New Insights about SGD

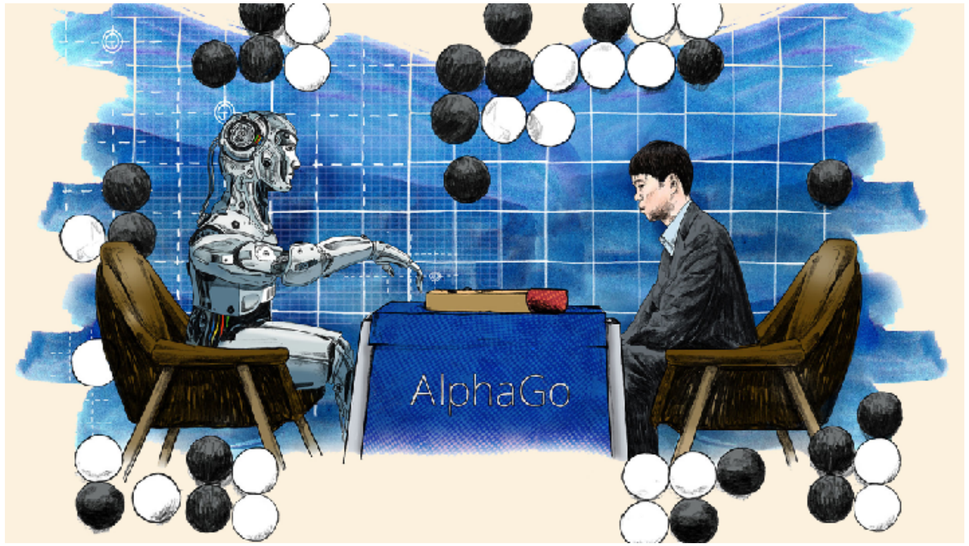
stepsize, risk convergence, and implicit regularization

Jingfeng Wu 10/2023

SIMONS INSTITUTE
for the Theory of Computing

Berkeley
UNIVERSITY OF CALIFORNIA

Modern machine learning

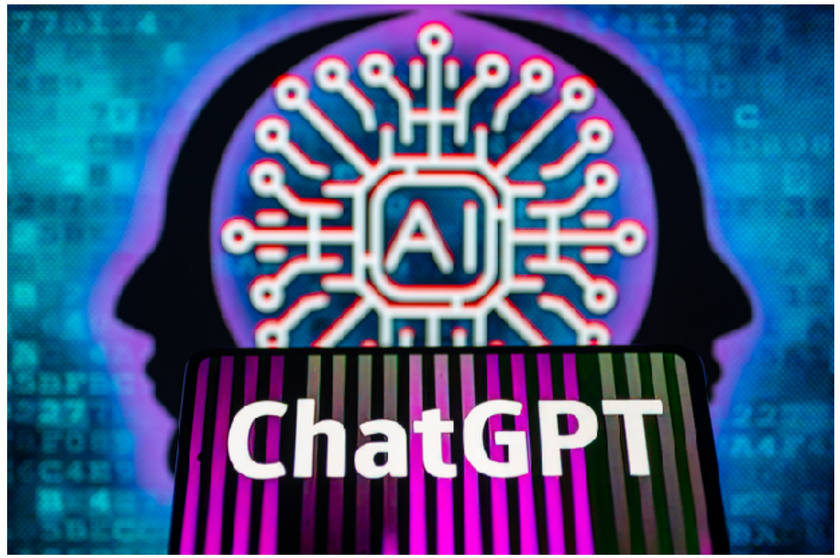


AlphaGo

Google assistant

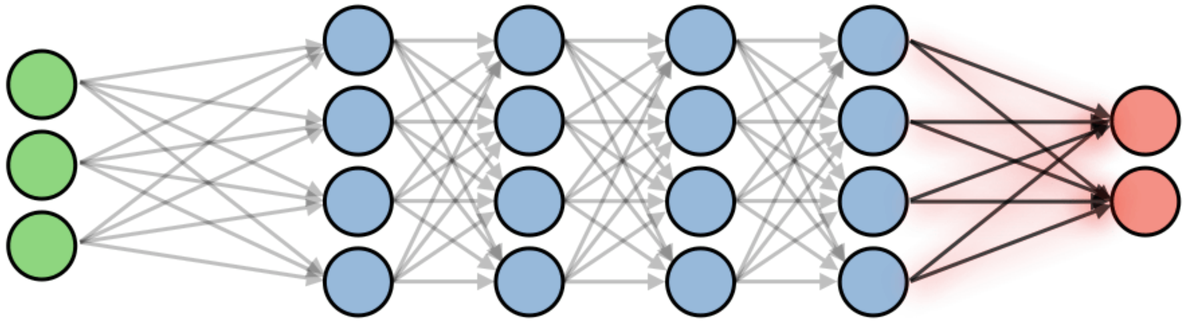
At their core is

deep learning



ChatGPT

Deep learning oversimplified



90% dog	100% dog
5% cat	0% cat
2% glasses	0% glasses
...	...

data x

model $f(\cdot)$
(with parameter w)

prediction
$f(x)$

label y

Repeat for TONs of (x, y) + some tips/tricks/tuning

Current DL practice

"Alchemy", largely trial & error

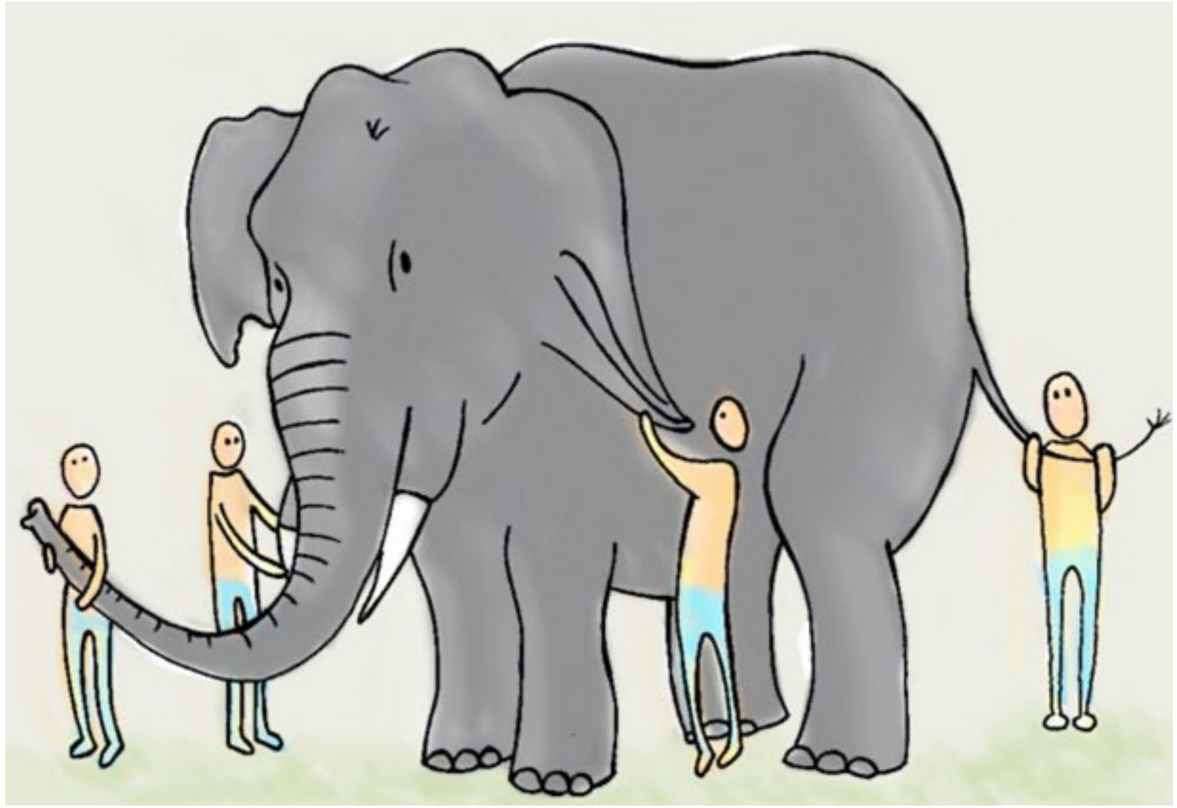
Usual formula:

1. one graduate student
2. one model
3. performance numbers on a few standard test sets
4. yay. we. rock.
5. one Ph.D.

