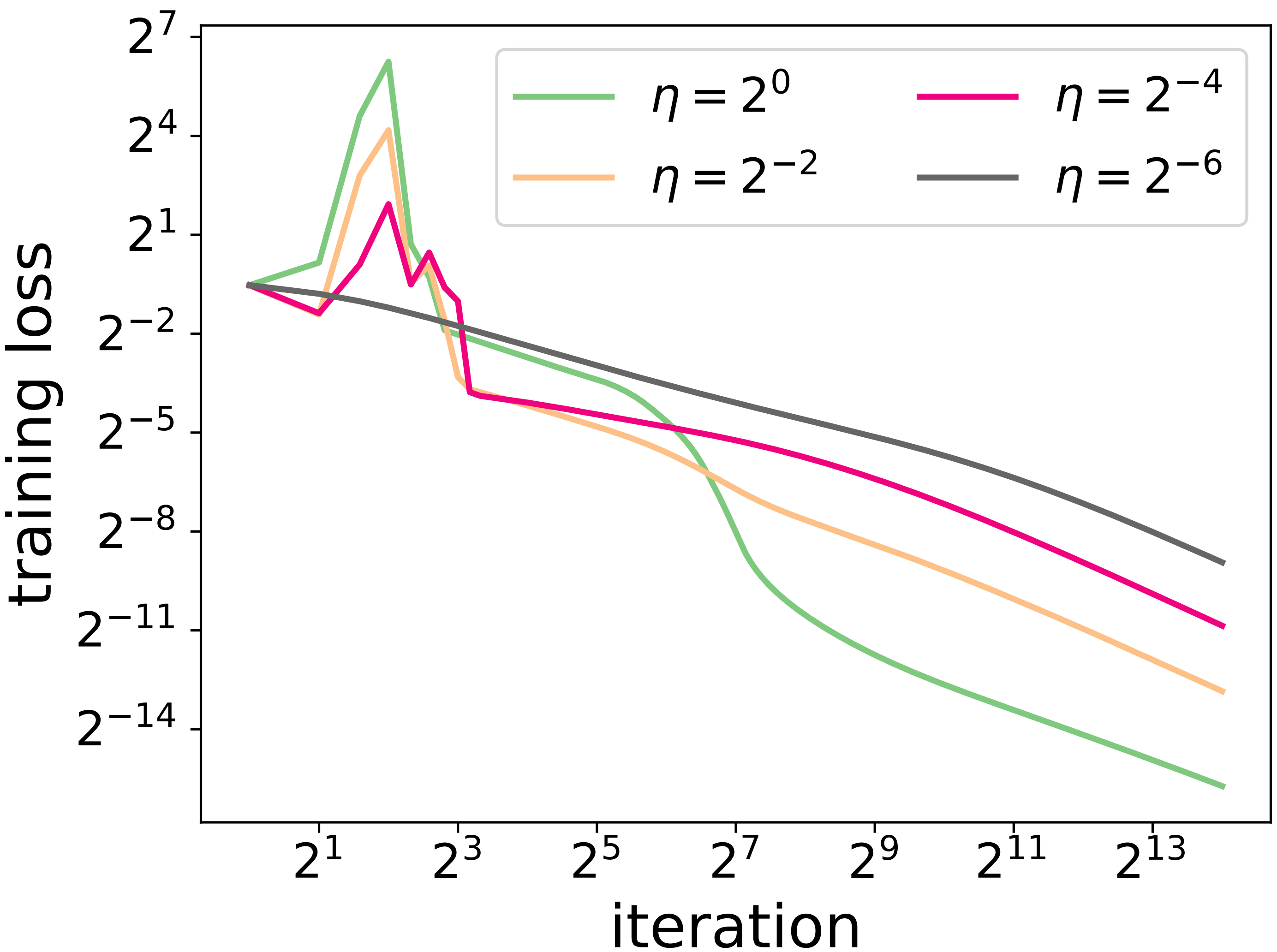
- no randomness
- no overflow
- convex landscape

but still unstable
(in efficient runs)

GD, 1k subset of MNIST "0" vs "8", logistic regression

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

Sandbox²: GD + linear model



- no randomness
- no overflow
- convex landscape

but still unstable
(in efficient runs)

🙀 — me in 2023

GD, 1k subset of MNIST "0" vs "8", logistic regression

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

Infinitesimal stepsize is stable

gradient descent $\qquad \theta_+ = \theta - \eta \nabla L(\theta)$

Infinitesimal stepsize is stable

gradient descent $\quad \theta_+ = \theta - \eta \nabla L(\theta)$

$$\Big\downarrow \; \eta \downarrow 0_+$$

gradient flow $\quad \mathrm{d}\theta = -\nabla L(\theta)\mathrm{d}t$

Infinitesimal stepsize is stable

gradient descent $\theta_+ = \theta - \eta \nabla L(\theta)$

$$\Big\downarrow \; \eta \downarrow 0_+$$

gradient flow $\mathrm{d}\theta = - \nabla L(\theta)\mathrm{d}t$

chain rule
$$
\begin{aligned}
\Rightarrow \;\; \mathrm{d}L(\theta) &= \nabla L(\theta)^\top \mathrm{d}\theta \\
&= -\|\nabla L(\theta)\|^2 \mathrm{d}t \\
&\leq 0
\end{aligned}
$$

Infinitesimal stepsize is stable

gradient descent $\theta_+ = \theta - \eta \nabla L(\theta)$

$$\Big\downarrow \; \eta \downarrow 0_+$$

gradient flow $\mathrm{d}\theta = -\nabla L(\theta)\mathrm{d}t$

chain rule
$$\Rightarrow \quad \mathrm{d}L(\theta) = \nabla L(\theta)^\top \mathrm{d}\theta$$
$$= -\|\nabla L(\theta)\|^2 \mathrm{d}t$$
$$\leq 0$$

integration $\Rightarrow \quad L(\theta) \downarrow$

Infinitesimal stepsize is stable

gradient descent $\theta_+ = \theta - \eta \nabla L(\theta)$

$$\Big\downarrow \; \eta \downarrow 0_+$$

gradient flow $\mathrm{d}\theta = -\nabla L(\theta)\mathrm{d}t$

chain rule $\Rightarrow \; \mathrm{d}L(\theta) = \nabla L(\theta)^\top \mathrm{d}\theta$

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

$$\leq 0$$

integration $\Rightarrow \; L(\theta) \downarrow$

GD with infinitesimal stepsize is stable

Infinitesimal stepsize is stable

GD $\to$ gradient flow

Infinitesimal stepsize is stable

GD → gradient flow

✅ **momentum.** GD with momentum → second order ODE

Su, Boyd, Candes. “A differential equation for modeling Nesterov's accelerated gradient method: theory and insights.” JMLR 2016

Infinitesimal stepsize is stable

GD → gradient flow

✅ **momentum.** GD with momentum → second order ODE

✅ **mini batch.** SGD → gradient flow + o(1) diffusion (SDE)

Su, Boyd, Candes. "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights." JMLR 2016

Li, Tai, E. "Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations." JMLR 2019

Infinitesimal stepsize is stable

GD → gradient flow

✅ **momentum.** GD with momentum → second order ODE

✅ **mini batch.** SGD → gradient flow + o(1) diffusion (SDE)

these ODE/SDEs minimize certain potential

Su, Boyd, Candes. "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights." JMLR 2016

Li, Tai, E. "Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations." JMLR 2019

Infinitesimal stepsize is stable

GD → gradient flow

✅ **momentum.** GD with momentum → second order ODE

✅ **mini batch.** SGD → gradient flow + o(1) diffusion (SDE)

these ODE/SDEs minimize certain potential

❓ **adaptivity.** Adam: unclear continuous limit

Su, Boyd, Candes. "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights." JMLR 2016

Li, Tai, E. "Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations." JMLR 2019

From infinitesimal to small stepsize

Descent lemma. For GD, $L(w_t)$ decreases monotonically if

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

From infinitesimal to small stepsize

Descent lemma. For GD, $L(w_t)$ decreases monotonically if

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

quadratics $\quad L(\theta) = \frac{1}{2}\lambda\theta^2$

From infinitesimal to small stepsize

Descent lemma. For GD, $L(w_t)$ decreases monotonically if

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

quadratics $\quad L(\theta) = \frac{1}{2}\lambda\theta^2$

Hessian $\quad \nabla^2 L(\theta) = \lambda$

From infinitesimal to small stepsize

Descent lemma. For GD, $L(w_t)$ decreases monotonically if

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

quadratics $L(\theta) = \frac{1}{2}\lambda\theta^2$

Hessian $\nabla^2 L(\theta) = \lambda$

GD
$$\begin{aligned} \theta_+ &= \theta - \eta \nabla L(\theta) \\ &= (1 - \lambda\eta)\theta \end{aligned}$$

From infinitesimal to small stepsize

Descent lemma. For GD, $L(w_t)$ decreases monotonically if

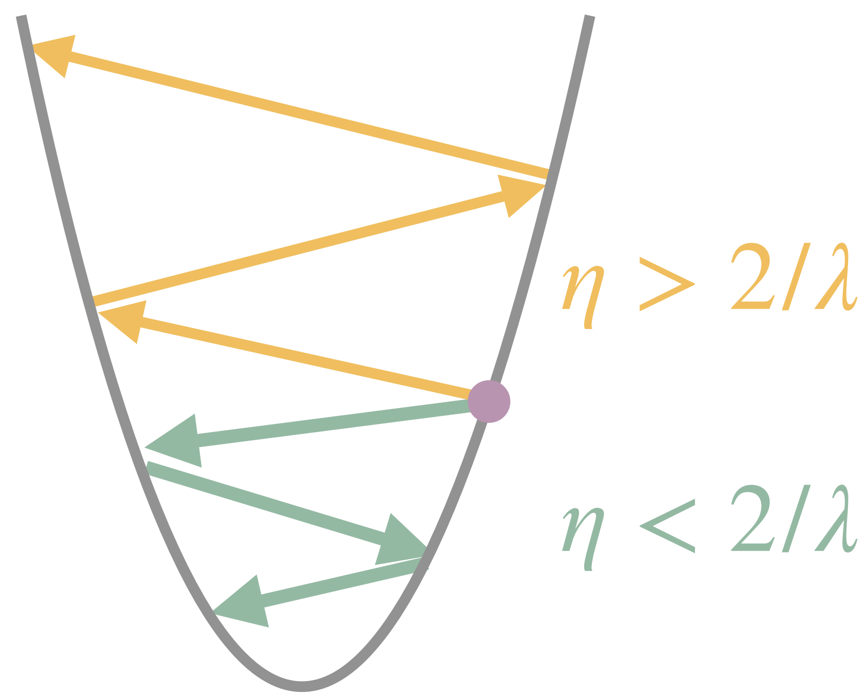
$$\eta < \frac{2}{\sup \|\nabla^2 L(\,\cdot\,)\|}$$

quadratics $\qquad L(\theta) = \frac{1}{2}\lambda\theta^2$

Hessian $\qquad \nabla^2 L(\theta) = \lambda$

GD $\qquad \theta_+ = \theta - \eta \nabla L(\theta)$

$\qquad\qquad = (1 - \lambda\eta)\theta$



From infinitesimal to small stepsize

Descent lemma. For GD, $L(w_t)$ decreases monotonically if

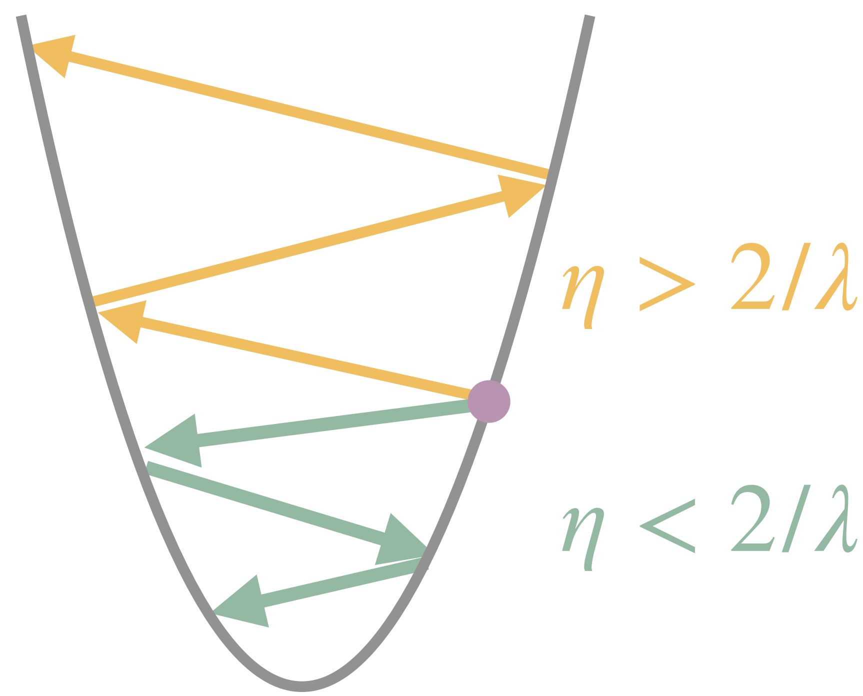
$$\eta < \frac{2}{\sup \|\nabla^2 L(\,\cdot\,)\|}$$

small stepsize implies descent

quadratics $L(\theta) = \frac{1}{2}\lambda\theta^2$

Hessian $\nabla^2 L(\theta) = \lambda$

GD
$$\theta_+ = \theta - \eta \nabla L(\theta)$$
$$= (1 - \lambda\eta)\theta$$



From infinitesimal to small stepsize

Descent lemma. For GD, $L(w_t)$ decreases monotonically if

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

small stepsize implies descent

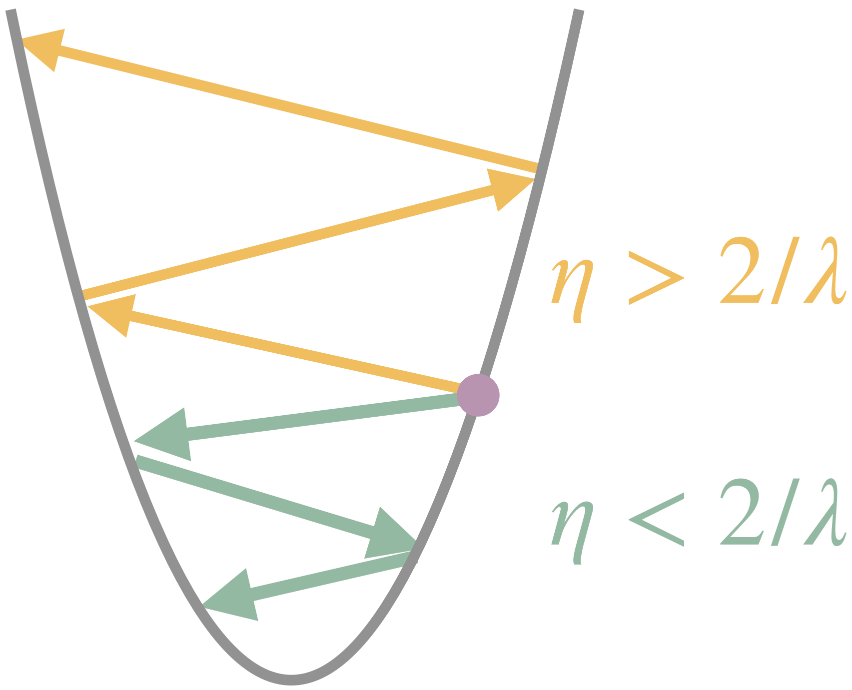
cornerstone of optimization theory

quadratics $\quad L(\theta) = \frac{1}{2}\lambda\theta^2$

Hessian $\quad \nabla^2 L(\theta) = \lambda$

GD $\quad \theta_+ = \theta - \eta \nabla L(\theta)$

$\quad = (1 - \lambda\eta)\theta$



From small to large stepsize

Large stepsize. A stepsize η is large for GD if

$$L(\theta_t) \text{ does not decrease monotonically}$$

From small to large stepsize

Large stepsize. A stepsize η is large for GD if

$$L(\theta_t) \text{ does not decrease monotonically}$$

Dynamical stability. If GD with large η converges to stationary point (why?), then in “regular” cases

From small to large stepsize

Large stepsize. A stepsize η is large for GD if

$$L(\theta_t) \text{ does not decrease monotonically}$$

Dynamical stability. If GD with large η converges to stationary point (why?), then in "regular" cases

$$\|\nabla^2 L(\theta_\infty)\| < \frac{2}{\eta}$$

Wu, Ma, E. "How SGD selects the global minima in over-parameterized learning: a dynamical stability perspective." NeurIPS 2018.

From small to large stepsize

Large stepsize. A stepsize η is large for GD if

$$L(\theta_t) \text{ does not decrease monotonically}$$

Dynamical stability. If GD with large η converges to stationary point (why?), then in "regular" cases

$$\|\nabla^2 L(\theta_\infty)\| < \frac{2}{\eta}$$

Intuition. Descent lemma is tight for quadratics

Wu, Ma, E. "How SGD selects the global minima in over-parameterized learning: a dynamical stability perspective." NeurIPS 2018.

From small to large stepsize

Large stepsize. A stepsize η is large for GD if

$$L(\theta_t) \text{ does not decrease monotonically}$$

Dynamical stability. If GD with large η converges to stationary point (why?), then in "regular" cases

$$\|\nabla^2 L(\theta_\infty)\| < \frac{2}{\eta}$$

Intuition. Descent lemma is tight for quadratics

alternative names: linear stability, Lyapunov stability...

Wu, Ma, E. "How SGD selects the global minima in over-parameterized learning: a dynamical stability perspective." NeurIPS 2018.

From small to large stepsize

Large stepsize. A stepsize η is large for GD if

$$L(\theta_t) \text{ does not decrease monotonically}$$

Dynamical stability. If GD with large η converges to stationary point (why?), then in "regular" cases

$$\|\nabla^2 L(\theta_\infty)\| < \frac{2}{\eta}$$

sharpness penalty

Intuition. Descent lemma is tight for quadratics

alternative names: linear stability, Lyapunov stability...

Wu, Ma, E. "How SGD selects the global minima in over-parameterized learning: a dynamical stability perspective." NeurIPS 2018.

From small to large stepsize

Sharpness penalty. If label-noise* SGD converges, under suitable assumptions,

$$\mathrm{tr}(\nabla^2 L(\theta_\infty)) < O(1/\eta)$$

Damian, Ma, Lee. "Label noise SGD provably prefers flat global minimizers." NeurIPS 2021

Li, Wang, Arora. "What happens after SGD reaches zero loss?—A mathematical framework." ICLR 2022

From small to large stepsize

Sharpness penalty. If label-noise* SGD converges, under suitable assumptions,

$$\mathrm{tr}(\nabla^2 L(\theta_\infty)) < O(1/\eta)$$

*for general SGD, the penalty also depends on noise covariance

Damian, Ma, Lee. "Label noise SGD provably prefers flat global minimizers." NeurIPS 2021

Li, Wang, Arora. "What happens after SGD reaches zero loss?—A mathematical framework." ICLR 2022

From small to large stepsize

Sharpness penalty. If label-noise* SGD converges, under suitable assumptions,

$$\mathrm{tr}(\nabla^2 L(\theta_\infty)) < O(1/\eta)$$

*for general SGD, the penalty also depends on noise covariance

training instability: $L(\theta_t)$ oscillates for $t = 1,2,\ldots$

minimizer flatness: $|L(\theta_\infty + \epsilon) - L(\theta_\infty)|$ is small

Damian, Ma, Lee. "Label noise SGD provably prefers flat global minimizers." NeurIPS 2021

Li, Wang, Arora. "What happens after SGD reaches zero loss?—A mathematical framework." ICLR 2022

From small to large stepsize

Sharpness penalty. If label-noise* SGD converges, under suitable assumptions,

$$\operatorname{tr}(\nabla^2 L(\theta_\infty)) < O(1/\eta)$$

*for general SGD, the penalty also depends on noise covariance

training instability: $L(\theta_t)$ oscillates for $t = 1,2,\ldots$

minimizer flatness: $|L(\theta_\infty + \epsilon) - L(\theta_\infty)|$ is small

large stepsize: less stable training, but flatter minima

Damian, Ma, Lee. "Label noise SGD provably prefers flat global minimizers." NeurIPS 2021

Li, Wang, Arora. "What happens after SGD reaches zero loss?—A mathematical framework." ICLR 2022

From small to large stepsize

Sharpness penalty. If label-noise* SGD converges, under suitable assumptions,

$$\mathrm{tr}(\nabla^2 L(\theta_\infty)) < O(1/\eta)$$

*for general SGD, the penalty also depends on noise covariance

training instability: $L(\theta_t)$ oscillates for $t = 1,2,\ldots$

minimizer flatness: $|L(\theta_\infty + \epsilon) - L(\theta_\infty)|$ is small

large stepsize: less stable training, but flatter minima

convergence? generalization?

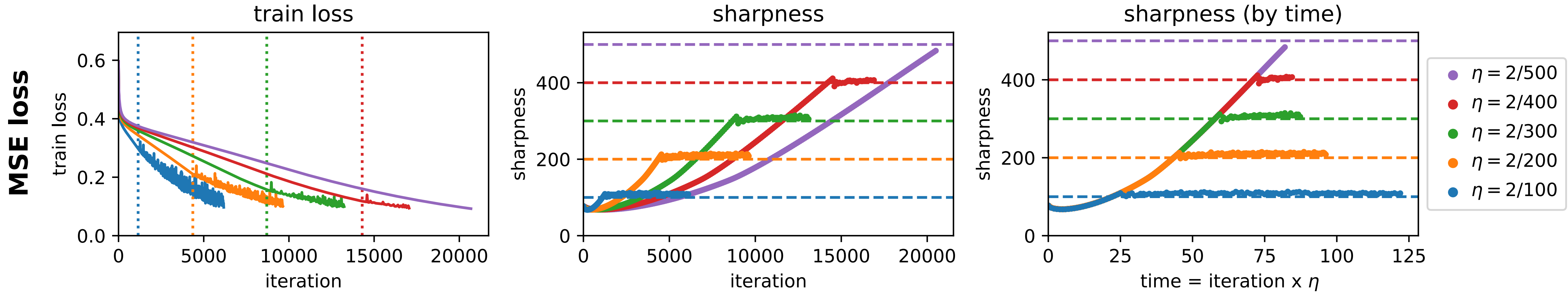
Damian, Ma, Lee. "Label noise SGD provably prefers flat global minimizers." NeurIPS 2021

Li, Wang, Arora. "What happens after SGD reaches zero loss?—A mathematical framework." ICLR 2022

From small to large stepsize

progressive sharpening

edge of stability



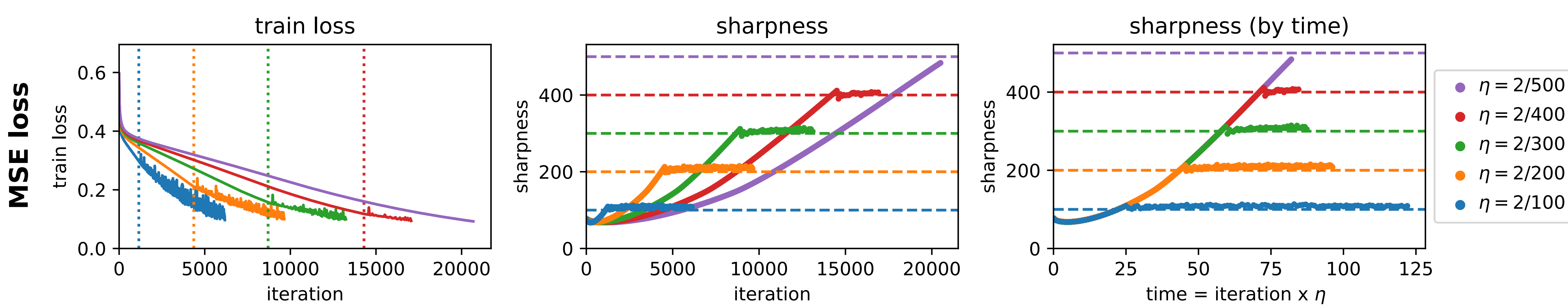
Cohen, Kaur, Li, Kolter, Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability." ICLR 2021

