Mitigating multiple descents:

A model-agnostic framework for risk monotonization

Pratik Patil

University of California Berkeley

Rutgers University Seminar

November 2023

Based on joint work with the following amazing collaborators:

- Arun Kuchibhotla (Carnegie Mellon University)
- Yuting Wei (University of Pennsylvania)
- Alessandro Rinaldo (University of Texas)
- Jin-Hong Du (Carnegie Mellon University)

References on risk monotonization: Subsampling, ensembling, and ridge regularization

1. Mitigating multiple descents: A model-agnostic framework for risk monotonization (joint with Arun Kuchibhotla, Yuting Wei, Alessandro Rinaldo) [benefits of subsampling]

2. Bagging in overparameterized learning: Risk characterization and risk monotonization (joint with Jin-Hong Du, Arun Kuchibhotla) [benefits of ensembling]

3. Generalized equivalences between subsampling and ridge regularization (joint with Jin-Hong Du) [connections to ridge]

References on risk monotonization: Subsampling, ensembling, and ridge regularization

1. Mitigating multiple descents: A model-agnostic framework for risk monotonization (joint with Arun Kuchibhotla, Yuting Wei, Alessandro Rinaldo) [benefits of subsampling]

2. Bagging in overparameterized learning: Risk characterization and risk monotonization (joint with Jin-Hong Du, Arun Kuchibhotla) [benefits of ensembling]

3. Generalized equivalences between subsampling and ridge regularization (joint with Jin-Hong Du) [connections to ridge]

References on risk monotonization: Subsampling, ensembling, and ridge regularization

1. Mitigating multiple descents: A model-agnostic framework for risk monotonization (joint with Arun Kuchibhotla, Yuting Wei, Alessandro Rinaldo) [benefits of subsampling]

2. Bagging in overparameterized learning: Risk characterization and risk monotonization (joint with Jin-Hong Du, Arun Kuchibhotla) [benefits of ensembling]

3. Generalized equivalences between subsampling and ridge regularization (joint with Jin-Hong Du) [connections to ridge]

References on risk monotonization: Subsampling, ensembling, and ridge regularization

1. Mitigating multiple descents: A model-agnostic framework for risk monotonization (joint with Arun Kuchibhotla, Yuting Wei, Alessandro Rinaldo) [benefits of subsampling]

2. Bagging in overparameterized learning: Risk characterization and risk monotonization (joint with Jin-Hong Du, Arun Kuchibhotla) [benefits of ensembling]

3. Generalized equivalences between subsampling and ridge regularization (joint with Jin-Hong Du) [connections to ridge]

Outline

Overparametrization 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.

Overparametrization 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.

Overparametrization 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.

Overparametrization 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.

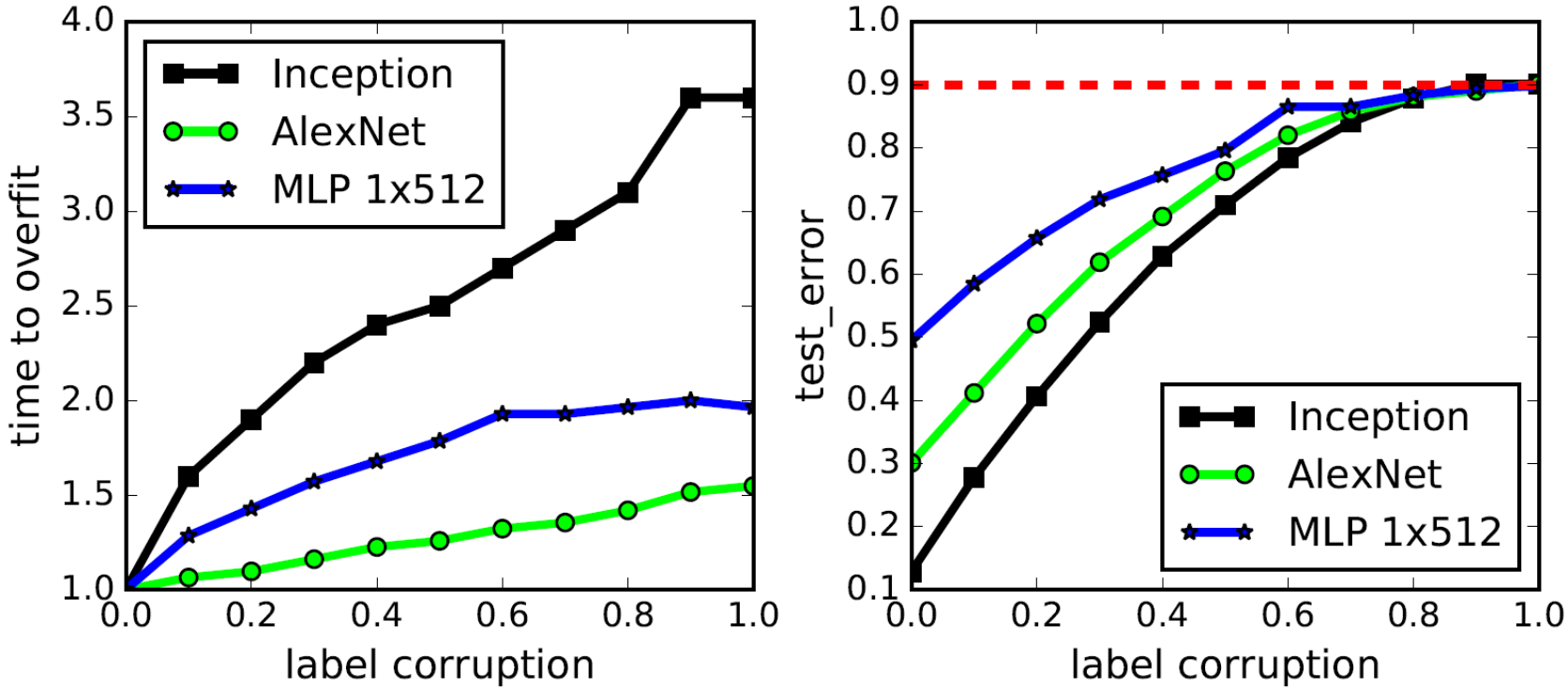
Overparametrization in machine learning

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.

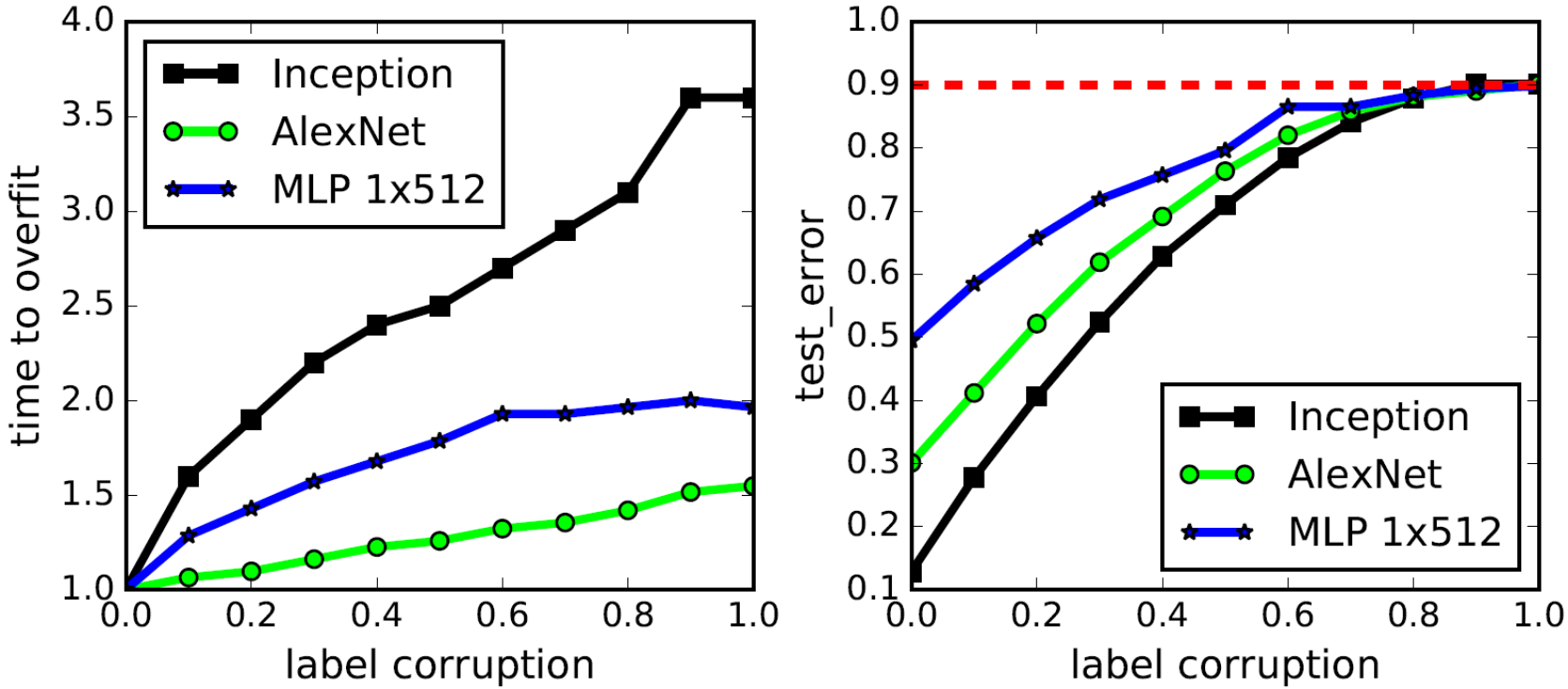
An influential experiment



"Understanding deep learning requires rethinking generalization"
Zhang, Bengio, Hardt, Recht, Vinyals, 2017

- CIFAR10 data (60,000 images [32 × 32]) with artificial label noise
- Three neural network architectures (with number of parameters): Inception (1,649,402), AlexNet (1,387,786), MLP 1x512 (1,209,866)

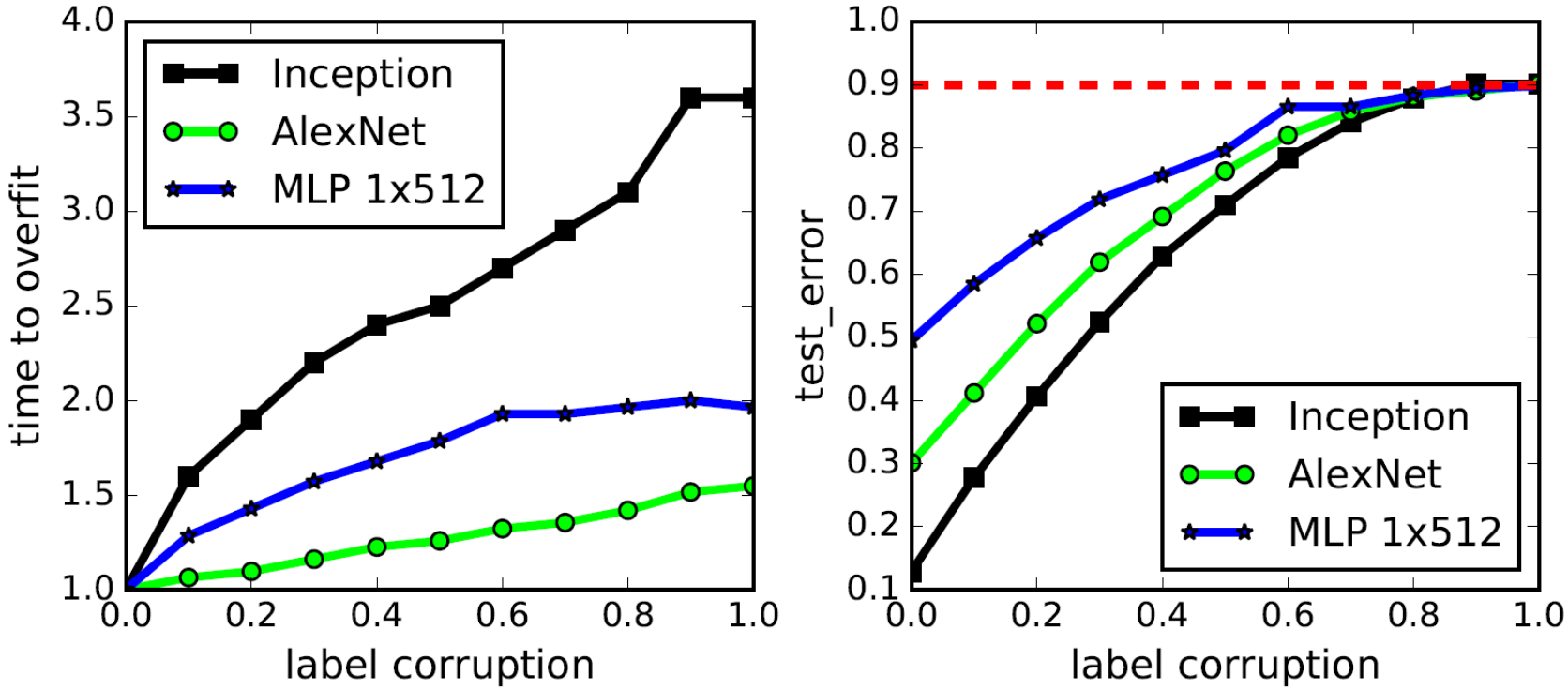
An influential experiment



"Understanding deep learning requires rethinking generalization"
Zhang, Bengio, Hardt, Recht, Vinyals, 2017

- CIFAR10 data (60,000 images [32 $\times$ 32]) with artificial label noise
- Three neural network architectures (with number of parameters): Inception (1,649,402), AlexNet (1,387,786), MLP 1x512 (1,209,866)

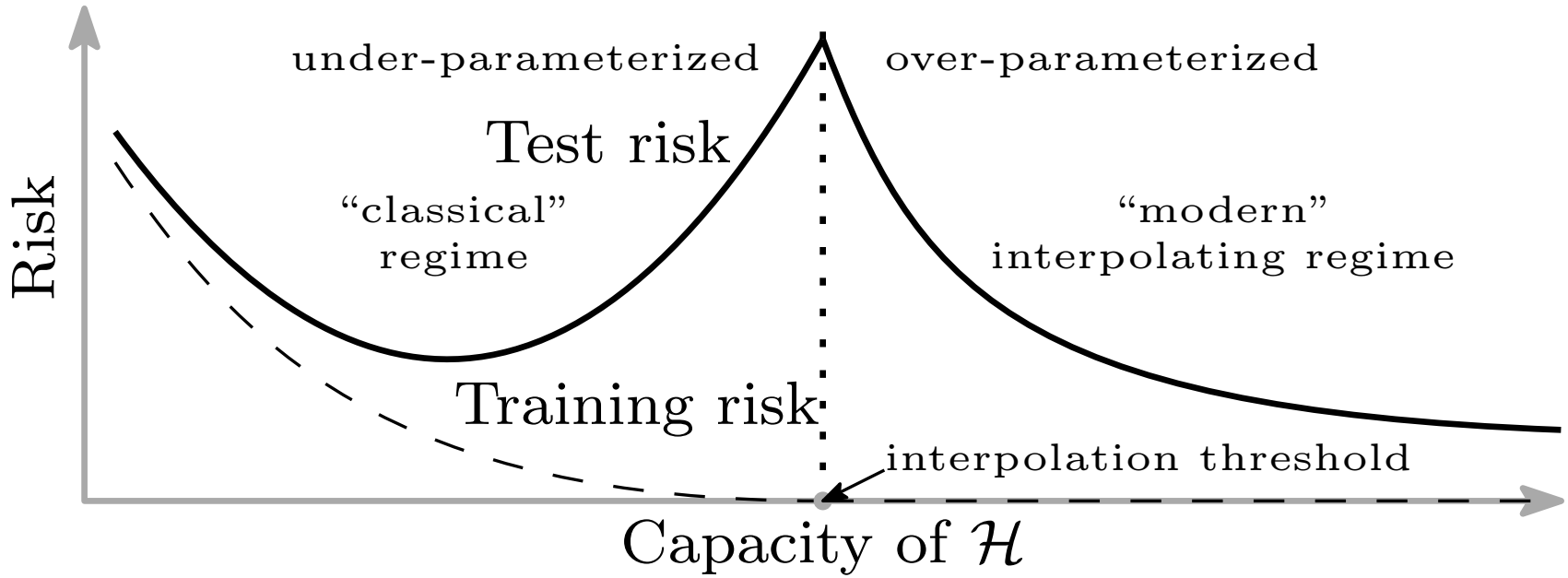
An influential experiment



"Understanding deep learning requires rethinking generalization"
Zhang, Bengio, Hardt, Recht, Vinyals, 2017

