

# Comparing Classes of Estimators: When does Gradient Descent Beat Ridge Regression in Linear Models?

Dominic Richards<sup>1</sup>   Edgar Dobriban<sup>2</sup>   Patrick Rebeschini<sup>1</sup>

<sup>1</sup>Department of Statistics  
University of Oxford

<sup>2</sup>Wharton Department of Statistics and Data Science  
University of Pennsylvania

February 20, 2022

# Collaborators



Dominic Richards



Patrick Rebeschini

# Overview

Introduction

Setup

Minimax Optimality for Orthogonal Designs

General Design Matrices

- Slowly Decaying Eigenvalues

- Quickly Decaying Eigenvalues

Proof Sketch for General Design Matrices

Conclusion

# Overview

## Introduction

## Setup

## Minimax Optimality for Orthogonal Designs

## General Design Matrices

## Proof Sketch for General Design Matrices

## Conclusion

# Modern Machine Learning

## Modern Machine Learning

- ▶ Big Datasets: ImageNet, Common Crawl, MS-Coco, ...
- ▶ Big Models: GPT-3, Megatron-Turing NLG, WideResNet, ...

Backbone of many algorithms is Stochastic Gradient Descent

$$\beta_0 \sim P_0 \quad \beta_{t+1} = \beta_t - \eta \hat{g}_t \quad \mathbb{E}[\hat{g}_t] = \nabla f(\beta_t) \quad t = 0, 1, \dots$$

# Motivations for (Stochastic) Gradient Descent

Can *generally* be split in two directions:

## Computation/Approximation

- ▶ Cheap gradient estimation through mini-batch sub-sampling
- ▶ Backpropagation + auto. diff allow for expressive models
- ▶ Accelerated computations with GPUs

## Statistics/Regularization (this work)

- ▶ Implicit Regularization: GD induces a *good* inductive bias
- ▶ Canonical example: on over-parametrized least squares, GD from  $\beta_0 = 0$  converges to *Least Norm* interpolating solution

# Shrinkage methods for Linear Models

Vast amount of work focusing on (non-sparse) shrinkage methods ...

- ▶ **Early works on shrinkage methods** - Tikhonov [1943], Stein [1956], James and Stein [1961], Hoerl and Kennard [1970], Stein [1981], ...
- ▶ **Inverse Problems (GD = Landweber Damping)** - Landweber [1951], Engl et al. [1996], Bissantz et al. [2007], Caponnetto and De Vito [2007], Yao et al. [2007], Bauer et al. [2007], Raskutti et al. [2014], Rosasco and Villa [2015], Blanchard and Mücke [2018], Pagliana and Rosasco [2019], Lin et al. [2020], ...

## Classical View / Takeaway

Explicit Regularization (e.g., Ridge Regression) is Gold Standard & GD is an Approximation

# Gradient Descent “Magic”

More recently, a focus on understanding (S)GD *Magic*

- ▶ (Early Stopped Grad Flow + Ridge)  $\gg$  Ridge [Skouras et al., 1994]
- ▶ Bias towards certain principal components - textbook result, see also Belkin et al, Wu et al. [2020]
- ▶ Mild sample inflation: Single-pass SGD vs Ridge - [Zou et al., 2021]
- ▶ ...



# Our perspective

- ▶ Modern methods **depend on many tuning parameters** (learning rate, batch size, regularization strength). Empirically, performance can depend strongly on them.
- ▶ Can we understand theoretically how performance is affected? Especially, what is the *sensitivity to suboptimal hyperparameters*?
- ▶ Our work that aims to answer this (in linear models) by comparing *classes of estimators* (e.g., GD & Ridge)

# Overview

Introduction

Setup

Minimax Optimality for Orthogonal Designs

General Design Matrices

Proof Sketch for General Design Matrices

Conclusion

# Setup

Standard Linear Model: observe features  $X \in \mathbb{R}^{n \times d}$ , outcome  $Y \in \mathbb{R}^n$

$$Y = X\beta_\star + \epsilon$$

**Assumptions:**

$$\mathbb{E}[\epsilon] = 0, \quad \mathbb{E}[\epsilon\epsilon^\top] = \sigma^2 I_n, \quad \mathbb{E}[\beta_\star\beta_\star^\top] = \frac{\psi}{d} I_d, \quad \epsilon \perp \beta_\star$$

Study average-case behavior  $\mathbb{E}_{\beta_\star}[L_{\beta_\star}(\hat{\beta})]$  where  $L_{\beta_\star}(\hat{\beta}) := \|\hat{\beta} - \beta_\star\|_2^2$

Random-effects model for  $\beta_\star$ : average-case analysis over problems where each feature has small effect.

