Optimal Ridge Regularization for Out-of-Distribution Prediction

Pratik Patil

University of California Berkeley

MDS 2024

Based on joint work with the following amazing collaborators:

- Jin-Hong Du (Carnegie Mellon University)
- Ryan Tibshirani (University of California Berkeley)

`https://pratikpatil.io/papers/ridge-ood.pdf`

Outline

Overparameterization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

Overparameterization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

Overparameterization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

Overparameterization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

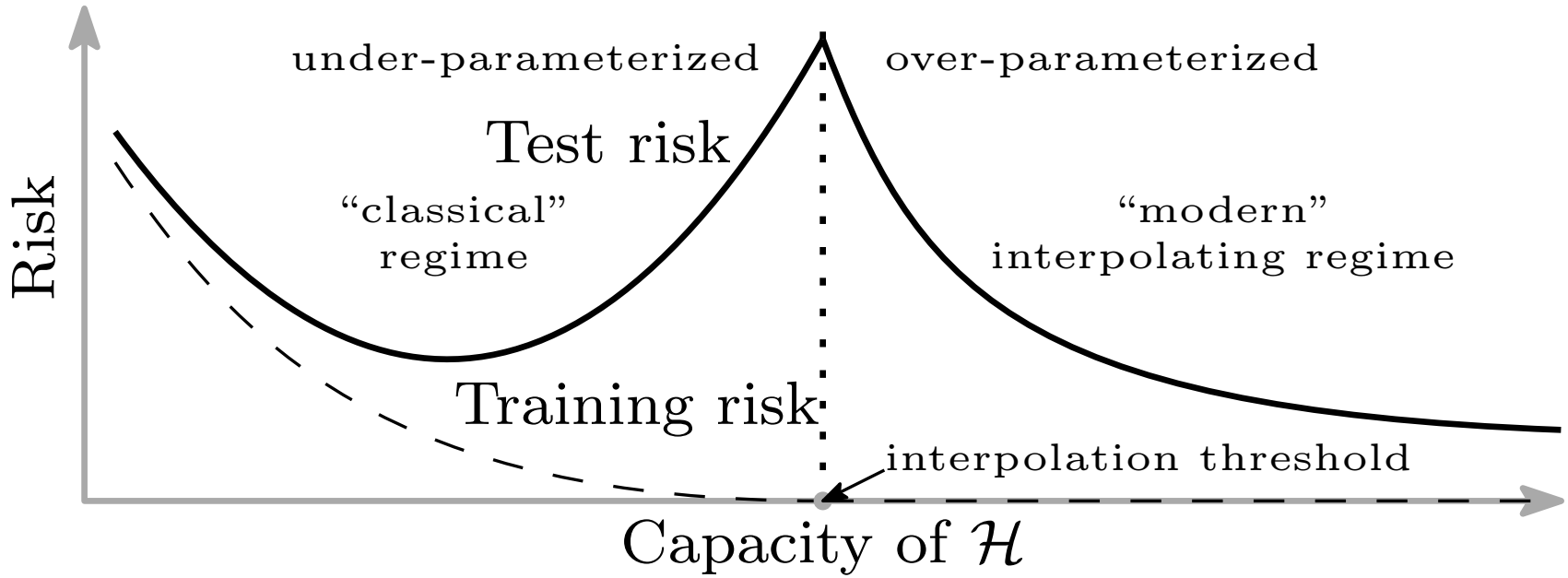
Overparameterization in machine learning

Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for:

- Representation: allows rich, expressive models for diverse real data
- Optimization: simple, local optimization methods often find near-optimal solutions to empirical risk minimization problem
- Generalization: despite overfitting, models generalize well in practice

This talk is about generalization aspect in overparameterized learning.

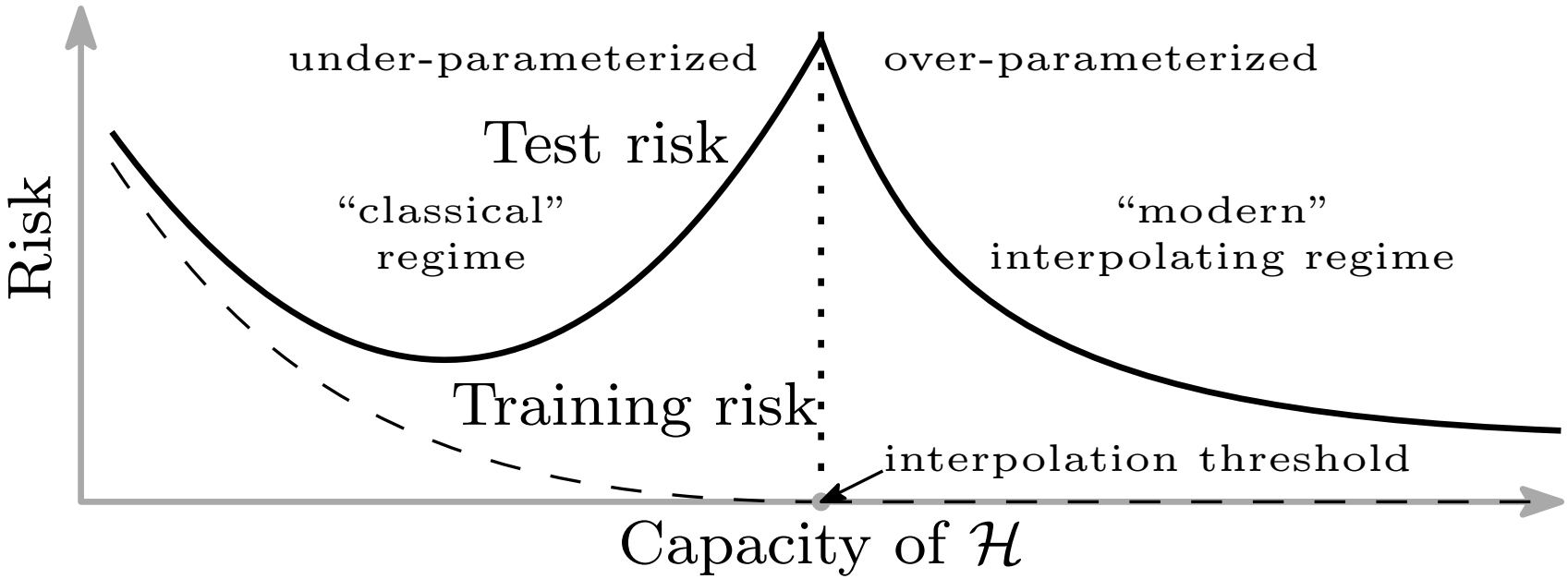
Peculiar generalization behavior: double descent



Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- The phenomenon is dubbed "double descent" in the risk curve.
- This trend holds for many model classes including linear regression, kernel regression, random forest, boosting, neural networks, etc.

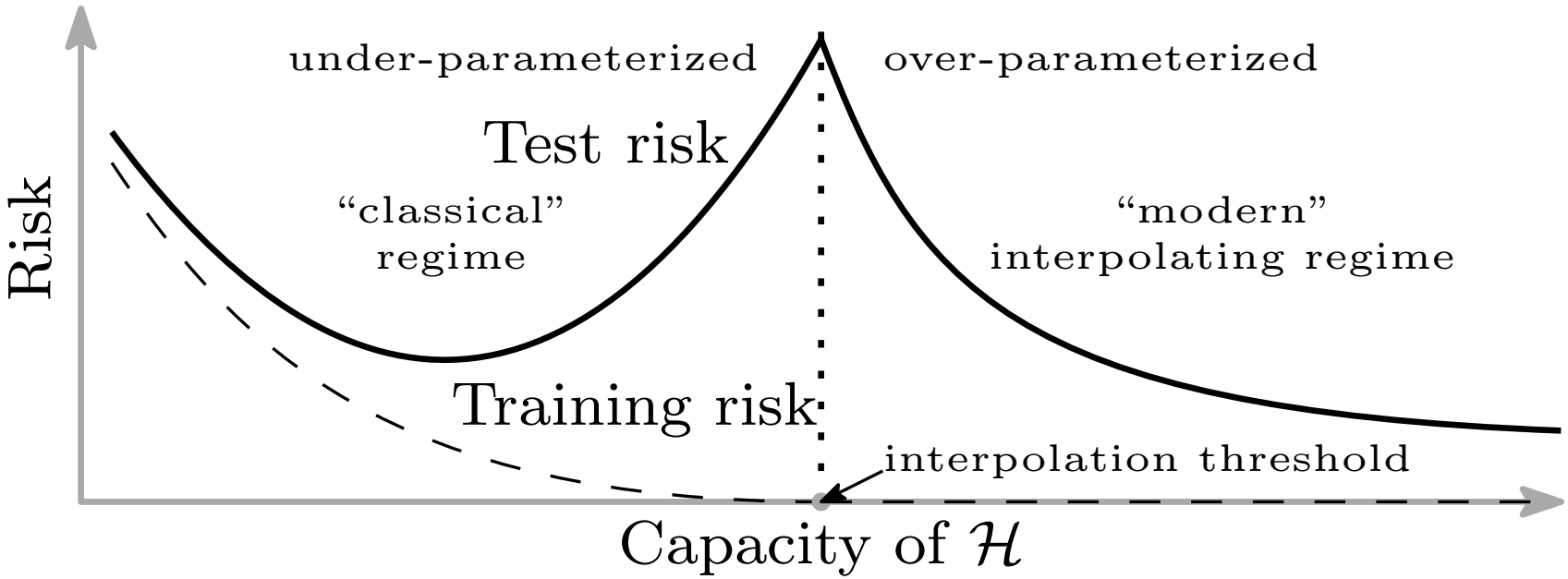
Peculiar generalization behavior: double descent



Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- The phenomenon is dubbed "double descent" in the risk curve.
- This trend holds for many model classes including linear regression, kernel regression, random forest, boosting, neural networks, etc.

Peculiar generalization behavior: double descent



Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- The phenomenon is dubbed "double descent" in the risk curve.
- This trend holds for many model classes including linear regression, kernel regression, random forest, boosting, neural networks, etc.

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

Recent theoretical developments

Understanding generalization of interpolators in simpler settings:

- Linear regression
  - Hastie, Montanari, Rosset, Tibshirani, 2019
  - Belkin, Hsu, Xu, 2019
  - Muthukumar, Vodrahalli, Sahai, 2019
  - Bartlett, Long, Lugosi, Tsigler, 2019
  - Mei, Montanari, 2019
- Kernel regression
  - Liang, Rakhlin, 2018
  - Liang, Rakhlin, Zhai, 2019
- Local methods
  - Belkin, Hsu, Mitra, 2018
  - Belkin, Rakhlin, Tsybakov, 2018
- and many more ...

What do we currently understand?

- In nearly all applications, current practice suggests we should design models to be massively overparameterized
- Once trained (typically by SGD), these models interpolate the training data (achieve zero training error)
- Still they are capable of having (often do have) good test error

Current understanding of this? In full theoretical rigor, not great.

However, the story is fairly well-understood for linear models, kernel models, and random feature models. See, e.g., nice monographs:

- Bartlett, Montanari, and Rakhlin (2021), "Deep learning: a statistical viewpoint"
- Belkin (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

What do we currently understand?

- In nearly all applications, current practice suggests we should design models to be massively overparameterized
- Once trained (typically by SGD), these models interpolate the training data (achieve zero training error)
- Still they are capable of having (often do have) good test error

Current understanding of this? In full theoretical rigor, not great.

However, the story is fairly well-understood for linear models, kernel models, and random feature models. See, e.g., nice monographs:

- Bartlett, Montanari, and Rakhlin (2021), "Deep learning: a statistical viewpoint"
- Belkin (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

What do we currently understand?

- In nearly all applications, current practice suggests we should design models to be massively overparameterized
- Once trained (typically by SGD), these models interpolate the training data (achieve zero training error)
- Still they are capable of having (often do have) good test error

Current understanding of this? In full theoretical rigor, not great.

However, the story is fairly well-understood for linear models, kernel models, and random feature models. See, e.g., nice monographs:

- Bartlett, Montanari, and Rakhlin (2021), "Deep learning: a statistical viewpoint"
- Belkin (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

What do we currently understand?

- In nearly all applications, current practice suggests we should design models to be massively overparameterized
- Once trained (typically by SGD), these models interpolate the training data (achieve zero training error)
- Still they are capable of having (often do have) good test error

Current understanding of this? In full theoretical rigor, not great.

However, the story is fairly well-understood for linear models, kernel models, and random feature models. See, e.g., nice monographs:

- Bartlett, Montanari, and Rakhlin (2021), "Deep learning: a statistical viewpoint"
- Belkin (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

What do we currently understand?

- In nearly all applications, current practice suggests we should design models to be massively overparameterized
- Once trained (typically by SGD), these models interpolate the training data (achieve zero training error)
- Still they are capable of having (often do have) good test error

Current understanding of this? In full theoretical rigor, not great.

However, the story is fairly well-understood for linear models, kernel models, and random feature models. See, e.g., nice monographs:

- Bartlett, Montanari, and Rakhlin (2021), "Deep learning: a statistical viewpoint"
- Belkin (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

What do we currently understand?

- In nearly all applications, current practice suggests we should design models to be massively overparameterized
- Once trained (typically by SGD), these models interpolate the training data (achieve zero training error)
- Still they are capable of having (often do have) good test error

Current understanding of this? In full theoretical rigor, not great.

However, the story is fairly well-understood for linear models, kernel models, and random feature models. See, e.g., nice monographs:

- Bartlett, Montanari, and Rakhlin (2021), "Deep learning: a statistical viewpoint"
- Belkin (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

What do we currently understand?

- In nearly all applications, current practice suggests we should design models to be massively overparameterized
- Once trained (typically by SGD), these models interpolate the training data (achieve zero training error)
- Still they are capable of having (often do have) good test error

Current understanding of this? In full theoretical rigor, not great.

However, the story is fairly well-understood for linear models, kernel models, and random feature models. See, e.g., nice monographs:

- Bartlett, Montanari, and Rakhlin (2021), "Deep learning: a statistical viewpoint"
- Belkin (2021), "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation"

Simplest linear analysis

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Given i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$y_i = \underbrace{x_i^\top \beta}_{f(x_i)} + \epsilon_i, \quad x_i^\top \beta \perp \epsilon_i$$

"Ridgeless" least squares estimator of y on X (which has rows x_i):

$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Let $\sigma^2 = \mathrm{Var}(\epsilon_i)$ [noise energy], $\rho^2 = \mathbb{E} f(x_i)^2$ [signal energy].
Under simplifying assumptions, as $n, p \to \infty$ with $p/n \to \gamma$:

$$\mathbb{E}(f(x_0) - x_0^\top \widehat{\beta})^2 \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \rho^2 \frac{\gamma - 1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \gamma > 1 \end{cases}$$

(Two terms: estimation bias, and estimation variance)

Simplest linear analysis

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Given i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$y_i = \underbrace{x_i^\top \beta}_{f(x_i)} + \epsilon_i, \quad x_i^\top \beta \perp \epsilon_i$$

"Ridgeless" least squares estimator of y on X (which has rows x_i):

$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Let $\sigma^2 = \mathrm{Var}(\epsilon_i)$ [noise energy], $\rho^2 = \mathbb{E} f(x_i)^2$ [signal energy].
Under simplifying assumptions, as $n, p \to \infty$ with $p/n \to \gamma$:

$$\mathbb{E}(f(x_0) - x_0^\top \widehat{\beta})^2 \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \rho^2 \frac{\gamma - 1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \gamma > 1 \end{cases}$$

(Two terms: estimation bias, and estimation variance)

Simplest linear analysis

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Given i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$y_i = \underbrace{x_i^\top \beta}_{f(x_i)} + \epsilon_i, \quad x_i^\top \beta \perp \epsilon_i$$

"Ridgeless" least squares estimator of y on X (which has rows x_i):

$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Let $\sigma^2 = \mathrm{Var}(\epsilon_i)$ [noise energy], $\rho^2 = \mathbb{E} f(x_i)^2$ [signal energy].
Under simplifying assumptions, as $n, p \to \infty$ with $p/n \to \gamma$:

$$\mathbb{E}(f(x_0) - x_0^\top \widehat{\beta})^2 \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \rho^2 \frac{\gamma - 1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \gamma > 1 \end{cases}$$

(Two terms: estimation bias, and estimation variance)

Simplest linear analysis

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Given i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$y_i = \underbrace{x_i^\top \beta}_{f(x_i)} + \epsilon_i, \quad x_i^\top \beta \perp \epsilon_i$$

