Risk Comparisons in Linear Regression
Implicit Regularization Dominates Explicit Regularization

Jingfeng Wu[1] Peter Bartlett*[13] Jason Lee*[1]
Sham Kakade*[23] Bin Yu*[1]
[1]UC Berkeley [2]Harvard [3]Google DeepMind

Linear regression

task. $x \sim \mathrm{N}(0, \Sigma)$, $y = x^\top w^* + \mathrm{N}(0,1)$ for $\|w^*\|_\Sigma \lesssim 1$ — problem determined by (Σ, w^*)

risk. $R(w) = \mathbb{E}(y - x^\top w)^2 - \mathbb{E}(y - x^\top w^*)^2$
$= \|w - w^*\|_\Sigma^2$

data. n iid samples $(x_1, y_1), \dots, (x_n, y_n)$ $\quad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

Algorithms

ridge.
$$w_\lambda^{\text{ridge}} = \arg\min \frac{1}{n}\sum_{i=1}^n \|x_i^\top w - y_i\|^2 + \lambda\|w\|^2 = (X^\top X + n\lambda I)^{-1} X^\top Y$$
hyperparameter: $\lambda \ge 0$

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: $t \ge 0$

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

Prior results

[Tsigler & Bartlett, 2023]

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

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

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

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

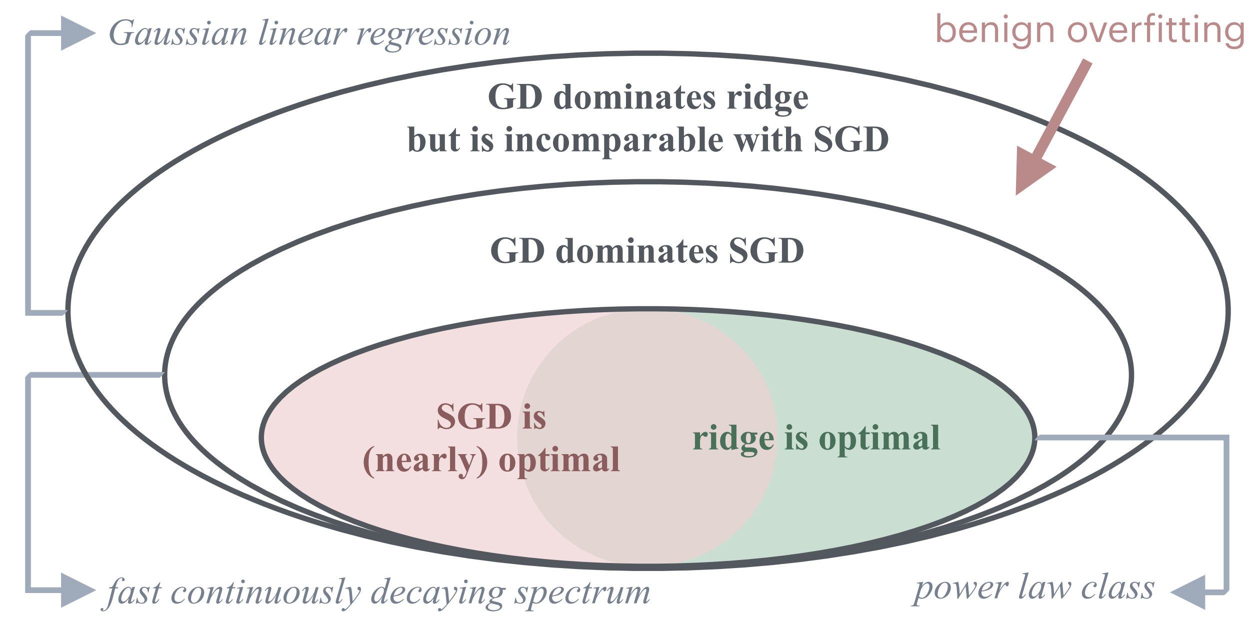
[Wu*, Zou*, Braverman, Gu, Kakade, 2022]