"This is Graduate Student Descent"

https://www.slideshare.net/npinto/highperformance-computing-needs-machine-learning-and-vice-versa-nips-2011-big-learning

Current DL theory



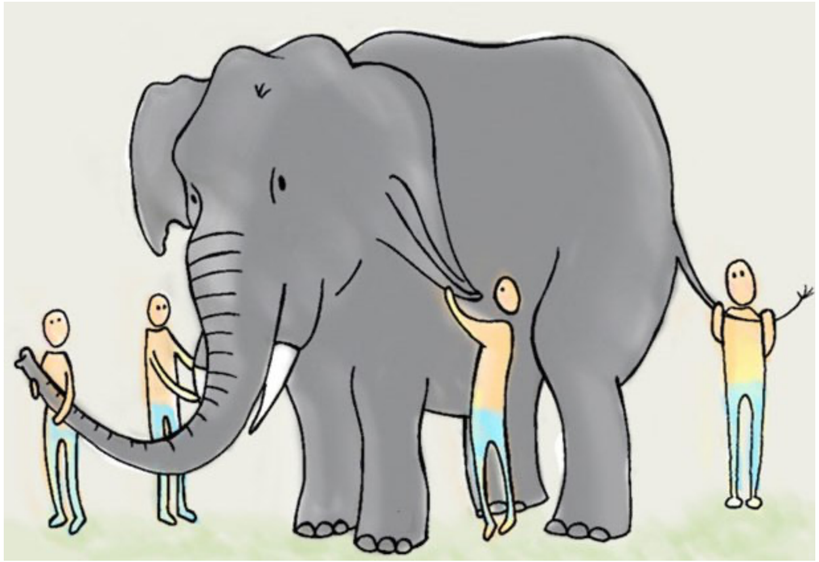
blind men and an elephant

- parameters: universal approximation, …
- random initialization: NTK, …
- low training error: benign overfitting, …
- SGD: implicit bias, …
- and many more

Would theory, from an incomplete picture, be useful?

This talk: better theory, better trial & error

only got two grad students
what to play with



the incomplete theory

all sorts of applications

its a feature not a bug
can try that

Part 1. Stepsize

(Stochastic) Gradient Descent

gradient descent

Cauchy, 1847

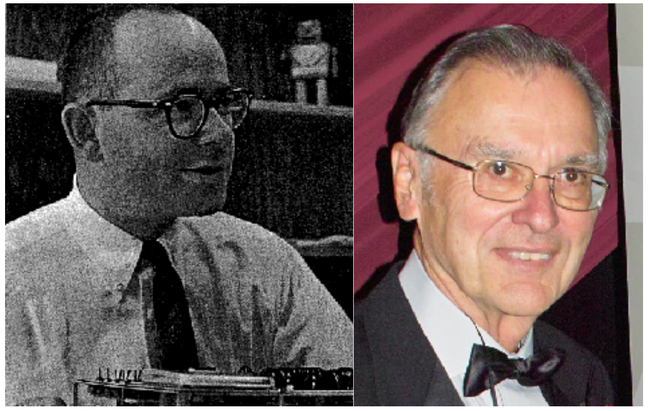
$$w_+ = w - \eta \cdot \nabla \ell(w)$$

making how much update?
AKA., stepsize / learning rate?

backpropagation

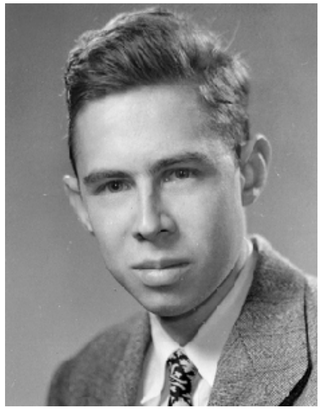
Werbos, 1974

adaptive linear neuron



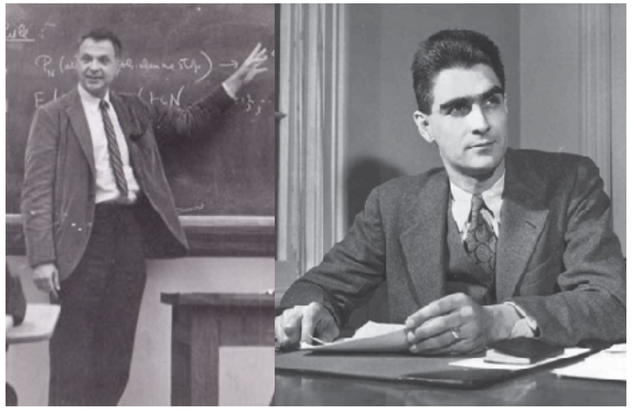
Widow & Hoff, 1960

perceptron



Rosenblatt, 1958

stochastic approximation



Robbins & Monro, 1951

Optimization theory oversimplified

$$\ell(w_+) = \ell(w - \eta \cdot \nabla \ell(w))$$

$$= \ell(w) - \eta \cdot \|\nabla \ell(w)\|^2 + \frac{\eta^2}{2} \cdot \nabla \ell(w)^\top \cdot \nabla^2 \ell(w) \cdot \nabla \ell(w) - O(\eta^3)$$

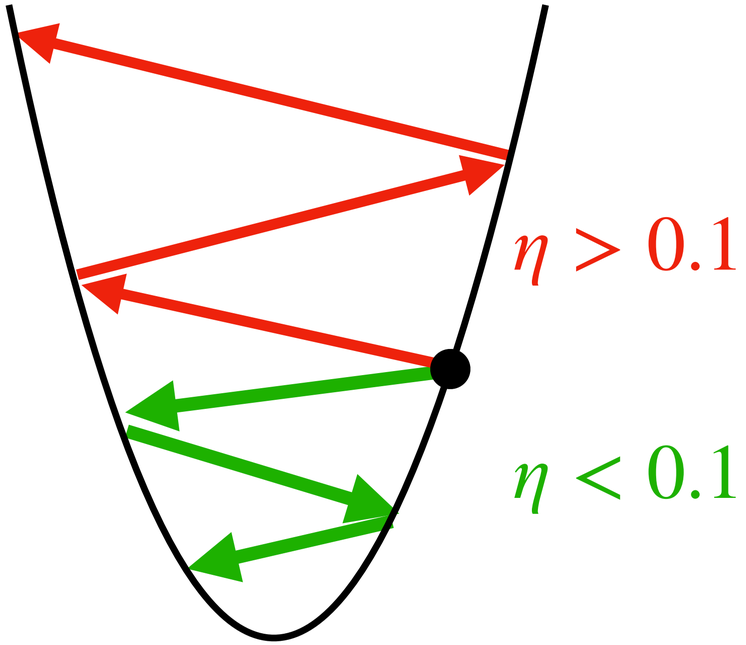
$$\leq \ell(w) - \eta \cdot \left(1 - \frac{\eta}{2} \cdot \|\nabla^2 \ell(w)\|_2\right) \cdot \|\nabla \ell(w)\|^2 - O(\eta^3)$$