From small to large stepsize

progressive sharpening

even starting satisfying descent lemma, sharpness increases along GD path until hitting $2/\eta$

edge of stability



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

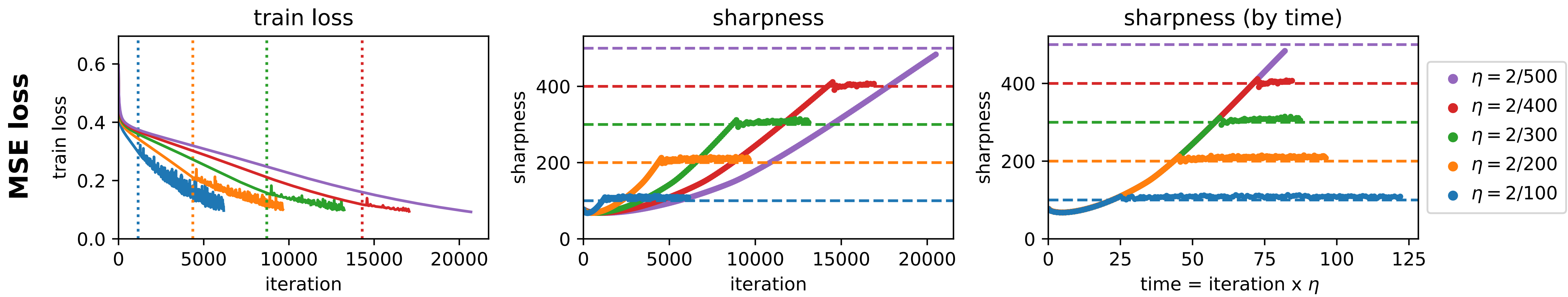
From small to large stepsize

progressive sharpening

even starting satisfying descent lemma, sharpness increases along GD path until hitting $2/\eta$

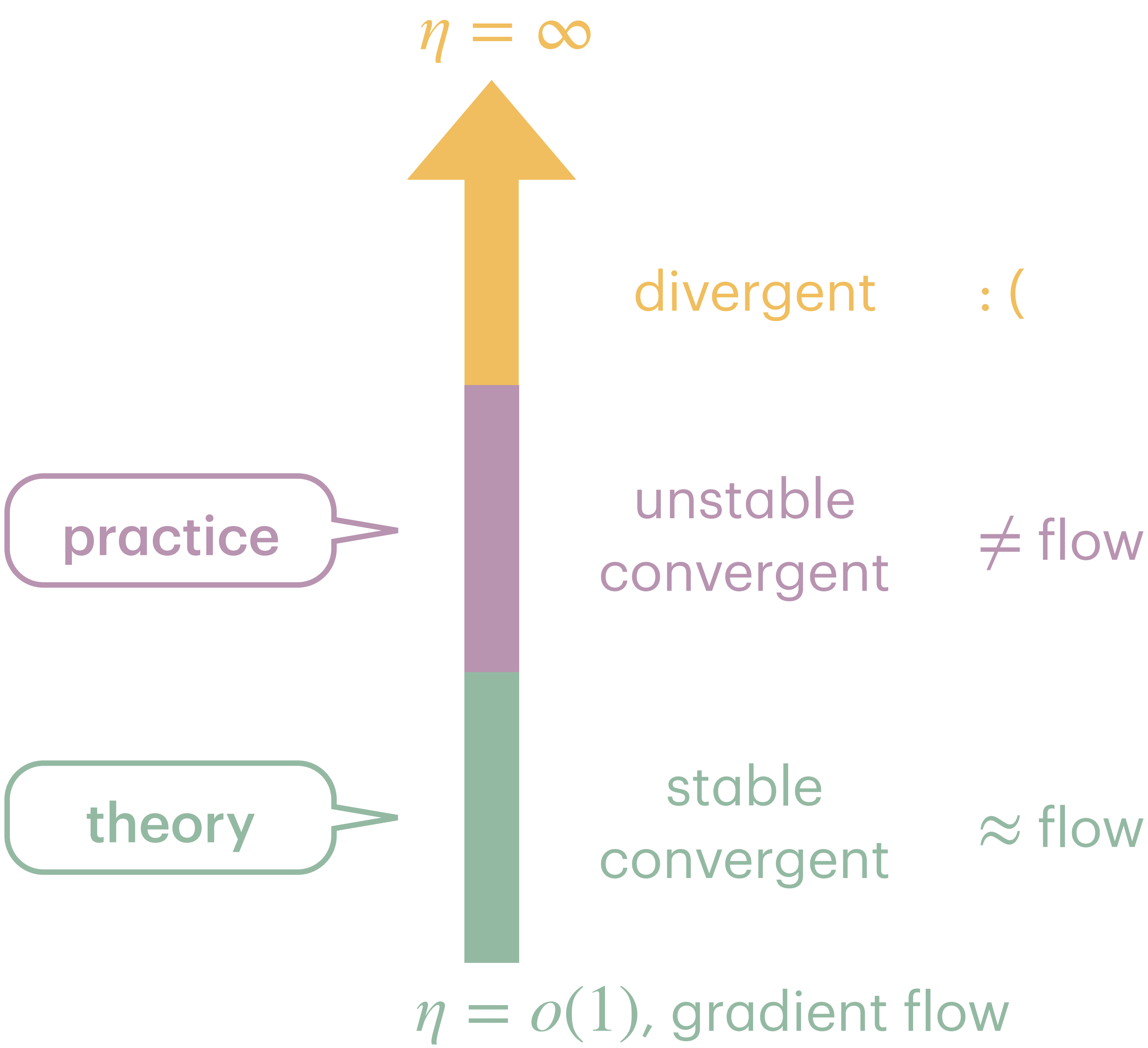
edge of stability

after **PS**, sharpness oscillates around $2/\eta$ for a while



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

From small to large stepsize



We will cover

Part 1: large stepsizes accelerate optimization

Part 2: large stepsizes prevent overfitting

We will cover

Part 1: large stepsizes accelerate optimization

Part 2: large stepsizes prevent overfitting

- **theory** & **insights** through clean **examples**

We will cover

Part 1: large stepsizes accelerate optimization

Part 2: large stepsizes prevent overfitting

- **theory** & **insights** through clean **examples**
- known results & open problems

We will cover

Part 1: large stepsizes accelerate optimization

Part 2: large stepsizes prevent overfitting

- **theory** & **insights** through clean **examples**
- known results & open problems
- why you should consider working on this!

We won’t cover but worth checking

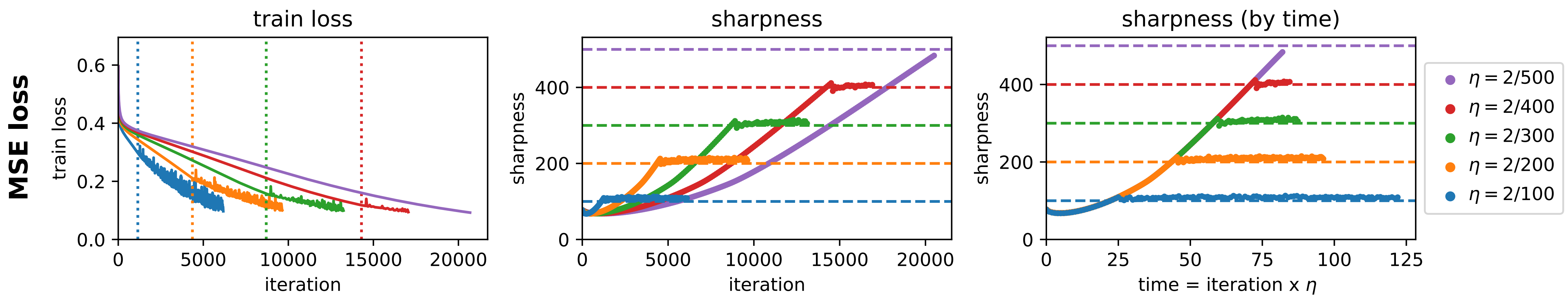
(1/many) experimental science of large stepsize

*check our website for more references

We won't cover but worth checking

(1/many) experimental science of large stepsize

📋 progressive sharpening & edge of stability

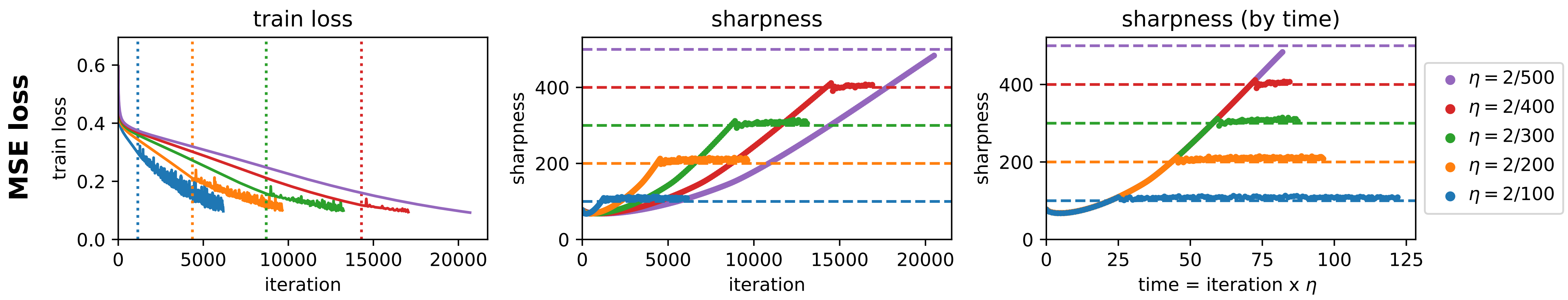


*check our website for more references

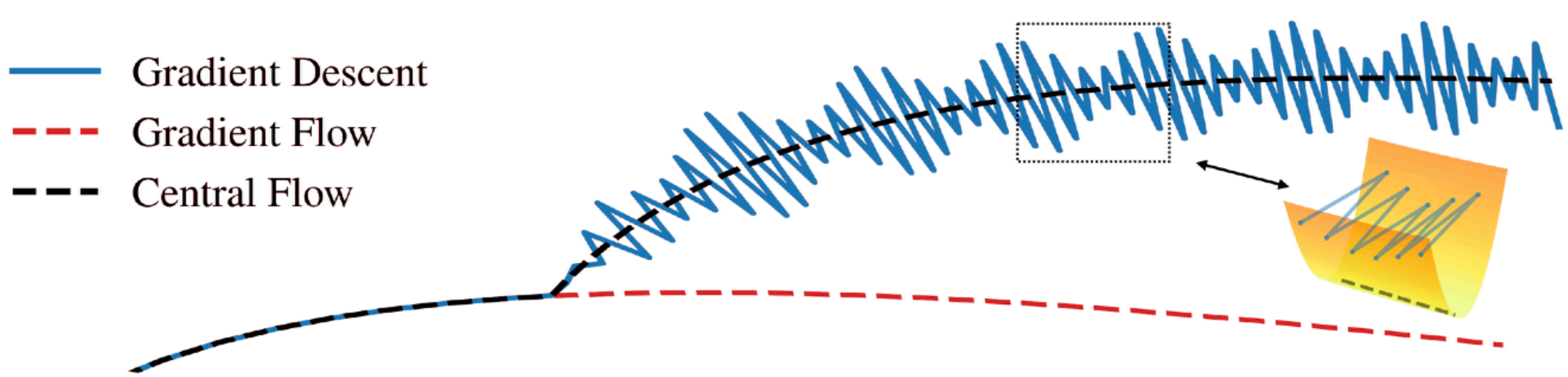
We won’t cover but worth checking

(1/many) experimental science of large stepsize

📋 progressive sharpening & edge of stability



📋 central flow: an approximation of the trajectory



*check our website for more references

Cohen, Damian, Talwalkar, Kolter, Lee. “Understanding optimization in deep learning with central flows.” ICLR 2025

We won’t cover but worth checking

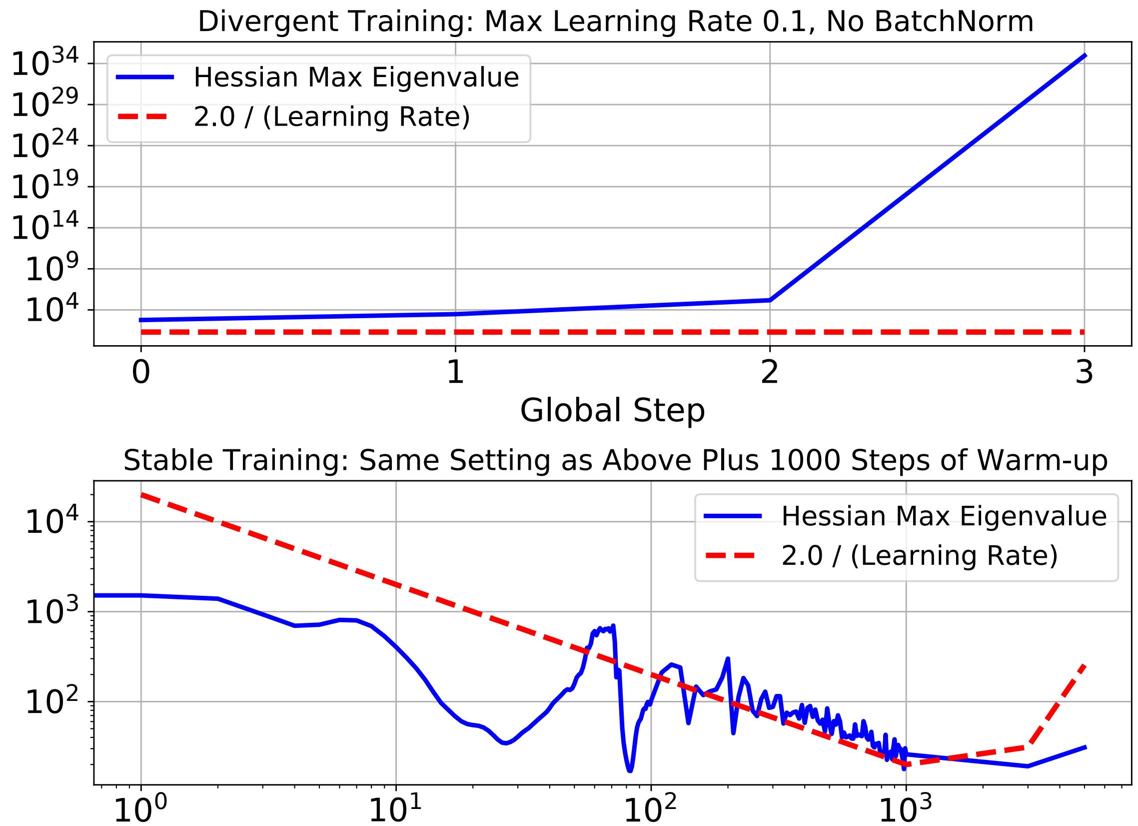
(2/many) optimizer-landscape codesign

*check our website for more references

We won't cover but worth checking

(2/many) optimizer-landscape codesign

📋 learning rate warmup navigates to flatter region



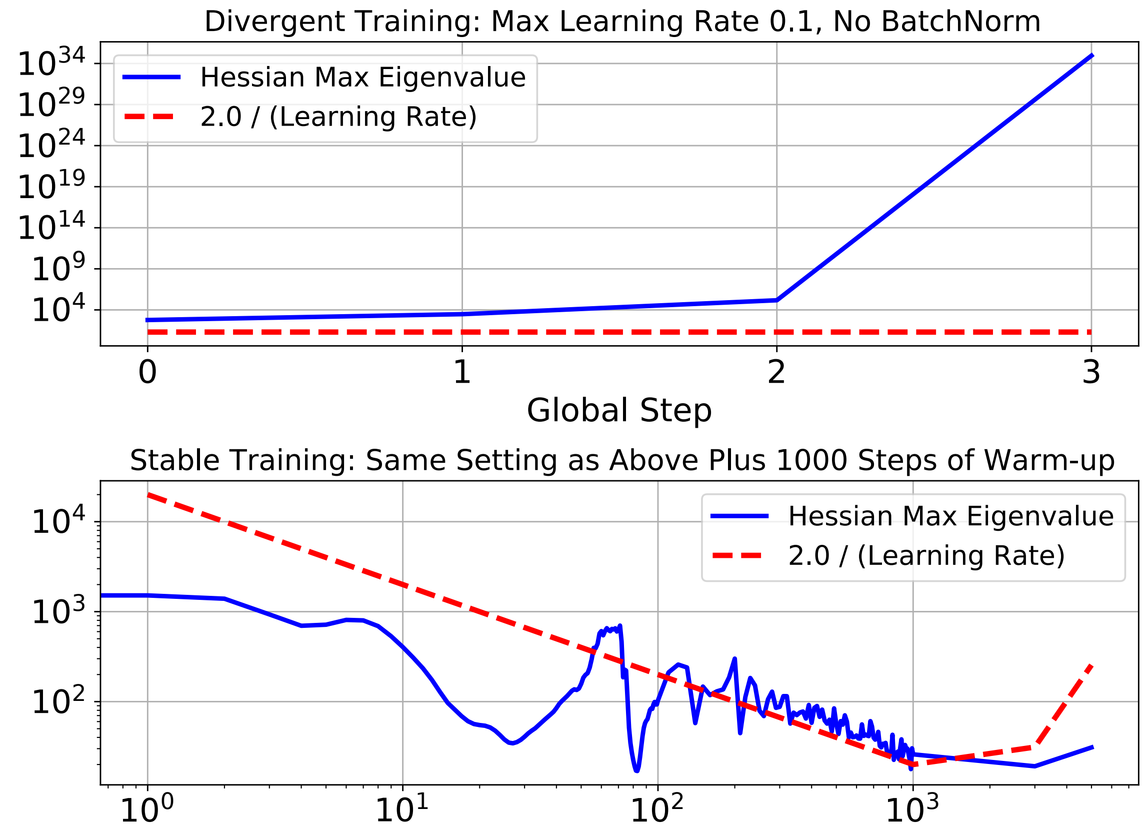
*check our website for more references

Gilmer, Ghorbani, Garg, et al. "A loss curvature perspective on training instability in deep learning." ICLR 2022

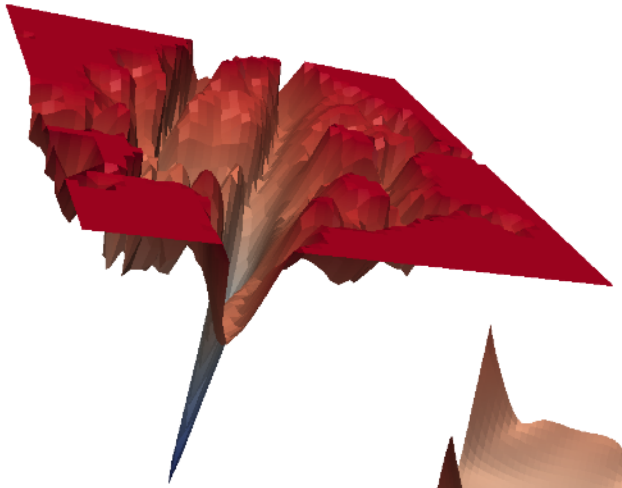
We won't cover but worth checking

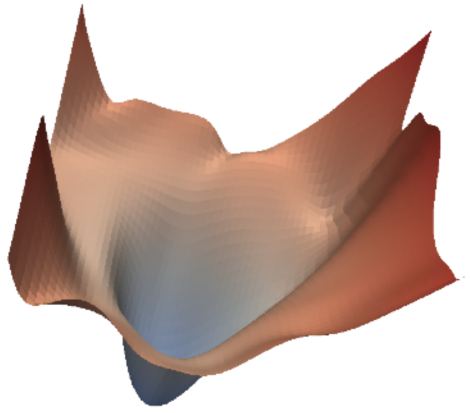
(2/many) optimizer-landscape codesign

📋 learning rate warmup navigates to flatter region



📋 sharpness-aware minimization





*check our website for more references

Gilmer, Ghorbani, Garg, et al. "A loss curvature perspective on training instability in deep learning." ICLR 2022

Foret, Kleiner, Mobahi, Neyshabur. "Sharpness-aware minimization for efficiently improving generalization." ICLR 2021

Part 1: optimization

Review: classical optimization theory

A modern take: acceleration via large stepsizes

Summary, open problems, Q&A

Part 2: generalization

Review: descent lemma

For GD, $L(\theta_t)$ decreases monotonically for small η such that

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

Review: descent lemma

For GD, $L(\theta_t)$ decreases monotonically for small η such that

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

Proof.

$$
\begin{aligned}
L(\theta_+) &= L(\theta - \eta \nabla L(\theta)) \\
&= L(\theta) - \eta \|\nabla L(\theta)\|^2 + \frac{\eta^2}{2} \nabla L(\theta)^\top \nabla^2 L(\nu) \nabla L(\theta) \\
&\leq L(\theta) - \eta \|\nabla L(\theta)\|^2 \left(1 - \frac{\eta}{2} \|\nabla^2 L(\nu)\|\right) \\
&\leq L(\theta)
\end{aligned}
$$

Review: descent lemma

For GD, $L(\theta_t)$ decreases monotonically for small η such that

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

Proof.

$$L(\theta_+) = L(\theta - \eta \nabla L(\theta))$$

GD step

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

$$\leq L(\theta) - \eta \|\nabla L(\theta)\|^2 \left(1 - \frac{\eta}{2} \|\nabla^2 L(\nu)\|\right)$$

$$\leq L(\theta)$$

Review: descent lemma

For GD, $L(\theta_t)$ decreases monotonically for small η such that

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

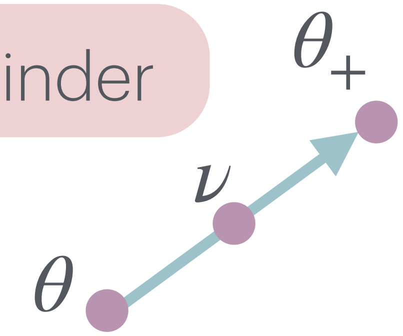
Proof.

$$L(\theta_+) = L(\theta - \eta \nabla L(\theta))$$

GD step

Taylor remainder

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



$$\leq L(\theta) - \eta \|\nabla L(\theta)\|^2 \left(1 - \frac{\eta}{2} \|\nabla^2 L(\nu)\|\right)$$

$$\leq L(\theta)$$

Review: descent lemma

For GD, $L(\theta_t)$ decreases monotonically for small η such that

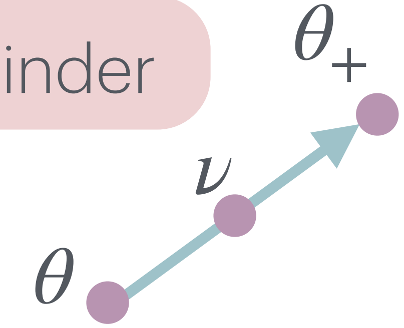
$$\eta < \frac{2}{\sup \|\nabla^2 L(\,\cdot\,)\|}$$

Proof.

$$L(\theta_+) = L(\theta - \eta \nabla L(\theta))$$

GD step

Taylor remainder



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

$$\leq L(\theta) - \eta \|\nabla L(\theta)\|^2 \left(1 - \frac{\eta}{2} \|\nabla^2 L(\nu)\|\right)$$

operator norm

$$\leq L(\theta)$$

Review: descent lemma

For GD, $L(\theta_t)$ decreases monotonically for small η such that

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

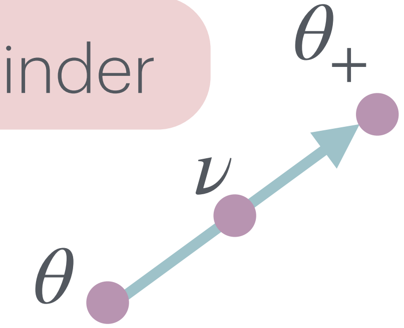
Proof.

$$L(\theta_+) = L(\theta - \eta \nabla L(\theta))$$

GD step

Taylor remainder

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



$$\leq L(\theta) - \eta \|\nabla L(\theta)\|^2 \left(1 - \frac{\eta}{2} \|\nabla^2 L(\nu)\|\right)$$

operator norm

$$\leq L(\theta)$$

small stepsize

Review: descent lemma

For GD, $L(\theta_t)$ decreases monotonically for small η such that

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

Proof.