- CIFAR10 data (60,000 images [32 $\times$ 32]) with artificial label noise
- Three neural network architectures (with number of parameters): Inception (1,649,402), AlexNet (1,387,786), MLP 1x512 (1,209,866)

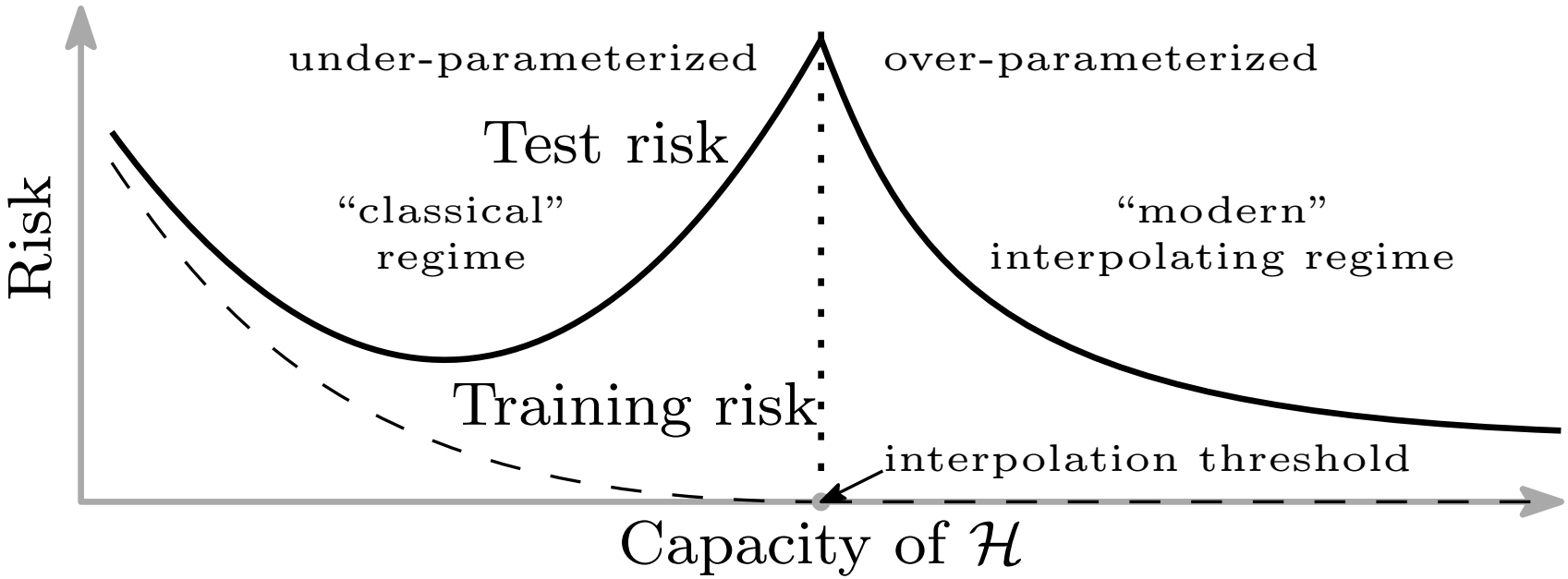
Peculiar generalization behavior: double descent



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.

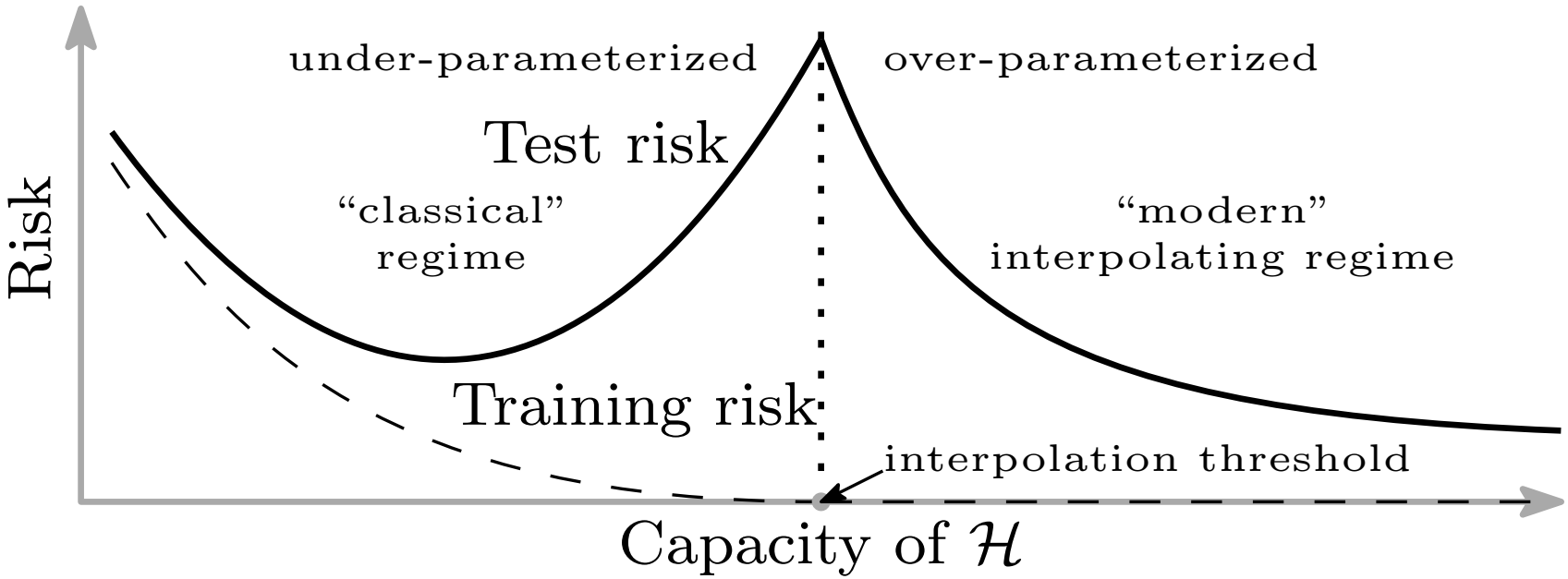
Peculiar generalization behavior: double descent



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 overparametrized
- 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 overparametrized
- 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 overparametrized
- 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 overparametrized
- 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 overparametrized
- 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 overparametrized
- 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 overparametrized
- 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):

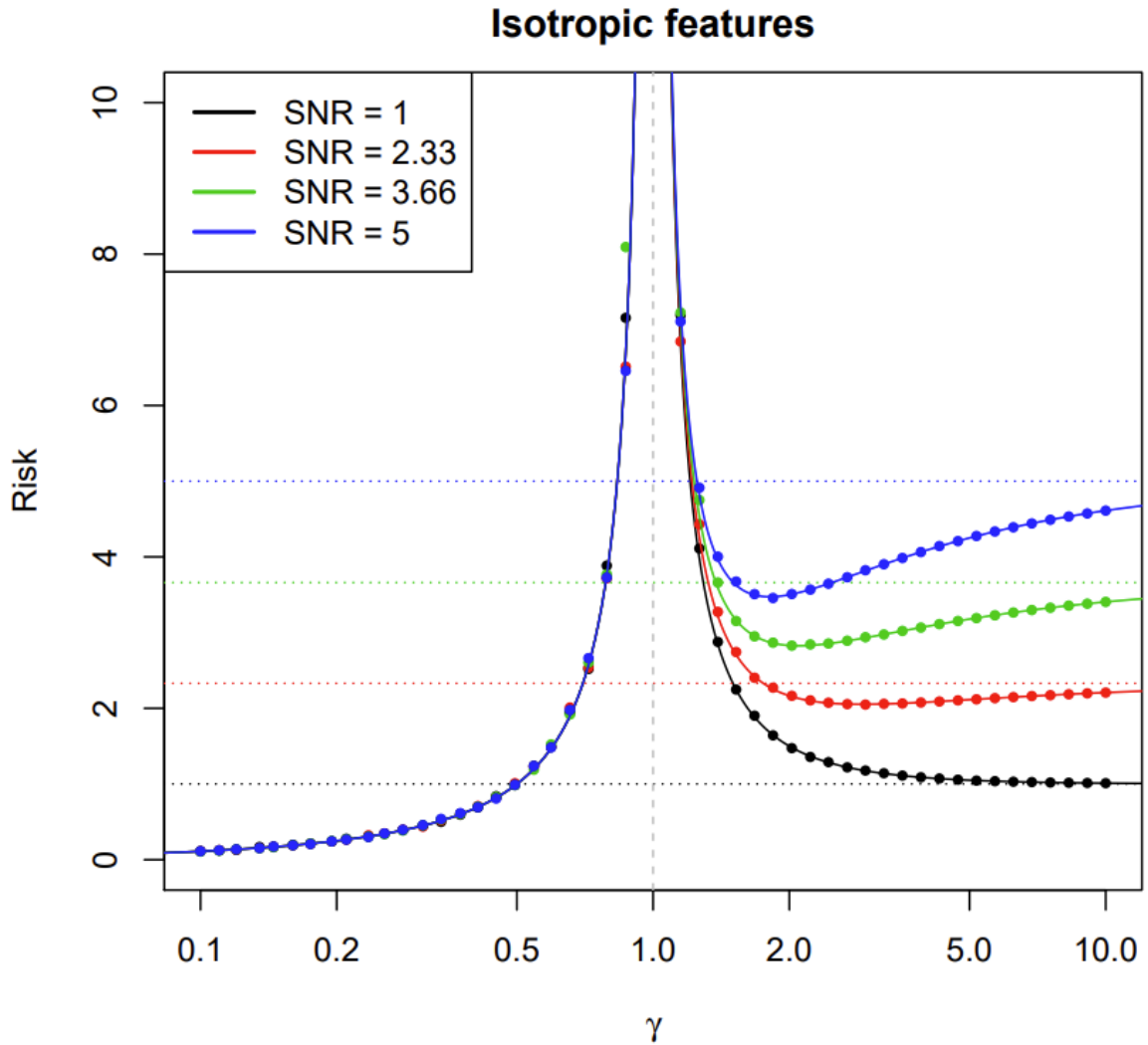
$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

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 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"

Double descent interpretations

- The risk first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double descent interpretations

- The risk first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double descent interpretations

- The risk first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double descent interpretations

- The risk first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double descent interpretations

- The risk first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Double descent interpretations

- The risk first increases as p/n increases up to some threshold and then decreases.
- There are two ways to view this:
  - If p is thought of as fixed (large value), this implies that as sample size increases the risk first decreases and then increases.
    More data hurts.
  - If n is thought of as fixed (large value), this implies that as the number of features/covariates increase the risk first increases and then decreases.
    More features do not hurt.
- We will focus on the first interpretation: more data can hurt.

Outline

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

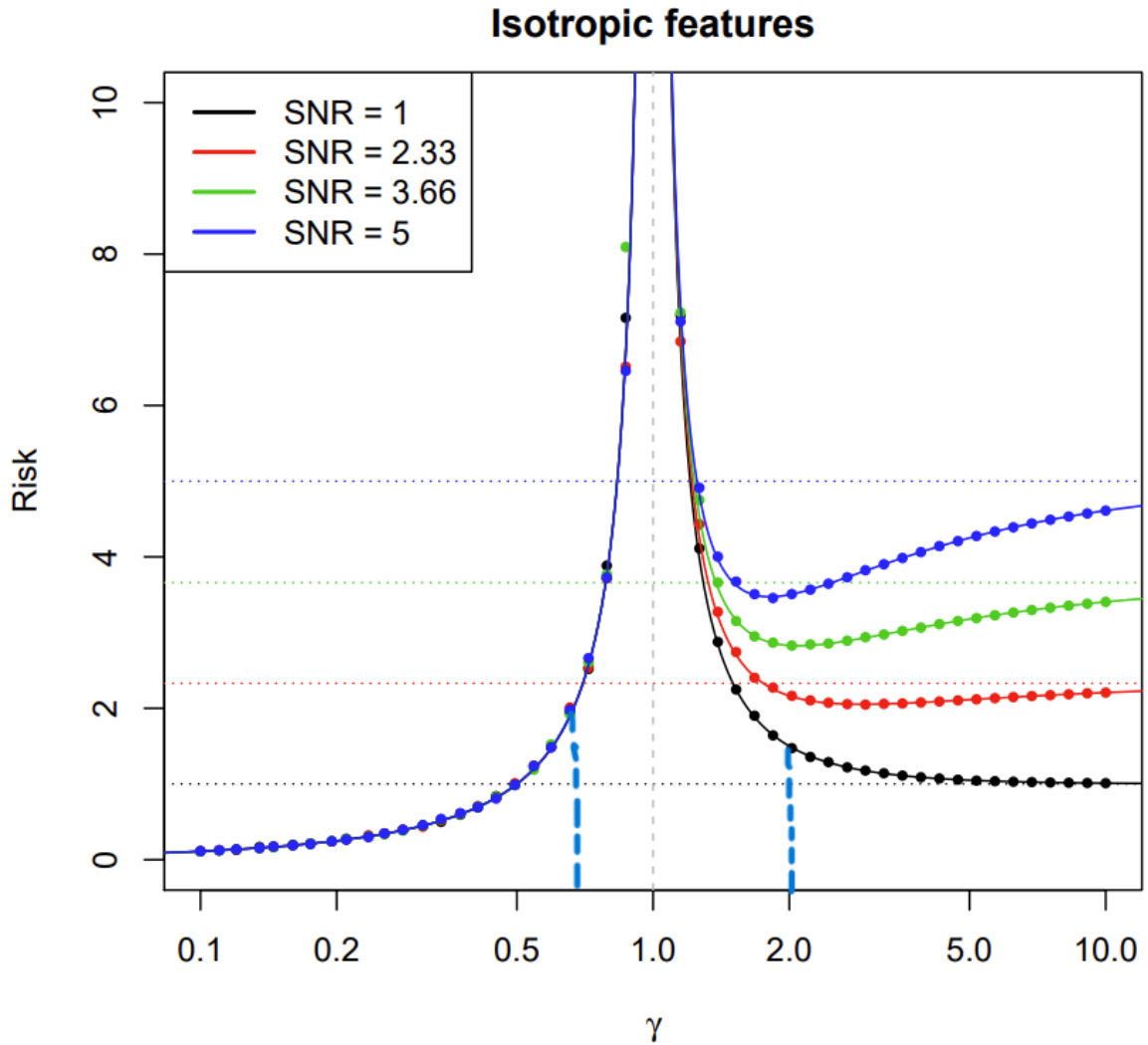
Motivation and main punchlines

- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure leading to worse risk as the number of observations increases is not using the data properly.

