(More on) the Notion of Double Descent

Norm Matloff
University of California at Davis

Bay Area R Users Group
December 15, 2020

URL for these slides (repeated on final slide):
`http://heather.cs.ucdavis.edu/BARUGdouble.pdf`

Norm Matloff University of California at Davis

Outlne of Talk

- Motivating issue: overfitting is "bad," but somehow it often seems to work in neural networks (NNs). Why?
- Possible answer: double descent curve.
- Related R software.
- Empirical illustration.
- So, was the motivating question answered?

Norm Matloff University of California at Davis

Motivation

- Neural networks (NNs) as "black boxes."
- $p >> n$, drastically overfit. NNs shouldn't work well.
- Why do neural networks work well (when they do), in spite of overfitting?

Norm Matloff University of California at Davis

Classical View

- Bias-variance tradeoff.
- Graph of mean loss should be U-shaped.
- As p first moves away from 0, bias decreases a lot, while variance increases a little; curve goes downward.
- Later variance overtakes bias; curve goes upward.
- Eventually (e.g. $p = n - 1$ for linear model), "perfect" fit of the training data — *interpolation* — but terrible at predicting new cases.
- Setting p past the interpolation point, or even near it, is considered overfitting.
- "Sweet spot" at bottom of the U.

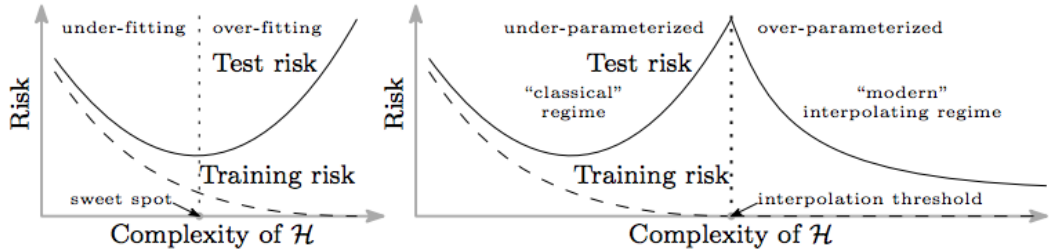
Norm Matloff University of California at Davis

A Bizarre Discovery

Norm Matloff University of California at Davis

A Bizarre Discovery

- Belkin *et al* discovered there is often a second U!



(a) U-shaped "bias-variance" risk curve (b) "double descent" risk curve

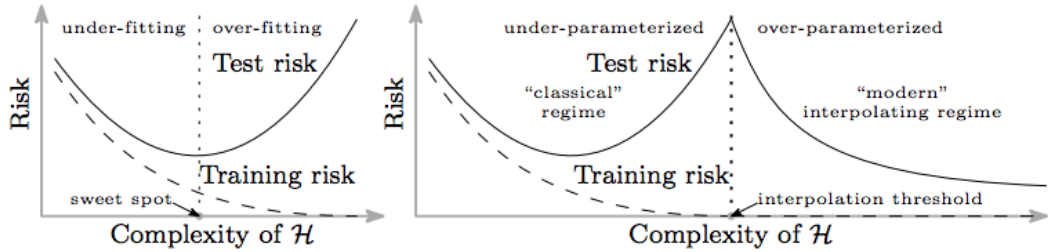
Figure 1: Curves for training risk (dashed line) and test risk (solid line). (a) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (b) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high complexity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

- Comes right after the interpolation point.

Norm Matloff University of California at Davis

A Bizarre Discovery

- Belkin *et al* discovered there is often a second U!



(a) U-shaped "bias-variance" risk curve (b) "double descent" risk curve

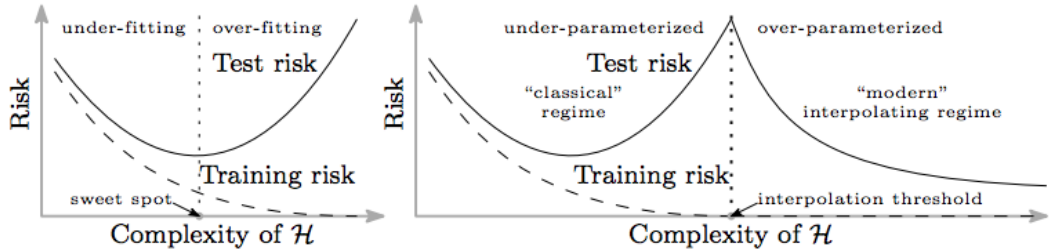
Figure 1: Curves for training risk (dashed line) and test risk (solid line). (a) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (b) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high complexity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

- Comes right after the interpolation point.
- "Yay, now we know why NNs work!"
- "Long live overfitting!"