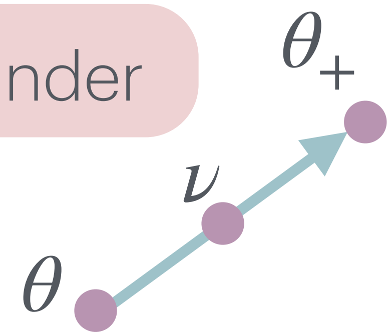
$$\begin{aligned}
L(\theta_+) &= L(\theta - \eta \nabla L(\theta)) \\
&= L(\theta) - \eta \|\nabla L(\theta)\|^2 + \frac{\eta^2}{2} \nabla L(\theta)^\top \nabla^2 L(\nu) \nabla L(\theta) \\
&\leq L(\theta) - \eta \|\nabla L(\theta)\|^2 \left(1 - \frac{\eta}{2} \|\nabla^2 L(\nu)\|\right) \\
&\leq L(\theta)
\end{aligned}$$

GD step

Taylor remainder

operator norm

small stepsize



this descent lemma can be generalized

Review: convergence rates

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer $w*$. For GD with $\eta = 1$, we have

descent lemma $L(\theta_t) \downarrow$

Review: convergence rates

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer $w*$. For GD with $\eta = 1$, we have

descent lemma $L(\theta_t) \downarrow$

convexity $L(\theta_t) - \min L \leq \dfrac{\|\theta_0 - \theta*\|^2}{2t}$

Review: convergence rates

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer $w*$. For GD with $\eta = 1$, we have

descent lemma $L(\theta_t) \downarrow$

convexity $$L(\theta_t) - \min L \leq \frac{\|\theta_0 - \theta*\|^2}{2t}$$

α-strong convexity $$L(\theta_t) - \min L \leq e^{-\alpha t}(L(\theta_0) - \min L)$$

Review: convergence rates

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer w^*. For GD with $\eta = 1$, we have

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

convexity $\quad L(\theta_t) - \min L \leq \dfrac{\|\theta_0 - \theta^*\|^2}{2t}$

α-strong convexity $\quad L(\theta_t) - \min L \leq e^{-\alpha t}(L(\theta_0) - \min L)$

number of steps to get ϵ-error:

$O\big(1/\epsilon\big)$ and $O\big(\kappa \log(1/\epsilon)\big)$

Review: convergence rates

Let L be 1-smooth ($\|\nabla^2 L\| \leq 1$) with finite minimizer w^*. For GD with $\eta = 1$, we have

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

convexity $\quad L(\theta_t) - \min L \leq \dfrac{\|\theta_0 - \theta^*\|^2}{2t}$

α**-strong convexity** $\quad L(\theta_t) - \min L \leq e^{-\alpha t}(L(\theta_0) - \min L)$

number of steps to get ϵ-error:

$$O\big(1/\epsilon\big) \text{ and } O\big(\kappa \log(1/\epsilon)\big)$$

$\kappa = 1/\alpha$, condition number

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \le \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \leq \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Proof.

step 1:

$$\mathrm{d}\frac{1}{2}\|\theta_t - \nu\|^2 = \langle \theta_t - \nu, \mathrm{d}\theta_t \rangle = \langle \theta_t - \nu, -\nabla L(\theta_t) \rangle \mathrm{d}t \leq L(\nu) - L(\theta_t)$$

step 2:

$$\frac{1}{2}\|\theta_t - \nu\|^2 - \frac{1}{2}\|\theta_0 - \nu\|^2 \leq \int_0^t L(\nu) - L(\theta_s)\mathrm{d}s \leq t\big(L(\theta) - L(\theta_t)\big)$$

step 3: rearranging terms

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \leq \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Proof.

step 1:

chain rule

$$\mathrm{d}\frac{1}{2}\|\theta_t - \nu\|^2 = \langle \theta_t - \nu, \mathrm{d}\theta_t \rangle = \langle \theta_t - \nu, -\nabla L(\theta_t) \rangle \mathrm{d}t \leq L(\nu) - L(\theta_t)$$

step 2:

$$\frac{1}{2}\|\theta_t - \nu\|^2 - \frac{1}{2}\|\theta_0 - \nu\|^2 \leq \int_0^t L(\nu) - L(\theta_s)\mathrm{d}s \leq t\big(L(\theta) - L(\theta_t)\big)$$

step 3: rearranging terms

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \le \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Proof.

step 1:

$$\mathrm{d}\frac{1}{2}\|\theta_t - \nu\|^2 \overset{\text{chain rule}}{=} \langle \theta_t - \nu, \mathrm{d}\theta_t \rangle \overset{\text{gradient flow}}{=} \langle \theta_t - \nu, -\nabla L(\theta_t) \rangle \mathrm{d}t \le L(\nu) - L(\theta_t)$$

step 2:

$$\frac{1}{2}\|\theta_t - \nu\|^2 - \frac{1}{2}\|\theta_0 - \nu\|^2 \le \int_0^t L(\nu) - L(\theta_s)\mathrm{d}s \le t\big(L(\theta) - L(\theta_t)\big)$$

step 3: rearranging terms

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \leq \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Proof.

step 1:

$$\mathrm{d}\frac{1}{2}\|\theta_t - \nu\|^2 \overset{\text{chain rule}}{=} \langle \theta_t - \nu, \mathrm{d}\theta_t \rangle \overset{\text{gradient flow}}{=} \langle \theta_t - \nu, -\nabla L(\theta_t)\rangle \mathrm{d}t \overset{\text{convexity}}{\leq} L(\nu) - L(\theta_t)$$

step 2:

$$\frac{1}{2}\|\theta_t - \nu\|^2 - \frac{1}{2}\|\theta_0 - \nu\|^2 \leq \int_0^t L(\nu) - L(\theta_s)\mathrm{d}s \leq t\big(L(\theta) - L(\theta_t)\big)$$

step 3: rearranging terms

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \leq \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Proof.

step 1:

$$\mathrm{d}\frac{1}{2}\|\theta_t - \nu\|^2 \overset{\text{chain rule}}{=} \langle \theta_t - \nu, \mathrm{d}\theta_t \rangle \overset{\text{gradient flow}}{=} \langle \theta_t - \nu, -\nabla L(\theta_t)\rangle \mathrm{d}t \overset{\text{convexity}}{\leq} L(\nu) - L(\theta_t)$$

step 2:

$$\frac{1}{2}\|\theta_t - \nu\|^2 - \frac{1}{2}\|\theta_0 - \nu\|^2 \overset{\text{integration}}{\leq} \int_0^t L(\nu) - L(\theta_s)\mathrm{d}s \leq t\big(L(\theta) - L(\theta_t)\big)$$

step 3: rearranging terms

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \le \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Proof.

step 1:

$$\mathrm{d}\frac{1}{2}\|\theta_t - \nu\|^2 \overset{\text{chain rule}}{=} \langle \theta_t - \nu, \mathrm{d}\theta_t \rangle \overset{\text{gradient flow}}{=} \langle \theta_t - \nu, -\nabla L(\theta_t) \rangle \mathrm{d}t \overset{\text{convexity}}{\le} L(\nu) - L(\theta_t)$$

step 2:

$$\frac{1}{2}\|\theta_t - \nu\|^2 - \frac{1}{2}\|\theta_0 - \nu\|^2 \overset{\text{integration}}{\le} \int_0^t L(\nu) - L(\theta_s)\mathrm{d}s \overset{\text{descent lemma}}{\le} t\big(L(\theta) - L(\theta_t)\big)$$

step 3: rearranging terms

Review: gradient flow analysis

For **convex** L and **gradient flow** $\mathrm{d}\theta_t = -\nabla L(\theta_t)\mathrm{d}t$, we have

$$L(\theta_t) - L(\nu) \leq \frac{\|\theta_0 - \nu\|^2}{2t} \quad \text{for all } \nu$$

Proof.

step 1:

$$\mathrm{d}\frac{1}{2}\|\theta_t - \nu\|^2 \overset{\text{chain rule}}{=} \langle \theta_t - \nu, \mathrm{d}\theta_t \rangle \overset{\text{gradient flow}}{=} \langle \theta_t - \nu, -\nabla L(\theta_t) \rangle \mathrm{d}t \overset{\text{convexity}}{\leq} L(\nu) - L(\theta_t)$$

step 2:

$$\frac{1}{2}\|\theta_t - \nu\|^2 - \frac{1}{2}\|\theta_0 - \nu\|^2 \overset{\text{integration}}{\leq} \int_0^t L(\nu) - L(\theta_s)\mathrm{d}s \overset{\text{descent lemma}}{\leq} t\big(L(\theta) - L(\theta_t)\big)$$

step 3: rearranging terms

for small stepsize, discretize this => GD analysis

Review: acceleration

number of steps to get ϵ-error

GD

$$\theta_+ = \theta - \eta \nabla L(\theta)$$

$$O\left(1/\epsilon\right) \text{ \& } O\left(\kappa \log(1/\epsilon)\right)$$

Review: acceleration

number of steps to get ϵ-error

GD

$$\theta_+ = \theta - \eta \nabla L(\theta)$$

$$O\big(1/\epsilon\big) \text{ \& } O\big(\kappa \log(1/\epsilon)\big)$$

Nesterov's momentum

$$\theta_+ = \nu - \eta \nabla L(\nu)$$

$$\nu_+ = \theta_+ + \beta(\theta_+ - \theta)$$

Review: acceleration

number of steps to get ϵ-error

GD

$$\theta_+ = \theta - \eta \nabla L(\theta)$$

$$O\big(1/\epsilon\big) \text{ \& } O\big(\kappa \log(1/\epsilon)\big)$$

Nesterov's momentum

$$\theta_+ = \nu - \eta \nabla L(\nu)$$

$$\nu_+ = \theta_+ + \beta(\theta_+ - \theta)$$

$$O\big(1/\sqrt{\epsilon}\big) \text{ \& } O\big(\sqrt{\kappa} \log(1/\epsilon)\big)$$

Review: acceleration

number of steps to get ϵ-error

GD

$$\theta_+ = \theta - \eta \nabla L(\theta)$$

$$O\big(1/\epsilon\big) \ \& \ O\big(\kappa \log(1/\epsilon)\big)$$

Nesterov's momentum

$$\theta_+ = \nu - \eta \nabla L(\nu)$$

$$\nu_+ = \theta_+ + \beta(\theta_+ - \theta)$$

$$O\big(1/\sqrt{\epsilon}\big) \ \& \ O\big(\sqrt{\kappa} \log(1/\epsilon)\big)$$

these rates are optimal

Review: acceleration

number of steps to get ϵ-error

GD

$$\theta_+ = \theta - \eta \nabla L(\theta)$$

$$O\big(1/\epsilon\big) \ \& \ O\big(\kappa \log(1/\epsilon)\big)$$

Nesterov's momentum

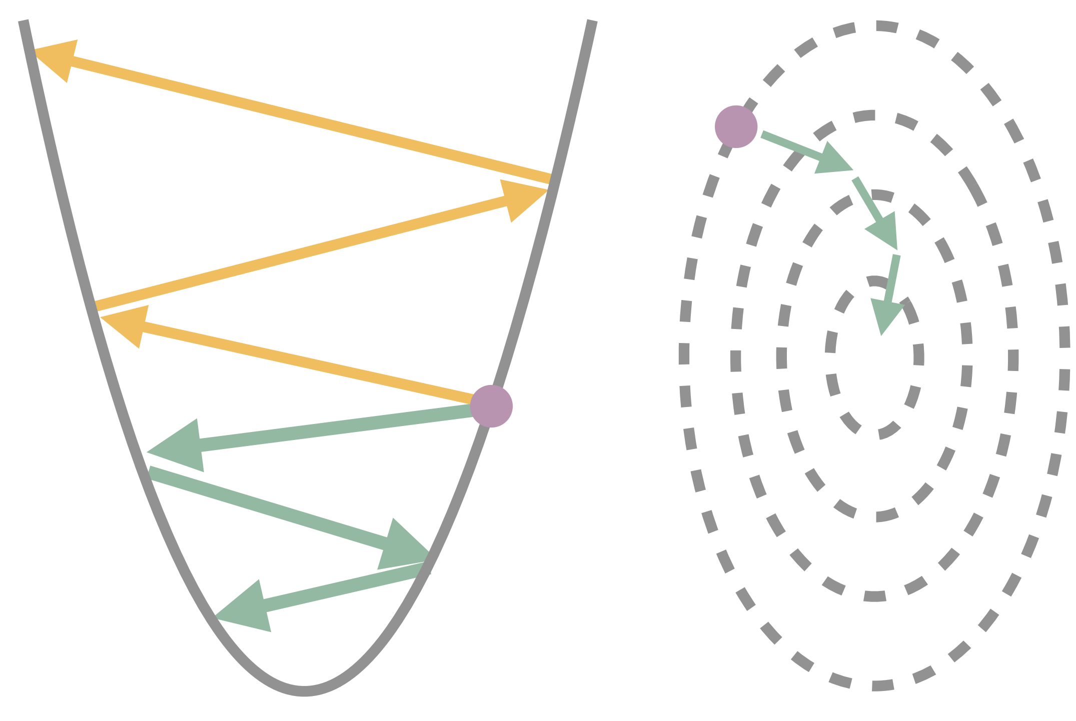
$$\theta_+ = \nu - \eta \nabla L(\nu)$$

$$\nu_+ = \theta_+ + \beta(\theta_+ - \theta)$$

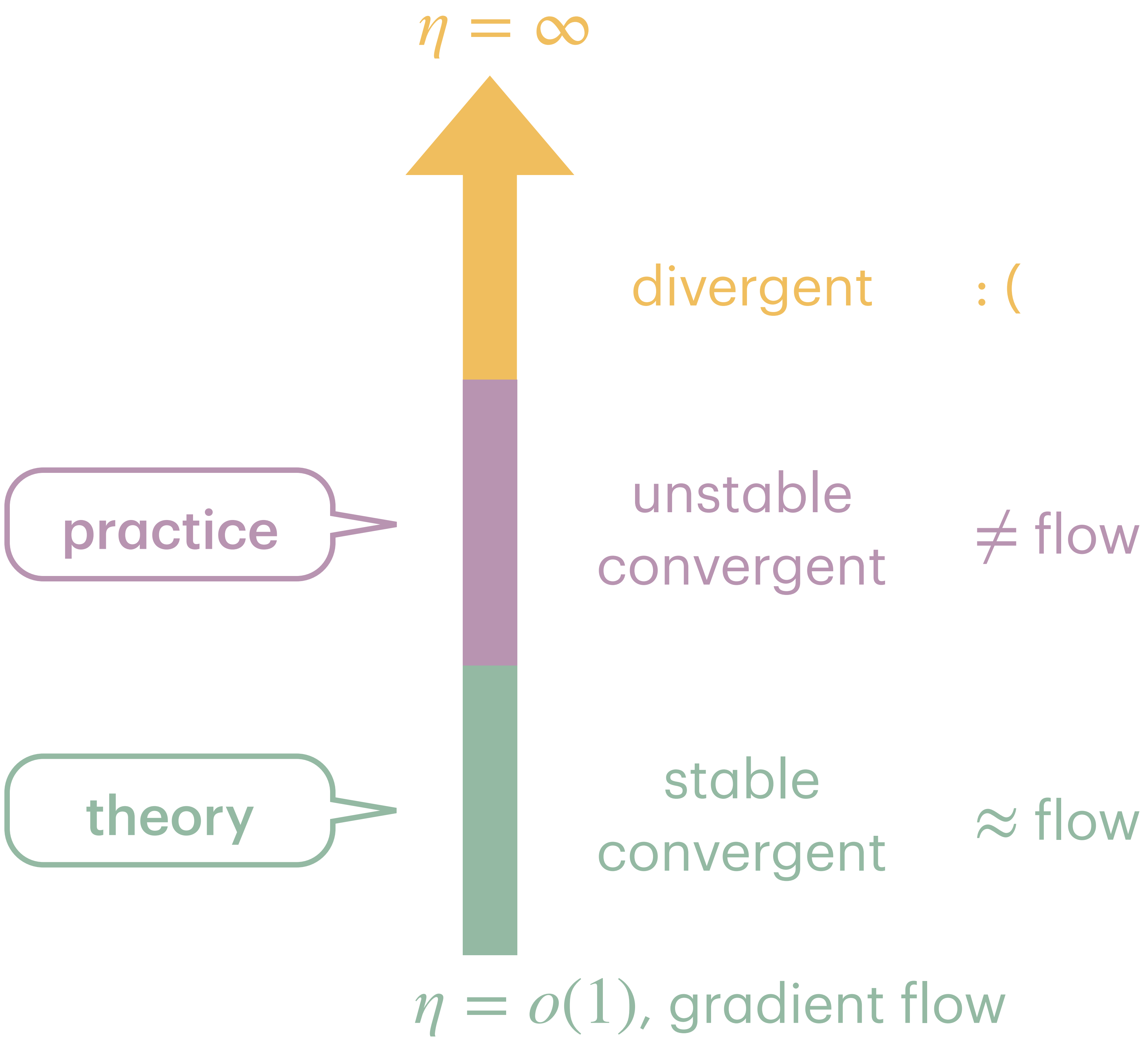
$$O\big(1/\sqrt{\epsilon}\big) \ \& \ O\big(\sqrt{\kappa} \log(1/\epsilon)\big)$$

these rates are optimal

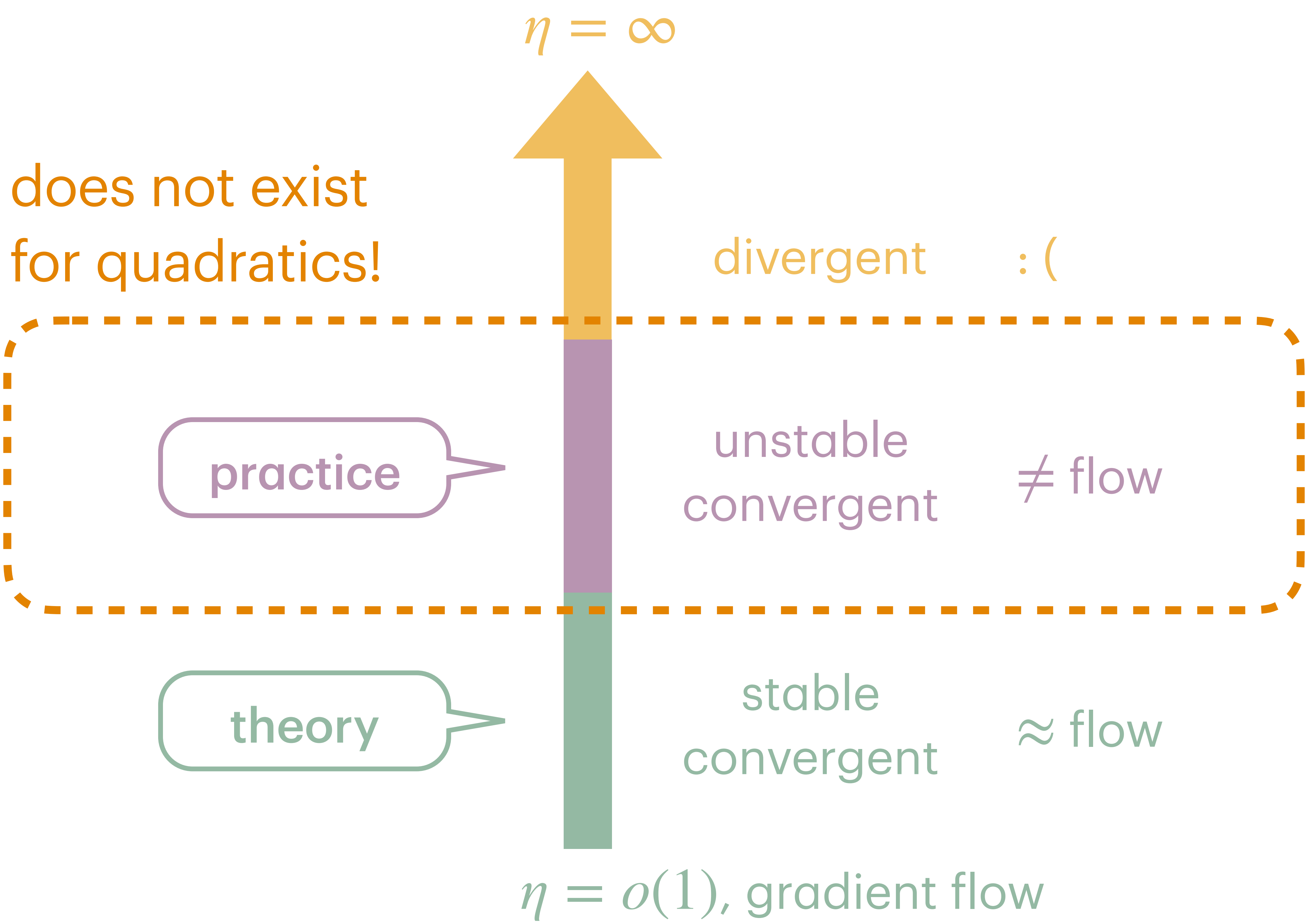
hard case: quadratics in high-dim



From small to large stepsize



From small to large stepsize

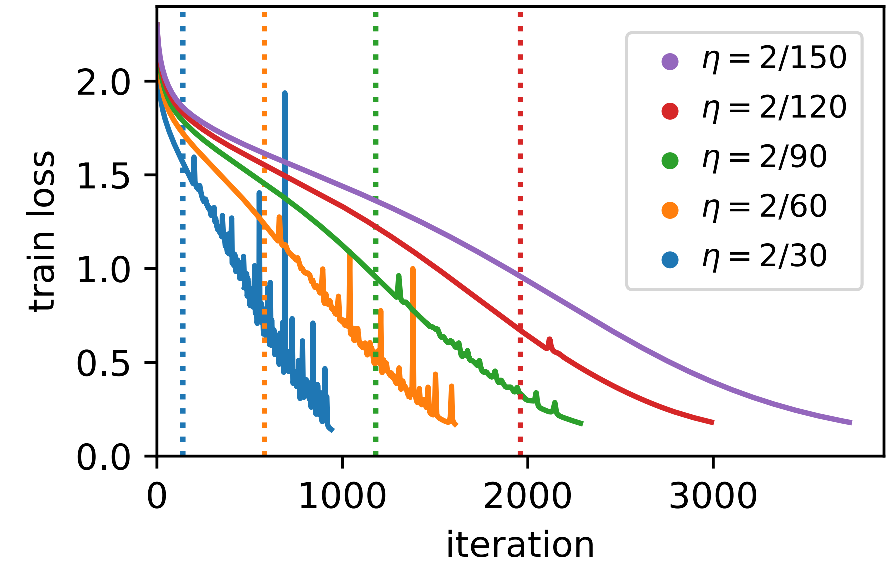


Alternative mental model

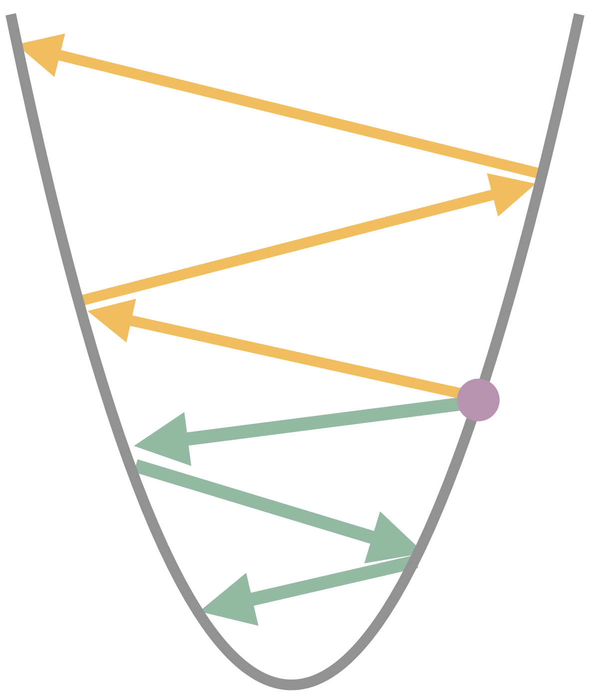
......