Key question: Can we modify any prediction procedure to mitigate the double or multiple descent behavior and achieve a monotonic risk behavior?

We propose two methods, dubbed zero-step and one-step, that take an input an arbitrary procedure and return a modified procedure that has a monotonic risk behavior. The main idea is that of subsampling.

Method overview and the problem



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

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

The problem

- Given a number of observations (n) and a number of features (p), how do we know if a lesser number of observations would actually yield a better risk?
- What is the best sample size to reduce the dataset in order to attain the best possible risk?

Solution: cross-validation.

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Basic idea of zero-step procedure

Given any arbitrary prediction procedure at a given aspect ratio $\gamma = p/n$:

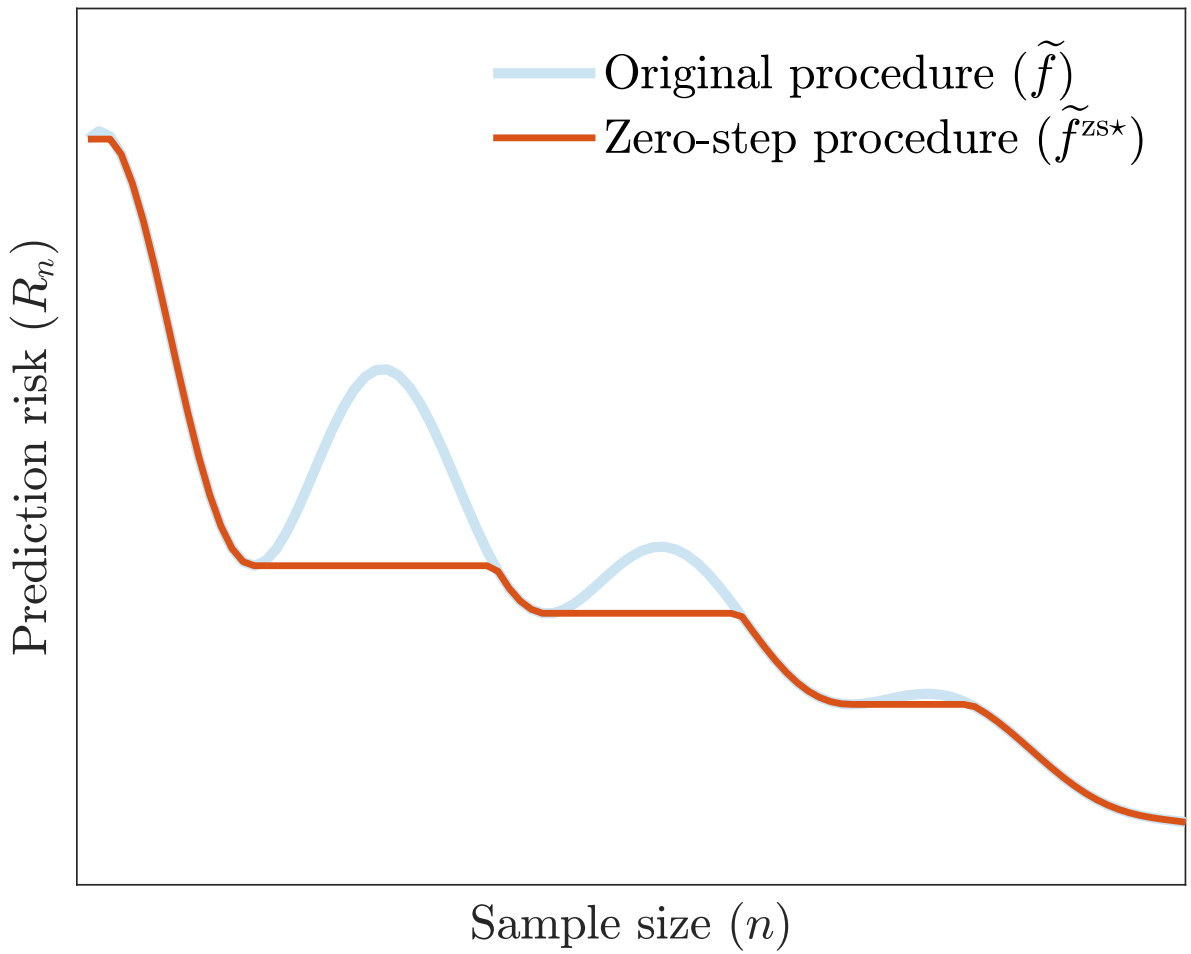
1. Risk estimation: construct a (dense grid of) aspect ratios $\geq \gamma$ by using datasets of sizes smaller than n, and estimate risks on test set
2. Model selection: select aspect ratio that delivers the smallest estimated risk and return the corresponding predictor
3. Risk monotonization: show that the risk profile of the resulting procedure is asymptotically monotone in the aspect ratio

Method highlights:

- applicable to generic (e.g black-box) prediction methods and common classification and regression loss functions
- model agnostic and requires minimal distributional assumptions
- works for procedures with diverging risks at some aspect ratios

Risk monotonization illustration

If R_n represents the "risk" of a procedure at sample size n, then by risk monotonization we mean a procedure with risk $\min_{m \leq n} R_m$.



Split sample cross-validation

- Given data $\mathcal{D}_n$ of n i.i.d. observations and a prediction procedure $\widetilde{f}$, split $\mathcal{D}_n$ into training data $\mathcal{D}_{\text{tr}}$ with $n(1 - 1/\log n)$ observations and test data $\mathcal{D}_{\text{te}}$ with $n/\log n$ observations.
- Note that
$$\lim_n \frac{p}{n} = \lim_n \frac{p}{n(1 - 1/\log n)}.$$
- For $n^{1/2} \leq k \leq |\mathcal{D}_{\text{tr}}|$, obtain a predictor $\widetilde{f}_k$ by training $\widetilde{f}$ on a subset of $\mathcal{D}_{\text{tr}}$ with k observations.
- If p/n converges to γ as $n \to \infty$, then
$$\left\{ \frac{p}{n^{1/2}}, \frac{p}{n^{1/2}+1}, \ldots, \frac{p}{|\mathcal{D}_{\text{tr}}|} \right\} \quad " \to " \quad [\gamma, \infty].$$
The set of aspect ratios for the predictors $\widetilde{f}_k$ covers $[\gamma, \infty]$.
- Choose one out of $\widetilde{f}_k$, $n^{1/2} \leq k \leq |\mathcal{D}_{\text{tr}}|$ using an estimate of out-of-sample risk computed from $\mathcal{D}_{\text{te}}$ This is split sample cross-validation.

Cross-validation risk estimate

- Traditionally, the risk of a predictor based on a test data is done via average loss. For example, with squared error loss, the traditional estimate of (prediction) risk of a predictor $\widetilde{f}_k$

$$\widehat{R}(\widetilde{f}_k) := \frac{1}{|\mathcal{D}_{\text{te}}|} \sum_{j \in \mathcal{D}_{\text{te}}} (Y_j - \widetilde{f}_k(X_j))^2.$$

- For a good performance simultaneously over $O(n)$ predictors and also to avoid strong tail assumptions on the loss, we also consider the median-of-means estimator.

- With either the average or median-of-means estimator of risk, we return the predictor $\widehat{f} := \widetilde{f}_{\widehat{k}}$ where

$$\widehat{k} := \operatorname*{argmin}_{n^{1/2} \le k \le |\mathcal{D}_{\text{tr}}|} \widehat{R}(\widetilde{f}_k).$$

- $\widehat{k}$ represents the "best" sample size to use for the given number of features in the dataset and $\widetilde{f}_{\widehat{k}}$ is what we call a zero-step predictor that achieves risk monotonization.

Risk monotonization guarantee

Theorem. Under the proportional asymptotics regime ($p/n \to \gamma$), and a mild assumption on the convergence of the prediction risk of $\widehat{f}$ trained on datasets with a limiting aspect ratio ζ converges to $R^{\text{det}}(\zeta; \widehat{f})$, we show:

$$R(\widehat{f}^{\text{cv}}) \;=\; \inf_{\zeta \in [\gamma, \infty]} R^{\text{det}}(\zeta; \widehat{f}) \;\times\; (1 + o_p(1)).$$

This shows that the zero-step predictor has a monotone risk in terms of the sample size and hence with respect to the limiting aspect ratio.

This is a model-free result in that no parametric model is assumed for the data. This is unlike most results in overparametrized learning which require stringent assumptions.

Risk monotonization guarantee

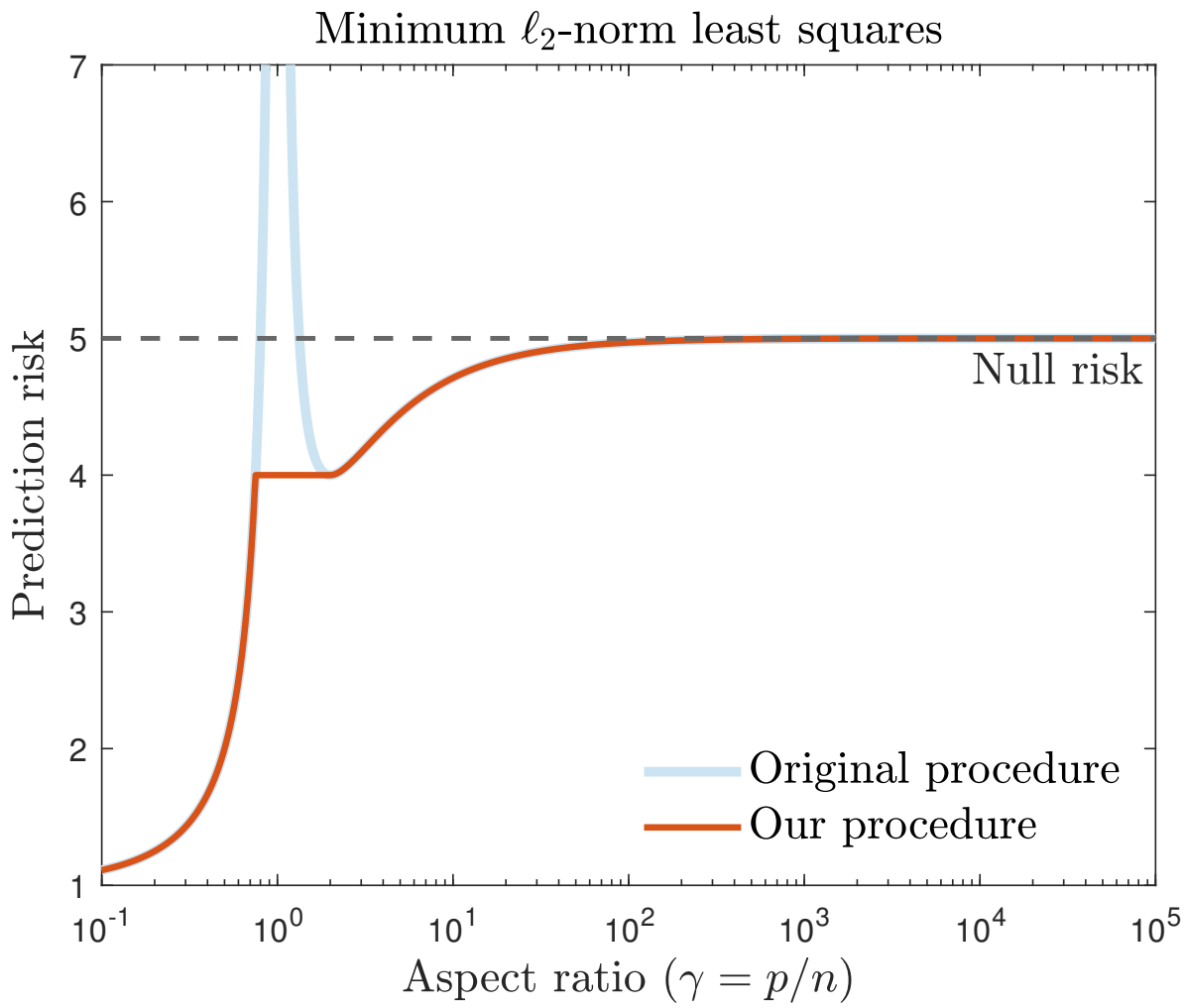
Theorem. Under the proportional asymptotics regime ($p/n \to \gamma$), and a mild assumption on the convergence of the prediction risk of $\widehat{f}$ trained on datasets with a limiting aspect ratio ζ converges to $R^{\text{det}}(\zeta; \widehat{f})$, we show:

$$R(\widehat{f}^{\text{cv}}) \;=\; \inf_{\zeta \in [\gamma, \infty]} R^{\text{det}}(\zeta; \widehat{f}) \;\times\; (1 + o_p(1)).$$

This shows that the zero-step predictor has a monotone risk in terms of the sample size and hence with respect to the limiting aspect ratio.

This is a model-free result in that no parametric model is assumed for the data. This is unlike most results in overparametrized learning which require stringent assumptions.

Risk monotonization (illustration)



Takeaways and extensions

Takeaways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Takeaways and extensions

Takeaways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Takeaways and extensions

Takeaways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Takeaways and extensions

Takeaways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Takeaways and extensions

Takeaways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Takeaways and extensions

Takeaways:

- We have introduced the zero-step prediction procedure that provably monotonizes the risk of a given predictor.
- The main idea is cross-validation based on test data, but with splitting done so as to maintain the limiting aspect ratio.

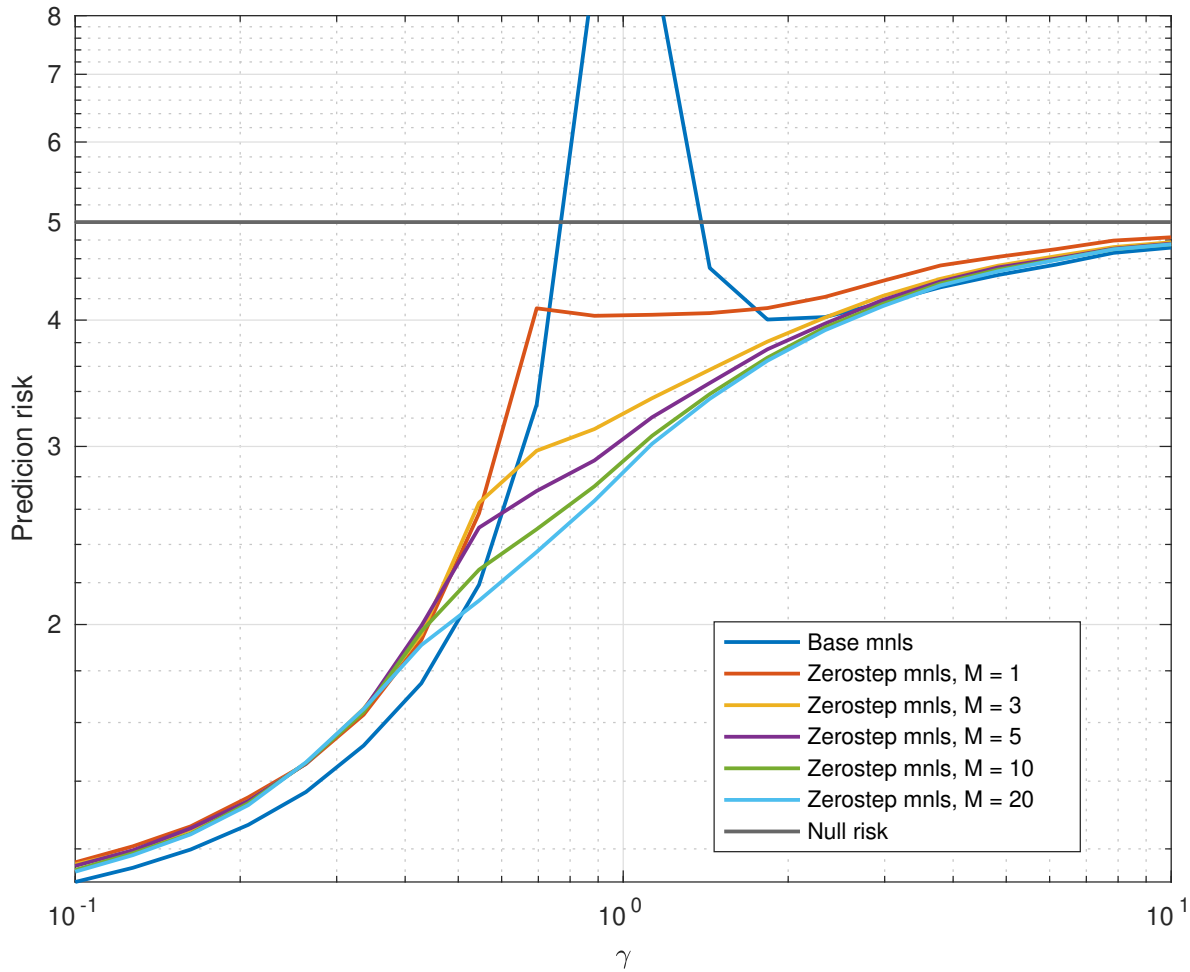
Extensions:

- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

Outline

Motivation beyond bagging analysis

Key question: How much improvement do we get if we use an ensemble of $M > 1$ subsampled datasets, rather than just a single subsampled dataset?