"Ridgeless" least squares estimator of y on X (which has rows x_i):

$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Let $\sigma^2 = \mathrm{Var}(\epsilon_i)$ [noise energy], $\rho^2 = \mathbb{E} f(x_i)^2$ [signal energy].
Under simplifying assumptions, as $n, p \to \infty$ with $p/n \to \gamma$:

$$\mathbb{E}(f(x_0) - x_0^\top \widehat{\beta})^2 \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \rho^2 \frac{\gamma - 1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \gamma > 1 \end{cases}$$

(Two terms: estimation bias, and estimation variance)

Simplest linear analysis

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Given i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$y_i = \underbrace{x_i^\top \beta}_{f(x_i)} + \epsilon_i, \quad x_i^\top \beta \perp \epsilon_i$$

"Ridgeless" least squares estimator of y on X (which has rows x_i):

$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Let $\sigma^2 = \mathrm{Var}(\epsilon_i)$ [noise energy], $\rho^2 = \mathbb{E} f(x_i)^2$ [signal energy].
Under simplifying assumptions, as $n, p \to \infty$ with $p/n \to \gamma$:

$$\mathbb{E}(f(x_0) - x_0^\top \widehat{\beta})^2 \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \rho^2 \frac{\gamma - 1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \gamma > 1 \end{cases}$$

(Two terms: estimation bias, and estimation variance)

Simplest linear analysis

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Given i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$y_i = \underbrace{x_i^\top \beta}_{f(x_i)} + \epsilon_i, \quad x_i^\top \beta \perp \epsilon_i$$

"Ridgeless" least squares estimator of y on X (which has rows x_i):

$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Let $\sigma^2 = \mathrm{Var}(\epsilon_i)$ [noise energy], $\rho^2 = \mathbb{E} f(x_i)^2$ [signal energy].
Under simplifying assumptions, as $n, p \to \infty$ with $p/n \to \gamma$:

$$\mathbb{E}(f(x_0) - x_0^\top \widehat{\beta})^2 \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \rho^2 \frac{\gamma - 1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \gamma > 1 \end{cases}$$

(Two terms: estimation bias, and estimation variance)

Simplest linear analysis

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Given i.i.d. training data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, where

$$y_i = \underbrace{x_i^\top \beta}_{f(x_i)} + \epsilon_i, \quad x_i^\top \beta \perp \epsilon_i$$

"Ridgeless" least squares estimator of y on X (which has rows x_i):

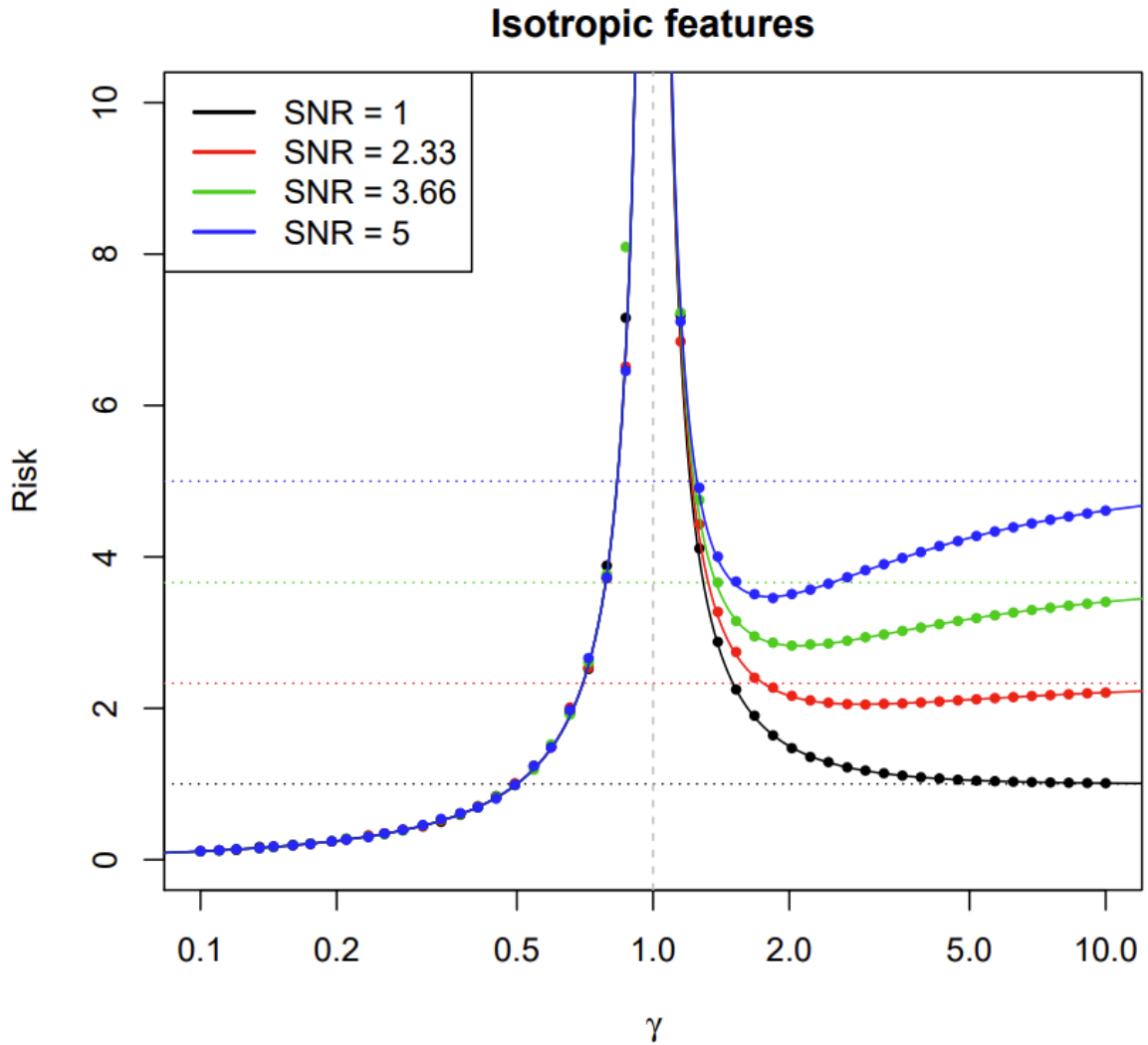
$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

Let $\sigma^2 = \mathrm{Var}(\epsilon_i)$ [noise energy], $\rho^2 = \mathbb{E} f(x_i)^2$ [signal energy].
Under simplifying assumptions, as $n, p \to \infty$ with $p/n \to \gamma$:

$$\mathbb{E}(f(x_0) - x_0^\top \widehat{\beta})^2 \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \gamma < 1 \\ \rho^2 \frac{\gamma - 1}{\gamma} + \sigma^2 \frac{1}{\gamma - 1} & \gamma > 1 \end{cases}$$

(Two terms: estimation bias, and estimation variance)

Double descent in linear regression



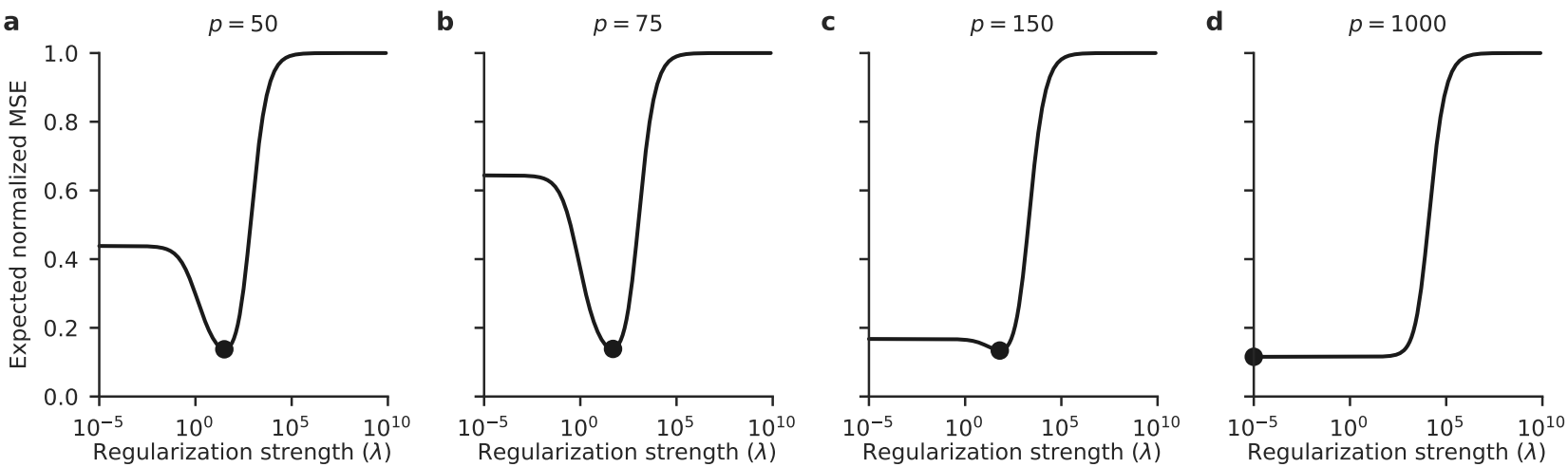
Here $\sigma^2 = 1$, thus signal-to-noise ratio (SNR) is ρ^2, and $\gamma = p/n$.

