

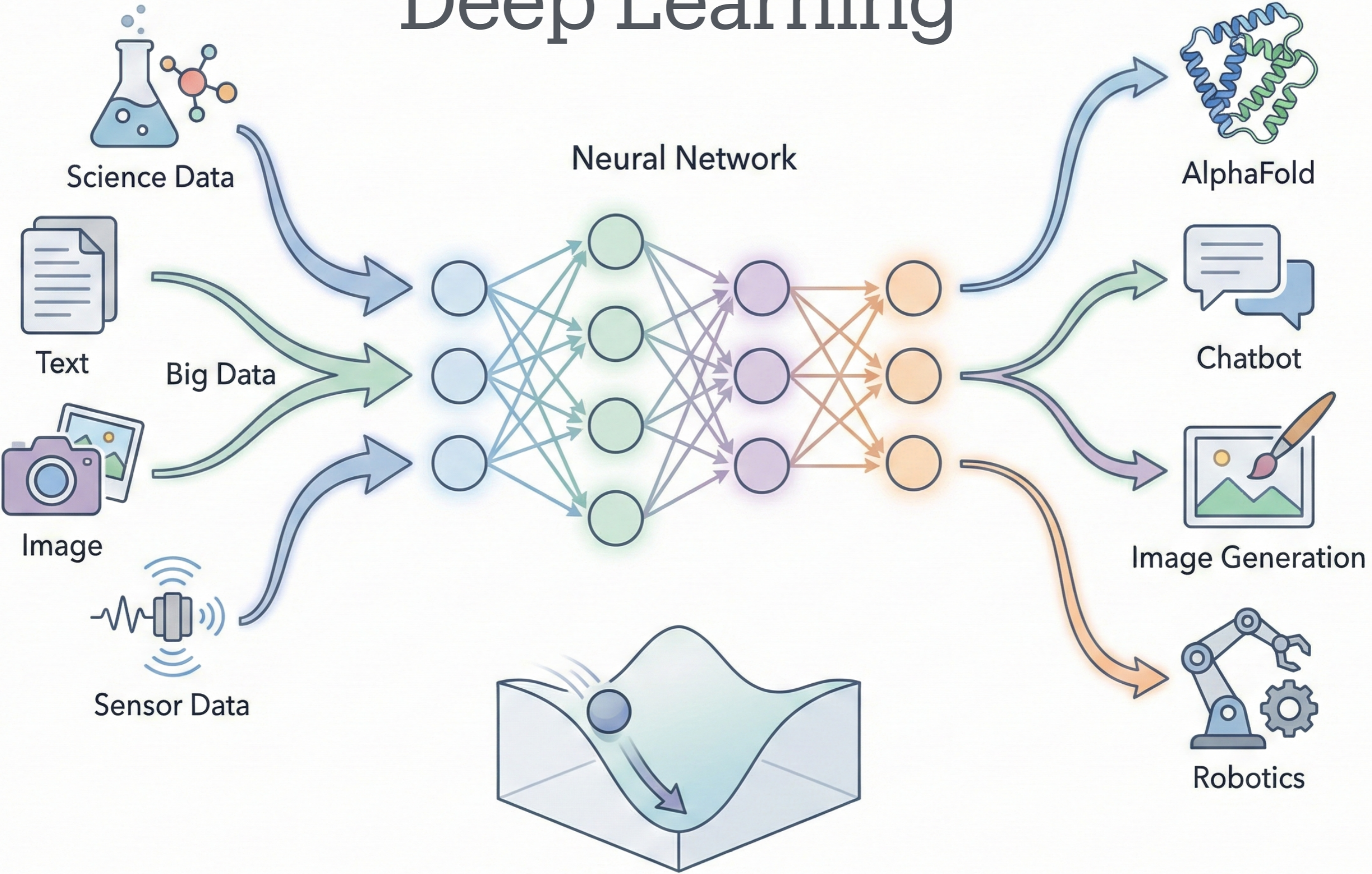
Towards a Less Conservative Theory of Machine Learning

Unstable Optimization & Implicit Regularization

Jingfeng Wu



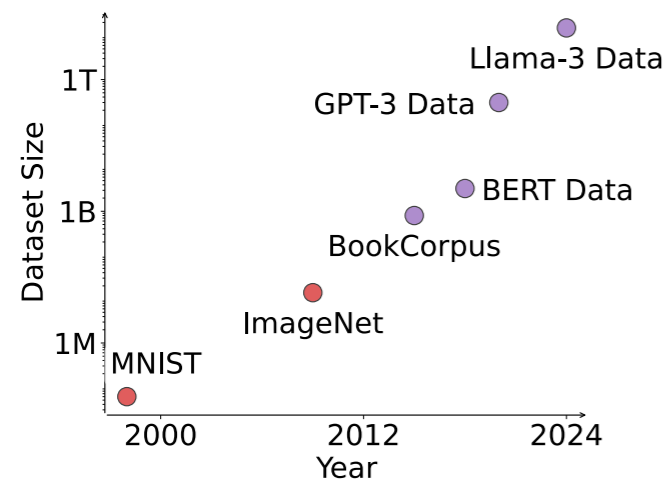
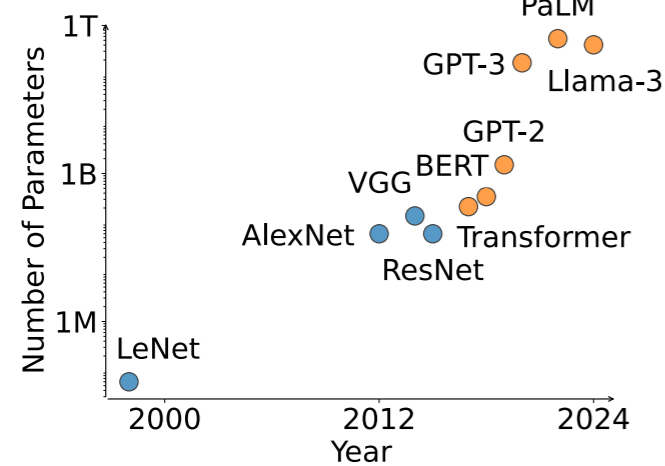
Deep Learning



Gradient-based Optimization

What makes deep learning thrive?