We provide precise risk characterization for ridgeless (and ridge) ensemles.

Ridge ensembles

- Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) \in \mathbb{R}^p \times \mathbb{R} : j \in [n]\}$ denote a dataset. The ridge estimator fitted on subsampled dataset $\mathcal{D}_I$ with $I \subseteq [n], |I| = k$ is:

$$\widehat{\beta}_k^{\lambda}(\mathcal{D}_I) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{k} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

- For $\lambda \geq 0$ fixed, ensemble ridge estimator is:

$$\widetilde{\beta}_{k,M}^{\lambda}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^M) := \frac{1}{M} \sum_{\ell \in [M]} \widehat{\beta}_k^{\lambda}(\mathcal{D}_{I_\ell}),$$

  with $I_1, \ldots, I_M \sim \mathcal{I}_k := \{\{i_1, \ldots, i_k\} : 1 \leq i_1 < \ldots < i_k \leq n\}$. The *full-ensemble* ridge estimator is defined by letting $M \to \infty$.
- The goal is to quantify and estimate the conditional prediction risk:

$$R_{k,M}^{\lambda} := \mathbb{E}[(y - \boldsymbol{x}^\top \widetilde{\beta}_{k,M}^{\lambda})^2 \mid \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M]$$

  under proportional asymptotics where $n, p, k \to \infty$, $p/n \to \phi$ and $p/k \to \phi_s$. Here, ϕ and ϕ_s are the *data* and *subsample* aspect ratios, respectively.

Ridge ensembles

- Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) \in \mathbb{R}^p \times \mathbb{R} : j \in [n]\}$ denote a dataset. The ridge estimator fitted on subsampled dataset $\mathcal{D}_I$ with $I \subseteq [n], |I| = k$ is:

$$\widehat{\beta}_k^{\lambda}(\mathcal{D}_I) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{k} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

- For $\lambda \geq 0$ fixed, ensemble ridge estimator is:

$$\widetilde{\beta}_{k,M}^{\lambda}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^M) := \frac{1}{M} \sum_{\ell \in [M]} \widehat{\beta}_k^{\lambda}(\mathcal{D}_{I_\ell}),$$

with $I_1, \ldots, I_M \sim \mathcal{I}_k := \{\{i_1, \ldots, i_k\} : 1 \leq i_1 < \ldots < i_k \leq n\}$. The *full-ensemble* ridge estimator is defined by letting $M \to \infty$.

- The goal is to quantify and estimate the conditional prediction risk:

$$R_{k,M}^{\lambda} := \mathbb{E}[(y - \boldsymbol{x}^\top \widetilde{\beta}_{k,M}^{\lambda})^2 \mid \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M]$$

under proportional asymptotics where $n, p, k \to \infty$, $p/n \to \phi$ and $p/k \to \phi_s$. Here, ϕ and ϕ_s are the *data* and *subsample* aspect ratios, respectively.

Ridge ensembles

- Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) \in \mathbb{R}^p \times \mathbb{R} : j \in [n]\}$ denote a dataset. The ridge estimator fitted on subsampled dataset $\mathcal{D}_I$ with $I \subseteq [n], |I| = k$ is:

$$\widehat{\beta}_k^{\lambda}(\mathcal{D}_I) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{k} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

- For $\lambda \geq 0$ fixed, ensemble ridge estimator is:

$$\widetilde{\beta}_{k,M}^{\lambda}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^M) := \frac{1}{M} \sum_{\ell \in [M]} \widehat{\beta}_k^{\lambda}(\mathcal{D}_{I_\ell}),$$

  with $I_1, \ldots, I_M \sim \mathcal{I}_k := \{\{i_1, \ldots, i_k\} : 1 \leq i_1 < \ldots < i_k \leq n\}$. The *full-ensemble* ridge estimator is defined by letting $M \to \infty$.

- The goal is to quantify and estimate the conditional prediction risk:

$$R_{k,M}^{\lambda} := \mathbb{E}[(y - \boldsymbol{x}^\top \widetilde{\beta}_{k,M}^{\lambda})^2 \mid \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M]$$

  under proportional asymptotics where $n, p, k \to \infty$, $p/n \to \phi$ and $p/k \to \phi_s$. Here, ϕ and ϕ_s are the *data* and *subsample* aspect ratios, respectively.

Ridge ensembles

- Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) \in \mathbb{R}^p \times \mathbb{R} : j \in [n]\}$ denote a dataset. The ridge estimator fitted on subsampled dataset $\mathcal{D}_I$ with $I \subseteq [n], |I| = k$ is:

$$\widehat{\beta}_k^{\lambda}(\mathcal{D}_I) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{k} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2.$$

- For $\lambda \geq 0$ fixed, ensemble ridge estimator is:

$$\widetilde{\beta}_{k,M}^{\lambda}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^M) := \frac{1}{M} \sum_{\ell \in [M]} \widehat{\beta}_k^{\lambda}(\mathcal{D}_{I_\ell}),$$

  with $I_1, \ldots, I_M \sim \mathcal{I}_k := \{\{i_1, \ldots, i_k\} : 1 \leq i_1 < \ldots < i_k \leq n\}$. The *full-ensemble* ridge estimator is defined by letting $M \to \infty$.
- The goal is to quantify and estimate the conditional prediction risk:

$$R_{k,M}^{\lambda} := \mathbb{E}[(y - \boldsymbol{x}^\top \widetilde{\beta}_{k,M}^{\lambda})^2 \mid \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M]$$

  under proportional asymptotics where $n, p, k \to \infty$, $p/n \to \phi$ and $p/k \to \phi_s$. Here, ϕ and ϕ_s are the *data* and *subsample* aspect ratios, respectively.

Data assumptions

1. Feature model:
   - Feature structure: $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector containing i.i.d. entries with mean 0, variance 1, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Covariance norm: There exist $r_{\text{min}}, r_{\text{max}}$ independent of p with $0 < r_{\text{min}} \leq r_{\text{max}} < \infty$ such that $r_{\text{min}}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\text{max}}\boldsymbol{I}_p$.
2. Response model:
   - Response structure: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$.
   - Noise structure: ϵ_i is an unobserved error that is assumed to be independent of $\boldsymbol{x}_i$ with mean 0, variance σ^2, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Signal norm: $\|\boldsymbol{\beta}_0\|_2$ uniformly bounded in p and $\lim_p \|\boldsymbol{\beta}_0\|_2^2 = \rho^2$.
3. Convergence of covariance and signal-weighted spectrums:
   - Covariance spectrum: $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$ is the eigenvalue decomposition.
   - Empirical spectrums: Assume there exist fixed distributions H and G such that the empirical spectral distributions satisfy

$$H_p(r) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} H,$$

$$G_p(r) := \frac{1}{\|\boldsymbol{\beta}_0\|_2^2}\sum_{i=1}^{p} (\boldsymbol{\beta}_0^\top \boldsymbol{w}_i)^2 \, \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} G.$$

Data assumptions

1. Feature model:
   - Feature structure: $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector containing i.i.d. entries with mean 0, variance 1, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Covariance norm: There exist $r_{\text{min}}, r_{\text{max}}$ independent of p with $0 < r_{\text{min}} \leq r_{\text{max}} < \infty$ such that $r_{\text{min}}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\text{max}}\boldsymbol{I}_p$.
2. Response model:
   - Response structure: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$.
   - Noise structure: ϵ_i is an unobserved error that is assumed to be independent of $\boldsymbol{x}_i$ with mean 0, variance σ^2, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Signal norm: $\|\boldsymbol{\beta}_0\|_2$ uniformly bounded in p and $\lim_p \|\boldsymbol{\beta}_0\|_2^2 = \rho^2$.
3. Convergence of covariance and signal-weighted spectrums:
   - Covariance spectrum: $\Sigma = \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$ is the eigenvalue decomposition.
   - Empirical spectrums: Assume there exist fixed distributions H and G such that the empirical spectral distributions satisfy

$$H_p(r) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} H,$$

$$G_p(r) := \frac{1}{\|\boldsymbol{\beta}_0\|_2^2}\sum_{i=1}^{p} (\boldsymbol{\beta}_0^\top \boldsymbol{w}_i)^2 \, \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} G.$$

Data assumptions

1. Feature model:
   - Feature structure: $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector containing i.i.d. entries with mean 0, variance 1, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Covariance norm: There exist $r_{\text{min}}, r_{\text{max}}$ independent of p with $0 < r_{\text{min}} \leq r_{\text{max}} < \infty$ such that $r_{\text{min}}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\text{max}}\boldsymbol{I}_p$.
2. Response model:
   - Response structure: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$.
   - Noise structure: ϵ_i is an unobserved error that is assumed to be independent of $\boldsymbol{x}_i$ with mean 0, variance σ^2, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Signal norm: $\|\boldsymbol{\beta}_0\|_2$ uniformly bounded in p and $\lim_p \|\boldsymbol{\beta}_0\|_2^2 = \rho^2$.
3. Convergence of covariance and signal-weighted spectrums:
   - Covariance spectrum: $\Sigma = \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$ is the eigenvalue decomposition.
   - Empirical spectrums: Assume there exist fixed distributions H and G such that the empirical spectral distributions satisfy

$$H_p(r) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} H,$$

$$G_p(r) := \frac{1}{\|\boldsymbol{\beta}_0\|_2^2}\sum_{i=1}^{p} (\boldsymbol{\beta}_0^\top \boldsymbol{w}_i)^2 \, \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} G.$$

Data assumptions

1. Feature model:
   - Feature structure: $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector containing i.i.d. entries with mean 0, variance 1, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Covariance norm: There exist $r_{\text{min}}, r_{\text{max}}$ independent of p with $0 < r_{\text{min}} \leq r_{\text{max}} < \infty$ such that $r_{\text{min}}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\text{max}}\boldsymbol{I}_p$.
2. Response model:
   - Response structure: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$.
   - Noise structure: ϵ_i is an unobserved error that is assumed to be independent of $\boldsymbol{x}_i$ with mean 0, variance σ^2, and bounded moment of order $4+\delta$ for some $\delta > 0$.
   - Signal norm: $\|\boldsymbol{\beta}_0\|_2$ uniformly bounded in p and $\lim_p \|\boldsymbol{\beta}_0\|_2^2 = \rho^2$.
3. Convergence of covariance and signal-weighted spectrums:
   - Covariance spectrum: $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$ is the eigenvalue decomposition.
   - Empirical spectrums: Assume there exist fixed distributions H and G such that the empirical spectral distributions satisfy

$$H_p(r) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} H,$$

$$G_p(r) := \frac{1}{\|\boldsymbol{\beta}_0\|_2^2}\sum_{i=1}^{p} (\boldsymbol{\beta}_0^\top \boldsymbol{w}_i)^2 \, \mathbb{1}_{\{r_i \leq r\}} \xrightarrow{\text{d}} G.$$

Risk characterization of bagged ridge predictors

Theorem. Under aforementioned assumptions, as $k, n, p \to \infty$ such that $\underbrace{p/n \to \phi \in (0, \infty)}_{\text{data aspect ratio}}$ and $\underbrace{p/k \to \phi_s \in [\phi, \infty]}_{\text{subsample aspect ratio}}$, the asymptotic risk $\mathscr{R}^{\text{sub}}_{\lambda,M}(\phi, \phi_s)$ is:

$$\mathscr{R}^{\text{sub}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \mathscr{B}^{\text{sub}}_{\lambda,M}(\phi, \phi_s) + \mathscr{V}^{\text{sub}}_{\lambda,M}(\phi, \phi_s),$$

where the bias and variance terms are given by

$$\mathscr{B}^{\text{sub}}_{\lambda,M}(\phi, \phi_s) = M^{-1} B_\lambda(\phi_s, \phi_s) + (1 - M^{-1}) B_\lambda(\phi, \phi_s),$$
$$\mathscr{V}^{\text{sub}}_{\lambda,M}(\phi, \phi_s) = M^{-1} V_\lambda(\phi_s, \phi_s) + (1 - M^{-1}) V_\lambda(\phi, \phi_s),$$

and the functions $B_\lambda(\cdot, \cdot)$ and $V_\lambda(\cdot, \cdot)$ are defined as

$$B_\lambda(\vartheta, \theta) = \rho^2 (1 + \widetilde{v}(-\lambda; \vartheta, \theta)) \widetilde{c}(-\lambda; \theta), \qquad V_\lambda(\vartheta, \theta) = \sigma^2 \widetilde{v}(-\lambda; \vartheta, \theta).$$

Here the non-negative constants $\widetilde{v}(-\lambda; \vartheta, \theta)$ and $\widetilde{c}(-\lambda; \theta)$ are defined as:

$$\widetilde{v}(-\lambda; \vartheta, \theta) = \frac{\vartheta \int r^2 (1 + v(-\lambda; \theta) r)^{-2} \, \mathrm{d}H(r)}{v(-\lambda; \theta)^{-2} - \vartheta \int r^2 (1 + v(-\lambda; \theta) r)^{-2} \, \mathrm{d}H(r)},$$
$$\widetilde{c}(-\lambda; \theta) = \int \frac{r}{(1 + v(-\lambda; \theta) r)^2} \, \mathrm{d}G(r).$$

Finally, $v(-\lambda; \theta)$ is the unique nonnegative solution to the fixed-point equation:

$$\frac{1}{v(-\lambda; \theta)} = \lambda + \theta \int \frac{r}{1 + v(-\lambda; \theta) r} \, \mathrm{d}H(r).$$

Bagged ridge risk characterization (illustration)