[descent lemma]

For **small** η, $\ell(w_t)$ decreases **monotonically**, GD works

For large η, GD does not work for quadratics

quadratic landscape

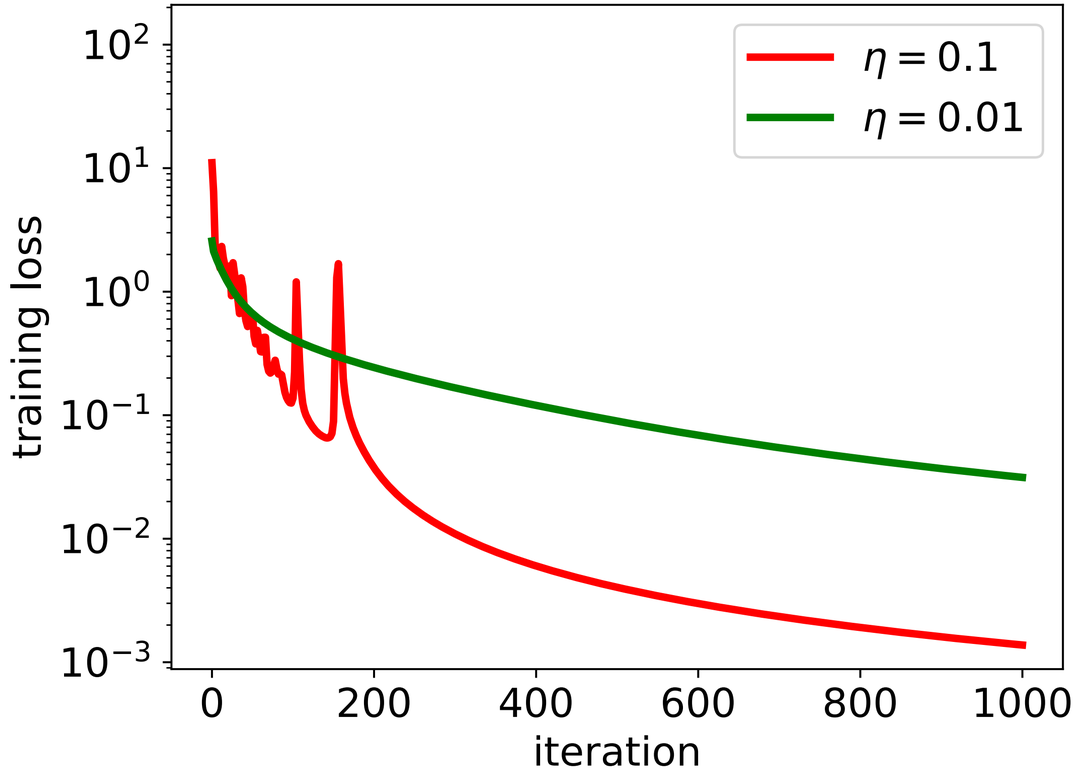


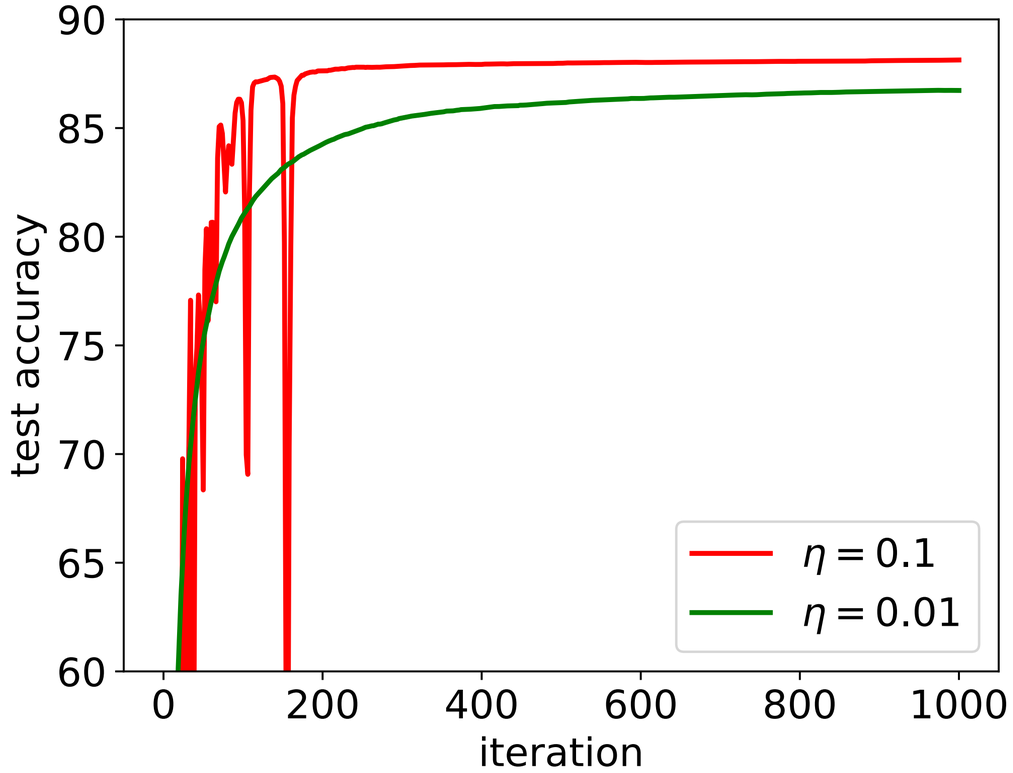
$$\ell(w) = 10w^2$$

$$w \leftarrow w - \eta \cdot 20w$$

Numbers in DL (classification problem)

