Navigating the Double Descent Landscape: Unraveling Model Complexity and Generalization in Finite Dimensions

Polly Pu

Abstract

This study investigates the double descent phenomenon using quasi-real data across finite dimensions, focusing on the dynamics of model risk in underparameterized and overparameterized regimes. By employing Singular Value Decomposition and varying model complexity, we analyze the impact of each term contributing to the test error. The empirical results highlight the significance of the interpolation threshold, singular values, and the alignment between training and test feature spaces. Furthermore, we explore the role of signal-to-noise ratio in shaping the double descent curve and the model's ability to capture the underlying data structure. Notably, in the absence of strong learners, the overparameterized regime demonstrates superior performance, particularly at higher signal-to-noise ratio levels. These findings provide valuable insights into the exploitation of overparameterization in machine learning. This study contributes to a deeper understanding of model complexity, generalization, and the effective utilization of overparameterization in machine learning.

1 Introduction

The research project at hand delves into the intricate dynamics of the double descent phenomenon—a concept that stands in stark contrast to the traditional wisdom within the field of machine learning. Typically, one would anticipate that as the complexity of a model increases, its performance on unseen data first improves and then deteriorates due to overfitting. However, double descent reveals a peculiar trend where, beyond a certain complexity threshold, further increasing the model's parameters can actually lead to an improvement in performance.

This anomaly is not just a theoretical curiosity but has significant practical implications. It compels us to reevaluate our strategies for model selection. Optimal model performance can no longer be reliably predicted by simply balancing bias and variance; the entire complexity spectrum must be considered. This includes what is known as the overparameterized regime, where models contain more parameters than training data points—an increasingly common scenario in contemporary deep learning endeavors.

Furthermore, our study concentrates on the signal-to-noise ratio (SNR), an integral measure that gauges the proportion of meaningful information to the background noise within a dataset. The importance of SNR extends well beyond machine learning, affecting disciplines as diverse as telecommunications and biological signal processing. By exploring the influence of SNR on double descent, we aim to unravel how the clarity of the underlying patterns in data affects model training across different complexity regimes.

Understanding these phenomena is paramount in an era dominated by big data and complicated algorithms, where selecting the appropriate model complexity is more challenging than ever. The findings from this research could potentially redefine how we approach the development and training of machine learning models, paving the way for novel regularization techniques that can exploit the benefits of high model complexity without succumbing to overfitting. This research stands to make a substantial impact on how we harness the full potential of machine learning models in practical, real-world scenarios.

2 Related Work

Recent papers have made significant contributions to our understanding of the double descent phenomenon in machine learning models. Hastie et al. [2] analyze the risk of minimum l_2 norm ("ridgeless") least squares regression under linear and nonlinear feature models, showing that the peak in test risk around the interpolation threshold and the descent of risk in the overparameterized regime are broadly present across these settings. While their analyses provide a wide range of insights under different modeling assumptions, they are asymptotic, potentially limiting finite-sample applicability.

Similarly, Belkin et al. [1] investigate min-norm interpolation in high-dimensional regression, recovering the ”double descent” behavior in both linear and nonlinear feature models. They provide both asymptotic and finite-sample results, with the latter being a strength, although requiring Gaussian assumptions. The paper demonstrates the essentiality of overparameterization for benign overfitting and reveals factors driving double descent, such as small but non-zero singular values in the feature matrix, substantial variation between test and training features, and residual errors from the best-in-class model.

Building upon these findings, Schaeffer et al. [5] delve into the factors causing double descent in ordinary linear regression, identifying three key elements that together create the phenomenon: small but non-zero singular values in the feature matrix, substantial projection of test features onto small right singular vectors of the feature matrix, and projection of best-in-class model residuals onto left singular vectors. They empirically demonstrate double descent on real datasets and ablate the three factors by varying the data sample size, showing that removing any of them prevents double descent. The study’s strengths lie in its intuitive explanation of key factors and empirical demonstrations on real data.

Nakkiran et al. [3] take a more empirical approach, observing the ubiquity of double descent behavior across several modern machine learning models and datasets, including deeper neural networks and random forests, as model complexity increases. While their work showcases the widespread presence of double descent in real-world settings, it does not provide mathematical analyses to explain the phenomenon.

These papers provide valuable context and insights for my work on analyzing double descent using quasi-real data. Hastie et al. and Belkin et al. establish the generality of double descent in asymptotic settings and identify key driving factors. Schaeffer et al. crystallize the explanation for double descent in linear regression, grounding the insights and demonstrating their empirical manifestation. Nakkiran et al. highlight the broad empirical presence of double descent.

My research aims to build upon these works by addressing some of their limitations. Most of the existing studies focus on asymptotic regimes and make distributional assumptions about the data generation process. In contrast, my work considers quasi-real data across finite dimensions to study the transition between regimes, varying the model complexity by changing the number of features fitted using the simple linear regression model without enforcing any assumptions on the model or the data. This approach allows for a more comprehensive analysis of the double descent phenomenon in practical settings, where the assumptions made in previous studies may not always hold.

