

Implicit Regularization

A Statistical View

Jingfeng Wu



SIMONS
INSTITUTE
for the Theory of Computing

Berkeley
UNIVERSITY OF CALIFORNIA

Machine learning

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

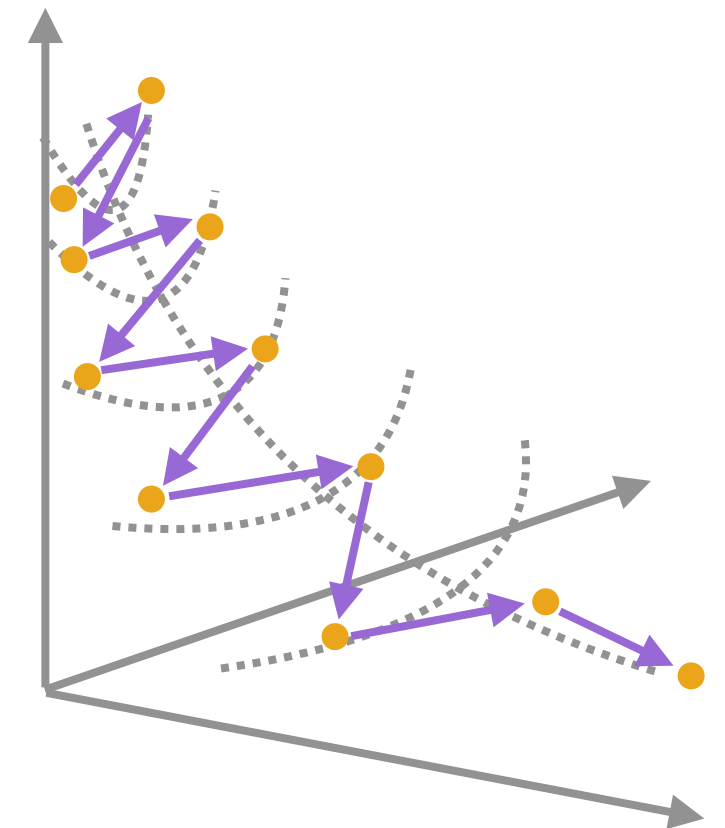
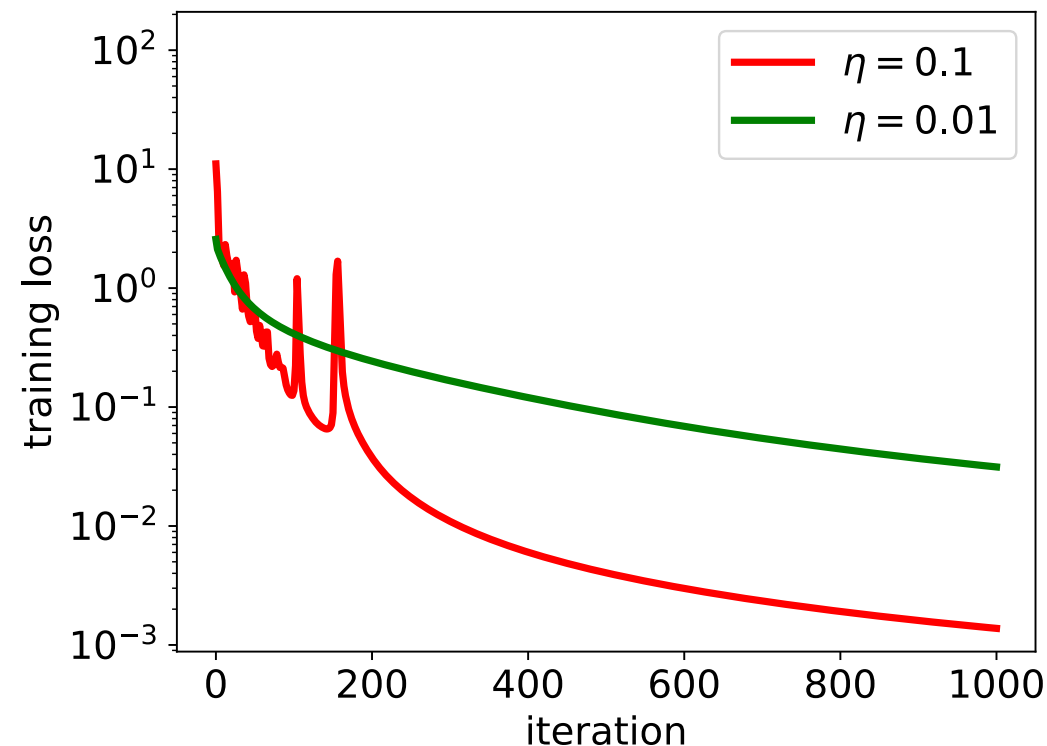
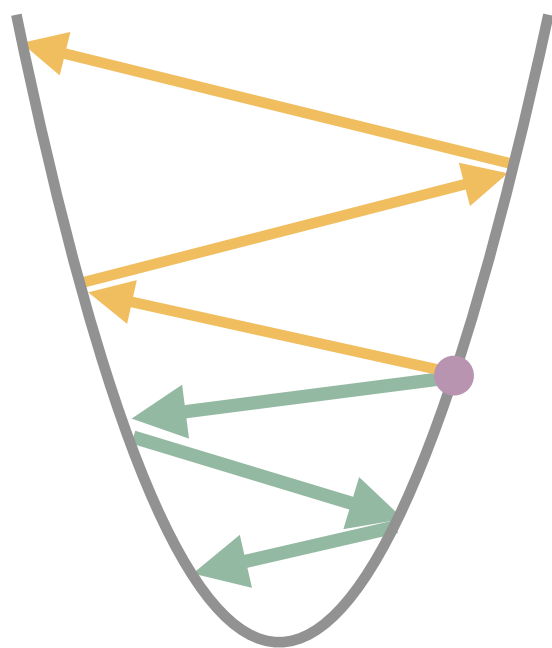
- optimization \leq gradient methods
- generalization \leq complexity control