3-layer net + 1,000 samples from MNIST+ GD with const-stepsize



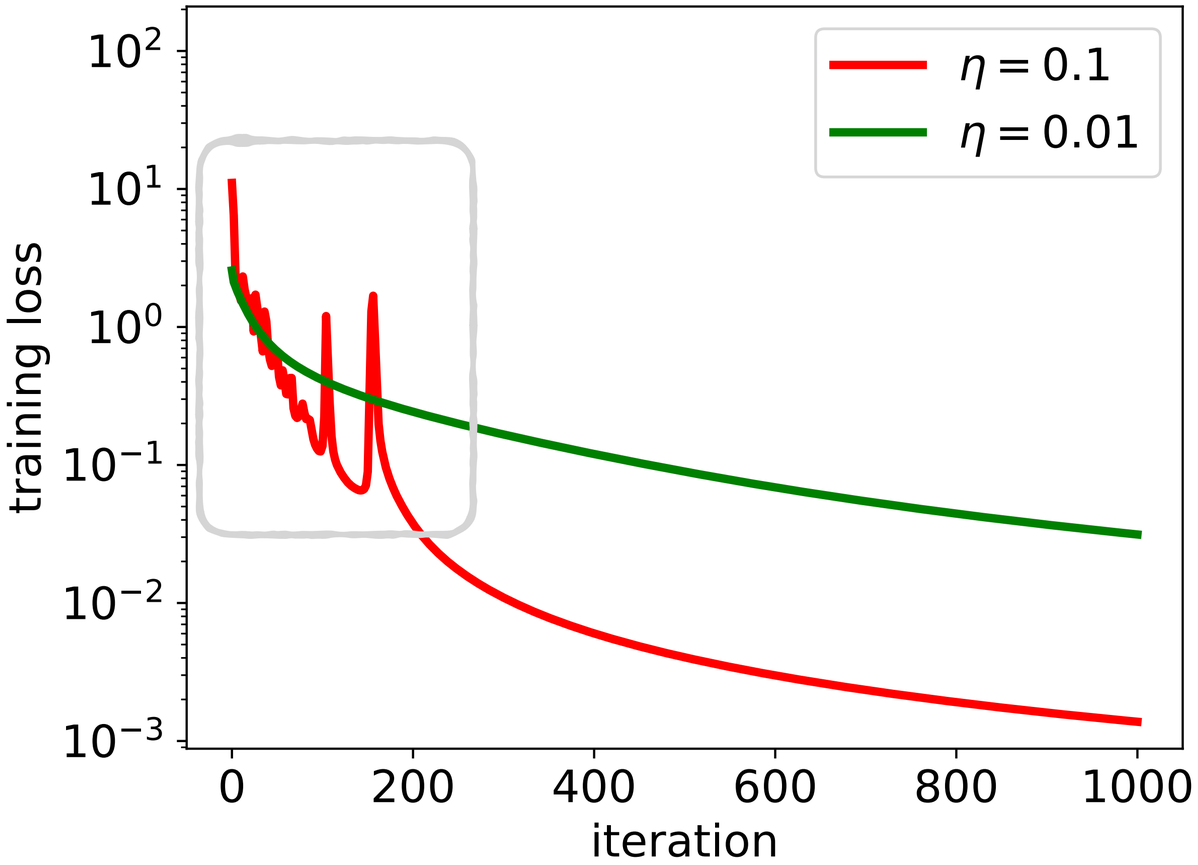


small stepsize works;

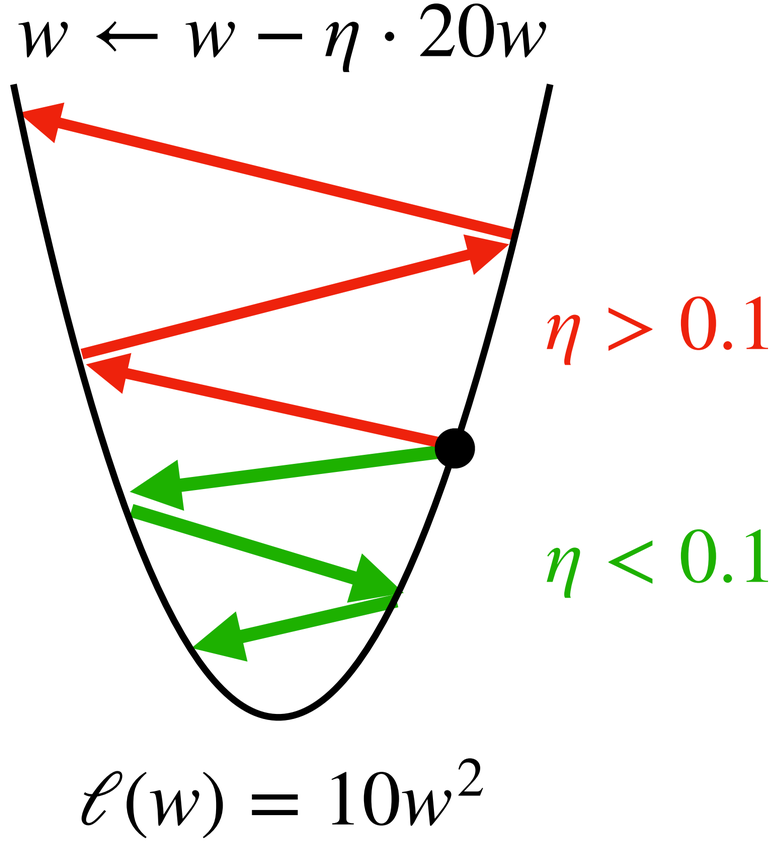
large stepsize also works and works better in the long run

Limitations of the old picture

DL classification (cross-entropy)



opt theory (quadratic)



old theory picture predicts green curve, but fails predicting red curve

Logistic regression, revisited

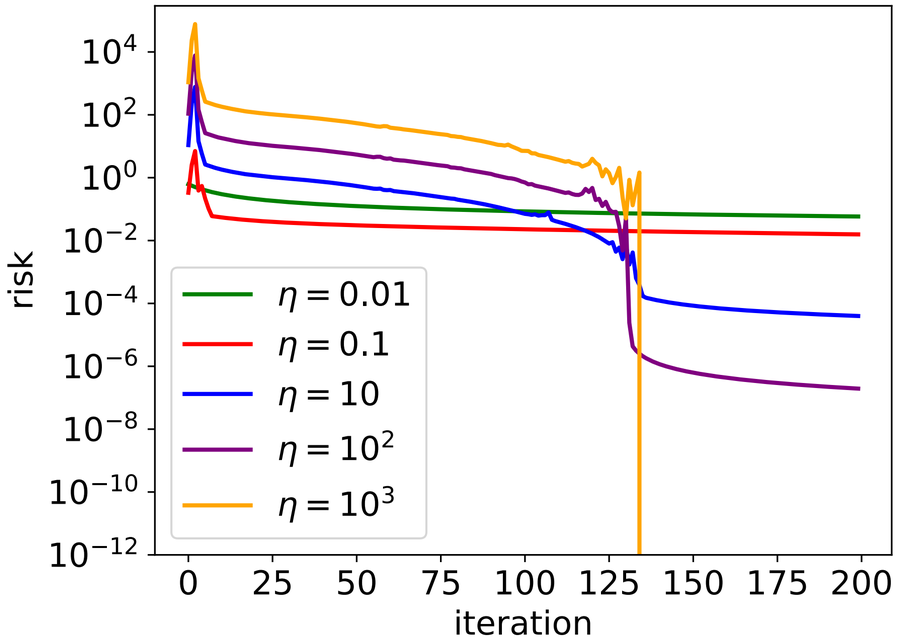
[WBL'23] In the NN training (minimizing a cross-entropy loss), if

