Mitigating multiple descents:

A model-agnostic framework for risk monotonization

Pratik Patil

University of California Berkeley

UC Berkeley Biostatistics Seminar
February 2024

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)

1. Thanks, Lexin! Hi everyone! It is good to be here. I have never been to this seminar series, so thanks for the invite. But I already know some people here, so it is a nice opportunity for me to say hello to them, and meet some new people.
2. Let me say a few words about myself. As Pierre said, I am Pratik. I am currently a postdoc at Berkeley. I was a PhD student at CMU.
3. The talk is broadly going to be about overparameterized learning. A part of it is based on work I did for my PhD. A part is some new extensions. It is broadly based on three papers.

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]

1. There are three parts to the talk, based on three papers. My goal is to try to convey three different points: one on subsampling, one on ensembling, and one on connections to ridge regression. Here are pointers to the papers:

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]

1. There are three parts to the talk, based on three papers. My goal is to try to convey three different points: one on subsampling, one on ensembling, and one on connections to ridge regression. Here are pointers to the papers:
2. The first one is about model-agnostic risk monotonization. This form the basis of the talk. The key takeaway here is the benefit of subsampling.

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]

1. There are three parts to the talk, based on three papers. My goal is to try to convey three different points: one on subsampling, one on ensembling, and one on connections to ridge regression. Here are pointers to the papers:
2. The first one is about model-agnostic risk monotonization. This form the basis of the talk. The key takeaway here is the benefit of subsampling.
3. The second one is about an explicit bagging analysis. The key takeaway here is the benefit of ensembling.

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]

1. There are three parts to the talk, based on three papers. My goal is to try to convey three different points: one on subsampling, one on ensembling, and one on connections to ridge regression. Here are pointers to the papers:
2. The first one is about model-agnostic risk monotonization. This form the basis of the talk. The key takeaway here is the benefit of subsampling.
3. The second one is about an explicit bagging analysis. The key takeaway here is the benefit of ensembling.
4. The third one is about some connections to ridge regression. The key takeaway here is certain equivalences to ridge regression.

Outline

Here's an outline of what I am going to be talking about. I will start by giving an overview of overparameterization and somee motivations behind risk monotonization. And then I will go in detail and tell you some of our results.

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.

1. To set the stage, we will mainly be focusing on overparameterized learning. As we all know, machine learning models these days fit a large number of parameters compared to the number of observations. By parameters here, I mean either raw features in the dataset, or learned features. Such overparameterization seems to be useful for a number of things.

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.

1. To set the stage, we will mainly be focusing on overparameterized learning. As we all know, machine learning models these days fit a large number of parameters compared to the number of observations. By parameters here, I mean either raw features in the dataset, or learned features. Such overparameterization seems to be useful for a number of things.
2. First, more parameters allow for rich classes of models that are capable of representing diverse set of real data.

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.

1. To set the stage, we will mainly be focusing on overparameterized learning. As we all know, machine learning models these days fit a large number of parameters compared to the number of observations. By parameters here, I mean either raw features in the dataset, or learned features. Such overparameterization seems to be useful for a number of things.
2. First, more parameters allow for rich classes of models that are capable of representing diverse set of real data.
3. Second, somewhat surprisingly, the optimization problem to fit these models simplifies dramatically in the overparametrized regime. So even simple, local optimization approaches often find near-optimal solutions.

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.

1. To set the stage, we will mainly be focusing on overparameterized learning. As we all know, machine learning models these days fit a large number of parameters compared to the number of observations. By parameters here, I mean either raw features in the dataset, or learned features. Such overparameterization seems to be useful for a number of things.
2. First, more parameters allow for rich classes of models that are capable of representing diverse set of real data.
3. Second, somewhat surprisingly, the optimization problem to fit these models simplifies dramatically in the overparametrized regime. So even simple, local optimization approaches often find near-optimal solutions.
4. Third, even more surprisingly, without any explicit regularization, the fitted models seem to perform well on unseen data in practice.

Overparametrization in machine learning

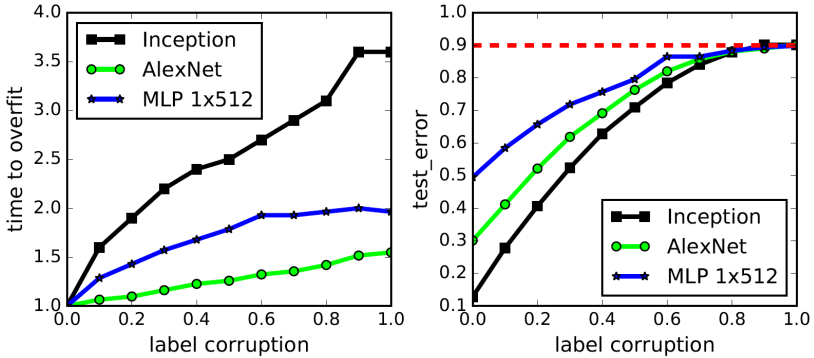
Modern machine learning models typically fit a huge number of parameters. Such overparameterization seems to be useful for:

- 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.

1. To set the stage, we will mainly be focusing on overparameterized learning. As we all know, machine learning models these days fit a large number of parameters compared to the number of observations. By parameters here, I mean either raw features in the dataset, or learned features. Such overparameterization seems to be useful for a number of things.
2. First, more parameters allow for rich classes of models that are capable of representing diverse set of real data.
3. Second, somewhat surprisingly, the optimization problem to fit these models simplifies dramatically in the overparametrized regime. So even simple, local optimization approaches often find near-optimal solutions.
4. Third, even more surprisingly, without any explicit regularization, the fitted models seem to perform well on unseen data in practice.
5. The focus of this talk will be on this third generalization aspect in overparameterized learning.

An influential experiment

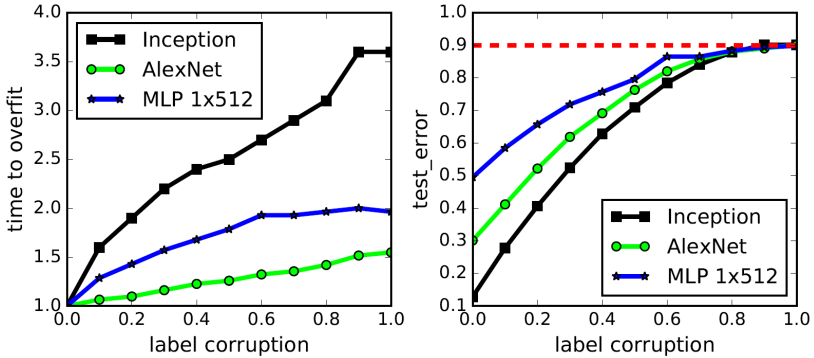


"Understanding deep learning requires rethinking generalization"
Zhang, Bengio, Hardt, Recht, Vinyals, 2017