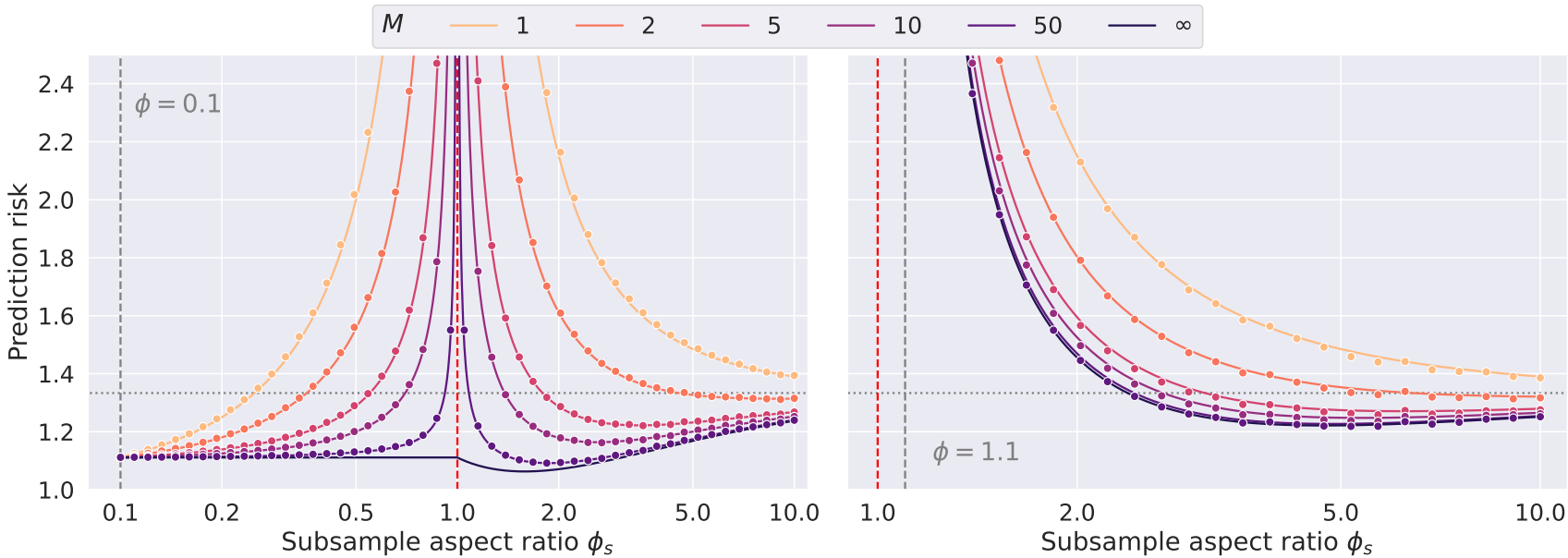


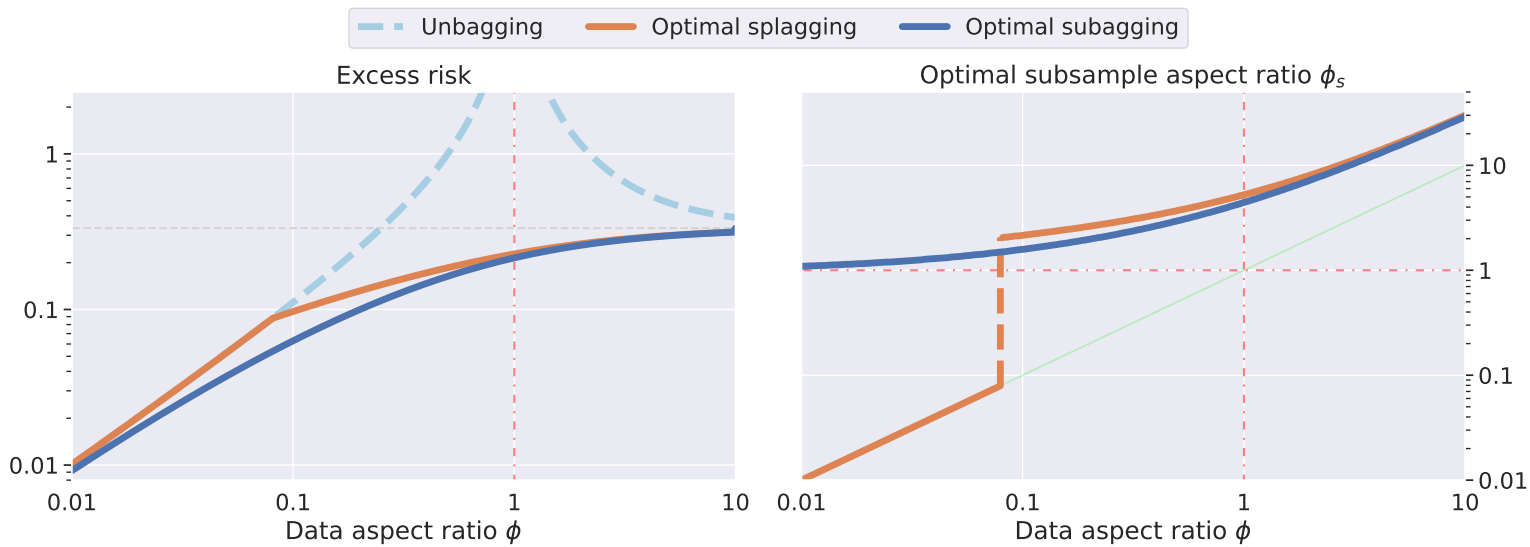
Figure: Asymptotic prediction risk curves for bagged ridgeless predictors ($\lambda = 0$), under AR1 model when $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = \lfloor p\phi \rfloor$ and $p = 500$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

Optimal bagged ridgeless predictor

Theorem. For any $\phi \geq 0$, the global minimum of $\phi_s \mapsto \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)$ is obtained in $\phi_s^* \in (1, \infty)$. That is

$$\sup_{M \in \mathbb{N}, \phi_s \in [\phi, \infty]} \mathscr{R}^{\text{sub}}_{0,M}(\phi, \phi_s) = \underbrace{\mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s^*)}_{\text{optimal bagged risk}}$$

$$< \min\{\underbrace{\mathscr{R}^{\text{sub}}_{0,1}(\phi, \phi)}_{\text{unbagged risk}}, \underbrace{\mathscr{R}^{\text{sub}}_{0,1}(\phi, \infty)}_{\text{null risk}}\}$$



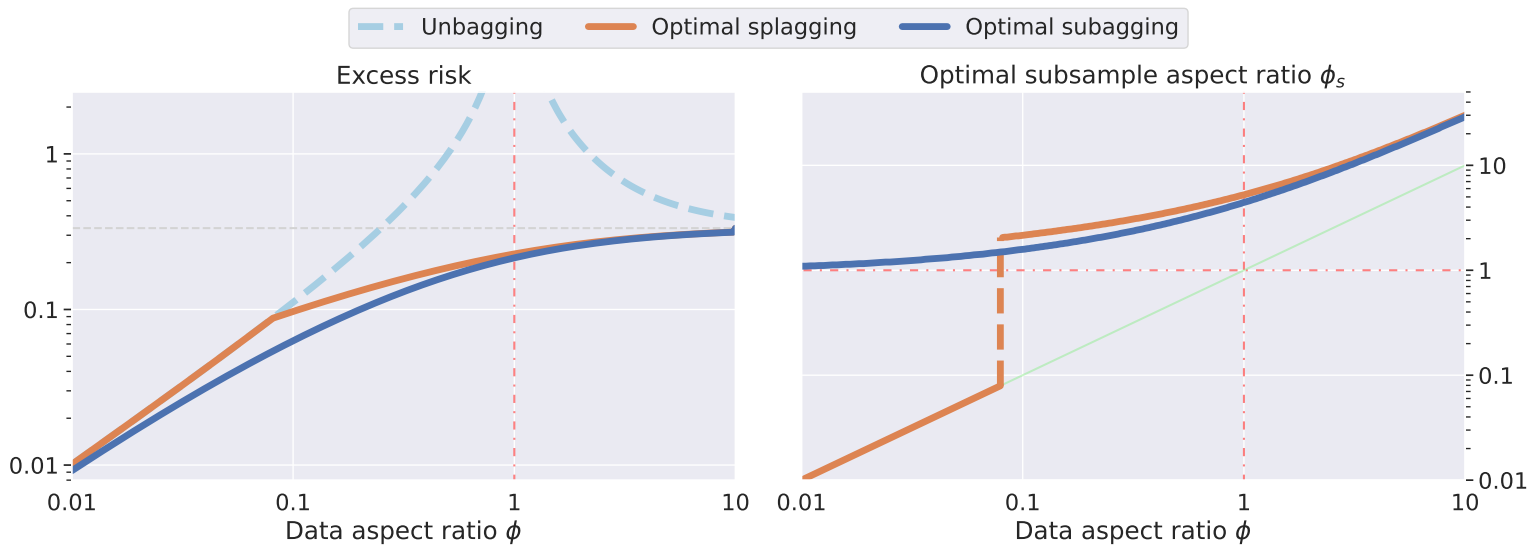
Subagged ridgeless *interpolators* always outperform subagged least squares, even when the full data has more observations than the number of features.

Optimal bagged ridgeless predictor

Theorem. For any $\phi \geq 0$, the global minimum of $\phi_s \mapsto \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)$ is obtained in $\phi_s^* \in (1, \infty)$. That is

$$\sup_{M \in \mathbb{N}, \phi_s \in [\phi, \infty]} \mathscr{R}^{\text{sub}}_{0,M}(\phi, \phi_s) = \underbrace{\mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s^*)}_{\text{optimal bagged risk}}$$

$$< \min\{\underbrace{\mathscr{R}^{\text{sub}}_{0,1}(\phi, \phi)}_{\text{unbagged risk}}, \underbrace{\mathscr{R}^{\text{sub}}_{0,1}(\phi, \infty)}_{\text{null risk}}\}$$



Subagged ridgeless *interpolators* always outperform subagged least squares, even when the full data has more observations than the number of features.

Back to risk monotonization

- Risk characterization $\rightarrow$ risk monotonization.
- Data splitting and cross-validation over subsample size.

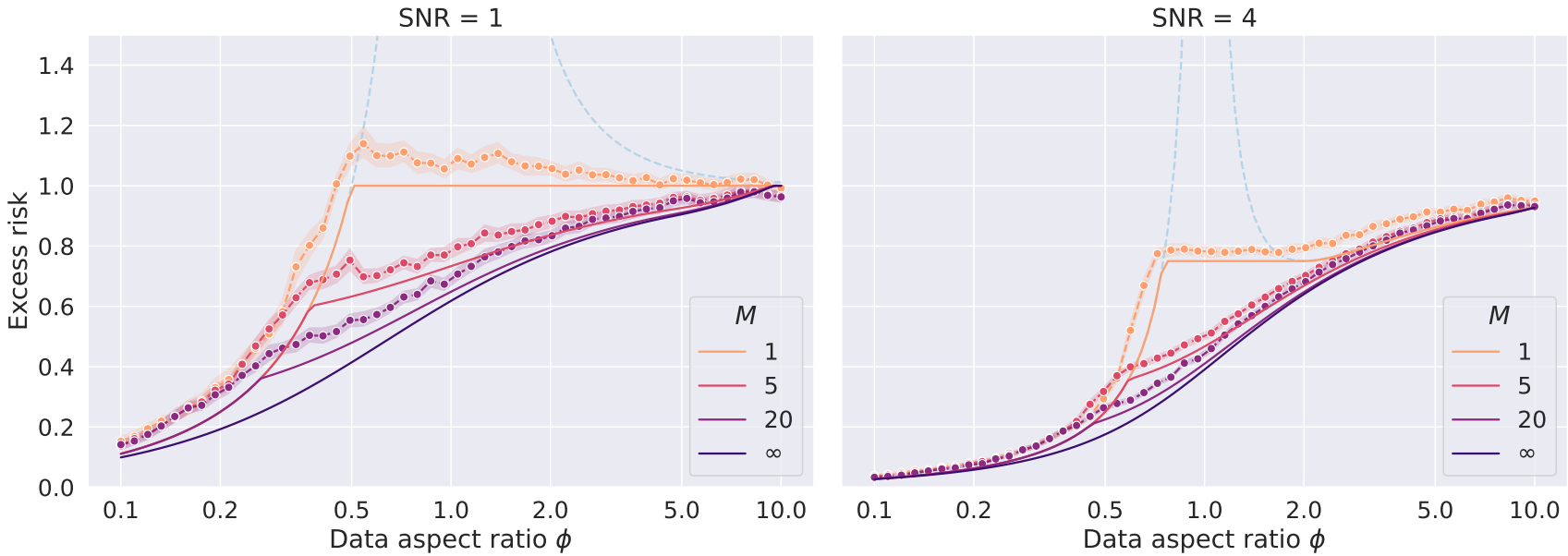
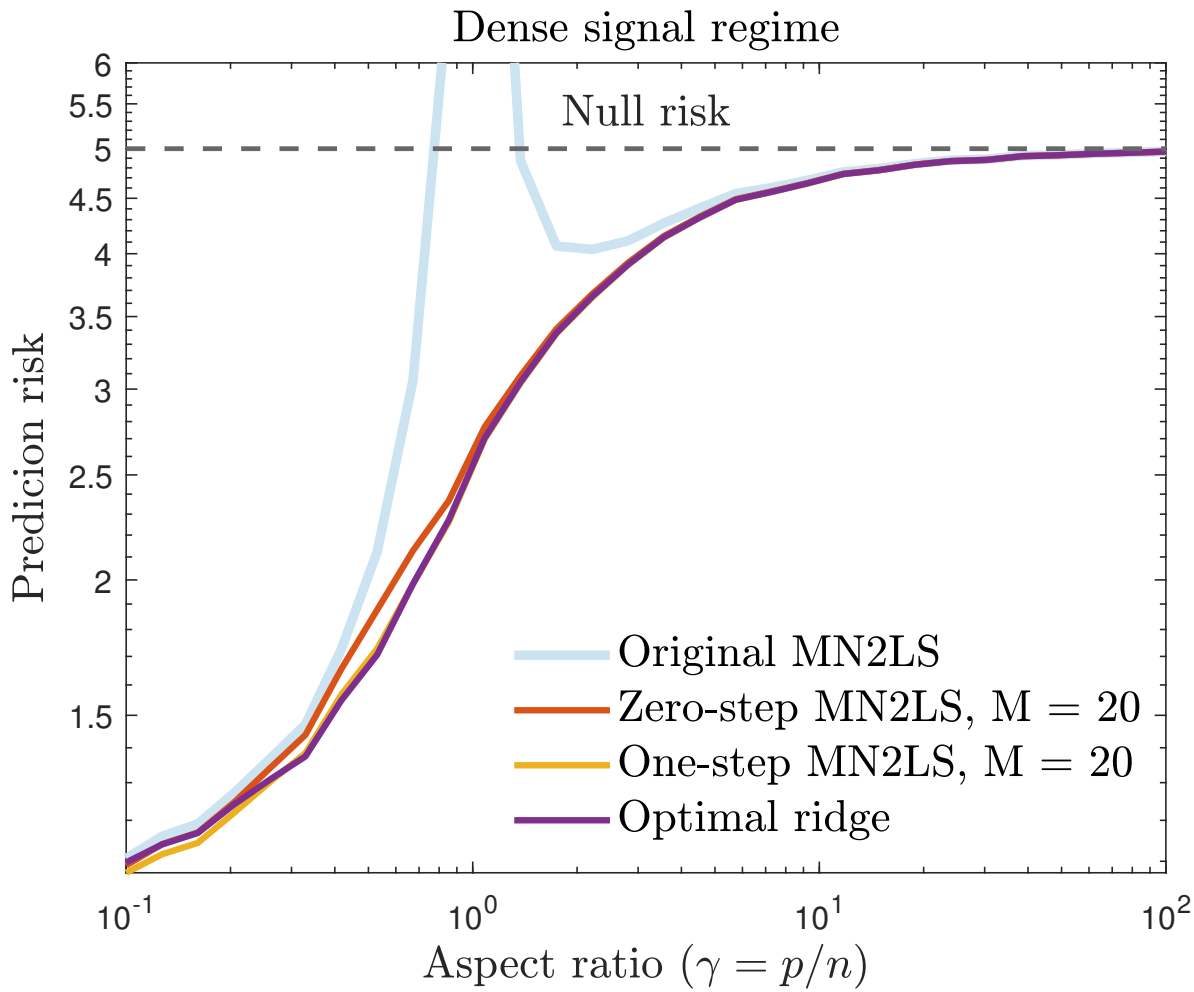