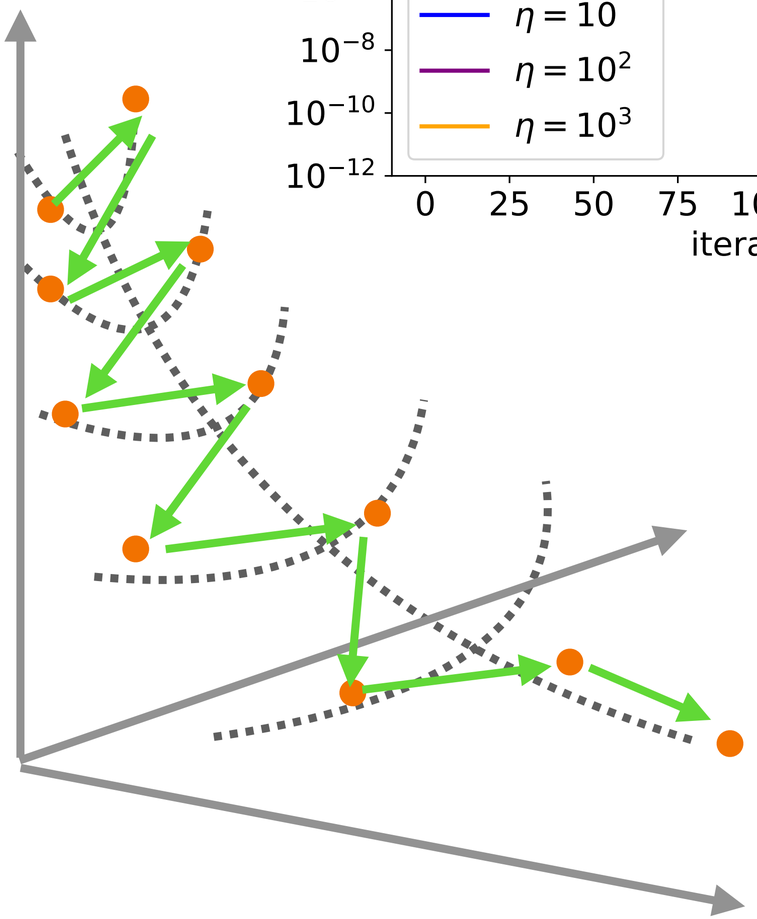
- the model is linear, i.e., $f(x) = x^\top w$,
- the datasets can be perfectly fitted,

Then GD, with **any constant stepsize**, minimizes the fitting error.

[open/ongoing]

Benefits of large stepsize?

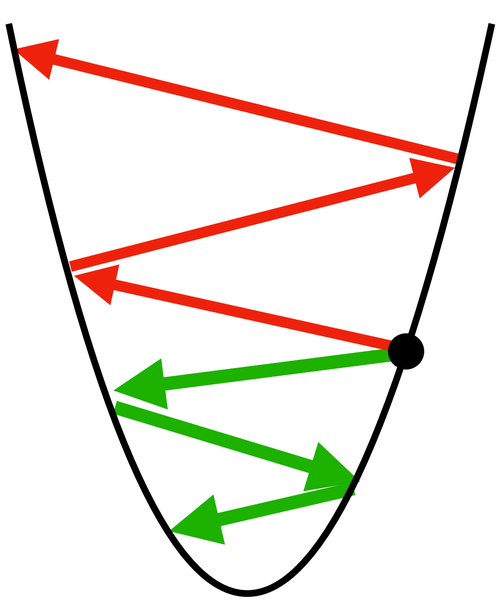


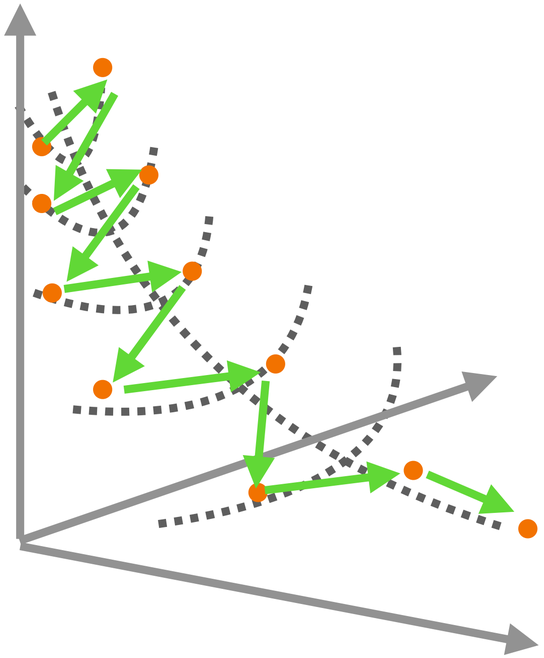


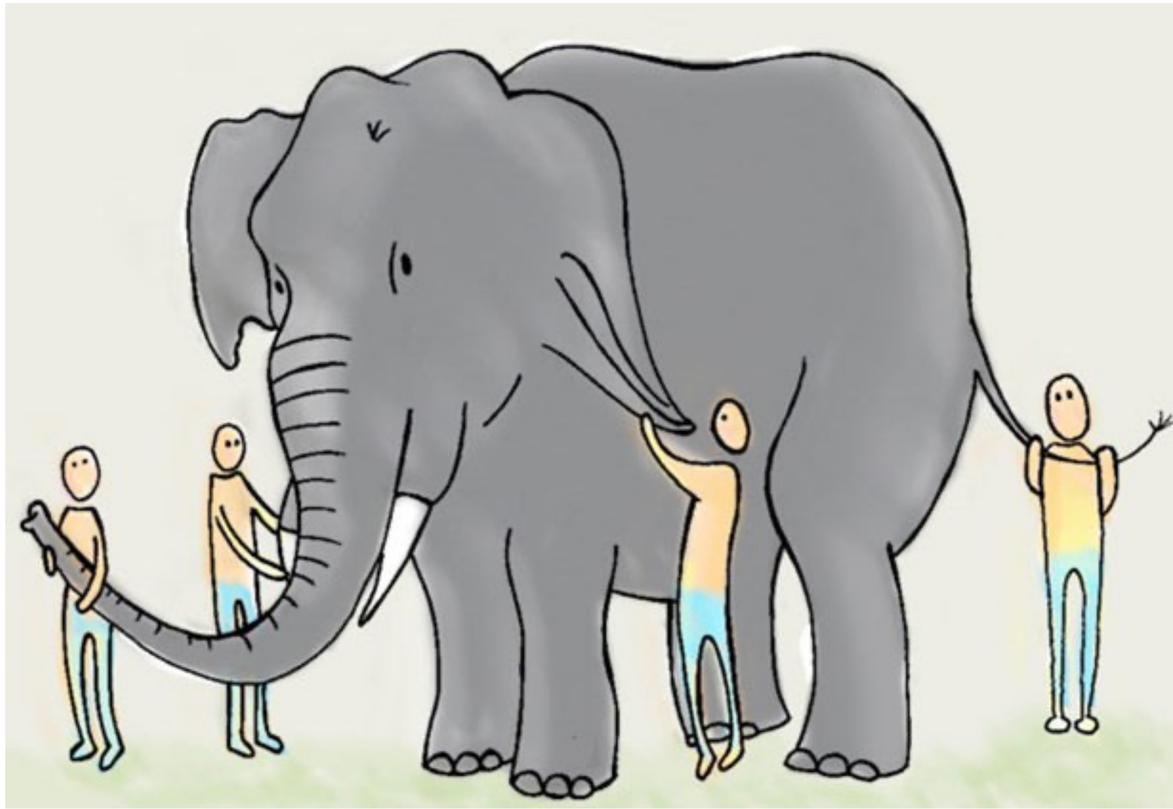
Wu, Jingfeng, Vladimir Braverman, and Jason D. Lee. "Implicit Bias of Gradient Descent for Logistic Regression at the Edge of Stability." NeurIPS 2023