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

What about ridge regression?

Why don't we just use ridge regression? That is, just take $\lambda > 0$? In part, because $\lambda = 0$ can actually be optimal in high dimensions!

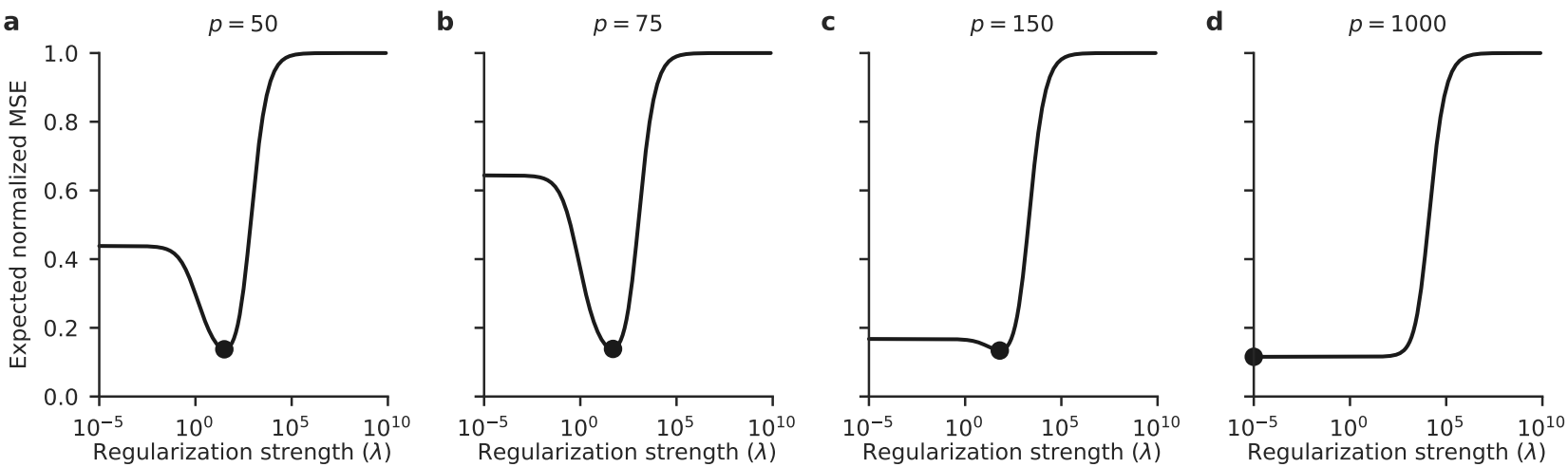
For example, from Kobak et al. (2020):



What about ridge regression?

Why don't we just use ridge regression? That is, just take $\lambda > 0$? In part, because $\lambda = 0$ can actually be optimal in high dimensions!

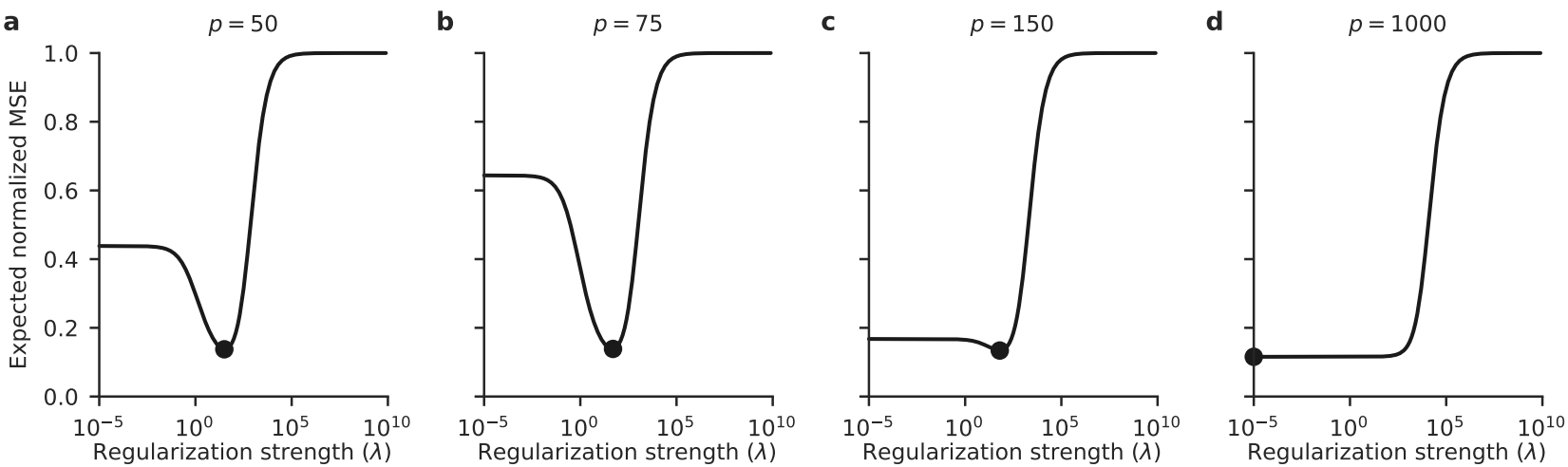
For example, from Kobak et al. (2020):



What about ridge regression?

Why don't we just use ridge regression? That is, just take $\lambda > 0$? In part, because $\lambda = 0$ can actually be optimal in high dimensions!

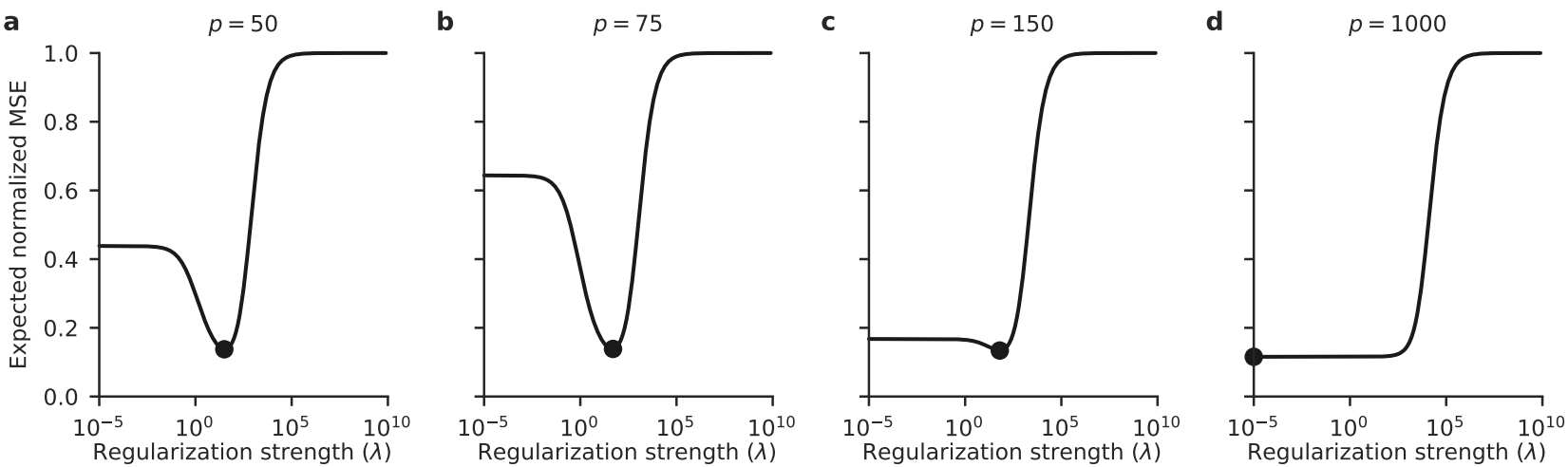
For example, from Kobak et al. (2020):



