Revisiting Model Complexity in the Wake of Overparameterized Learning

Pratik Patil

Carnegie Mellon University

TOPML 2022

Based on joint work with Ryan Tibshirani

Overparameterization in machine learning

Modern machine learning involves fitting 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 on training data
- Generalization: despite overfitting, models generalize well in practice

This talk is broadly about the generalization aspect.

Overparameterization in machine learning

Modern machine learning involves fitting 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 on training data
- Generalization: despite overfitting, models generalize well in practice

This talk is broadly about the generalization aspect.

Overparameterization in machine learning

Modern machine learning involves fitting 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 on training data
- Generalization: despite overfitting, models generalize well in practice

This talk is broadly about the generalization aspect.

Overparameterization in machine learning

Modern machine learning involves fitting 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 on training data
- Generalization: despite overfitting, models generalize well in practice

This talk is broadly about the generalization aspect.

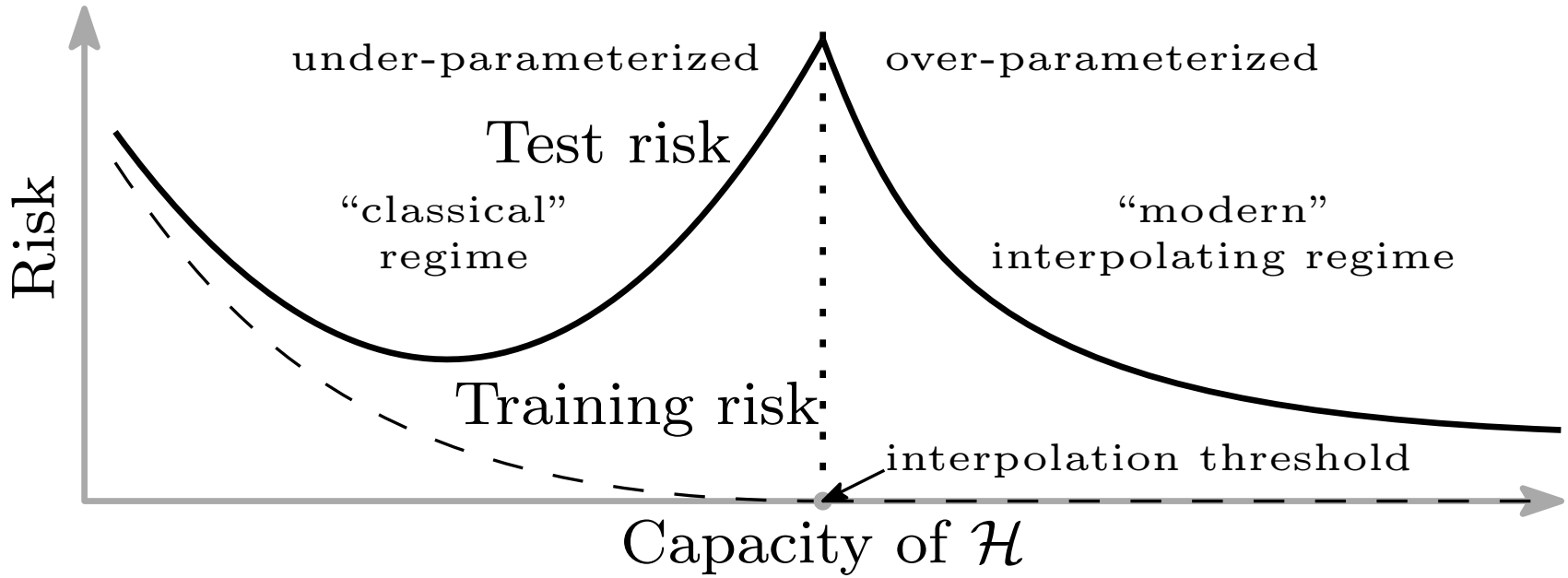
Overparameterization in machine learning

Modern machine learning involves fitting 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 on training data
- Generalization: despite overfitting, models generalize well in practice

This talk is broadly about the generalization aspect.

The double/multiple descent phenomenon

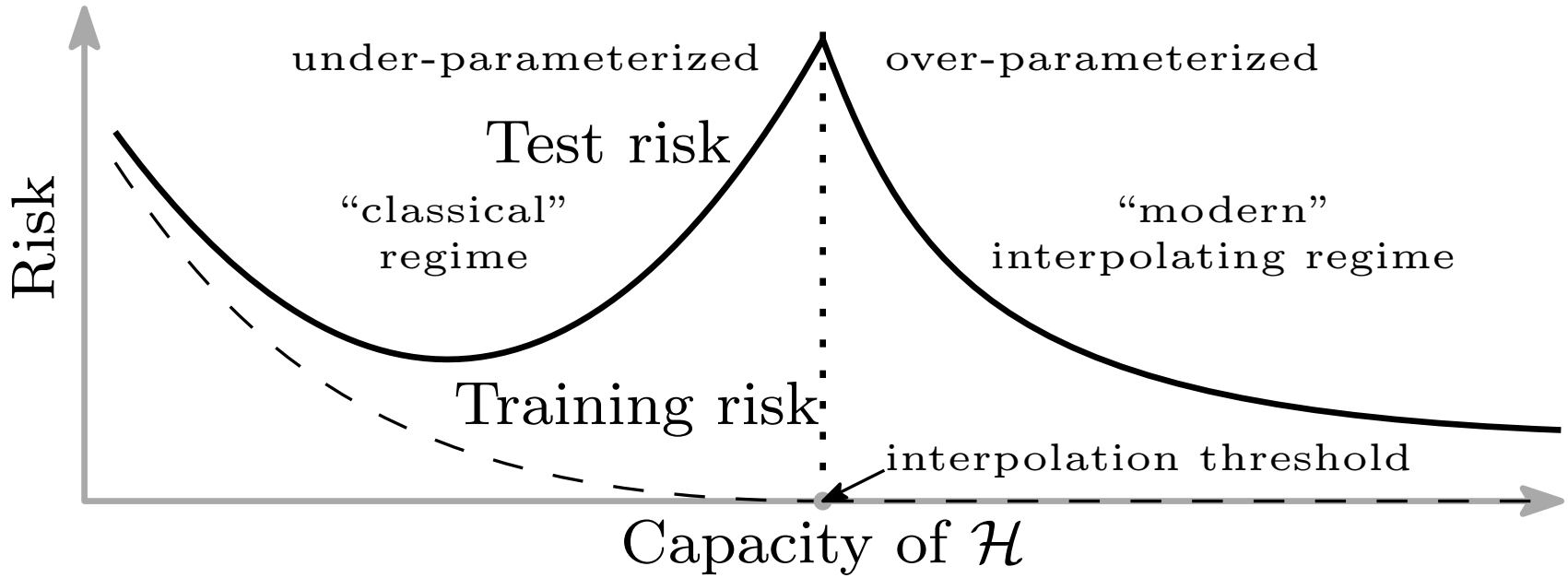


"Reconciling modern machine learning practice and the bias variance tradeoff"
Belkin, Hsu, Ma, Mandal, 2018

- 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.

Is there an equivalent "reparameterization" of "overparameterization"?