Rethinking optimization theory

- in DL (classification), Taylor approx. is limited (b/c its local)
- “valley” is more stable than “quadratic basin”
- descent lemma isn’t a golden rule (more likely a bad rule)







Part 2. Implicit Regularization

Statistical learning theory oversimplified

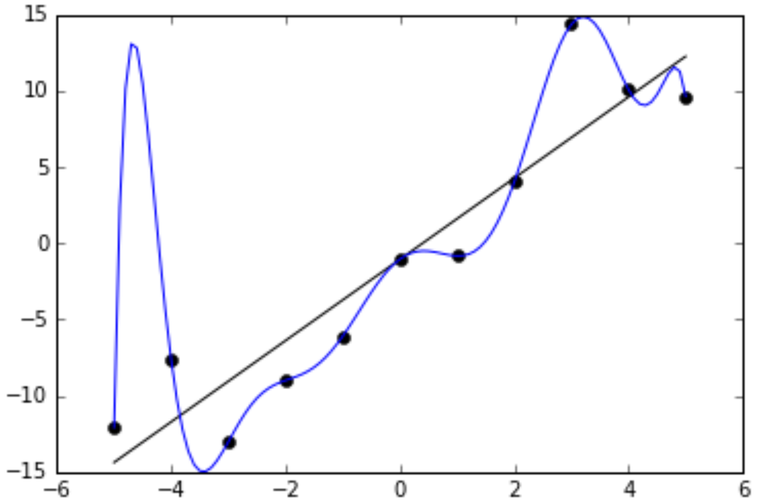
$$R(w) - R(w^*) = R(w) - \hat{R}_n(w) + \hat{R}_n(w) - \hat{R}_n(w^*) + \hat{R}_n(w^*) - R(w^*)$$
$$\leq \texttt{Empirical} + 2 \cdot \sup_{w \in \mathscr{H}} |R(w) - \hat{R}_n(w)|$$

1. Generalization: empirical fit vs. complexity control

1 is (always) a good advice, 2 is debatable

$$\sup_{w \in \mathscr{H}} |R(w) - \hat{R}_n(w)| \lesssim \sqrt{\frac{\log |\mathscr{H}|}{n}}$$

2. #Param (uniformly) controls complexity



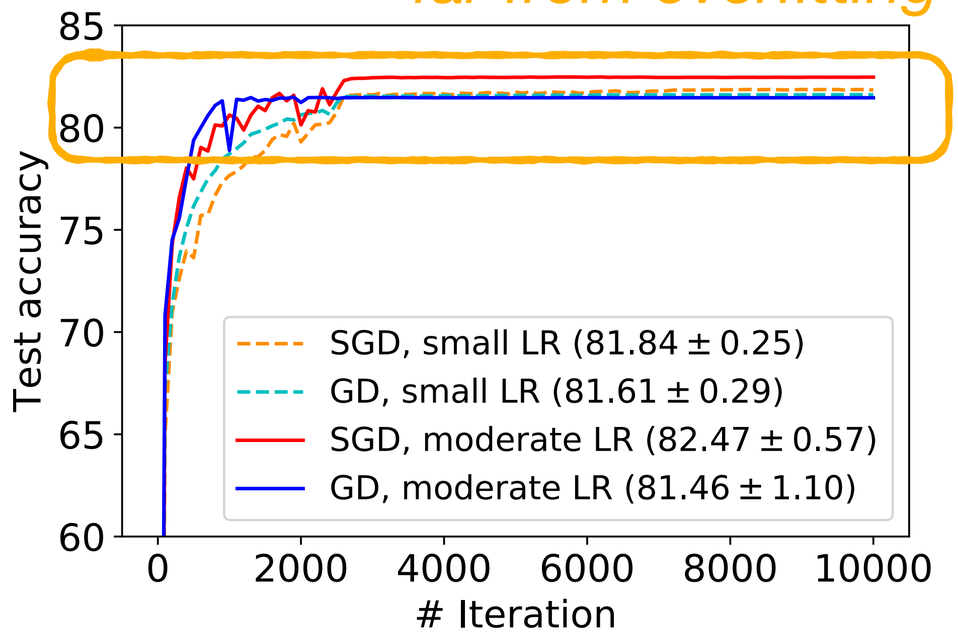
complexity uncontrolled
=> overfitting

What controls complexity in DL?

1. hypothesis class size
   - #params
2. explicit regularization
   - weight decay, dropout,…
3. implicit regularization
   - SGD, a simple learning rule

How does SGD regularize the model?

far from overfitting



*NN with **11,330** parameters*

*fitting **2,000** samples from FashionMNIST*

***without** explicit regularization*

Wu, Jingfeng, Difan Zou, Vladimir Braverman, and Quanquan Gu. "Direction Matters: On the Implicit Bias of Stochastic Gradient Descent with Moderate Learning Rate." ICLR 2021

Linear / ReLU regression, revisited

$$\text{minimize} \quad R(\mathbf{w}) = \mathbb{E}\left(\phi(\mathbf{x}^\top \mathbf{w}) - y\right)^2, \quad \mathbf{w} \in \mathbb{R}^d$$

$$\phi = \texttt{id} \text{ or } \max\{\,\cdot\,, 0\}$$

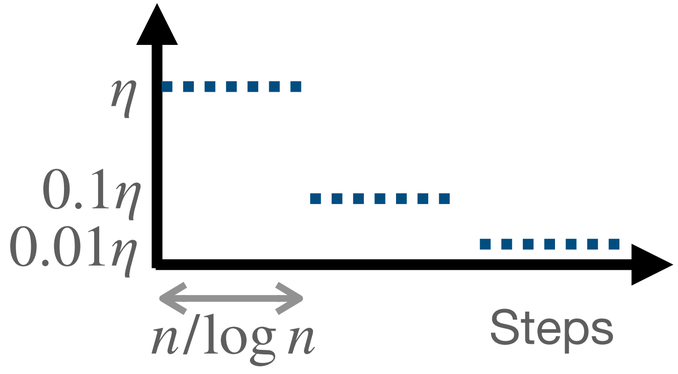
with n iid samples: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$

Distributional assumption (simplified)

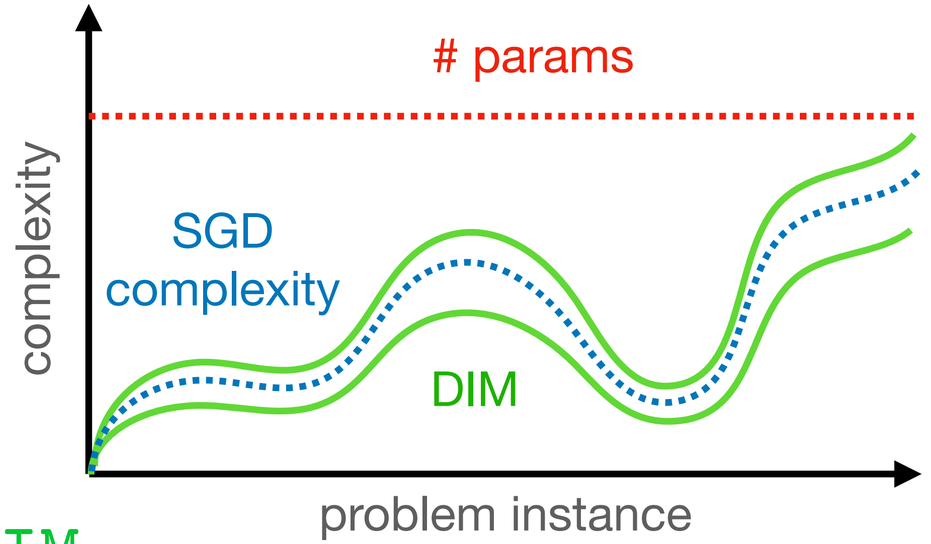
$$y = \phi(\mathbf{x}^\top \mathbf{w}_*) + \mathcal{N}(0,1), \quad \|\mathbf{w}_*\|_2 \le 1, \quad \mathbf{x} \sim \mathcal{N}(0, \mathbf{H})$$