3 Data Sources and Experimental Design

3.1 Data Source

The study utilizes the ”Communities and Crime” dataset [4] sourced from the UCI Machine Learning Repository, which integrates socio-economic data from the 1990 U.S. Census, law enforcement data from the 1990 Law Enforcement Management and Administrative Statistics (LEMAS) survey, and crime data from the 1995 FBI Uniform Crime Report (UCR). This dataset comprises 1,994 instances and 128 attributes that capture various aspects of community characteristics and crime statistics, emphasizing demographics, economic factors, and law enforcement metrics such as median family income and per capita police presence. The data within the dataset is normalized to a decimal range from 0.00 to 1.00 using an unsupervised, equal-interval binning method. This normalization approach maintains the distribution characteristics within each attribute while truncating values beyond three standard deviations to 1.00 or 0.00, effectively minimizing the influence of outliers. This method does not preserve the relationships between different attributes, which can be advantageous in analyses where features are abundant but individually insufficient to capture the complex structure of the underlying data distribution, a common scenario in machine learning practices.

3.2 Data Transformation

After an initial review of the ”Communities and Crime” dataset, we identified and removed six categorical columns, such as ’community’ and ’community name’, that are non-predictive and do not contribute to our analysis. We also addressed missing values, which were primarily concentrated in 23 columns related to the LEMAS survey. These columns were excluded, as their absence does not impact the integrity of our analysis, reducing our feature set to 99 columns.

To expand the dimensionality of our feature space and enhance the model's ability to capture complex patterns, we applied a Hermite transformation to each of the remaining 99 features. Hermite transformation involves generating polynomial features up to a specified degree—in this case, up to the 10th degree—for each feature. This method is particularly useful in uncovering non-linear relationships that might be present in the data by considering higher-order terms of the variables. Importantly, Hermite polynomials produce a basis that is orthogonal with respect to the weight function of the Gaussian distribution, which can improve the stability and efficiency of our regression model by reducing issues related to multicollinearity among transformed features.

Additionally, we introduced interaction terms between the original 99 columns. Interaction terms are essential for modeling the effects of interacting predictors and can provide insights into how the combined influence of two features affects the dependent variable, which in this context is the crime rate. This approach allows us to explore synergistic or antagonistic effects between different community and law enforcement characteristics.

Through these transformations, the final dataset encompasses 5,940 columns.

3.3 Experimental Design

In our experimental design, given the expanded and transformed dataset, we employ the LASSO technique to refine our linear model and establish a baseline for feature importance. The selection of features is guided by the LASSO path, which plots the trajectory of each coefficient against different values of the regularization parameter α. Features are ranked based on the value of α at which their coefficients shrink to zero, as illustrated in Figure 1. A higher α value indicates that the feature's coefficient remains non-zero longer, signifying greater importance.

To implement this, we use a LassoCV model configured with five-fold cross-validation and without an intercept to determine the most predictive features. The optimal model yielded 33 non-zero coefficients at an α value of 0.0134. Using this model configuration, we simulate quasi-real data by constructing a new response variable $\vec{y}$, calculated as $\vec{y} = \mathbf{X}\beta_{\text{LASSO}} + \vec{\epsilon}_{\text{LASSO}}$. Here, $\vec{\epsilon}_{\text{LASSO}}$ represents the noise term, sampled from the empirical distribution of residuals derived from our best-fitting LASSO model.

This quasi-simulated dataset serves as the basis for further analysis, allowing us to assess the robustness and generalizability of our findings under controlled conditions that mimic real-world data variations. This approach ensures that our evaluations and conclusions are not only based on the observed data but are also tested against potential variations within the same structural framework.

4 Linear Regression Under Two Regimes

We begin by defining our supervised dataset with N training data points, each with Q features:

$$\mathbf{X} := \begin{bmatrix} \vec{x}_1 & \vec{x}_2 & \dots & \vec{x}_N \end{bmatrix}^\top \quad \text{and} \quad \vec{y} := \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}^\top$$

where $\vec{x}_n \in \mathbb{R}^Q$ and $y_n \in \mathbb{R}$.