# Classes of Estimators

Focusing on two classes of estimators:

- **Ridge Regression:** for *regularization parameters*  $\lambda > 0$ , minimize  $\frac{1}{2n} \|X\beta - Y\|_2^2 + \frac{\lambda}{2} \|\beta\|_2^2$ , leading to

$$\hat{\beta}_\lambda := \left( \frac{X^\top X}{n} + \lambda I_d \right)^{-1} \frac{X^\top Y}{n}$$

- **Gradient Descent:** for  $\eta > 0$ ,  $\hat{\beta}_{\eta,0} = 0$ , and for *iterates*  $t \geq 0$

$$\begin{aligned} \hat{\beta}_{\eta,t+1} &= \hat{\beta}_{\eta,t} - \eta \nabla_{\hat{\beta}_{\eta,t}} \left( \frac{1}{2n} \|X\hat{\beta}_{\eta,t} - Y\|_2^2 \right) \\ &= \hat{\beta}_{\eta,t} - \frac{\eta}{n} X^\top (X\hat{\beta}_{\eta,t} - Y) \\ &= \sum_{\ell=0}^t \eta \left( I_d - \eta \frac{X^\top X}{n} \right)^\ell \frac{X^\top Y}{n} \end{aligned}$$

## A key classical Lemma

Optimal amount of ridge regularization  $\lambda_\star := \frac{\sigma^2 p}{\psi_n} = \frac{1}{SNR} \frac{p}{n}$ .

GD (and other estimators  $\Phi(\frac{X^\top X}{n})\frac{X^\top Y}{n}$  where  $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ ) cannot beat opt. tuned ridge regression in expectation

### Lemma

Under the random-effects model, for any  $\eta > 0$ ,  $t > 0$

$$\underbrace{\mathbb{E}_{\epsilon, \beta_\star}[L_{\beta_\star}(\hat{\beta}_{\eta, t})]}_{\text{Error of GD}} \geq \underbrace{\mathbb{E}_{\epsilon, \beta_\star}[L_{\beta_\star}(\hat{\beta}_{\lambda_\star})]}_{\text{Opt. Tuned Ridge Regression}}$$

But  $\lambda_\star$  **depends upon the unknown**  $SNR = \psi/\sigma^2$

In practice *classes of models* are considered, e.g.,

$$\Gamma_{\text{ridge}} = \{\hat{\beta}_\lambda : \lambda \in \{0, \delta, 2\delta, \dots, 1\}\} \quad \delta > 0$$

then picking  $\operatorname{argmin}_{\hat{\beta} \in \Gamma_{\text{ridge}}} \hat{\mathbb{E}}_{\epsilon, \beta_\star}[L_{\beta_\star}(\hat{\beta})]$  e.g., via cross-validation

# Overview

Introduction

Setup

Minimax Optimality for Orthogonal Designs

General Design Matrices

Proof Sketch for General Design Matrices

Conclusion

# Minimax Optimality for a *Class of Models*

- Evaluate a class of models  $\Gamma = \{\Phi_1, \dots, \Phi_k\}$ , where  $\Phi_j : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is a *spectral shrinker*, so  $\hat{\beta}_{\Phi_j} = \Phi_j(\frac{X^\top X}{n}) \frac{X^\top Y}{n}$ , via the smallest excess risk

$$\min_{\hat{\beta} \in \Gamma} \left( \mathbb{E}_{\beta_\star, \epsilon} [L_{\beta_\star}(\hat{\beta})] - \mathbb{E}_{\beta_\star, \epsilon} [L_{\beta_\star}(\hat{\beta}_{\lambda_\star})] \right)$$

- Consider a parameter space  $\Theta$  of  $\theta = (\psi, \sigma, X)$  (recall  $\mathbb{E}[\epsilon \epsilon^\top] = \sigma^2 I_n$ ,  $\mathbb{E}[\|\beta_\star\|^2] = \psi$ )
- What is the minimax optimal class of  $k$  models? Find

$$\mathcal{M}_k = \inf_{\Gamma = \{\Phi_1, \dots, \Phi_k\}} \sup_{(\psi, \sigma, X) \in \Theta} \min_{\hat{\beta} \in \Gamma} \left( \mathbb{E}_{\beta_\star, \epsilon} [L_{\beta_\star}(\hat{\beta})] - \mathbb{E}_{\beta_\star, \epsilon} [L_{\beta_\star}(\hat{\beta}_{\lambda_\star})] \right)$$

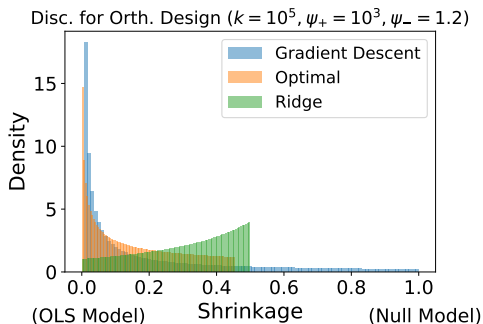
Ridge? GD? something else?

# Minimax Optimality for Orthogonal Designs

1. Orthogonal design:  $X^\top X/n = I_p$
2. Wlog fix  $\sigma = 1$  (normalize noise level:  $\mathbb{E}[\epsilon\epsilon^\top] = I_n$ )
3. Parameter space: unknown signal strength  $\mathbb{E}[\|\beta_\star\|^2] = \psi \in [\psi_-, \psi_+]$  belongs to some fixed interval.