SGD / Perceptron

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t \cdot \left(\phi(\mathbf{x}_t^\top \mathbf{w}_{t-1}) - y_t\right) \cdot \mathbf{x}_t, \quad t = 1, \ldots, n$$



Implicit complexity control



[ZWBGK'21, WZBGK'22, WZCBGK'23]

$$\texttt{Bias} + \frac{\texttt{DIM}}{n} \lesssim \mathbb{E}R(\mathbf{w}_n) - \min R \lesssim \texttt{Bias} + \frac{\texttt{DIM}}{n}$$

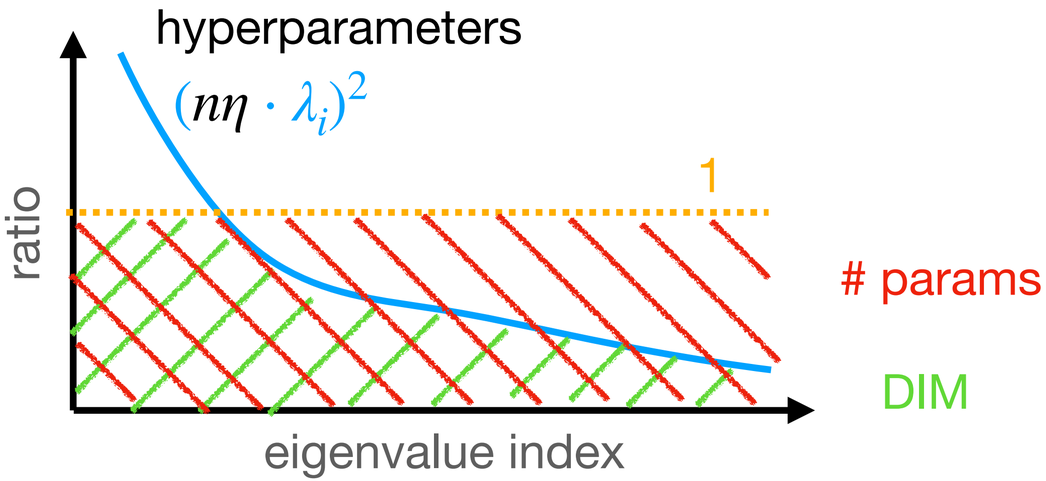
$\texttt{DIM} \leq d$, a function of n, η, and $\mathbf{H}$, controls the complexity

- Zou, Difan, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham Kakade. "Benign overfitting of constant-stepsize sgd for linear regression." COLT 2021.
- Wu, Jingfeng, Difan Zou, Vladimir Braverman, Quanquan Gu, and Sham Kakade. "Last iterate risk bounds of sgd with decaying stepsize for overparameterized linear regression." ICML 2022
- Wu, Jingfeng, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. "Finite-Sample Analysis of Learning High-Dimensional Single ReLU Neuron." ICML 2023

Implicit complexity control

$$\texttt{DIM} := \#\left\{ \lambda_i \geq \frac{1}{n\eta} \right\} + n^2\eta^2 \cdot \sum_{\lambda_i < \frac{1}{n\eta}} \lambda_i^2$$

$(\lambda_i)_{i\geq 1}$ denote eigenvalues of $\mathbf{H} = \mathbb{E}[\mathbf{x}\mathbf{x}^\top]$



DIM is small when eigenvalues decay fast

SGD exploits a “low-dim” structure in data

Excess Risk = Bias + Noise * Dim / Steps

	SGD	Ridge Regression [TB 2020]
Bias	$\left\| \prod_{t=1}^{n} (\mathbf{I} - \eta_t \mathbf{H}) \mathbf{w}^* \right\|_{\mathbf{H}}^2$	$\frac{(\lambda + \sum_{i>k^*} \lambda_i)^2}{n^2} \cdot \|\mathbf{w}^*\|^2_{\mathbf{H}^{-1}_{0:k^*}} + \|\mathbf{w}^*\|^2_{\mathbf{H}_{k^*:\infty}}$
Noise	σ^2	σ^2
"Dim"	$k^* + \eta^2 K^2 \sum_{i>k^*} \lambda_i^2$	$k^* + \frac{n^2}{(\lambda + \sum_{i>k^*} \lambda_i)^2} \sum_{i>k^*} \lambda_i^2$
"Steps"	$K = n/\log n$	n
Optimal k^*	$\lambda_1 \geq \ldots \geq \lambda_{k^*} \approx \frac{1}{\eta K} \geq \ldots$	$\lambda_1 \geq \ldots \geq \lambda_{k^*} \approx \frac{\lambda + \sum_{i>k^*} \lambda_i}{n} \geq \ldots$

1. ridge with large λ <=> SGD with small η

2. SGD has smaller bias in the head

Implicit vs. explicit regularization

[ZWBGFK'21] Fix a set of "natural" least square problems (opt is not hard & label is noisy).

Assume an oracle tunes both SGD and ridge to their best on each problem instance. Let m and n be their sample complexity to attain the same rate.

- For **every** problem instance: $m \lesssim n \log^2(n)$
- There **exist** a problem instance: $n \geq m^2/\log^4(m)$

tuned SGD inflates **at most poly-log** more samples than tuned ridge

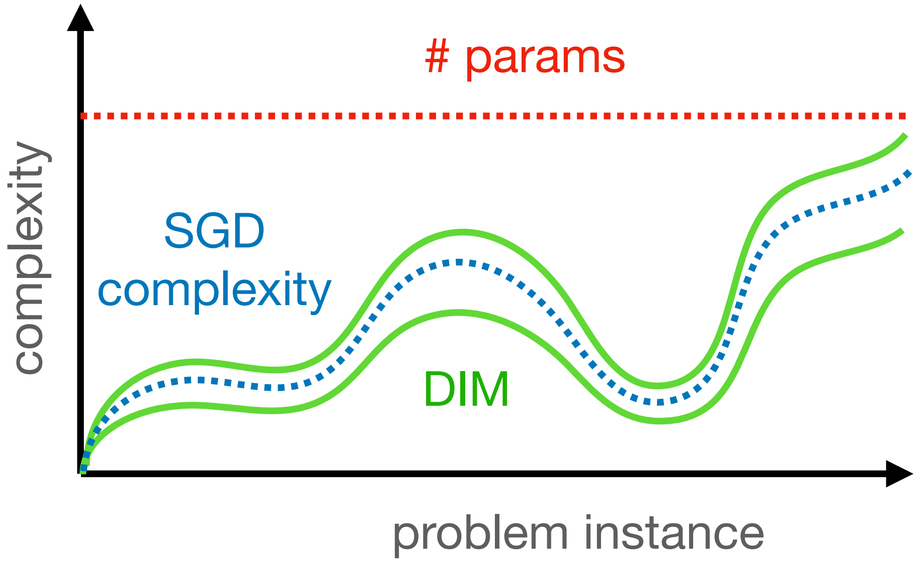
tuned ridge can inflate **poly** more samples than tuned SGD

