Ridgeless Regression in Overparameterized Linear Models

Tyler Schmidt

University of Iowa

12/8/2025

Introduction

- High-dimensional data is increasingly common in genomics, climate modeling, artificial intelligence, finance, and many other fields.
- When $p \geq n$ or the predictors are multicollinear, the matrix $X^\top X$ becomes singular. Thus the usual OLS closed-form solution

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

  is not defined.
- This motivates the need for regularization techniques such as Ridge, Lasso, and Ridgeless (minimum-norm) regression.

Ridge and Lasso — quick comparison

- The ℓ_p norm of a vector x is defined as $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$.
- **Ridge** (ℓ_2 penalty)

$$\hat{\beta}_\lambda = \arg\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad \Rightarrow \quad \hat{\beta}_\lambda = (X^\top X + \lambda I)^{-1} X^\top y$$

  - Stabilizes ill-conditioned $X^\top X$ by shrinking coefficients.
  - $\lambda > 0$ prevents blow-up of directions with small singular values.

- **Lasso** (ℓ_1 penalty)

$$\hat{\beta}_\lambda = \arg\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

  - Encourages sparsity, but can only select a maximum of n variables (James et al., 2023).

- OLS (and lasso/ridge when $\lambda = 0$) fails when $p \geq n$. This leads us to generalized inverses.

Generalized Inverses

- A **generalized inverse** G of a matrix A satisfies:

$$AGA = A.$$

- Key facts:
  - The generalized inverse is not unique.
  - Different generalized inverses correspond to different solutions of $X^\top X\hat{\beta} = X^\top y$.
  - Among all solutions, the minimum-norm solution is often preferred.

Moore–Penrose Pseudoinverse

- The **Moore–Penrose pseudoinverse** A^+ is the unique matrix satisfying the four Penrose conditions:

$$AA^+A = A, \qquad A^+AA^+ = A^+,$$

$$(AA^+)^\top = AA^+, \qquad (A^+A)^\top = A^+A.$$

- Important properties:
  - Exists for every matrix (square or rectangular).
  - Provides the minimum-norm solution to $Ax = b$:

$$x^* = A^+b = \arg\min\{\|x\|_2 : Ax = b\}.$$

  - If A is full column rank: $A^+ = (A^TA)^{-1}A^T$.
  - If A is full row rank: $A^+ = A^T(AA^T)^{-1}$ (MacAusland, n.d.).

Computing the Pseudoinverse via SVD

- Any matrix $A \in \mathbb{R}^{n \times p}$ has a singular value decomposition

$$A = U \Sigma V^\top,$$

where:

  - U contains orthonormal left singular vectors,
  - V contains orthonormal right singular vectors,
  - $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

- Singular values measure how much X stretches vectors. Formally, the nonzero singular values of X are the square roots of the nonzero eigenvalues of both $X^\top X$ and $X X^\top$.
- The pseudoinverse is a p x n matrix constructed by inverting the nonzero singular values:

$$A^+ = V \Sigma^+ U^\top, \qquad \Sigma^+ = \mathrm{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_r} \right).$$

- If a singular value $\sigma_i = 0$, the corresponding direction lies in the null space of A and the pseudoinverse sets its coefficient to 0.

Ridgeless Regression

- Ridgeless regression is the minimum-norm least squares solution for $p \geq n$:

$$\hat{\beta}_{\text{ridgeless}} = X^{+} y$$

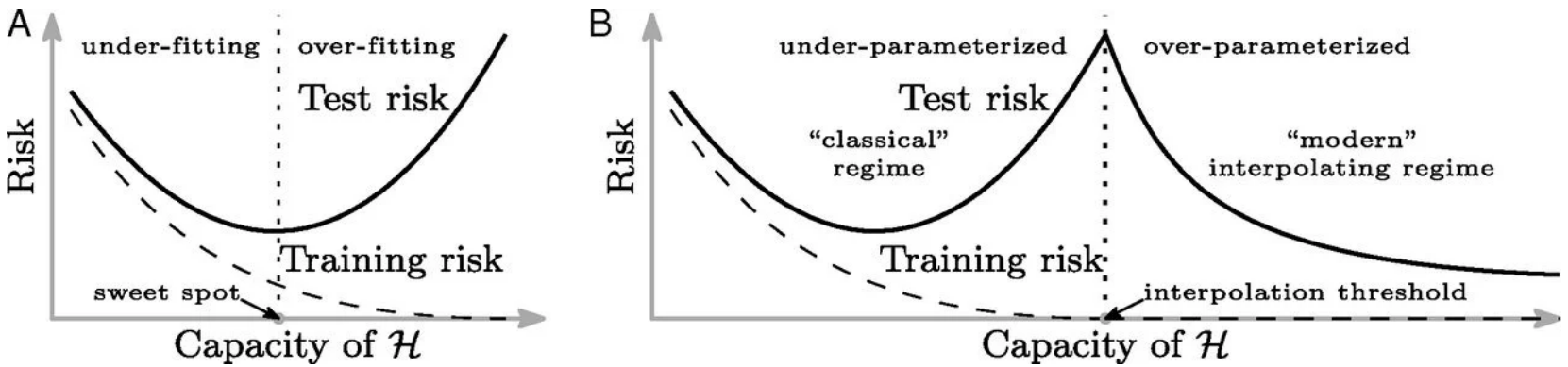
  where X^{+} is the Moore–Penrose pseudoinverse of X.
- Can be seen as the limit of Ridge as $\lambda \to 0^{+}$:

$$\hat{\beta}_{\text{ridgeless}} = \lim_{\lambda \to 0^{+}} (X^{\top} X + \lambda I)^{-1} X^{\top} y$$

- Key points:
  - Fits the training data perfectly (zero training error)
  - Implicitly regularizes by discarding directions in the null space of X
- Crucially it exhibits the double descent phenomena, where Test MSE initially descends, then spikes at the interpolation threshold ($p \approx n$), and then descends again

The Interpolation Threshold

- The interpolation threshold occurs when the number of parameters matches the number of observations $p \approx n$
- At this point the training data can be fitted perfectly.
- The matrix $X^\top X$ becomes nearly singular.
- Small singular values of X cause coefficients to explode.
- Ridgeless regression selects the minimum-norm interpolator.



History of double descent

- 1970s–2000s: Ridge and Lasso regression popularized for stabilizing OLS in multicollinear or high-dimensional settings. Classic "U"-shaped bias-variance tradeoff idea popular.
- 2010s–present: Massively overparametrized neural networks such as Alexnet became state of the art.
- 2019: The seminal paper "Reconciling modern machine learning practice and the bias-variance trade-of" (Belkin et al., 2019) formalized the idea of the double descent phenomena.
- 2022: The paper "Surprises in high-dimensional ridgeless least squares interpolation" (Hastie et al., 2022) directly connected this phenoma to linear regression.
- Ridgeless regression (minimum-norm solution) became a central theoretical model for understanding interpolation in deep learning.

Connections with Gradient Descent

- Initialize $\beta^{(0)} = 0$ and run gradient descent on the least squares loss:

$$\beta^{(k)} = \beta^{(k-1)} + tX^\top(y - X\beta^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

  with step size $t > 0$ sufficiently small.

- Gradient descent converges to the minimum-norm least squares solution:

$$\lim_{k \to \infty} \beta^{(k)} = \beta_{\text{ridgeless}}.$$

- Intuition:
  - Each update $\beta^{(k)}$ lies in the row space of X.
  - The limit therefore also lies in the row space.
  - The minimum-norm solution is the unique solution in the row space (Tibshirani, 2023).
- This connection also extends to stochastic gradient descent, widely used in training neural networks.

Simulation Setup

- Inspired by the first simulation done in (Hastie et al., 2022)
- Generate a design matrix $X \in \mathbb{R}^{n \times p}$ with independent Gaussian entries:

$$X_{ij} \sim N(0, 1), \quad i = 1, \ldots, n, \; j = 1, \ldots, p.$$

- Draw random $\beta \sim N(0, I_p)$ and we rescale to achieve the desired signal-to-noise ratio (SNR):

$$\beta^* := \beta \frac{\sqrt{\mathrm{SNR}}\,\sigma}{\|\beta\|_2}.$$

  Here, $\sigma^2 := 1$ is the variance of the noise.
- Finally, to generate the responses,