Machine learning

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

- optimization \leq gradient methods

past work: large stepsize accelerates GD for logistic regression



Machine learning

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

- optimization \leq gradient methods
- generalization \leq complexity control

this talk: generalization, done together with optimization

Complexity control

classical answer: **explicit control**

- model family
- norm regularization
- ...

deep learning: **implicit controls via opt algo**

- early stopping
- stochastic averaging
- ...

how good is implicit regularization?

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

One of our results

For all Gaussian linear regression problems:

early stopping is

- always no worse
- sometimes much better

than ℓ_2 -regularization.

Our approach

Instance-wise risk comparison ← instead of minimax

- GD vs ridge regression

← high dimension

- GD vs (online) SGD



Peter Bartlett



Jason Lee



Sham Kakade



Bin Yu

Wu, Bartlett*, Lee*, Kakade*, Yu*. "Risk comparisons in linear regression: implicit regularization dominates explicit regularization." arXiv 2025

Linear regression

finite signal-to-noise ratio



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

problem determined by (Σ, w^*)

excess risk / prediction error

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

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

ridge regression

hyperparameter: $\lambda \geq 0$

$$\begin{aligned} w_{\lambda}^{\text{ridge}} &= \arg \min \frac{1}{n} \sum_{i=1}^n \|x_i^{\top} w - y_i\|^2 + \lambda \|w\|^2 \\ &= (X^{\top} X + n\lambda I)^{-1} X^{\top} Y \end{aligned}$$

gradient descent

hyperparameter: $t \geq 0$

- $w_0 = 0$
- for $s = 1, \dots, t$,

$$w_s = w_{s-1} - \frac{\eta}{n} X^{\top} (X w_{s-1} - Y)$$

- $w_t^{\text{gd}} = w_t$

Notation

- SVD

$$\Sigma = \sum_{i \geq 1} \lambda_i u_i u_i^\top \quad \lambda_1 \geq \lambda_2 \geq \dots$$