The double/multiple descent phenomenon

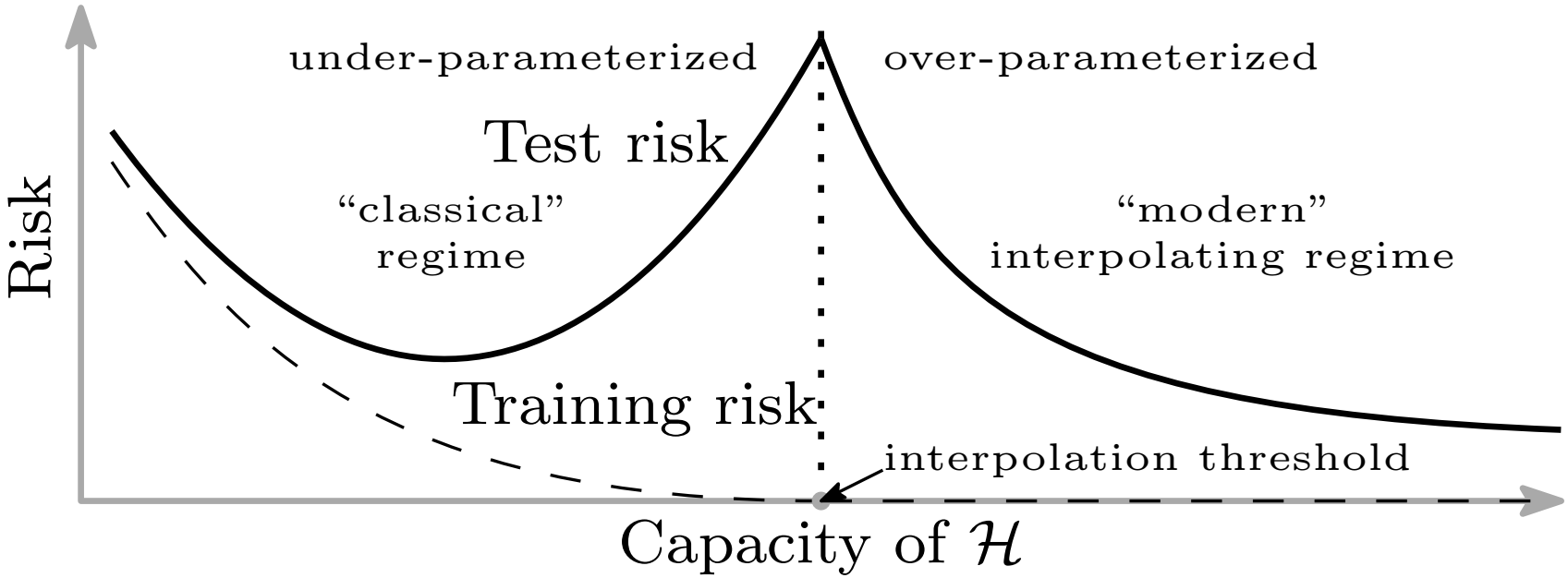


"Reconciling modern machine learning practice and the bias variance tradeoff"
Belkin, Hsu, Ma, Mandal, 2018

- 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.

Is there an equivalent "reparameterization" of "overparameterization"?

The double/multiple descent phenomenon

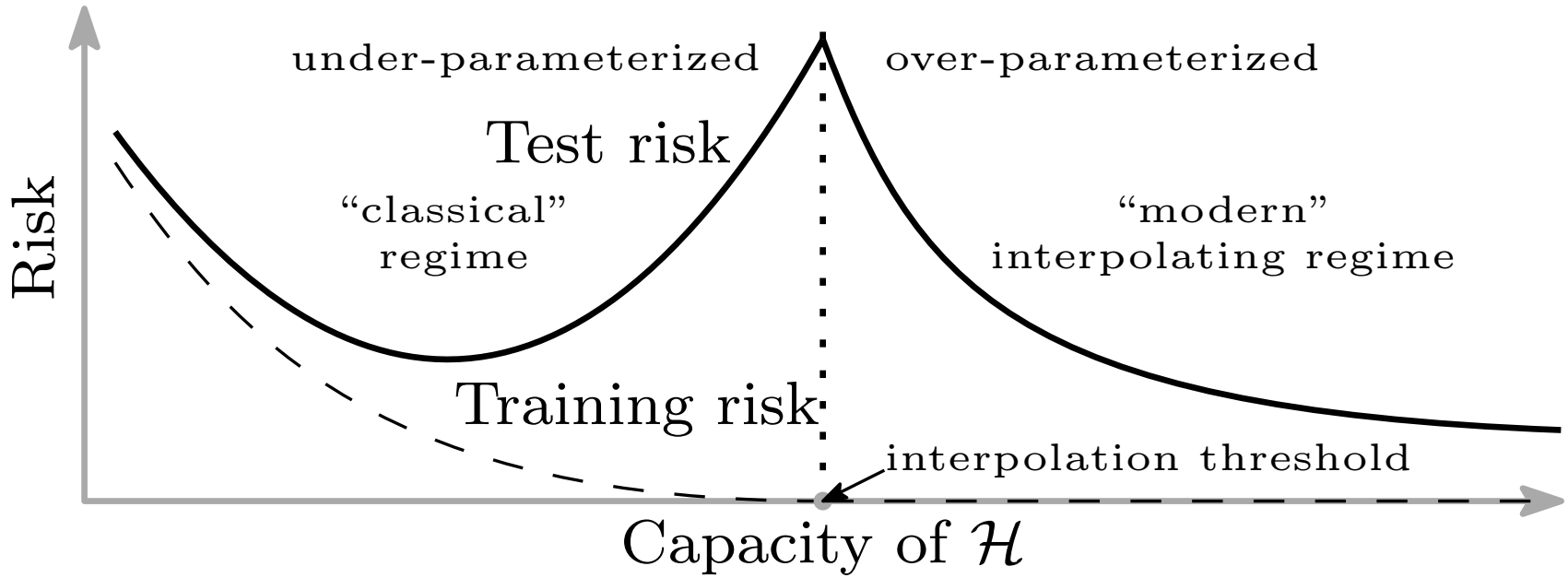


"Reconciling modern machine learning practice and the bias variance tradeoff"
Belkin, Hsu, Ma, Mandal, 2018

- 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.

Is there an equivalent "reparameterization" of "overparameterization"?

The double/multiple descent phenomenon

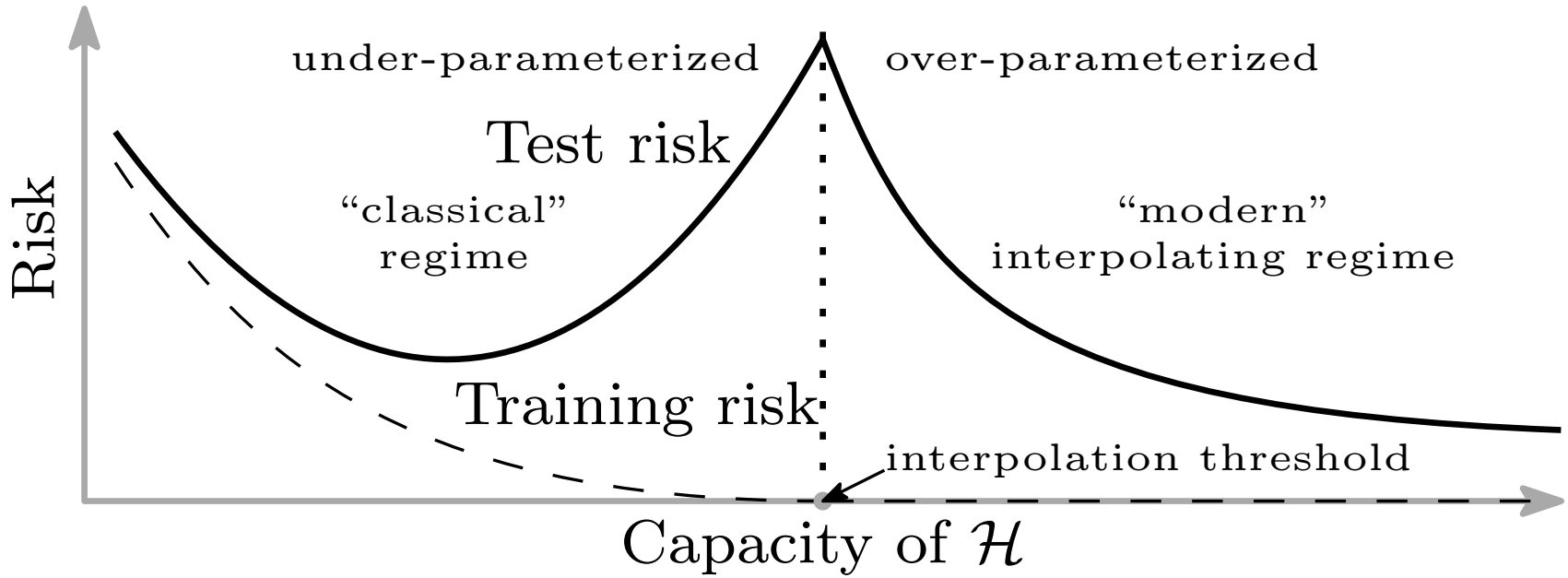


"Reconciling modern machine learning practice and the bias variance tradeoff"
Belkin, Hsu, Ma, Mandal, 2018

- 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.

Is there an equivalent "reparameterization" of "overparameterization"?

The double/multiple descent phenomenon



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

- 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.

Is there an equivalent "reparameterization" of "overparameterization"?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Conceptual and theoretical challenges

Current machine learning practice suggests:

- we should design models to be highly overparametrized
- train until nearly zero training error (i.e., interpolate training data)
- the trained models still can generalize well and have good test error

Main statistical challenges for us:

- benign overfitting conceptually breaks classical bias-variance tradeoff
- trained models fall outside the realm of uniform convergence
- one of core questions is how to correctly measure model complexity

This work attempts to answer the following questions:

- Is there a good measure of model complexity for predictive models?
- How to compare model complexity of different (near) interpolators?