Our goal is to identify the best linear estimator by minimizing the expected loss:

$$\hat{f}(\cdot) := \arg\min_f \mathbb{E}[l(y_{\text{new}}, f(\vec{x}_{\text{new}}))]$$

In the Ordinary Least Squares (OLS) setting, we assume f is a linear function:

$$y_{\text{new}} \approx \vec{x}_{\text{new}}^\top \hat{\beta}$$

where $\hat{\beta}$ is defined as:

$$\hat{\beta} := \arg\min_\beta \|\mathbf{X}\beta - \vec{y}\|_2^2$$

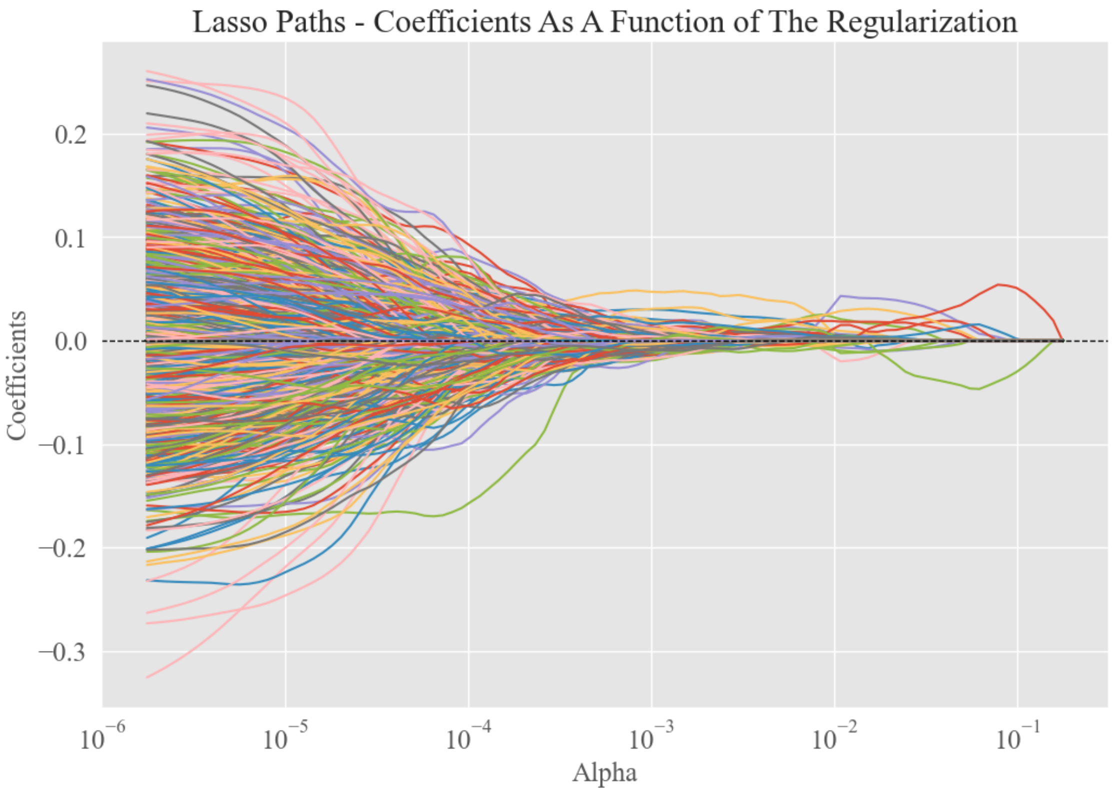

Figure 1: LASSO Paths

When $P \leq N$ and $\mathbf{X}$ has full column rank, $\hat{\beta}$ has a unique solution. In contrast, the overparameterized scenario where $P > N$ leads to infinitely many solutions. Here, we focus on the least-norm solution:

$$\min_{\beta \in S} \|\beta\|_2^2 \quad \text{where} \quad S := \arg\min_{\beta} \|\mathbf{X}\beta - \vec{y}\|_2^2$$

This solution set S can be expressed using the Moore-Penrose pseudoinverse ($\mathbf{X}^\dagger$) as:

$$S = \{\mathbf{X}^\dagger \vec{y} + \vec{z} \mid \vec{z} \in \mathcal{N}(\mathbf{X})\}$$

To simplify calculations and generalize results, especially when $\mathbf{X}$ does not have full rank, we employ Singular Value Decomposition (SVD):

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^\top$$

where $\text{rank}(\mathbf{X}) = r$. The minimization problem transforms to:

$$\min_{\beta} \|\mathbf{X}\beta - \vec{y}\|_2^2 = \min_{\beta} \|\Sigma\mathbf{V}^\top\beta - \mathbf{U}^\top\vec{y}\|_2^2$$

Setting $\gamma = \mathbf{V}^\top\beta$ and $\vec{z} = \mathbf{U}^\top\vec{y}$, the minimized function becomes:

$$\|\Sigma\gamma - \vec{z}\|_2^2 = \sum_{i=1}^{r} (\sigma_i\gamma_i - z_i)^2 + \sum_{i=r+1}^{\min\{N,P\}} z_i^2$$

The optimal solution for γ_i is:

$$\gamma_i = \begin{cases} \frac{z_i}{\sigma_i} & \text{if } i \leq r \\ 0 & \text{if } i > r \end{cases}$$