- head and tail divided by k

$$\Sigma_{0:k} = \sum_{i \leq k} \lambda_i u_i u_i^\top \quad \Sigma_{k:\infty} = \sum_{i > k} \lambda_i u_i u_i^\top$$

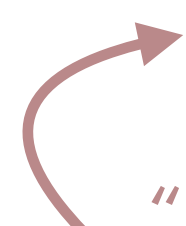
- matrix M , vector v

$$M^{-1} = \text{pseudoinverse of } M \quad \|v\|_M^2 = v^\top M v$$

Bounds for ridge

*possible to pin down constants via RMT

Theorem. For all $\lambda \geq 0$, in expectation


$$\mathbb{E}R(w_{\lambda}^{\text{ridge}}) \gtrsim \tilde{\lambda}^2 \|w^*\|_{\Sigma_{0:k^*}^{-1}}^2 + \|w^*\|_{\Sigma_{k^*:\infty}}^2 + \min \left\{ \frac{D}{n}, 1 \right\}$$

“ \mathbb{E} ” can be made “w.h.p.”

same upper bound holds w.h.p.

critical index

$$k^* = \min \left\{ k : \lambda + \frac{\sum_{i>k} \lambda_i}{n} \geq c\lambda_{k+1} \right\}$$

effective regularization

$$\tilde{\lambda} = \lambda + \frac{\sum_{i>k^*} \lambda_i}{n}$$

effective dimension

$$D = k^* + \frac{1}{\tilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2$$

A ridge-type bound for GD

Theorem [WBKLY'25]. For all $0 < \eta \lesssim 1/\text{tr}(\Sigma)$ and $t \geq 0$, w.h.p.

$$R(w_t^{\text{gd}}) \lesssim \tilde{\lambda}^2 \|w^*\|_{\Sigma_{0:k^*}^{-1}}^2 + \|w^*\|_{\Sigma_{k^*:\infty}}^2 + \frac{D}{n}$$

was $\min \left\{ \frac{D}{n}, 1 \right\}$

critical index

$$k^* = \min \left\{ k : \frac{1}{\eta t} + \frac{\sum_{i>k} \lambda_i}{n} \geq c \lambda_{k+1} \right\}$$

effective regularization

$$\tilde{\lambda} = \frac{1}{\eta t} + \frac{\sum_{i>k^*} \lambda_i}{n}$$

was λ

effective dimension

$$D = k^* + \frac{1}{\tilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2$$

GD is no worse than ridge.

Proof. If $D > n$, set $t = 0$; otherwise, set $t = 1/(\eta\lambda)$.

GD dominates ridge

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

Theorem [WBKLY'25]. For **every** Gaussian linear regression, $n \geq 1$, and $\lambda \geq 0$, there is t such that: w.h.p.

$$R(w_t^{\text{gd}}) \lesssim \mathbb{E}R(w_\lambda^{\text{ridge}})$$

Prior work. Assume an isotropic prior, $\mathbb{E}w^{*\otimes 2} \propto I$

$$\inf_\lambda \mathbb{E}R(w_\lambda^{\text{ridge}}) \leq \mathbb{E}R(w_t^{\text{gd}}) \leq 1.69 \mathbb{E}R(w_\lambda^{\text{ridge}})$$

