Exploration of Overparameterization in Linear Regression

Andreas Oliveira
B.Sc Computer Engineering, Boston University
andoliv@bu.edu

December 13, 2022

1 Introduction

In contrast with the classical bias vs variance trade off view, Neural Networks have shown an impressive ability to overfit and still provide good generalization. Therefore, a flurry of research activity emerged in how to analyze these extremely over-parameterized models that were generalizing well. One interesting subfield of this research has turned to the simpler case of overparameterized linear regression. The intent is that by understanding how over-parameterization in linear regression can lead to good results, will allow us to build intuition on how this generalizes to more complex models. Notable works in this subfield include the papers by Hsu et. al. [3], Bartlet et. al. [1] and Sahai et. al. [9]. This project will focus on understanding, giving intuition and summarizing some of the main results from them. I will also show examples to empirically verify their results. Finally, the concept of effective rank in the paper by Bartlet is crucial to the quality of the interpolating solution and it also generalizes to any algorithm. Therefore, I will fetch the effective rank of two datasets where deep learning has overfitted while providing good generalization, to see if it is important in determining the goodness of an interpolating solution for neural networks as well.

2 Related Work

There has been a lot of active research in understanding double descent curves or the tension between over-fitting and generalization. The double descent phenomena was first coined by Belkin et. al. [5] to point out that the classical bias-variance tradeoff was maybe incorrect for neural networks. Additionally, studies like Zhang et al. [11] have shown the ability of neural networks to interpolate and generalize well or even interpolate noise, furthering this notion that deep learning requires overthinking generalization. From the perspective of Kernel Learning, Belkin et al. [4] showed that several Kernel machines trained to have zero regression error or classification loss can also perform very well on test data, even if it is corrupted with noise. Or with respect to bagging methods, Wyner et al. [10] proposes to explain the success of classifiers like Random Forests and Adaboost due to their interpolative nature (which is what happens with over-parameterized models). Now on the topic of over-parameterized linear regression. A paper by Hastie et. al. [7] analyzed the asymptotic risk of over-parameterized linear regression when both the dimension and sample size go to infinity under a fixed ratio γ. Mitra [8] also studies the case of over-parameterized linear regression in the l_1, l_2 min norm cases and obtains explicit analytical formulas on the risk of these min norm interpolating solutions. Bibas et al. [6], analyzes the case of online linear regression to show that under certain conditions of the correlation matrix the overparameterized solution can attain good generalization error. Finally on the three works that I will analyze, Hsu et al. [2] is mainly focused on describing the expected accuracy of a predictor $\hat{\beta}$ under the fact that the features are standard normal and uncorrelated. Sahai et al.

[9] proves several theorems, related to different assumptions on some whitened matrix, but ultimately they are all upper or lower bounds on the risk of an interpolative solution (not necessarily the min norm) with some probability guarantee. Bartlet et al. [1] gives lower bounds on the risk of the min norm interpolator based on some weaker assumptions of the distribution in comparison to Hsu et al. and Sahai et al.

3 Problem Formulation

This project will focus on the supervised learning setting of regression. That is, we assume there exists a distribution $\mathcal{D}$ which generates points (x, y). The point $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Furthermore, we assume both x, y have mean zero and $y = \beta^T x + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then we consider the problem of receiving a set S_n of n iid samples $(x, y) \sim \mathcal{D}$ and make a linear predictor $\hat{\beta}$, to minimize $L_{S_n}(\beta) = \frac{1}{n}\Sigma_{i=1}^m (y_i - \beta^T x_i)^2$. The generalizability of $\hat{\beta}$ is measured with respect to the $L_{\mathcal{D}}(\hat{\beta}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - \hat{\beta}^T x)^2]$. Further we will assume that some transformation on x is made to make a new set $a(x)$ where the prediction rule will act. Now the main question of this study is: how accurate is the predictor $\hat{\beta}$ if the number of dimensions it has is larger than the number of samples in the set S_n? First we go through some background on linear regression and then we go into the works that address this question.

3.1 Under-parameterized Linear Regression

If the number of parameters in β is smaller than the number of training samples n, then we call the system undetermined, because there is a unique solution to minimizing $L_{S_n}(\beta)$. The reason being we can consider the optimization of $L_{S_n}(\beta)$ as a problem of minimizing the distance between $||\mathbf{Y} - \mathbf{X}\beta||_2^2$ such that $\mathbf{Y}$ is the column vector of $y_i's \in S_n$ and $\mathbf{X}$ is the matrix where each row is an $x_i \in S_n$, which is an undetermined system if the dimension β is smaller than the dimension of $\mathbf{Y}$. Now, the minimization can be seen as trying to find the orthogonal projection of $\mathbf{Y}$ onto the subspace spanned by the columns of $\mathbf{X}$. Hence by the orthogonal projection principle:

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) = 0 \tag{1}$$

which implies $\beta = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{Y})$ where here if the sample covariance matrix is not invertible then we can use the pseudo inverse or some ridge regression, but for the purposes of the simulations we used the pseudo inverse.