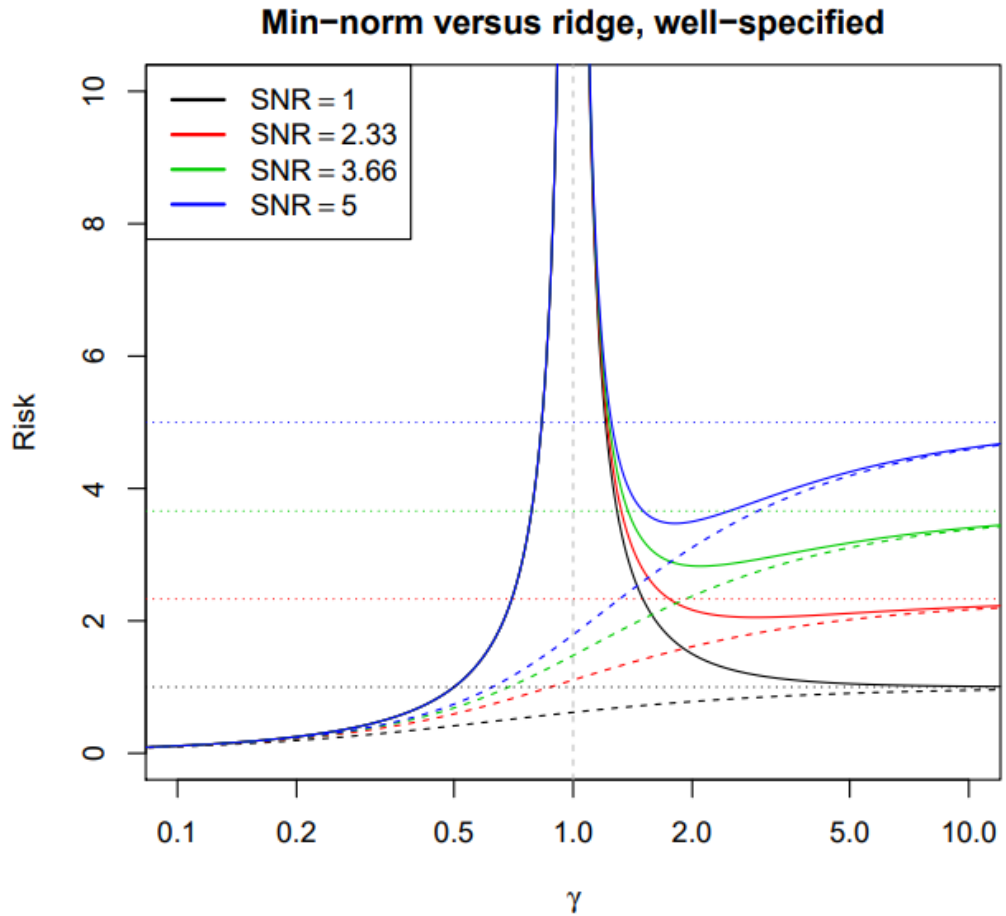
What about ridge regression?

Why don't we just use ridge regression? That is, just take $\lambda > 0$? In part, because $\lambda = 0$ can actually be optimal in high dimensions!

For example, from Kobak et al. (2020):



No double descent with optimal ridge regression?



Here $\sigma^2 = 1$, thus signal-to-noise ratio (SNR) is ρ^2, and $\gamma = p/n$.

Hastie, Montanari, Rosset, Tibshirani, 2019: "Surprises in high-dimensional ridgeless least squares interpolation"

Outline

Ridge regression

- Ordinary ridge regression problem:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \|y - X\beta\|_2^2/n + \lambda\|\beta\|_2^2$$

  - $\lambda \geq 0$ is a tuning parameter
  - optimization problem convex in β and has an explicit solution:

$$\widehat{\beta}(\lambda) = (X^T X/n + \lambda I_p)^{-1} X^T y/n$$

  - extend for any $\lambda \in \mathbb{R}$ using Moore-Penrose pseudo-inverse:

$$\widehat{\beta}(\lambda) = (X^T X/n + \lambda I_p)^{+} X^T y/n$$

- Role of λ:
  - $\lambda = 0$: least squares estimate
  - $\lambda = \infty$: null estimate
  - $\lambda \in (0, \infty)$: fitting liner model versus shrinking the coefficients

Optimal ridge regression (in-distribution setting)

The goal is to study the behavior of asymptotic prediction risk:

$$R(\widehat{\beta}^{\lambda}) = \mathbb{E}_{x_0,y_0}[(y_0 - x_0^{\top}\widehat{\beta}^{\lambda})^2 \mid X, y] \to \mathscr{R}(\lambda, \phi)$$

as the feature size p and the sample size n diverge proportionally to an *aspect ratio* $p/n \to \phi \in (0, \infty)$.

Two questions in the in-distribution (IND) setting when the test point (x_0, y_0) has the same distribution as train data (X, y):

(Q1) What is the behavior of the *optimal ridge penalty*, $\arg\min_{\lambda \geq \lambda_{\min}} \mathscr{R}(\lambda, \phi)$, as a function of parameters such as signal-to-noise ratio, data aspect ratio, feature correlations, and signal structure?

(Q2) What is the behavior of the *optimally tuned ridge risk*, $\min_{\lambda \geq \lambda_{\min}} \mathscr{R}(\lambda, \phi)$, as a function of these same problem parameters?

Ridge prediction risk: bias-variance decomposition

Prediction risk is governed by bias and variance terms.

Let $\widehat{\Sigma} = X^\top X/n$ denote the sample covariance matrix.

Let $\Sigma = \mathbb{E}[\widehat{\Sigma}]$ denote the population covariance matrix.

- Variance term:
$$\sigma^2 \operatorname{tr}(\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}\Sigma)/n$$
  - does not depend on β_0 (true in general for linear smoothers!)
  - decreasing function of λ (variance reduction from ridge shrinkage)

- Bias term:
$$\lambda^2 \beta_0^\top (\widehat{\Sigma} + \lambda I)^{-1}\Sigma(\widehat{\Sigma} + \lambda I)^{-1}\beta_0$$
  - quadratic form in β_0 and strictly non-negative
  - when $p \leq n$, achieves minimum at $\lambda = 0$
  - when $p > n$, may not be minimized at $\lambda = 0$

Ridge prediction risk analysis landscape (IND)

- isotropic Σ, non-random β_0 [Serdobolski '83, Tulino and Verdu '04]
  - follows from [Marchenko-Pastur '67]
  - optimal λ positive
- arbitrary Σ, random β_0, isotropic prior [Dobriban and Wager '18]
  - follows from [Ledoit, Peche '12]
  - optimal λ still positive
- arbitrary Σ, random β_0, source function prior [Richards, et al. '20]
  - follows from [Bao, Pan, Zhou '15] and [Rubio and Mestre '12]
  - optimal λ can be zero
- arbitrary Σ, random β_0, general prior [Wu and Xu '20]
  - extension of [Rubio and Mestre '12]
  - optimal λ can be negative
- arbitrary Σ, non-random β_0 [Hastie, Montanari, Rosset, Tibs. '20]
  - leverage recent random matrix theory [Knowles and Yin '18]
  - optimal λ can be anything

Optimal regularization landscape (IND)

Table 1: **Optimal regularization landscape in ridge regression.** Here, '○' indicates either an isotropic feature or signal covariance, and '⊗' indicates anisotropic features or signal covariance. For the data aspect ratio ϕ, 'all' indicates $\phi \in (0, \infty)$, 'under' indicates $\phi \in (0, 1)$ for the underparameterized regime, and 'over' indicates $\phi \in (1, \infty)$ for the overparameterized regime. For the minimum penalty $\lambda_{\min}$, 'neg' and 'more neg' respectively indicate the naive (loose) and exact lower bound on the negative values (Definition 2.3). For the optimal penalty λ^*, green and red contrast the cases when the sign changes. 'Arb. Mod.', 'Arb. SNR.', and 'Arb. Spec.' indicate allowing for arbitrary response model, signal-to-noise ratio, and feature covariance spectrum, respectively. Please see Table 6 for the reference key.