Takaway points

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise
- Results when applied to some illustrative examples:
  - min ℓ_2/ℓ_1-norm interpolators: the complexity measures maximized when $n \approx p$ and typically decreases as p increases beyond n
  - we can reparameterize every overparameterized model into an equivalent underparameterized model in terms of risk behavior

Based on extension of ideas from optimism theory and degrees of freedom.

Takaway points

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise
- Results when applied to some illustrative examples:
  - min ℓ_2/ℓ_1-norm interpolators: the complexity measures maximized when $n \approx p$ and typically decreases as p increases beyond n
  - we can reparameterize every overparameterized model into an equivalent underparameterized model in terms of risk behavior

Based on extension of ideas from optimism theory and degrees of freedom.

Takaway points

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise
- Results when applied to some illustrative examples:
  - min ℓ_2/ℓ_1-norm interpolators: the complexity measures maximized when $n \approx p$ and typically decreases as p increases beyond n
  - we can reparameterize every overparameterized model into an equivalent underparameterized model in terms of risk behavior

Based on extension of ideas from optimism theory and degrees of freedom.

Takaway points

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise
- Results when applied to some illustrative examples:
  - min ℓ_2/ℓ_1-norm interpolators: the complexity measures maximized when $n \approx p$ and typically decreases as p increases beyond n
  - we can reparameterize every overparameterized model into an equivalent underparameterized model in terms of risk behavior

Based on extension of ideas from optimism theory and degrees of freedom.

Takaway points

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise
- Results when applied to some illustrative examples:
  - min ℓ_2/ℓ_1-norm interpolators: the complexity measures maximized when $n \approx p$ and typically decreases as p increases beyond n
  - we can reparameterize every overparameterized model into an equivalent underparameterized model in terms of risk behavior

Based on extension of ideas from optimism theory and degrees of freedom.

Takaway points

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise
- Results when applied to some illustrative examples:
  - min ℓ_2/ℓ_1-norm interpolators: the complexity measures maximized when $n \approx p$ and typically decreases as p increases beyond n
  - we can reparameterize every overparameterized model into an equivalent underparameterized model in terms of risk behavior

Based on extension of ideas from optimism theory and degrees of freedom.

Takaway points

- Propose measures of model complexity that are:
  - algorithm-specific and applies for any prediction algorithm
  - produce a number between 0 and n (the number of observations)
- Two variants of model complexities are:
  - emergent model complexity that depends on the prediction algorithm as well as underlying the regression function
  - intrinsic model complexity that depends on the prediction algorithm only and its adaptability to pure noise
- Results when applied to some illustrative examples:
  - min ℓ_2/ℓ_1-norm interpolators: the complexity measures maximized when $n \approx p$ and typically decreases as p increases beyond n
  - we can reparameterize every overparameterized model into an equivalent underparameterized model in terms of risk behavior

Based on extension of ideas from optimism theory and degrees of freedom.

Outline

Degrees of freedom at large

Degrees of freedom means different things in different fields, but they're more or less similar. There is a core concept behind it.

It essentially has to do with the dimension or effective number of parameters of "something".

- In mechanics, that something = mechanical system.
- In physics and chemistry, something = physical system.
- In statistics, something = prediction procedure.

A nice commonality is that we can usually guess the degrees of freedom based on intuition, at least qualitatively.

Degrees of freedom at large

Degrees of freedom means different things in different fields, but they're more or less similar. There is a core concept behind it.

It essentially has to do with the dimension or effective number of parameters of "something".

- In mechanics, that something = mechanical system.
- In physics and chemistry, something = physical system.
- In statistics, something = prediction procedure.

A nice commonality is that we can usually guess the degrees of freedom based on intuition, at least qualitatively.

Degrees of freedom at large

Degrees of freedom means different things in different fields, but they're more or less similar. There is a core concept behind it.

It essentially has to do with the dimension or effective number of parameters of "something".

- In mechanics, that something = mechanical system.
- In physics and chemistry, something = physical system.
- In statistics, something = prediction procedure.

A nice commonality is that we can usually guess the degrees of freedom based on intuition, at least qualitatively.

Degrees of freedom at large

Degrees of freedom means different things in different fields, but they're more or less similar. There is a core concept behind it.

It essentially has to do with the dimension or effective number of parameters of "something".

- In mechanics, that something = mechanical system.
- In physics and chemistry, something = physical system.
- In statistics, something = prediction procedure.

A nice commonality is that we can usually guess the degrees of freedom based on intuition, at least qualitatively.

Degrees of freedom at large

Degrees of freedom means different things in different fields, but they're more or less similar. There is a core concept behind it.

It essentially has to do with the dimension or effective number of parameters of "something".

- In mechanics, that something = mechanical system.
- In physics and chemistry, something = physical system.
- In statistics, something = prediction procedure.

A nice commonality is that we can usually guess the degrees of freedom based on intuition, at least qualitatively.

Degrees of freedom at large

Degrees of freedom means different things in different fields, but they're more or less similar. There is a core concept behind it.

It essentially has to do with the dimension or effective number of parameters of "something".

- In mechanics, that something = mechanical system.
- In physics and chemistry, something = physical system.
- In statistics, something = prediction procedure.

A nice commonality is that we can usually guess the degrees of freedom based on intuition, at least qualitatively.

Degrees of freedom in statistics

Degrees of freedom in statistics is defined, intuitively, as the effective number of parameters used by a prediction procedure.

While this seems vague, it has a precise definition for a broad class of estimation problems. We will define this shortly.

Why is this an important concept? Why you would ever go to the trouble of describing degrees of freedom?

Essentially, it provides a way to put two different prediction procedures on equal footing.

Degrees of freedom in statistics

Degrees of freedom in statistics is defined, intuitively, as the effective number of parameters used by a prediction procedure.

While this seems vague, it has a precise definition for a broad class of estimation problems. We will define this shortly.

Why is this an important concept? Why you would ever go to the trouble of describing degrees of freedom?

Essentially, it provides a way to put two different prediction procedures on equal footing.

Degrees of freedom in statistics

Degrees of freedom in statistics is defined, intuitively, as the effective number of parameters used by a prediction procedure.

While this seems vague, it has a precise definition for a broad class of estimation problems. We will define this shortly.

Why is this an important concept? Why you would ever go to the trouble of describing degrees of freedom?

Essentially, it provides a way to put two different prediction procedures on equal footing.

Degrees of freedom in statistics

Degrees of freedom in statistics is defined, intuitively, as the effective number of parameters used by a prediction procedure.

While this seems vague, it has a precise definition for a broad class of estimation problems. We will define this shortly.

Why is this an important concept? Why you would ever go to the trouble of describing degrees of freedom?

Essentially, it provides a way to put two different prediction procedures on equal footing.

Degrees of freedom in statistics

Degrees of freedom in statistics is defined, intuitively, as the effective number of parameters used by a prediction procedure.

While this seems vague, it has a precise definition for a broad class of estimation problems. We will define this shortly.

Why is this an important concept? Why you would ever go to the trouble of describing degrees of freedom?

Essentially, it provides a way to put two different prediction procedures on equal footing.

Definition under squared error loss

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{dfF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\tilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error} =: \mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error} =: \mathrm{ErrT}(\widehat{f})} = 2\sigma^2 \mathrm{dfF}(\widehat{f})$$

Definition under squared error loss

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{dfF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\tilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error} =: \mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error} =: \mathrm{ErrT}(\widehat{f})} = 2\sigma^2 \mathrm{dfF}(\widehat{f})$$

Definition under squared error loss

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{dfF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\tilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error} =: \mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error} =: \mathrm{ErrT}(\widehat{f})} = 2\sigma^2 \mathrm{dfF}(\widehat{f})$$

Definition under squared error loss

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{dfF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\tilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error} =: \mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error} =: \mathrm{ErrT}(\widehat{f})} = 2\sigma^2 \mathrm{dfF}(\widehat{f})$$

Definition under squared error loss

Consider data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ such that $y_i = f(x_i) + \varepsilon_i$ where $f : \mathbb{R}^p \to \mathbb{R}$ is regression function, ε_i has mean 0 and variance σ^2.