# Histograms of shrinkers for orthogonal design



- ▶ Histograms of  $(\phi_j)_{j=1,\dots,k}$ ,  $\phi_j = 1 - \Phi_j$  for GD, Ridge, Optimal shrinker.
- ▶ Ridge does not place any shrinkers above  $1/2$ , as  $\phi_j = 1 - 1/(j/k + 1)$ .
- ▶ GD shrinkers stop short of 0, since shrinkers are of the form  $(1 - c/k)^j$  and thus,  $\min_{j \leq k} (1 - c/k)^j \gtrsim \exp(-c)$  (for large  $k$ ).
- ▶ Density of GD & optimal grids decreases around 0 (OLS model). In contrast, density of ridge grid increases.

# Minimax Optimality for Orthogonal Designs

1. Orthogonal design:  $X^\top X/n = I_p$
2. Wlog fix  $\sigma = 1$  (normalize noise level:  $\mathbb{E}[\epsilon\epsilon^\top] = I_n$ )
3. Parameter space: unknown signal strength  $\mathbb{E}[\|\beta_\star\|^2] = \psi \in [\psi_-, \psi_+]$  belongs to some fixed interval.
4. Then minimax excess risk  $M_k$  depends on  $k$ ,  $[\psi_-, \psi_+]$ . How?
  - 4.1 Should decrease with  $k$
  - 4.2 Should be zero when  $\psi_- = \psi_+$
  - 4.3 Should increase as  $[\psi_-, \psi_+]$  “expands”

# Minimax Optimality for Orthogonal Designs

## Theorem

Wlog fix  $\sigma = 1$ . Suppose  $\psi \in [\psi_-, \psi_+]$  and  $X^\top X/n = I_p$ . Then

$$\mathcal{M}_k = \frac{1}{k^2} \left( \frac{1}{\sqrt{1+\psi_-}} - \frac{1}{\sqrt{1+\psi_+}} \right)^2.$$

Optimal class  $\Gamma = \{\Phi_1, \dots, \Phi_k\}$  where  $\Phi_j = 1 - \phi_j$  for  $j = 1, \dots, k$  is

$$\phi_j = \left[ \frac{1}{x_+} + \left( j - \frac{1}{2} \right) c \right]^2 - \frac{c^2}{4}, \quad c = \frac{1}{k} \left( \frac{1}{x_-} - \frac{1}{x_+} \right), \quad x_{\pm} = \sqrt{1+\psi_{\pm}}$$

# Minimax Optimality for Orthogonal Designs

## Theorem

Wlog fix  $\sigma = 1$ . Suppose  $\psi \in [\psi_-, \psi_+]$  and  $X^\top X/n = I_p$ . Then

$$\mathcal{M}_k = \frac{1}{k^2} \left( \frac{1}{\sqrt{1+\psi_-}} - \frac{1}{\sqrt{1+\psi_+}} \right)^2.$$

Optimal class  $\Gamma = \{\Phi_1, \dots, \Phi_k\}$  where  $\Phi_j = 1 - \phi_j$  for  $j = 1, \dots, k$  is

$$\phi_j = \left[ \frac{1}{x_+} + \left( j - \frac{1}{2} \right) c \right]^2 - \frac{c^2}{4}, \quad c = \frac{1}{k} \left( \frac{1}{x_-} - \frac{1}{x_+} \right), \quad x_{\pm} = \sqrt{1+\psi_{\pm}}$$

1. Should decrease with  $k$  ✓

# Minimax Optimality for Orthogonal Designs

## Theorem

Wlog fix  $\sigma = 1$ . Suppose  $\psi \in [\psi_-, \psi_+]$  and  $X^\top X/n = I_p$ . Then

$$\mathcal{M}_k = \frac{1}{k^2} \left( \frac{1}{\sqrt{1 + \psi_-}} - \frac{1}{\sqrt{1 + \psi_+}} \right)^2.$$

Optimal class  $\Gamma = \{\Phi_1, \dots, \Phi_k\}$  where  $\Phi_j = 1 - \phi_j$  for  $j = 1, \dots, k$  is

$$\phi_j = \left[ \frac{1}{x_+} + \left( j - \frac{1}{2} \right) c \right]^2 - \frac{c^2}{4}, \quad c = \frac{1}{k} \left( \frac{1}{x_-} - \frac{1}{x_+} \right), \quad x_{\pm} = \sqrt{1 + \psi_{\pm}}$$

1. Should decrease with  $k$  ✓
2. Should be zero when  $\psi_- = \psi_+$  ✓

# Minimax Optimality for Orthogonal Designs

## Theorem

Wlog fix  $\sigma = 1$ . Suppose  $\psi \in [\psi_-, \psi_+]$  and  $X^\top X/n = I_p$ . Then