next: GD can be much better than ridge

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

Power law class

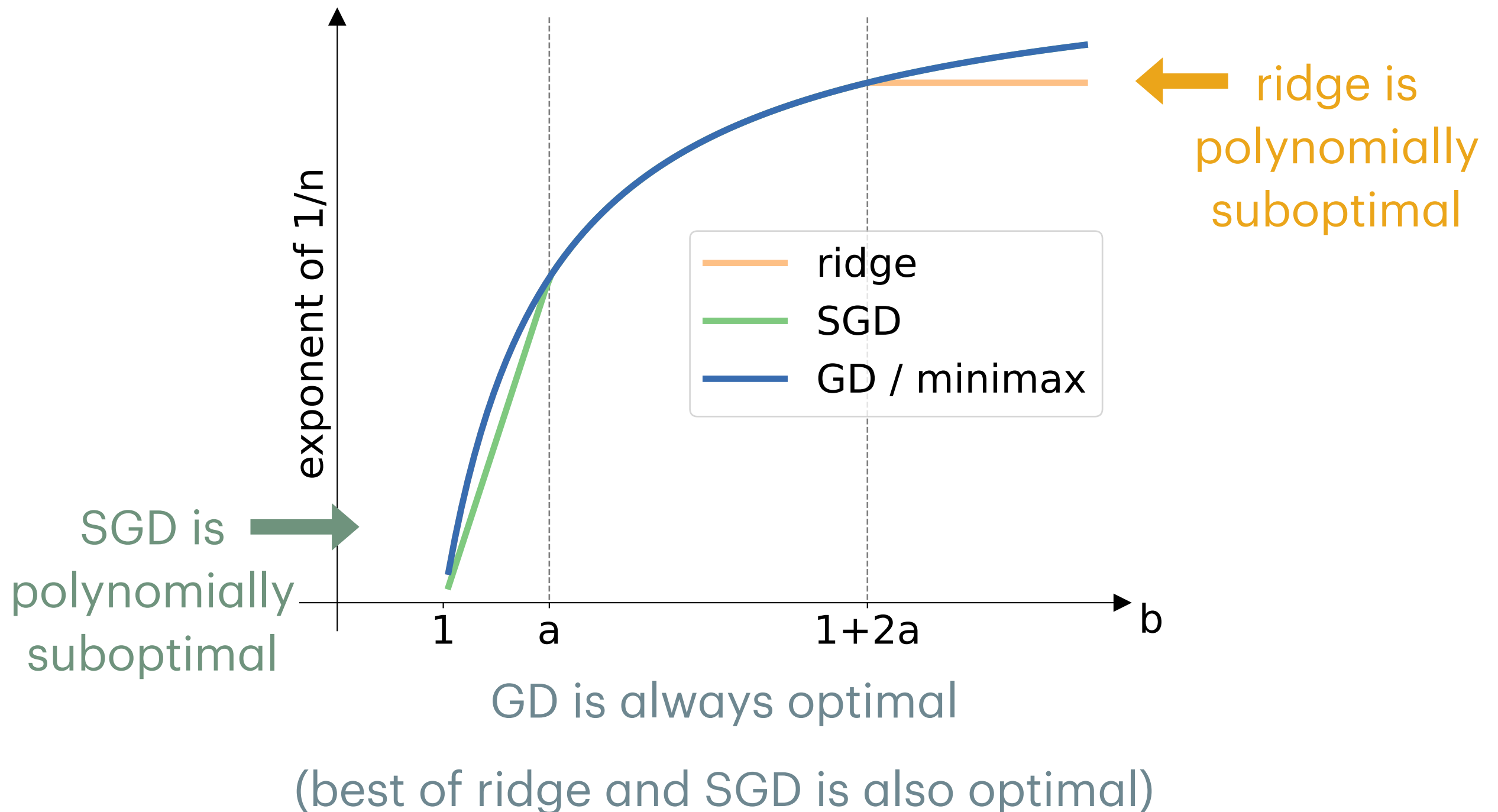
$$\lambda_i \approx i^{-a} \quad \lambda_i (u_i^\top w^*)^2 \approx i^{-b} \quad \text{for } a, b > 1$$

	$1 < b < a$	$a < b < 1 + 2a$	$b > 1 + 2a$
ridge	$O\left(n^{-\frac{b-1}{b}}\right)$		$\Omega\left(n^{-\frac{2a}{1+2a}}\right)$
SGD	$\tilde{\Omega}\left(n^{-\frac{b-1}{a}}\right)$	$\tilde{O}\left(n^{-\frac{b-1}{b}}\right)$	
GD	$O\left(n^{-\frac{b-1}{b}}\right)$		
minimax	$\Omega\left(n^{-\frac{b-1}{b}}\right)$		

GD is always optimal
 ridge/SGD is only partially optimal

Power law class

$$\lambda_i \approx i^{-a} \quad \lambda_i (u_i^\top w^*)^2 \approx i^{-b} \quad \text{for } a, b > 1$$