3.2 Over-parameterized Linear Regression

If the number of parameters in β is larger than the number of training samples n, then we call the system overdetermined or overparameterized. The reason being the minimization of $||\mathbf{Y} - \mathbf{X}\beta||_2^2$ with respect to β has infinitely many solutions.

3.2.1 $\mathcal{L}_2$ Min-Norm Solution

One way to consider the minimization of $L_S(\beta)$ in an overdetermined system is finding:

$$\begin{aligned} \hat{\beta} = argmin_{\beta\in\mathbb{R}^d} \quad & ||\beta||_2 \\ \text{s.t.} \quad & \mathbf{X}\beta = \mathbf{Y} \end{aligned} \tag{2}$$

As it turns out this problem is well defined, meaning there is only a unique solution. In order to prove this notice that the zero vector is in the nullspace of $\mathbf{X}$ and since the nullspace and rowspace are orthogonal complements, then the vector in the rowspace of $\mathbf{X}$, namely $\mathbf{X}^T w$, that yields $\mathbf{X}\mathbf{X}^T w = \mathbf{Y}$ (which must exist if $\mathbf{Y}$ isn't just the zero vector) will be the argmin solution. Therefore, we solve the system of equations (2) by finding $w = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{Y}$ and then setting $\hat{\beta} = \mathbf{X}^T w$.

Which means the solution is:

$$\hat{\beta} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{Y}.$$

4 Comparison between Over and Under Parametrization

Now that we understand the under and over-parameterized linear regression models we turn to the

works of Hsu, Bartlet and Sahai to classify how do the solutions for them compare.

4.1 Hsu's Paper

In his paper Hsu focuses on the case where the features $x \sim \mathcal{N}(0, I_d)$ and again we have $y = \beta^T x + \epsilon$ where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Now the task we are focused on is in the setting where we are given a set of n samples to form $\mathbf{X} \in \mathbb{R}^{n \times d}$ but $d >> n$. Furthermore, we want to learn a predictor $\hat{\beta}$ for every $p \leq d$, in order to understand how the generalizability of $\hat{\beta}$ evolves as it has more and more dimensions. Hence, if $p < d$, then we must learn $\hat{\beta}$ by either choosing a set A of p random features from the original training matrix $\mathbf{X}$, and forming a new one $\mathbf{X}_A$ composing only of the features selected. Or by having a prescient model that chooses the p features from the d available and then forming the new matrix. There is also a small caveat with the equation $L_{\mathcal{D}}(\hat{\beta}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - \hat{\beta}^T x)^2]$ which is that if our $\hat{\beta}$ has less dimensions than x then we can't take its inner product with x. In order to overcome this caveat simply append zeros on the dimensions that weren't used to make $\hat{\beta}$. Finally, let β_A be the features of β that are in the set A and β_{A^c} be the features that are not in A. Now we can proceed to provide the main results.

4.1.1 Main Results

Theorem 1. *If the subset A, with cardinality p, is picked under the prescient model, and then $\hat{\beta}$ is learned, then the:*

$$\mathbb{E}_{(x,y),S_n}[(y-\hat{\beta}^T x)^2] = \begin{cases} (\|\beta_{A^c}\|^2 + \sigma^2)(1 + \frac{p}{n-p-1}) \text{ if} \\ p \leq n-2 \\ +\infty \text{ if } n-1 \leq p \leq n+1 \\ \|\beta_A\|^2(1 - \frac{n}{p}) + \\ (\|\beta_{A^c}\|^2 + \sigma^2)(1 + \frac{n}{p-n-1}) o.w \end{cases} \tag{3}$$

If the subset A, with cardinality p, is picked randomly, and then $\hat{\beta}$ is learned, then the:

$$\mathbb{E}_{(x,y),S_n}[(y-\hat{\beta}^T x)^2] = \begin{cases} ((1 - \frac{p}{d})\|\beta\|^2 + \sigma^2)(1 + \frac{p}{n-p-1}) \\ \text{if } p \leq n-2 \\ \|\beta\|^2(1 - \frac{n}{D}(2 - \frac{D-n-1}{p-n-1})) + \\ \sigma^2(1 + \frac{n}{p-n-1}) \text{ if } p \geq n+2 \end{cases} \tag{4}$$