Σ	β	Σ_0	β_0	$\phi \lessgtr 1$	$\lambda_{\min}$	Arb. Mod.	Arb. SNR	Arb. Spec.	Additional Specific Data Geometry Conditions	λ^*	Reference
In-distribution											
⊗	○	Σ	β	all	zero	✗	✓	✗		+	[DW, Thm. 2.1]
○	⊗	Σ	β	all	zero	✗	✓	✗		+	[HMRT, Cor. 5]
				under	neg	✗	✓	✗		+	[WX, Prop. 6]
				over	neg	✗	✗	✗	Strict misalignment of (Σ, β)	+	[WX, Thm. 4]
				over	neg	✗	✗	✗	Strict alignment of (Σ, β)	−	[WX, Thm. 4, Prop. 7]
⊗	⊗	Σ	β	over	zero	✗	✗	✗	and/or special feature model	0	[RMR, Cor. 2]
				under	more neg	✓	✓	✓		+	**Theorem 3.1** (1)
				over	more neg	✓	✓	✓	General alignment of $(\Sigma, \beta, \sigma^2)$	−	**Theorem 3.1** (2)

Ridge regression under distribution shifts

We consider two types of distribution shifts:

(i) *Covariate shift*: where $P_{x_0} \neq P_x$ but $P_{y_0|x_0} = P_{y|x}$.
(ii) *Regression shift*: where $P_{y_0|x_0} \neq P_{y|x}$ but $P_{x_0} = P_x$.

Again two questions under the out-of-distribution (OOD) setting:

(Q1′) *How does distribution shift alter optimal regularization* λ^**?*
(Q2′) *How does distribution shift alter optimal risk behavior* $\mathscr{R}(\lambda^*, \phi)$*?*

Ridge regression under distribution shifts

We consider two types of distribution shifts:

(i) *Covariate shift*: where $P_{x_0} \neq P_x$ but $P_{y_0|x_0} = P_{y|x}$.
(ii) *Regression shift*: where $P_{y_0|x_0} \neq P_{y|x}$ but $P_{x_0} = P_x$.

Again two questions under the out-of-distribution (OOD) setting:

(Q1′) *How does distribution shift alter optimal regularization* λ^**?*
(Q2′) *How does distribution shift alter optimal risk behavior* $\mathscr{R}(\lambda^*, \phi)$*?*

Optimal regularization landscape (OOD)

Table 1: **Optimal regularization landscape in ridge regression.** Here, '○' indicates either an isotropic feature or signal covariance, and '⊗' indicates anisotropic features or signal covariance. For the data aspect ratio ϕ, 'all' indicates $\phi \in (0, \infty)$, 'under' indicates $\phi \in (0, 1)$ for the underparameterized regime, and 'over' indicates $\phi \in (1, \infty)$ for the overparameterized regime. For the minimum penalty $\lambda_{\min}$, 'neg' and 'more neg' respectively indicate the naive (loose) and exact lower bound on the negative values (Definition 2.3). For the optimal penalty λ^*, green and red contrast the cases when the sign changes. 'Arb. Mod.', 'Arb. SNR.', and 'Arb. Spec.' indicate allowing for arbitrary response model, signal-to-noise ratio, and feature covariance spectrum, respectively. Please see Table 6 for the reference key.

Σ	β	Σ_0	β_0	$\phi \lessgtr 1$	$\lambda_{\min}$	Arb. Mod.	Arb. SNR	Arb. Spec.	Additional Specific Data Geometry Conditions	λ^*	Reference
In-distribution											
⊗	○	Σ	β	all	zero	✗	✓	✗		+	[DW, Thm. 2.1]
○	⊗	Σ	β	all	zero	✗	✓	✗		+	[HMRT, Cor. 5]
				under	neg	✗	✓	✗		+	[WX, Prop. 6]
				over	neg	✗	✗	✗	Strict misalignment of (Σ, β)	+	[WX, Thm. 4]
				over	neg	✗	✗	✗	Strict alignment of (Σ, β)	−	[WX, Thm. 4, Prop. 7]
⊗	⊗	Σ	β	over	zero	✗	✗	✗	and/or special feature model	0	[RMR, Cor. 2]
				under	more neg	✓	✓	✓		+	**Theorem 3.1** (1)
				over	more neg	✓	✓	✓	General alignment of $(\Sigma, \beta, \sigma^2)$	−	**Theorem 3.1** (2)
Out-of-distribution											
⊗	○	Σ_0	β	all	more neg	✓	✓	✓		+	**Proposition 3.2**
⊗	⊗	Σ_0	β	under	more neg	✓	✓	✓		+	**Theorem 3.3** (1)
⊗	⊗	I	β	over	more neg	✓	✓	✓		+	**Theorem 3.3** (2)
○	⊗	Σ_0	β	over	more neg	✓	✓	✓	General alignment of $(\Sigma_0, \beta, \sigma^2)$	−	**Theorem 3.3** (3)
				under	more neg	✓	✓	✓	General alignment of (Σ, β, β_0)	−	**Theorem 3.4** (1), (39)
⊗	⊗	Σ	β_0	under	more neg	✓	✓	✓	General misalignment of (Σ, β, β_0)	+	**Theorem 3.4** (1), (39)
				over	more neg	✓	✓	✓	General alignment of $(\Sigma, \beta, \beta_0, \sigma^2)$	−	**Theorem 3.4** (2)

Data assumptions and lower bound on regularization

Data assumptions:

- Covariate: Each feature vector x_i for $i \in [n]$ can be decomposed as $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ contains i.i.d. entries z_{ij} for $j \in [p]$ with mean 0, variance 1, and bounded $4 + \mu$ moments for some $\mu > 0$.
- Response: Each response variable y_i for $i \in [n]$ has mean 0, and bounded $4 + \mu$ moments.

Lower bound on λ:

- Let $\mu_{\min} \in \mathbb{R}$ be the unique solution, satisfying $\mu_{\min} > -r_{\min}$, to the equation:

$$1 = \phi \, \bar{\mathrm{tr}}[\Sigma^2 (\Sigma + \mu_{\min} I)^{-2}],$$

  and let $\lambda_{\min}(\phi)$ be given by:

$$\lambda_{\min}(\phi) = \mu_{\min} - \phi \, \bar{\mathrm{tr}}[\Sigma (\Sigma + \mu_{\min} I)^{-1}].$$

Data assumptions and lower bound on regularization

Data assumptions:

- Covariate: Each feature vector x_i for $i \in [n]$ can be decomposed as $x_i = \Sigma^{1/2} z_i$, where $z_i \in \mathbb{R}^p$ contains i.i.d. entries z_{ij} for $j \in [p]$ with mean 0, variance 1, and bounded $4 + \mu$ moments for some $\mu > 0$.
- Response: Each response variable y_i for $i \in [n]$ has mean 0, and bounded $4 + \mu$ moments.

Lower bound on λ:

- Let $\mu_{\min} \in \mathbb{R}$ be the unique solution, satisfying $\mu_{\min} > -r_{\min}$, to the equation:

$$1 = \phi \, \bar{\mathrm{tr}}[\Sigma^2 (\Sigma + \mu_{\min} I)^{-2}],$$

  and let $\lambda_{\min}(\phi)$ be given by:

$$\lambda_{\min}(\phi) = \mu_{\min} - \phi \, \bar{\mathrm{tr}}[\Sigma (\Sigma + \mu_{\min} I)^{-1}].$$

Out-of-distribution ridge risk characterization

The OOD risk decomposes into:

$$\mathscr{R}(\lambda,\phi) := \underbrace{\mathscr{B}(\lambda,\phi)}_{\text{bias}} + \underbrace{\mathscr{V}(\lambda,\phi)}_{\text{variance}} + \underbrace{\mathscr{E}(\lambda,\phi)}_{\text{extra bias}} + \underbrace{\kappa^2}_{\text{irreducible error}},$$

where

$$\begin{aligned}
\mathscr{B} &= \mu^2 \cdot \beta^\top (\Sigma + \mu I)^{-1} (\widetilde{\nu}\Sigma + \Sigma_0)(\Sigma + \mu I)^{-1}\beta, \\
\mathscr{V} &= \sigma^2 \widetilde{\nu}, \\
\mathscr{E} &= 2\mu \cdot \beta^\top (\Sigma + \mu I)^{-1}\Sigma_0(\beta_0 - \beta), \\
\kappa^2 &= (\beta_0 - \beta)^\top \Sigma_0 (\beta_0 - \beta) + \sigma_0^2.
\end{aligned}$$