Norm Matloff University of California at Davis

A Bizarre Discovery

- Belkin *et al* discovered there is often a second U!



(a) U-shaped "bias-variance" risk curve (b) "double descent" risk curve

Figure 1: Curves for training risk (dashed line) and test risk (solid line). (a) The classical *U-shaped risk curve* arising from the bias-variance trade-off. (b) The *double descent risk curve*, which incorporates the U-shaped risk curve (i.e., the "classical" regime) together with the observed behavior from using high complexity function classes (i.e., the "modern" interpolating regime), separated by the interpolation threshold. The predictors to the right of the interpolation threshold have zero training risk.

- Comes right after the interpolation point.
- "Yay, now we know why NNs work!"
- "Long live overfitting!"
- Well, not so fast. Let's look a little closer.

Norm Matloff University of California at Davis

Why a Double U?

Norm Matloff University of California at Davis

Why a Double U?

- First, why might a double U occur? Say, for a linear model.

Norm Matloff University of California at Davis

Why a Double U?

- First, why might a double U occur? Say, for a linear model.
- Prior to interpolation point $p = n - 1$, fit is unique, and there is no bias.

Norm Matloff University of California at Davis

Why a Double U?

- First, why might a double U occur? Say, for a linear model.
- Prior to interpolation point $p = n - 1$, fit is unique, and there is no bias.
- For any $p \geq n$, infinitely many solutions — all biased, but some have small variance. Thus mean risk may decrease for a while as p continues to increase.

Norm Matloff University of California at Davis

Minimum-Norm Solution

Norm Matloff University of California at Davis

Minimum-Norm Solution

- Of the infinitely-many solutions for any particular $p \geq n$, we might choose the one with smallest l_2 norm.

Norm Matloff University of California at Davis

Minimum-Norm Solution

- Of the infinitely-many solutions for any particular $p \geq n$, we might choose the one with smallest l_2 norm. (Hope small norm $\Rightarrow$ small variance.)

Norm Matloff University of California at Davis

Minimum-Norm Solution

- Of the infinitely-many solutions for any particular $p \geq n$, we might choose the one with smallest l_2 norm. (Hope small norm $\Rightarrow$ small variance.)
- Gradient-based approches tend to produce minimum-norm solutions. (Inferred from the form of the iterative updating equation.)

Norm Matloff University of California at Davis

Minimum-Norm Solution

- Of the infinitely-many solutions for any particular $p \geq n$, we might choose the one with smallest l_2 norm. (Hope small norm $\Rightarrow$ small variance.)
- Gradient-based approches tend to produce minimum-norm solutions. (Inferred from the form of the iterative updating equation.)
- Since NNs use SGD, this suggests a possible reason why NNs often work well in spite of overfitting.

Norm Matloff University of California at Davis

Findings of Hastie *et al*

Findings of Hastie *et al*

Theoretical paper by Hastie, Montanari, Rosset and Tibshirani (2019):

- Linear model, asymptotic analysis, minimum-norm.
- Depends on SNR. Second U has a minimum if SNR > 1. Etc.
- Optimal ridge regression beats min-norm; leave-1-out CV yields optimal.

Norm Matloff University of California at Davis

Related R Functions

Norm Matloff University of California at Davis

Related R Functions

- Do try this at home!
- **MASS::ginv()**
  Moore-Penrose inverse, $X \to X^+$.
  Min-norm solution to least-square problem.
  $\widehat{\beta} = X^+ Y$
  Reduces to the usual $(X'X)^{-1}Y$ if $p < n$.
- **glmnet::cv.glmnet()**
  Finds optimal ridge (with **alpha = 0**).
- Optional conveniences for the polynomial illustration below:
  regtools::qePoly(), **regtools::ridgePoly()**

Norm Matloff University of California at Davis

Empirical Illustration

Norm Matloff University of California at Davis

Empirical Illustration

- Million Song dataset, UCI; 90 features (audio measurements), 500K songs.
- Use the first p features, quadratic model; vary p.