- 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)

1. As an illustration, let's look at a real example. This is a plot from a paper from a few of years ago.

An influential experiment

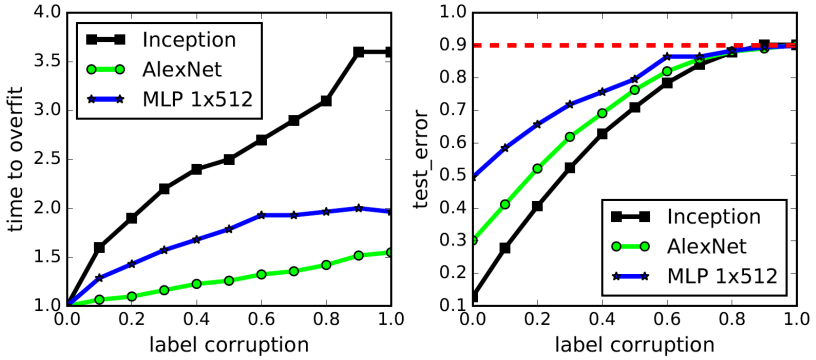


"Understanding deep learning requires rethinking generalization"
Zhang, Bengio, Hardt, Recht, Vinyals, 2017

- 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)

1. As an illustration, let's look at a real example. This is a plot from a paper from a few of years ago.
2. This is for a image classification problem using the CIFAR10 dataset, which has 10 categories, and about 60,000 observations.

An influential experiment

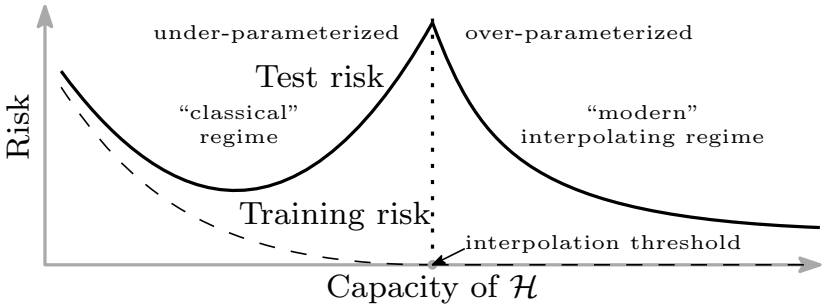


"Understanding deep learning requires rethinking generalization"
Zhang, Bengio, Hardt, Recht, Vinyals, 2017

- 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)

1. As an illustration, let's look at a real example. This is a plot from a paper from a few of years ago.
2. This is for a image classification problem using the CIFAR10 dataset, which has 10 categories, and about 60,000 observations.
3. The experiment adds increasing levels of artificial noise to the labels and trains three neural network architectures which are highly overparametrized (with parameters on the order of 20-30 times the number of observations).
4. The x-axis of the left plot shows the amount of label corruption. The y-axis show how much time is needed to fit completely with 0 training error to such label-corrupted data.
5. As we can observe, the network takes longer to fit with label corruption, but is still able to fit completely indicating that the models are rich enough to fit increasing levels of noise.
6. The right plot shows the test error with for the same models that are completely overfit at each label corruption level.
7. The striking thing about this plot is that even with the models trained with 0 error on highly noisy data, the prediction performance is still reasonably well.
8. For example, at noise level 0.2, the test error of models that completely overfit to the training data is still below the random chance which is 0.9 for this 10-category classification problem. The test error smoothly increasing with the amount of label corruption.

Peculiar generalization behavior: double descent

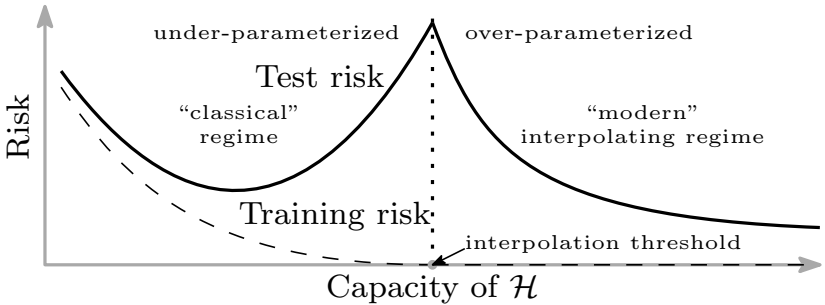


Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- 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.

1. Overparameterized models exhibit peculiar generalization behavior that's illustrated in this plot. This plot is from a paper by Belkin, Hsu, Ma, and Mandal.
2. They looked at various model classes and their generalization performance as you increase the model capacity, typically measured in terms of the number of model parameters.
3. They observed that the prediction risk obeys the classical bias-variance trade-off until the point of interpolation, but beyond interpolation, it again decreases.

Peculiar generalization behavior: double descent

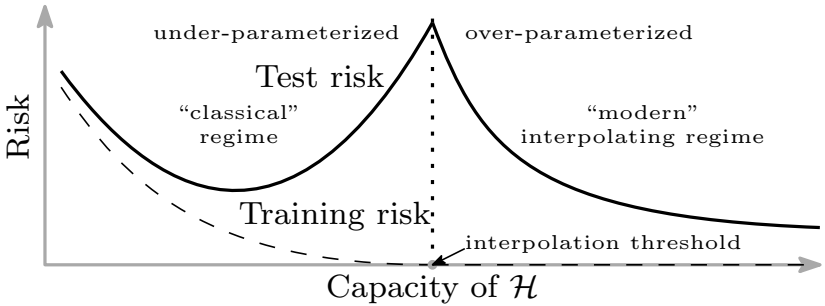


Belkin, Hsu, Ma, Mandal, 2018: "Reconciling modern machine learning practice and the bias variance tradeoff"

- 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.

1. Overparameterized models exhibit peculiar generalization behavior that's illustrated in this plot. This plot is from a paper by Belkin, Hsu, Ma, and Mandal.
2. They looked at various model classes and their generalization performance as you increase the model capacity, typically measured in terms of the number of model parameters.
3. They observed that the prediction risk obeys the classical bias-variance trade-off until the point of interpolation, but beyond interpolation, it again decreases.
4. This phenomenon is called double descent in the risk curve as a function of model capacity.

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.

1. Overparameterized models exhibit peculiar generalization behavior that's illustrated in this plot. This plot is from a paper by Belkin, Hsu, Ma, and Mandal.
2. They looked at various model classes and their generalization performance as you increase the model capacity, typically measured in terms of the number of model parameters.
3. They observed that the prediction risk obeys the classical bias-variance trade-off until the point of interpolation, but beyond interpolation, it again decreases.
4. This phenomenon is called double descent in the risk curve as a function of model capacity.
5. They found that such trend holds more generally for many classes of models beyond neural networks including kernel methods, random forests, boosting, etc.
6. The interesting aspect here is that the minimum of the prediction risk can happen on the right side of the curve, i.e., in the overparametrized regime.