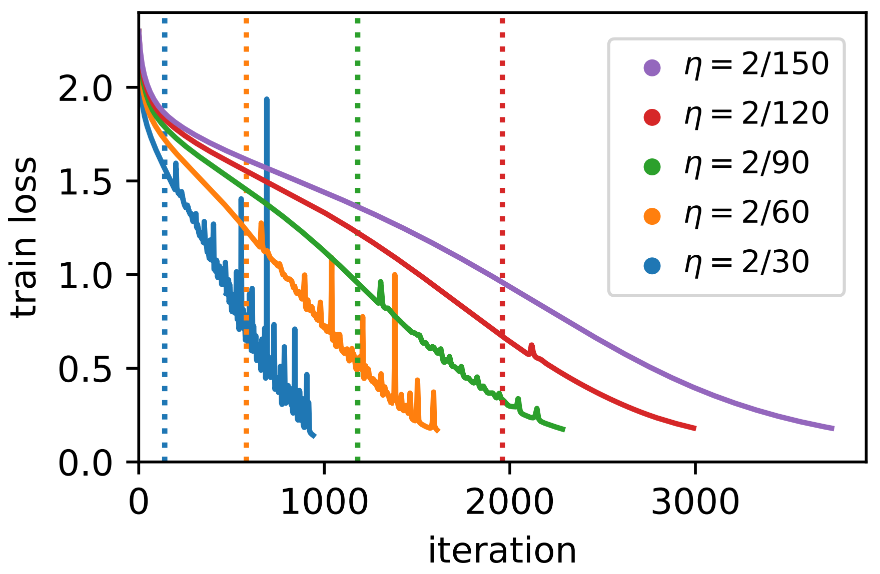
deep learning

unstable convergence observed



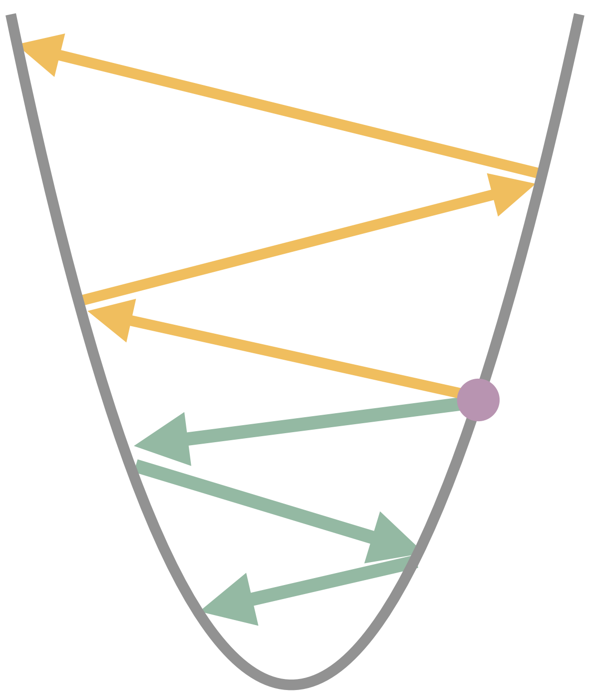
Alternative mental model

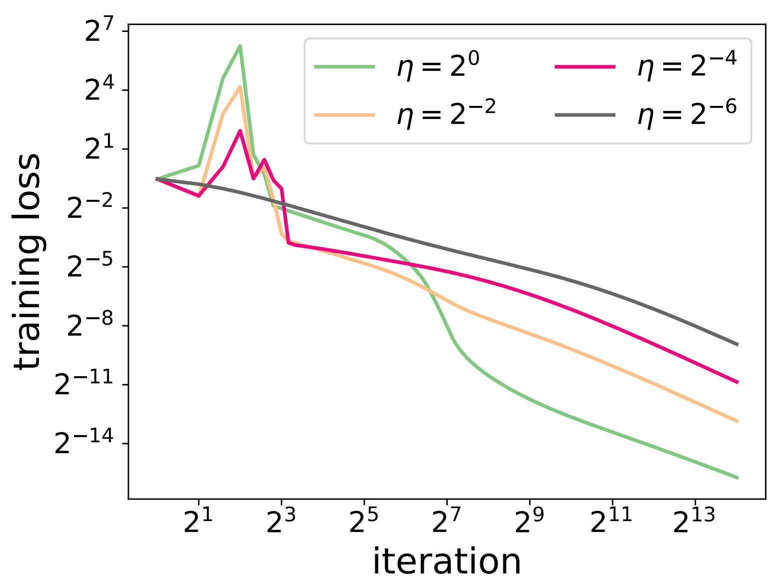


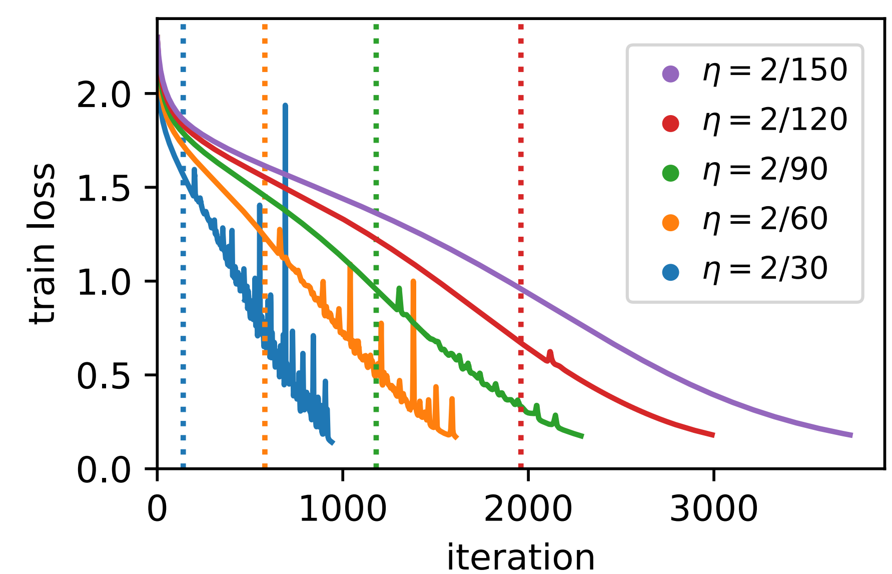


Alternative mental model

linear regression	logistic regression	......	deep learning
unstable convergence impossible	**observable & provable**		unstable convergence observed







(1/3) Logistic regression

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

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

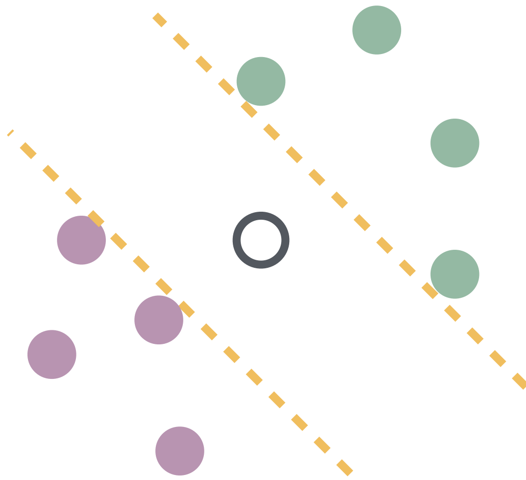
(1/3) Logistic regression

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

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

Assumption (bounded + separable)

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



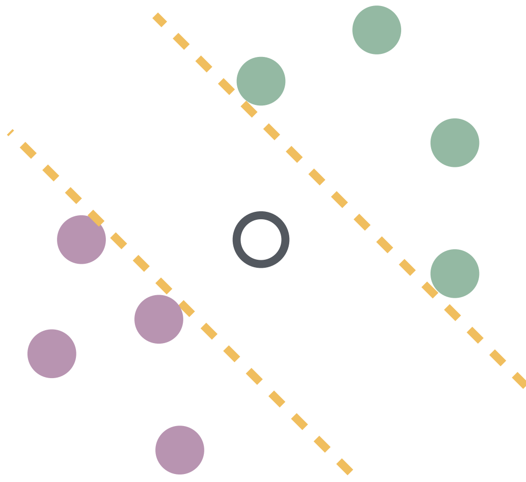
(1/3) Logistic regression

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

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

Assumption (bounded + separable)

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



"almost surely" if overparameterized

(1/3) Logistic regression

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

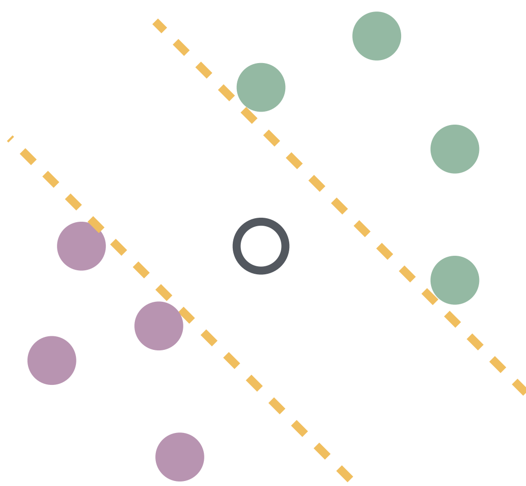
smooth, convex
non-strongly convex

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

Assumption (bounded + separable)

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

"almost surely" if overparameterized



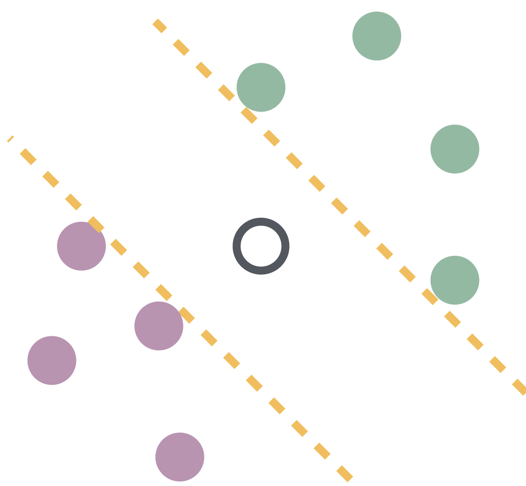
(1/3) Logistic regression

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

smooth, convex
non-strongly convex

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

Assumption (bounded + separable)



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

"almost surely" if overparameterized

Classical theory

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

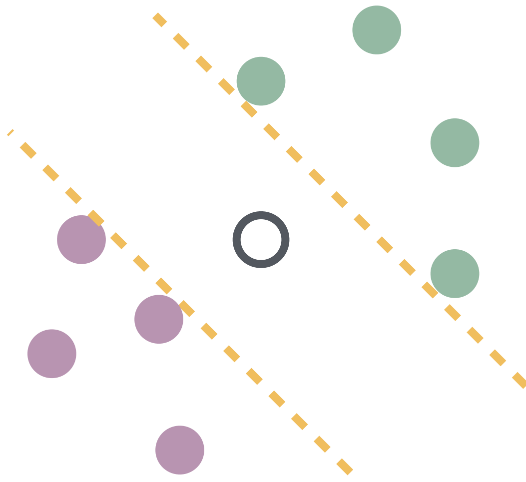
(1/3) Logistic regression

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

smooth, convex
non-strongly convex

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

Assumption (bounded + separable)



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

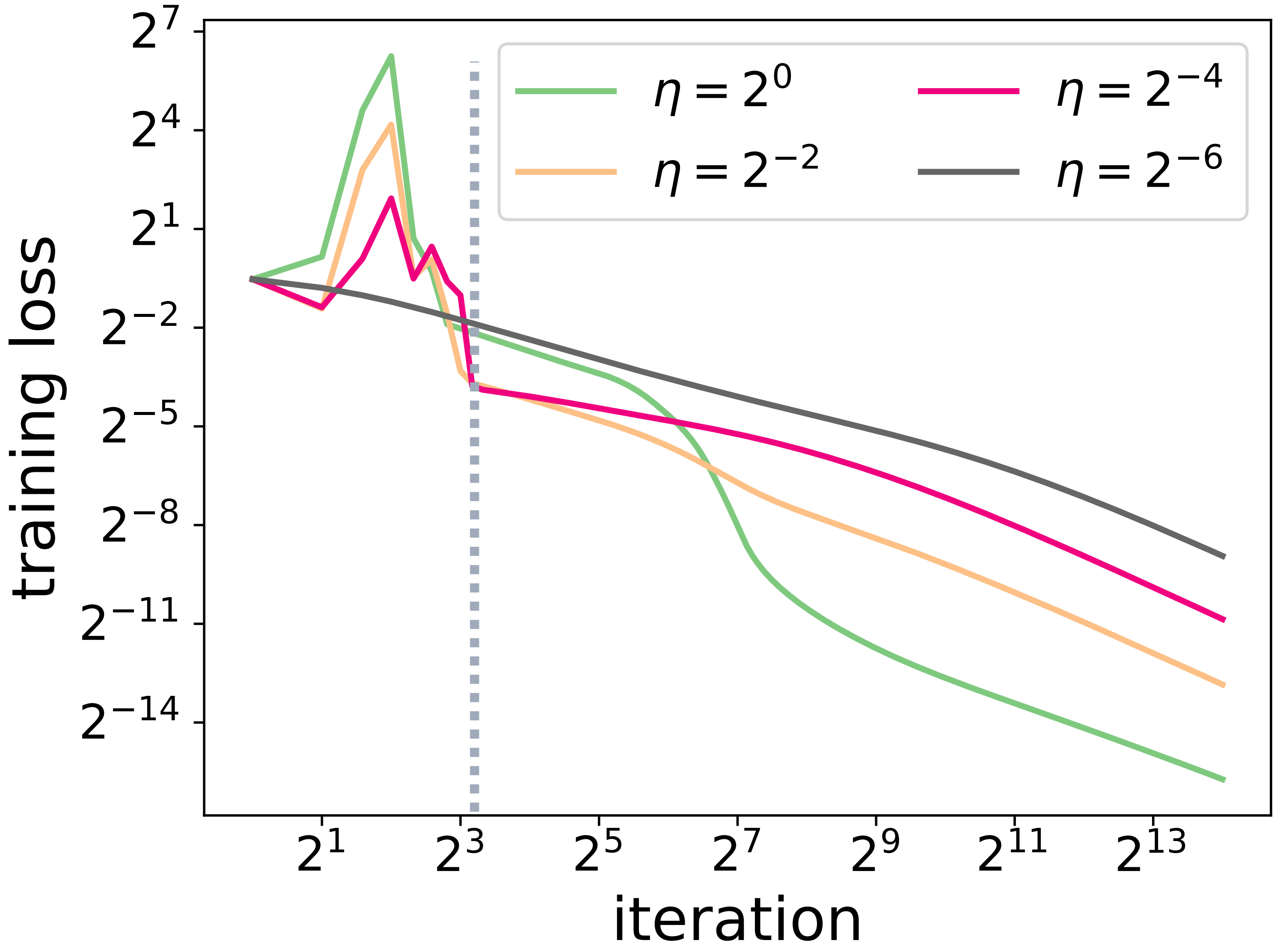
"almost surely" if overparameterized

Classical theory

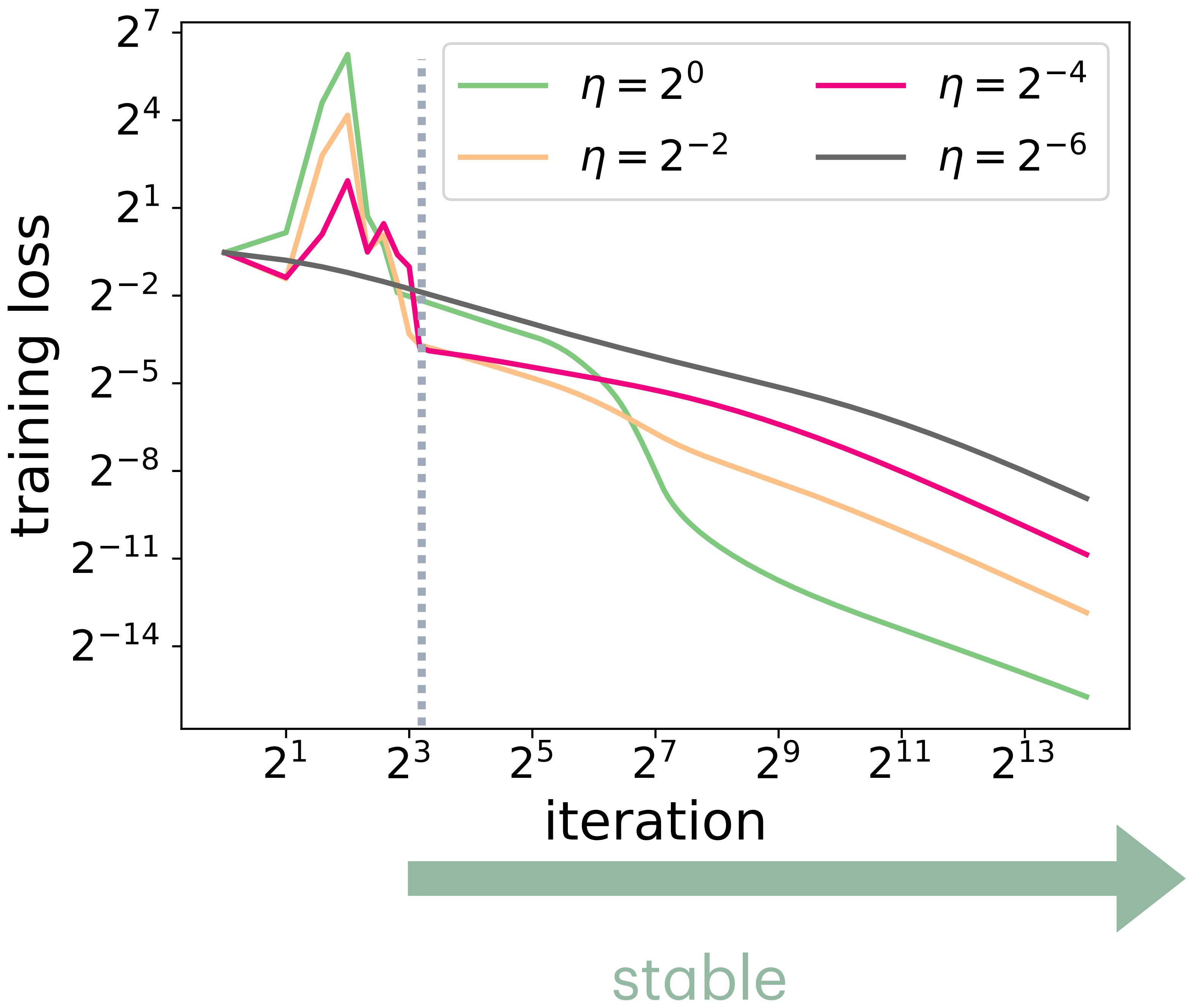
For $\eta = \Theta(1)$, $L(\theta_t) \downarrow$ and $L(\theta_t) = \tilde{O}(1/t)$

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

(1/3) Logistic regression

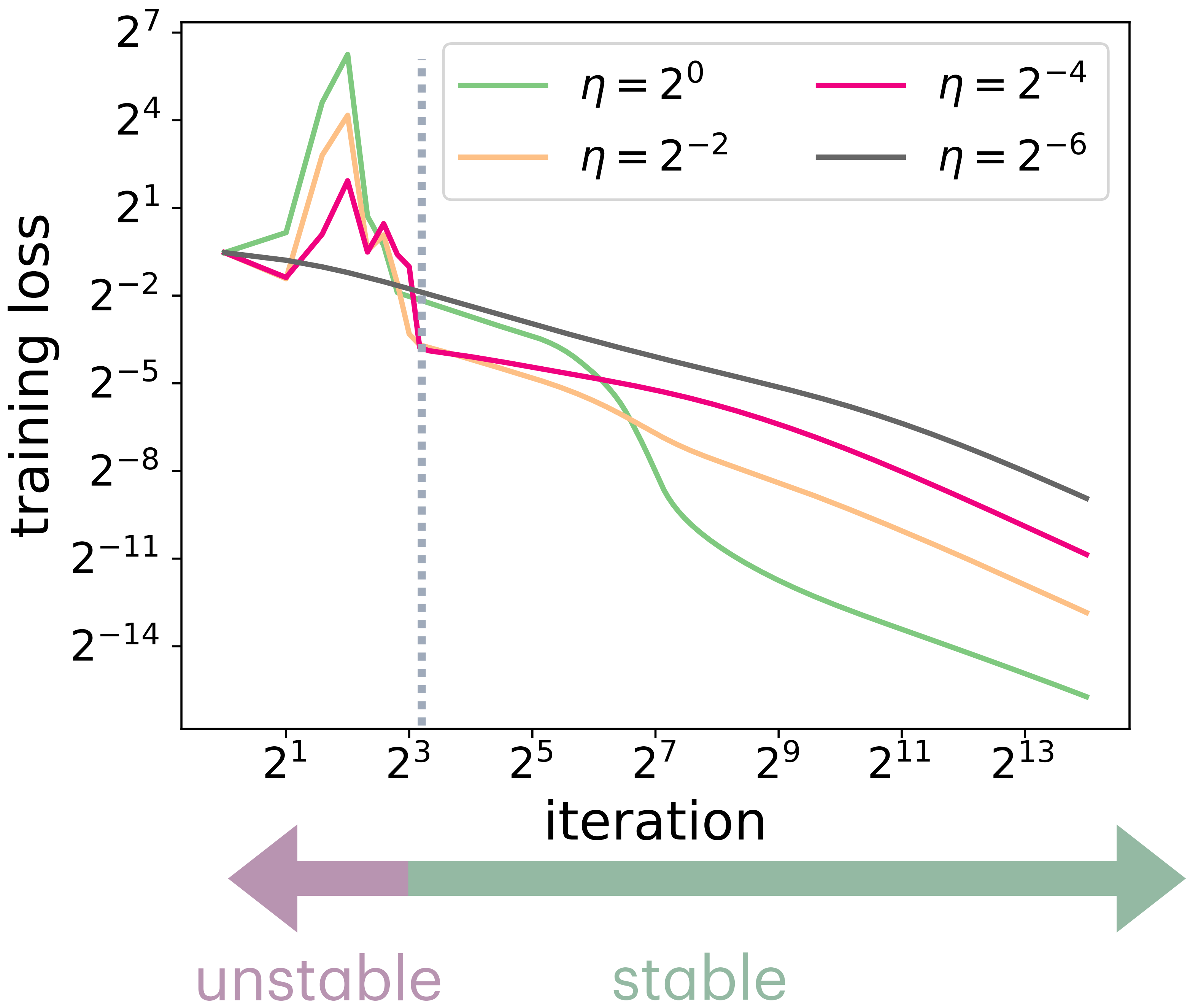


(1/3) Logistic regression



s-th step is in stable phase if $L(\theta_t) \downarrow$ for all $t \geq s$

(1/3) Logistic regression



s-th step is in stable phase if $L(\theta_t) \downarrow$ for all $t \geq s$

unstable phase if otherwise

(1/3) Theorem

Unstable phase.

Phase transition.

Stable phase.

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

(1/3) Theorem

Unstable phase.

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

Phase transition.

Stable phase.

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

(1/3) Theorem

Unstable phase.

tendency to decrease

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

Phase transition.

Stable phase.

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

(1/3) Theorem

Unstable phase.

tendency to decrease

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

Phase transition.

GD exits unstable phase in τ steps for

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

Stable phase.

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

(1/3) Theorem

Unstable phase.

tendency to decrease

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

Phase transition.

GD exits unstable phase in τ steps for

$\tau = \Theta(\eta)$

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

Stable phase.

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

(1/3) Theorem

Unstable phase.

tendency to decrease

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

Phase transition.

GD exits unstable phase in τ steps for

$\tau = \Theta(\eta)$

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

Stable phase.

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

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

(1/3) Theorem

Unstable phase.

tendency to decrease

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

Phase transition.

GD exits unstable phase in τ steps for

$\tau = \Theta(\eta)$

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

Stable phase.

"flow rate"

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

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

(1/3) Effects of large stepsize

1. Asymptotic $1/(\eta t)$ rate $\Rightarrow$ 2x stepsize 2x faster

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

(1/3) Effects of large stepsize

1. Asymptotic $1/(\eta t)$ rate $\Rightarrow$ 2x stepsize 2x faster

2. Phase transition in $\Theta(\eta)$ steps $\Rightarrow$ longer unstable phase

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

(1/3) Effects of large stepsize

1. Asymptotic $1/(\eta t)$ rate $\Rightarrow$ 2x stepsize 2x faster

2. Phase transition in $\Theta(\eta)$ steps $\Rightarrow$ longer unstable phase

3. Given #steps $T \geq \Theta(n)$, if choose $\eta = \Theta(T)$, then

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

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

(1/3) Effects of large stepsize

1. Asymptotic $1/(\eta t)$ rate $\Rightarrow$ 2x stepsize 2x faster

2. Phase transition in $\Theta(\eta)$ steps $\Rightarrow$ longer unstable phase

3. Given #steps $T \geq \Theta(n)$, if choose $\eta = \Theta(T)$, then

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

A lower bound. There exists a separable dataset, if η is such that $L(\theta_t) \downarrow$ for all t, then

$$L(\theta_t) = \Omega(1/t)$$

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

(1/3) Effects of large stepsize

1. Asymptotic $1/(\eta t)$ rate $\Rightarrow$ 2x stepsize 2x faster

2. Phase transition in $\Theta(\eta)$ steps $\Rightarrow$ longer unstable phase

3. Given #steps $T \geq \Theta(n)$, if choose $\eta = \Theta(T)$, then

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

A lower bound. There exists a separable dataset, if η is such that $L(\theta_t) \downarrow$ for all t, then