scaling



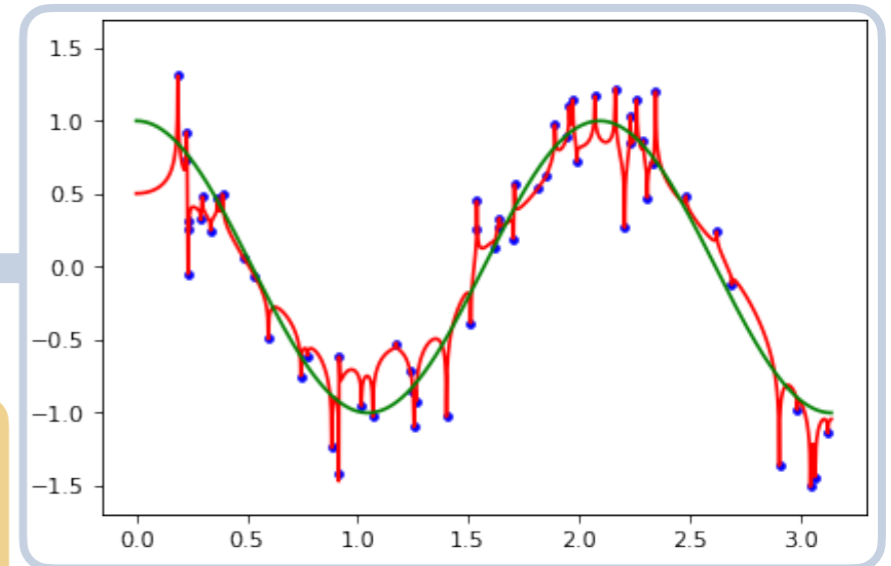
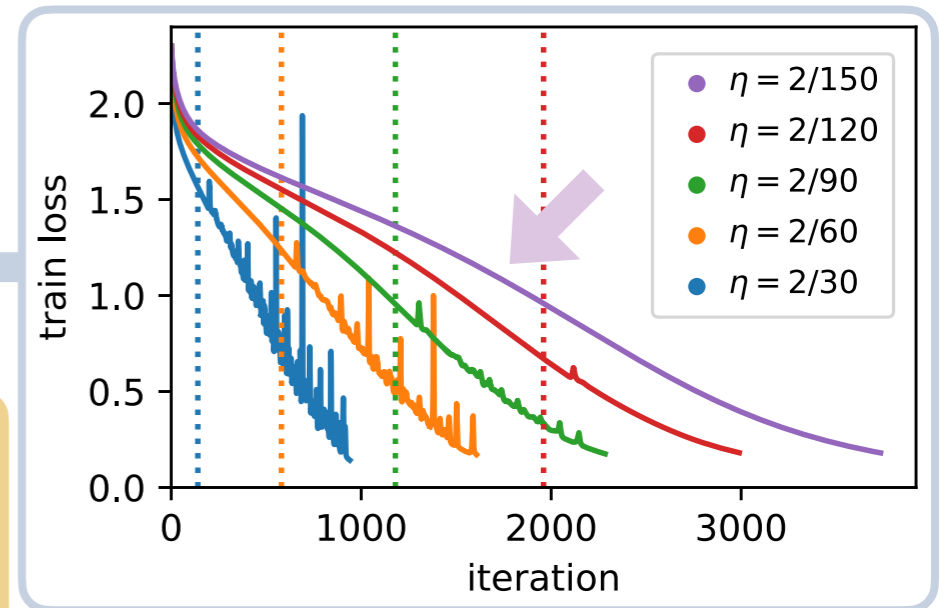
new mechanisms

optimization
training instability
"edge of stability"
...

large stepsize
unstable optimization

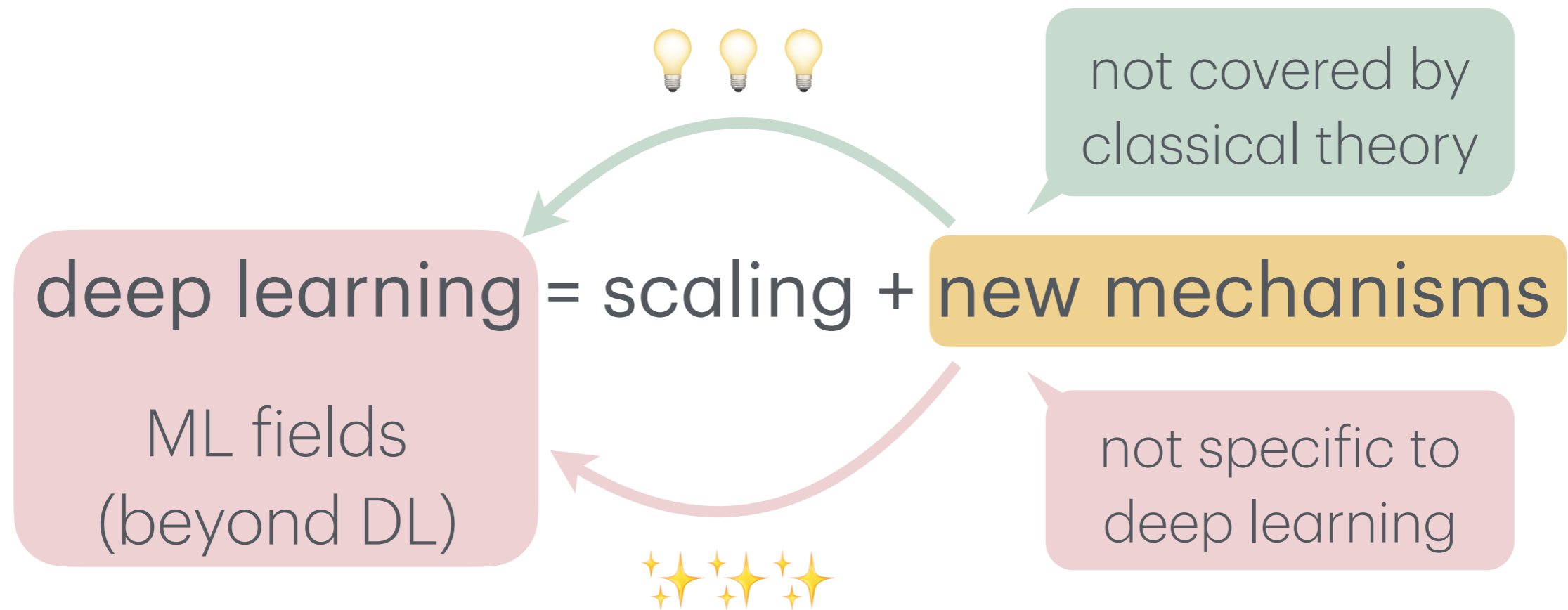
generalization
overparameterization
"benign overfitting"
...

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



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

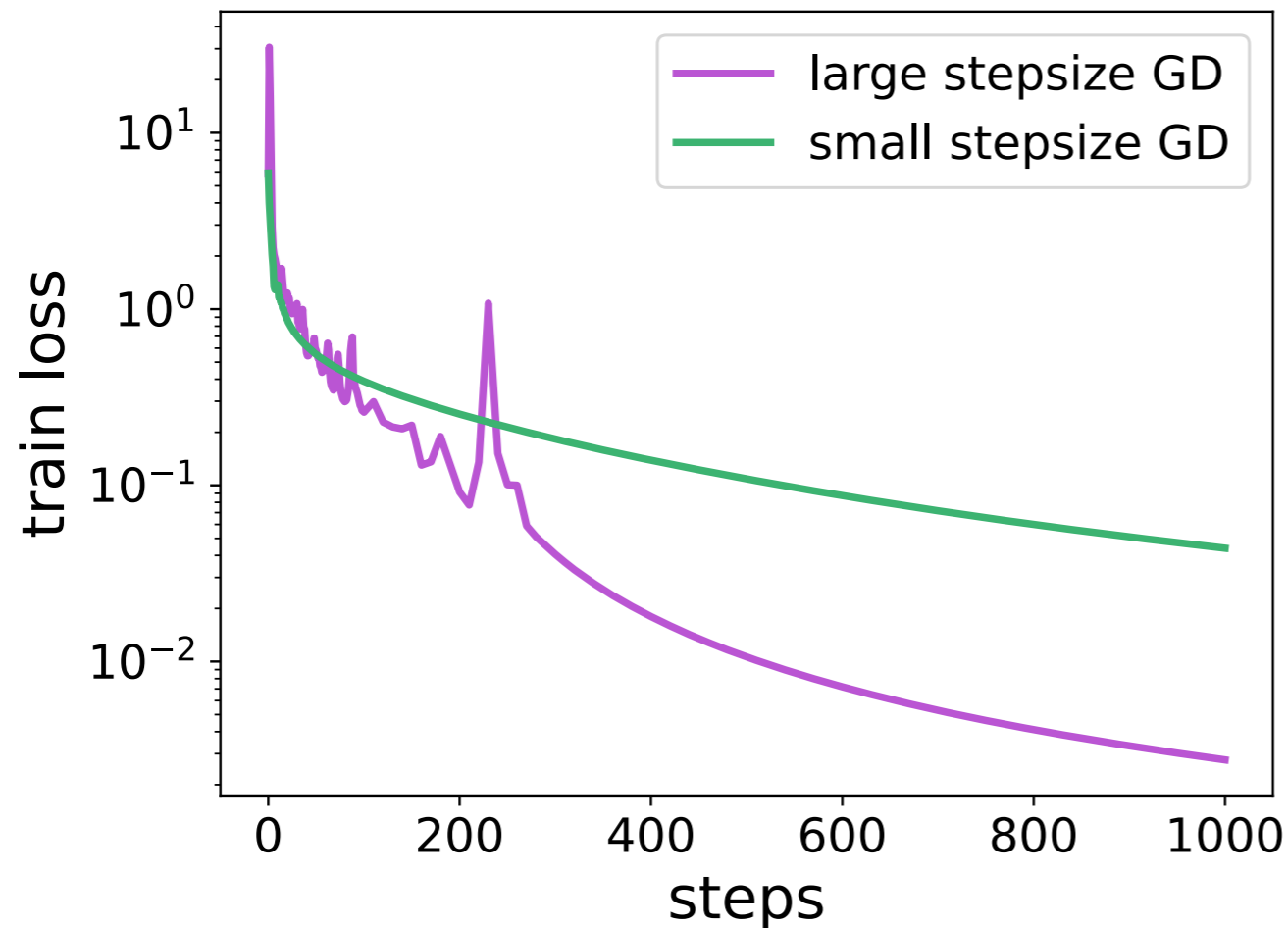
large stepsize accelerates gradient descent in logistic regression