4.1.2 Discussion of Main Results

Looking simply at equation 3 we see that the error of a predictor in a setting where $p \leq n-2$, is a factor of the noise in the model plus the norm of the features that your predictor did not take into account. This makes sense since you expect to achieve near optimal prediction on the features you did learn as $n \to \infty$ in the under-parameterized setting but you can never make any guarantees on the features you did not try to learn. For the overparameterized case where $p > n+1$ we see that the accuracy of a predictor $\hat{\beta}$ is a function of the norm of the features that you did learn and a factor of the norm of the features that you did not learn plus the noise in the system. In fact if we let $p \to \infty$ then the expression for the over-parameterized setting implies we would always predict $\hat{\beta} = 0$, since, the cost of the predictor would be $\|\beta\|^2 + \sigma^2$ and by the fact that:

$$\mathbb{E}[(y - x^T\hat{\beta})^2] = \|\beta - \hat{\beta}\|^2 + \sigma^2$$

(see appendix for the above reduction) it implies $\hat{\beta} = 0^p$, which is quite an interesting implication.

Also note that getting the second part of the previous theorem is simply a matter of noting that $\|\beta_A\| = \frac{p}{n}\|\beta\|$ and $\|\beta_{A^c}\| = (1 - \frac{p}{n})\|\beta\|$ under a random uniform selection model of the features.

4.1.3 Some curves show double descent, some not.

Overall these three equations don't seem all that intuitive but if one plots the results reported by Hsu under the prescient model and the random uniform model then we get the following pictures.

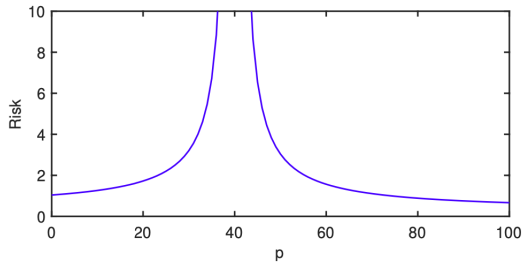


Figure 1: Plot of risk as a function of p, under the random selection model of T, Here, $\|\beta\|_2^2 = 1$,$\beta_j^2 \propto \frac{1}{j^2}$, $\sigma^2 = \frac{1}{25}$, d=100, and n=40

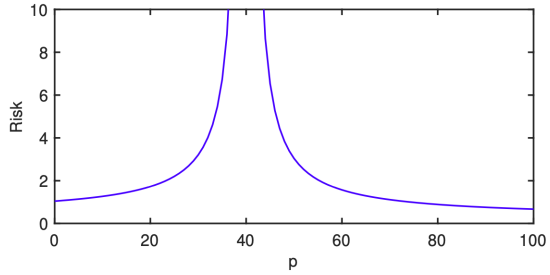


Figure 2: Plot of risk as a function of p, under the "prescient" selection model of T, Here, $\|\beta\|_2^2 = 1$,$\beta_j^2 \propto \frac{1}{j^2}$, $\sigma^2 = \frac{1}{25}$, d=100, and n=40

Though these pictures are for cherry picked values of β. In fact we can also get the figures 3 and 4 by picking different values of β. One curious thing about figures 3 and 4 is that the "prescient" model being used is always including the most signal it can by selecting an additional dimension. That is if p = 1, it picks the feature that is most important which is feature 1, if p = 2, it picks the first and the second, which are the most important, and so on. Yet even though it does that the mse only increases as it selects more features. Why is that? To help answer this, it is helpful to look at Sahai's Fourier perspective on overparameterization.

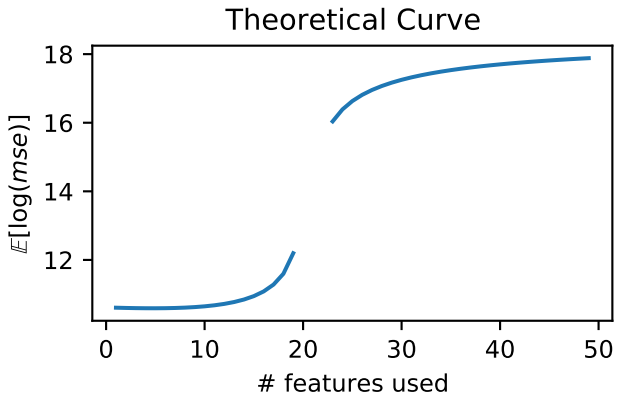


Figure 3: Plot of theoretical risk as a function of p, under a "prescient" model, where β is the vector that has value 10000 as its first component, and the rest are just a countdown from 49 to 1. Also n = 20, d = 50, $\sigma^2 = \frac{1}{25}$.