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 ...

1. Motivated by this, there has been some work in the last couple years understanding generalization of interpolators in different settings using a variety of techniques.

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 ...

1. Motivated by this, there has been some work in the last couple years understanding generalization of interpolators in different settings using a variety of techniques.
2. For linear regression, there's been work understanding the risk behavior of the min-norm least squares interpolator. There are some results both in an asymptotic setting where the number of features grow with the number of observations and also in finite setting trying to understand the phenomenon of benign overfitting.

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 ...

1. Motivated by this, there has been some work in the last couple years understanding generalization of interpolators in different settings using a variety of techniques.
2. For linear regression, there's been work understanding the risk behavior of the min-norm least squares interpolator. There are some results both in an asymptotic setting where the number of features grow with the number of observations and also in finite setting trying to understand the phenomenon of benign overfitting.
3. Beyond linear regression, there's also been work on kernel methods with special kinds of kernels: kernels which are non-linear functions of inner product kernel, and Laplace kernels.

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 ...

1. Motivated by this, there has been some work in the last couple years understanding generalization of interpolators in different settings using a variety of techniques.
2. For linear regression, there's been work understanding the risk behavior of the min-norm least squares interpolator. There are some results both in an asymptotic setting where the number of features grow with the number of observations and also in finite setting trying to understand the phenomenon of benign overfitting.
3. Beyond linear regression, there's also been work on kernel methods with special kinds of kernels: kernels which are non-linear functions of inner product kernel, and Laplace kernels.
4. There's been work on nearest neighbor rules and simplicial interpolation. There's also work on kernel smoothing with singular kernels.

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 ...

1. Motivated by this, there has been some work in the last couple years understanding generalization of interpolators in different settings using a variety of techniques.
2. For linear regression, there's been work understanding the risk behavior of the min-norm least squares interpolator. There are some results both in an asymptotic setting where the number of features grow with the number of observations and also in finite setting trying to understand the phenomenon of benign overfitting.
3. Beyond linear regression, there's also been work on kernel methods with special kinds of kernels: kernels which are non-linear functions of inner product kernel, and Laplace kernels.
4. There's been work on nearest neighbor rules and simplicial interpolation. There's also work on kernel smoothing with singular kernels.
5. And there's been many interesting papers understanding the risk behavior of interpolators in different settings.

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"

1. So in nutshell: the current practice suggests that in nearly all applications, we should design our models to be massively overparametrized.

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"

1. So in nutshell: the current practice suggests that in nearly all applications, we should design our models to be massively overparametrized.
2. Once trained fully, these models will typically nearly interpolate the training data, that is achieve near zero training error.

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"

1. So in nutshell: the current practice suggests that in nearly all applications, we should design our models to be massively overparametrized.
2. Once trained fully, these models will typically nearly interpolate the training data, that is achieve near zero training error.
3. And still the trained models can and often do have good test error.

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"

1. So in nutshell: the current practice suggests that in nearly all applications, we should design our models to be massively overparametrized.
2. Once trained fully, these models will typically nearly interpolate the training data, that is achieve near zero training error.
3. And still the trained models can and often do have good test error.
4. Now, how much of this do we understand theoretically?

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"

1. So in nutshell: the current practice suggests that in nearly all applications, we should design our models to be massively overparametrized.
2. Once trained fully, these models will typically nearly interpolate the training data, that is achieve near zero training error.
3. And still the trained models can and often do have good test error.
4. Now, how much of this do we understand theoretically?
5. Not a whole lot in full generality.
6. There has been a flurry of work in the last few years, and we do understand some of this simplified models. But there is still a lot that we do not understand.

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"

1. So in nutshell: the current practice suggests that in nearly all applications, we should design our models to be massively overparametrized.
2. Once trained fully, these models will typically nearly interpolate the training data, that is achieve near zero training error.
3. And still the trained models can and often do have good test error.
4. Now, how much of this do we understand theoretically?
5. Not a whole lot in full generality.
6. There has been a flurry of work in the last few years, and we do understand some of this simplified models. But there is still a lot that we do not understand.
7. But, there has been a flurry of work in the last few years understanding generalization of near interpolators, and we do understand this story for simplified models, such as ridge regression, kernel regression, etc.

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"

1. So in nutshell: the current practice suggests that in nearly all applications, we should design our models to be massively overparametrized.
2. Once trained fully, these models will typically nearly interpolate the training data, that is achieve near zero training error.
3. And still the trained models can and often do have good test error.
4. Now, how much of this do we understand theoretically?
5. Not a whole lot in full generality.
6. There has been a flurry of work in the last few years, and we do understand some of this simplified models. But there is still a lot that we do not understand.
7. But, there has been a flurry of work in the last few years understanding generalization of near interpolators, and we do understand this story for simplified models, such as ridge regression, kernel regression, etc.
8. There are two nice survey papers on this by Bartlett, Montanari, Rakhlin and Belkin in Acta Numerica that summarize the current results on these, which I recommend for those who are interested in these topics.

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 $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, 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):

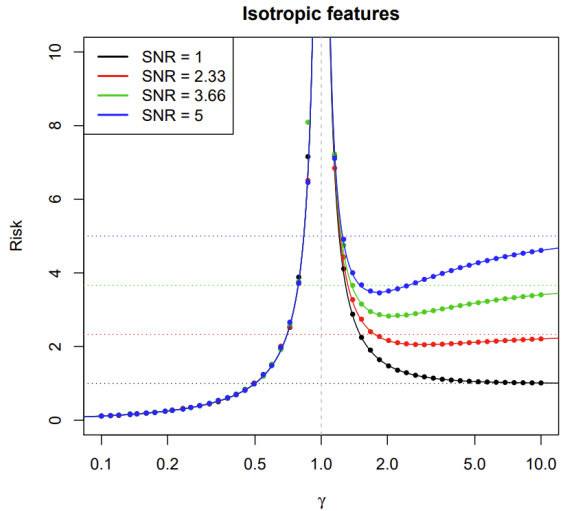
$$\widehat{\beta} = (X^\top X)^\dagger X^\top y = \lim_{\lambda \to 0^+} \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \frac{1}{n} \|y - Xb\|_2^2 + \lambda \|b\|_2^2 \right\}$$

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"

Just to recall: this is the double descent behaviour for the min ℓ_2-norm interpolator as a function of the aspect ratio p/n denotedy by γ.

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.

1. In particular, the so-called double or multiple descent risk behavior, in which the risk first increases as the ratio p/n increases up to certain phase transition threshold, and then decreases again.