For all $0 < \eta \lesssim 1/\mathrm{tr}(\Sigma)$, in expectation
$$\mathbb{E}R\big(w_\eta^{\text{sgd}}\big) \eqsim \left\| \prod_{i=1}^n (I - \eta_i \Sigma) w^* \right\|_\Sigma^2 + \frac{D}{N}$$

effective steps $\quad N = n/\log n$

critical index $\quad k^* := \min\left\{\frac{1}{\eta N} \ge c\lambda_{k+1}\right\}$

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



"dominance": always no worse, sometimes much better

GD dominates ridge

Theorem

For all $0 < \eta \lesssim 1/\mathrm{tr}(\Sigma)$ and $t \ge 0$, w.h.p.
$$R\big(w_t^{\text{gd}}\big) \lesssim \tilde\lambda^2 \|w^*\|^2_{\Sigma^{-1}_{0:k^*}} + \|w^*\|^2_{\Sigma_{k^*:\infty}} + \frac{D}{n}$$

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

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

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

Corollary

For *every* Gaussian linear regression, $n \ge 1$, and $\lambda \ge 0$, there is t such that: w.h.p.
$$R\big(w_t^{\text{gd}}\big) \lesssim \mathbb{E}R\big(w_\lambda^{\text{ridge}}\big)$$

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

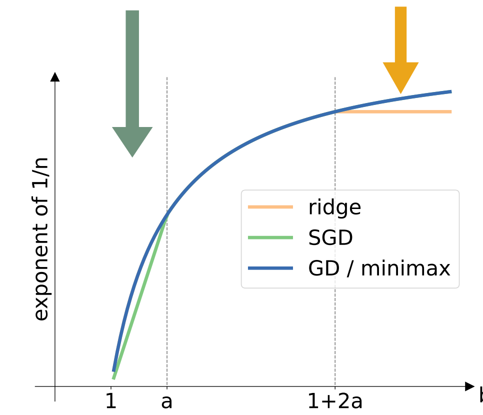
Power law class

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

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

SGD is polynomially suboptimal

ridge is polynomially suboptimal



GD is always optimal

GD is incomparable with SGD

Theorem

For all $0 < \eta \lesssim 1/\mathrm{tr}(\Sigma)$ and $t \ge 0$
$$\mathbb{E}R\big(w_t^{\text{gd}}\big) \gtrsim \left(\frac{\sum_{i>\ell^*}\lambda_i}{n}\right)^2 \|w^*\|^2_{\Sigma^{-1}_{0:\ell^*}} + \|w^*\|^2_{\Sigma_{\ell^*:\infty}} + \min\left\{\frac{D}{n}, 1\right\}$$

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

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

Corollary

$n \ge 1$. For a sequence of d-dim problems
$$d \ge n^2 \qquad w^* = \begin{bmatrix} n^{0.45} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} n^{-0.9} & & & \\ & 1/d & & \\ & & \ddots & \\ & & & 1/d \end{bmatrix}$$

we have $\|w^*\|_\Sigma^2 \lesssim 1$, moreover

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

GD dominates SGD in a significant subset

Theorem

For all $0 < \eta \lesssim 1/\mathrm{tr}(\Sigma)$ and $0 \le t \lesssim n$, w.h.p.
$$R\big(w_t^{\text{gd}}\big) \lesssim \left\|(I - \eta\Sigma)^{t/2} w^*\right\|_\Sigma^2 + \frac{D}{n} + \left(\frac{D_1}{n}\right)^2$$

critical index $\quad k^* := \min\left\{\frac{1}{\eta t} \ge c\lambda_{k+1}\right\}$

effective dimension $\quad D = k^* + \eta^2 t^2 \sum_{i>k^*}\lambda_i^2$

order-1 effective dim $\quad D_1 = k^* + \eta t \sum_{i>k^*}\lambda_i$

Assumption

Spectrum decays fast and continuously:
$$\text{for all } \tau > 1, \quad \tau \sum_{\lambda_i < 1/\tau} \lambda_i \lesssim \#\{\lambda_i \ge 1/\tau\}$$

Corollary

For *every* Gaussian linear regression satisfying the above, $n \ge 1$, and $0 \le \eta \lesssim 1$, there is t such that
$$\mathbb{E}R\big(w_t^{\text{gd}}\big) \lesssim \mathbb{E}R\big(w_\eta^{\text{sgd}}\big)$$

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