This choice minimizes $\|\gamma\|_2^2$, providing the least norm solution:

$$\hat{\beta} = \mathbf{V}\hat{\gamma} = \sum_{i=1}^{r} \frac{1}{\sigma_i} \vec{v}_i \vec{u}_i^\top \vec{y} = \mathbf{X}^\dagger \vec{y}$$

where $\mathbf{X}^\dagger = \mathbf{V}\Sigma^\dagger \mathbf{U}^\top$ is the Moore-Penrose pseudoinverse of $\mathbf{X}$. This approach acts as an implicit form of regularization. By focusing on solutions with the smallest Euclidean norm, SVD naturally dampens the influence of less significant components, specifically those corresponding to smaller singular values. This selective attenuation helps to stabilize the solution by reducing the impact of noise and potential overfitting, especially relevant in high-dimensional settings where the number of features exceeds the number of observations. In the context of the double descent phenomenon, this approach can mitigate the peak of test error typically observed at the interpolation threshold—where the model complexity exactly matches the number of training samples. By implicitly regularizing through SVD, the model's complexity is effectively managed, potentially smoothing out the double descent curve and ensuring more robust generalization even as model complexity continues to increase. Consequently, our analysis consistently applies SVD regression across both underparameterized and overparameterized regimes.

5 Analyzing the Test Errors Under Two Regimes

Assume the optimal linear model is defined by $\vec{y} = \mathbf{X}\beta^* + \vec{e}$, where $\vec{e}$ captures errors inherent to the best linear model. These errors may include random, irreducible noise or deterministic patterns that linear regression fails to capture for the data population.

Define $\vec{y}^* := \mathbf{X}\beta^*$ as the deterministic component of the model. The prediction error analysis for both underparameterized and overparameterized regimes, especially when evaluated along the orthogonal singular modes of $\mathbf{X}$, can be expressed as:

$$\hat{y}_{\text{test}} - y^*_{\text{test}} = \vec{x}_{\text{test}}^\top (\hat{\beta} - \beta^*) = \vec{x}_{\text{test}}^\top (\mathbf{X}^\dagger \vec{y} - \beta^*)$$

This expands to:

$$\vec{x}_{\text{test}}^\top (\mathbf{X}^\dagger \vec{y} - \beta^*) = \vec{x}_{\text{test}}^\top \mathbf{V}\Sigma^\dagger \mathbf{U}^\top (\mathbf{X}\beta^* + \vec{e}) - \vec{x}_{\text{test}}^\top \beta^*$$

Simplifying further, we get:

$$\vec{x}_{\text{test}}^\top \mathbf{V}\Sigma^\dagger \mathbf{U}^\top \vec{e} = \sum_{i=1}^{r} \frac{1}{\sigma_i} (\vec{x}_{\text{test}}^\top \vec{v}_i)(\vec{u}_i^\top \vec{e})$$

This equation quantifies the projection of the test vector $\vec{x}_{\text{test}}$ onto the orthogonal singular modes of $\mathbf{X}$, multiplied by the respective singular values and the projection of the error vector $\vec{e}$. The components $(\vec{x}_{\text{test}}^\top \vec{v}_i)$ represent the alignment of the test data with the singular vectors, and $(\vec{u}_i^\top \vec{e})$ measure the contribution of the error aligned along each singular mode.

This formulation matches the variance term derivation in the analysis from Belkin et al. [1] and Schaeffer et al. [5], providing a theoretical foundation for understanding the error dynamics in different regimes of parameterization in terms of singular value decomposition. This highlights how errors propagate through the dimensions retained in the model, thus illuminating the potential limitations and strengths of linear models in high-dimensional settings.

Analysis of Each Term's Influence on Test Error

1. $\frac{1}{\sigma_i}$**:** This term is the inverse of the i-th singular value, σ_i, of the matrix $\mathbf{X}$. Singular values measure the "strength" or "influence" of their corresponding singular vectors in representing the dataset. Smaller singular values correspond to directions in which the data has less variance. The inverse of these values amplifies the impact of errors in directions where the data is naturally less spread out. Increasing the number of features in high-dimensional space typically leads to smaller singular values (σ_i), as more dimensions often mean capturing more noise or irrelevant variations. This decrease in σ_i results in larger values of $\frac{1}{\sigma_i}$, hence potentially increasing the impact of each term.

2. $(\vec{x}_{\textbf{test}}^{\top}\vec{v}_i)$**:** This component represents the projection of the test vector $\vec{x}_{\text{test}}$ onto the i-th right singular vector $\vec{v}_i$ of $\mathbf{X}$. The right singular vectors form an orthonormal basis for the feature space, and this projection quantifies how much of the test data lies along the direction defined by $\vec{v}_i$. In high-dimensional spaces, where feature vectors can have complex interactions, the alignment of $\vec{x}_{\text{test}}$ with certain singular vectors can significantly determine how those specific dimensions influence the prediction. If the additional features increase the alignment of test data with less significant modes, the impact of the term on the test error will be larger.