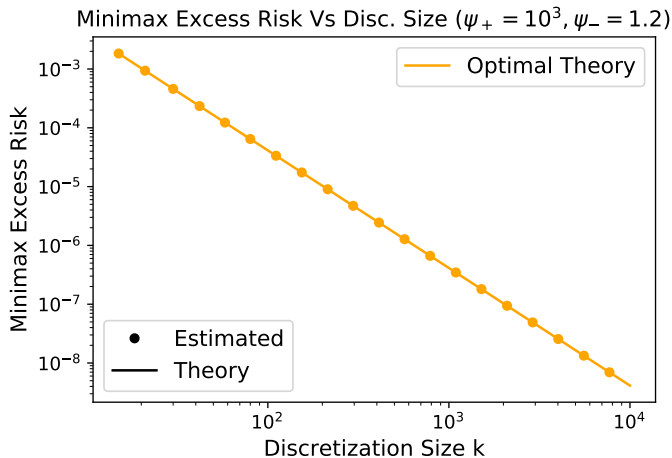
$$\mathcal{M}_k = \frac{1}{k^2} \left( \frac{1}{\sqrt{1 + \psi_-}} - \frac{1}{\sqrt{1 + \psi_+}} \right)^2.$$

Optimal class  $\Gamma = \{\Phi_1, \dots, \Phi_k\}$  where  $\Phi_j = 1 - \phi_j$  for  $j = 1, \dots, k$  is

$$\phi_j = \left[ \frac{1}{x_+} + \left( j - \frac{1}{2} \right) c \right]^2 - \frac{c^2}{4}, \quad c = \frac{1}{k} \left( \frac{1}{x_-} - \frac{1}{x_+} \right), \quad x_{\pm} = \sqrt{1 + \psi_{\pm}}$$

1. Should decrease with  $k$  ✓
2. Should be zero when  $\psi_- = \psi_+$  ✓
3. Should increase as  $[\psi_-, \psi_+]$  “expands” ✓

# Minimax Excess Risk



- Perfect match of theory and numerics.

# Excess risks of specific classes

Comparing shrinkage performance with  $k$  models

- ▶ **Minimax Grid:** We have

$$\text{Minimax excess risk} \sim \frac{1}{k^2}$$

- ▶ **Ridge Regression:** with uniform discretization

$$\text{Excess risk decays as } \begin{cases} \sim 1/k & \text{for } k \text{ small enough} \\ \sim 1/k^2 & \text{for } k \text{ large enough} \end{cases}$$

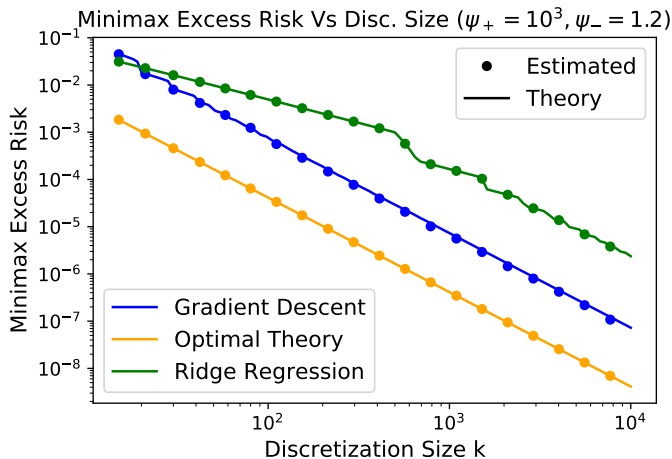
(also have “switching” behavior)

- ▶ **Gradient Descent:** With  $\eta \sim 1/k$ ,

$$\text{Excess risk} \sim \frac{1}{k^2}$$



# Comparing Excess Risks



- ▶ Excess risk of Ridge Regression initially decays at the slow  $O(1/k)$  rate
- ▶ Excess risks of GD and optimal method both decay at fast  $O(1/k^2)$  rate.

## Minimax Excess Risk: Proof sketch

- ▶ After some algebra, equivalent to

$$\inf_{\phi_1, \dots, \phi_k \in \mathbb{R}} \sup_{x \in [x_-, x_+]} \min_{j \in [k]} \left| x\phi_j - \frac{1}{x} \right|.$$

- ▶ By the monotonicity properties of  $g : (0, \infty) \times (0, \infty) \mapsto [0, \infty)$ ,  $g(x, l) = |xl - \frac{1}{x}|$ , for fixed  $x_+^{-2} \leq \phi_1 \leq \dots \leq \phi_k \leq x_-^{-2}$ , the sup is

$$\max \left( \frac{1}{x_-} - x_- \phi_k, \max_{j=1, \dots, k-1} \frac{\phi_{j+1} - \phi_j}{\sqrt{2(\phi_{j+1} + \phi_j)}}, x_+ \phi_1 - \frac{1}{x_+} \right).$$

- ▶ If any two terms are not equal, can decrease the inf by changing some  $\phi_j$ . By the compactness of  $x_+^{-2} \leq \phi_1 \leq \dots \leq \phi_k \leq x_-^{-2}$ , the min is achieved; and all terms must be equal (say, to  $\gamma$ ).
- ▶ Solve recursion  $\phi_{j+1} - \phi_j = \gamma \cdot \sqrt{2(\phi_{j+1} + \phi_j)}$ , with boundary conditions.