Results so far

GD dominates ridge

- always no worse
- sometimes much better

remark (computation)

multi-pass SGD (sample with replacement)

- multi-pass SGD is no better than GD
- with correct stepsizes, multi-pass SGD \approx GD

Why not known earlier?

fixed design is easy [DFKU'13, 6 pages]

but random design is hard

- instance-wise, not worst-case
- high-dim is surprising [BLLT'20, 44 pages]
- right tools 2019+

more surprise: GD vs (online) SGD

Dhillon, Foster, Kakade, Unga. “A risk comparison of ordinary least squares vs ridge regression.” *JMLR* 2013

Bartlett, Long, Lugosi, Tsigler. “Benign overfitting in linear regression.” *PNAS* 2020

Batch / online

gradient descent

- $w_0 = 0$
- for $s = 1, \dots, t$,
$$w_s = w_{s-1} - \frac{\eta}{n} X^\top (X w_{s-1} - Y)$$
- $w_t^{\text{gd}} = w_t$

hyperparameter: $t \geq 0$

stochastic gradient descent

- $w_0 = 0, \eta_0 = \eta, N = n/\log n$
- for $i = 1, \dots, n$,
$$\eta_i = \begin{cases} 0.1\eta_{i-1} & \text{if } i \% N = 0 \\ \eta_{i-1} & \text{else} \end{cases}$$
$$w_i = w_{i-1} - \eta_i (x_i^\top w_{i-1} - y_i) x_i$$
- $w_\eta^{\text{sgd}} = w_n$

hyperparameter: $0 < \eta \lesssim 1/\text{tr}(\Sigma)$

compare implicit regularization: batch vs online

Bounds for SGD

Theorem. For all $0 < \eta \lesssim 1/\text{tr}(\Sigma)$, in expectation

$$\mathbb{E}R(w_{\eta}^{\text{sgd}}) \approx \left\| \prod_{i=1}^n (I - \eta_i \Sigma) w^* \right\|_{\Sigma}^2 + \frac{D}{N}$$

matching upper / lower bounds

effective steps

$$N = n / \log n$$

“ N ” can be made “ n ”

critical index

$$k^* := \min \left\{ \frac{1}{\eta N} \geq c \lambda_{k+1} \right\}$$

effective dimension

$$D = k^* + \eta^2 N^2 \sum_{i > k^*} \lambda_i^2$$

effective
regularization

Zou*, Wu*, Braverman, Gu, Kakade. “Benign overfitting of constant-stepsize SGD for linear regression.” COLT 2021

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

SGD vs ridge

$$\text{excess risk} = \text{bias} + D/N$$

	SGD	ridge
<i>bias</i>	$\ e^{-\Theta(\eta N)\Sigma_{0:k^*}} w^*\ _{\Sigma_{0:k^*}}^2 + \ w^*\ _{\Sigma_{k^*:\infty}}^2$ bias decays faster	$\tilde{\lambda}^2 \ w^*\ _{\Sigma_{0:k^*}^{-1}}^2 + \ w^*\ _{\Sigma_{k^*:\infty}}^2$
<i>effective steps</i>	$N = n/\log n$	$N = n$
<i>critical index</i>	$\lambda_{k^*} \gtrsim \frac{1}{\eta N} \gtrsim \lambda_{k^*+1}$	$\lambda_{k^*} \gtrsim \lambda + \frac{\sum_{i>k^*} \lambda_i}{n} \gtrsim \lambda_{k^*+1}$
<i>effective regularization</i>	$\tilde{\lambda} = \frac{1}{\eta N}$ constraint	$\tilde{\lambda} = \lambda + \frac{\sum_{i>k^*} \lambda_i}{n}$ constraint
<i>effective dimension</i>	$\eta \lesssim 1/\text{tr}(\Sigma)$ $D = k^* + \frac{1}{\tilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2$	heavy tail

GD dominates ridge; would GD dominate SGD?