3. $(\vec{u}_i^{\top}\vec{e})$**:** This term is the projection of the error vector $\vec{e}$ onto the i-th left singular vector $\vec{u}_i$. Since the left singular vectors correspond to the data space (rows of $\mathbf{X}$), this projection measures how much of the error is expressed along the dimension associated with $\vec{u}_i$. In scenarios where the error components align significantly with these vectors, especially those corresponding to smaller singular values, their contribution to overall prediction errors can be disproportionately high.

6 Empirical Analysis

6.1 Methods

Building upon the empirical analysis methodology detailed in Section 3.3, we delve into an examination of the test error dynamics by selecting a subset of 100 samples from the quasi-simulated data. Our procedure involves carrying out SVD regression on a spectrum of features, ranging from 1 to 800, chosen sequentially in the order determined by the LASSO paths. The impact of each term contributing to the test error, as explicated in the prior section, is scrutinized through this process. Employing 10-fold cross-validation for each regression iteration aids in estimating the predictive risk accurately.

For the empirical analysis of the first component, $\frac{1}{\sigma_i}$, we graph the smallest non-zero singular values of the training matrix $\mathbf{X}$ within each dimensional setting $\mathbf{X} \in \mathbb{R}^{N \times P}$ where $P \in \{1, \ldots, 800\}$. This visualization aids in understanding how the singularity of $\mathbf{X}$ evolves with increasing feature space and its consequent effect on the test error.

Regarding the second term $(\vec{x}_{\text{test}}^{\top}\vec{v}_i)$, we employ principal angle analysis between the subspaces spanned by the training and test sets' features. The principal angle is defined as the smallest angle θ between two subspaces $\mathcal{U}$ and $\mathcal{V}$, and can be determined by finding unit vectors in these subspaces that are maximally aligned. This analysis is achieved by constructing orthogonal bases Q_U and Q_V for $\mathcal{U}$ and $\mathcal{V}$, respectively, followed by computing the inner product matrix $R = Q_U^T Q_V$. The SVD of R then yields the principal angles, with the smallest principal angle signified by $\cos(\theta) = \sigma_{max}$, the largest singular value. By performing this analysis, we can gauge the alignment between the test features and the dominant modes of variation in the training data, hence appraising the influence of this term on the test error.

For the third term $(\vec{u}_i^{\top}\vec{e})$, the approach involves leveraging the error term $\vec{\epsilon}_{\text{LASSO}}$ inferred from the LASSO model as the best linear model error. We calculate the maximal projections of this error onto the left singular vectors U derived from the SVD of the training matrix, i.e., $\max(U^T\vec{\epsilon}_{\text{LASSO}})$, and record the corresponding singular values associated with the maximally projected $\vec{u}_i$. This procedure helps in quantifying the extent to which the directions with the most significant errors contribute to the overall test error, thus providing insights into the model's predictive performance in the presence of noise and model mis-specification.

6.2 Results

The results of the empirical analysis elucidate intriguing dynamics of model risk across different feature set sizes. As depicted in Figure 2, the training risk exhibits a consistent decline with the increasing number of features (P), descending to zero sharply at the interpolation threshold—where the number of features equals the number of samples. Contrarily, the test risk demonstrates a divergent behavior, escalating to its apex at the interpolation threshold before diminishing as we progress into the overparameterized regime.

Interestingly, in the underparameterized regime, the inclusion of merely four features yields the lowest test error at 0.0736. This minimal error can be attributed to the fact that these features are likely integral to the true data-generating process, as they are derived from the optimal LASSO results used to create the quasi-simulated target. Therefore, these classical regime features are highly informative, leading to superior