Figure: Asymptotic excess risk curves for cross-validated bagged ridgeless predictors ($\lambda = 0$), under the isotopic model when $\rho^2 = 1$ for varying $\mathtt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M with replacement. For each value of M, the points denote finite-sample risks and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\mathrm{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

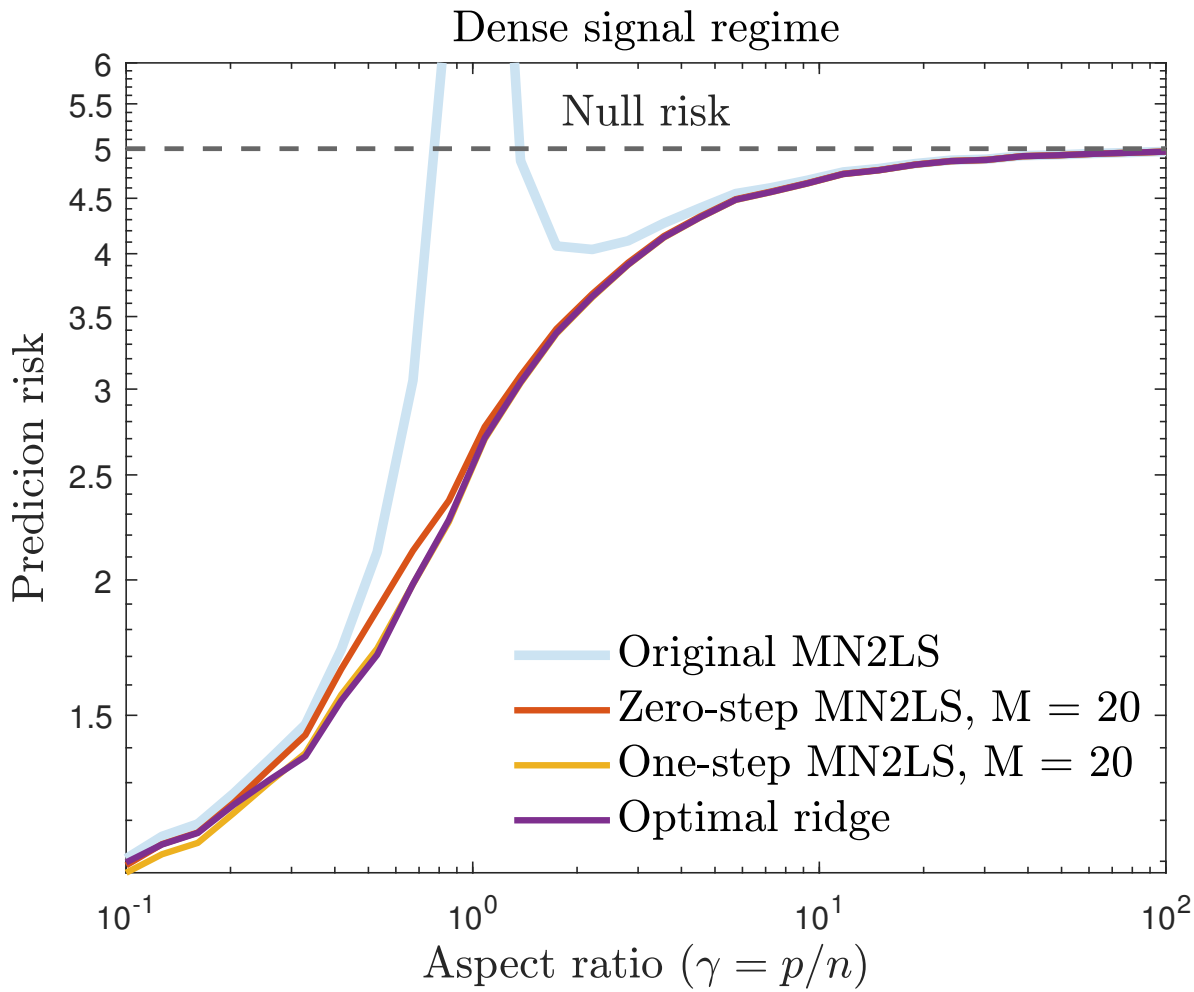
Comparison with optimal ridge regularization



Recall here $\gamma = p/n$ is the aspect ratio. The base predictor is ridgeless.

Key question: Is the connection to ridge regularization just coincedental?

Comparison with optimal ridge regularization



Recall here $\gamma = p/n$ is the aspect ratio. The base predictor is ridgeless.

Key question: Is the connection to ridge regularization just coincedental?

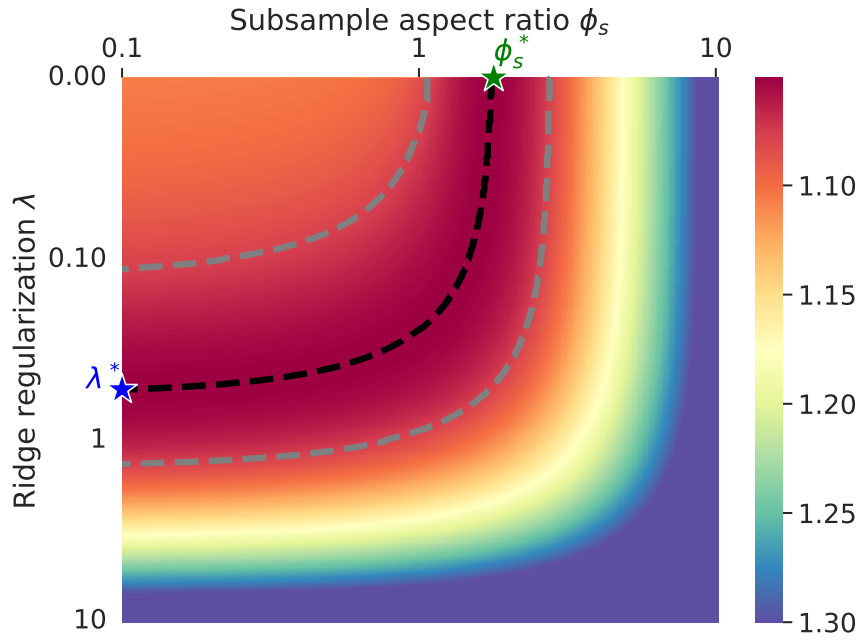
Outline

Prediction risk equivalence

- As $p/n \to \phi$ and $p/k \to \phi_s$, the prediction risk in the full ensemble ($M = \infty$) converges:

$$R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s).$$

- For $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure in the log-log scale.



- Risk equivalence (Theorem 2.3):

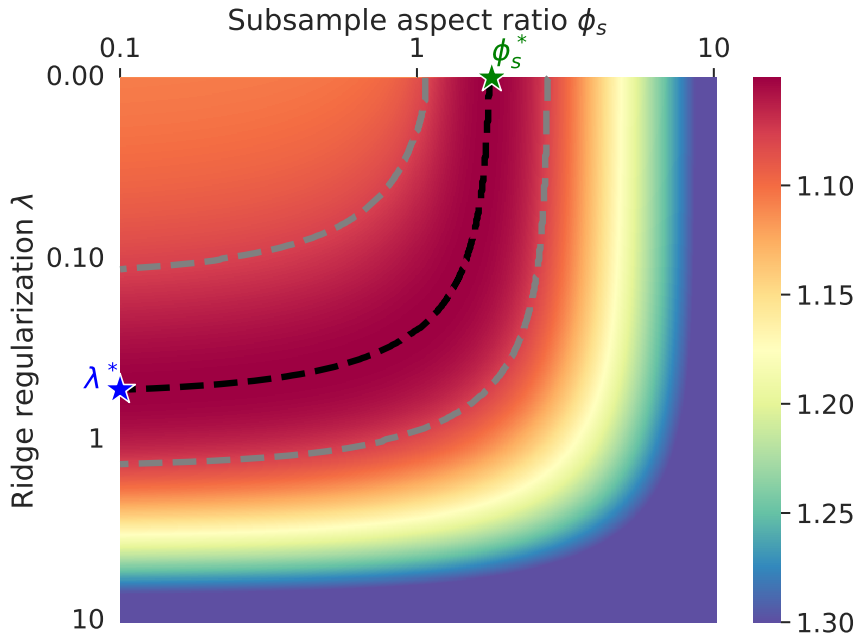
$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda \geq 0}} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s)}_{\text{opt. ridge ensemble}}.$$

Prediction risk equivalence

- As $p/n \to \phi$ and $p/k \to \phi_s$, the prediction risk in the full ensemble ($M = \infty$) converges:

$$R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s).$$

- For $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure in the log-log scale.



- Risk equivalence (Theorem 2.3):

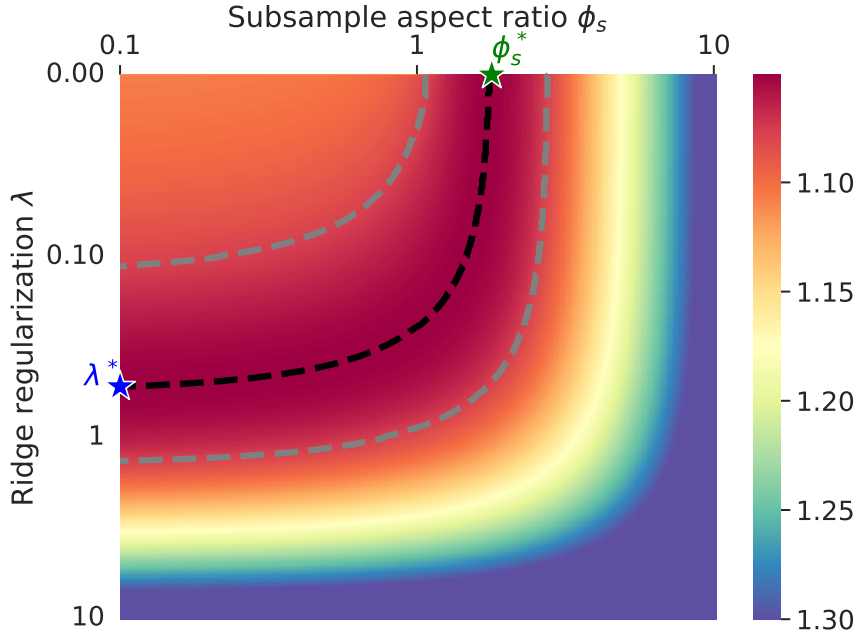
$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda \geq 0}} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s)}_{\text{opt. ridge ensemble}}.$$

Prediction risk equivalence

- As $p/n \to \phi$ and $p/k \to \phi_s$, the prediction risk in the full ensemble ($M = \infty$) converges:

$$R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s).$$

- For $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure in the log-log scale.



- Risk equivalence (Theorem 2.3):

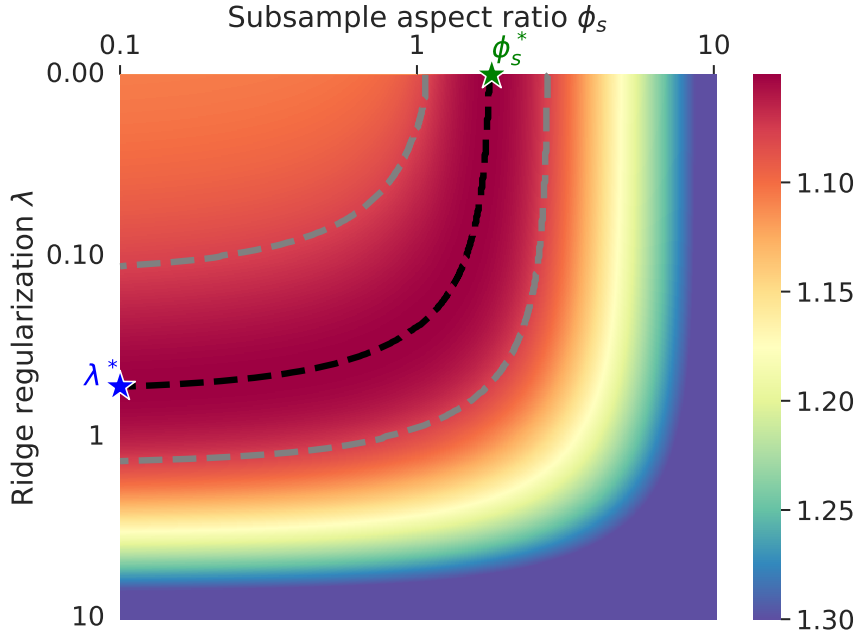
$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda \geq 0}} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s)}_{\text{opt. ridge ensemble}}.$$

Prediction risk equivalence

- As $p/n \to \phi$ and $p/k \to \phi_s$, the prediction risk in the full ensemble ($M = \infty$) converges:

$$R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s).$$

- For $\phi = 0.1$, the risk profile as a function of (λ, ϕ_s) is shown in the figure in the log-log scale.



- Risk equivalence (Theorem 2.3):

$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda \geq 0}} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s)}_{\text{opt. ridge ensemble}}.$$

Generalized risk

- Let $\boldsymbol{\beta}_0 = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\mathbb{E}[\boldsymbol{x}y]$ be the best linear projection of y onto $\boldsymbol{x}$
- For a linear functional $L(\boldsymbol{\beta}) = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{b}$, we study generalized risks:

$$R(\widehat{\beta}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0) = \frac{1}{\mathrm{nrow}(\boldsymbol{A})} \|L(\widehat{\beta} - \boldsymbol{\beta}_0)\|_2^2, \tag{1}$$

under proportional asymptotics where $n, p, k \to \infty$, $p/n \to \phi$ and $p/k \to \psi$. Here, ϕ and ψ are the *data* and *subsample aspect ratios*, respectively.

Statistical learning problem	$L(\widehat{\beta} - \beta_0)$	$\boldsymbol{A}$	$\boldsymbol{b}$	$\mathrm{nrow}(\boldsymbol{A})$
vector coefficient estimation	$\widehat{\beta} - \beta_0$	$\boldsymbol{I}_p$	0	p
projected coefficient estimation	$\boldsymbol{a}^\top(\widehat{\beta} - \beta_0)$	$\boldsymbol{a}^\top$	0	1
training error estimation	$\boldsymbol{X}\widehat{\beta} - \boldsymbol{y}$	$\boldsymbol{X}$	$-\boldsymbol{f}_{\mathrm{NL}}$	n
in-sample prediction	$\boldsymbol{X}(\widehat{\beta} - \beta_0)$	$\boldsymbol{X}$	0	n
out-of-sample prediction	$\boldsymbol{x}_0^\top\widehat{\beta} - y_0$	$\boldsymbol{x}_0^\top$	$-\epsilon_0$	1

Generalized risk

- Let $\boldsymbol{\beta}_0 = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\mathbb{E}[\boldsymbol{x}y]$ be the best linear projection of y onto $\boldsymbol{x}$
- For a linear functional $L(\boldsymbol{\beta}) = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{b}$, we study generalized risks:

$$R(\widehat{\beta}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0) = \frac{1}{\mathrm{nrow}(\boldsymbol{A})} \|L(\widehat{\beta} - \boldsymbol{\beta}_0)\|_2^2, \tag{1}$$

under proportional asymptotics where $n, p, k \to \infty$, $p/n \to \phi$ and $p/k \to \psi$. Here, ϕ and ψ are the *data* and *subsample aspect ratios*, respectively.