Norm Matloff University of California at Davis

Empirical Illustration

- Million Song dataset, UCI; 90 features (audio measurements), 500K songs.
- Use the first p features, quadratic model; vary p.
- 10 replications, take mean over the 10.
- Predictors not interchangeable, so a better investigation would be to randomly permute the features in each rep.

Norm Matloff University of California at Davis

Empirical Illustration

- Million Song dataset, UCI; 90 features (audio measurements), 500K songs.
- Use the first p features, quadratic model; vary p.
- 10 replications, take mean over the 10.
- Predictors not interchangeable, so a better investigation would be to randomly permute the features in each rep.
- Training set, random n rows of the dataset; vary n.

Norm Matloff University of California at Davis

Empirical Illustration

- Million Song dataset, UCI; 90 features (audio measurements), 500K songs.
- Use the first p features, quadratic model; vary p.
- 10 replications, take mean over the 10.
- Predictors not interchangeable, so a better investigation would be to randomly permute the features in each rep.
- Training set, random n rows of the dataset; vary n.
- Find Mean Absolute Prediction Error.

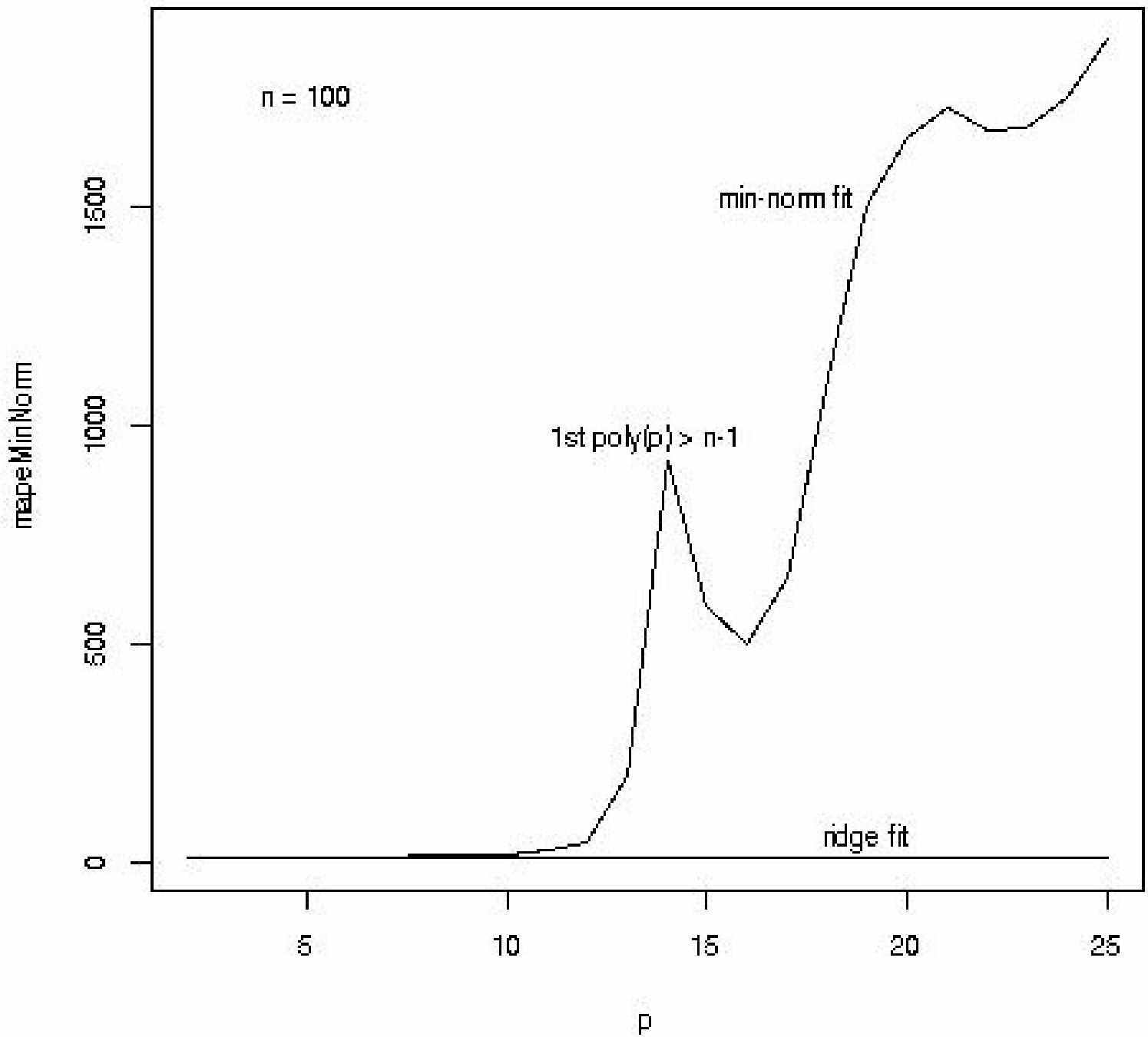
Empirical Illustration

- Million Song dataset, UCI; 90 features (audio measurements), 500K songs.
- Use the first p features, quadratic model; vary p.
- 10 replications, take mean over the 10.
- Predictors not interchangeable, so a better investigation would be to randomly permute the features in each rep.
- Training set, random n rows of the dataset; vary n.
- Find Mean Absolute Prediction Error.
- Interested in:
  - Does the double-U show up?
  - Does ridge do better?