$$L(\theta_t) = \Omega(1/t)$$

acceleration by large stepsize

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

(1/3) A “non-quadratic” picture

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

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

(1/3) A “non-quadratic” picture

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

$L(\theta) = \hat{\mathbb{E}} \ln\big(1 + \exp(-yx^\top\theta)\big)$

***minimizer at* ∞**

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

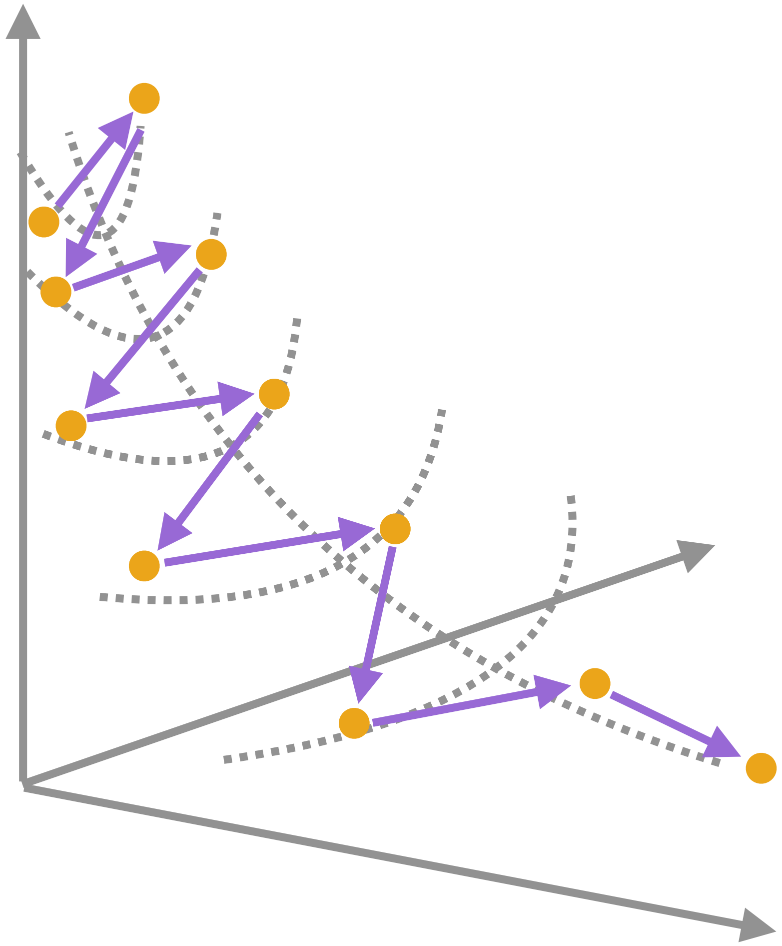
(1/3) A "non-quadratic" picture

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

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

minimizer at ∞

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



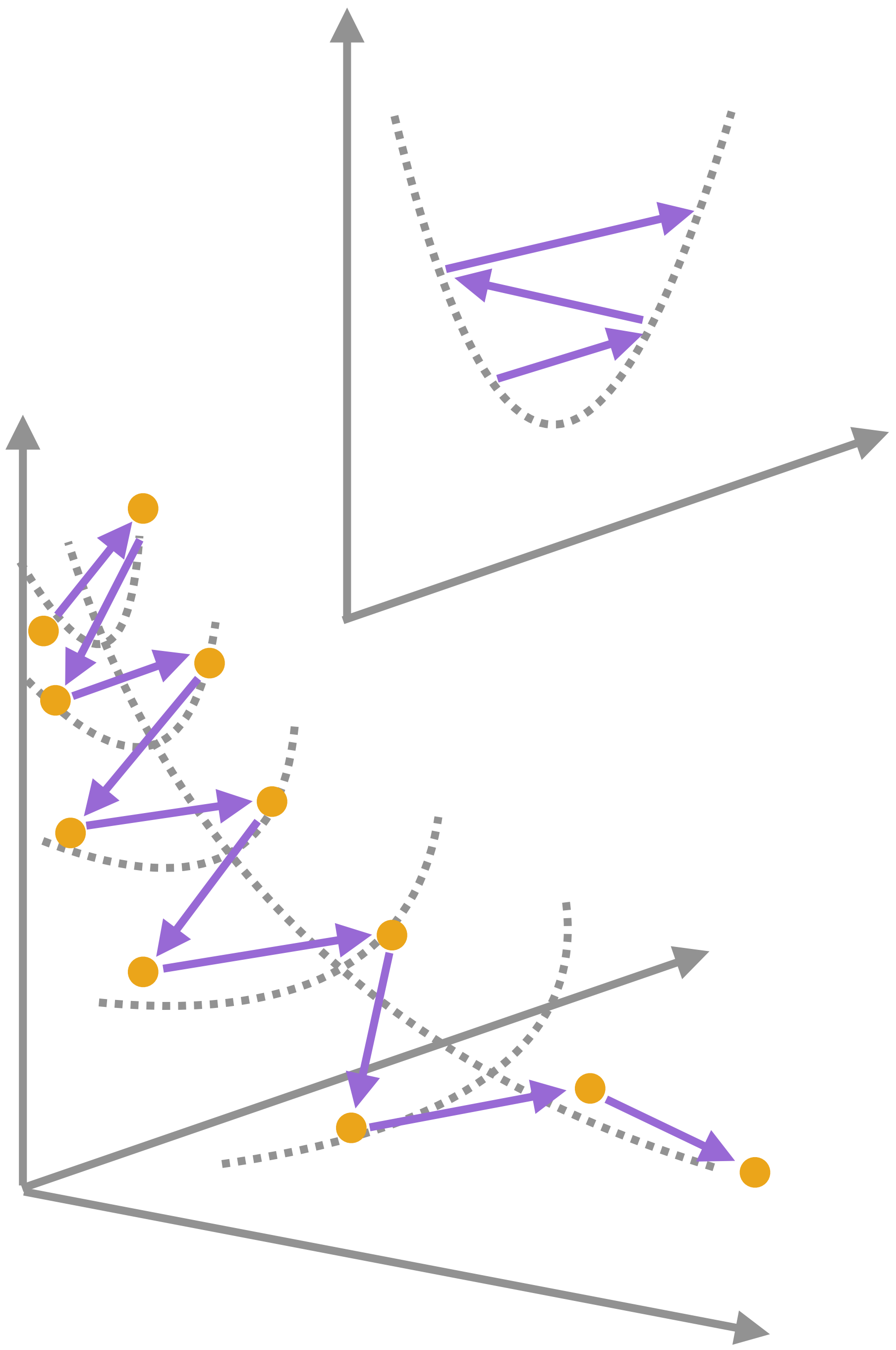
(1/3) A "non-quadratic" picture

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

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

minimizer at ∞

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



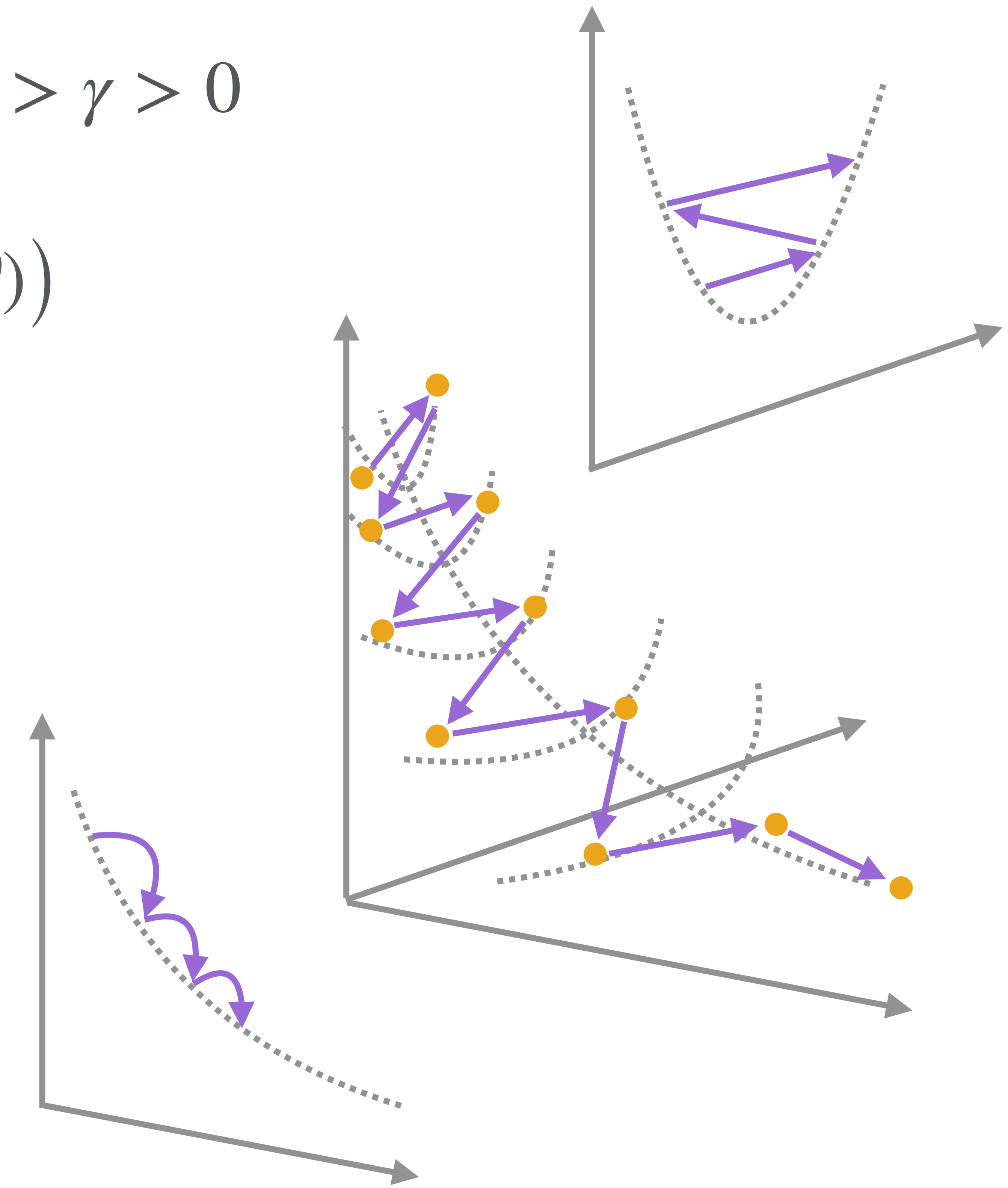
(1/3) A “non-quadratic” picture

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

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

minimizer at ∞

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



(1/3) A "non-quadratic" picture

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

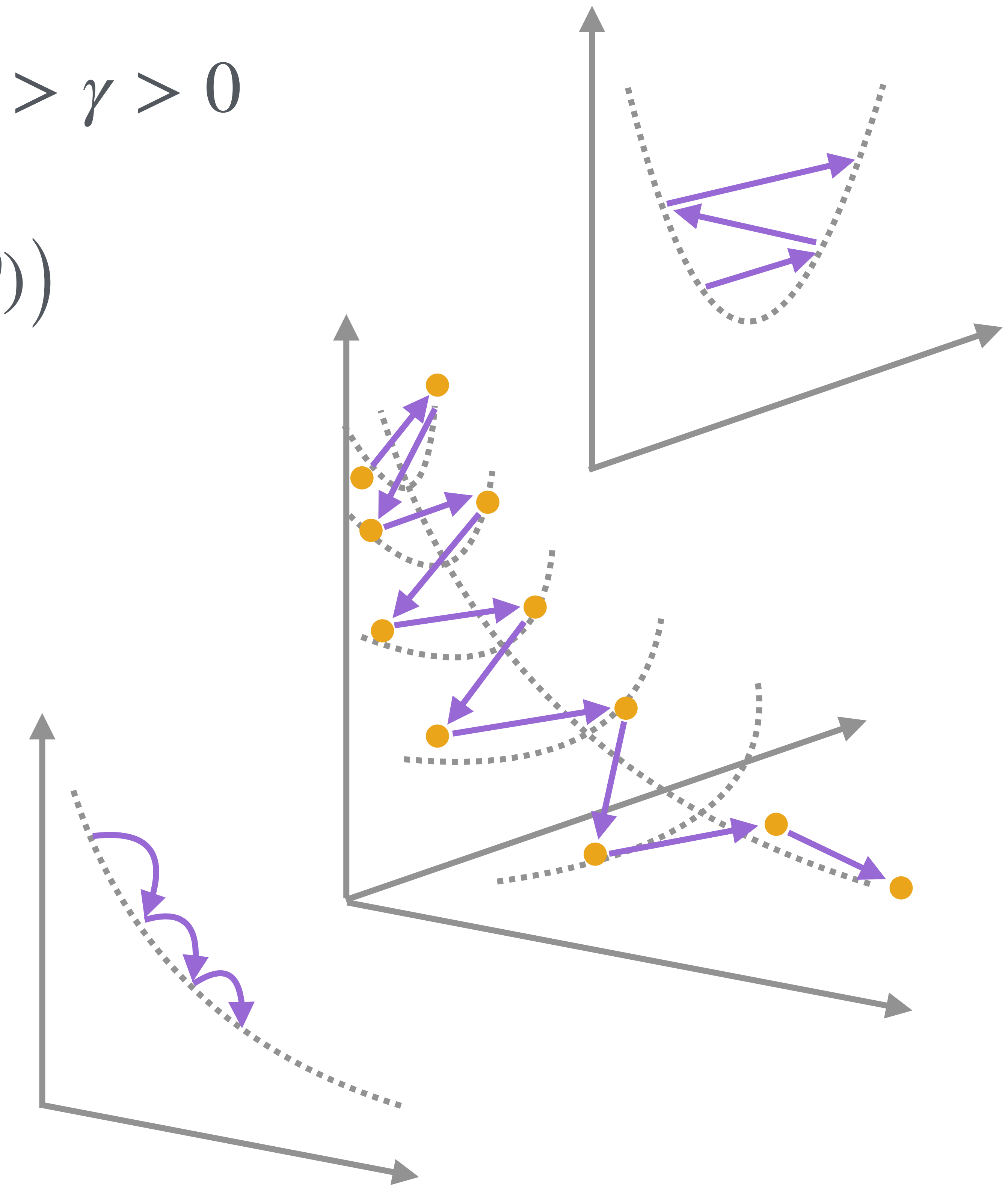
$L(\theta) = \hat{\mathbb{E}} \ln\left(1 + \exp(-y x^\top \theta)\right)$

minimizer at ∞

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

self-bounded

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



Two extensions

minimizer at ∞

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

self-bounded

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

Two extensions

minimizer at ∞

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

finite minimizer

e.g. regularization

self-bounded

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

Two extensions

minimizer at ∞

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

finite minimizer

e.g. regularization

self-bounded

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

enabling "tricks"

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

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

Two extensions

minimizer at ∞

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

finite minimizer

e.g. regularization

unstable convergence under finite minimizer

self-bounded

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

enabling "tricks"

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

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

Two extensions

minimizer at ∞

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

finite minimizer

e.g. regularization

unstable convergence under finite minimizer

self-bounded

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

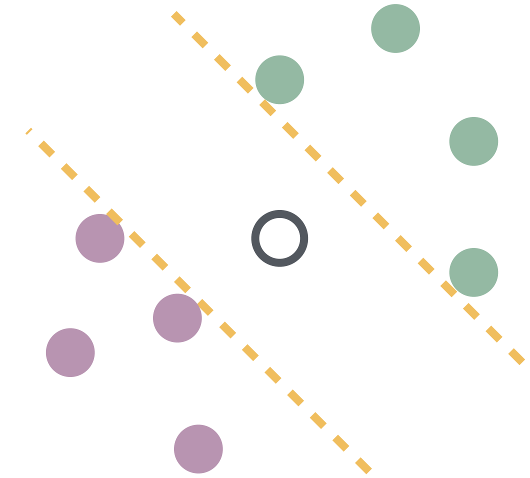
enabling "tricks"

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

large stepsizes for GD variants

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

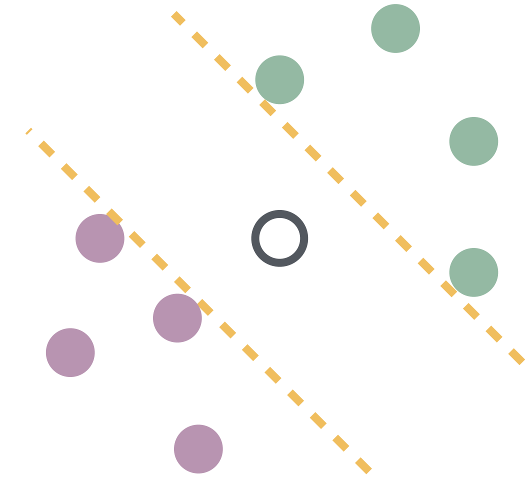
(2/3) Large, adaptive stepsize



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

$$\begin{aligned}\theta_{t+1} &= \theta_t - \eta\big((-\ell^{-1})' \circ L(\theta_t)\big)\,\nabla L(\theta_t) \\ &\approx \theta_t - \frac{\eta}{L(\theta_t)}\nabla L(\theta_t)\end{aligned}$$

(2/3) Large, adaptive stepsize



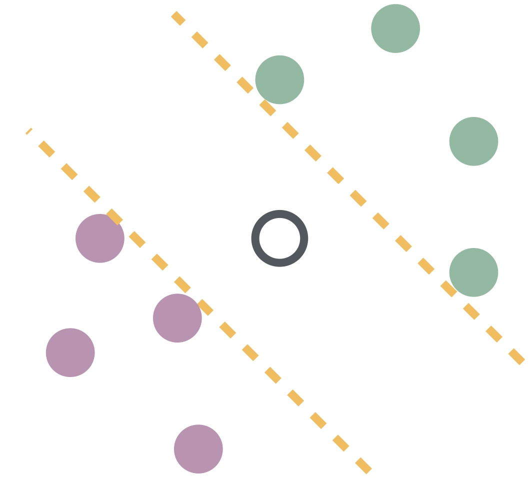
$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i x_i^\top \theta) \qquad \ell(t) = \ln\big(1 + \exp(-t)\big)$$

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

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

adapt to curvature

(2/3) Large, adaptive stepsize



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

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

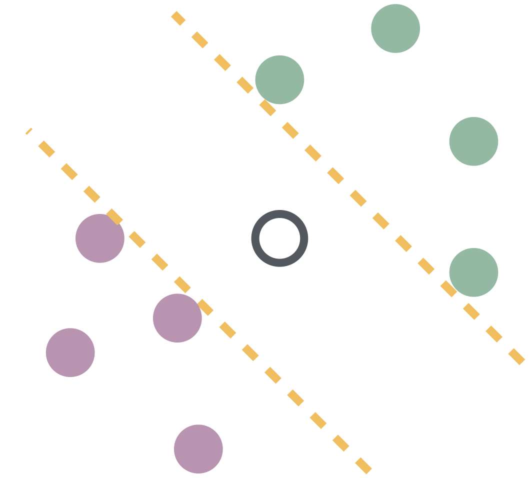
adapt to curvature

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

$$\theta_{t+1} = \theta_t - \eta \nabla \phi(\theta_t) \qquad \phi(\theta) = -\,\ell^{-1}(L(\theta))$$

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

(2/3) Large, adaptive stepsize



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

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

adapt to curvature

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

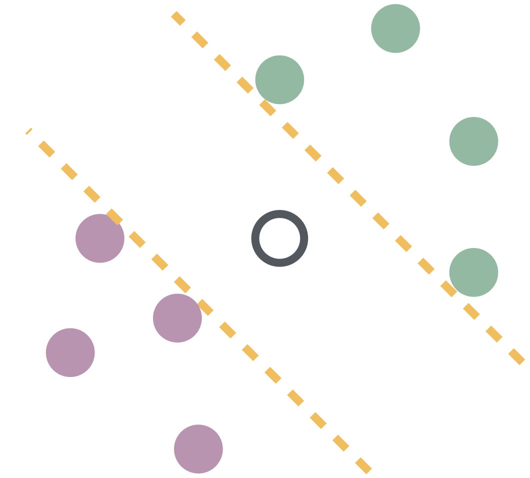
$$\theta_{t+1} = \theta_t - \eta\,\nabla\phi(\theta_t) \qquad \phi(\theta) = -\,\ell^{-1}(L(\theta))$$

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

[Ji & Telgarsky, 2021]

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

(2/3) Large, adaptive stepsize



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

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

adapt to curvature

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

$$\theta_{t+1} = \theta_t - \eta\,\nabla\phi(\theta_t) \qquad \phi(\theta) = -\,\ell^{-1}(L(\theta))$$

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

[Ji & Telgarsky, 2021]

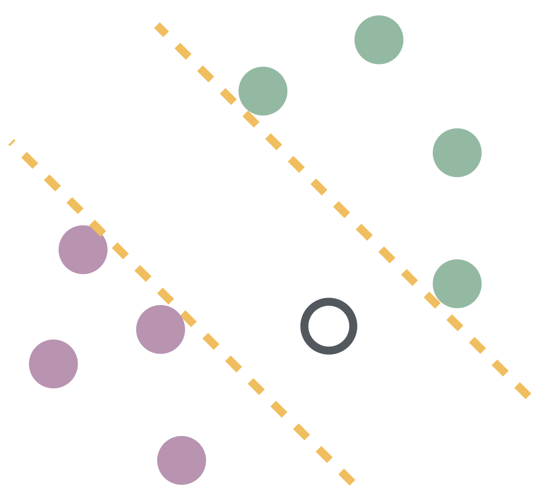
For $\eta = \Theta(1)$, $L(\theta_t) \downarrow$ and $L(\theta_t) \le \exp(-\Theta(t))$

large stepsize makes adaptive GD even faster

(2/3) Theorem

Assume separability with margin γ. For $t \geq 1/\gamma^2$ and every η

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

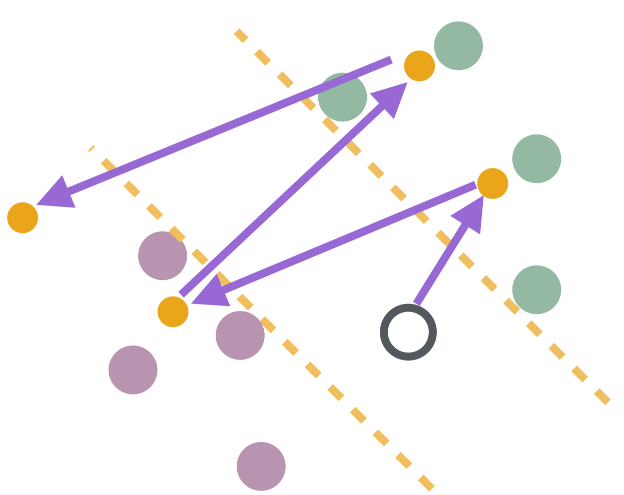


Zhang, **Wu**, Lin, Bartlett. "Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes." ICML 2025

(2/3) Theorem

Assume separability with margin γ. For $t \geq 1/\gamma^2$ and every η

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

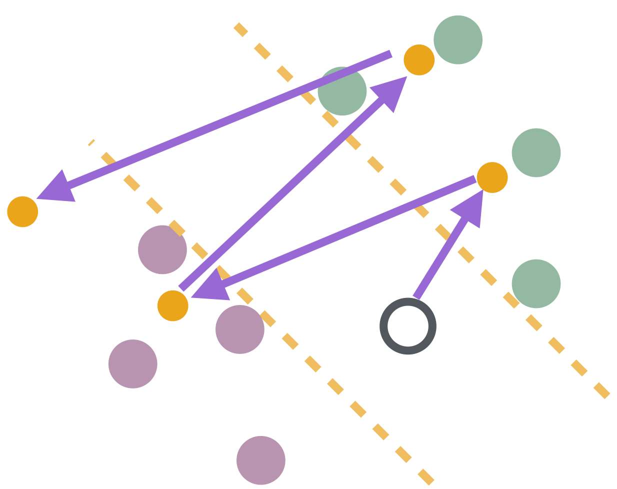


Zhang, **Wu**, Lin, Bartlett. "Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes." ICML 2025

(2/3) Theorem

Assume separability with margin γ. For $t \geq 1/\gamma^2$ and every η

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



arbitrarily small error in $1/\gamma^2$ steps

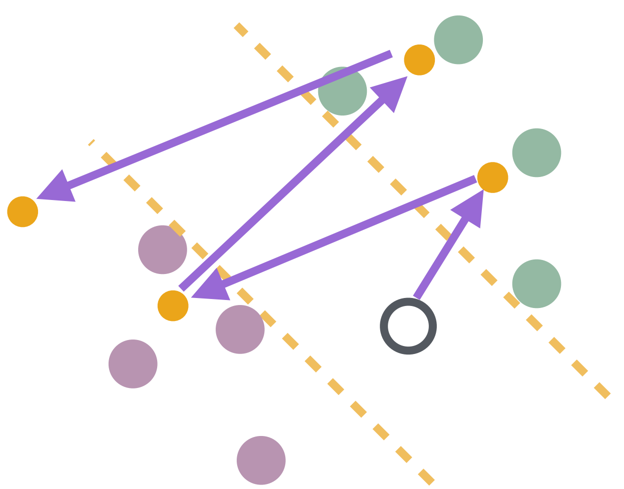
$$\lim_{\eta \to \infty} L(\bar{\theta}_t) = 0 \quad \text{for} \quad t = 1/\gamma^2$$

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

(2/3) Theorem

Assume separability with margin γ. For $t \geq 1/\gamma^2$ and every η

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



arbitrarily small error in $1/\gamma^2$ steps

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

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

(2/3) Theorem (lower bound)

$\forall \theta_0$, $\exists\ (x_i, y_i)_{i=1}^n$ 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

(2/3) Theorem (lower bound)

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

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

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

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

(2/3) Theorem (lower bound)

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

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

first-order batch method:

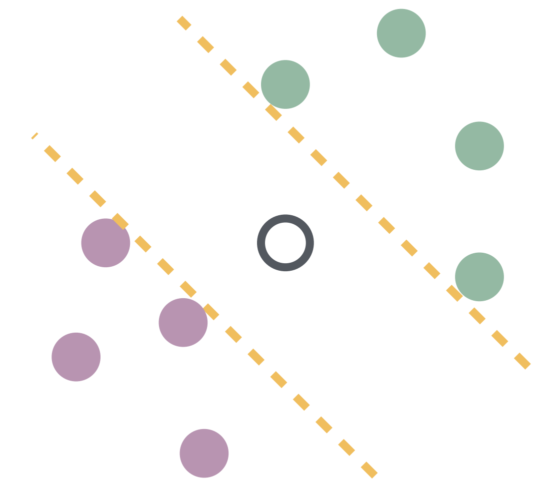
$$\theta_t \in \theta_0 + \mathsf{span}\{\nabla L(\theta_0), \ldots, \nabla L(\theta_{t-1})\}$$

where $L(\theta) = \hat{\mathbb{E}} \ell(y x^\top \theta)$ for any ℓ

large, adaptive stepsizes = minimax optimal

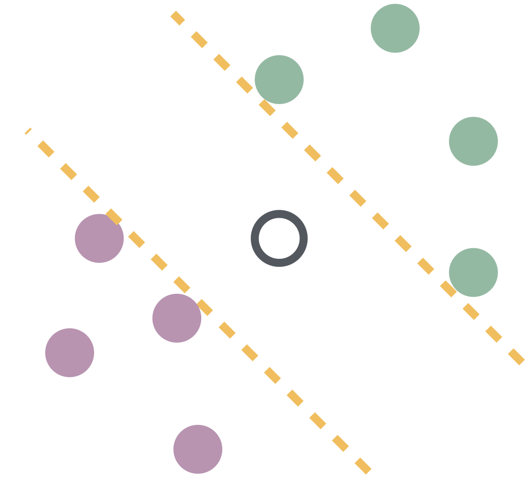
Zhang, **Wu**, Lin, Bartlett. "Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes." ICML 2025