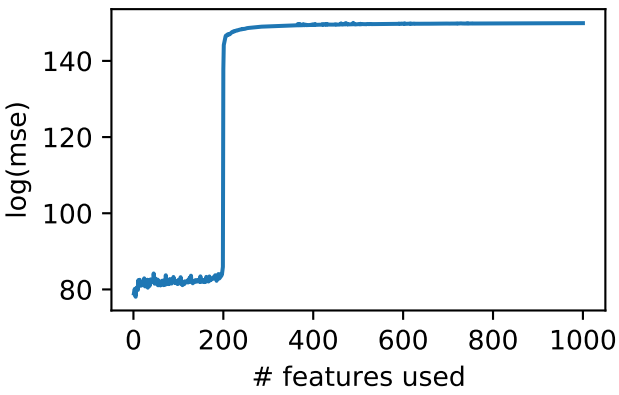


Figure 4: Plot of experimental risk as a function of p, under a "prescient" model, where β is the vector that has value 100000 as its first component, and the rest are just a countdown from 999 to 1. Also n = 200, d = 1000, $\sigma^2 = \frac{1}{25}$ and testing the predictor on 5000 points

4.2 Sahai's Paper

Sahai's setup is very similar to the one we have described before the only difference being he tackles the issue of overparameterization from a Fourier Perspective. That is we don't consider a linear prediction rule of the form $\beta^T x$ but of the form $\beta^T a(x)$ where a(x) is the frequency transformation of x. Furthermore the risk of β is evaluated with: $\mathbb{E}_{(x,y)}[(y - \beta^T x)^2] - \sigma^2$, subtracting the irreducible noise in the system. (A note, I will not be covering the main theorems that Sahai presents but only some of the intuition he gives in the Fourier perspective of how overparameterization can lead to poor generalizability or good generalizability)

4.2.1 Signal Contamination

First consider the data where $\mathbf{X}$ is just a discrete set of n regularly spaced points in the interval $[0, 1]$ and $\mathbf{Y}$ is just the constant signal 1. Now since we are working with the fourier perspective we want the estimator to interpolate the data not with a simple linear regression model but with a linear combination of Fourier features. That is we are mapping $x \to a(x)$ where to begin with we let $a(x) = [1, e^{2\pi i \frac{1}{n}}, ..., e^{2\pi i \frac{n-1}{n}}]$. Then we will create a predictor β that minimizes $L_S(\beta) = \Sigma_{i=1}^m (y_i - \beta^T a(x)_i)^2$. From the description of the signal the feature $a_0(x) = 1$ would perfectly describe the data and if we assume that the sample y is always constant then we have created the optimal predictor. But what happens if I include higher dimensional features in $a(x)$, that is I include an alias so that $a(x) = [1, e^{2\pi i \frac{1}{n}}, ..., e^{2\pi i \frac{n-1}{n}}, e^{2\pi i \frac{n}{n}}]$. In that case there would be infinitely many solutions of β that have $L_S(\beta) = \Sigma_{i=1}^m (y_i - \beta^T a(x)_i)^2$ being zero, including $\beta = [1, 0, 0..., 0]$, $\beta = [0, 0, 0..., 1]$, $\beta = [\frac{1}{2}, 0, 0..., \frac{1}{2}]$. If we were looking for the l_2 norm solution then in fact $\beta = [\frac{1}{2}, 0, 0..., \frac{1}{2}]$ is the unique solution. What does this mean? Well in an effort to minimize the l_2 norm of β the signal has bled out from $a_0(x)$ onto $a_n(x)$ and even though they both describe the $\mathbf{X}, \mathbf{Y}$, if we consider the $\mathbf{Y}$ to always be constant no matter what than including the higher frequency $a_n(x)$ harms the generalizability of β. In general if there are $d = mn$ features in $a(x)$ then in an effort to minimize the l_2 norm the signal of each $a_i(x)$ such that $i < n$ will be split among its aliases a_{i+ln} for $l = 1, ..., \frac{d}{n}$.

4.2.2 Noise absorption

Although this aliasing may be a problem for preserving the true signal of data, it may be a benefit if you want to absorb noise. For example if we consider $\mathbf{Y}$ as an n dimensional noise vector where $y \sim \mathcal{N}(0, \sigma^2)$ and we use the same $\mathbf{X}, a(\mathbf{X})$ as we described before, then what can we infer about β if we don't use aliases in comparison to when we do. Well, since the signal only has n points and we are initially using n dimensions on $a(x)$ then we can perfectly predict the data (make $\mathbf{Y} = a(\mathbf{X})\beta$) with some β. In fact if we evaluate the mse on the β that perfectly predicts the data we get:

$$\begin{aligned} \mathbb{E}_{(x,y)}[(y - \beta^T a(x))^2] - \sigma^2 = \\ \mathbb{E}_{(x,y)}[y^2 - \|\beta^T a(x)\|^2 - 2y\beta^T a(x)] - \sigma^2 = \\ \mathbb{E}_{(x,y)}[\|\beta^T a(x)\|^2] = \mathbb{E}_{(x,y)}[\|\beta\|^2] = \sigma^2 \end{aligned} \tag{5}$$

where the second to last equality was achieved by Parseval's relation, considering that the data is perfectly fit so the energy in the real domain is equivalent to the energy in the frequency domain which is in turn just β. In fact due to the isotropy of the Gaussian nose, in expectation the magnitude of each feature β_k is $\frac{\sigma^2}{n}$. Now if we include aliased features let's say $\frac{d}{n}$ aliased features for every feature in the original $a(x)$ n-dimensional vector then what happens to the original $|\beta_k|$? Well again we would have an overparameterized system, and if we were trying to get the min norm parameters, then $|\beta_k|$ would get divided into every $\frac{d}{n}$ aliases. That is if the original $|\hat{\beta}_k| = \frac{\sigma^2}{n}$, then $|\hat{\beta}_{new_k}| = \frac{\sigma^2}{d}$.

4.2.3 Discussion

Now this paints an interesting picture. In the first scenario we had a low dimesional signal that was accurately predicted with no aliases and then was contaminated when we included aliases. However, in

the second scenario we actually benefitted from using aliases, if we did not want the low dimensional frequencies to absorb all the noise in the system. This is the key idea of overparameterization, how can we use the extra parameters to absorb the noise while also preserving the low dimensional structure of the signal?

4.2.4 Signal Preservation with Weights

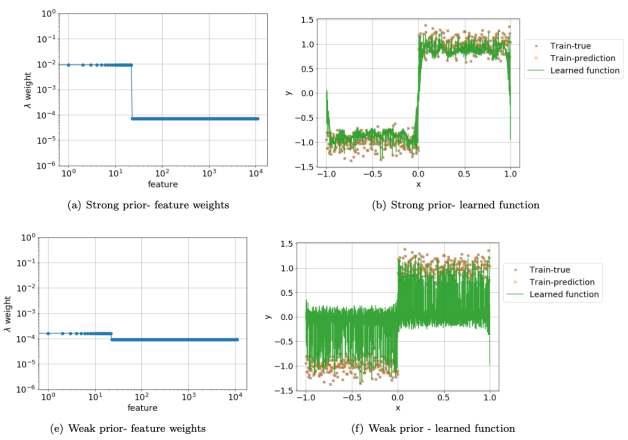
One way to do so is by imposing restrictions on the coefficients of each a_k. That is we might want a set of fourier features that interpolates the data and has some argmin condition with respect to some explicit weights w_k. More precisely, let there be a signal $(\mathbf{X}, \mathbf{Y})$ then we want:

$$\begin{aligned} \hat{\beta} = argmin_{\beta \in \mathbb{R}^d} \quad & \|\beta\|_2^2 \\ \text{s.t.} \quad & \Sigma \beta_k w_k a_k(x_j) = y_j \text{ for all j } \in [n] \end{aligned} \tag{6}$$

Which is equivalent to the following system if $\alpha_k = \frac{\beta_k}{w_k}$:

$$\begin{aligned} \hat{\alpha} = argmin_{\alpha \in \mathbb{R}^d} \quad & \Sigma_{k=0}^{d-1} \frac{\alpha_k^2}{w_k^2} \\ \text{s.t.} \quad & \Sigma \alpha_k a_k(x_j) = y_j \text{ for all j } \in [n] \end{aligned} \tag{7}$$

Now intuitively if the weight w_k is higher for smaller k then it means α_k can be larger for smaller k as well. This implies that we can have an explicit preference for the low dimensional features while also choosing high dimensional ones. This is shown very well by the figure (5) in Sahai's paper.



(a) Strong prior- feature weights (b) Strong prior- learned function

(e) Weak prior- feature weights (f) Weak prior - learned function

Figure 5: Effect of different priors on weighted l_2 norm interpolation with n = 500, d = 11000 when the true signal is the sign function.

The figure shows that when you have strong priors on low dimensional frequencies you can achieve a lot of the interpolative behavior you get from being in the overparameterized setting while still preserving the low dimensional structure of the sign function. On the other hand for the weakened priors, while you do get a lot of the interpolative behavior, the low dimensional signal of the sign function is not preserved, which is because the high dimensional frequencies are allowed to play as big a role as low dimensional ones.