Statistical learning problem	$L(\widehat{\beta} - \boldsymbol{\beta}_0)$	$\boldsymbol{A}$	$\boldsymbol{b}$	$\mathrm{nrow}(\boldsymbol{A})$
vector coefficient estimation	$\widehat{\beta} - \boldsymbol{\beta}_0$	$\boldsymbol{I}_p$	0	p
projected coefficient estimation	$\boldsymbol{a}^\top(\widehat{\beta} - \boldsymbol{\beta}_0)$	$\boldsymbol{a}^\top$	0	1
training error estimation	$\boldsymbol{X}\widehat{\beta} - \boldsymbol{y}$	$\boldsymbol{X}$	$-\boldsymbol{f}_{\mathrm{NL}}$	n
in-sample prediction	$\boldsymbol{X}(\widehat{\beta} - \boldsymbol{\beta}_0)$	$\boldsymbol{X}$	0	n
out-of-sample prediction	$\boldsymbol{x}_0^\top\widehat{\beta} - y_0$	$\boldsymbol{x}_0^\top$	$-\epsilon_0$	1

Asymptotic equivalence and relaxed assumptions

Asymptotic equivalence:

- Let $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ be sequences of (additively) conformable matrices of arbitrary dimensions (including vectors and scalars).
- We say that $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ are *asymptotically equivalent*, denoted as $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$, if $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ almost surely for any sequence of random matrices $\boldsymbol{C}_p$ with bounded trace norm that are (multiplicatively) conformable and independent of $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$.
- Note that for sequences of scalar random variables, the definition simply reduces to the typical almost sure convergence of sequences of random variables involved.

Data assumptions:

- Feature distribution: Each feature vector $\boldsymbol{x}_i$ for $i \in [n]$ can be decomposed as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ 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 distribution: Each response variable y_i for $i \in [n]$ has mean 0, and bounded $4 + \mu$ moments.

Asymptotic equivalence and relaxed assumptions

Asymptotic equivalence:

- Let $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ be sequences of (additively) conformable matrices of arbitrary dimensions (including vectors and scalars).
- We say that $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ are *asymptotically equivalent*, denoted as $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$, if $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ almost surely for any sequence of random matrices $\boldsymbol{C}_p$ with bounded trace norm that are (multiplicatively) conformable and independent of $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$.
- Note that for sequences of scalar random variables, the definition simply reduces to the typical almost sure convergence of sequences of random variables involved.

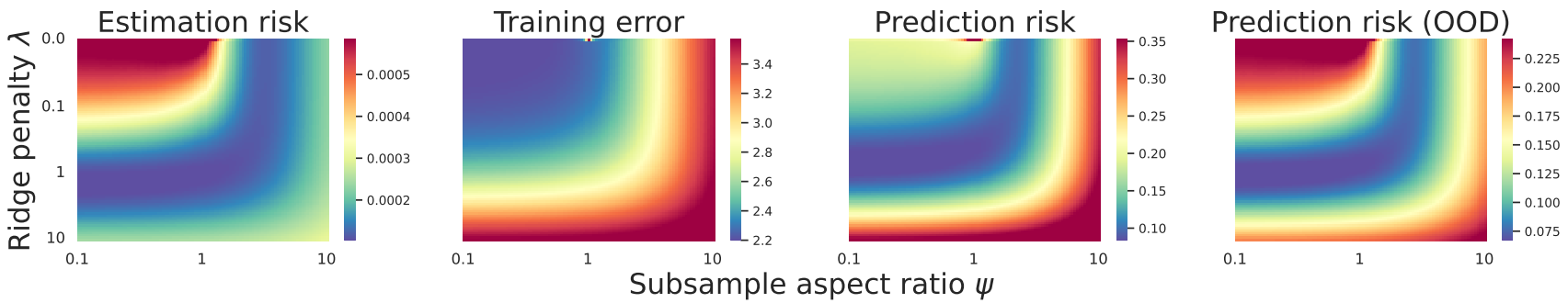
Data assumptions:

- Feature distribution: Each feature vector $\boldsymbol{x}_i$ for $i \in [n]$ can be decomposed as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ 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 distribution: Each response variable y_i for $i \in [n]$ has mean 0, and bounded $4 + \mu$ moments.

Generalized risk equivalences

Theorem. For any $\bar{\psi} \in [\phi, +\infty]$, let $\bar{\lambda}$ be as defined in (4). Then, for any pair of (λ_1, ψ_1) and (λ_2, ψ_2) on the path $\mathcal{P}(\bar{\lambda}; \phi, \bar{\psi})$ as defined in (5), the generalized risk functionals (1) of the full-ensemble estimator are asymptotically equivalent:

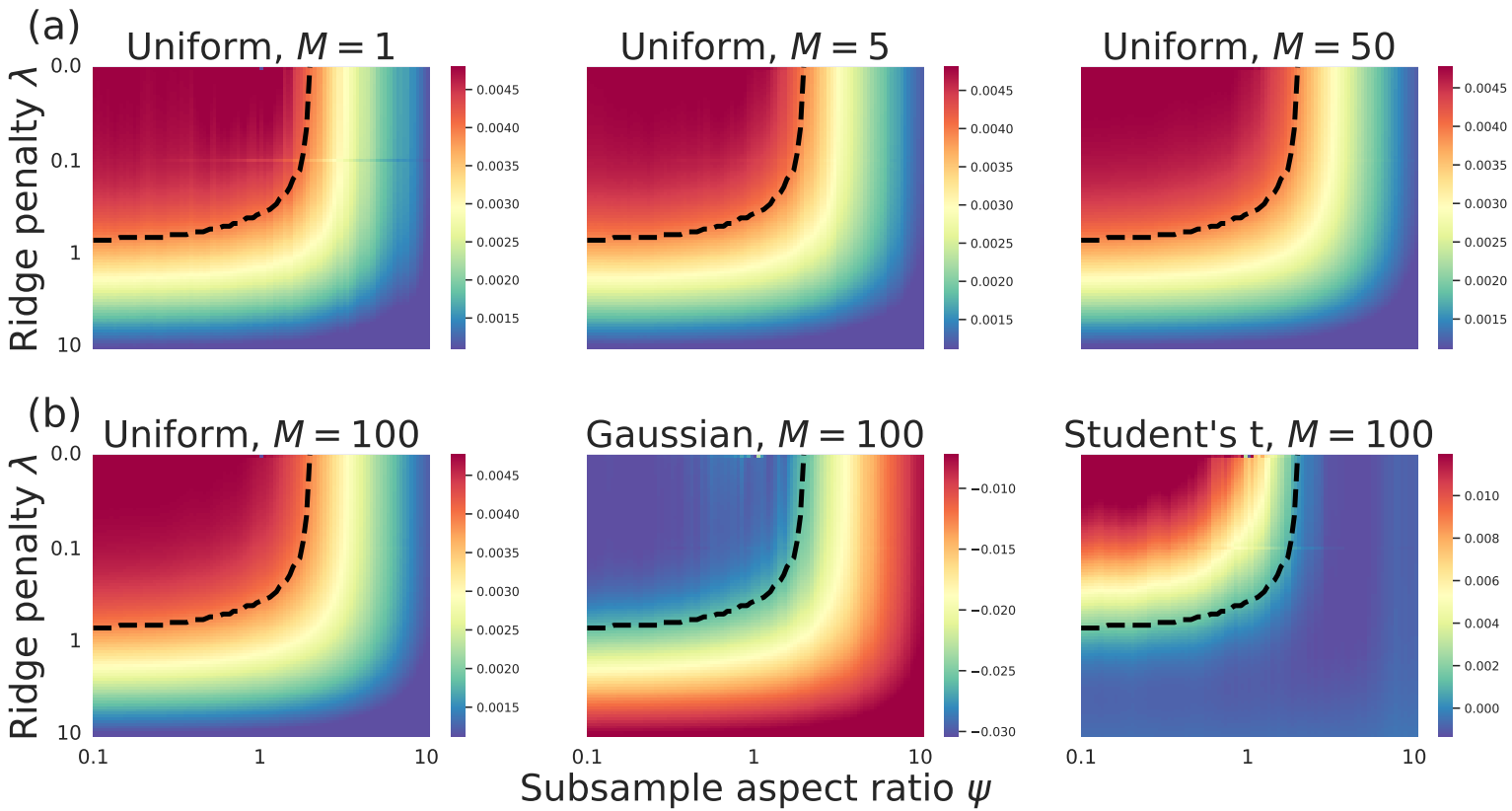
$$R\big(\widehat{\beta}^{\lambda_1}_{\lfloor p/\psi_1 \rfloor, \infty}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0\big) \;\simeq\; R\big(\widehat{\beta}^{\lambda_2}_{\lfloor p/\psi_2 \rfloor, \infty}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0\big). \tag{2}$$



Structural equivalences

Theorem. For any $\bar{\psi} \in [\phi, +\infty]$, let $\bar{\lambda}$ be as in (4). Then, for any $M \in \mathbb{N} \cup \{\infty\}$ and any pair of (λ_1, ψ_1) and (λ_2, ψ_2) on the path (5), the M-ensemble estimators are asymptotically equivalent:

$$\widehat{\beta}^{\lambda_1}_{\lfloor p/\psi_1 \rfloor, M} \simeq \widehat{\beta}^{\lambda_2}_{\lfloor p/\psi_2 \rfloor, M}, \qquad \forall (\lambda_1, \psi_1), (\lambda_2, \psi_2) \in \mathcal{P}(\bar{\lambda}; \phi, \bar{\psi}). \tag{3}$$



Equivalence paths

- Given $\phi \in (0, \infty)$ and $\bar{\psi} \in [\phi, \infty]$, our statement of equivalences between different ensemble estimators is defined through certain paths characterized by two endpoints $(0, \bar{\psi})$ and $(\bar{\lambda}, \phi)$.
- Let H_p be the empirical spectral distribution of $\boldsymbol{\Sigma}$: $H_p(r) = p^{-1} \sum_{i=1}^{p} \mathbb{1}_{\{r_i \le r\}}$, where r_i's are the eigenvalues of $\boldsymbol{\Sigma}$. Consider the following system of equations in $\bar{\lambda}$ and v:

$$\frac{1}{v} = \bar{\lambda} + \phi \int \frac{r}{1 + vr} \mathrm{d}H_p(r), \quad \text{and} \quad \frac{1}{v} = \bar{\psi} \int \frac{r}{1 + vr} \mathrm{d}H_p(r). \tag{4}$$

- Now, define a path $\mathcal{P}(\bar{\lambda}; \phi, \bar{\psi})$ that passes through the endpoints $(0, \bar{\psi})$ and $(\bar{\lambda}, \phi)$:

$$\mathcal{P}(\bar{\lambda}; \phi, \bar{\psi}) = \left\{ (1 - \theta) \cdot (\bar{\lambda}, \phi) + \theta \cdot (0, \bar{\psi}) \mid \theta \in [0, 1] \right\}. \tag{5}$$

- For any $M \in \mathbb{N} \cup \{\infty\}$, let $\bar{\lambda}_n$ be the value that satisfies the following equation in ensemble ridgeless and ridge gram matrices:

$$\frac{1}{M} \sum_{\ell=1}^{M} \frac{1}{k} \operatorname{tr} \left[\left(\frac{1}{k} \boldsymbol{L}_{I_\ell} \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{L}_{I_\ell} \right)^+ \right] = \frac{1}{n} \operatorname{tr} \left[\left(\frac{1}{n} \boldsymbol{X} \boldsymbol{X}^\top + \bar{\lambda}_n \boldsymbol{I}_n \right)^{-1} \right]. \tag{6}$$

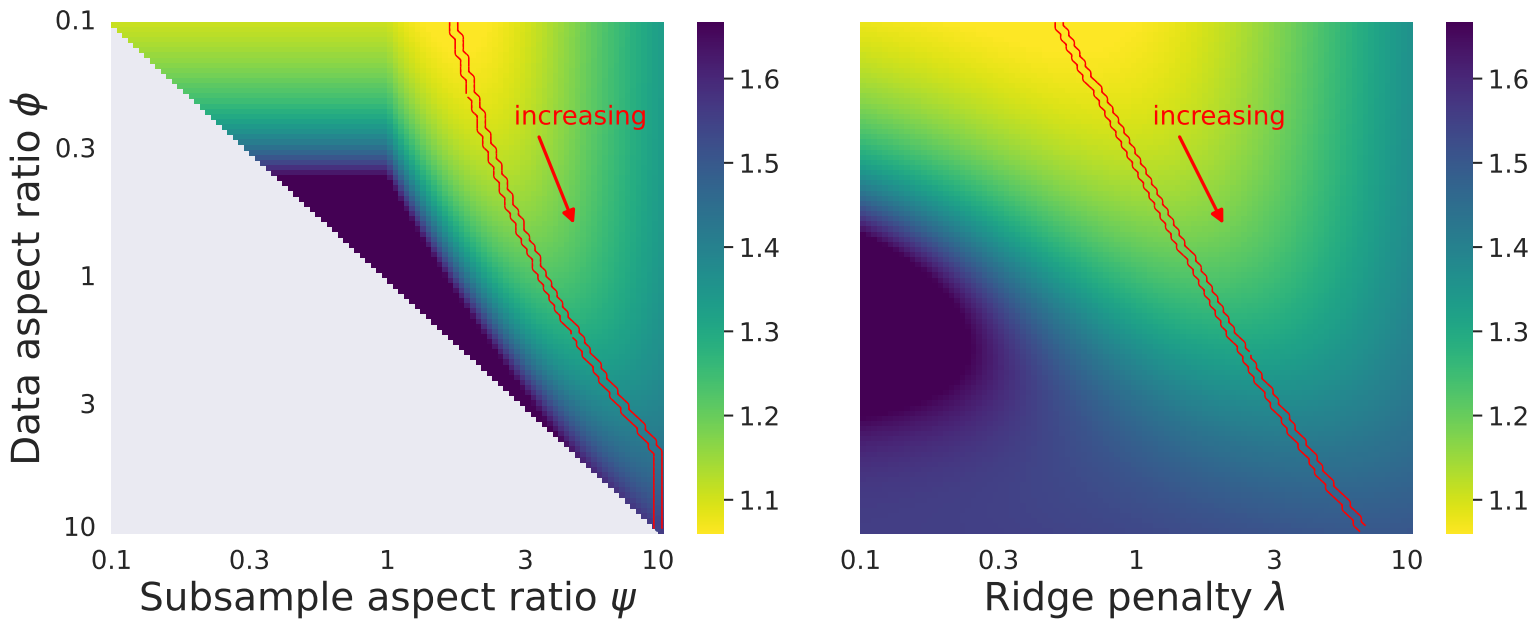
  Define the data-dependent path $\mathcal{P}_n = \mathcal{P}(\bar{\lambda}_n; \phi_n, \bar{\psi}_n)$.

Implications: Monotonicity of optimal ridge

- An open problem raised by Nakkiran et al. (2021) asks whether the prediction risk of ridge regression with optimal ridge penalty λ^* is monotonically increasing in the data aspect ratio $\phi = p/n$.
- Our equivalences imply that the prediction risk of an optimally-tuned ridge estimator is monotonically increasing in the data aspect ratio under mild regularity conditions.
- Under proportional asymptotics, our result settles a recent open question raised by Conjecture 1 of Nakkiran et al. (2021) concerning the monotonicity of optimal ridge regression under anisotropic features and general data models while maintaining a regularity condition that preserves the linearized signal-to-noise ratios across regression problems.