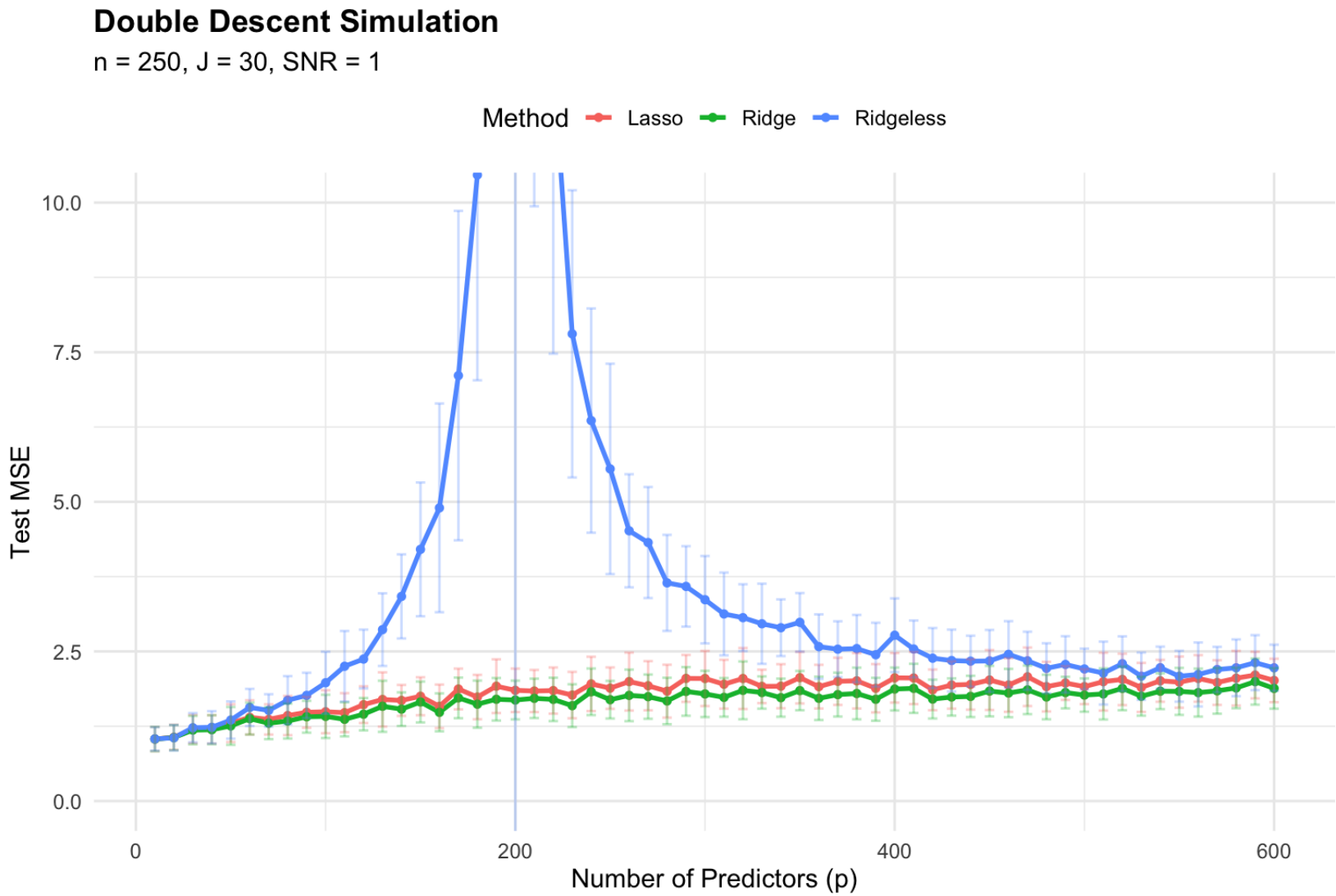
$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