(3/3) ℓ_2-regularization



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

(3/3) ℓ_2-regularization

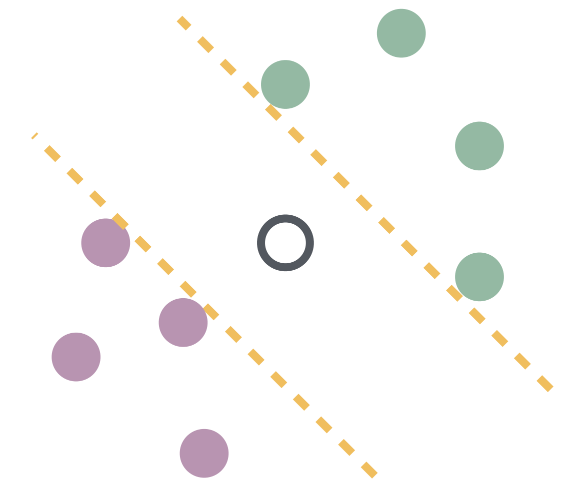


$\Theta(1)$-smooth, λ-strongly convex

condition number $\kappa = \Theta(1/\lambda)$

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

(3/3) ℓ_2-regularization



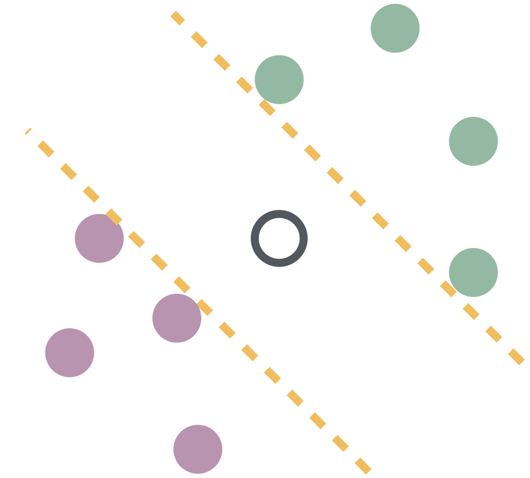
$\Theta(1)$-smooth, λ-strongly convex

condition number $\kappa = \Theta(1/\lambda)$

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

finite minimizer $\tilde{\theta}$ with norm $||\tilde{\theta}|| = O(\ln \kappa)$

(3/3) ℓ_2-regularization



$\Theta(1)$-smooth, λ-strongly convex

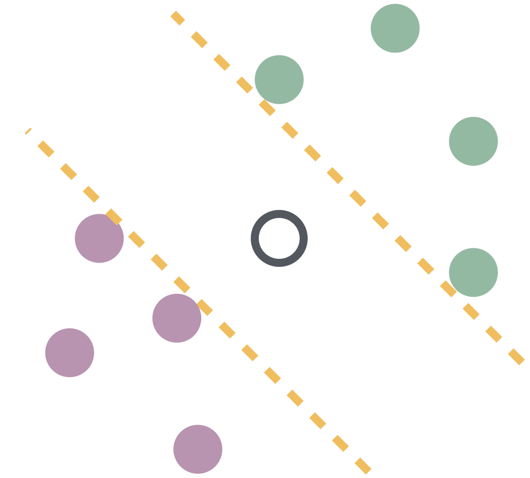
condition number $\kappa = \Theta(1/\lambda)$

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

finite minimizer $\tilde{\theta}$ with norm $\|\tilde{\theta}\| = O(\ln \kappa)$

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

(3/3) ℓ_2-regularization



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

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

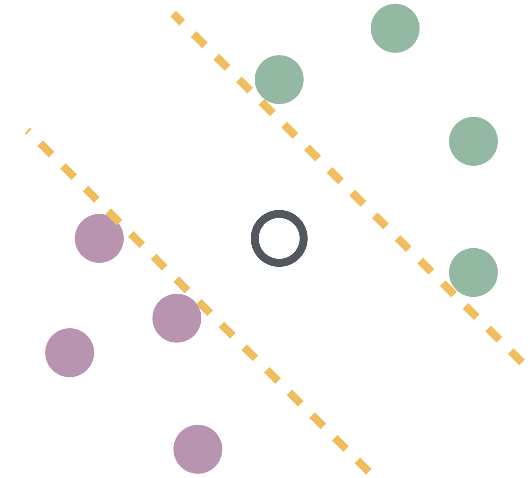
finite minimizer $\tilde{\theta}$ with norm $\|\tilde{\theta}\| = O(\ln \kappa)$

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

Classical theory

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

(3/3) ℓ_2-regularization



$\Theta(1)$-smooth, λ-strongly convex

condition number $\kappa = \Theta(1/\lambda)$

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

finite minimizer $\tilde{\theta}$ with norm $\|\tilde{\theta}\| = O(\ln \kappa)$

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

Classical theory

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

improved to $\tilde{O}(\sqrt{\kappa})$ by Nesterov

(3/3) Theorem

Let $\kappa = 1/\lambda$. Assume separability and

$$\lambda \leq \Theta\left(\frac{1}{n \ln n}\right) \qquad \eta \leq \Theta\big(\min\{\sqrt{\kappa}, \kappa/n\}\big)$$

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

(3/3) Theorem

Let $\kappa = 1/\lambda$. Assume separability and

$\eta_{\max} = \Theta(\sqrt{\kappa})$

$$\lambda \leq \Theta\left(\frac{1}{n \ln n}\right) \qquad \eta \leq \Theta\big(\min\{\sqrt{\kappa}, \kappa/n\}\big)$$

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

(3/3) Theorem

Let $\kappa = 1/\lambda$. Assume separability and

$\eta_{\max} = \Theta(\sqrt{\kappa})$

$$\lambda \leq \Theta\left(\frac{1}{n \ln n}\right) \qquad \eta \leq \Theta\big(\min\{\sqrt{\kappa}, \kappa/n\}\big)$$

Phase transition. GD exits unstable phase in τ steps for

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

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

(3/3) Theorem

Let $\kappa = 1/\lambda$. Assume separability and

$\eta_{\max} = \Theta(\sqrt{\kappa})$

$$\lambda \leq \Theta\left(\frac{1}{n \ln n}\right) \qquad \eta \leq \Theta\big(\min\{\sqrt{\kappa}, \kappa/n\}\big)$$

Phase transition. GD exits unstable phase in τ steps for

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

$\tau = \Theta(\sqrt{\kappa})$

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

(3/3) Theorem

Let $\kappa = 1/\lambda$. Assume separability and

$\eta_{\max} = \Theta(\sqrt{\kappa})$

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

Phase transition. GD exits unstable phase in τ steps for

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

$\tau = \Theta(\sqrt{\kappa})$

Stable phase. $\tilde{L}(\theta_{\tau+t}) \downarrow$ and

$$\tilde{L}(\theta_{\tau+t}) - \min \tilde{L} \lesssim \exp(-t\eta/\kappa)$$

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

(3/3) Theorem

Let $\kappa = 1/\lambda$. Assume separability and

$\eta_{\max} = \Theta(\sqrt{\kappa})$

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

Phase transition. GD exits unstable phase in τ steps for

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

$\tau = \Theta(\sqrt{\kappa})$

Stable phase. $\tilde{L}(\theta_{\tau+t}) \downarrow$ and

$t = \Theta(\sqrt{\kappa} \ln(1/\epsilon))$

$$\tilde{L}(\theta_{\tau+t}) - \min \tilde{L} \lesssim \exp(-t\eta/\kappa)$$

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

(3/3) Theorem

Let $\kappa = 1/\lambda$. Assume separability and

$\eta_{\max} = \Theta(\sqrt{\kappa})$

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

Phase transition. GD exits unstable phase in τ steps for

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

$\tau = \Theta(\sqrt{\kappa})$

Stable phase. $\tilde{L}(\theta_{\tau+t}) \downarrow$ and

$t = \Theta(\sqrt{\kappa} \ln(1/\epsilon))$

$$\tilde{L}(\theta_{\tau+t}) - \min \tilde{L} \lesssim \exp(-t\eta/\kappa)$$

from $\tilde{O}(\kappa)$ to $\tilde{O}(\sqrt{\kappa})$: acceleration via large stepsize

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

(3/3) Theorem

for $\lambda \le \Theta(1)$, improvement is $\tilde{O}(\kappa^{2/3})$

Let $\kappa = 1/\lambda$. Assume separability and

$\eta_{\max} = \Theta(\sqrt{\kappa})$

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

Phase transition. GD exits unstable phase in τ steps for

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

$\tau = \Theta(\sqrt{\kappa})$

Stable phase. $\tilde{L}(\theta_{\tau+t}) \downarrow$ and

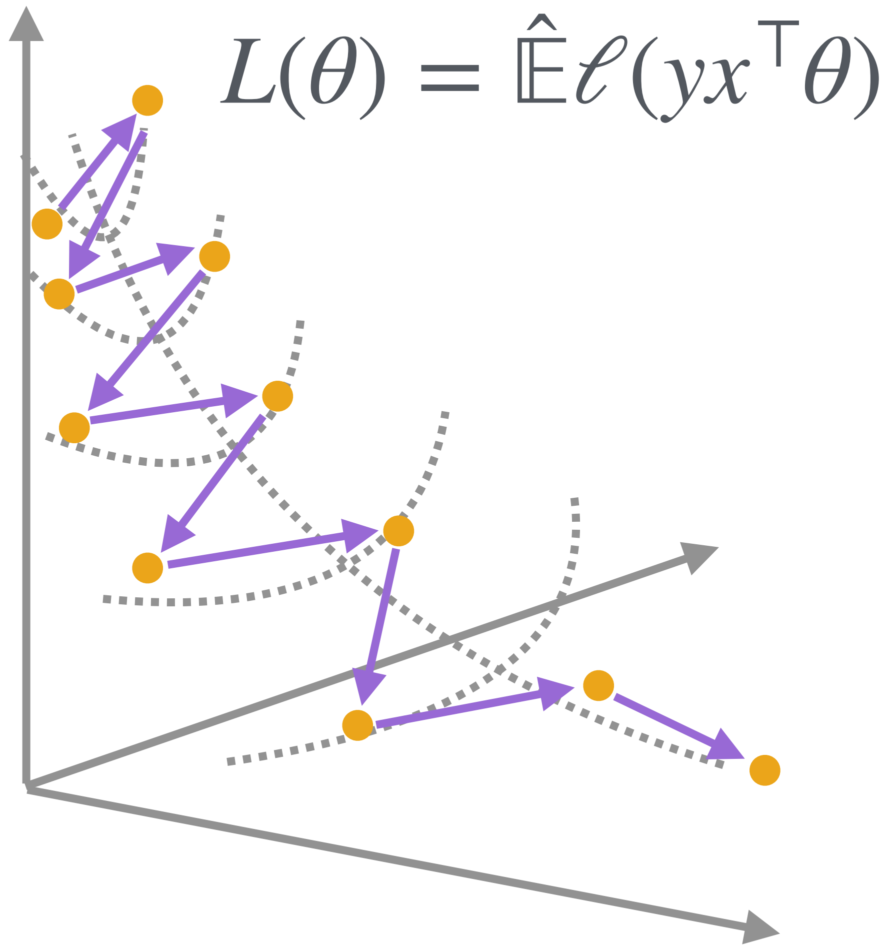
$t = \Theta(\sqrt{\kappa} \ln(1/\epsilon))$

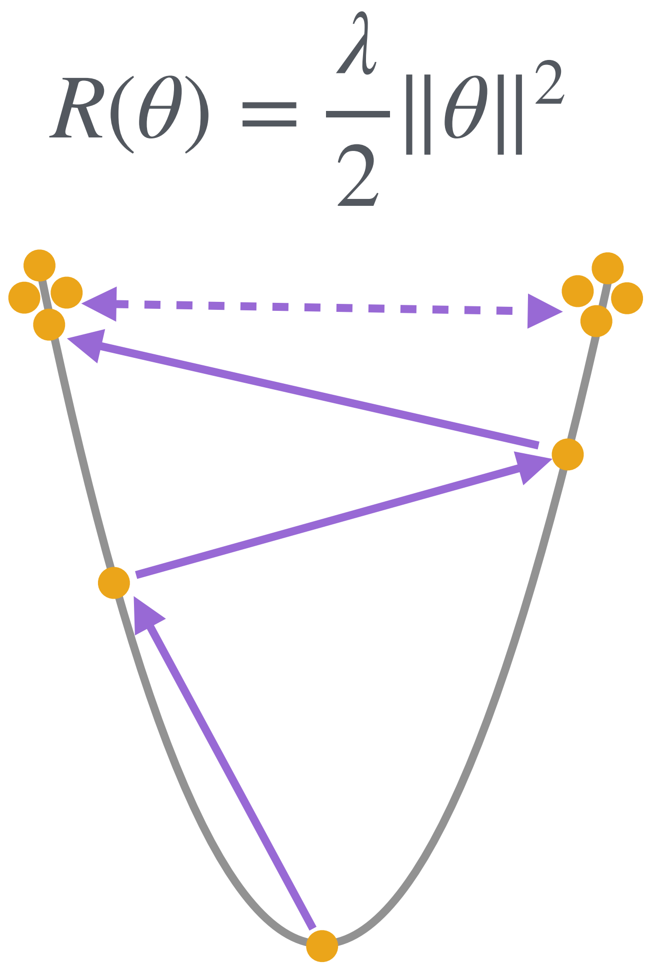
$$\tilde{L}(\theta_{\tau+t}) - \min \tilde{L} \lesssim \exp(-t\eta/\kappa)$$

from $\tilde{O}(\kappa)$ to $\tilde{O}(\sqrt{\kappa})$: acceleration via large stepsize

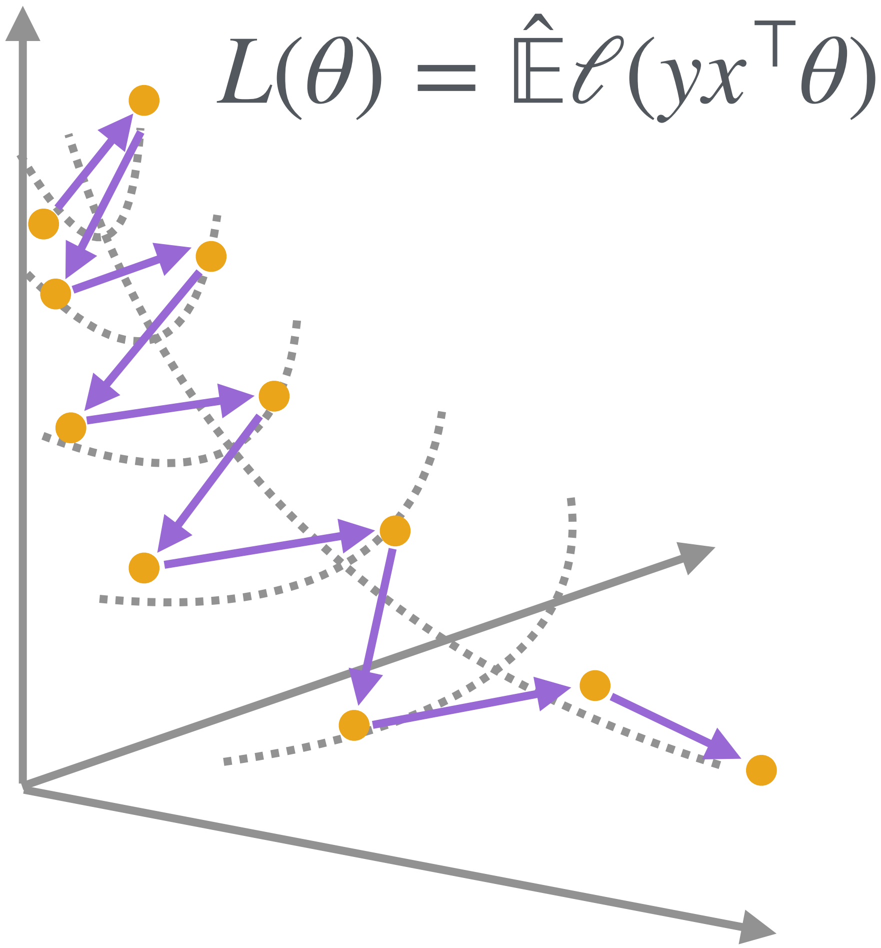
Wu, Marion, Bartlett. "Large stepsizes accelerate gradient descent for regularized logistic regression." NeurIPS 2025

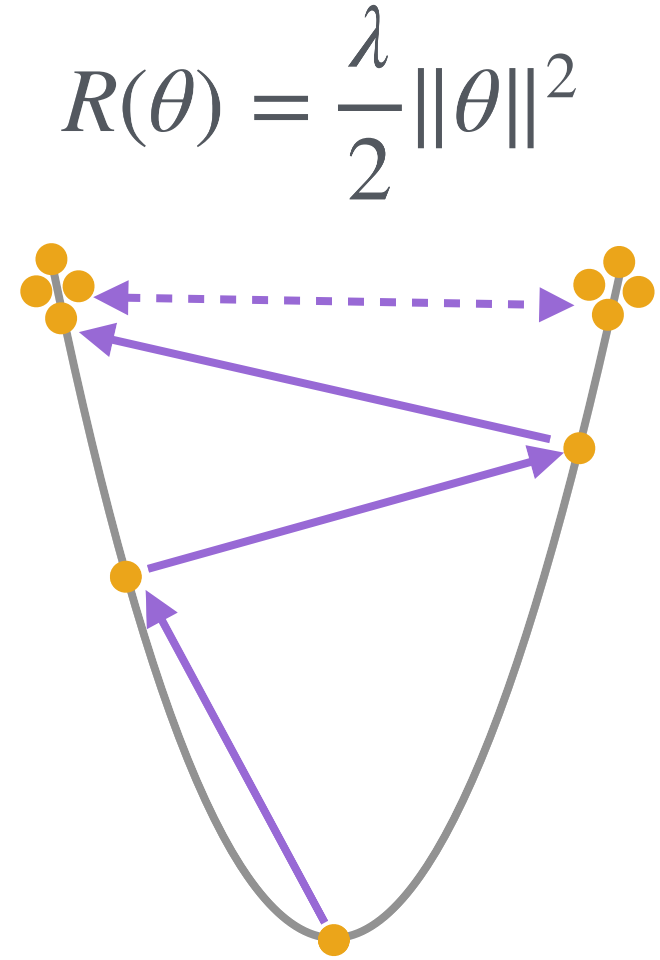
(3/3) Picture: valley + basin





(3/3) Picture: valley + basin

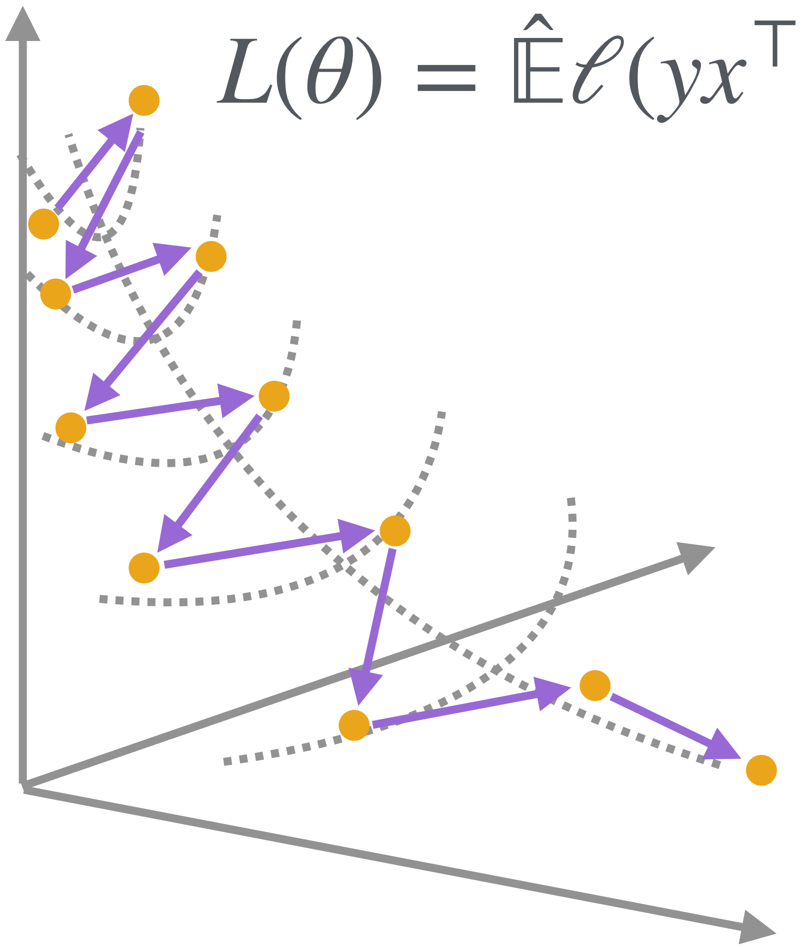




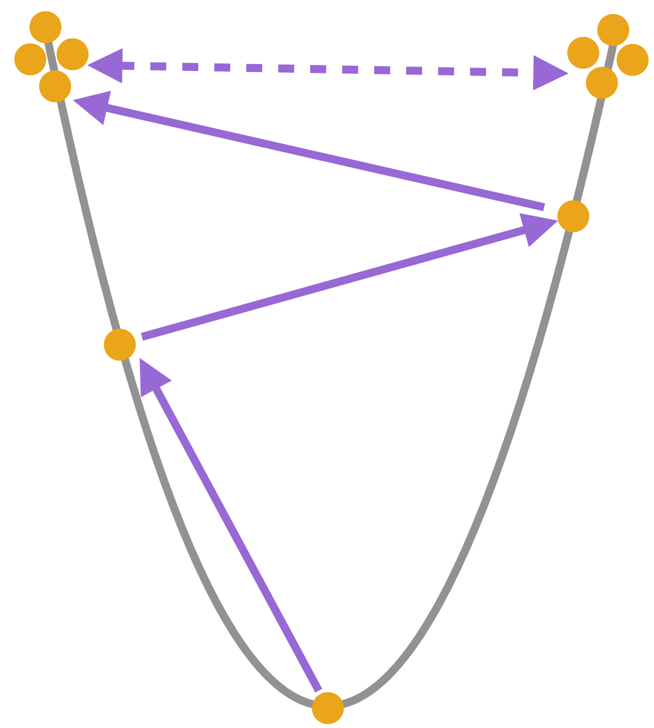
Unstable. $\tilde{L} \approx L$, $R \leq \Theta(1)$, "overshoot"

(3/3) Picture: valley + basin

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



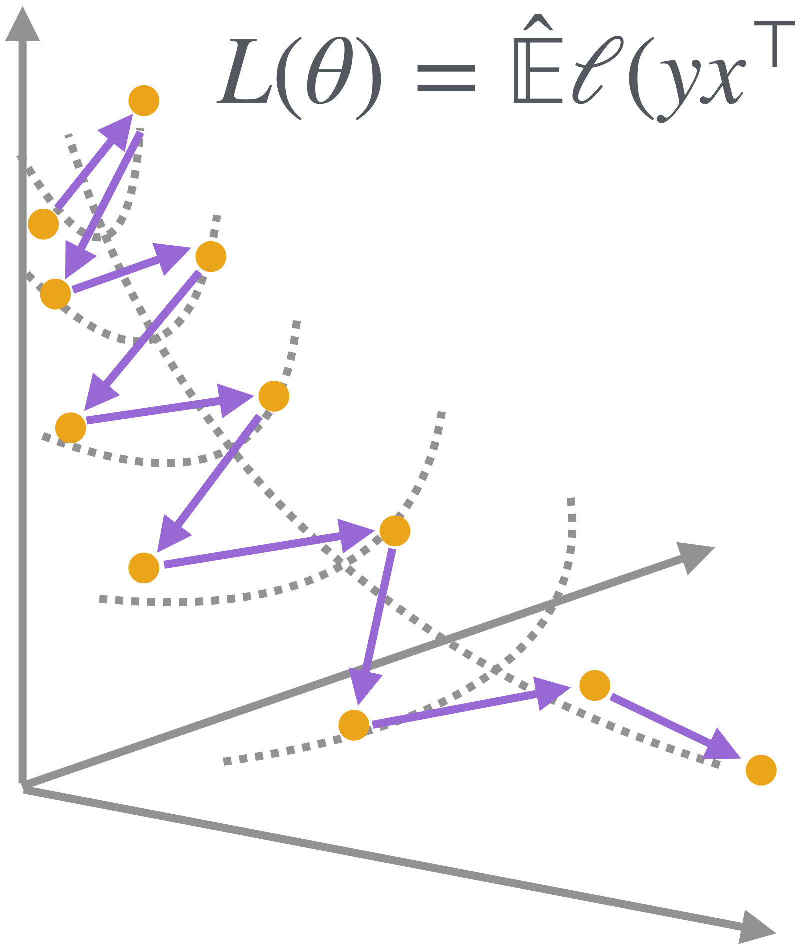
$$R(\theta) = \frac{\lambda}{2}\|\theta\|^2$$