Let $\mathcal{A}$ be any fitting algorithm that maps $\{(x_i, y_i)\}_{i=1}^n \overset{\mathcal{A}}{\mapsto} \widehat{f}$.
The degrees of freedom of predictor $\widehat{f}$ is defined as

$$\mathrm{dfF}(\widehat{f}) = \sum_{i=1}^{n} \mathrm{Cov}(y_i, \widehat{f}(x_i))/\sigma^2 = \mathrm{tr}\left[\mathrm{Cov}(y, \widehat{f}(X))\right]/\sigma^2,$$

where y: response vector, X: feature matrix, $\widehat{f}(X)$: predicted response

Where does squared error loss come into play?

$$\underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(\tilde{y}_i - \widehat{f}(x_i))^2\Big]}_{\text{fixed-X prediction error} =: \mathrm{ErrF}(\widehat{f})} - \underbrace{\mathbb{E}\Big[\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2\Big]}_{\text{expected training error} =: \mathrm{ErrT}(\widehat{f})} = 2\sigma^2 \mathrm{dfF}(\widehat{f})$$

Degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinarly least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{dfF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinarly least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{dfF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinarly least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{dfF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinarly least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{dfF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinarly least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{dfF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinarly least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{dfF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Degrees of freedom in linear regression

- Suppose $p \leq n$ and X has full (column) rank, and we take $\widehat{f}$ to be ordinarly least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \, \|y - X\beta\|_2^2 = (X^T X)^{-1} X^T y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X(X^T X)^{-1} X^T]/\sigma^2 = p.$$

- Suppose $p \geq n$ and X has full (row) rank, and we take $\widehat{f}$ to be min ℓ_2-norm least squares predictor $\widehat{f}(X) = X\widehat{\beta}$, where

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : X\beta = y\} = (X^\top X)^\dagger X^\top y.$$

$$\mathrm{dfF}(\widehat{f}) = \mathrm{tr}[\mathrm{Cov}(y, X\widehat{\beta})]/\sigma^2 = \mathrm{tr}[\sigma^2 X^\top (XX^\top)^{-1} X]/\sigma^2 = n.$$

Thus, $\mathrm{dfF}(\widehat{f})$ is p for $p \leq n$, but is always n for $p \geq n$ (not meaningful).

This is fixed-X degrees of freedom. How to extend for random-X setting?

Outline

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{dfF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{dfF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{dfF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{dfF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{dfF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{dfF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{dfF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{dfF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \tag{dfF}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{dfF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{dfF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Re-interpreting fixed-X degrees of freedom

Fixed-X degrees of freedom is a standard algorithm specific measure of complexity, but no notion of random-X degrees of freedom we know of.

We cast fixed-X degrees of freedom from a different perspective.

- Define fixed-X optimism of $\widehat{f}$ by $\mathrm{OptF}(\widehat{f}) = \mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$.
- Consider the following family of "reference" models:
  - $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
  - (U_k, v) is random design with k features, and noise with level σ^2.
- Recall that $\mathrm{dfF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) = k$ so long as $\mathrm{rank}(U_k) = k$.
- Thus, for a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{dfF}(\widehat{f})$ is also equal to the value of k that satisfy the following relation:

$$\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfF)}$$

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Emergent random-X degrees of freedom

"Matching optimism" interpretation can be extended to random-X setting and leads to the definition of random-X degrees of freedom.

- Define random-X optimism of $\widehat{f}$ by $\mathrm{OptR}(\widehat{f}) = \mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f})$, where $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2]$ is the random-X prediction error.
- We thus define the random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, emergent)}$$

Recall here:

- $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm,
- (U_k, v) is random design with k features, and noise with level σ^2.

We call the measure emergent random-X degrees of freedom.

Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by dfR^i, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to "pure noise".

Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by dfR^i, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to "pure noise".

Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by dfR^i, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to "pure noise".

Intrisic random-X degrees of freedom

- The emergent random-X degrees of freedom, $\mathrm{dfR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f.
- When matching optimisms, the observed random-X optimism of $\widehat{f}$ consists of bias, which may inflate the degrees of freedom.
- We thus also define intrinsic random-X degrees of freedom, denoted by dfR^i, as the k for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)) \qquad \text{(dfR, intrinsic)}$$

The intrinsic random-X degrees of freedom measures the inherent complexity of the predictor $\widehat{f}$ in terms of overfitting to "pure noise".

Computing random-X degrees of freedom

We have defining equations for random-X degrees of freedom, but how do we actually compute the solution in practice?

We need three numbers:

- $\mathrm{OptR}(\mathcal{A}(X, y))$: emergent random-X optimism
  Either supplied by user or via cross or held-out validation
- $\mathrm{OptR}(\mathcal{A}(X, v))$: intrinstic random-X optimism
  Simulate at noise level if known, otherwise average
- $\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(X, v))$: reference optimism
  Either simulate or use invariant asymptotic limit that holds under quite generic conditions on the random design matrix and noise:

$$\frac{\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v))}{\sigma^2} \to \frac{1-(1-\xi)^2}{1-\xi} \text{ as } n, p \to \infty \text{ and } p/n \to \xi \in (0, 1),$$

Computing random-X degrees of freedom

We have defining equations for random-X degrees of freedom, but how do we actually compute the solution in practice?

We need three numbers:

- $\mathrm{OptR}(\mathcal{A}(X, y))$: emergent random-X optimism
  Either supplied by user or via cross or held-out validation
- $\mathrm{OptR}(\mathcal{A}(X, v))$: intrinstic random-X optimism
  Simulate at noise level if known, otherwise average
- $\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(X, v))$: reference optimism
  Either simulate or use invariant asymptotic limit that holds under quite generic conditions on the random design matrix and noise:

$$\frac{\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v))}{\sigma^2} \to \frac{1-(1-\xi)^2}{1-\xi} \text{ as } n, p \to \infty \text{ and } p/n \to \xi \in (0, 1),$$

Computing random-X degrees of freedom

We have defining equations for random-X degrees of freedom, but how do we actually compute the solution in practice?

We need three numbers:

- $\mathrm{OptR}(\mathcal{A}(X, y))$: emergent random-X optimism
  Either supplied by user or via cross or held-out validation
- $\mathrm{OptR}(\mathcal{A}(X, v))$: intrinstic random-X optimism
  Simulate at noise level if known, otherwise average
- $\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(X, v))$: reference optimism
  Either simulate or use invariant asymptotic limit that holds under quite generic conditions on the random design matrix and noise:

$$\frac{\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v))}{\sigma^2} \to \frac{1-(1-\xi)^2}{1-\xi} \text{ as } n, p \to \infty \text{ and } p/n \to \xi \in (0, 1),$$

Computing random-X degrees of freedom

We have defining equations for random-X degrees of freedom, but how do we actually compute the solution in practice?

We need three numbers:

- $\mathrm{OptR}(\mathcal{A}(X, y))$: emergent random-X optimism
  Either supplied by user or via cross or held-out validation
- $\mathrm{OptR}(\mathcal{A}(X, v))$: intrinstic random-X optimism
  Simulate at noise level if known, otherwise average
- $\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(X, v))$: reference optimism
  Either simulate or use invariant asymptotic limit that holds under quite generic conditions on the random design matrix and noise:

$$\frac{\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v))}{\sigma^2} \to \frac{1-(1-\xi)^2}{1-\xi} \text{ as } n, p \to \infty \text{ and } p/n \to \xi \in (0, 1),$$

Computing random-X degrees of freedom

We have defining equations for random-X degrees of freedom, but how do we actually compute the solution in practice?

We need three numbers:

- $\mathrm{OptR}(\mathcal{A}(X, y))$: emergent random-X optimism
  Either supplied by user or via cross or held-out validation
- $\mathrm{OptR}(\mathcal{A}(X, v))$: intrinstic random-X optimism
  Simulate at noise level if known, otherwise average
- $\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(X, v))$: reference optimism
  Either simulate or use invariant asymptotic limit that holds under quite generic conditions on the random design matrix and noise:

$$\frac{\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v))}{\sigma^2} \to \frac{1-(1-\xi)^2}{1-\xi} \text{ as } n, p \to \infty \text{ and } p/n \to \xi \in (0, 1),$$

Computing random-X degrees of freedom

Under the asymptotic limit, solving for the random-X degrees of freedom by matching optimisms leads:

$$\mathrm{dfR}(\widehat{f})/n \to 1 + \frac{\psi}{2} - \sqrt{1 + \frac{\psi^2}{4}}.$$

where ψ is the normalized random-X optimism of the given predictor.

Remarks:

- There is a unique number in $[0, n]$ satisfying the desired relations.
- This is an interpretable range for random-X degrees of freedom:
  - The least complex predictor has dfR of 0,
  - The most complex predictor has dfR of n, as if saturated model.

Computing random-X degrees of freedom

Under the asymptotic limit, solving for the random-X degrees of freedom by matching optimisms leads:

$$\mathrm{dfR}(\widehat{f})/n \to 1 + \frac{\psi}{2} - \sqrt{1 + \frac{\psi^2}{4}}.$$

where ψ is the normalized random-X optimism of the given predictor.

Remarks:

- There is a unique number in $[0, n]$ satisfying the desired relations.
- This is an interpretable range for random-X degrees of freedom:
  - The least complex predictor has dfR of 0,
  - The most complex predictor has dfR of n, as if saturated model.