GD does not dominate SGD

Theorem [WBKLY'25]. $n \geq 1$. For a sequence of d -dim problems

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

we have $\|w^*\|_{\Sigma}^2 \leq 1$, moreover

- for all $0 < \eta \lesssim 1$ and $t \geq 0$, $\mathbb{E}R(w_t^{\text{gd}}) = \Omega(n^{-0.2})$
- for $\eta \approx 1$, $\mathbb{E}R(w_{\eta}^{\text{sgd}}) = O(\log(n)/n)$

in high-dimensional
online learning could be poly better than batch!

A lower bound for GD

Theorem [WBKLY'25]. For all $0 < \eta \lesssim 1/\text{tr}(\Sigma)$ and $t \geq 0$

$$\mathbb{E}R(w_t^{\text{gd}}) \gtrsim \left(\frac{\sum_{i>\ell^*} \lambda_i}{n} \right)^2 \|w^*\|_{\Sigma_{0:\ell^*}^{-1}}^2 + \|w^*\|_{\Sigma_{\ell^*:\infty}}^2 + \min \left\{ \frac{D}{n}, 1 \right\}$$

effective dimension $D = k^* + \frac{1}{\tilde{\lambda}^2} \sum_{i>k^*} \lambda_i^2$ as before...

benign overfitting index $\ell^* = \min \left\{ k : \frac{\sum_{i>k} \lambda_i}{n} \geq c \lambda_{k+1} \right\}$

GD variance = ridge variance

GD bias \geq OLS bias

OLS bias can be large
in high-dimensional

when would GD dominate SGD?

A SGD-type bound for GD

Theorem [WBKLY'25]. For all $0 < \eta \lesssim 1/\text{tr}(\Sigma)$ and $0 \leq t \lesssim n$, w.h.p.

$$R(w_t^{\text{gd}}) \lesssim \left\| (I - \eta \Sigma)^{t/2} w^* \right\|_{\Sigma}^2 + \frac{D}{n} + \left(\frac{D_1}{n} \right)^2$$

critical index

$$k^* := \min \left\{ \frac{1}{\eta t} \geq c \lambda_{k+1} \right\} \quad \leftarrow \text{same as SGD}$$

effective dimension

$$D = k^* + \eta^2 t^2 \sum_{i > k^*} \lambda_i^2 \quad \leftarrow \text{when } t = \Theta(N)$$

order-1 effective dim

$$D_1 = k^* + \eta t \sum_{i > k^*} \lambda_i$$

- $D \leq D_1$, always
- in the hard example, $D \ll D_1$

when would $D_1 \lesssim D$?

Spectrum condition

Assumption. Spectrum decays *fast and continuously*

$$\text{for all } \tau > 1, \quad \tau \sum_{\lambda_i < 1/\tau} \lambda_i \lesssim \#\{\lambda_i \geq 1/\tau\}$$

satisfied by

- $\lambda_i \approx a^{-i}$ for $a > 1$
- $\lambda_i \approx i^{-a}$ for $a > 1$

- rules out benign overfitting
- implies $D_1 \lesssim k^* \leq D$

violated by

- $\lambda_i \approx i^{-1} \log^{-a}(i)$ for $a > 1$
- $(\lambda_i)_{i \geq 1}$ in the hard example

$$(n^{-0.9}, 1/d, \dots, 1/d) \text{ for } d \geq n^2$$

GD dominates SGD in a subclass

Assumption. Spectrum decays fast and continuously

$$\text{for all } \tau > 1, \quad \tau \sum_{\lambda_i < 1/\tau} \lambda_i \lesssim \#\{\lambda_i \geq 1/\tau\}$$