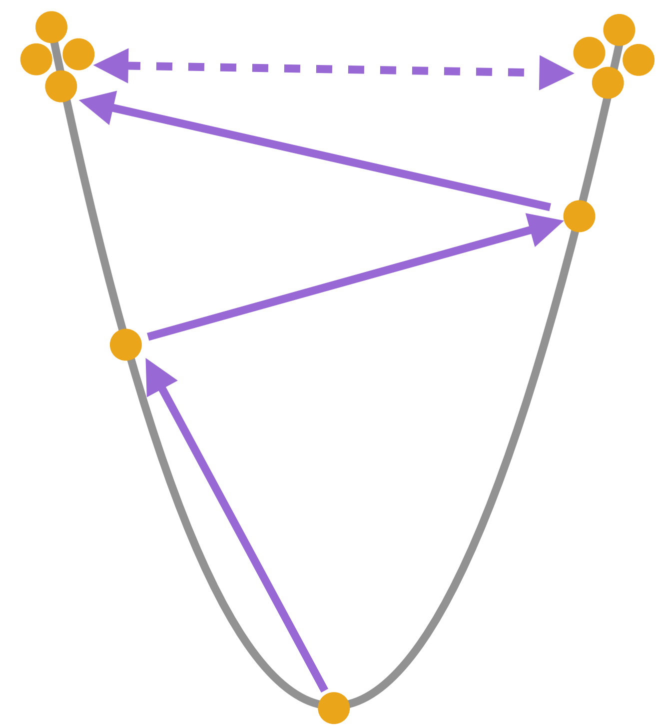
Unstable. $\tilde{L} \approx L$, $R \leq \Theta(1)$, "overshoot" $\|\tilde{\theta}\| = O(\ln\kappa)$

(3/3) Picture: valley + basin

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



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

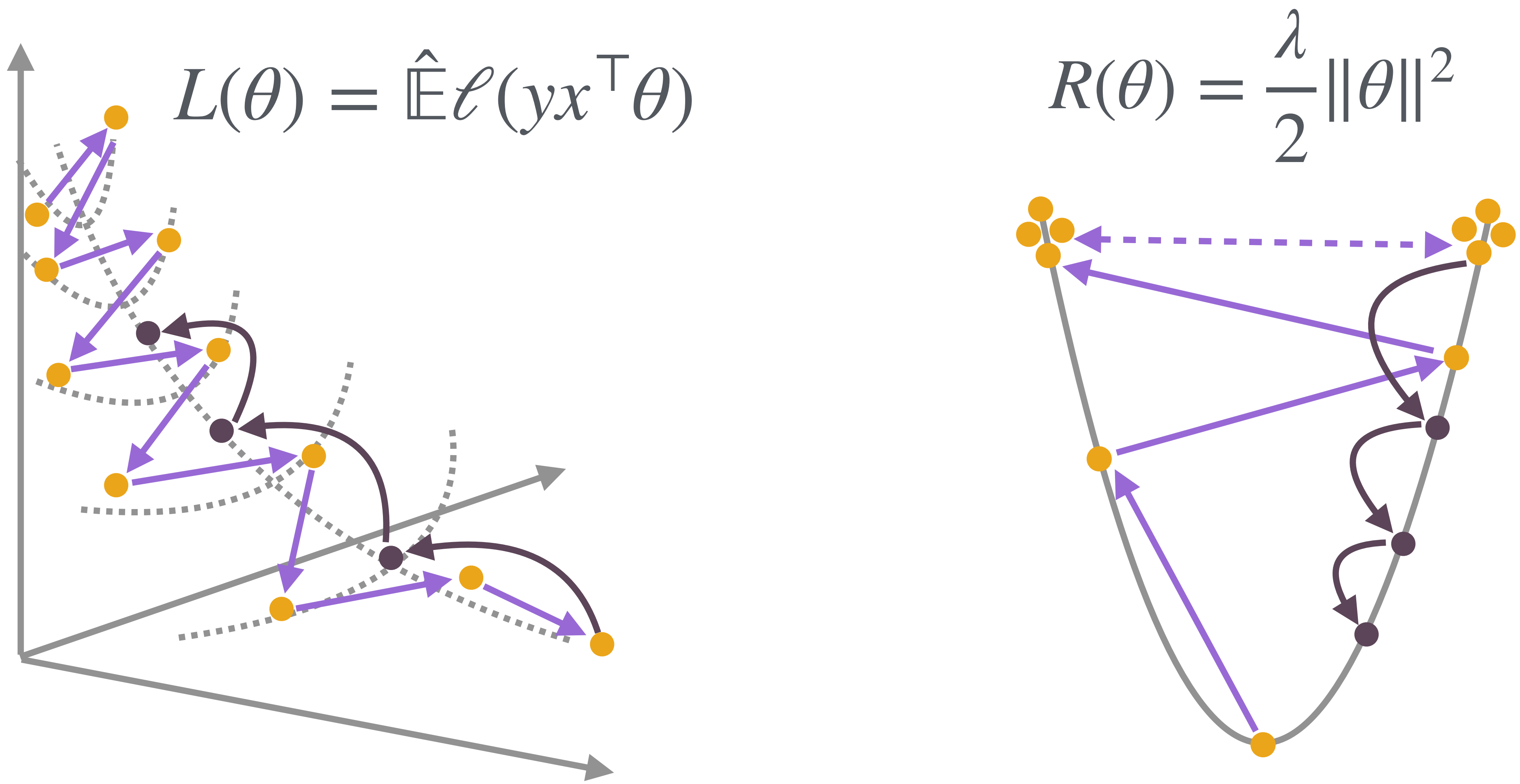


Unstable. $\tilde{L} \approx L$, $R \leq \Theta(1)$, "overshoot"

$\|\tilde{\theta}\| = O(\ln \kappa)$

$\sup \|\theta_t\| = \Theta(\eta) = \mathrm{poly}(\kappa)$

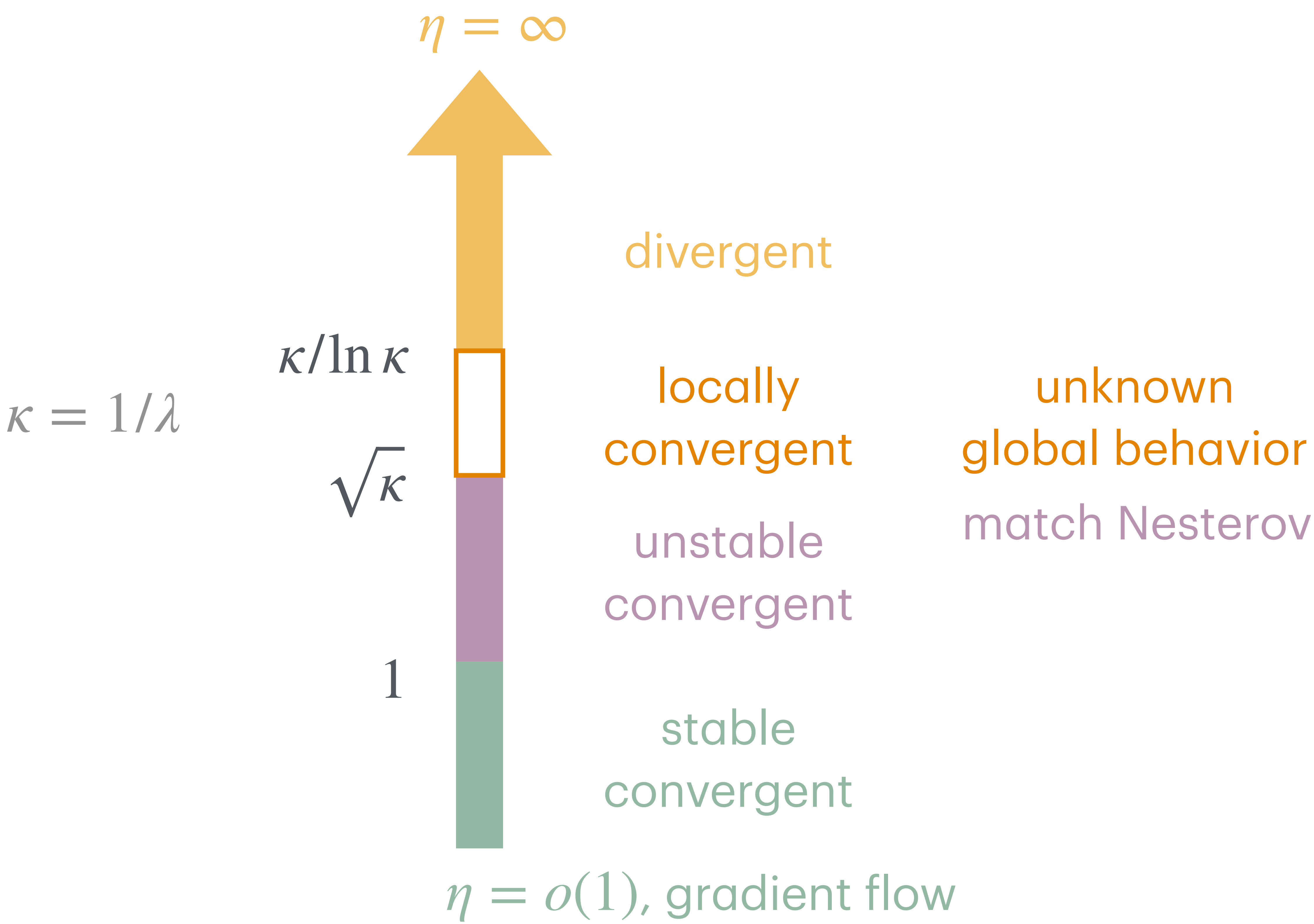
(3/3) Picture: valley + basin



Unstable. $\tilde{L} \approx L$, $R \leq \Theta(1)$, “overshoot” $\qquad \|\tilde{\theta}\| = O(\ln \kappa)$

Stable. “move back” $\qquad \sup \|\theta_t\| = \Theta(\eta) = \mathrm{poly}(\kappa)$

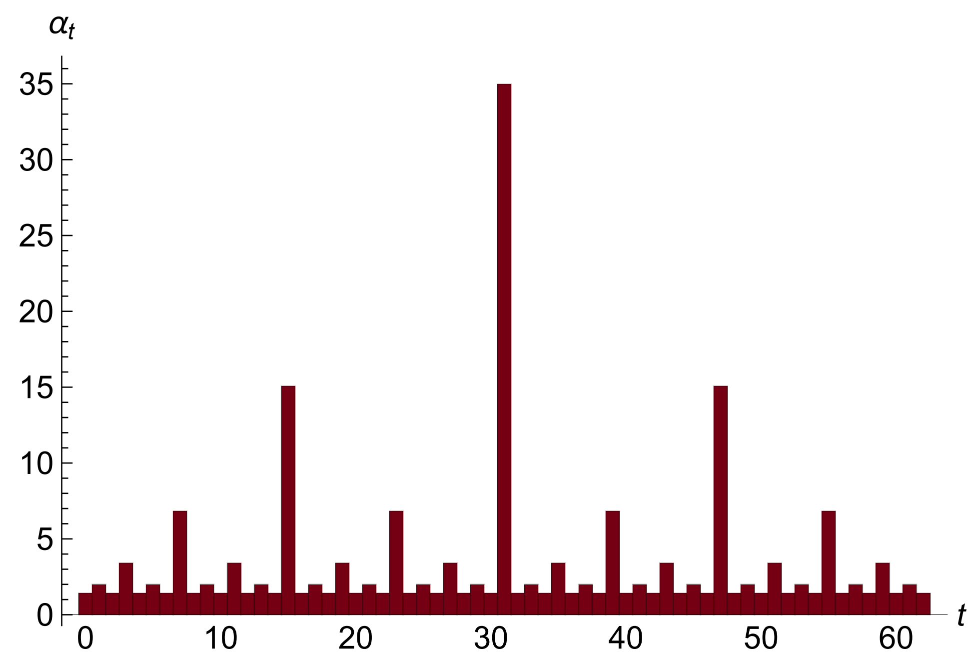
(3/3) Stepsize diagram



Related: long steps

Theorem. Let L be convex and smooth. For GD with *silver stepsize scheduler* $(\alpha_s)_{s\geq 0}$ and $t = 2^k - 1$, we have

$$L(\theta_t) - \min L = O(1/t^{1.27})$$



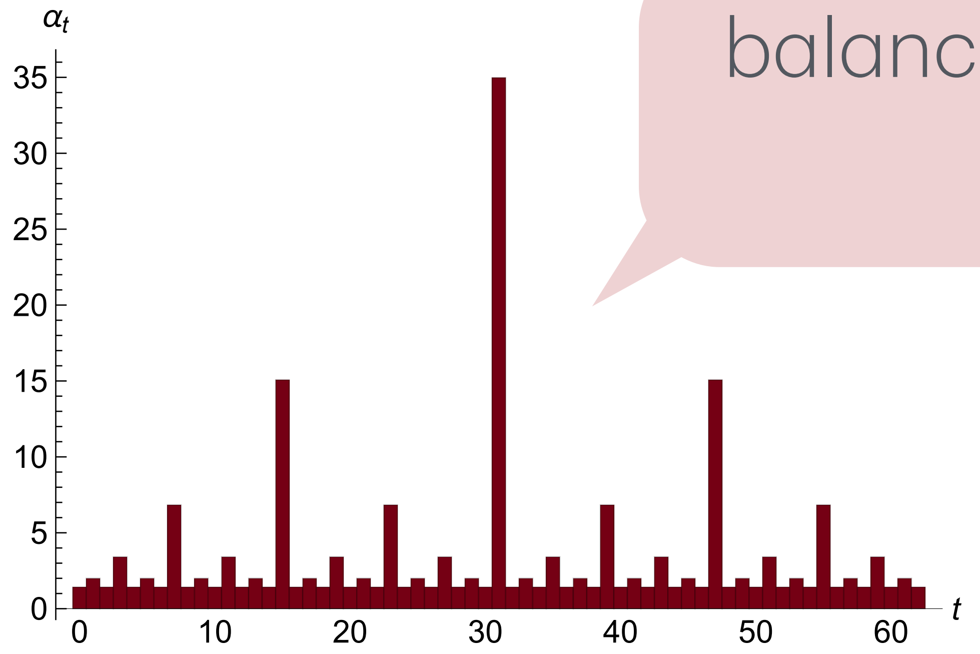
Altschuler, Parrilo. "Acceleration by stepsize hedging II: silver stepsize schedule for smooth convex optimization." Mathematical Programming 2024

Grimmer, Shu, Wang. "Composing optimized stepsize schedules for gradient descent." Mathematics of Operations Research 2025

Related: long steps

Theorem. Let L be convex and smooth. For GD with *silver stepsize scheduler* $(\alpha_s)_{s\geq 0}$ and $t = 2^k - 1$, we have

$$L(\theta_t) - \min L = O(1/t^{1.27})$$



balance performance in high/low curvatures:
$0.5\theta^2$ vs Huber function

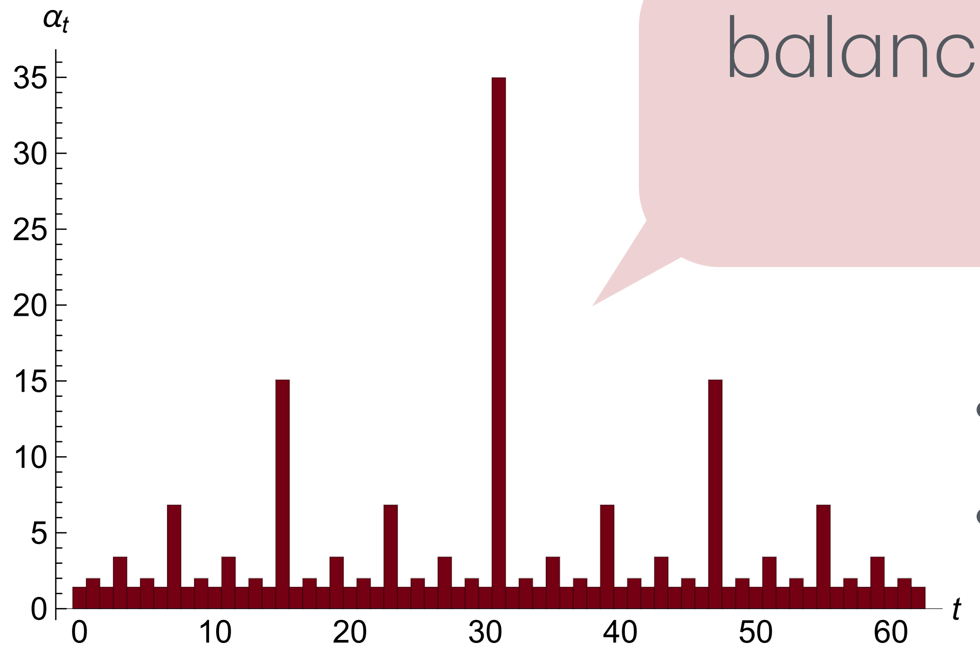
Altschuler, Parrilo. "Acceleration by stepsize hedging II: silver stepsize schedule for smooth convex optimization." Mathematical Programming 2024

Grimmer, Shu, Wang. "Composing optimized stepsize schedules for gradient descent." Mathematics of Operations Research 2025

Related: long steps

Theorem. Let L be convex and smooth. For GD with *silver stepsize scheduler* $(\alpha_s)_{s \geq 0}$ and $t = 2^k - 1$, we have

$$L(\theta_t) - \min L = O(1/t^{1.27})$$



balance performance in high/low curvatures:
$0.5\theta^2$ vs Huber function

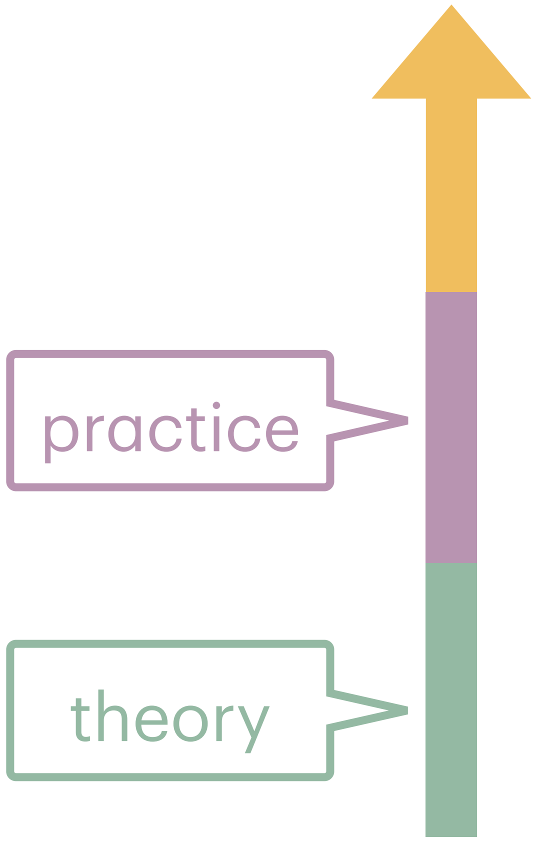
- cover more problems, e.g., quadratics
- less practical stepsize scheduler

Altschuler, Parrilo. "Acceleration by stepsize hedging II: silver stepsize schedule for smooth convex optimization." Mathematical Programming 2024

Grimmer, Shu, Wang. "Composing optimized stepsize schedules for gradient descent." Mathematics of Operations Research 2025

Summary

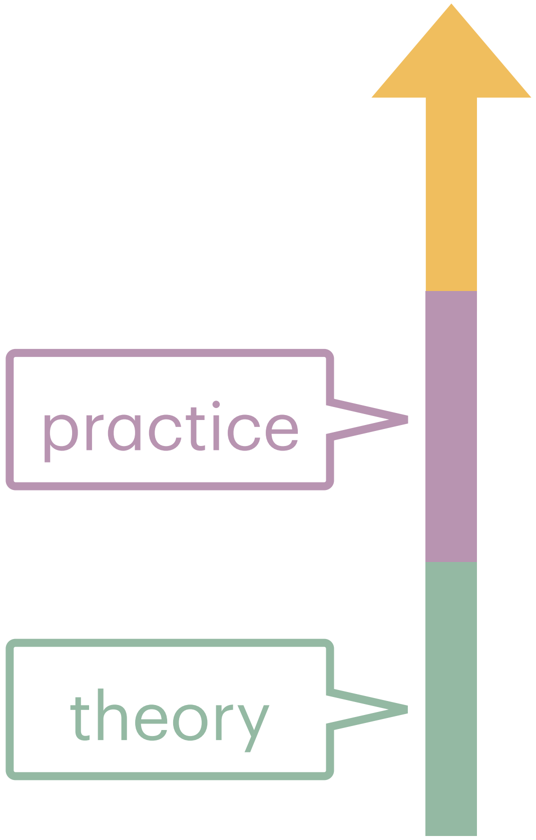
training instability caused by large stepsize