# Overview

Introduction

Setup

Minimax Optimality for Orthogonal Designs

General Design Matrices

- Slowly Decaying Eigenvalues

- Quickly Decaying Eigenvalues

Proof Sketch for General Design Matrices

Conclusion

# General Design Matrices - Setup

**Design Matrix:** eigenvalues of  $X^\top X/n$  are  $s_1 \geq s_2 \geq \dots \geq s_r > 0$

**Assumptions:** Suppose  $\psi = 1$  and that

$$1 \geq \frac{d}{n} \sigma^2 = \lambda_\star \geq \lambda_{\min} > 0$$


where  $\lambda_{\min} > 0$  is known. Equivalently,  $\sigma \geq \sigma_{\min} > 0$  where  $\sigma_{\min}$  known.<sup>1</sup>

**Model Classes:** size  $k > 1$

$$\begin{aligned} \mathcal{C}^{\text{GD}}(\eta) &:= \{\hat{\beta}_{\eta,j} : 1 \leq j \leq k\}, \\ \mathcal{C}^{\text{Ridge}}(\lambda_{\min}) &:= \{\hat{\beta}_\lambda : \lambda \in \{\lambda_{\min}, \lambda_{\min} + \delta, \lambda_{\min} + 2\delta, \dots, 1\}, \} \end{aligned}$$

where  $\delta = (1 - \lambda_{\min})/(k - 1)$

---

<sup>1</sup>Differs from previous parametrization with fixed  $\sigma$ , but equivalent to it. 

# Comparing Classes of Models

## Definition

For any two finite classes of estimators  $\mathcal{C}_1 = \{\hat{\beta}^u\}_{u \in U_1}$  and  $\mathcal{C}_2 = \{\hat{\beta}^u\}_{u \in U_2}$ , define the *relative sub-optimality ratio*<sup>a</sup>

$$\mathcal{S}(\mathcal{C}_1, \mathcal{C}_2) := \frac{\min_{\hat{\beta} \in \mathcal{C}_1} \mathbb{E}_{\beta_*, \epsilon}[L_{\beta_*}(\hat{\beta})] - \mathbb{E}_{\beta_*, \epsilon}[L_{\beta_*}(\hat{\beta}_{\lambda_*})]}{\min_{\hat{\beta} \in \mathcal{C}_2} \mathbb{E}_{\beta_*, \epsilon}[L_{\beta_*}(\hat{\beta})] - \mathbb{E}_{\beta_*, \epsilon}[L_{\beta_*}(\hat{\beta}_{\lambda_*})]}.$$

---

<sup>a</sup>for  $x > 0$ ,  $x/0 = \infty$ , and  $0/0$  is undefined

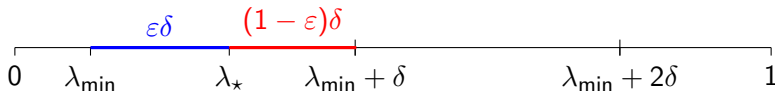
- ▶ If  $\mathcal{S}(\mathcal{C}_1, \mathcal{C}_2) < 1$ , best model in  $\mathcal{C}_1$  outperforms best model in  $\mathcal{C}_2$ .
- ▶ Study  $\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min}))$

# General Design Matrices - Setup

- ▶ Provide general results based on eigenvalue decay
- ▶ Specialize them to slow decay (power law  $i^{-\alpha}$ ,  $\alpha < 1$ ); and fast decay (power law  $i^{-\alpha}$ ,  $\alpha > 1$ ; exponential). Focus presentation on this

## General Design Matrices - Slow Decay Result

Let  $\Gamma = \{\lambda_{\min}, \lambda_{\min} + \delta, \lambda_{\min} + 2\delta, \dots, 1\}$  &  $\text{Dist}_\delta(\lambda_\star, \Gamma) = \frac{1}{\delta} \min_{\lambda \in \Gamma} |\lambda - \lambda_\star|$



$\text{Dist}_\delta(\lambda_\star, \Gamma) = \min\{\varepsilon, 1 - \varepsilon\}$  if  $\lambda_\star = \lambda_{\min} + (j + \varepsilon)\delta$  for  $j \leq k - 2, \varepsilon \in (0, 1)$ .

### Theorem (Slow Decay)

Suppose  $s_i = i^{-\alpha}$  for  $\alpha \in (0, 1)$ . There exists  $r_{\alpha, \lambda_\star, k, \lambda_{\min}} > 1$  such that if

$$k \gtrsim \frac{1}{\lambda_\star \lambda_{\min}} \log\left(1 + \frac{1}{\lambda_\star}\right) \quad r \geq r_{\alpha, \lambda_\star, k, \lambda_{\min}}$$

then with  $\eta = 1/(k\lambda_{\min})$

$$\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min})) \simeq \frac{1}{\text{Dist}_\delta(\lambda_\star, \Gamma)^2} \left(\frac{d}{n}\right)^2 \left(\frac{\sigma^2}{\sigma_{\min}}\right)^4.$$

# General Design Matrices w/ Slow Decay - Result

If eigenvalues decay slowly,  $r$  is large and  $\text{Dist}_\delta(\lambda_\star, \Gamma) = \Omega(1)$  then

$$\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min})) \simeq \left(\frac{d}{n}\right)^2 \left(\frac{\sigma^2}{\sigma_{\min}}\right)^4.$$