Computing random-X degrees of freedom

Under the asymptotic limit, solving for the random-X degrees of freedom by matching optimisms leads:

$$\mathrm{dfR}(\widehat{f})/n \to 1 + \frac{\psi}{2} - \sqrt{1 + \frac{\psi^2}{4}}.$$

where ψ is the normalized random-X optimism of the given predictor.

Remarks:

- There is a unique number in $[0, n]$ satisfying the desired relations.
- This is an interpretable range for random-X degrees of freedom:
  - The least complex predictor has dfR of 0,
  - The most complex predictor has dfR of n, as if saturated model.

Computing random-X degrees of freedom

Under the asymptotic limit, solving for the random-X degrees of freedom by matching optimisms leads:

$$\mathrm{dfR}(\widehat{f})/n \to 1 + \frac{\psi}{2} - \sqrt{1 + \frac{\psi^2}{4}}.$$

where ψ is the normalized random-X optimism of the given predictor.

Remarks:

- There is a unique number in $[0, n]$ satisfying the desired relations.
- This is an interpretable range for random-X degrees of freedom:
  - The least complex predictor has dfR of 0,
  - The most complex predictor has dfR of n, as if saturated model.

Outline

Ordinary, min ℓ_2-norm and min ℓ_1-norm least squares

Given (X, y), consider predictors of the form $\widehat{f}(x) = x^\top \widehat{\beta}$, where

- Ordinary least squares

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2$$

- Min ℓ_2-norm least squares (mn2ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : y = X\beta\}$$

- Min ℓ_1-norm least squares (mn1ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_1 : y = X\beta\}$$

Ordinary, min ℓ_2-norm and min ℓ_1-norm least squares

Given (X, y), consider predictors of the form $\widehat{f}(x) = x^\top \widehat{\beta}$, where

- Ordinary least squares

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2$$

- Min ℓ_2-norm least squares (mn2ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : y = X\beta\}$$

- Min ℓ_1-norm least squares (mn1ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_1 : y = X\beta\}$$

Ordinary, min ℓ_2-norm and min ℓ_1-norm least squares

Given (X, y), consider predictors of the form $\widehat{f}(x) = x^\top \widehat{\beta}$, where

- Ordinary least squares

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2$$

- Min ℓ_2-norm least squares (mn2ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : y = X\beta\}$$

- Min ℓ_1-norm least squares (mn1ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_1 : y = X\beta\}$$

Ordinary, min ℓ_2-norm and min ℓ_1-norm least squares

Given (X, y), consider predictors of the form $\widehat{f}(x) = x^\top \widehat{\beta}$, where

- Ordinary least squares

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2$$

- Min ℓ_2-norm least squares (mn2ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : y = X\beta\}$$

- Min ℓ_1-norm least squares (mn1ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_1 : y = X\beta\}$$

Ordinary, min ℓ_2-norm and min ℓ_1-norm least squares

Given (X, y), consider predictors of the form $\widehat{f}(x) = x^\top \widehat{\beta}$, where

- Ordinary least squares

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|y - X\beta\|_2^2$$

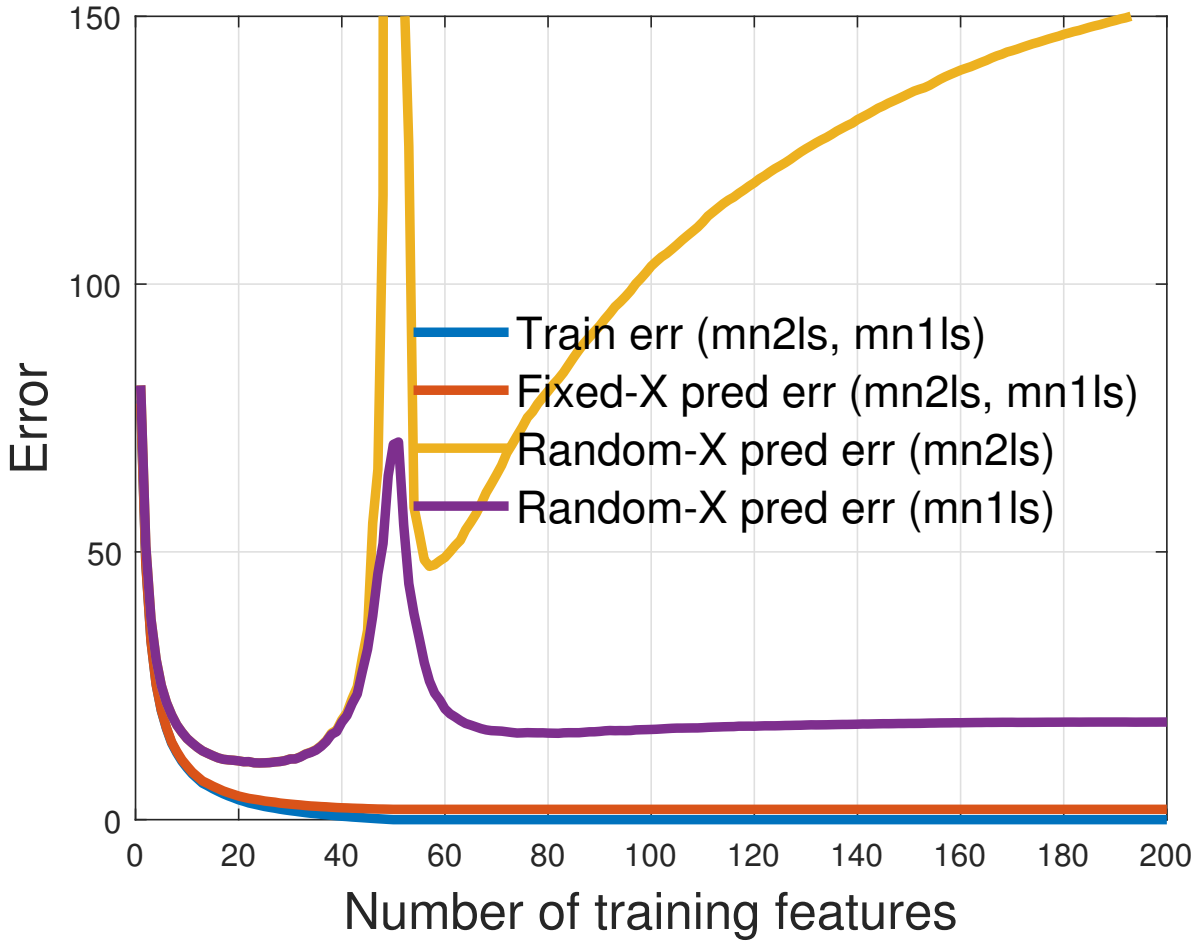
- Min ℓ_2-norm least squares (mn2ls)

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_2 : y = X\beta\}$$

- Min ℓ_1-norm least squares (mn1ls)