The optimal regularization is defined as:

$$\lambda^* \in \operatorname*{arg\,min}_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda,\phi).$$

Out-of-distribution ridge risk characterization

The OOD risk decomposes into:

$$\mathscr{R}(\lambda,\phi) := \underbrace{\mathscr{B}(\lambda,\phi)}_{\text{bias}} + \underbrace{\mathscr{V}(\lambda,\phi)}_{\text{variance}} + \underbrace{\mathscr{E}(\lambda,\phi)}_{\text{extra bias}} + \underbrace{\kappa^2}_{\text{irreducible error}},$$

where

$$\begin{aligned}
\mathscr{B} &= \mu^2 \cdot \beta^\top (\Sigma + \mu I)^{-1} (\widetilde{\nu}\Sigma + \Sigma_0)(\Sigma + \mu I)^{-1}\beta, \\
\mathscr{V} &= \sigma^2 \widetilde{\nu}, \\
\mathscr{E} &= 2\mu \cdot \beta^\top (\Sigma + \mu I)^{-1}\Sigma_0(\beta_0 - \beta), \\
\kappa^2 &= (\beta_0 - \beta)^\top \Sigma_0 (\beta_0 - \beta) + \sigma_0^2.
\end{aligned}$$

The optimal regularization is defined as:

$$\lambda^* \in \operatorname*{arg\,min}_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda,\phi).$$

Out-of-distribution ridge risk characterization

The OOD risk decomposes into:

$$\mathscr{R}(\lambda, \phi) := \underbrace{\mathscr{B}(\lambda, \phi)}_{\text{bias}} + \underbrace{\mathscr{V}(\lambda, \phi)}_{\text{variance}} + \underbrace{\mathscr{E}(\lambda, \phi)}_{\text{extra bias}} + \underbrace{\kappa^2}_{\text{irreducible error}},$$

where

$$\begin{aligned}
\mathscr{B} &= \mu^2 \cdot \beta^\top (\Sigma + \mu I)^{-1} (\widetilde{\nu}\Sigma + \Sigma_0)(\Sigma + \mu I)^{-1} \beta, \\
\mathscr{V} &= \sigma^2 \widetilde{\nu}, \\
\mathscr{E} &= 2\mu \cdot \beta^\top (\Sigma + \mu I)^{-1} \Sigma_0 (\beta_0 - \beta), \\
\kappa^2 &= (\beta_0 - \beta)^\top \Sigma_0 (\beta_0 - \beta) + \sigma_0^2.
\end{aligned}$$

The optimal regularization is defined as:

$$\lambda^* \in \underset{\lambda \geq \lambda_{\min}(\phi)}{\arg\min}\, \mathscr{R}(\lambda, \phi).$$

Optimal regularization sign characterization (IND)

Theorem. Sign of optimal regularization (in-distribution setting):

1. *(Underparameterized)* When $\phi < 1$, we have $\lambda^* \geq 0$.
2. *(Overparameterized)* When $\phi > 1$, if for all $v < 1/\mu(0, \phi)$, the following general alignment holds:

$$\frac{\bar{\mathrm{tr}}[B\Sigma(v\Sigma + I)^{-2}] + \sigma^2}{\bar{\mathrm{tr}}[B\Sigma(v\Sigma + I)^{-3}] + \sigma^2} > \frac{\bar{\mathrm{tr}}[\Sigma(v\Sigma + I)^{-2}]}{\bar{\mathrm{tr}}[\Sigma(v\Sigma + I)^{-3}]}, \tag{1}$$

where $B = \beta\beta^\top$, then we have $\lambda^* < 0$.

- Alignment condition (1) captures how well the signal B is aligned with the feature covariance Σ.
- λ^* could be negative in the overparameterized regime when $p > n$.

Illustration for optimal regularization sign (IND)

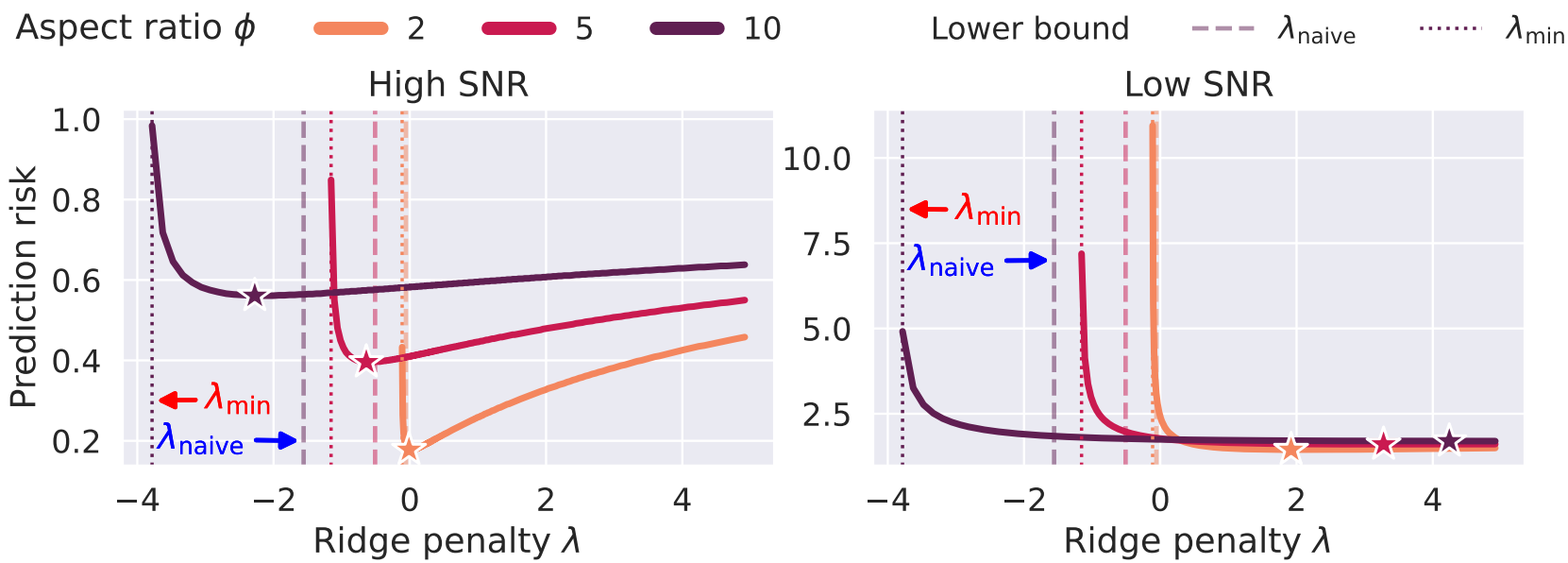


Figure: Illustration of negative or positive optimal regularization under general alignment.

- λ^* can be smaller than the previous bound.
- The more the alignment (seen as a function of SNR), the lower λ^*; the more the misalignment, the higher λ^* (seen as a function of SNR).

Optimal regularization sign characterization (OOD, I)

Theorem. Sign of optimal regularization (covariate shift setting):

1. *(Underparameterized)* When $\phi < 1$, we have $\lambda^* \geq 0$.
2. *(Overparameterized)* When $\phi > 1$, if $\Sigma_0 = I$ (corresponding to the estimation risk), then we have $\lambda^* \geq 0$.
3. *(Overparameterized)* When $\phi > 1$, if $\Sigma = I$ and

$$\bar{\mathrm{tr}}[\Sigma_0 B] > \bar{\mathrm{tr}}[\Sigma_0]\left(\bar{\mathrm{tr}}[B] + \frac{(1+\mu(0,\phi))^3}{\mu(0,\phi)^3}\sigma^2\right), \tag{2}$$

where $B = \beta\beta^\top$, then we have $\lambda^* < 0$.

- The isotropic test covariance case ($\Sigma_0 = I$) is similar to underparameterized cases.
- Alignment condition (2) captures how the well the signal B aligned with covariance matrix of test features Σ_0.
- λ^* can be negative even in the isotropic train covariance case ($\Sigma = I$).