- ▶ If  $d \simeq n$  and  $\sigma_{\min} = \Omega(\sigma)$  then  $\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min})) \simeq \sigma^4$
- ▶ (roughly) GD outperforms ridge regression in High-Dim. + Low Noise!
- ▶ For instance, if  $r = d = n^q$  for  $q \in (g(\alpha), 1)$ , then<sup>2</sup>

$$\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min})) \simeq \frac{(\sigma^2/\sigma_{\min})^4}{n^{2(1-q)}}$$

Tends to zero as  $n, d \rightarrow \infty$

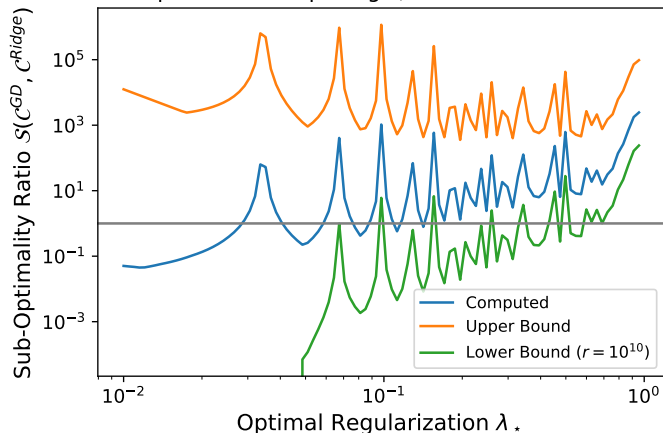
---

<sup>2</sup> $g(\alpha) = (\alpha + 1)/(1 + \alpha(2 - \alpha))$



# General Design Matrices w/ Slow Decay - Experiment

Sub-Opt. Ratio Vs Opt Reg. ( $k=100, r=10^6, \alpha=0.5$ )



- ▶ Numerically compute  $\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min}))$  for fixed  $r$ .
- ▶ Evaluate our upper and lower bounds (constants = 1)
- ▶ Bounds capture ↗ trend and spiking due to discretization as a function of  $\lambda_* = p\sigma^2/n$ .

# General Design Matrices w/ Fast Decay - Result

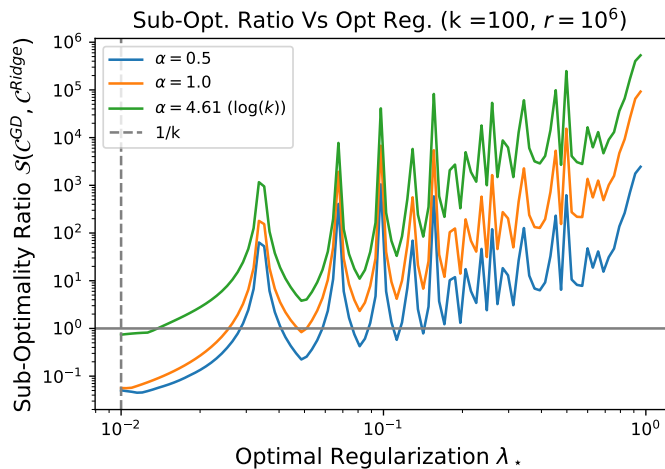
## Theorem (Fast Decay)

Suppose  $s_i = i^{-\alpha}$  for  $\alpha > 1$  and  $s_r \leq \lambda_*$ , and  $k \gtrsim \frac{1}{\lambda_* \lambda_{\min}} \log(1 + \frac{1}{\lambda_*})$ .  
Then with  $\eta = 1/(k\lambda_{\min})$

$$\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min})) \gtrsim \frac{\exp(-O(\alpha))}{\text{Dist}_{\delta}(\lambda_*, \Gamma)^2} \left(\frac{d}{n}\right)^2 k^2 \sigma^4.$$

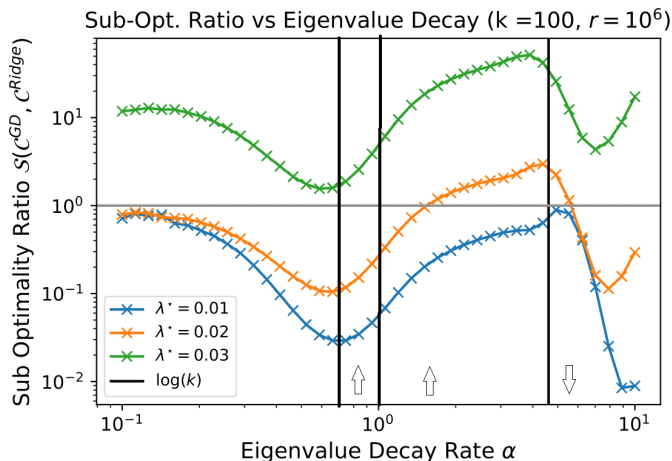
- ▶ Intuition: Ridge Reg. outperforms GD, so  $\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min})) \gg 1$ , when eigenvalues decay quickly.
- ▶ When  $d \sim n$  and  $\sigma$  is a constant, nontrivial when  $1 < \alpha \lesssim \log(k)$
- ▶ Also holds for exponentially decaying eigenvalues  $s_i = e^{-\rho i}$  for  $\rho > 0$ .