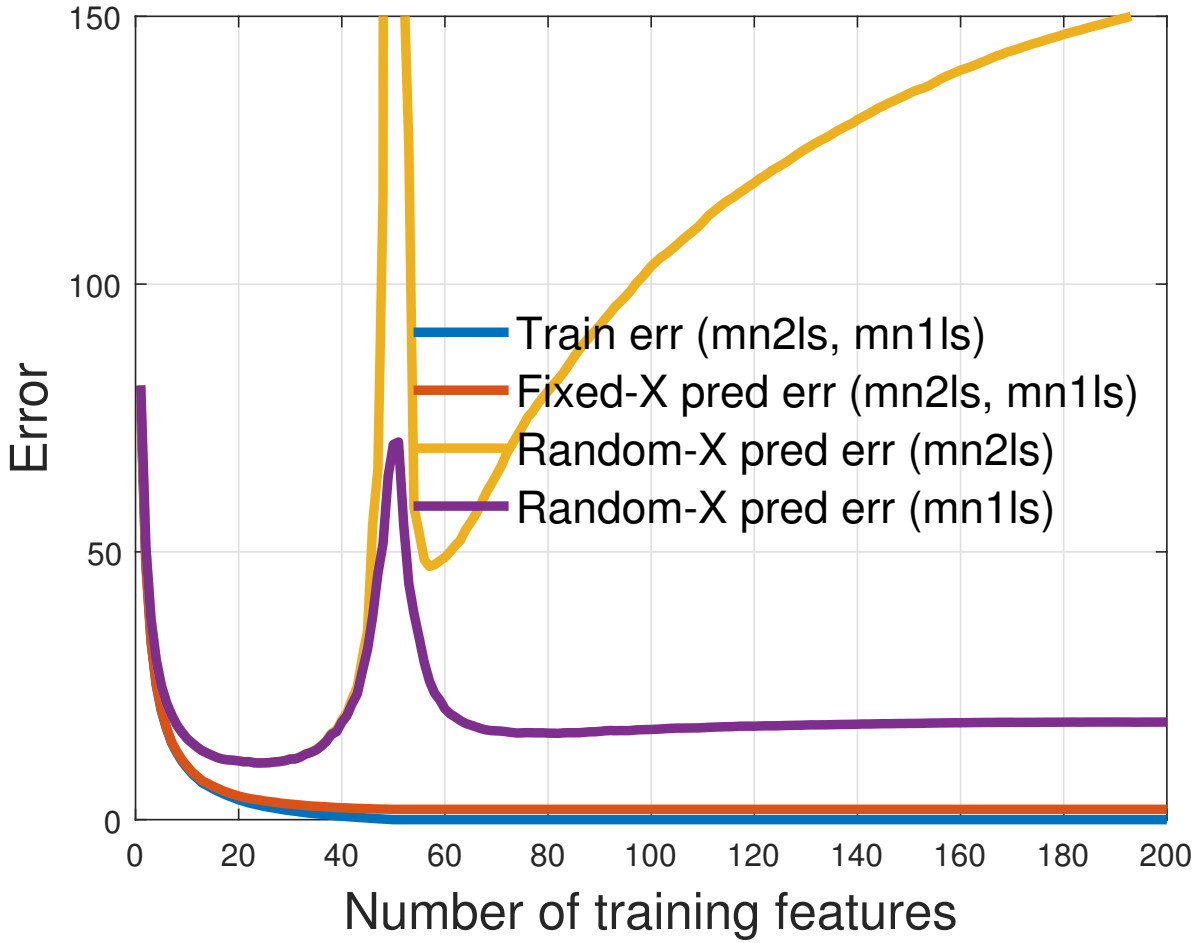
$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{\|\beta\|_1 : y = X\beta\}$$

Errors (train, fixed-X, random-X) versus model size



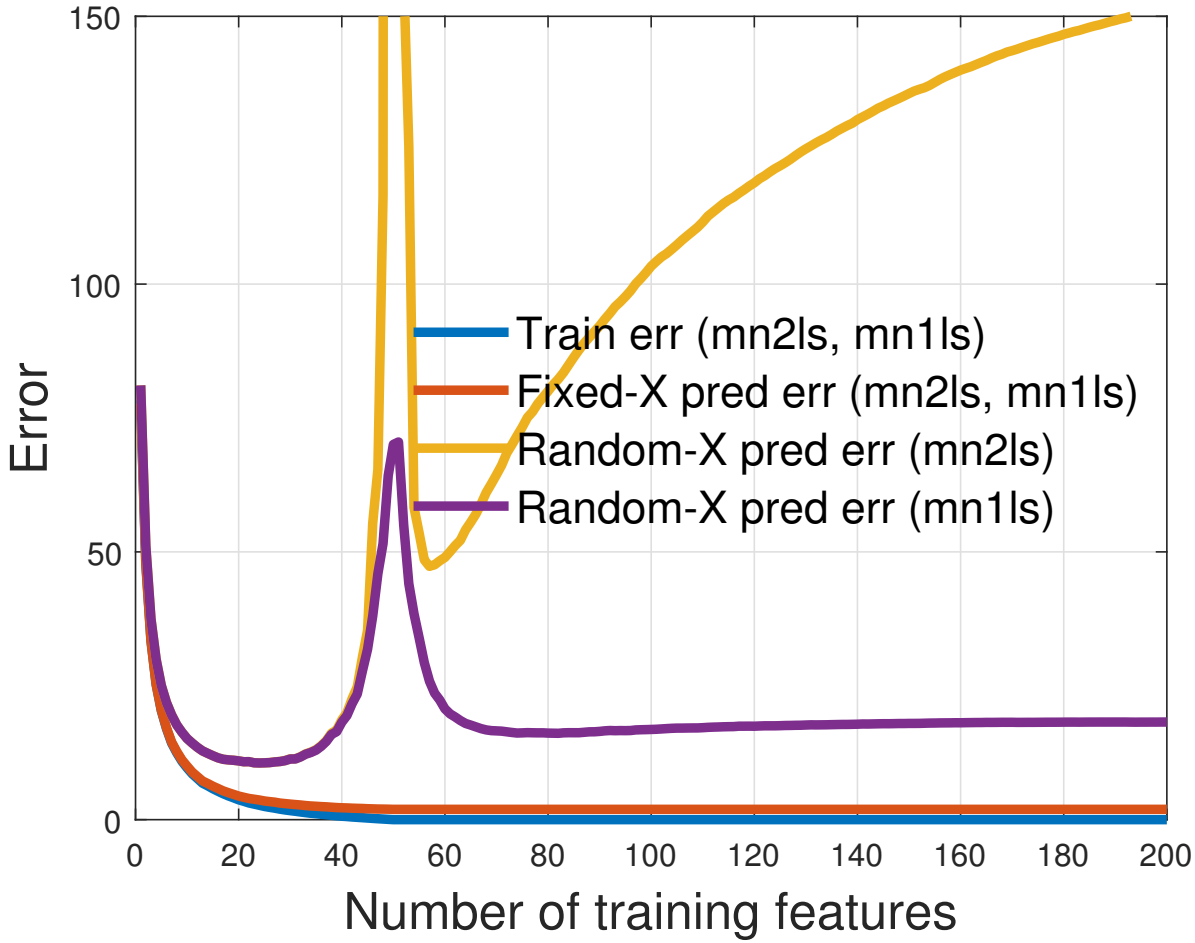
- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 features
- Train, fixed-X: descent then constant; random-X: double descent

Errors (train, fixed-X, random-X) versus model size



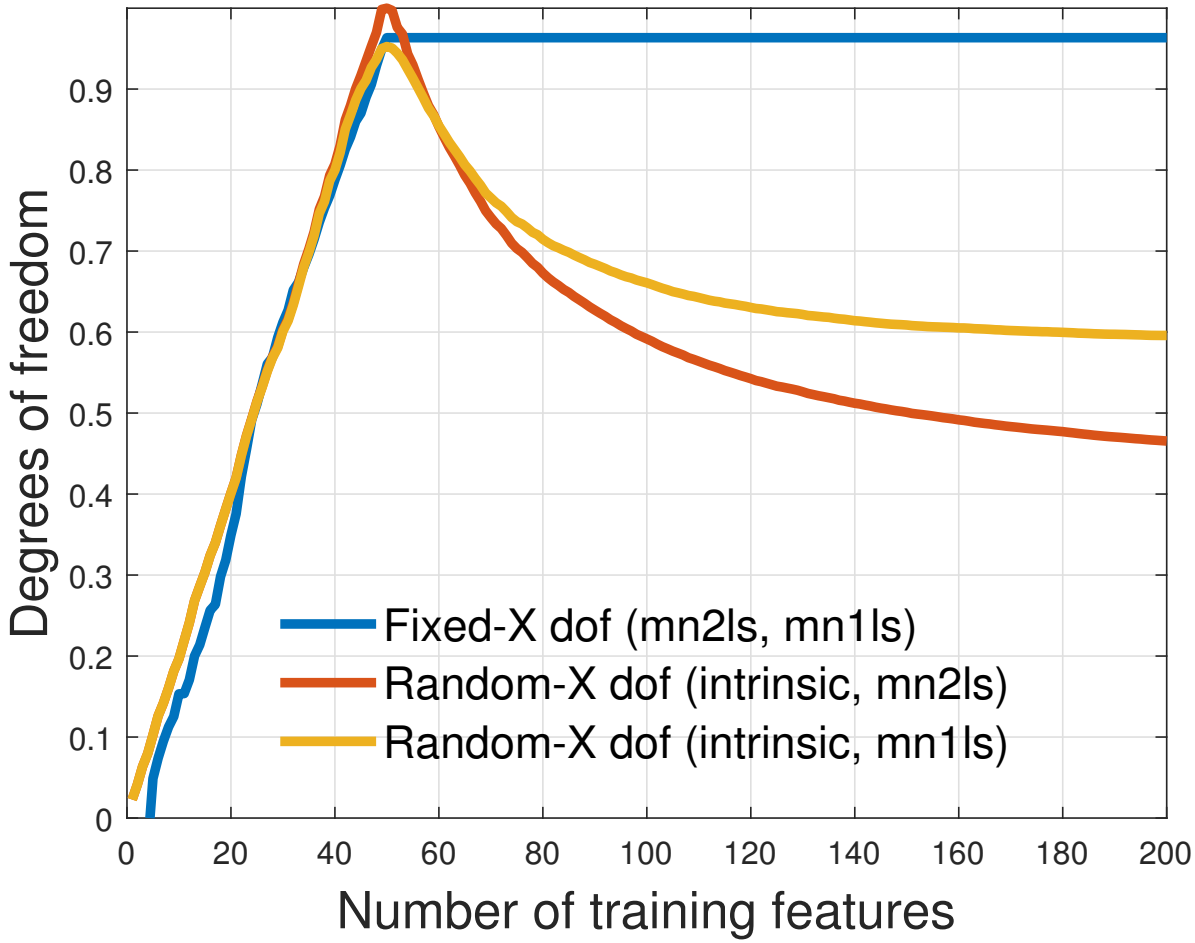
- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 features
- Train, fixed-X: descent then constant; random-X: double descent