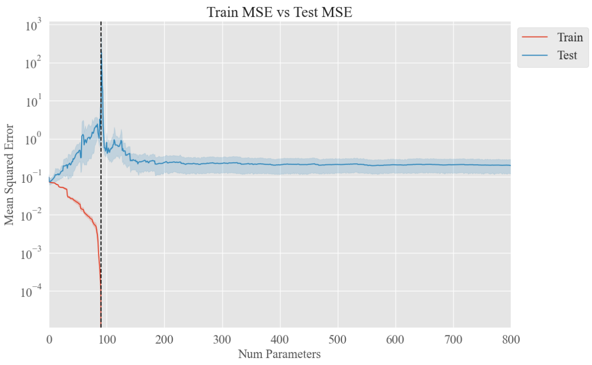

Figure 2: Train MSE vs. Test MSE

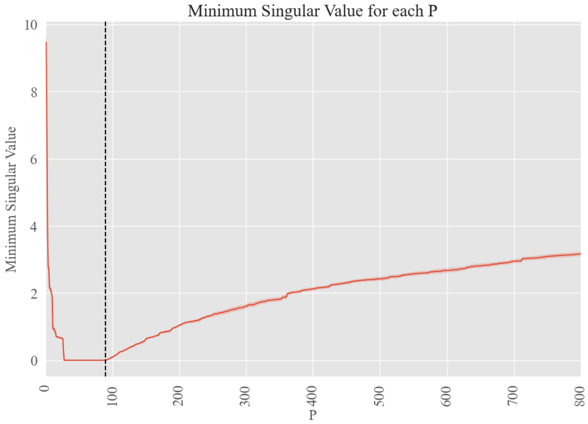

Figure 3: Minimum Singular Values

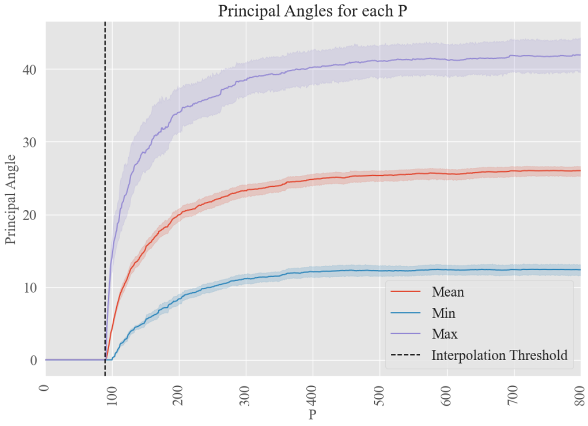

Figure 4: Principal Angles

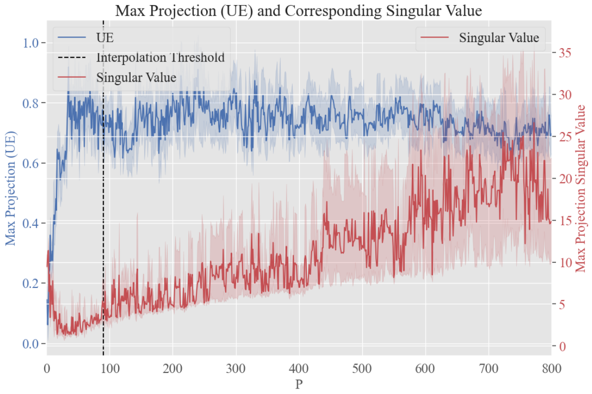

Figure 5: $(U^{\top}\vec{\epsilon}_{LASSO})_{i=max}$ vs $\sigma_{i=max}$

performance compared to the overparameterized setting, where $P = 470$ features produce a risk of 0.1960. This suggests that within the classical regime, a model with the true data-generating features is able to capture the underlying structure more effectively than a model with excessive, possibly redundant features.

In Figure 3, the minimum singular values from the training sets follow a downward trajectory towards zero as we approach the interpolation threshold. This descent implies that as more features are added, the data structure captured by the model becomes less stable, inflating the test error. After surpassing the threshold, however, these singular values begin to increase, which correlates with the observed decrease in the test error within the overparameterized regime. This phenomenon indicates that the additional features beyond the interpolation threshold enhance the model's ability to generalize, despite the complexity introduced by a higher-dimensional feature space.

The principal angle analysis, shown in Figure 4, reveals that the angles between the subspaces spanned by the training and test sets initially remain zero, indicating perfect alignment. As more features are added beyond the interpolation threshold, these angles start to widen. This widening is a direct consequence of the increased dimensionality, which grants more freedom for the subspaces to diverge, potentially leading to a greater discrepancy between the training and test spaces. This divergence could partially explain the persistence of high test risk as P grows since it could lead to models that capture noise specific to the training data rather than the underlying data structure.

Figure 5 showcases the maximum projection of the best linear model error onto the left singular vectors and their corresponding singular values. It is observed that the maximal projection heightens with an increasing number of features, aligning with smaller singular values. This pattern implies that as more features are added, the directions of least variance begin to dominate the error structure, inflating the test error up to the interpolation threshold. Once we enter the overparameterized regime, these projections stabilize and show a slight decline, with corresponding singular values growing larger. This trend suggests that the model starts to discount the less significant, noisy directions in favor of more stable structural features of the data, which could contribute to improved prediction accuracy despite the higher dimensionality.

7 Adjusting Signal Noise Levels

The signal-to-noise ratio (SNR) plays a pivotal role in the analysis of the double descent phenomenon. It provides a quantifiable measure of the ratio between the strength of the relevant information (signal) and the level of irrelevant or random variation (noise) present in the dataset. A higher SNR implies that the signal is more pronounced compared to the noise, thereby facilitating models to discern patterns more effectively and generalize better prior to reaching the interpolation threshold. Conversely, as model complexity increases, a lower SNR can lead to a conflation of noise with the signal, culminating in overfitting at the interpolation threshold. However, when moving into the overparameterized regime, models with excess capacity might begin to disentangle the noise from the signal once again, thereby reducing the test error in a second descent. In the subsequent section, we delve into the impact of varying SNR levels on the double descent curve through the analysis of quasi-simulated samples.

We define SNR as follows:

$$\text{SNR} = \frac{\text{Var}(\mathbf{X}\beta_{\text{LASSO}})}{\sigma^2_{\text{error}}}$$

where $\boldsymbol{\beta}$ represents the true underlying coefficients that generate the data for the optimal linear model as estimated by LASSO, and σ^2_{error} denotes the variance of the errors under this model. The current SNR stands at 1.8131.