4.2.5 From Fourier Features to Real features

Now the picture Sahai paints with the Fourier features is very clear, however that doesn't necessarily translate to the regular linear regression model. What we can say is that from the experiments that broke the double descent curve (where we had some true $\beta = [100000, 999, 998, ..., 1] \in \mathbb{R}^{1000}$ and $y = \beta^T x + \epsilon$ where $x \sim \mathcal{N}(0, I^{1000})$ and $\epsilon \sim \mathcal{N}(0, \frac{1}{25})$), we can observe the signal bleeding effect talked in Sahai by plotting the magnitude of the first feature being predicted and seeing that as the number of features used to make a predictor $\hat{\beta}$ increases the value $\hat{\beta}_0$ only decreases.

Now a key limitation of some of the results I presented from Sahai (he has other results that give up-

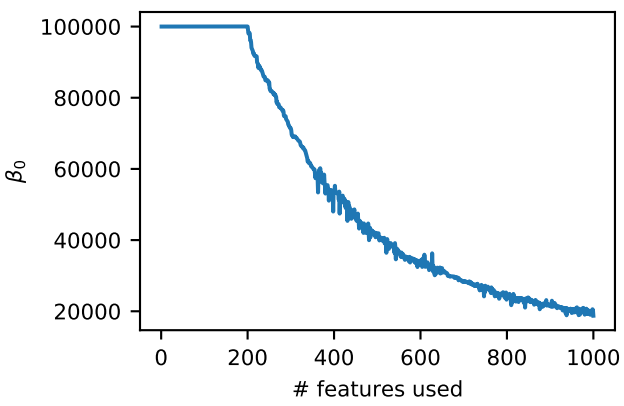


Figure 6: Signal bleed from the feature β_0 in the setup that breaks the double descent curve

per bounds on the error of interpolative solutions (not restricted to min norm) with respect to d and n) is that in order for interpolative solutions to be good there needs to be some explicit priors that dictate the behavior of these solutions. However one of the remarkable results that Bartlet achieves, is that in fact you don't need this explicit weighing to get an understanding of how accurate your min norm interpolative solution is. In fact all you need is some implicit information given by the covariance matrix of your problem's distribution.

4.3 Bartlet et. al. Work

4.3.1 Setting

Bartlett's setup is a bit different from the one I have presented and hence I will point it out in full. A linear regression problem in a separable Hilbert space H is defined by a random covariate vector $x \in \mathbb{H}$ and outcome $y \in \mathbb{R}$. Then Bartlet assumes:

- the covariance operator $\Sigma = \mathbb{E}[xx^T]$
- x and y are mean zero
- $x = V\Lambda^{\frac{1}{2}}z$, where $\Sigma = V\Lambda^{\frac{1}{2}}V^T$ is the spectral decomposition of Σ and z has components that are independent σ_x^2-subgaussian with σ_x a positive constant. That is, for all $\lambda \in \mathbb{H}$,

$$\mathbb{E}[exp(\lambda^T z)] \leq exp(\frac{\sigma_x^2||\lambda||^2}{2})$$

, where $||\cdot||$ is the norm in the Hilbert space $\mathbb{H}$

- $y - x^T\theta_*$ is σ_y^2-subgaussian, conditionally on x, that is for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}[exp(\lambda(y - x^T\theta_*))|x] \leq exp(\frac{\sigma_y^2\lambda^2}{2})$$

- almost surely, the projection of the data X on the space orthogonal to any eigenvector of Σ spans a space of dimension n.

Now the risk of an estimator θ is defined similarly to Sahai but they do not assume some distribution on the noise, hence the risk $R(\theta)$ is defined as:

$$R(\theta) = \mathbb{E}_{(x,y)}[(y - x^T\theta)^2 - (y - x^T\theta_*)^2]. \quad (8)$$

4.3.2 Main Results

Theorem 2. *For any σ_x there are $b, c, c_1 > 1$ for which the following holds. Consider a linear regression problem as defined previously. Define:*

$$k_* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$$

where the minimum of this empty set is defined to be ∞. Suppose $\delta < 1$ with $log(\frac{1}{\delta}) < \frac{n}{c}$. If $k^* \geq \frac{n}{c_1}$, then $\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}$. Otherwise,

$$R(\hat{\theta}) \leq c\left(||\theta^*||^2||\Sigma||max\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \frac{log(\frac{1}{\delta})}{n}\}\right) + clog\left(\frac{1}{\delta}\right)\sigma_y^2\left(\frac{k_*}{n} + \frac{n}{R_{k_*}(\Sigma)}\right) \quad (9)$$

with probability at least $1 - \delta$, and

$$\mathbb{E}[R(\hat{\theta})] \geq \frac{\sigma^2}{c}\left(\frac{k_*}{n} + \frac{n}{R_{k_*}(\Sigma)}\right) \quad (10)$$

Now these results all include the terms $r_k(\Sigma)$ and $R_k(\Sigma)$. So here is their definition:

$$r_k(\Sigma) = \frac{\Sigma_{i>k}\lambda_i}{\lambda_{k+1}} \tag{11}$$

$$R_k(\Sigma) = \frac{(\Sigma_{i>k}\lambda_i)^2}{\Sigma_{i>k}\lambda_i^2} \tag{12}$$

where here λ_i is the ith eigen value of Σ.