Errors (train, fixed-X, random-X) versus model size



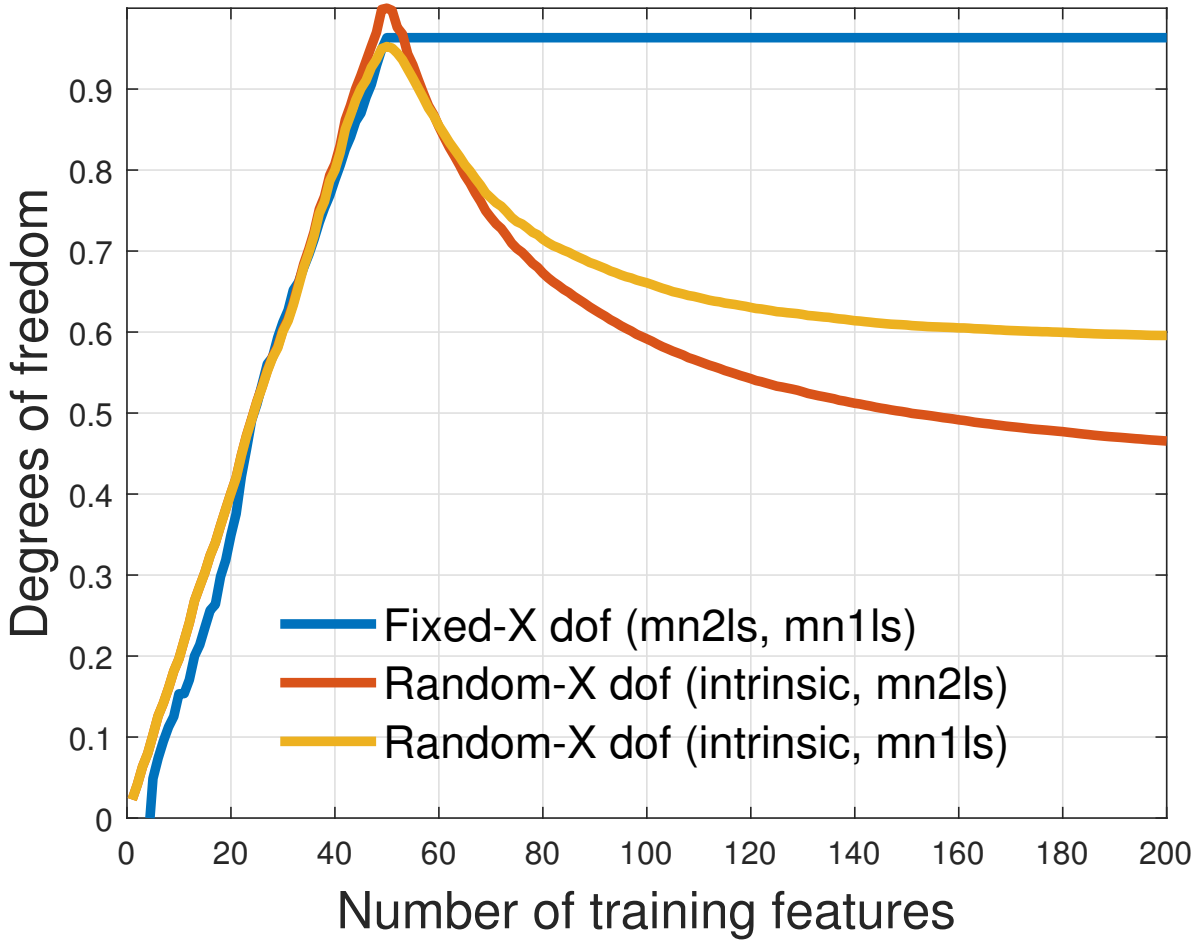
- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 features
- Train, fixed-X: descent then constant; random-X: double descent

Degrees of freedom (fixed-X, random-X) versus model size



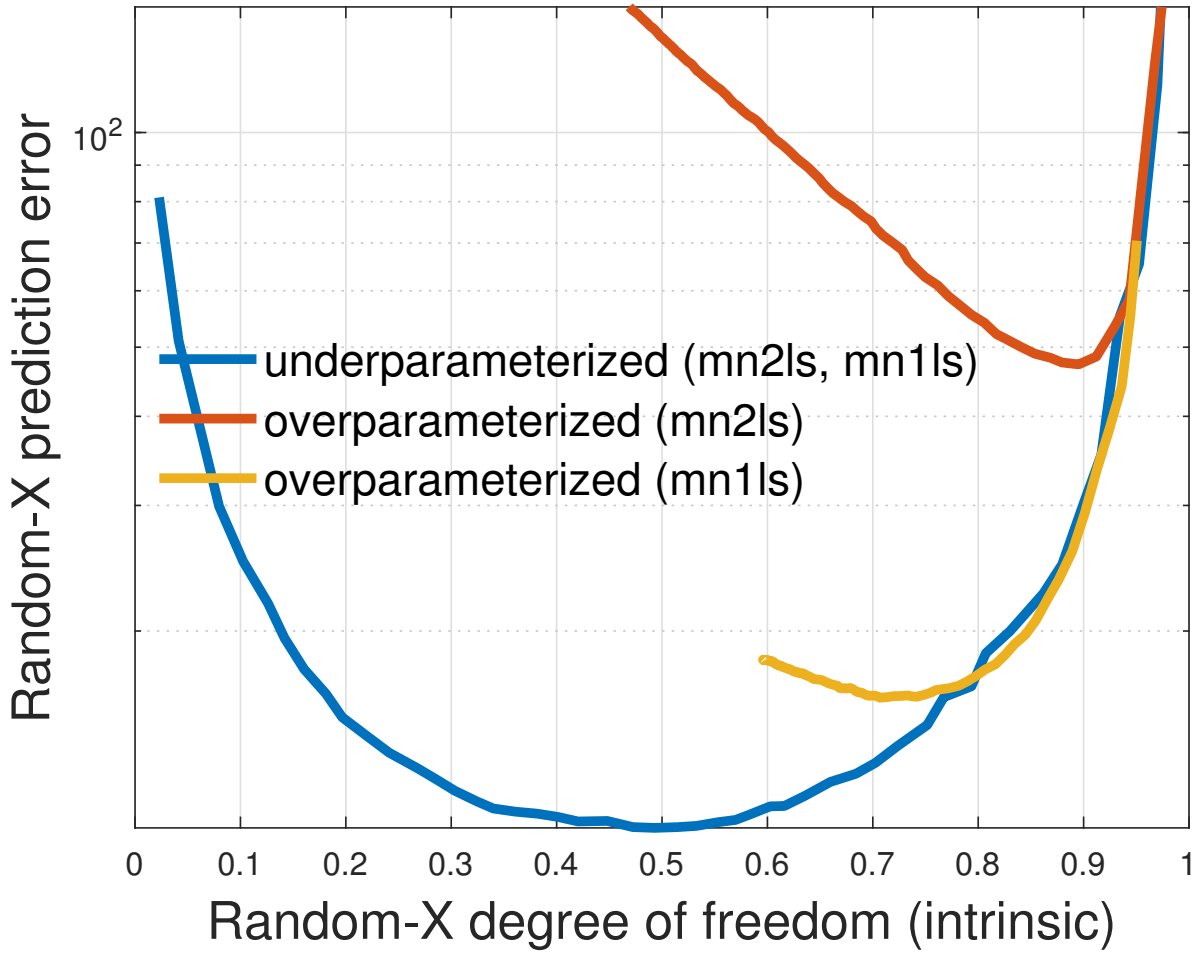
- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 feaures
- Fixed-X: increase then constant; random-X: increase then decrease