Norm Matloff University of California at Davis

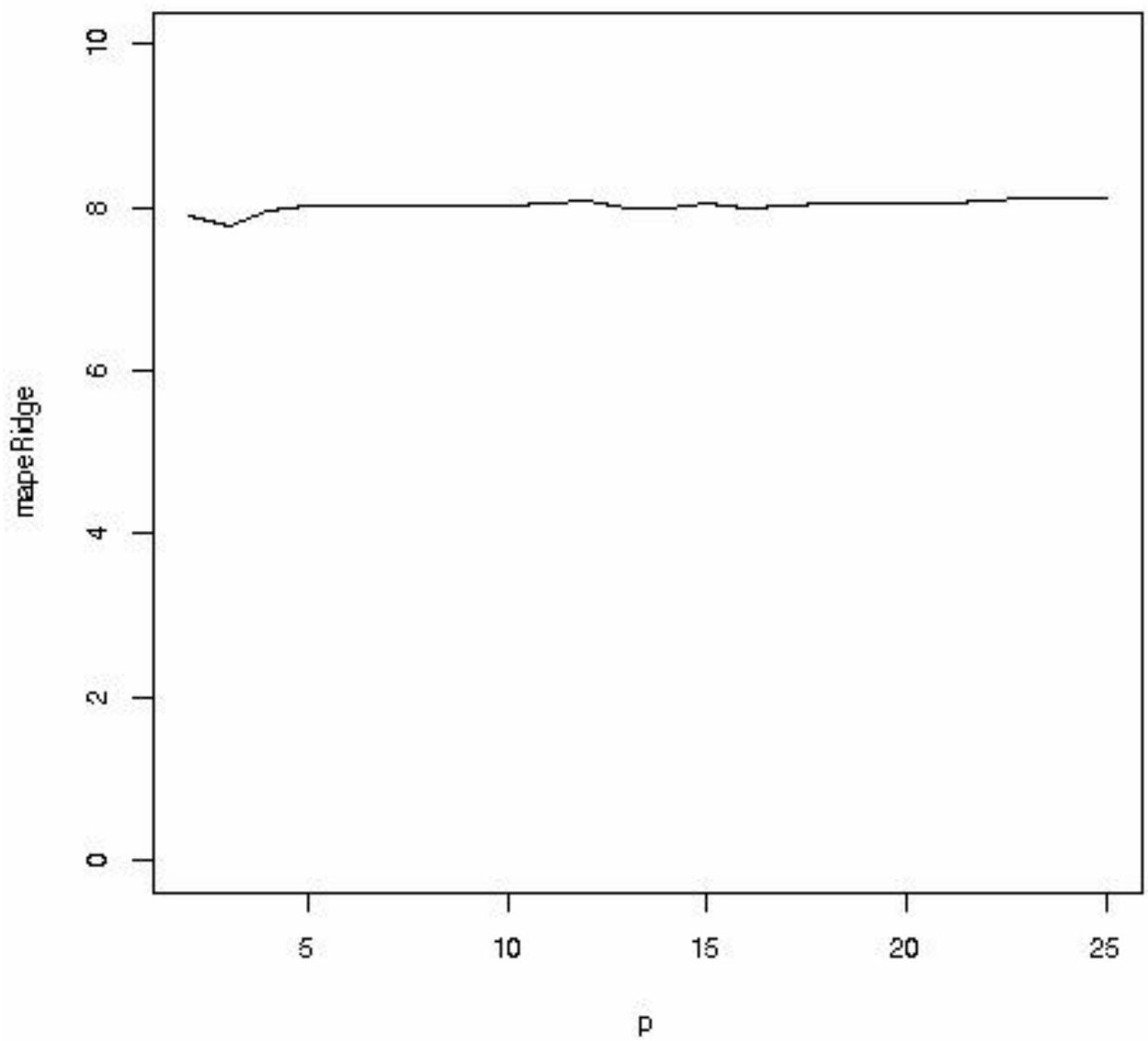
Results

Norm Matloff
University of California at Davis



Norm Matloff
University of California at Davis

Wait, *what* was that ridge graph again?



Norm Matloff University of California at Davis

Notes on the Graphs

Norm Matloff University of California at Davis

Notes on the Graphs

- We do see double-descent. (Even triple?)

Norm Matloff University of California at Davis

Notes on the Graphs

- We do see double-descent. (Even triple?)
- As predicted, ridge beat minimum-norm —REALLY beat it.

Norm Matloff University of California at Davis

Notes on the Graphs

- We do see double-descent. (Even triple?)
- As predicted, ridge beat minimum-norm —REALLY beat it.
- Ridge exhibited a classical U shape, though the U is surprisingly shallow. Maybe due to:

Norm Matloff University of California at Davis

Notes on the Graphs

- We do see double-descent. (Even triple?)
- As predicted, ridge beat minimum-norm —REALLY beat it.
- Ridge exhibited a classical U shape, though the U is surprisingly shallow. Maybe due to:
- Prediction power somewhat weak; best MAPE was 7.8, $l_1 = 8.2$.