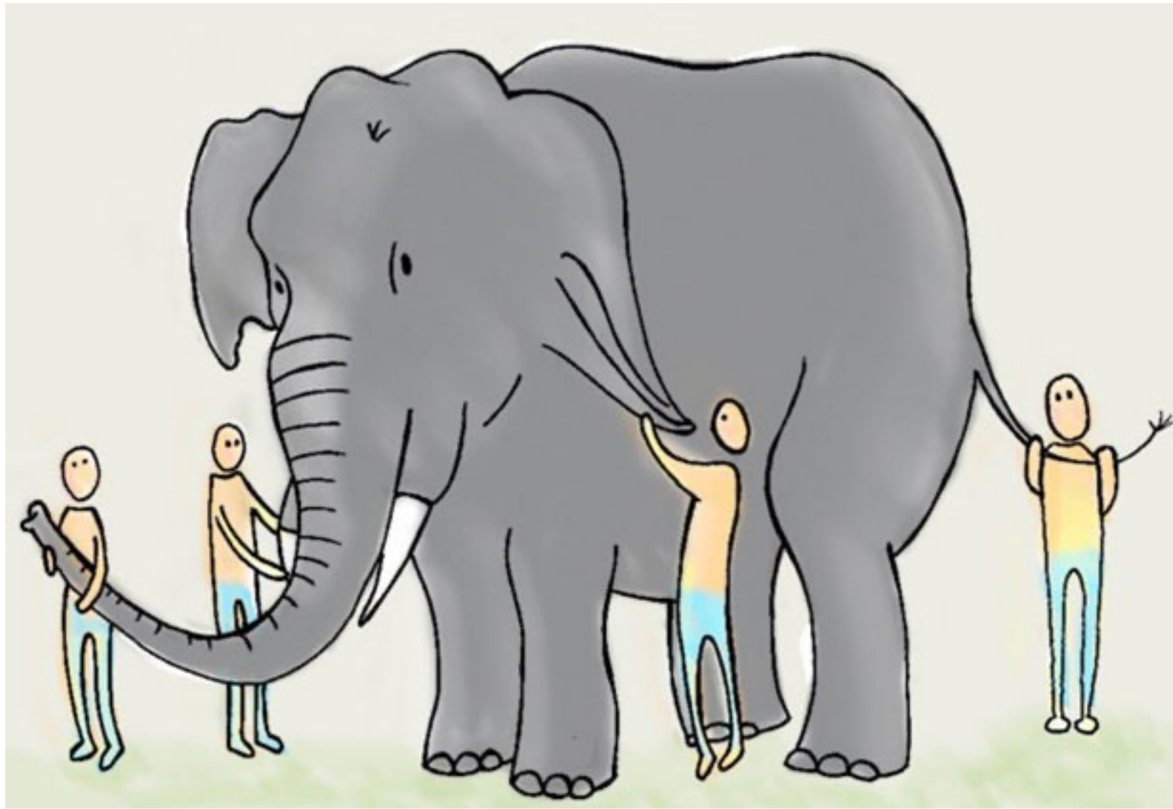
tune SGD before tuning weight decay :)

Zou, Difan, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P. Foster, and Sham Kakade. "The benefits of implicit regularization from sgd in least squares problems." NeurIPS 2021

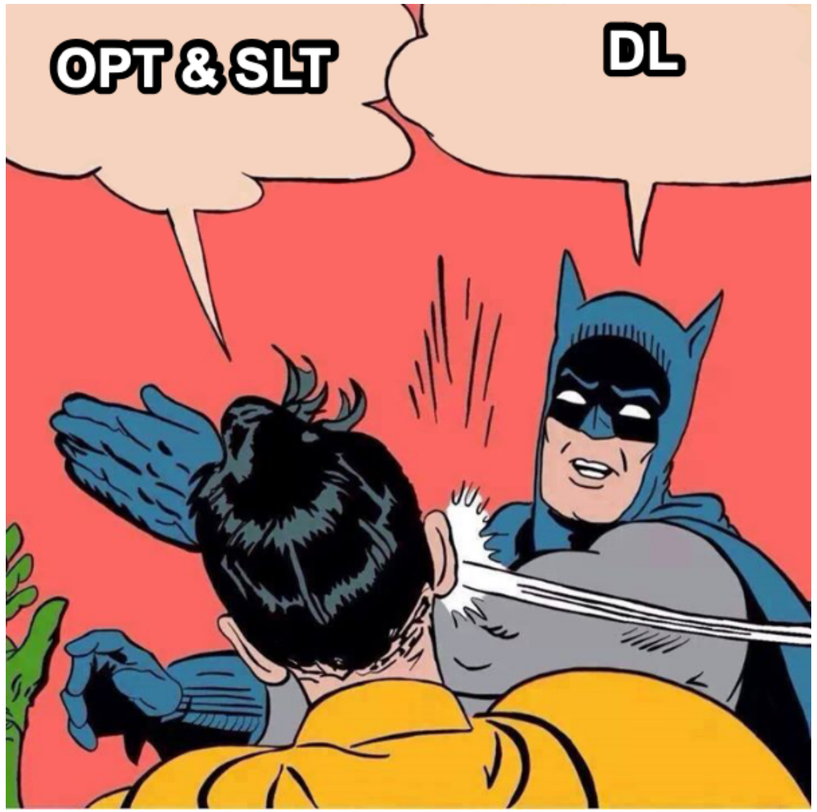
Rethinking statistical learning theory

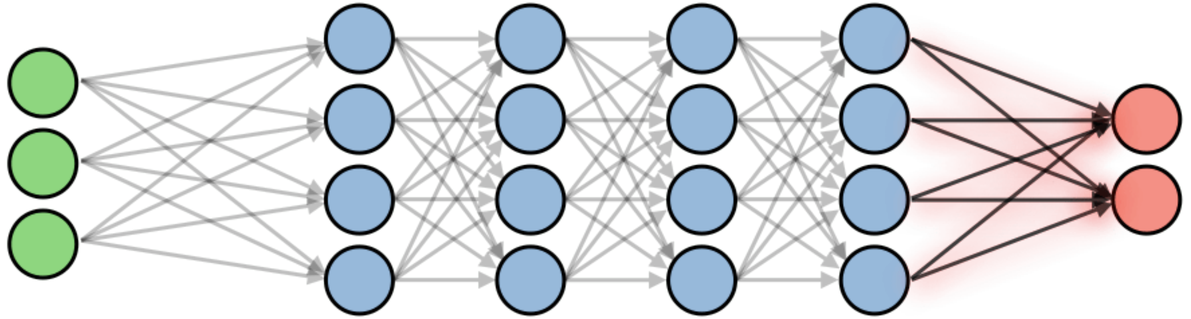
- in DL, complexity isn't controlled by #params
- complexity is largely controlled by simple learning rules
- stepsize balances empirical fit and complexity control





Conclusions





- **optimization** and **statistical learning theory** are still insightful
- but need revisions in **deep learning**
- large stepsize, implicit regularization
- **more effective trial & error** if things are understood better

my website