Degrees of freedom (fixed-X, random-X) versus model size



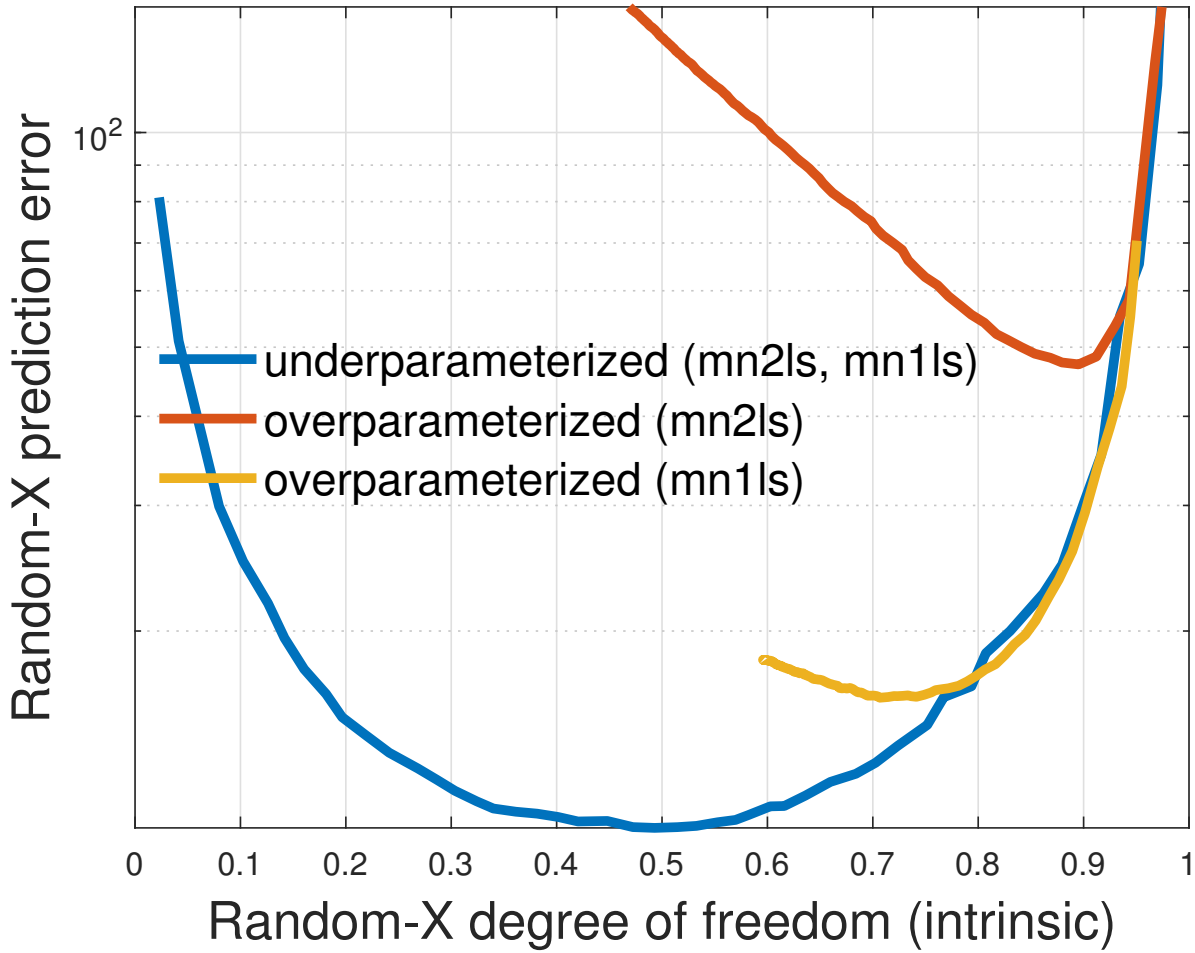
- Fixed data with $n = 50$ and response non-linear in $p = 200$ features
- Model class: estimators fitted on nested subsets of 1 to 200 feaures
- Fixed-X: increase then constant; random-X: increase then decrease

Prediction error (random-X) versus model size



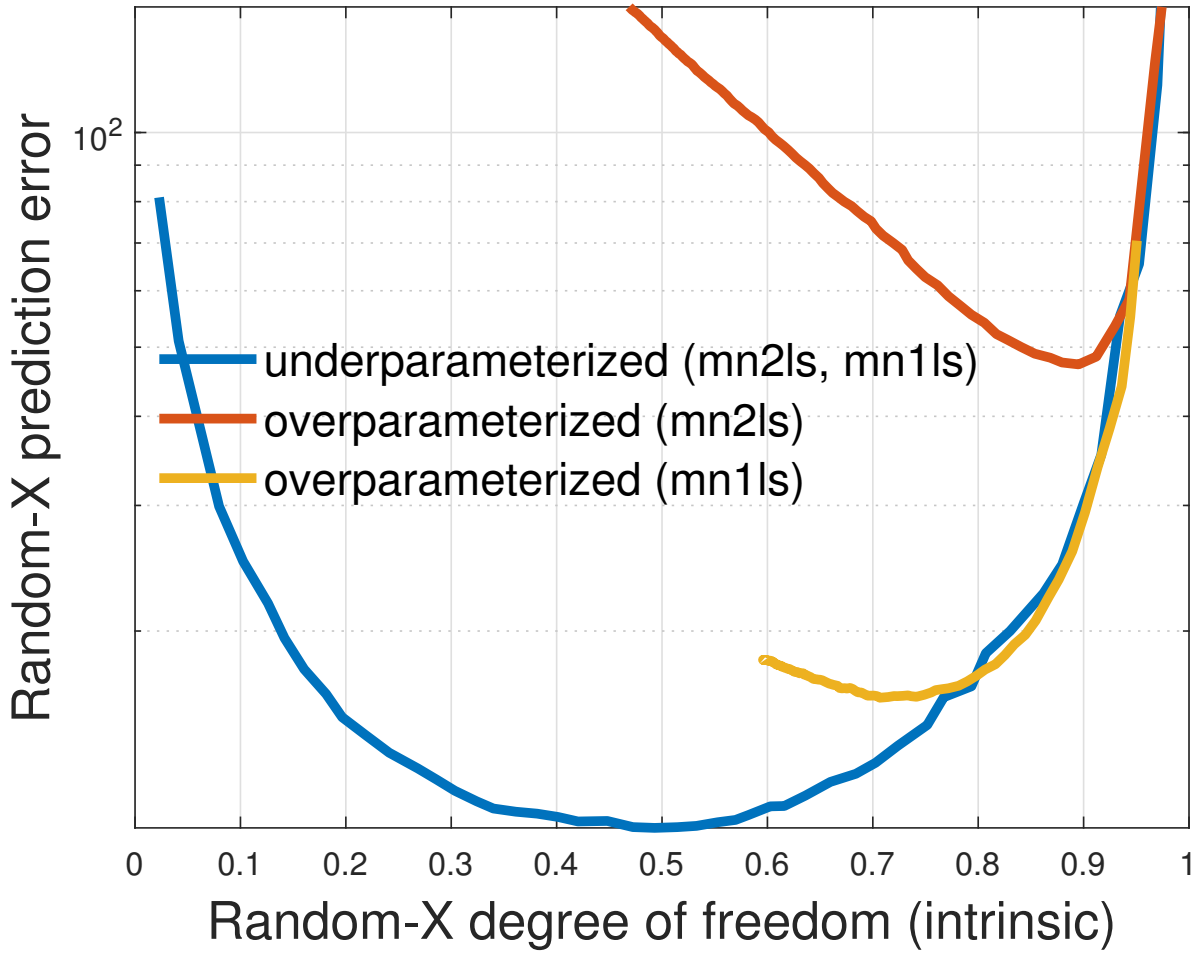
- Same setup as before
- Underparameterized: *U*-curve; overparameterized: also *U*-curve!
- Punchline: map from overparameterized to underparameterized

Prediction error (random-X) versus model size



- Same setup as before
- Underparameterized: *U*-curve; overparameterized: also *U*-curve!
- Punchline: map from overparameterized to underparameterized

Prediction error (random-X) versus model size



- Same setup as before
- Underparameterized: *U*-curve; overparameterized: also *U*-curve!
- Punchline: map from overparameterized to underparameterized

Outline

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond . . .

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Discussion and future directions

A high-level view of the work:

- Suppose we are given a family of models for which we want a complexity measure under a specific error metric.
- Construct a family of "reference" models spanning same optimisms.
- Find the model in the reference family that is closest to the observed optimism. Declare complexity as complexity of that reference model.

Key relation:

$$\mathrm{OptR}(\widehat{f}) = \mathrm{OptR}(\widehat{f}^{\mathrm{ref}})$$

Going beyond ...

- Attribute total complexity to various components: bias, variance, covariate shift, etc.
- Other error metrics beyond squared error
- Unsupervised setting?

Thanks for listening!

Questions/comments/thoughts?