Norm Matloff University of California at Davis

Notes on the Graphs

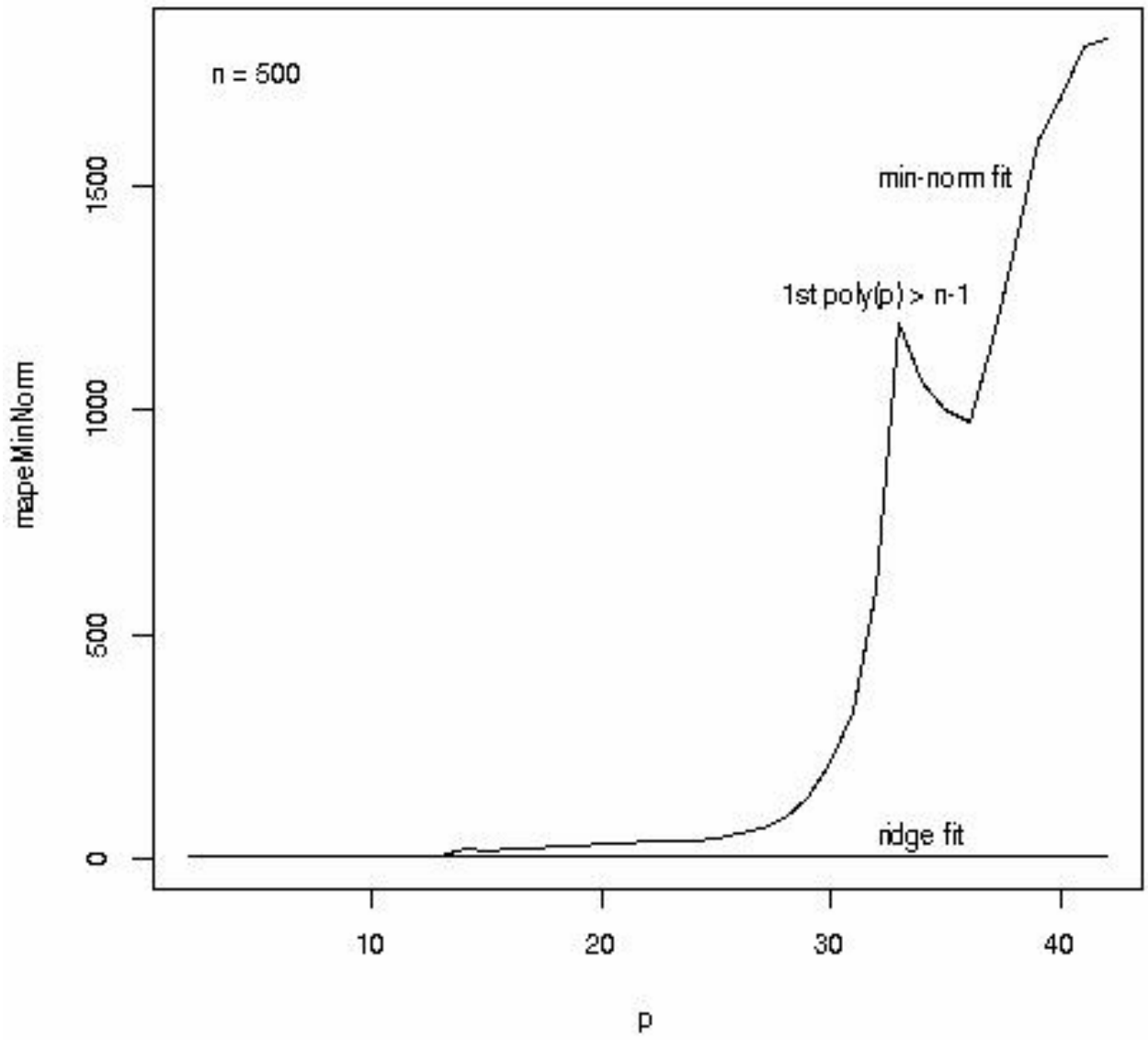
- We do see double-descent. (Even triple?)
- As predicted, ridge beat minimum-norm —REALLY beat it.
- Ridge exhibited a classical U shape, though the U is surprisingly shallow. Maybe due to:
- Prediction power somewhat weak; best MAPE was 7.8, $l_1 = 8.2$.
- No double descent for ridge, and there shouldn't be. (See intuition on why double-U for min-norm.)

Norm Matloff University of California at Davis

$$n = 500$$

Norm Matloff University of California at Davis

n = 500



Norm Matloff University of California at Davis

n = 500, cont'd.

- Same pattern.
- Interpolation now at $p = 32$.
- Larger sample size; best MAPE now down to 7.4.

Norm Matloff University of California at Davis

Anything to See Here?

- So, does double-descent, min-norm explain why NNs often well in spite of overfitting?
- Our work (Cheng *et al*, 2018) argues that NNs essentially perform polynomial regression.
- But (Hastie *et al*) say min-norm linear models (that includes poly regression) don't do as well as ridge.
- Maybe this new look into the NN black box isn't so insightful after all.

Norm Matloff University of California at Davis

So Why Do NNs Get Away with Overfitting?

Norm Matloff University of California at Davis

So Why Do NNs Get Away with Overfitting?

- So, if min-norm and double descent don't explain the (sometime) success of NNs in spite of overfitting, what does?
- Probably regularization. But where, how? My take on it:
  - Min-norm is regularization, but that explanation seems not to work. (Ridge regularization far better.)
  - Dropout.
  - Data runs a gauntlet of ReLUs, with many/most paths stopped short. This in effect reduces dimension.
  - Results from next-to-last layer are averaged to produce a final value. Averaging is a form of regularization.
  - Or...maybe they are not overfitting after all?

Norm Matloff University of California at Davis

Are NNs Indeed Overfitting?

Norm Matloff University of California at Davis

Are NNs Indeed Overfitting?

- From Krizhevsky *et al* (2012) (AlexNet paper):
- 60 million weights.
- Data augmentation factor 2048.
- Dropout factor 0.5.
- E.g. MNIST: n expanded from 65K to 130M. So 60M weights is technically not overfitting. After dropout, 30M.
- Granted, true n is not 130M, but all this suggests that they are not overfitting after all.

Norm Matloff University of California at Davis

Access These Slides

URL for these slides:

`http://heather.cs.ucdavis.edu/BARUGdouble.pdf`