- “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

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

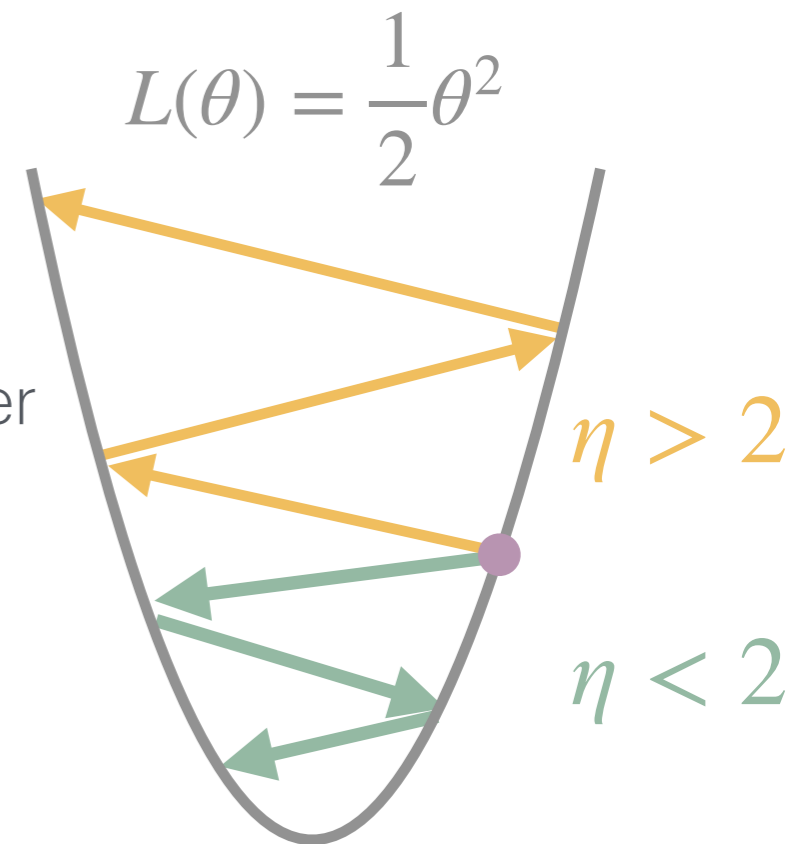
**classical theory fails to
predict best stepsize**

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 $t = O(\kappa \ln(1/\epsilon))$

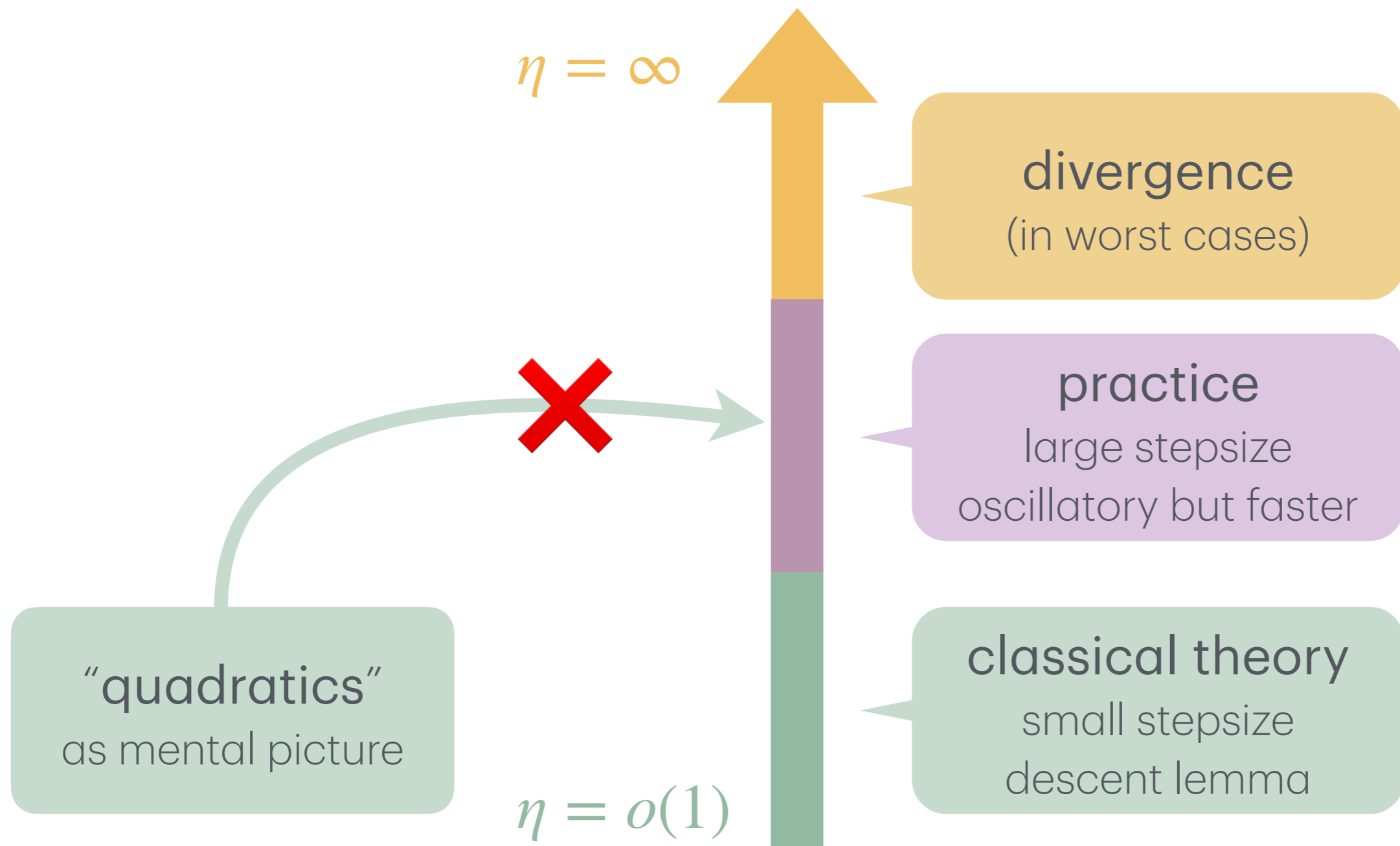
condition number

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

acceleration by Nesterov's momentum:

$$O(1/t^2) \text{ \& } O(\sqrt{\kappa} \ln(1/\epsilon))$$

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

linear
regression

logistic
regression

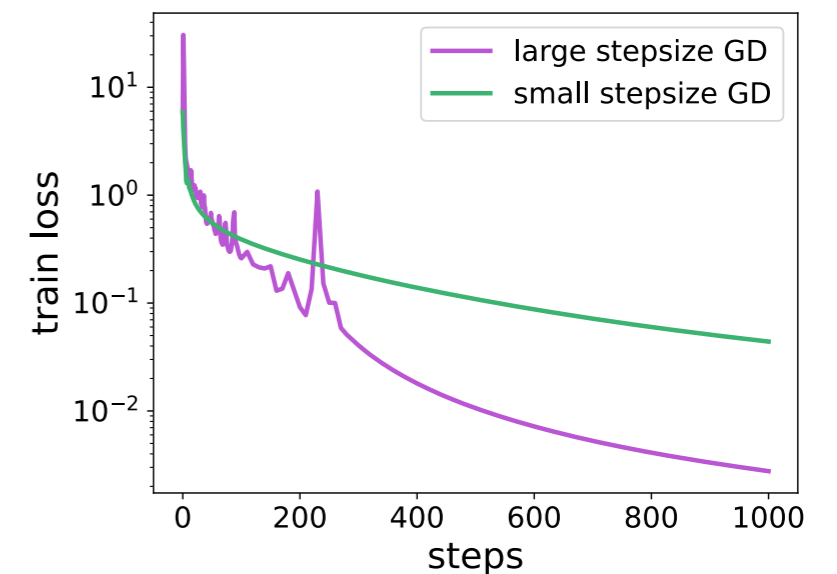
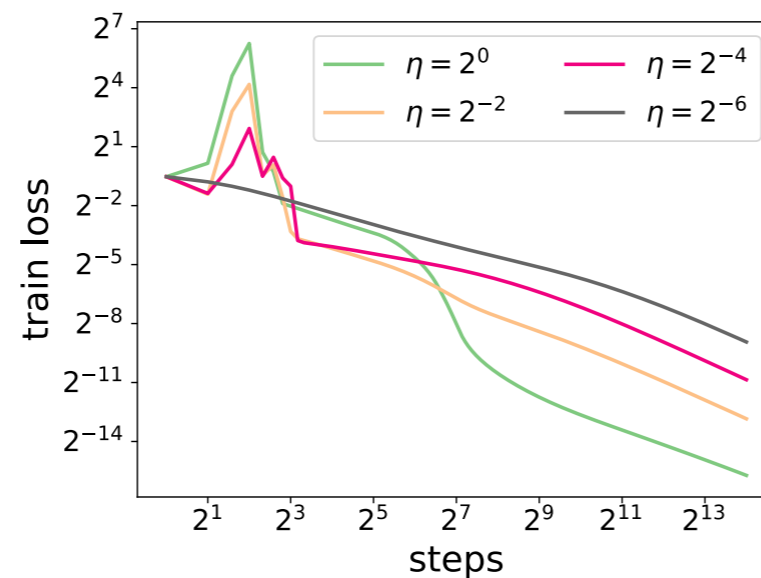
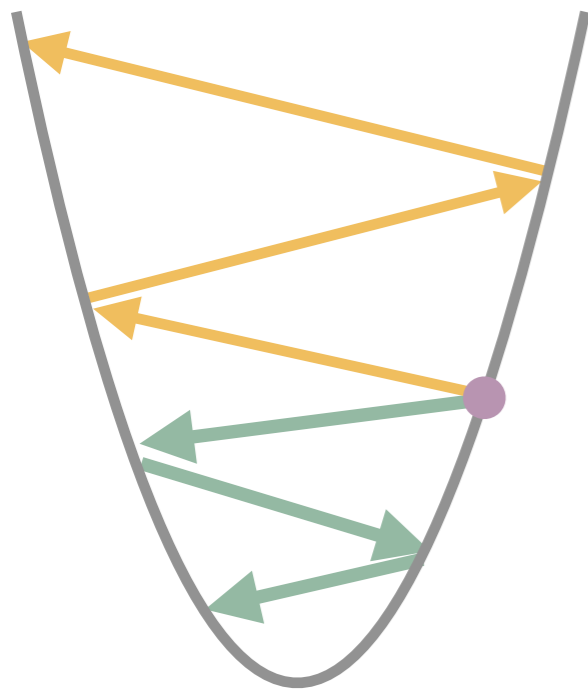
.....

deep
learning

unstable
optimization
impossible

observable
& provable

unstable
optimization
observed



Logistic regression

smooth, convex
non-strongly convex

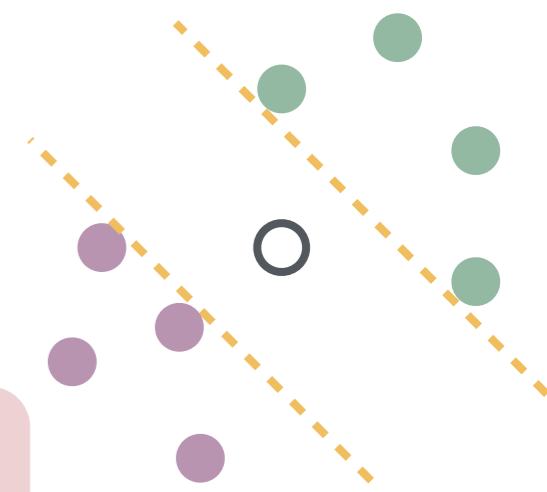
empirical risk $L(\theta) = \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y_i x_i^\top \theta))$