Implications of equivalences: illustration

Theorem. Let $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \psi \in [\phi, \infty]$. Then, for $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{b} = \boldsymbol{0}$, the optimal risk of the ridgeless ensemble, $\min_{\psi \geq \phi} \mathscr{R}(0; \phi, \psi)$, is monotonically increasing in ϕ. Consequently, the optimal risk of the ridge predictor, $\min_{\geq 0} \mathscr{R}(;\phi, \phi)$, is also monotonically increasing in ϕ.

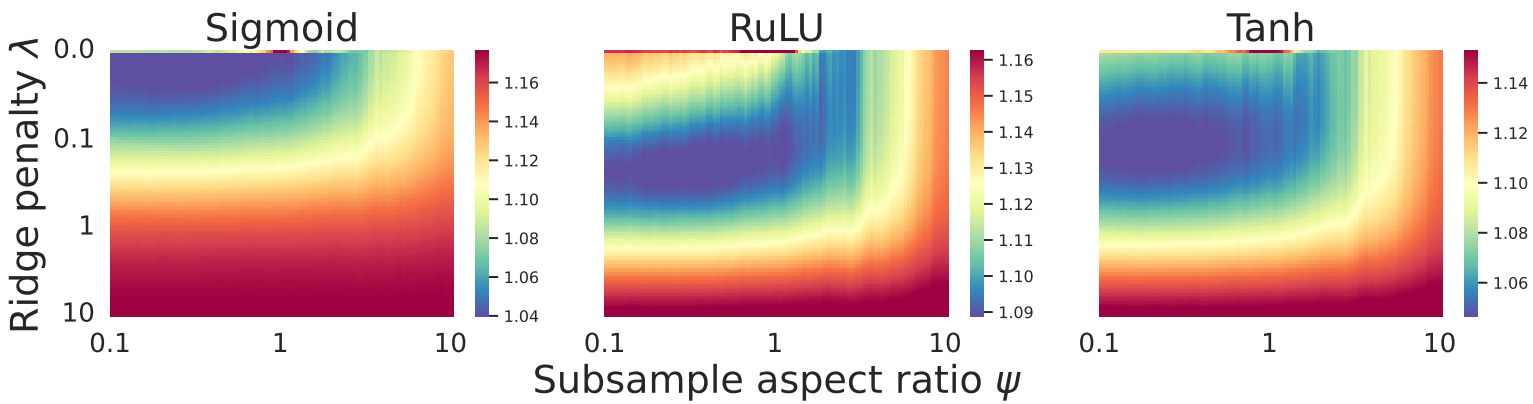


Extension 1: Equivalences for random features

Conjecture. Define $\phi_n = p/n$. Let $k \le n$ be the subsample size and denote by $\bar{\psi}_n = p/k$. Suppose φ satisfies certain regularity conditions. For any $M \in \mathbb{N} \cup \{\infty\}$, let $\bar{\lambda}_n$ be the value that satisfies

$$\frac{1}{M}\sum_{\ell=1}^{M}\frac{1}{k}\operatorname{tr}\left[\left(\frac{1}{k}\varphi(\boldsymbol{L}_{I_\ell}\boldsymbol{X}\boldsymbol{F}^\top)\varphi(\boldsymbol{L}_{I_\ell}\boldsymbol{X}\boldsymbol{F}^\top)^\top\right)^+\right] = \frac{1}{n}\operatorname{tr}\left[\left(\frac{1}{n}\varphi(\boldsymbol{X}\boldsymbol{F}^\top)\varphi(\boldsymbol{X}\boldsymbol{F}^\top)^\top + \bar{\lambda}_n \boldsymbol{I}_n\right)^{-1}\right].$$

Define the data-dependent path $\mathcal{P}_n = \mathcal{P}(\bar{\lambda}_n; \phi_n, \bar{\psi}_n)$. Then similar equivalences continue to hold along $\mathcal{P}_n$.

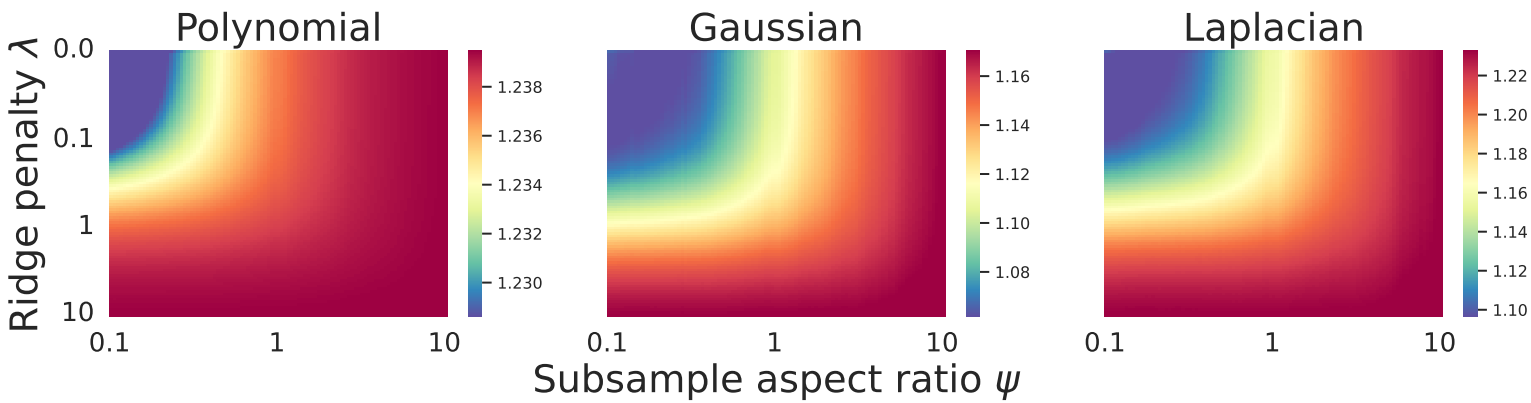


Extension 2: Equivalences for kernel features

Conjecture. Define $\phi_n = p/n$. Suppose the kernel K satisfies certain regularity conditions. Let $k \leq n$ be the subsample size and denote by $\bar{\psi}_n = p/k$. For any $M \in \mathbb{N} \cup \{\infty\}$, let $\bar{\lambda}_n$ be a solution to

$$\frac{1}{M}\sum_{\ell=1}^{M} \mathrm{tr}\left[\boldsymbol{K}_{I_\ell}^{+}\right] = \mathrm{tr}\left[\left(\boldsymbol{K}_{[n]} + \frac{n}{p}\bar{\lambda}_n \boldsymbol{I}_n\right)^{-1}\right].$$

Define the data-dependent path $\mathcal{P}_n = \mathcal{P}(\bar{\lambda}_n; \phi_n, \bar{\psi}_n)$. Then similar equivalences continue to hold along $\mathcal{P}_n$.



Outline

Main takeaways

1. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

2. Ensembling helps significantly near the interpolator threshold. Subagged ridgeless interpolators always outperform subagged least squares, even when the full data has more observations than the number of features.

3. There are connections between the implicit regularization induced by subsampling and explicit ridge regularization for subsampled ridge ensembles.

Main takeaways

1. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

2. Ensembling helps significantly near the interpolator threshold. Subagged ridgeless interpolators always outperform subagged least squares, even when the full data has more observations than the number of features.

3. There are connections between the implicit regularization induced by subsampling and explicit ridge regularization for subsampled ridge ensembles.

Main takeaways

1. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

2. Ensembling helps significantly near the interpolator threshold. Subagged ridgeless interpolators always outperform subagged least squares, even when the full data has more observations than the number of features.

3. There are connections between the implicit regularization induced by subsampling and explicit ridge regularization for subsampled ridge ensembles.

Main takeaways

1. It is possible to modify any given prediction procedure to mitigate double descent behavior and achieve a monotonic risk behavior through subsampling and cross-validation.

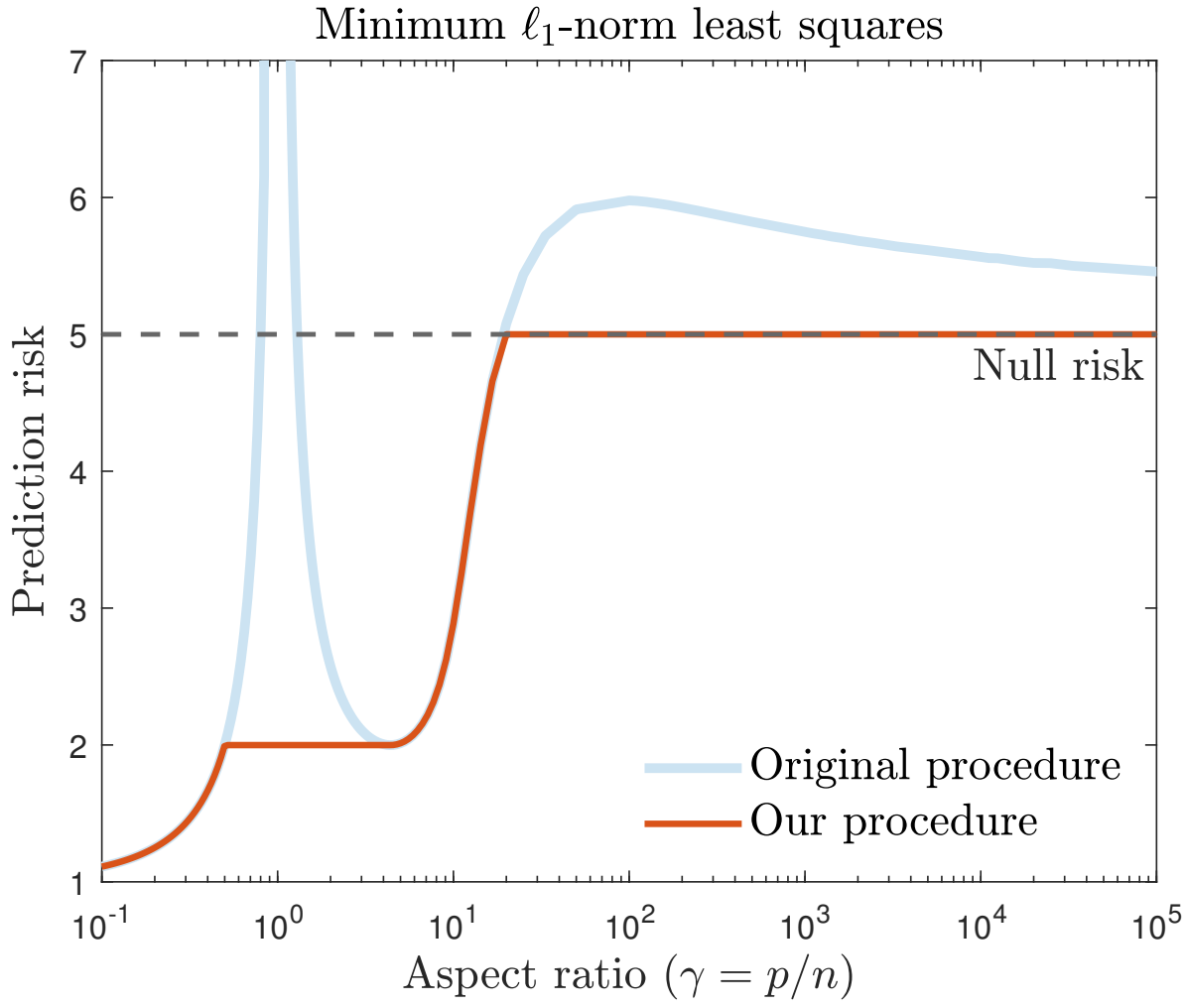
2. Ensembling helps significantly near the interpolator threshold. Subagged ridgeless interpolators always outperform subagged least squares, even when the full data has more observations than the number of features.

3. There are connections between the implicit regularization induced by subsampling and explicit ridge regularization for subsampled ridge ensembles.

Thanks for listening!

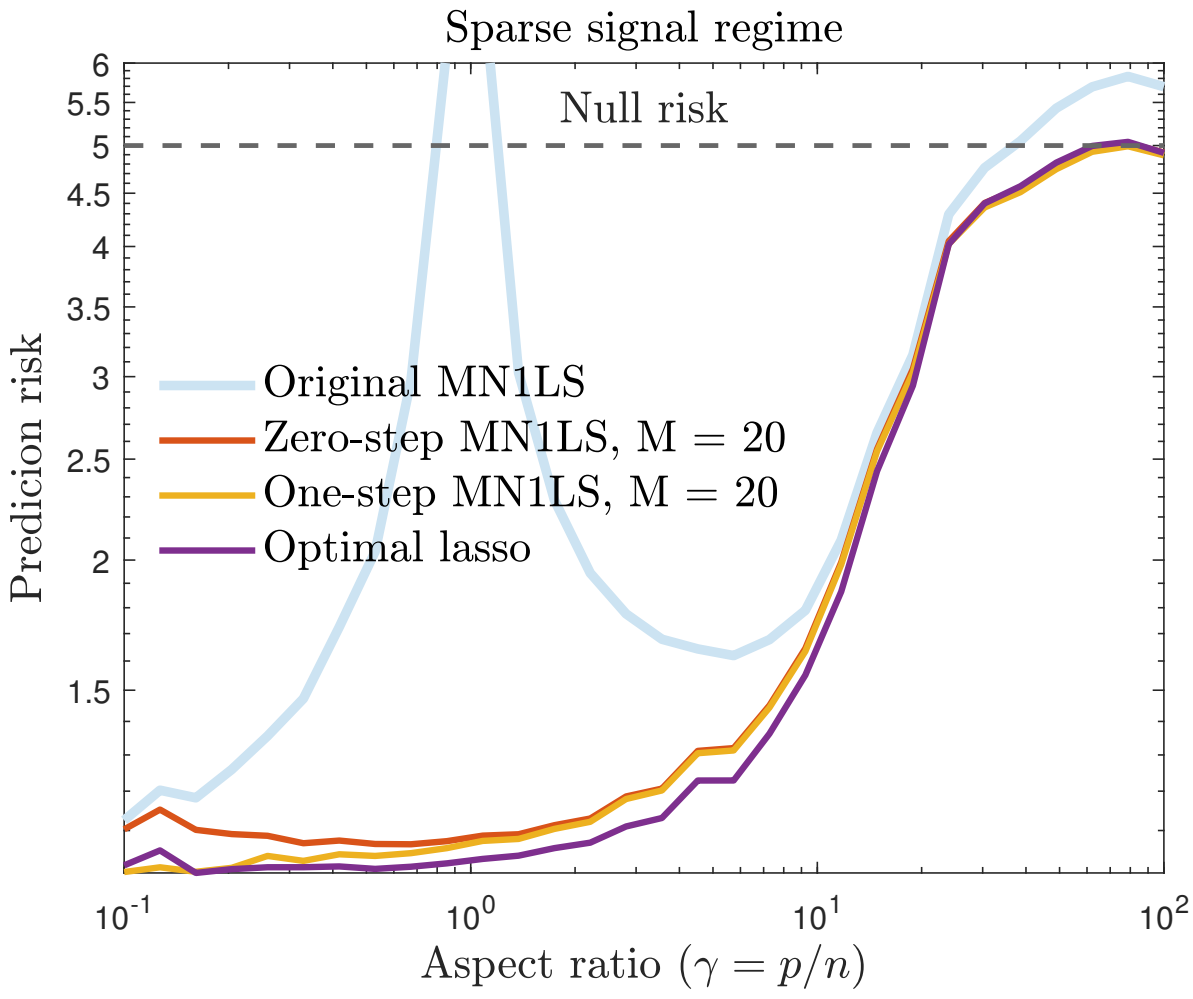
Questions/comments/thoughts?

What about lasso?



"Mitigating multiple descents: A model-agnostic framework for risk monotonization"
P., Kuchibhotla, Wei, Rinaldo, 2021

What about lasso?



More empirical evidence for lasso