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.

1. In particular, the so-called double or multiple descent risk behavior, in which the risk first increases as the ratio p/n increases up to certain phase transition threshold, and then decreases again.
2. There are two ways to view this.

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.

1. In particular, the so-called double or multiple descent risk behavior, in which the risk first increases as the ratio p/n increases up to certain phase transition threshold, and then decreases again.
2. There are two ways to view this.
3. One is by thinking the data dimension, p, as fixed. Then, double descent implies that as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.

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.

1. In particular, the so-called double or multiple descent risk behavior, in which the risk first increases as the ratio p/n increases up to certain phase transition threshold, and then decreases again.
2. There are two ways to view this.
3. One is by thinking the data dimension, p, as fixed. Then, double descent implies that as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.
4. Another is by thinking the number of observations n, as fixed. Then, double descent implies that as the data dimension p increases, the risk first increases, and then decreases. So in this sense, more features do not 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.

1. In particular, the so-called double or multiple descent risk behavior, in which the risk first increases as the ratio p/n increases up to certain phase transition threshold, and then decreases again.
2. There are two ways to view this.
3. One is by thinking the data dimension, p, as fixed. Then, double descent implies that as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.
4. Another is by thinking the number of observations n, as fixed. Then, double descent implies that as the data dimension p increases, the risk first increases, and then decreases. So in this sense, more features do not hurt.
5. We will focus on this first interpretation that more data can hurt the prediction procedure.

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.

1. In particular, the so-called double or multiple descent risk behavior, in which the risk first increases as the ratio p/n increases up to certain phase transition threshold, and then decreases again.
2. There are two ways to view this.
3. One is by thinking the data dimension, p, as fixed. Then, double descent implies that as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.
4. Another is by thinking the number of observations n, as fixed. Then, double descent implies that as the data dimension p increases, the risk first increases, and then decreases. So in this sense, more features do not hurt.
5. We will focus on this first interpretation that more data can hurt the prediction procedure.

Outline

Now we will switch to the risk monotonziation aspect.

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.

1. In general, when we have i.i.d. data, we expect that more data will help in prediction or estimation.

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.

1. In general, when we have i.i.d. data, we expect that more data will help in prediction or estimation.
2. One consequence of the double descent behaviour is that, if you fix feature size p, then as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.

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.

1. In general, when we have i.i.d. data, we expect that more data will help in prediction or estimation.
2. One consequence of the double descent behaviour is that, if you fix feature size p, then as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.
3. And thus, a procedure that leads to worse risk as number of observations increases is somehow not using the data properly, and is "suboptimal" in some sense.

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.

1. In general, when we have i.i.d. data, we expect that more data will help in prediction or estimation.
2. One consequence of the double descent behaviour is that, if you fix feature size p, then as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.
3. And thus, a procedure that leads to worse risk as number of observations increases is somehow not using the data properly, and is "suboptimal" in some sense.
4. The key question that we ask is if it is possible to modify an arbitrary prediction procedure so that it has a monotonic risk behavior?

Motivation and main punchlines

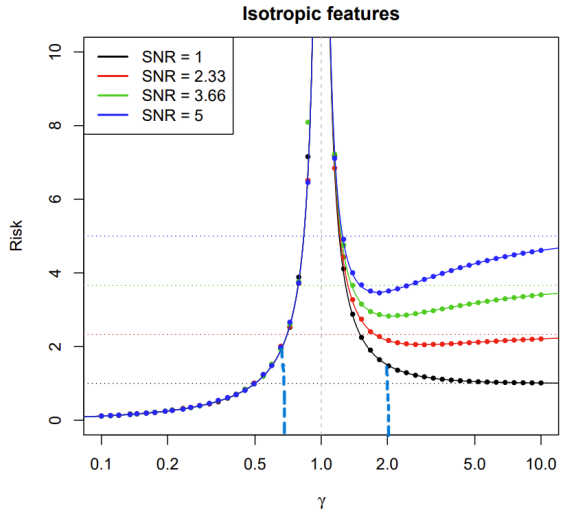
- When the data comprises of i.i.d. observations, we expect that more data will help in prediction or estimation.
- Double or multiple descent behaviour implies that for fixed feature size p (large value), as sample size increases the risk first decreases and then increases. More data can hurt!
- A procedure 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.

1. In general, when we have i.i.d. data, we expect that more data will help in prediction or estimation.
2. One consequence of the double descent behaviour is that, if you fix feature size p, then as the sample size increases, the risk first decreases, but then increases. So in a sense, more data hurts.
3. And thus, a procedure that leads to worse risk as number of observations increases is somehow not using the data properly, and is "suboptimal" in some sense.
4. The key question that we ask is if it is possible to modify an arbitrary prediction procedure so that it has a monotonic risk behavior?
5. We show in this work that it is possible. In particular, we propose two methods, think of them as wrapper methods, dubbed zero-step and one-step, that can take as input any arbitrary prediction procedure and return modified procedures who monotonic risk and avoid double or multiple descent risk behaviour. This is one via subsampling and simply using less observations for better risk behavior.

Method overview and the problem



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

This is the double descent behaviour for the min ℓ_2-norm interpolator as a function of the aspect ratio p/n denotedy by γ. If we are operating at an aspect ration of say 0.8, then it is better to move to a higher aspect ratio of say 2, in terms of risk behaviour.

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.

1. Now this simple strategy will work if we knew the oracle risk profile. But in order to implement such a strategy in practice using available data, the main problem is how do we know if a smaller n will actually lead to a better risk?

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.

1. Now this simple strategy will work if we knew the oracle risk profile. But in order to implement such a strategy in practice using available data, the main problem is how do we know if a smaller n will actually lead to a better risk?
2. And moreover, what is the best sample size to reduce the dataset to in order to attain the best possible risk?

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.