Optimal regularization sign characterization (OOD, II)

Theorem. Sign of optimal regularization (label shift setting):

1. *(Underparameterized)* When $\phi < 1$, if $\sigma^2 = o(1)$ and for all $\mu \geq 0$, the following general alignment holds:

$$\bar{\mathrm{tr}}[B_0\Sigma^2(\Sigma+\mu I)^{-2}] > \bar{\mathrm{tr}}[B\Sigma^2(\Sigma+\mu I)^{-2}], \tag{3}$$

   where $B = \beta\beta^\top$ and $B_0 = \beta_0\beta^\top$, then we have $\lambda^* < 0$.

2. *(Overparameterized)* When $\phi > 1$, if the general alignment conditions (1) and (3) hold, then we have $\lambda^* < 0$.

- λ^* can be negative even if the design is underparameterized!

Illustration (optimal OOD regularization)

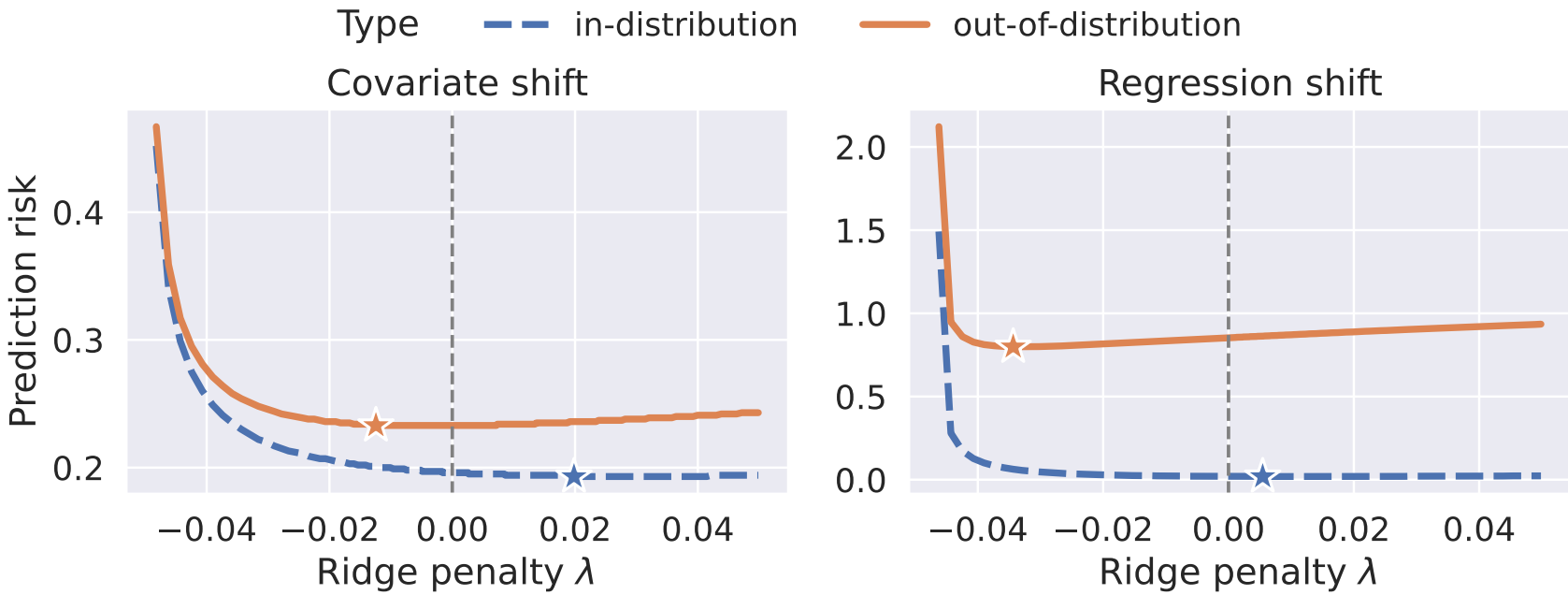


Figure: Covariate and regression shift can lead to negative optimal regularization in both underparameterized and overparameterized regimes.

- The design is isotropic on the left.
- The design is underparameterized on the right.

Optimal risk monotonicity

Theorem. Monotonicity of optimal ridge risk (both IND and OOD settings): The map $\phi \mapsto \min_{\lambda \geq \lambda_{\min}(\phi)} \mathscr{R}(\lambda, \phi)$ is monotonically increasing in ϕ.

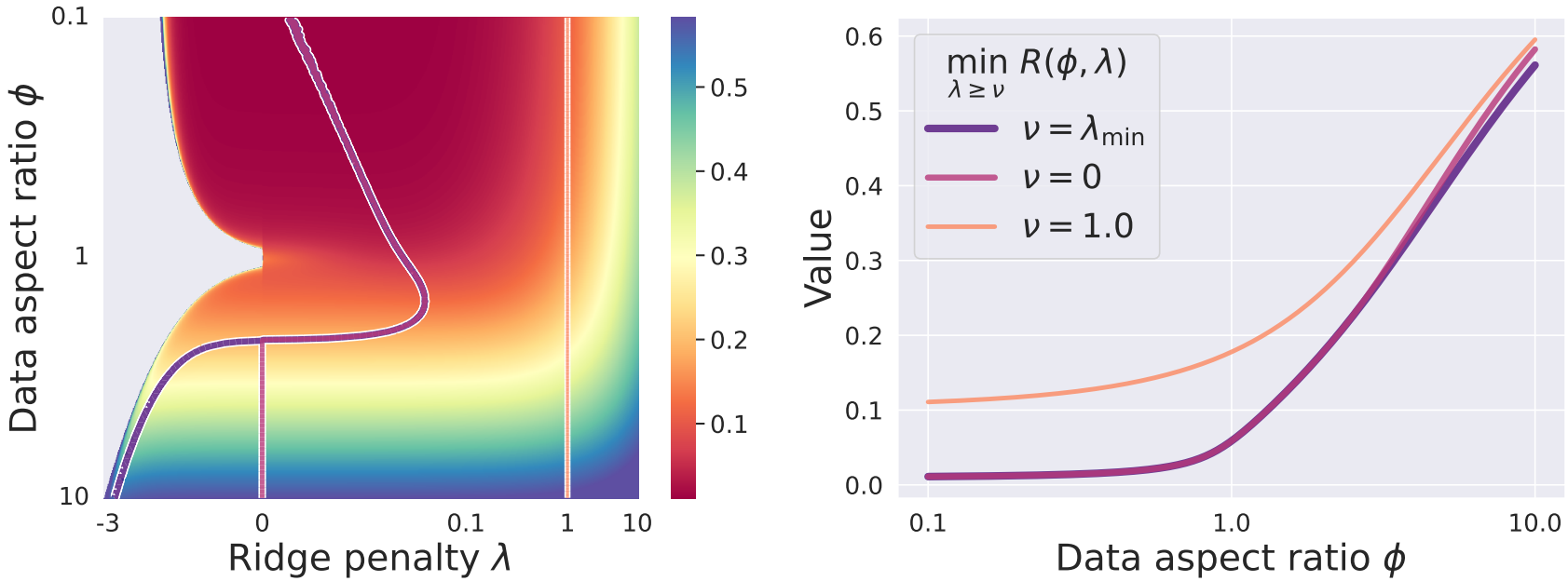


Figure: Ridge regression optimized over $\lambda \geq \nu$ for different thresholds ν has monotonic risk profile.

Outline

Takeaways

- Optimal regularization behavior:
  - The optimal ridge regularization level can be positive, zero, or negative, depending on the alignment between train and test distributions.
  - Negative regularization can be optimal under certain covariate and regression shifts.
- Risk monotonicity:
  - The optimally tuned ridge risk is monotonic in the data aspect ratio (ϕ), even in the OOD setting.
  - This monotonicity holds for both the in-distribution (IND) and OOD scenarios, provided optimal regularization is used.
- General conditions:
  - The established conditions do not rely on specific models for train or test distributions, allowing for arbitrary shifts and a wide range of regularization levels.

Future directions

- Lasso and other penalized estimators:
  - Extend the analysis to the lasso and other convex penalized M-estimators.
  - Empirical evidence suggests similar monotonic risk behavior under optimal tuning.

- Data-dependent regularization tuning:
  - Develop practical methods for data-dependent tuning of ridge penalties in the OOD setting.
  - Investigate techniques to ensure optimal regularization in real-world applications.

Thanks for listening!

Questions/comments/thoughts?