Simulation Algorithm: Double Descent Study

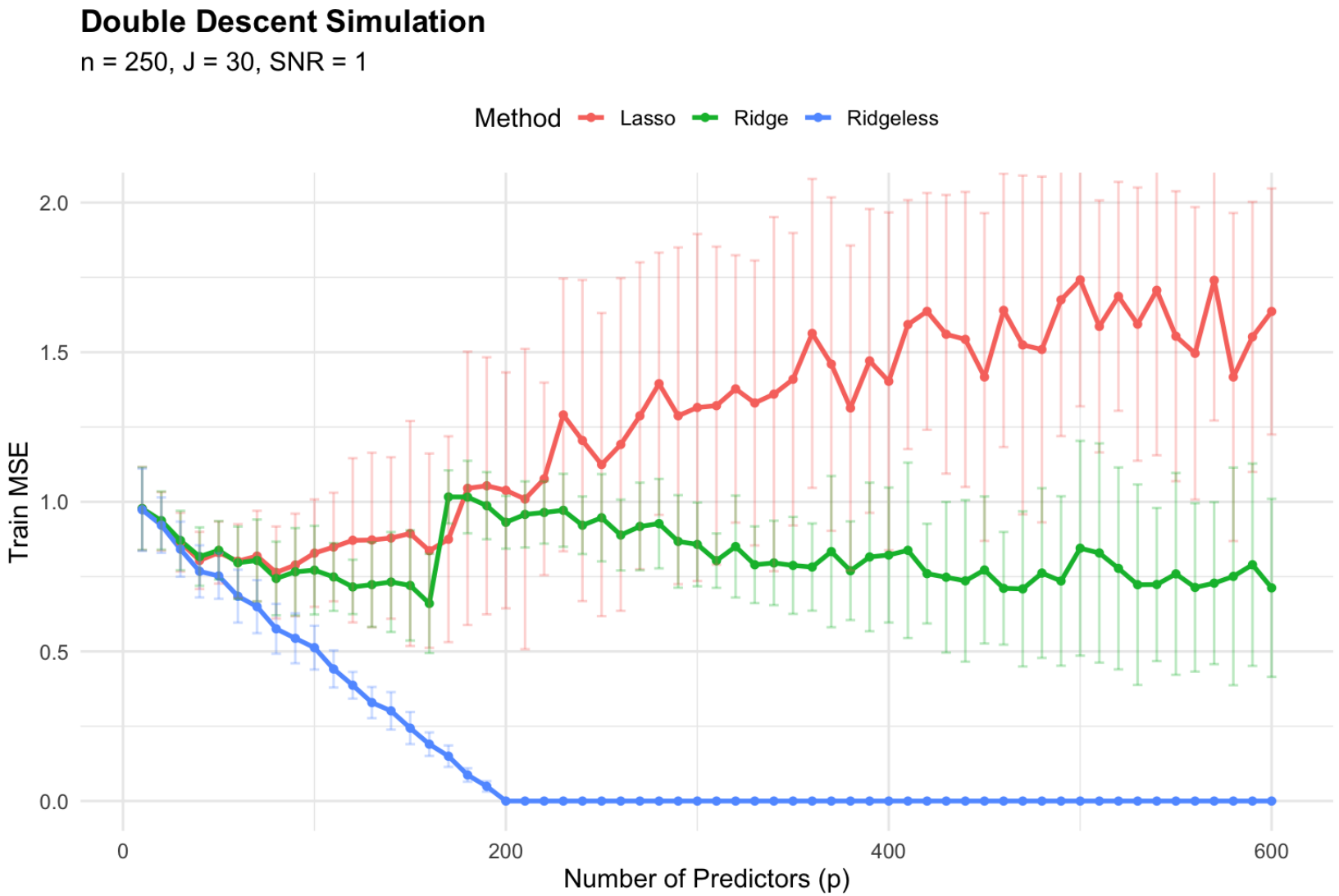
Algorithm 1 Double Descent Simulation with Fixed True Sparsity

Inputs: Sample size $n = 250$, Dimension sequence $P = \{10, 20, \ldots, 600\}$, Number of simulations $J = 30$, SNR, σ^2
for $p \in P$ **do**
    for $j = 1$ to J **do**
        Generate data with dimension p with given SNR, σ^2, and seed $100 + j$
        Split sample into 80/20 training/test sets
        Fit: Ridge, Lasso, and Ridgeless on training data
        Compute test MSE for each method
    end for
    Compute average test MSE across simulations $j = 1, \ldots, J$
end for
Output: For each p, mean test MSE of Ridge, Lasso, and Ridgeless