# General Design Matrices - Experiment w/ $\alpha$



- ▶ Numerically compute  $\mathcal{S}(\mathcal{C}^{\text{GD}}(\eta), \mathcal{C}^{\text{Ridge}}(\lambda_{\min}))$  for fixed  $r$ .
- ▶  $\log(k)$  seems to be roughly the threshold where  $RR \gg GD$  uniformly for  $\lambda_* \geq 1/k$  (where our theory holds).

# General Design Matrices - Full range of $\alpha$



- rough agreement with our theory on ranges  $[c, 1]$ ,  $[1, \log k]$ ,  $[\log k, C]$

# Overview

Introduction

Setup

Minimax Optimality for Orthogonal Designs

General Design Matrices

Proof Sketch for General Design Matrices

Conclusion

# Proof Sketch - Setup

For a class of models  $\mathcal{C}$ , optimal excess risk is

$$\mathcal{E}(\mathcal{C}) := \min_{\hat{\beta} \in \mathcal{C}} \left\{ \mathbb{E}_{\beta_*, \epsilon} [L_{\beta_*}(\hat{\beta})] - \mathbb{E}_{\beta_*, \epsilon} [L_{\beta_*}(\hat{\beta}_{\lambda_*})] \right\}.$$

Consider one estimator  $\hat{\beta}_{\Phi} = \Phi(\frac{X^{\top}X}{n})\frac{X^{\top}Y}{n}$ . With  $M_{\Phi}(s) = 1 - \Phi(s)s$

$$\mathcal{E}(\hat{\beta}_{\Phi}) = \frac{\lambda_*}{d} \sum_{i=1}^r \left( \frac{1}{\lambda_*} + \frac{1}{s_i} \right) \left( M_{\Phi}(s_i) - \frac{\lambda_*}{\lambda_* + s_i} \right)^2.$$

Consider the classes  $\mathcal{C} = (\Phi_1, \dots, \Phi_k)$  and  $\delta = (1 - \lambda_{\min})/(k - 1)$

- ▶ **Ridge Regression:**  $M_{\Phi_j}(s) = \lambda_j/(s + \lambda_j)$  where  $\lambda_j = \lambda_{\min} + (j - 1)\delta$
- ▶ **GD:**  $M_{\Phi_j}(s) = (1 - \eta s)^j$  for  $j = 1, \dots, k$ .

# Proof Sketch - Ridge Regression

Recall that  $\lambda_\star = \lambda_{\min} + (j + \varepsilon)\delta$  for some  $j \in \{0, 1, \dots, k-2\}, \varepsilon \in (0, 1)$

$$\begin{aligned}\mathcal{E}(\mathcal{C}^{\text{Ridge}}(\lambda_{\min}, k)) &= \lambda_\star \min_{t=1, \dots, k} \frac{1}{d} \sum_{i=1}^r \left( \frac{1}{\lambda_\star} + \frac{1}{s_i} \right) \left( \frac{\lambda_t}{\lambda_t + s_i} - \frac{\lambda_\star}{\lambda_\star + s_i} \right)^2 \\ &= \lambda_\star \min_{\kappa \in \{-\varepsilon, 1-\varepsilon\}} \frac{1}{d} \sum_{i=1}^r \left( \frac{1}{\lambda_\star} + \frac{1}{s_i} \right) \frac{s_i^2 \kappa^2 \delta^2}{(\lambda_\star + \kappa \delta + s_i)^2 (\lambda_\star + s_i)^2}\end{aligned}$$

Suppose  $\varepsilon = 1/2$  and  $\delta \leq \lambda_\star$ , so the summands are  $\frac{s_i}{\lambda_\star} \frac{\delta^2}{(s_i + \lambda_\star)^3}$ ; then

$$\mathcal{E}(\mathcal{C}^{\text{Ridge}}(\lambda_{\min}, k)) \simeq \frac{\lambda_\star \delta^2}{d} \left( \sum_{s_i > \lambda_\star} \frac{\lambda_\star}{s_i^2} + \underbrace{\sum_{s_i \leq \lambda_\star} \frac{s_i}{\lambda_\star^3}}_{\text{Tail of Eigenvalue Distribution}} \right). \quad (1)$$

# Proof Sketch - Gradient Descent

Plugging in the class of models

$$\mathcal{E}(\mathcal{C}^{\text{GD}}(\eta, k)) = \lambda_{\star} \min_{t=1, \dots, k} \frac{1}{d} \sum_{i=1}^r \left( \frac{1}{\lambda_{\star}} + \frac{1}{s_i} \right) \left( (1 - \eta s_i)^t - \frac{\lambda_{\star}}{\lambda_{\star} + s_i} \right)^2.$$

Switch perspective: define **per-eigenvalue optimal number of iterations**  $t^*(s) = \log(1 + s/\lambda_{\star}) / \log(1/(1 - \eta s))$  so that

$$(1 - \eta s)^{t^*(s)} = \frac{\lambda_{\star}}{\lambda_{\star} + s} \quad \lim_{s \rightarrow 0} t^*(s) = \frac{1}{\eta \lambda_{\star}}$$