1. Now this simple strategy will work if we knew the oracle risk profile. But in order to implement such a strategy in practice using available data, the main problem is how do we know if a smaller n will actually lead to a better risk?
2. And moreover, what is the best sample size to reduce the dataset to in order to attain the best possible risk?
3. And one solution to this is via 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

1. Here's the basic idea what we call the zero-step method.

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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ

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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ
3. We first construct a dense grid of aspect ratios higher than the given aspect ratio γ by considering datasets of samples sizes smaller than n and estimate the risk profile on those aspect ratios using a test set

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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ
3. We first construct a dense grid of aspect ratios higher than the given aspect ratio γ by considering datasets of samples sizes smaller than n and estimate the risk profile on those aspect ratios using a test set
4. We then select the aspect ratio and the corresponding predictor fitted at that aspect ratio that gives the best estimated risk

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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ
3. We first construct a dense grid of aspect ratios higher than the given aspect ratio γ by considering datasets of samples sizes smaller than n and estimate the risk profile on those aspect ratios using a test set
4. We then select the aspect ratio and the corresponding predictor fitted at that aspect ratio that gives the best estimated risk
5. And finally, we show that such modified procedure has a risk that's monotone in the aspect ratio

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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ
3. We first construct a dense grid of aspect ratios higher than the given aspect ratio γ by considering datasets of samples sizes smaller than n and estimate the risk profile on those aspect ratios using a test set
4. We then select the aspect ratio and the corresponding predictor fitted at that aspect ratio that gives the best estimated risk
5. And finally, we show that such modified procedure has a risk that's monotone in the aspect ratio
6. Some highlight of this procedure are:

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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ
3. We first construct a dense grid of aspect ratios higher than the given aspect ratio γ by considering datasets of samples sizes smaller than n and estimate the risk profile on those aspect ratios using a test set
4. We then select the aspect ratio and the corresponding predictor fitted at that aspect ratio that gives the best estimated risk
5. And finally, we show that such modified procedure has a risk that's monotone in the aspect ratio
6. Some highlight of this procedure are:
7. The method can applied for any generic starting procedure, along with common loss functions of interest.

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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ
3. We first construct a dense grid of aspect ratios higher than the given aspect ratio γ by considering datasets of samples sizes smaller than n and estimate the risk profile on those aspect ratios using a test set
4. We then select the aspect ratio and the corresponding predictor fitted at that aspect ratio that gives the best estimated risk
5. And finally, we show that such modified procedure has a risk that's monotone in the aspect ratio
6. Some highlight of this procedure are:
7. The method can applied for any generic starting procedure, along with common loss functions of interest.
8. The method is model agnostic and require minimal distributional assumptions to show the monotonicity property.

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

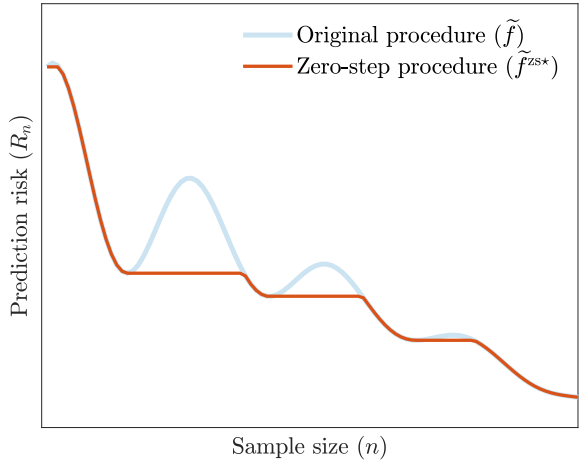
Method highlights:

- 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

1. Here's the basic idea what we call the zero-step method.
2. Suppose we are given any arbitrary prediction procedure that's operating at some given aspect ratio γ
3. We first construct a dense grid of aspect ratios higher than the given aspect ratio γ by considering datasets of samples sizes smaller than n and estimate the risk profile on those aspect ratios using a test set
4. We then select the aspect ratio and the corresponding predictor fitted at that aspect ratio that gives the best estimated risk
5. And finally, we show that such modified procedure has a risk that's monotone in the aspect ratio
6. Some highlight of this procedure are:
7. The method can applied for any generic starting procedure, along with common loss functions of interest.
8. The method is model agnostic and require minimal distributional assumptions to show the monotonicity property.
9. Moreover, the method works even there are diverging risks at various aspect ratios, which is common in the overparameterized settings

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 $\min_{m \leq n} R_m$.



Here's a cartoon illustration of what the risk monotization looks like. For every n, we will return a predictor whose risk is no more than the risk at any smaller sample size. So in this sense, the resulting risk profile would the largest monotone function below the original risk profile.

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 $n^{1/2} \le k \le |\mathcal{D}_{\text{tr}}|$, 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
$$\left\{ \frac{p}{n^{1/2}}, \frac{p}{n^{1/2}+1}, \ldots, \frac{p}{|\mathcal{D}_{\text{tr}}|} \right\} \;\text{ " } \to \text{ " }\; [\gamma, \infty].$$
The set of aspect ratios for the predictors $\widetilde{f}_k$ covers $[\gamma, \infty]$.
- Choose one out of $\widetilde{f}_k$, $n^{1/2} \le k \le |\mathcal{D}_{\text{tr}}|$ using an estimate of out-of-sample risk computed from $\mathcal{D}_{\text{te}}$ This is split sample cross-validation.

I will skip the details how we actually do the cross-validation in the interest of time, but I am happy to discuss more either towards the end or later offline.

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.

1. Here's an informal statement that we can show. If the original prediction procedure has a cerain risk profile, then the risk profile of the zero-step procedure would be a monotonized version of that risk profile.

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:

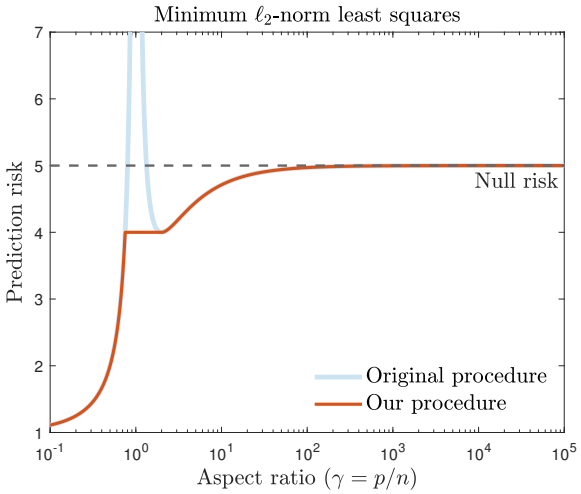
$$R(\widehat{f}^{\text{cv}}) \;=\; \inf_{\zeta \in [\gamma, \infty]} R^{\text{det}}(\zeta; \widehat{f}) \;\times\; (1 + o_p(1)).$$

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.

1. Here's an informal statement that we can show. If the original prediction procedure has a cerain risk profile, then the risk profile of the zero-step procedure would be a monotonized version of that risk profile.
2. This result requires minimal ditributional assumptions, that's unlike other results in overparameterized litetaure, which require strong assumptions on the data generating distribution.

Risk monotonization (illustration)



As an illustration, for the min ℓ_2-norm and ℓ_1-norm interpolators, here's how the risk monotonization by the zero-step procedure looks like.

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)

1. So let me start wrapping up by giving an overall summary of our contributions in this work.

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)

1. So let me start wrapping up by giving an overall summary of our contributions in this work.
2. We have introduced a general-purpose method, dubbed zero-step, that can provably monotonizes the risk of any given predictor.