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

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

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

implied by
overparameterization



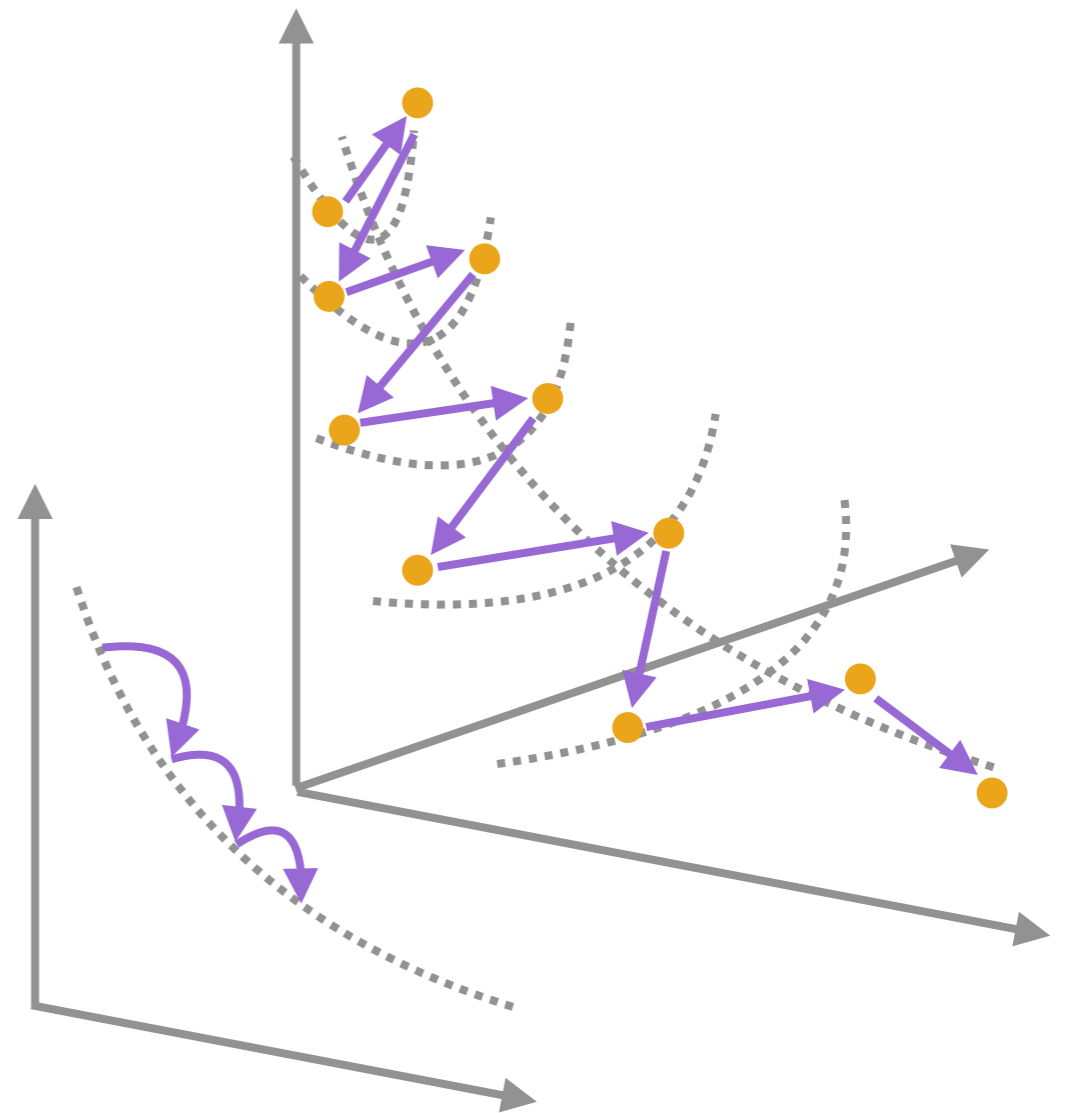
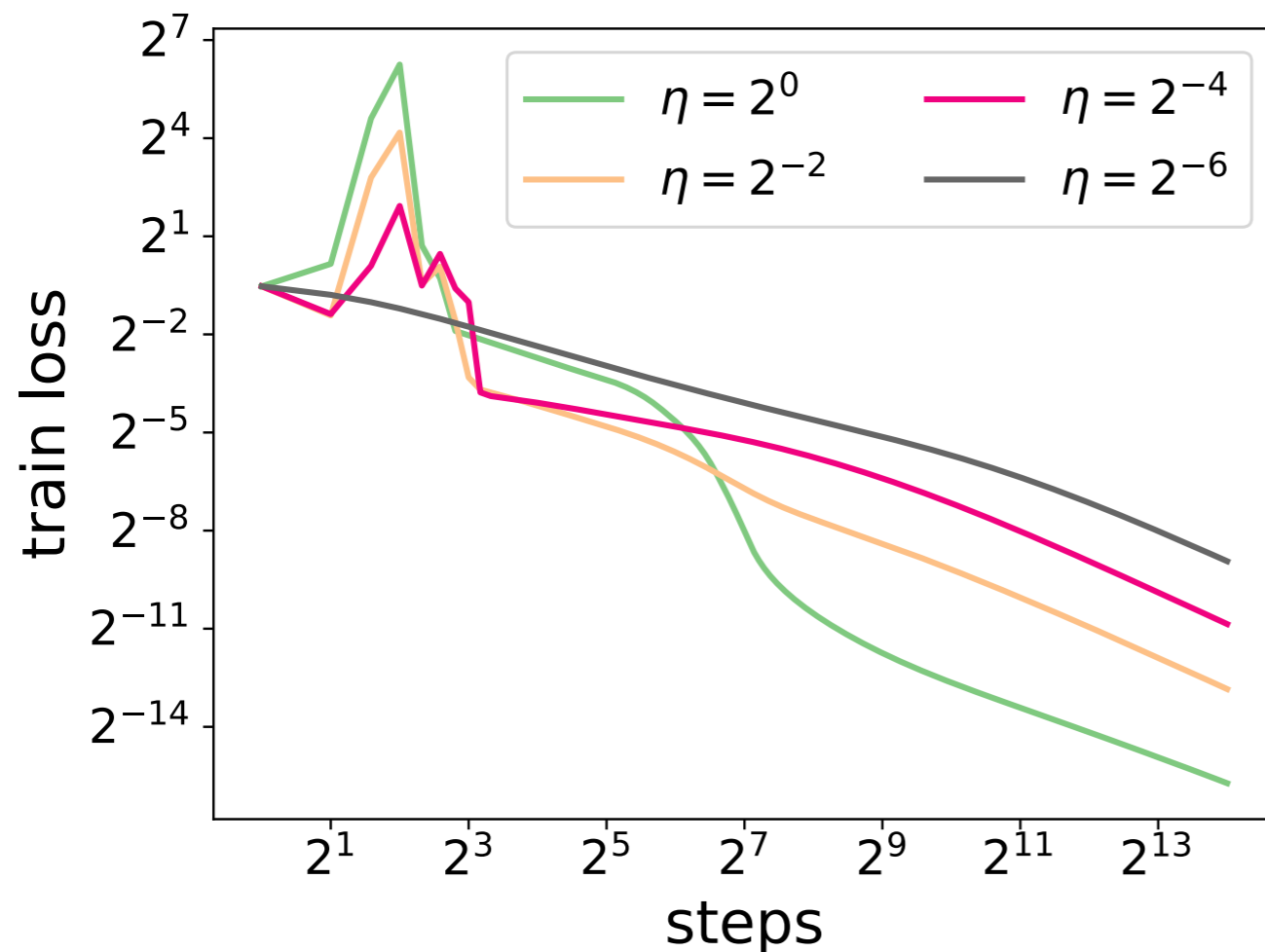
Classical theory. For $\eta = \Theta(1), L(\theta_t) \downarrow$ and $L(\theta_t) = \tilde{O}(1/t)$