Choosing **global number of iterations**  $t = \lceil (\eta \lambda_{\star})^{-1} \rceil$ , can bound

$$\mathcal{E}(\mathcal{C}^{\text{GD}}(\eta, k)) \lesssim \frac{\lambda_{\star}}{d} \left( \sum_{i: s_i > \lambda_{\star}} \frac{\lambda_{\star}}{s_i^2} + \sum_{i: s_i \leq \lambda_{\star}} \frac{1}{s_i} \left( 1 - (1 - \eta s_i)^{\lceil (\eta \lambda_{\star})^{-1} \rceil - t^*(s_i)} \right)^2 \right)$$

**Key calculation:** bounding  $\lceil (\eta \lambda_{\star})^{-1} \rceil - t^*(s_i)$



# Proof Sketch - Gradient Descent

We have

$$\frac{1}{\eta\lambda_\star} - t^\star(s) = t^\star(0) - t^\star(s) \leq \frac{3}{2} \frac{s}{\eta\lambda_\star^2}$$

yielding  $1 - (1 - \eta s_i)^{\lceil (\eta\lambda_\star)^{-1} \rceil - t^\star(s_i)} \lesssim \eta s (1 + \frac{s}{\eta\lambda_\star^2})$  and thus

$$\mathcal{E}(\mathcal{C}^{\text{GD}}(\eta, k)) \lesssim \frac{\lambda_\star}{d} \left( \sum_{i: s_i > \lambda_\star} \frac{\lambda_\star}{s_i^2} + \sum_{i: s_i \leq \lambda_\star} \max \left\{ \eta^2 s_i, \frac{s_i^3}{\lambda_\star^4} \right\} \right). \quad (2)$$

Compare Tail Dependence  $s_i < \eta\lambda_\star^2$

► **Ridge Regression:**  $\frac{\delta^2 s}{\lambda_\star^3}$

► **Gradient Descent:**  $\lambda_\star \eta^2 s$

Eigenvalue decay slow: Dominant error is in the tails, so

$$\frac{\lambda_\star \eta^2 s_i}{(\delta^2 s_i)/(\lambda_\star^3)} \approx \frac{\lambda_\star^4}{k^2 \lambda_{\min}^2} \frac{1}{(1/k^2)} = \frac{\lambda_\star^4}{\lambda_{\min}^2} = \left(\frac{d}{n}\right)^2 \left(\frac{\sigma^2}{\sigma_{\min}}\right)^2$$

where  $\eta = 1/(k\lambda_{\min})$ ,  $\delta = O(1/k)$ .

# Overview

Introduction

Setup

Minimax Optimality for Orthogonal Designs

General Design Matrices

Proof Sketch for General Design Matrices

Conclusion

# Conclusion

Aim: compare classes of estimators accounting for (discrete) hyperparameters.  
Some intriguing results in linear models.

## Orthogonal Designs:

- ▶ Min-Max-Min excess risk

$$\underbrace{\inf_{\Gamma=\{\Phi_1, \dots, \Phi_k\}}}_{\text{Model Class}} \underbrace{\sup_{\psi, \sigma, X}}_{\text{Nature}} \underbrace{\min_{\hat{\beta} \in \Gamma}}_{\text{Model Selection}} \underbrace{\left( \mathbb{E}_{\beta_*, \epsilon} [L_{\beta_*}(\hat{\beta})] - \mathbb{E}_{\beta_*, \epsilon} [L_{\beta_*}(\hat{\beta}_{\lambda_*})] \right)}_{\text{Excess risk}}$$

- ▶ For  $k$  models, optimal class and GD:  $O(1/k^2)$
- ▶ Ridge: “switching” between  $O(1/k)$  &  $O(1/k^2) \implies \text{GD} \gg \text{Ridge}$

**General Designs:** compare GD & Ridge via relative suboptimality

- ▶ Fast eigenvalue decay  $1 < \alpha \lesssim \log(k)$ :  $\text{Ridge} \gg \text{GD}$
- ▶ Slow eigenvalue decay  $0 < \alpha < 1$ , small noise, high dim:  $\text{GD} \gg \text{Ridge}$

# References I

- Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.
- Nicolai Bissantz, Thorsten Hohage, Axel Munk, and Frits Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45(6):2610–2636, 2007.
- Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4): 971–1013, 2018.
- Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3): 331–368, 2007.
- Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.
- Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

# References II

- W James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 1961.
- Louis Landweber. An iteration formula for fredholm integral equations of the first kind. *American journal of mathematics*, 73(3):615–624, 1951.
- Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020.
- Nicolò Pagliana and Lorenzo Rosasco. Implicit regularization of accelerated methods in hilbert spaces. *Advances in Neural Information Processing Systems*, 32: 14481–14491, 2019.
- Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and nonparametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.
- Lorenzo Rosasco and Silvia Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.
- K Skouras, C Goutis, and MJ Bramson. Estimation in linear models using gradient descent with early stopping. *Statistics and Computing*, 4(4):271–278, 1994.

## References III

- Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.
- Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.
- Andrey Nikolayevich Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.
- Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. In *International Conference on Learning Representations*, 2020.
- Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.
- Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, Dean P Foster, and Sham M Kakade. The benefits of implicit regularization from SGD in least squares problems. *arXiv preprint arXiv:2108.04552*, 2021.