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)

1. So let me start wrapping up by giving an overall summary of our contributions in this work.
2. We have introduced a general-purpose method, dubbed zero-step, that can provably monotonizes the risk of any given predictor.
3. The main idea behind our approach is that of cross-validation with careful splitting of data.

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)

1. So let me start wrapping up by giving an overall summary of our contributions in this work.
2. We have introduced a general-purpose method, dubbed zero-step, that can provably monotonizes the risk of any given predictor.
3. The main idea behind our approach is that of cross-validation with careful splitting of data.

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)

1. So let me start wrapping up by giving an overall summary of our contributions in this work.
2. We have introduced a general-purpose method, dubbed zero-step, that can provably monotonizes the risk of any given predictor.
3. The main idea behind our approach is that of cross-validation with careful splitting of data.
4. We also have a one-step variant that improves on the zero-step procedure and also has a monotone risk behavior. This is akin to boosting.

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:

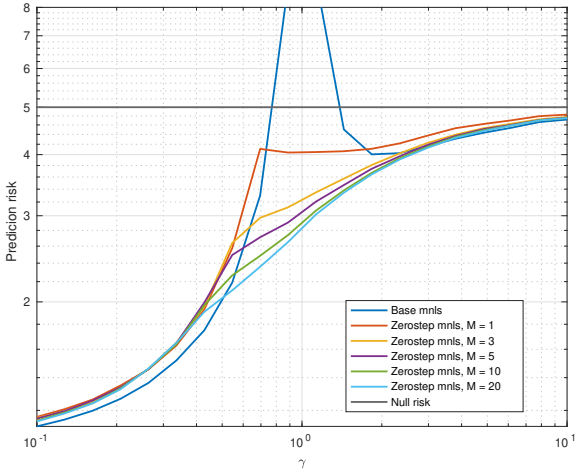
- We also introduce a one-step prediction procedure inspired by classical one-step estimator that improves on zero-step procedure (similar to boosting)
- Both zero-step and one-step procedures can be further improved by multiple subsamplings and averaging (similar to bagging)

1. So let me start wrapping up by giving an overall summary of our contributions in this work.
2. We have introduced a general-purpose method, dubbed zero-step, that can provably monotonizes the risk of any given predictor.
3. The main idea behind our approach is that of cross-validation with careful splitting of data.
4. We also have a one-step variant that improves on the zero-step procedure and also has a monotone risk behavior. This is akin to boosting.
5. We also consider subsampling more than once and averaging predictors fitted on different subsamples. This is akin 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 $r_{\min}, r_{\max}$ independent of p with $0 < r_{\min} \leq r_{\max} < \infty$ such that $r_{\min}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$.
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

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

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

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 $r_{\min}, r_{\max}$ independent of p with $0 < r_{\min} \leq r_{\max} < \infty$ such that $r_{\min}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$.
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 $r_{\min}, r_{\max}$ independent of p with $0 < r_{\min} \leq r_{\max} < \infty$ such that $r_{\min}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$.
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

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

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

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 $r_{\min}, r_{\max}$ independent of p with $0 < r_{\min} \leq r_{\max} < \infty$ such that $r_{\min}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$.
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)

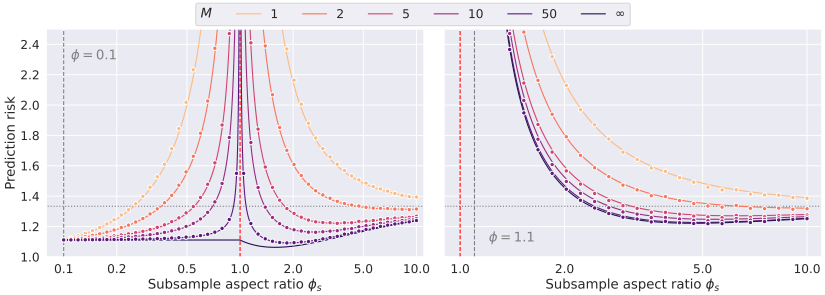
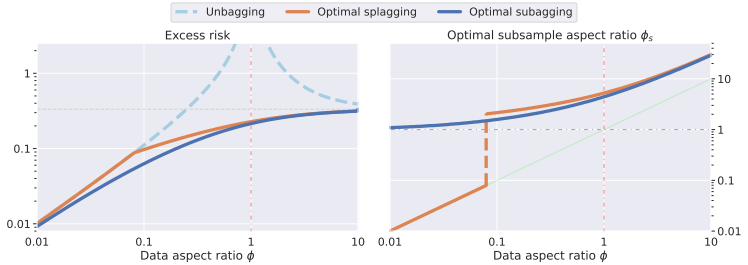


Figure: Asymptotic prediction risk curves for bagged ridgeless predictors ($\lambda = 0$), under AR1 model when $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The null risk is marked as a dotted line. For each value of 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

$$\sup_{M \in \mathbb{N}, \phi_s \in [\phi, \infty]} \mathscr{R}^{\text{sub}}_{0,M}(\phi, \phi_s) = \underbrace{\mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s^*)}_{\text{optimal bagged risk}}$$
$$< \min\{\underbrace{\mathscr{R}^{\text{sub}}_{0,1}(\phi, \phi)}_{\text{unbagged risk}}, \underbrace{\mathscr{R}^{\text{sub}}_{0,1}(\phi, \infty)}_{\text{null risk}}\}$$

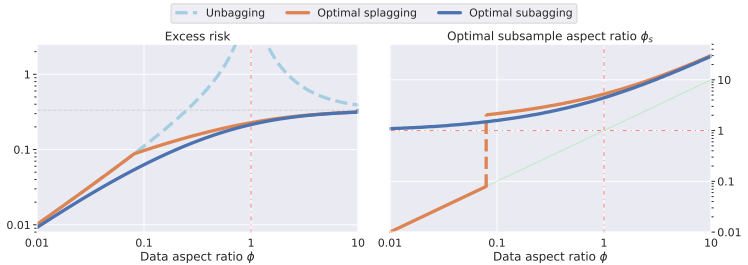


Subagged ridgeless *interpolators* always outperform subagged least squares, even when the full data has more observations than the number of features.

Optimal bagged ridgeless predictor

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

$$\sup_{M \in \mathbb{N}, \phi_s \in [\phi, \infty]} \mathscr{R}^{\texttt{sub}}_{0,M}(\phi, \phi_s) = \underbrace{\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi, \phi_s^*)}_{\text{optimal bagged risk}}$$
$$< \min\{\underbrace{\mathscr{R}^{\texttt{sub}}_{0,1}(\phi, \phi)}_{\text{unbagged risk}}, \underbrace{\mathscr{R}^{\texttt{sub}}_{0,1}(\phi, \infty)}_{\text{null risk}}\}$$



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.