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

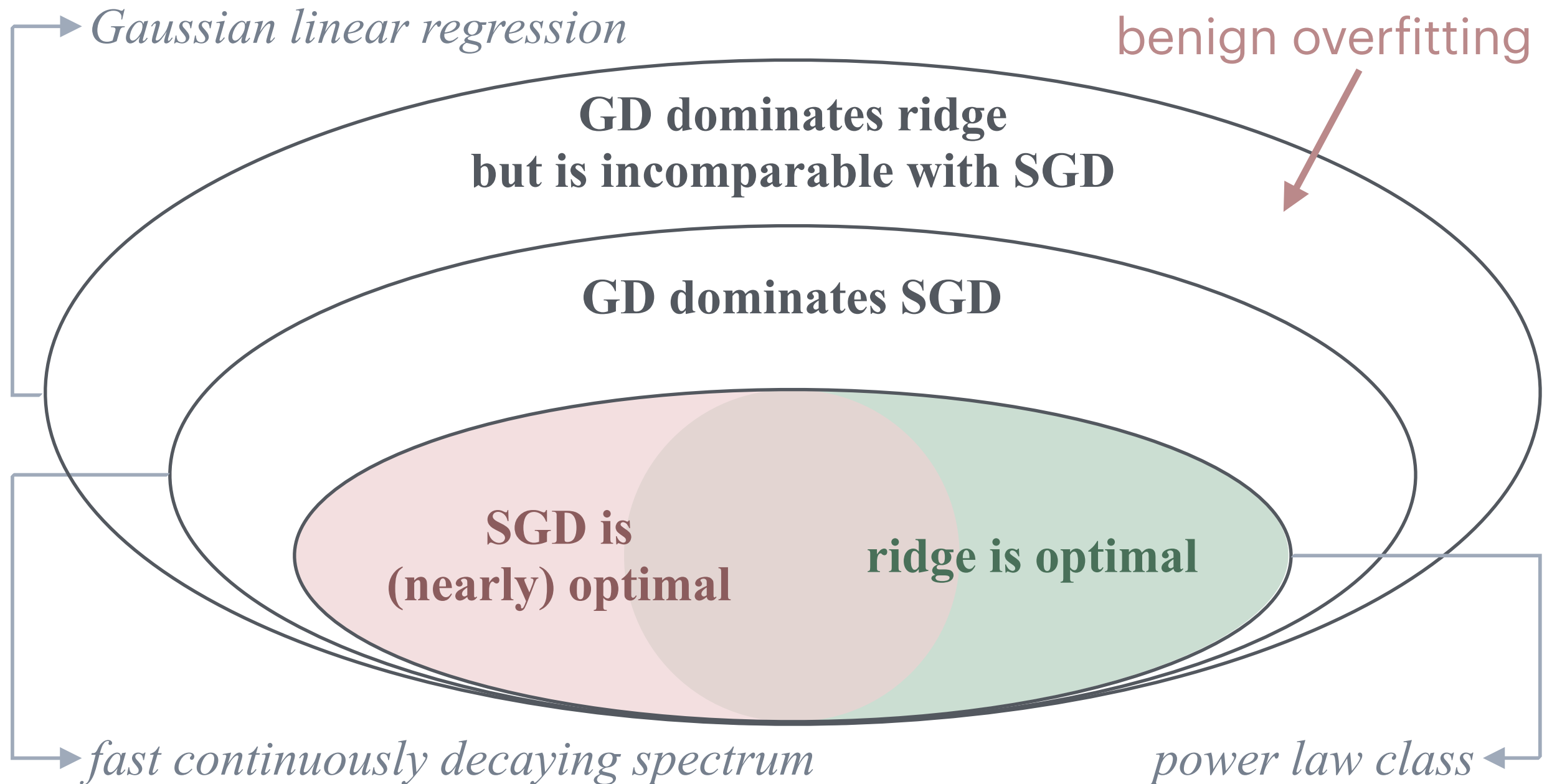
Theorem [WBKLY'25]. For every Gaussian linear regression satisfying the above, $n \geq 1$, and $0 \leq \eta \lesssim 1$, there is t such that

$$\mathbb{E}R(w_t^{\text{gd}}) \lesssim \mathbb{E}R(w_\eta^{\text{sgd}})$$

Proof. Assumption implies $D_1 \lesssim k^* \leq D$.

no constraint on w^*

Contributions



“dominance”: always no worse, sometimes much better