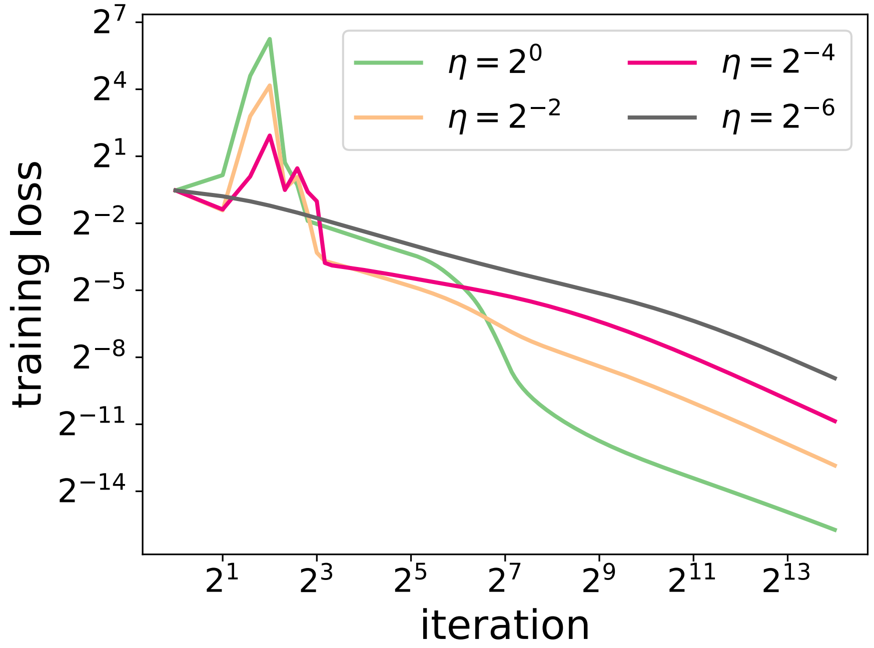


Summary

training instability caused by large stepsize

acceleration via large stepsize: three ML examples



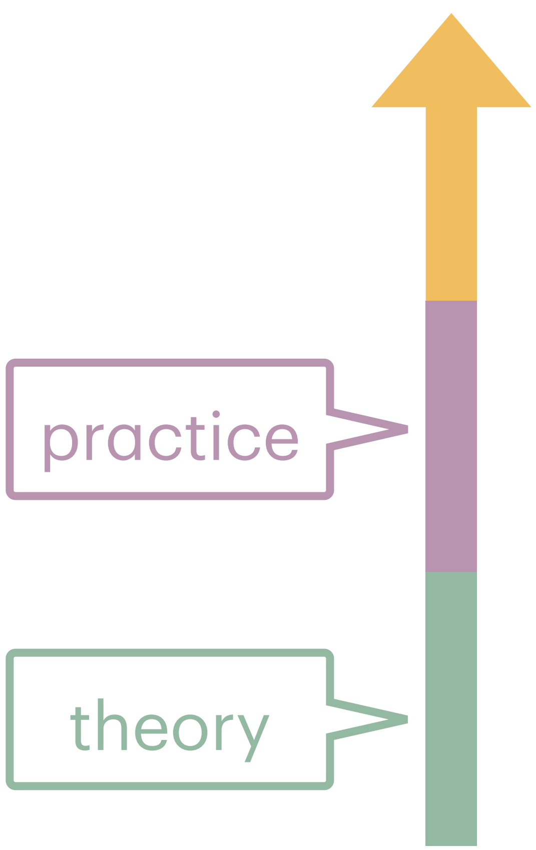


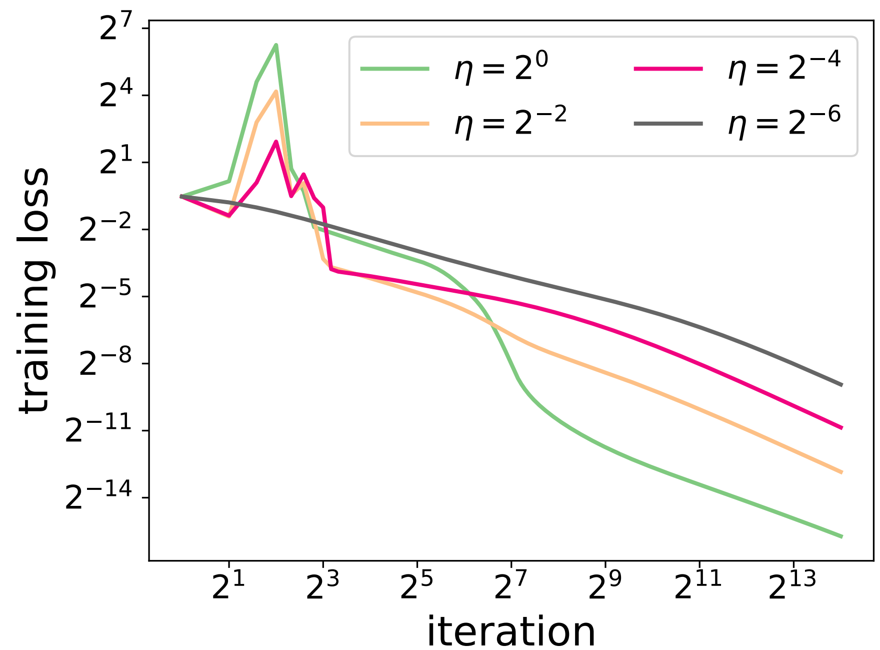
Summary

training instability caused by large stepsize

acceleration via large stepsize: three ML examples

general losses





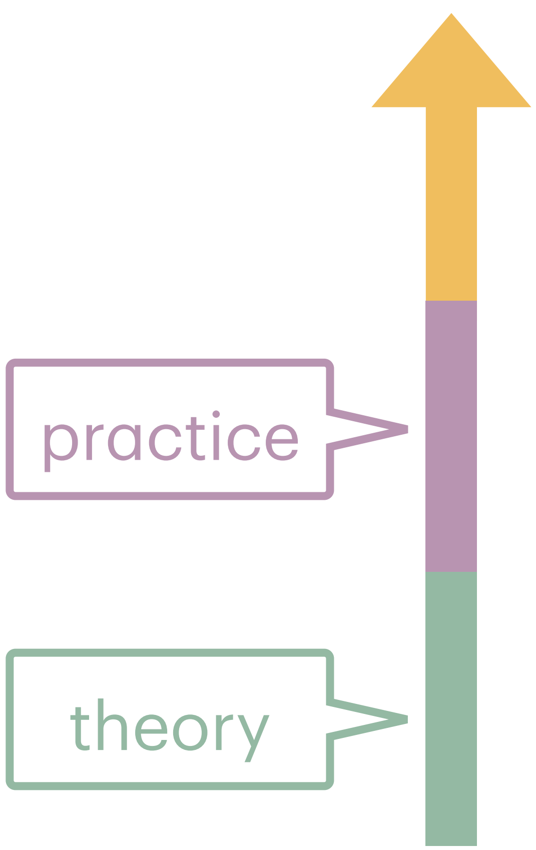
Summary

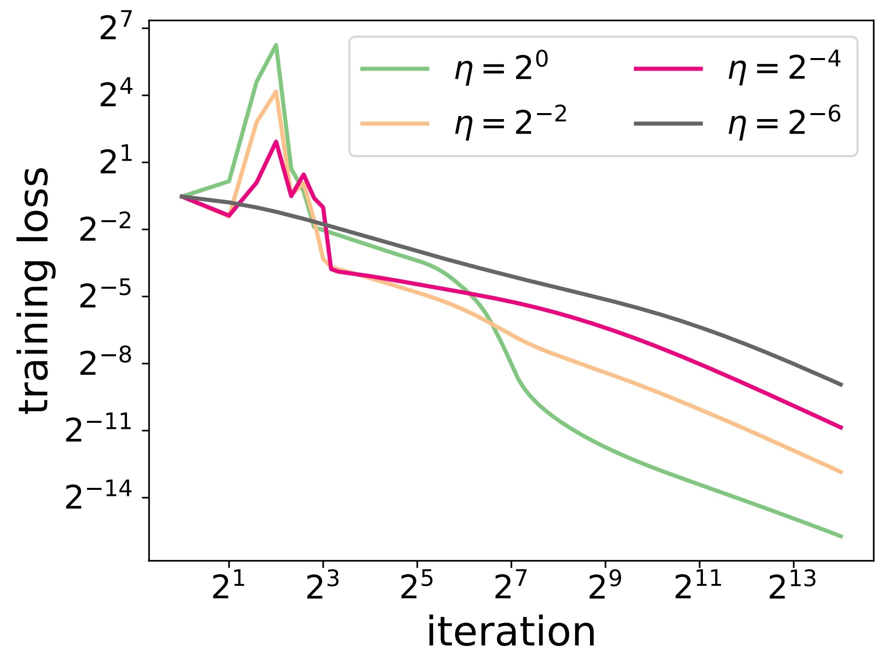
training instability caused by large stepsize

acceleration via large stepsize: three ML examples

general losses

neural networks





Summary

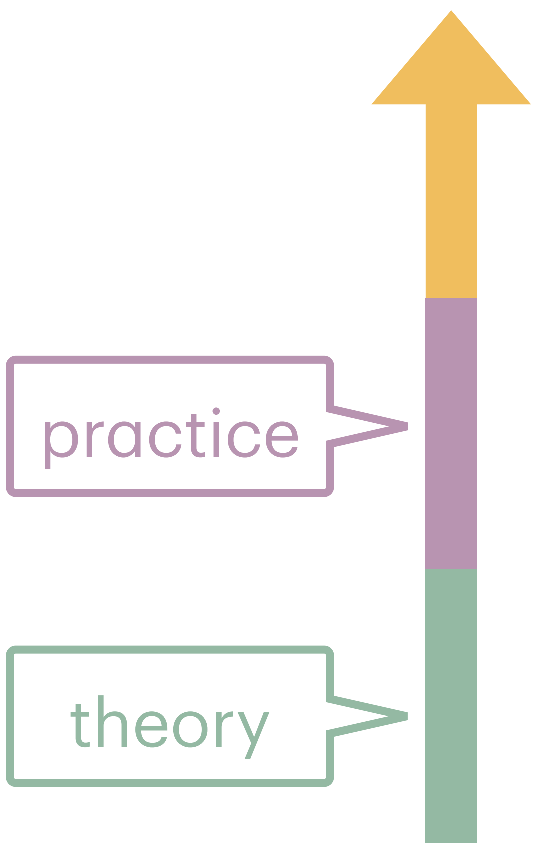
training instability caused by large stepsize

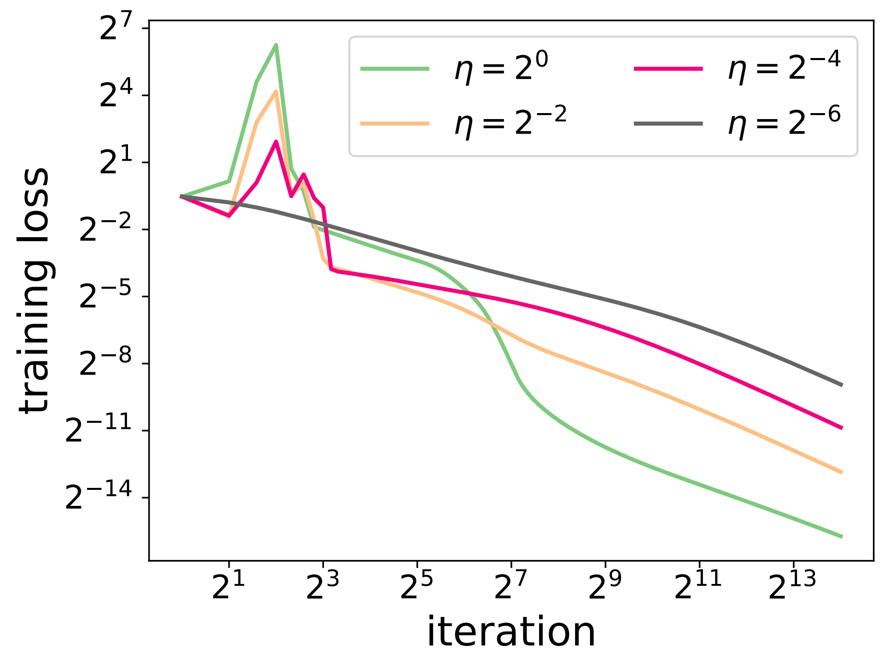
acceleration via large stepsize: three ML examples

general losses

neural networks

implicit bias, generalization





Summary

training instability caused by large stepsize

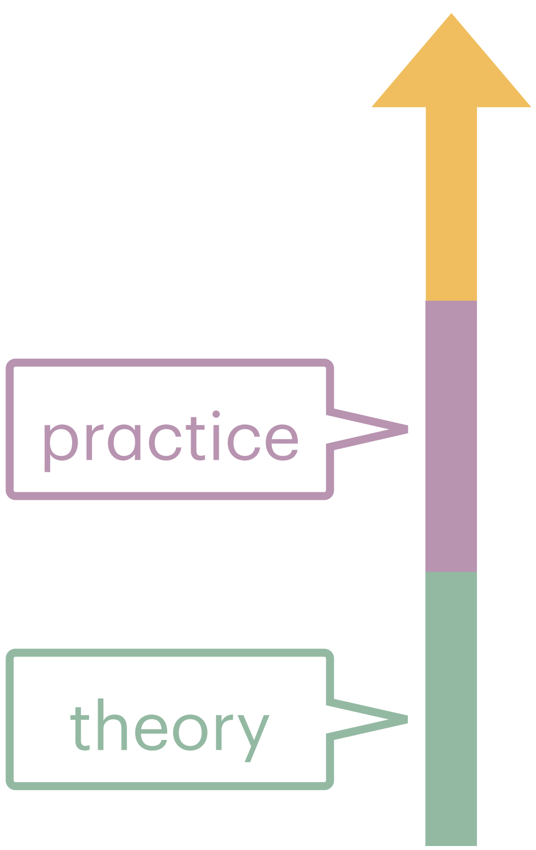
acceleration via large stepsize: three ML examples

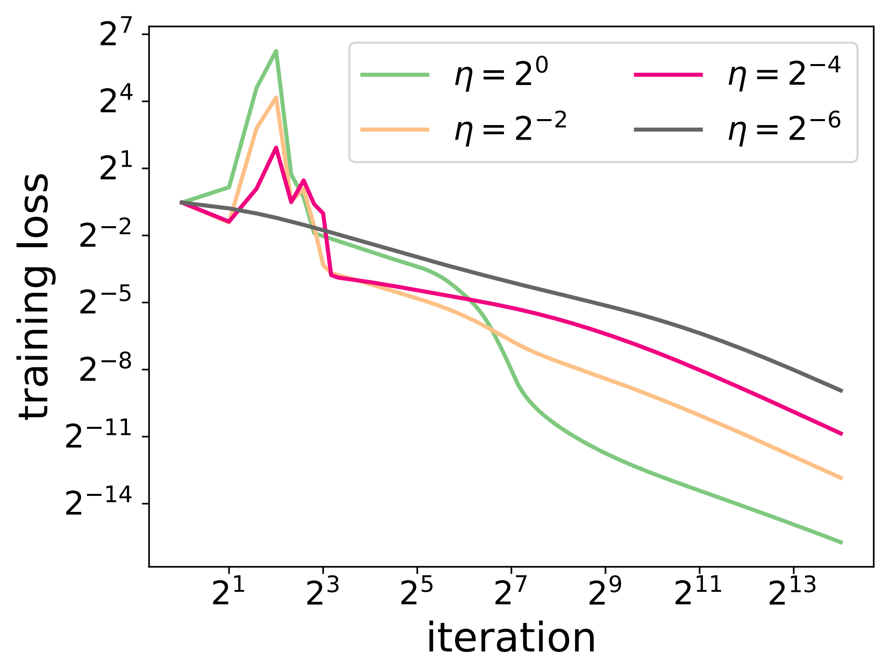
general losses

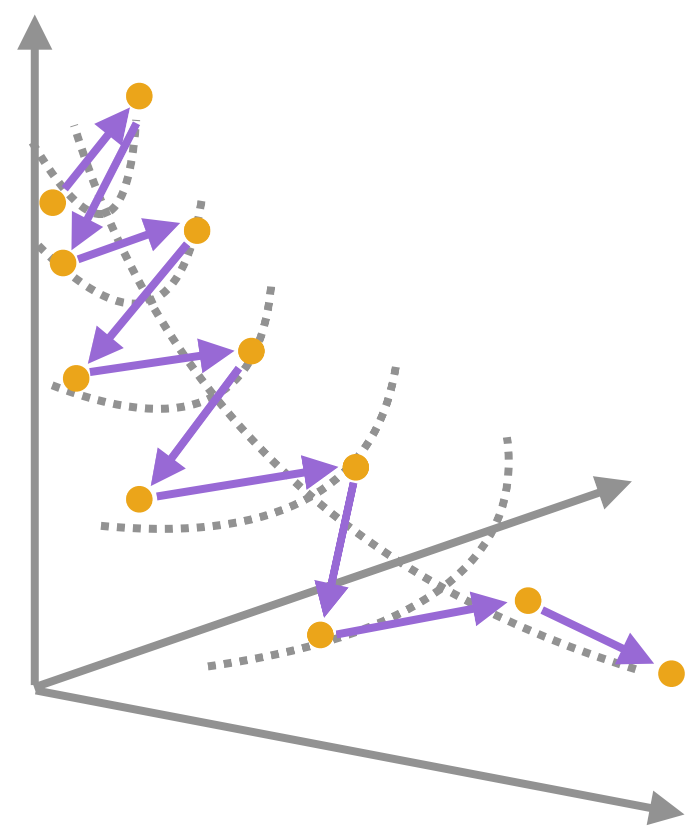
neural networks

implicit bias, generalization

new mental picture: valley







Summary

training instability caused by large stepsize

acceleration via large stepsize: three ML examples

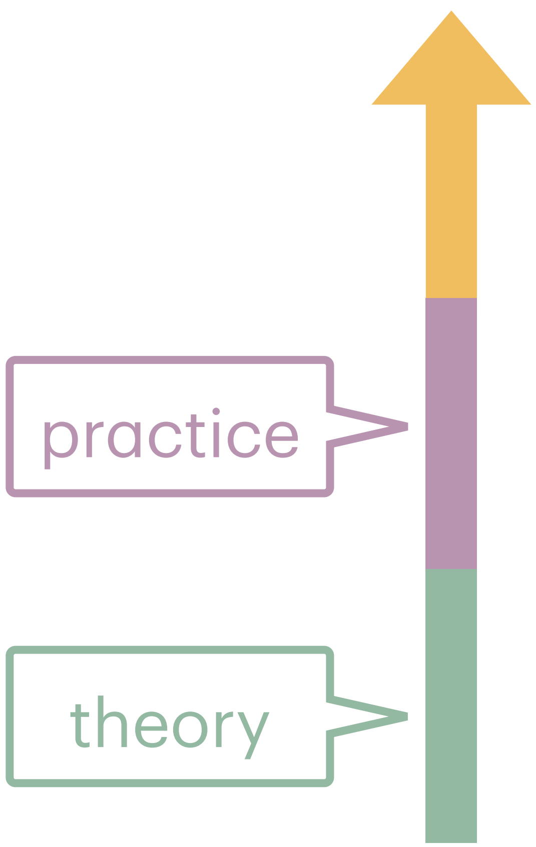
general losses

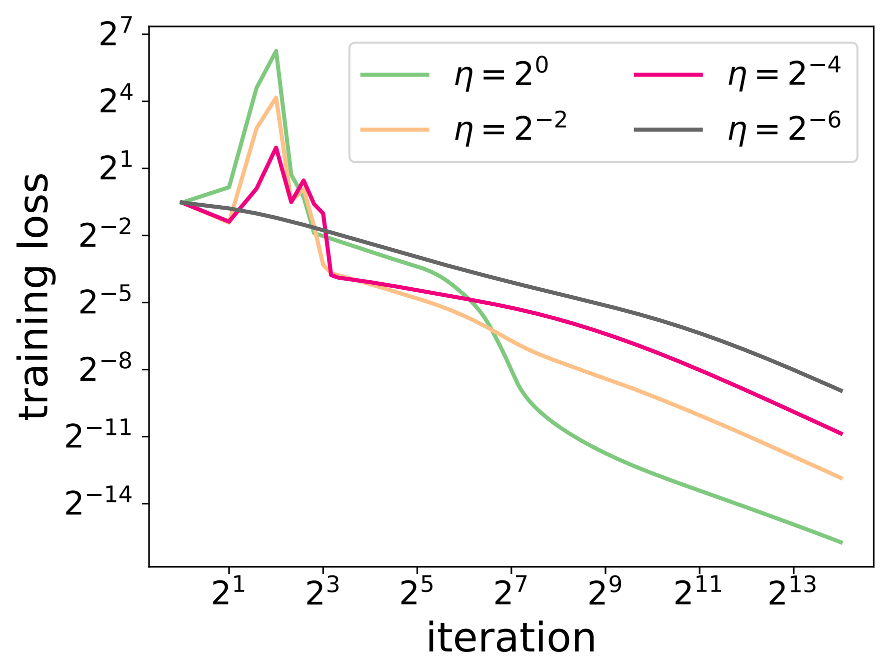
neural networks

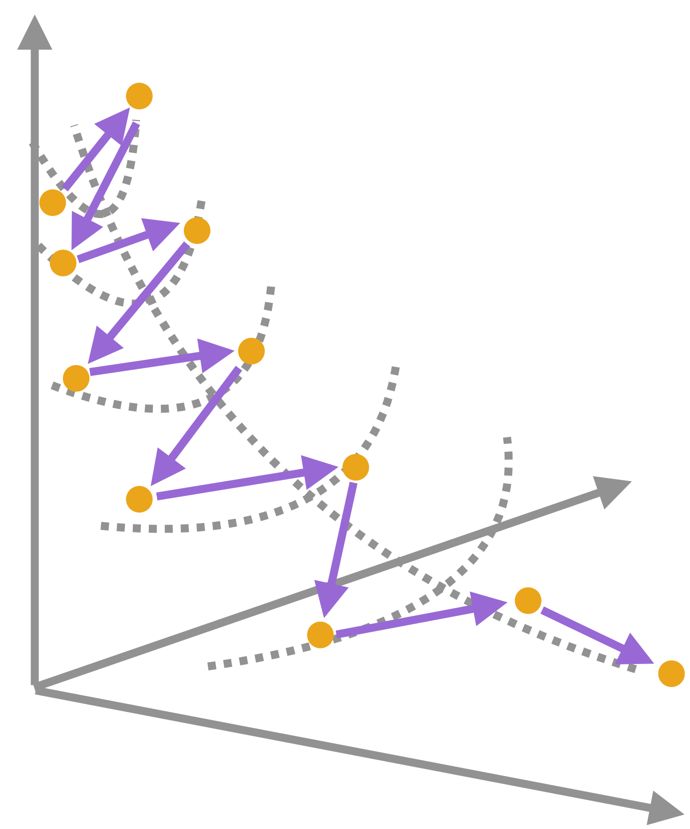
implicit bias, generalization

new mental picture: valley

cross entropy, attention, ...







Open problems (set 1/2)

Call for clear, rigorous understanding on

Open problems (set 1/2)

Call for clear, rigorous understanding on

🤔 what functional property enables large stepsize?

Open problems (set 1/2)

Call for clear, rigorous understanding on

🤔 what functional property enables large stepsize?

🤔 trackable measures of trajectory: sharpness? local mean?

Open problems (set 1/2)

Call for clear, rigorous understanding on

🤔 what functional property enables large stepsize?

🤔 trackable measures of trajectory: sharpness? local mean?

🤔 early-phase feature learning, especially against NTK?

Open problems (set 1/2)

Call for clear, rigorous understanding on

🤔 what functional property enables large stepsize?

🤔 trackable measures of trajectory: sharpness? local mean?

🤔 early-phase feature learning, especially against NTK?

🤔 large stepsize for other optimizers, e.g., SGD, Adam?

Open problems (set 2/2)

Call for useful, heuristic insights on

Open problems (set 2/2)

Call for useful, heuristic insights on

💡 better stepsize schedulers, e.g., warmup, stepsize decaying?

Open problems (set 2/2)

Call for useful, heuristic insights on

💡 better stepsize schedulers, e.g., warmup, stepsize decaying?

💡 better optimizer, e.g., preconditioning, normalization?

Open problems (set 2/2)

Call for useful, heuristic insights on

💡 better stepsize schedulers, e.g., warmup, stepsize decaying?

💡 better optimizer, e.g., preconditioning, normalization?

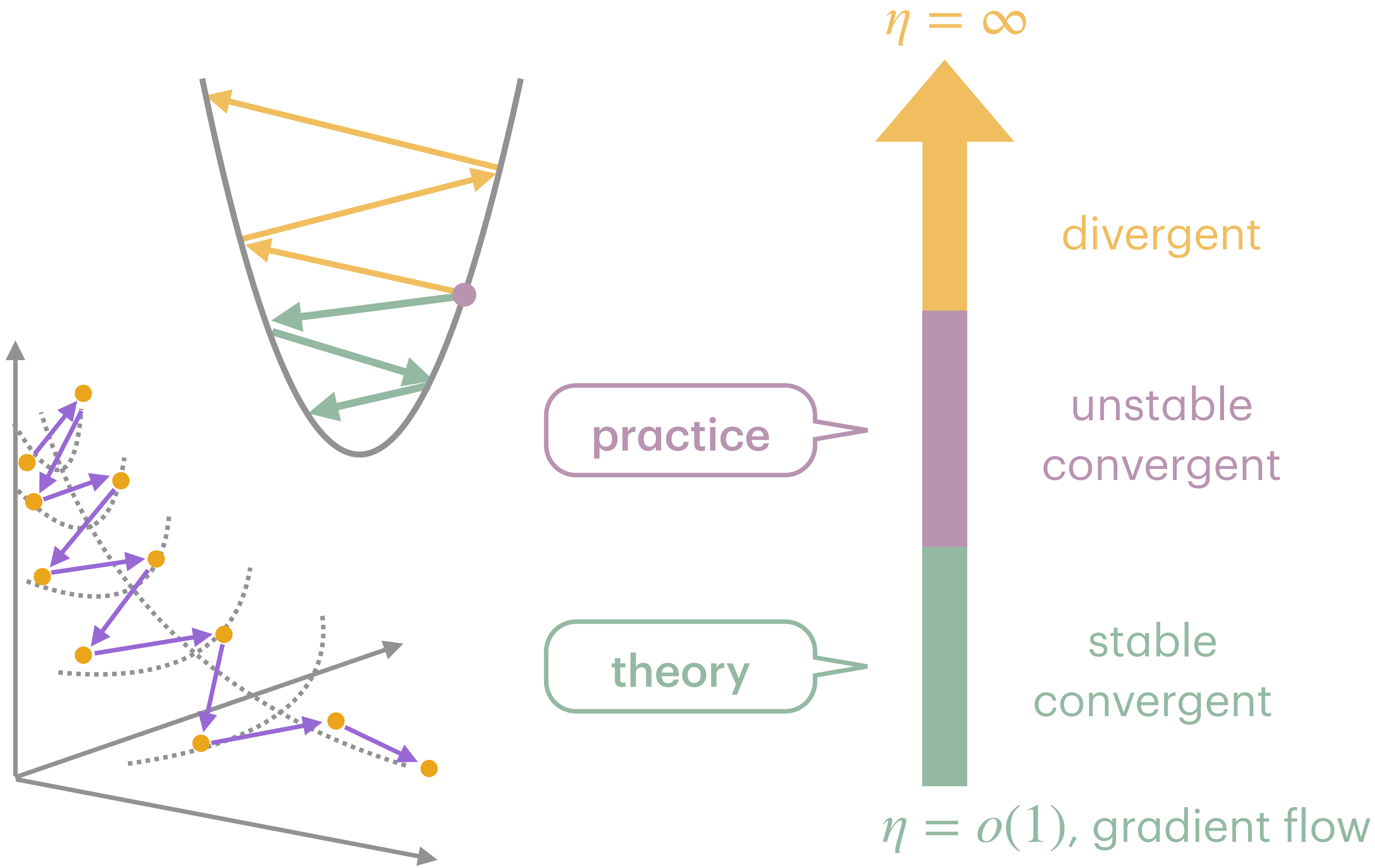
💡 interplay between stepsize vs structure, e.g., attention, depth?

Open problems (set 2/2)

Call for useful, heuristic insights on

- 💡 better stepsize schedulers, e.g., warmup, stepsize decaying?
- 💡 better optimizer, e.g., preconditioning, normalization?
- 💡 interplay between stepsize vs structure, e.g., attention, depth?
- 💡 how to understand other instabilities, e.g., data, precision?

Q & A



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