To modulate the SNR, we adjust the variance of the error term in the model:

$$\vec{y} = \mathbf{X}\beta_{\text{LASSO}} + \vec{\tilde{\epsilon}} \tag{1}$$

where $\vec{\tilde{\epsilon}} \sim \mathcal{N}(0, \sigma^2_{\text{error}})$. The model alteration entails scaling σ^2_{error} by a factor r within the set $[0.5, 1, 8]$, thus redefining the error variance as $\sigma^2_{\text{error}} := \frac{\text{Var}(\vec{\epsilon}_{\text{LASSO}})}{r}$. Additionally, we characterize the null risk as the MSE of a baseline model that consistently predicts the mean of the training samples, i.e., $\vec{\hat{y}} = \bar{y}_{\text{train}}\vec{1}$.

Upon revising the error variances, we obtain SNRs of $[0.9066, 1.8131, 14.5052]$ for each corresponding value of r. As illustrated in Figure 6, augmenting the SNR engenders a downward shift of the test MSE and a more pronounced minimum within the classical regime. The magnitude of the test MSE reduction attenuates as the SNR increases, suggesting that when the signal becomes increasingly discernible, the model can capture the intrinsic data structure with a smaller set of features. Any supplementary features beyond this critical number are likely to align with noise rather than signal, exacerbating overfitting. Therefore, a higher SNR signifies that the model's performance not only improves relative to the null risk across both regimes but also indicates that the model's capacity is more effectively utilized in capturing the underlying data structure without being unduly influenced by noise.

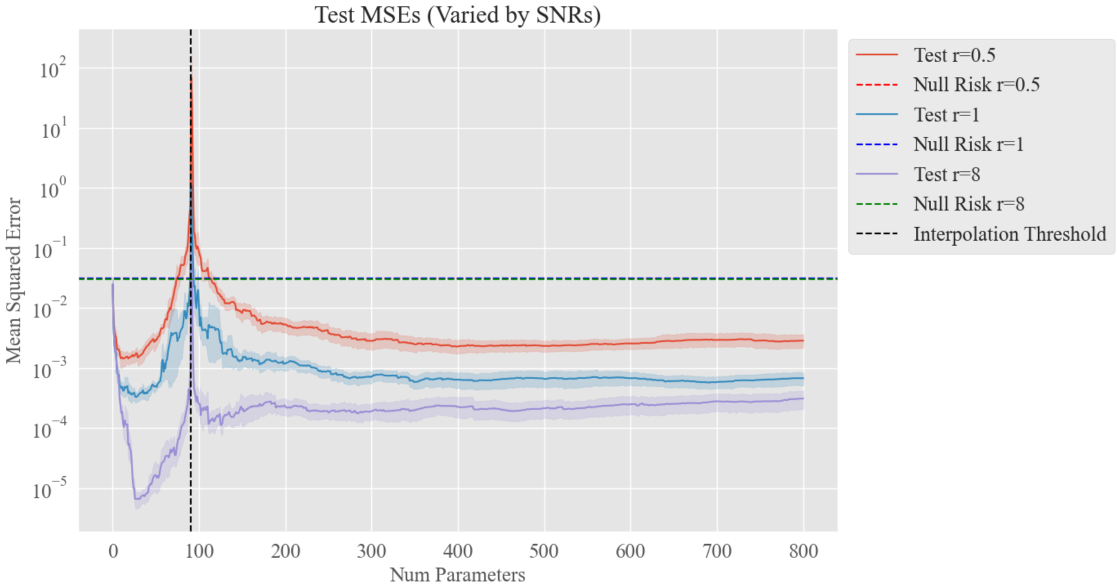

Figure 6: Test MSE Varied by SRNs

7.1 Mis-Specified Models

However, the analyses conducted thus far have confirmed the existence of a second descent in the overparameterized regime, but they have not demonstrated any local minima within this regime that surpass the performance seen in the classical (underparameterized) regime. As articulated in earlier discussions, the classical regime may exhibit superior performance due to the initial inclusion of true model features in the early stages of training. This scenario typifies a well-specified model where some features act as strong learners, capturing significant aspects of the underlying data distribution. Conversely, when the model lacks these strong learners—referred to as a misspecified model—the double descent curve and varying SNR can shed light on the system's behavior in a more ambiguous feature space where each predictor contributes modestly, akin to a weak learner.

Table 1: Comparison of Local Minima in Different Regimes

SNR	**Underparameterized Regime Local Min**	**Overparameterized Regime Local Min**
14.5052	0.0046	0.0028
1.8131	0.0042	0.0009
0.9066	0.0032	0.0004

We proceed by excluding the first 33 features, which are known to constitute the true data-generating mechanism, and reevaluating the analysis while manipulating the SNR ratios. The results, depicted in Figure 7, reveal a compelling shift: in the absence of the true model features, the overparameterized regime significantly outstrips the underparameterized regime. An increase in SNR enhances the performance within the overparameterized setting, emphasizing the model's ability to capitalize on the higher dimensionality when strong learners are absent. This observation is quantitatively captured in Table 1, highlighting the local minima across both regimes under varying SNR conditions.