Now we are ready to interpret the results from Bartlett (I will focus on equation 9 since that is the one that is interesting to good min norm interpolators). On a surface level the theorem says that the risk of any min norm interpolator is dependent on the covariance matrix of the distribution which you are drawing from. Furthermore it says that in order to have the risk of a min norm interpolator be small, then:

$$\lim_{n\to\infty}\frac{r_0(\Sigma)}{n} = \lim_{n\to\infty}\frac{k_*}{n} = \lim_{n\to\infty}\frac{n}{R_{k_*}(\Sigma)} = 0$$

. If these conditions are met for the covariance matrix then we call it a benign matrix. Now the first and third limits don't have a really intuitive meaning but if we look at the limit with respect to k_* then we can get a picture of what a good interpolative solution will represent. In sum, k_* is the cutoff of the heavy directions in your covariance matrix. In order to realize this go back to the definition that $k_* = min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. Now observe that all the eigen values before the position k_* are so large that the sum of the ones that come after it, divided by it, is still less than a some multiple of n. Hence k_* denotes how many heavy directions there are in your covariance matrix. Now if we want $\frac{k_*}{n}$ to be very small then it means a good interpolator should lie in a low dimensional subspace and have a lot of dimensions that are meaningless and can fit noise. Which is the idea we talked about before. Now that we have explained a little bit of what the result implies we have to consider how does a covariance matrix get the three limits to go to zero?

Well a simple and unrealistic way is if we allow the matrix to change as the number of samples in your training matrix changes. Precisely we want to have the covariance matrix grow with complexity $O(nlog(n))$ and we want the eigen values to be a function of n. One case, would be where the first eigen value of this changing covariance matrix is $\lambda_1 = n^2log(n)$ and the remaining are $\lambda_i = nlog(n)$. Then:

$$r_0(\Sigma) = O(\frac{(nlog(n))(nlog(n))}{n^2log(n)}) = O(log(n)) \Rightarrow \lim_{n\to\infty}\frac{r_0(\Sigma)}{n} = 0 \tag{13}$$

. Furthermore it is easy to see that k_* will be 1, meaning the sum of all eigen values after 1 over the second eigen value will be larger than the sample size. Therefore:

$$R_{k_*}(\Sigma) = O(\frac{(n^2log^2(n))^2}{n^3log^3(n)}) = O(nlog(n)) \Rightarrow \lim_{n\to\infty}\frac{n}{R_{k_*}(\Sigma)} = 0 \tag{14}$$

. Hence we have constructed an example where the linear interpolator only gets better as n grows, though this only happens because the Covariance matrix is growing and changing with the number of points in the training set, which can't happen if the underlying data distribution for your problem is set. Now for that reason, Bartlet proves when does fixing the size of the covariate can yield optimal min norm interpolators as $n \to \infty$ (in other words Σ is benign). The result is two-fold, one is for the case Σ that lies in an infinite dimensional subspace and the second is for Σ_n lying in a finite dimensional space.

Theorem 3.

- *If* $\lambda_k(\Sigma) = k^{-\alpha}ln^{-\beta}(k+1)$, *then* Σ *is benign if and only* $\alpha = 1$ *and* $\beta > 1$
- *If*

$$\lambda_k(\Sigma_n) = \begin{cases}\gamma_k + \epsilon_n \text{ if } k \leq p_n,\\ 0 \text{ otherwise}\end{cases}$$

and $\gamma_k = \Theta(exp(\frac{-k}{\tau}))$*, then* Σ_n *is benign iff* $p_n = w(n)$ *and* $ne^{-o(n)} = \epsilon_n p_n = o(n)$*. Furthermore, for* $p_n = \Omega(n)$ *and* $\epsilon_n p_n = ne^{-o(n)}$*,*

$$R(\hat{\theta}) = O\left(\frac{\epsilon_n p_n + 1}{n} + \frac{ln(\frac{n}{\epsilon_n p_n})}{n} + max\{\frac{1}{n}, \frac{n}{p_n}\}\right)$$