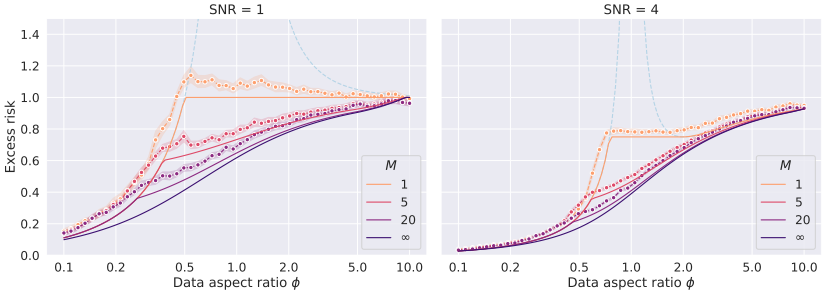
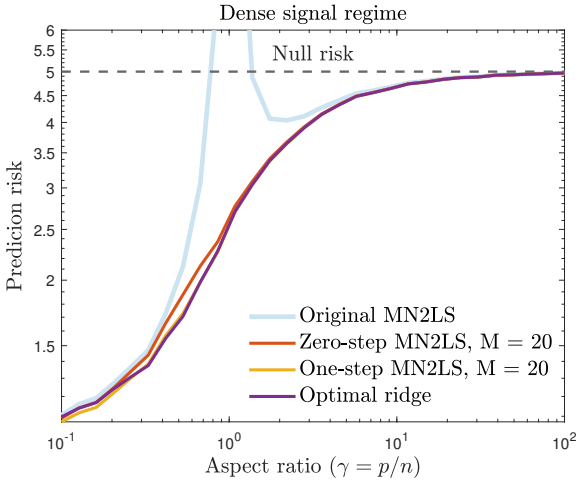


Figure: Asymptotic excess risk curves for cross-validated bagged ridgeless predictors ($\lambda = 0$), under the isotopic model when $\rho^2 = 1$ for varying $\texttt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M with replacement. For each value of M, the points denote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

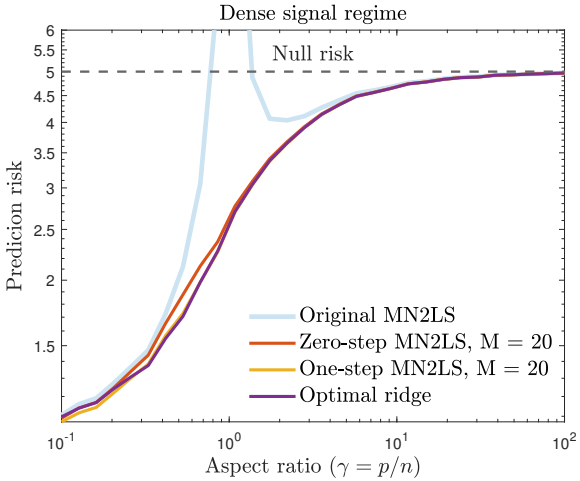
Comparison with optimal ridge regularization



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?

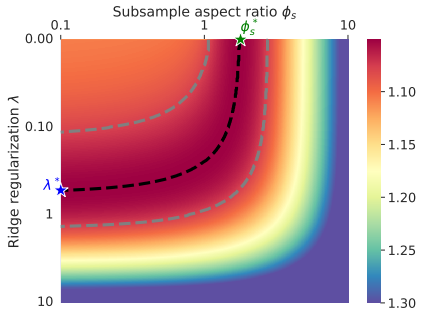
Outline

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):

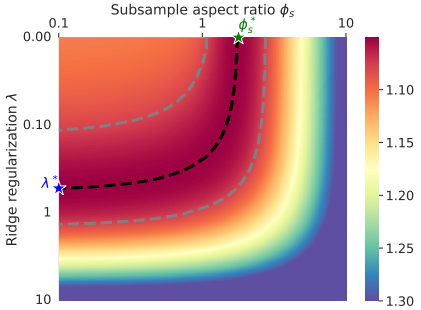
$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda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

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):

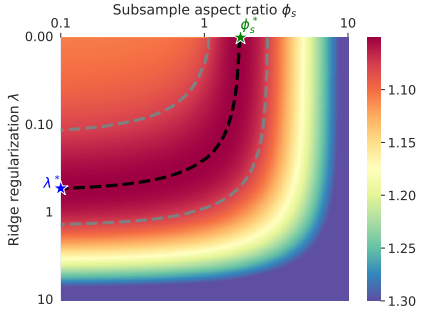
$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda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

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):

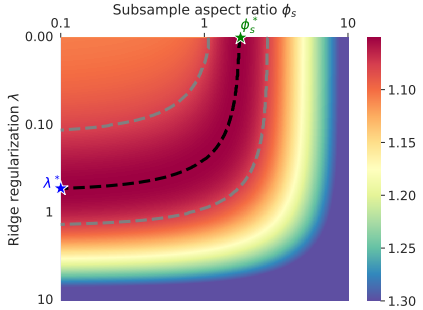
$$\underbrace{\min_{\phi_s \geq \phi} \mathscr{R}^{\text{sub}}_{0,\infty}(\phi, \phi_s)}_{\text{opt. ridgeless ensemble}} = \underbrace{\min_{\lambda \geq 0} \mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi)}_{\text{opt. ridge predictor}} = \underbrace{\min_{\substack{\phi_s \geq \phi, \\ \lambda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

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 $\beta_0 = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\mathbb{E}[\boldsymbol{x}y]$ 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:

$$R(\widehat{\beta}; \boldsymbol{A}, \boldsymbol{b}, \beta_0) = \frac{1}{\mathrm{nrow}(\boldsymbol{A})}\|L(\widehat{\beta} - \beta_0)\|_2^2, \tag{1}$$

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 $\beta_0 = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\mathbb{E}[\boldsymbol{x}y]$ 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:

$$R(\widehat{\beta}; \boldsymbol{A}, \boldsymbol{b}, \beta_0) = \frac{1}{\mathrm{nrow}(\boldsymbol{A})} \|L(\widehat{\beta} - \beta_0)\|_2^2, \tag{1}$$

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

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.