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

Large stepsize accelerates GD

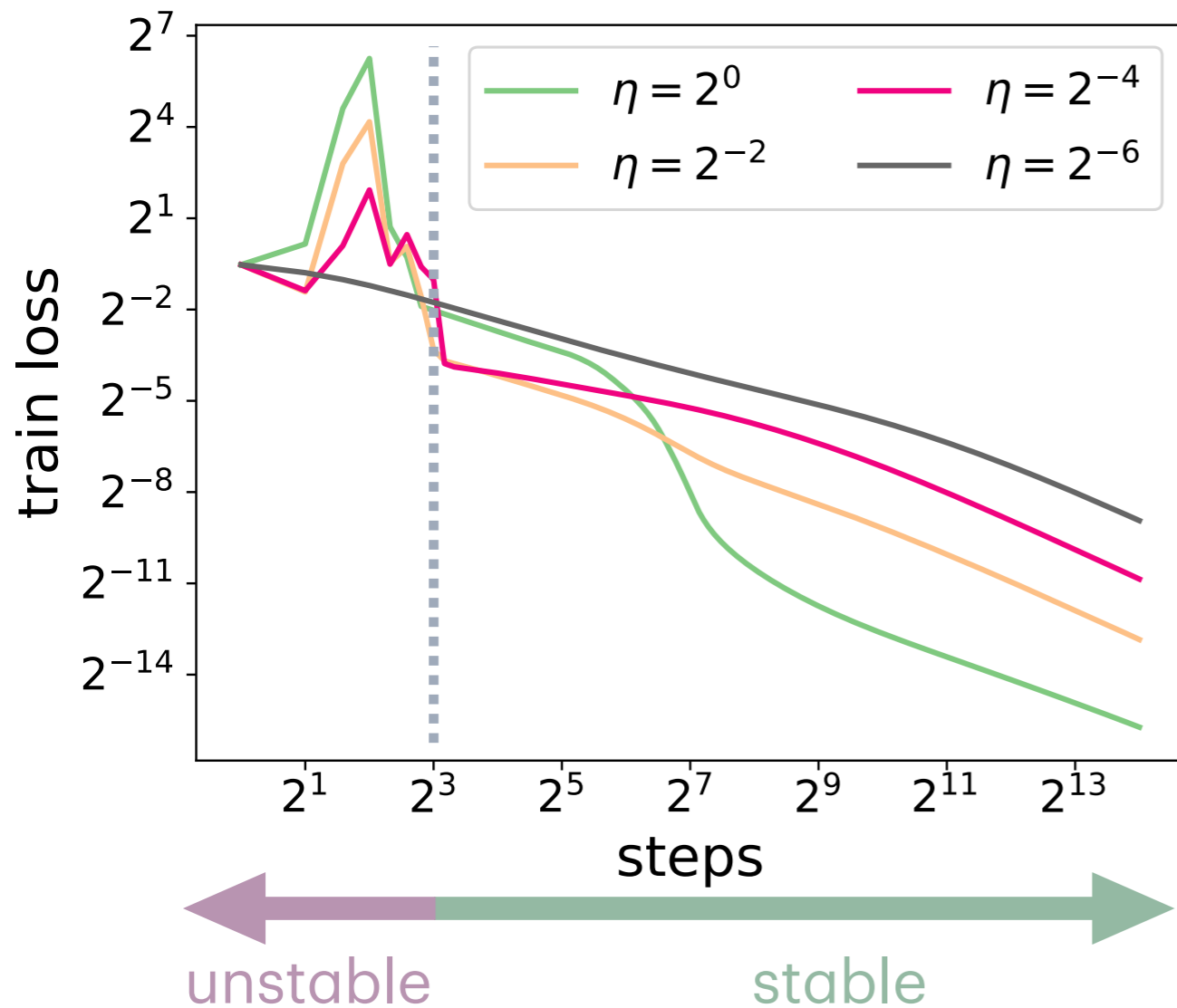
Theorem. Let #steps be $T \geq \Theta(1)$. For some $\eta = \Theta(T)$, we have

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



“open valley” as mental picture

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

Regularized logistic regression

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

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

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

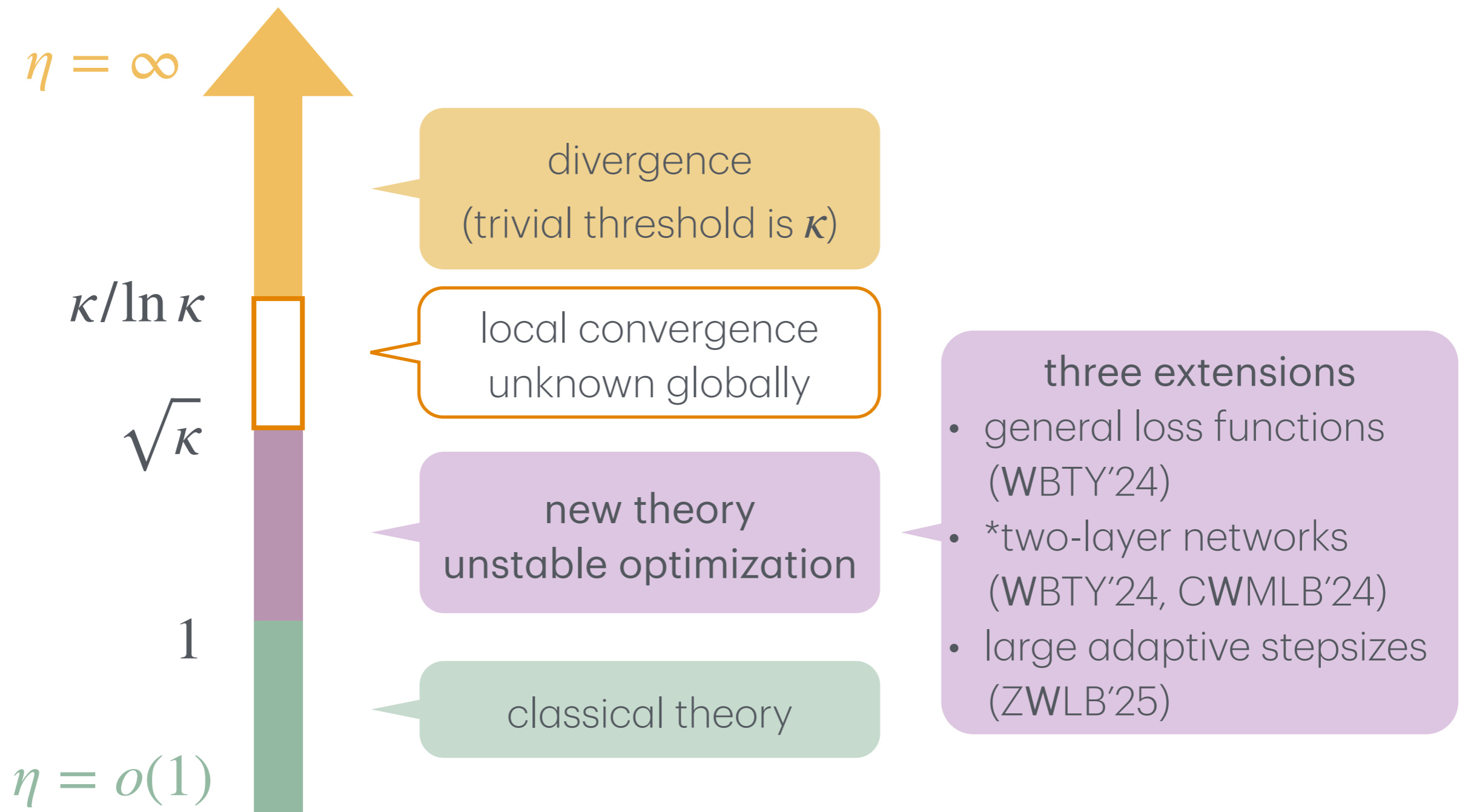
Classical theory. For $\eta = \Theta(1)$, we have $\tilde{L}(\theta_t) \downarrow$ and

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

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

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

Stepsize diagram revisited



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

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

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

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

Challenges:
non-convexity
confounders
...
Sandbox:
linear regression

“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 \geq 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)