This insight is particularly valuable when the true data-generating features are unknown or when no single feature stands out as a strong predictor. In such circumstances, an overparameterized model, especially at higher SNR levels, may become advantageous by integrating a multitude of weak learners to approximate the underlying data structure more faithfully than a model constrained by fewer features.

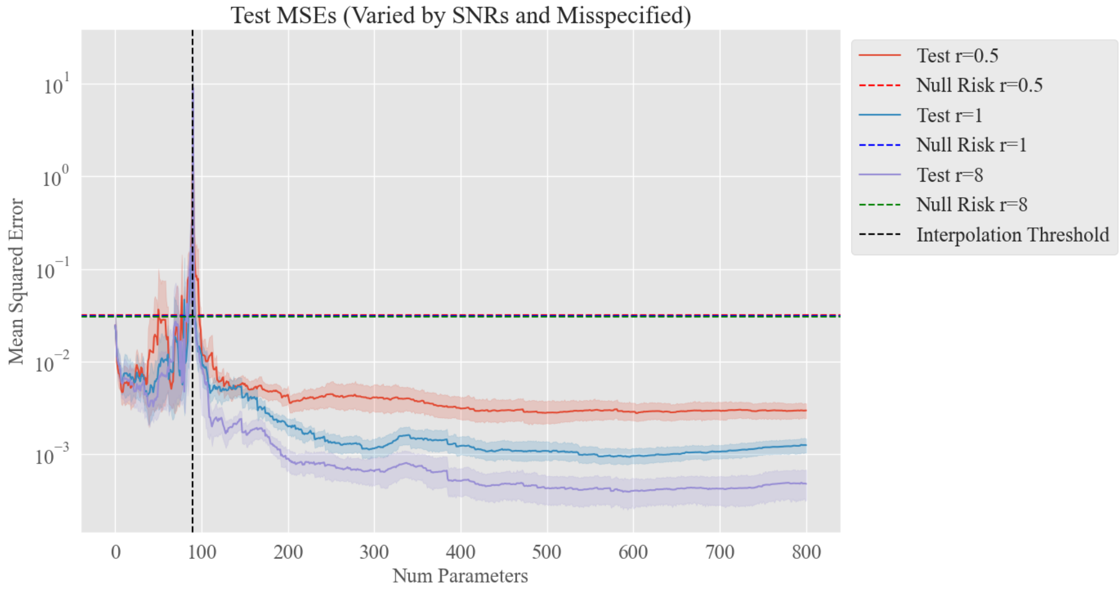

Figure 7: Test MSE Varied by SRNs (Mis-specified)

8 Conclusion and Future Works

In conclusion, this study has provided a comprehensive analysis of the double descent phenomenon using quasi-real data across finite dimensions. By varying model complexity and examining the impact of each term contributing to the test error, we have gained insights into the dynamics of model risk in both underparameterized and overparameterized regimes. The empirical results highlight the importance of the interpolation threshold, the influence of singular values, and the alignment between training and test feature spaces. Furthermore, by adjusting signal-to-noise ratios, we have demonstrated the significant role of SNR in shaping the double descent curve and the model's ability to capture the underlying data structure. Notably, in the absence of strong learners, the overparameterized regime has shown superior performance, particularly at higher SNR levels.

Future research could focus on extending these findings to larger data sample sizes and exploring their direct applications to machine learning. A clear understanding of how the overparameterized regime can be exploited for machine learning is crucial. Additionally, as all the error metrics used in the current analyses are based on the l_2 norm, investigating the behavior of the overparameterized regime under different l_p norms may yield interesting results and provide a more comprehensive understanding of the double descent phenomenon. These future research directions have the potential to further advance our knowledge of model complexity, generalization, and the effective utilization of overparameterization in machine learning.

References

[1] Mikhail Belkin, Daniel Hsu, and Ji Xu. “Two models of double descent for weak features”. In: (Mar. 2019).

[2] Trevor Hastie et al. “Surprises in high-dimensional ridgeless least squares interpolation”. In: *The Annals of Statistics* 50 (Apr. 2022). DOI: 10.1214/21-AOS2133.

[3] Preetum Nakkiran et al. “Deep double descent: where bigger models and more data hurt*”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2021 (Dec. 2021), p. 124003. DOI: 10.1088/1742-5468/ac3a74.

[4] Michael Redmond. *Communities and Crime Unnormalized*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5PC8X. 2011.

[5] Rylan Schaeffer et al. “Double Descent Demystified: Identifying, Interpreting Ablating the Sources of a Deep Learning Puzzle”. In: (Mar. 2023).