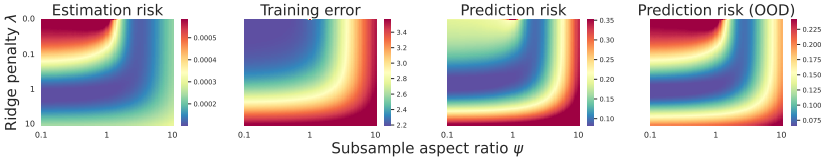
Data assumptions:

- 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:

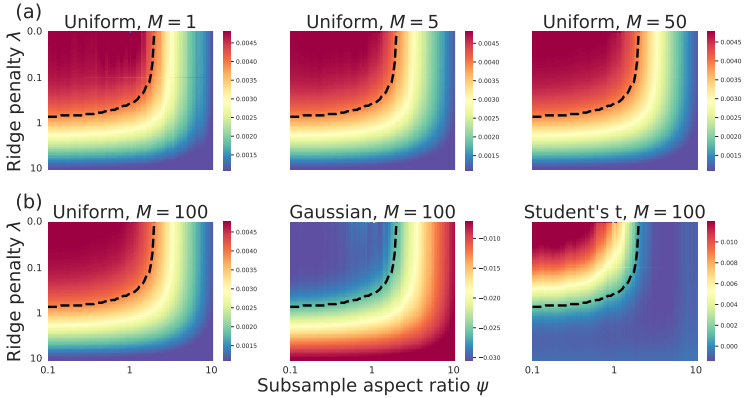
$$R\big(\widehat{\beta}^{\lambda_1}_{\lfloor p/\psi_1 \rfloor, \infty}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0\big) \;\simeq\; R\big(\widehat{\beta}^{\lambda_2}_{\lfloor p/\psi_2 \rfloor, \infty}; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\beta}_0\big). \tag{2}$$



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:

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

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

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

- 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:

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

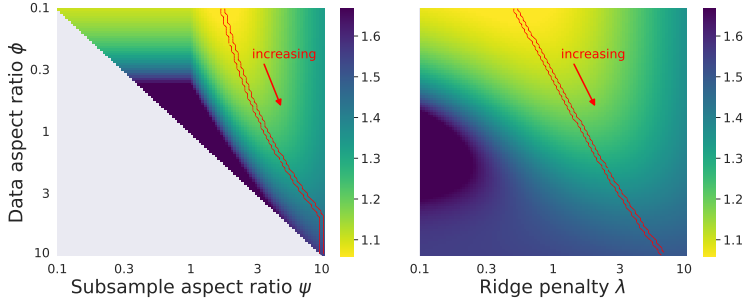
Define the data-dependent path $\mathcal{P}_n = \mathcal{P}(\bar{\lambda}_n; \phi_n, \bar{\psi}_n)$.

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.

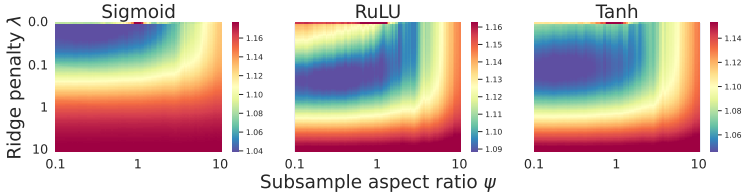


Extension 1: Equivalences for random features

Conjecture. Define $\phi_n = p/n$. Let $k \leq n$ 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 $\mathcal{P}_n$.

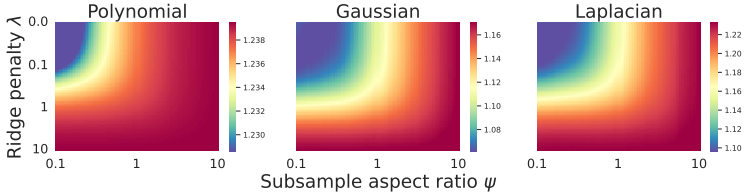


Extension 2: Equivalences for kernel features

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.

1. Alright, to conclude my talk: in this thesis, we studied three aspects of overparameterized learning: cross-validation, risk monotonization, and model complexity. And the takeaways are:

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.

1. Alright, to conclude my talk: in this thesis, we studied three aspects of overparameterized learning: cross-validation, risk monotonization, and model complexity. And the takeaways are:
2. One that cross-validation still works in the overparameterized regime for ridge regression, even when the regularization is 0 via suitable analytic continuation.

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.

1. Alright, to conclude my talk: in this thesis, we studied three aspects of overparameterized learning: cross-validation, risk monotonization, and model complexity. And the takeaways are:
2. One that cross-validation still works in the overparameterized regime for ridge regression, even when the regularization is 0 via suitable analytic continuation.
3. Second that it is possible to modify any arbitrary prediction procedure so that it has monotonic risk behavior via suitable subsampling and cross-validation.

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.

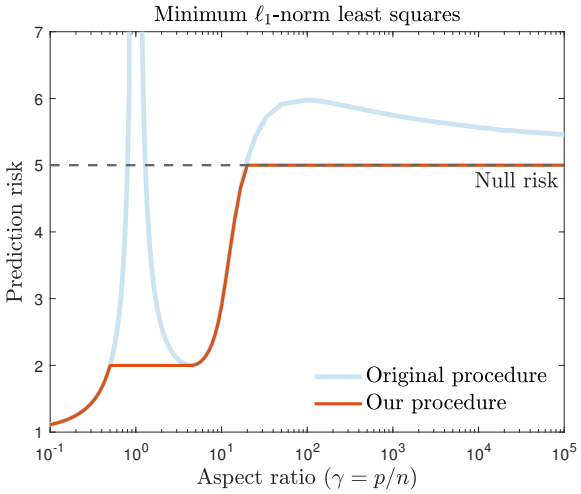
3. There are connections between the implicit regularization induced by subsampling and explicit ridge regularization for subsampled ridge ensembles.

1. Alright, to conclude my talk: in this thesis, we studied three aspects of overparameterized learning: cross-validation, risk monotonization, and model complexity. And the takeaways are:
2. One that cross-validation still works in the overparameterized regime for ridge regression, even when the regularization is 0 via suitable analytic continuation.
3. Second that it is possible to modify any arbitrary prediction procedure so that it has monotonic risk behavior via suitable subsampling and cross-validation.
4. And three that there is a principled measure of model complexity in the overparameterized regime in the form of random-X degrees of freedom.

Thanks for listening!

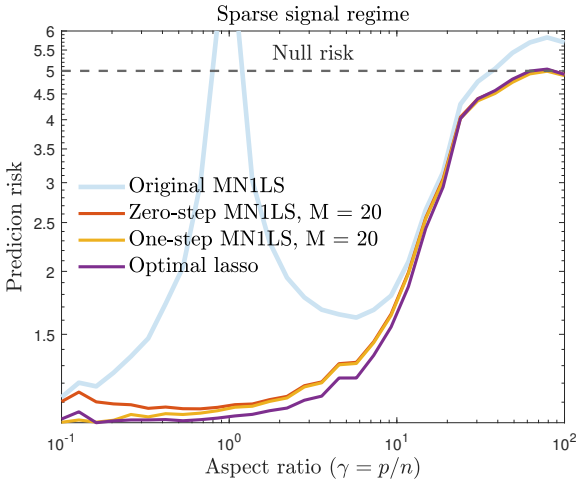
Questions/comments/thoughts?

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