“James-Stein dominates OLS” \Rightarrow machine learning

Linear regression

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

problem instance given by (Σ, θ^*)

nontrivial label noise

(excess) test error / prediction risk

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

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

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

Explicit / implicit regularization

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

ridge regression

hyperparameter: $\lambda \geq 0$

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

gradient descent

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

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

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

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

natural problems
e.g., power-law

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

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

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

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

for $\eta t = 1/\lambda$

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

polynomial filter

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

more effective when
 θ^* decays fast

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

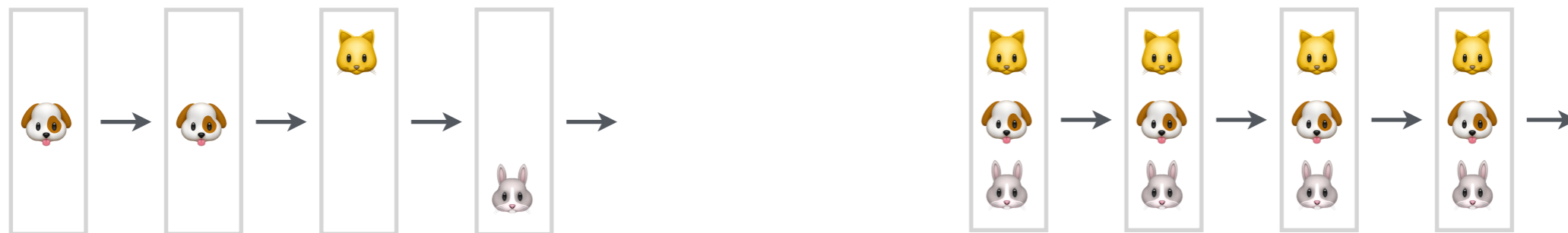
exponential filter

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

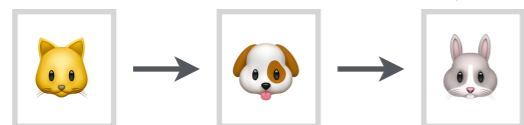
(“SGD” dominates) GD dominates ridge

Theorem. In linear regression, GD dominates ridge;

multi-pass SGD (with replacement) = GD



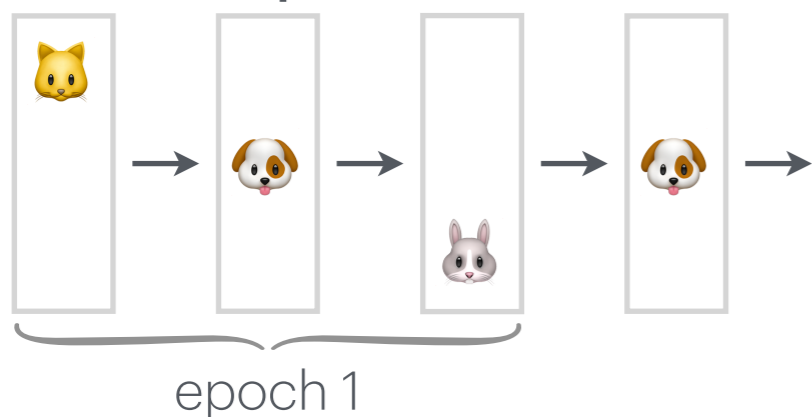
online SGD \preceq GD



high-dim; related to benign overfitting

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

multi-epoch SGD (without replacement) dominates GD



the SGD variant
used in deep learning

Contribution 3: from theory to practice

principled parallelization method for training language models

- “Seesaw: accelerating training by balancing learning rate and batch size scheduling”

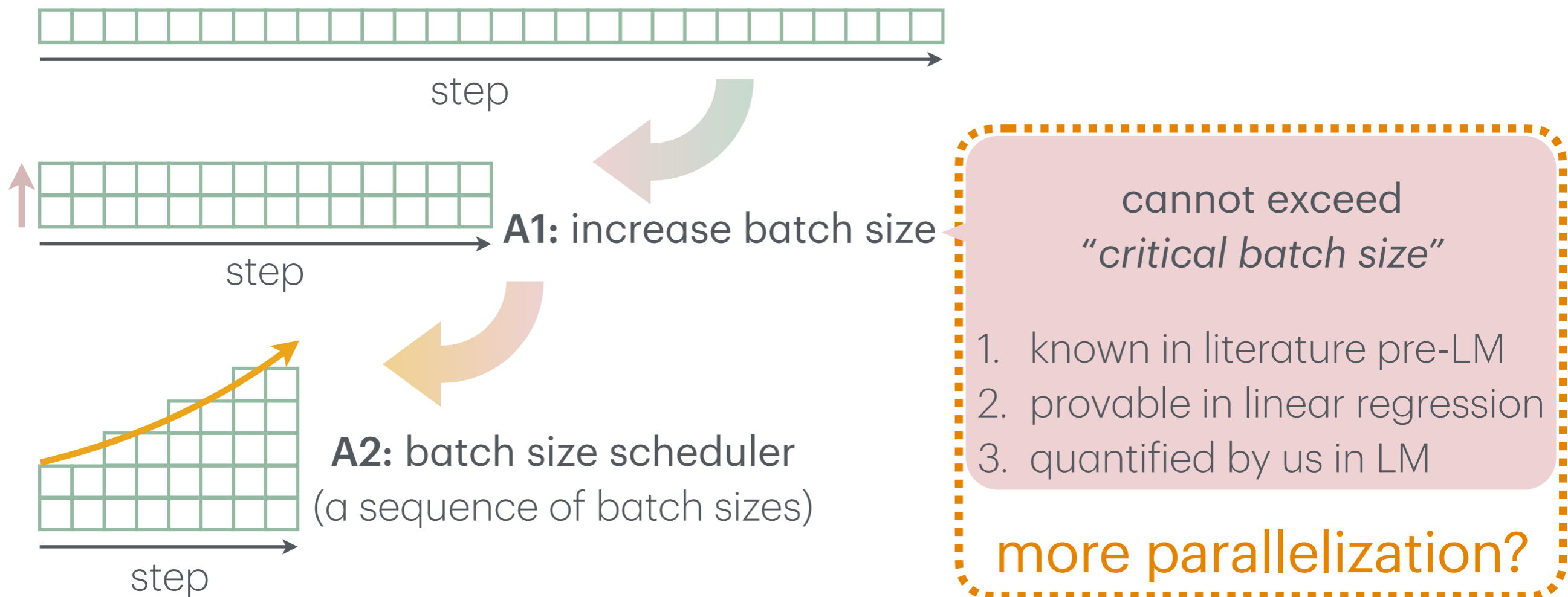
Alexandru Meterez*, Depen Morwani*, **W**, Costin-Andrei Oncescu, Cengiz Pehlevan, Sham Kakade

ICLR 2026

Language Model (LM) training

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

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



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?