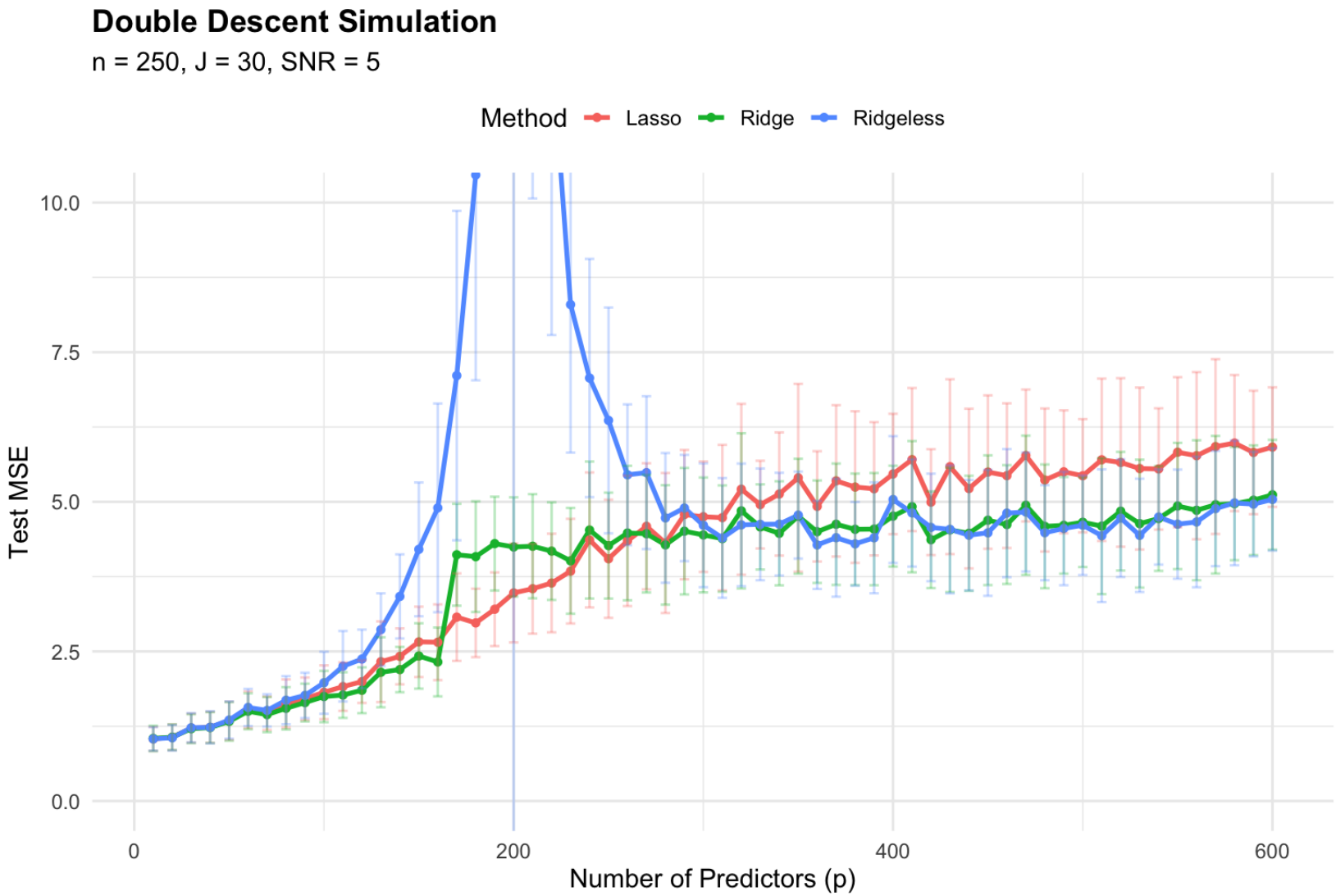
Results 1/3


Double Descent Simulation
n = 250, J = 30, SNR = 1
Method
Lasso
Ridge
Ridgeless
Test MSE
10.0
7.5
5.0
2.5
0.0
0
200
400
600
Number of Predictors (p)

Results 2/3


Double Descent Simulation
n = 250, J = 30, SNR = 1
Method
Lasso
Ridge
Ridgeless
Train MSE
2.0
1.5
1.0
0.5
0.0
0
200
400
600
Number of Predictors (p)

Results 3/3


Double Descent Simulation
n = 250, J = 30, SNR = 5
Method
Lasso
Ridge
Ridgeless
Test MSE
10.0
7.5
5.0
2.5
0.0
0
200
400
600
Number of Predictors (p)

Conclusion

- High-Dimensional data is becoming increasingly common but poses a challenge when utilizing linear models.
- Ridge stabilizes solutions by shrinking all directions and excels when there is multicollinearity.
- Lasso induces sparsity and does well in that domain.
- Ridgeless regression interpolates the data and chooses the minimum norm solution. It appears unstable near $p = n$, but improves again for $p > n$
- Double descent provides a unified view linking classical statistics and modern deep learning.

References

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116. URL https://doi.org/10.1073/pnas.1903070116.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, April 2022. doi: 10.1214/21-AOS213. URL https://doi.org/10.1214/21-AOS213.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer, Berlin, 2 edition, 2023. doi: 10.1007/978-1-0716-1418-1. URL https://doi.org/10.1007/978-1-0716-1418-1.

Ross MacAusland. The moore–penrose inverse and least squares. Lecture Notes for MATH 420: Advanced Topics in Linear Algebra, University of Puget Sound, n.d.

Ryan Tibshirani. Overparametrized regression: Ridgeless interpolation. Lecture Notes for Advanced Topics in Statistical Learning, Spring 2023, 2023.

Link to Github repo with code:
https://github.com/tyler3schmidt/STAT5200_computing_project