sandbox:
normalized SGD

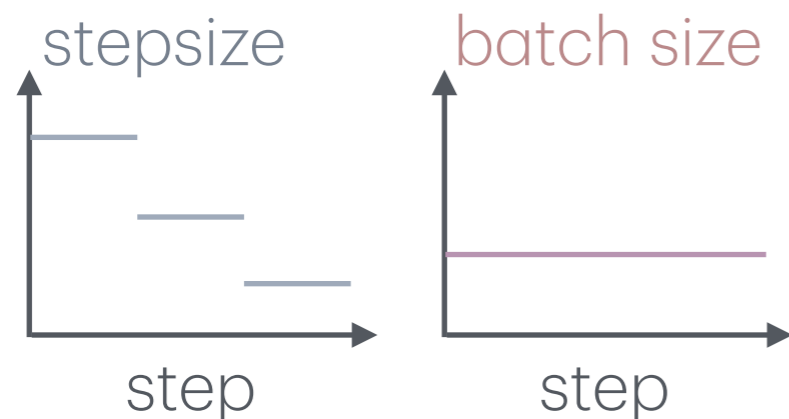
compatible
with Adam

works with
language models

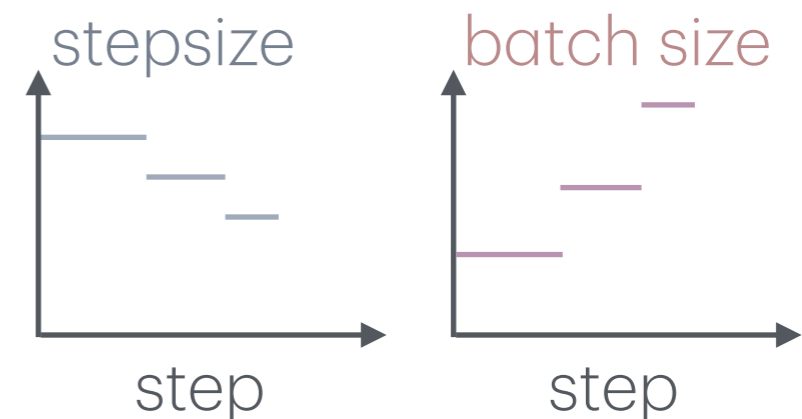
sandbox:
linear regression

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

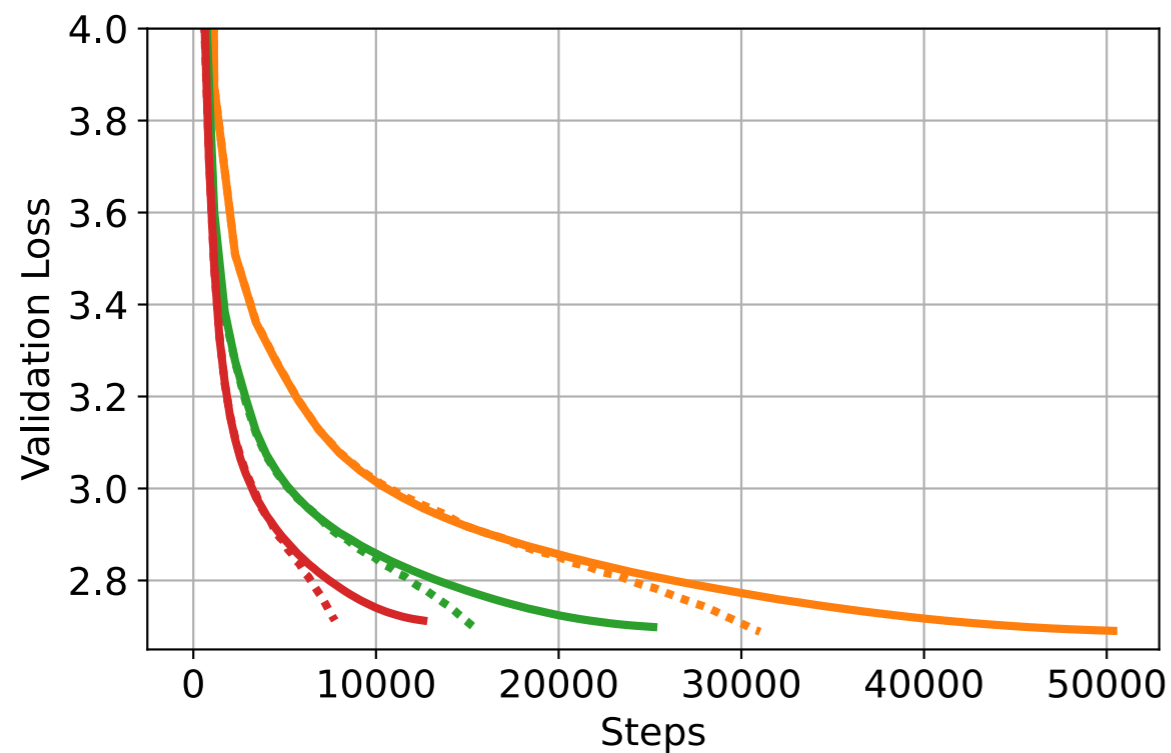
default: stepsize scheduler
 $\eta \rightarrow \eta/1.1, B$ fixed



Seesaw: joint scheduler
 $\eta \rightarrow \eta/\sqrt{1.1}, B \rightarrow 1.1B$




Same error ($\pm 0.17\%$), 36% fewer steps



transformer (600M), Adam, C4

Seesaw

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

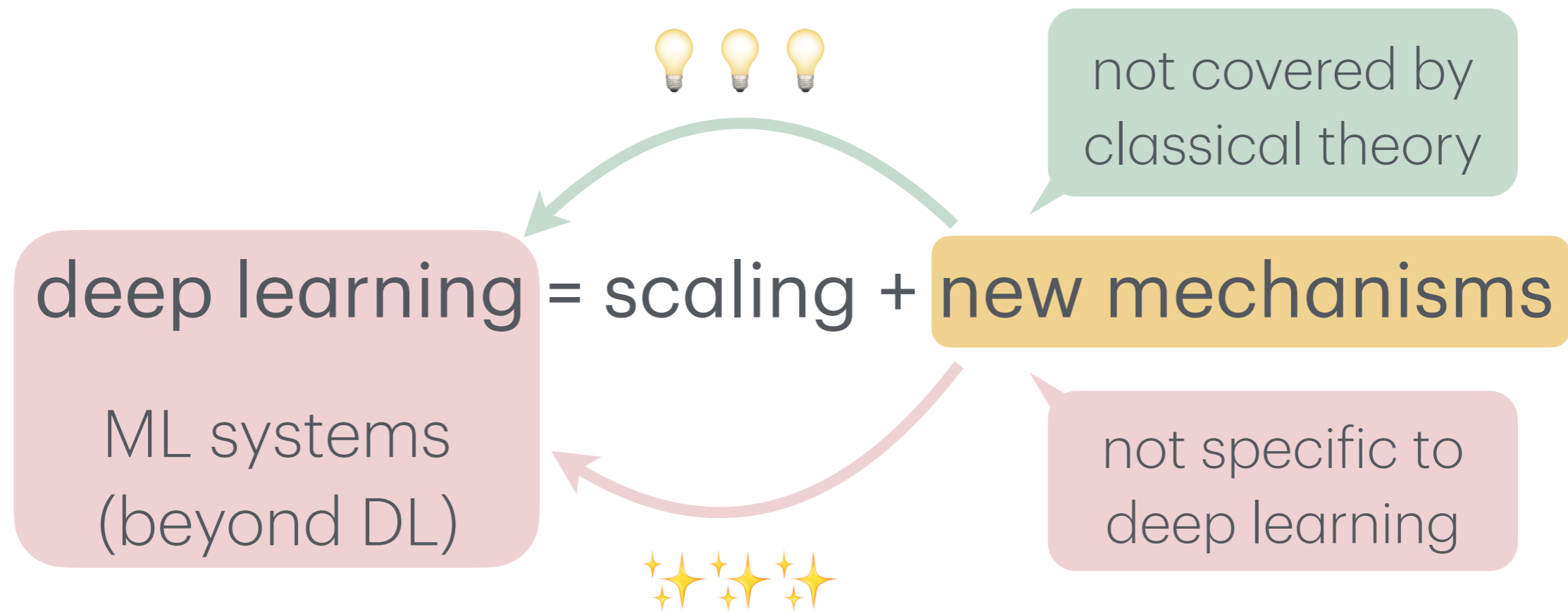
simple, meaningful sandbox
can be predictive!

initial batch size: 2^8 , 2^9 , 2^{10} (= CBS)

solid curve: default (fixed batch size, cosine stepsize scheduler)

dotted curve: Seesaw (ours)

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

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

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