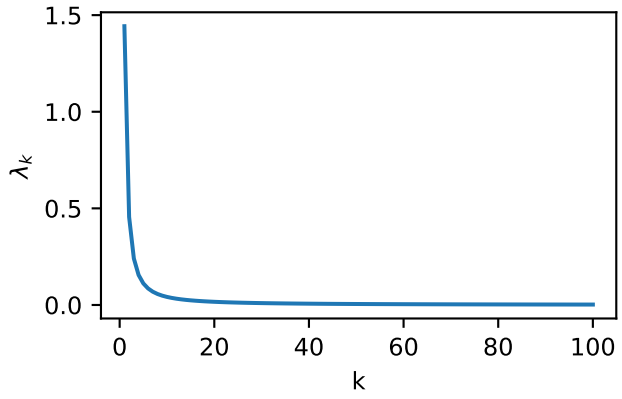


Figure 7: Plot of the eigen values of a Benign matrix in a infinite dimensional subspace.

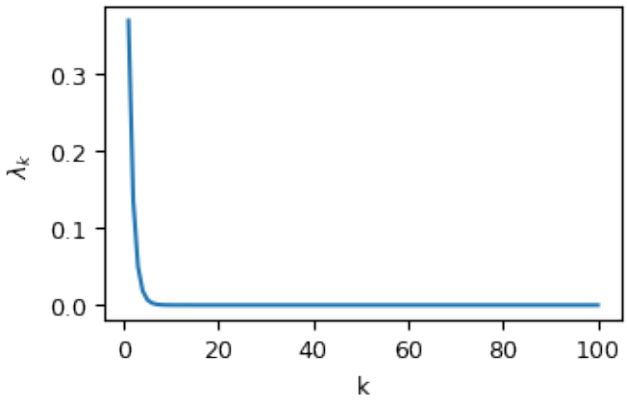


Figure 8: Plot of the eigen values of a Benign matrix in a finite dimensional subspace, where n = 60

From Figures 7 and 8 and from the formula we can observe that in order for a min norm interpolator to be accurate we need a lot of the eigen vectors in the covariance matrix to have small magnitude but not be too small either. In fact like it was talked before we want to have a small number of heavy directions but the other features should again be virtually uselesss so you can fit the noise to their directions.

Furthermore, what is interesting about Bartlet's results is that it also gives some intuition behind why the min norm risk interpolator for the counter-example of the double descent we have is not good. In our counter example the covariance matrix was simply the identity distribution, hence the eigen values of our covariate matrix were not at all distributed like the ideal benign Σ for a good min norm interpolator. Now if we ran the counter example under the same β, σ,d,n but with a covariance matrix that has the following eigenvalues:

$$\lambda_k(\Sigma) = \begin{cases} e^{-100k} + \epsilon_n \text{ if } k \leq p_n, \\ 0 \text{ otherwise} \end{cases}$$

and the features are all uncorrelated. Then we would get the following picture:

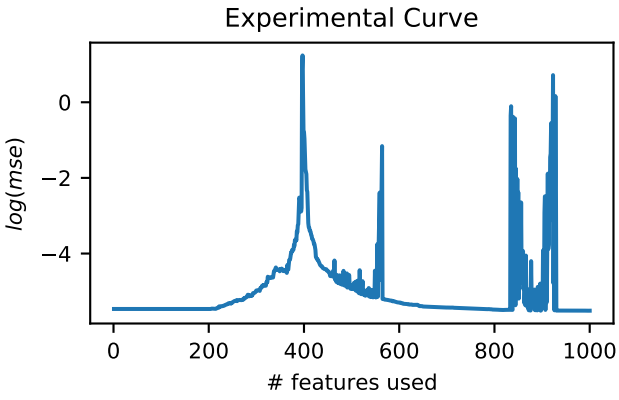


Figure 9: Running the same example as figure 4 but with a covariance matrix that would achieve a better min norm interpolator

Which shows a double descent behavior. This is not necesarily explained by Bartlet's result but that it should have a lower mse for the min norm interpolator is.

5 Other experiments

5.1 Effective Ranks of MNIST and CIFAR-10

Since Bartlett's effective rank condition can be extended to any problem (as it is only a factor of the covariance matrix, the number of samples and dimensions in your dataset), I fetched the CIFAR-10 and MNIST dataset and by considering the dataset as the entire distribution (in order to make sense of how to get the covariance matrix), plotted the first and second effective rank of the two. In doing so, I found that the scaling of the first and second effective rank that yield a good norm interpolator in regression do not explain the accuracy of ANN's in these architectures. This happens because even though the first effective rank is small compared to the sample size of both datasets, therefore the term $\frac{r_0(\Sigma)}{n}$ is small, the term $\frac{n}{R_i(\Sigma)}$ is never small since the second effective rank never exceeds 1000 even though both datasets have 10K+ samples.

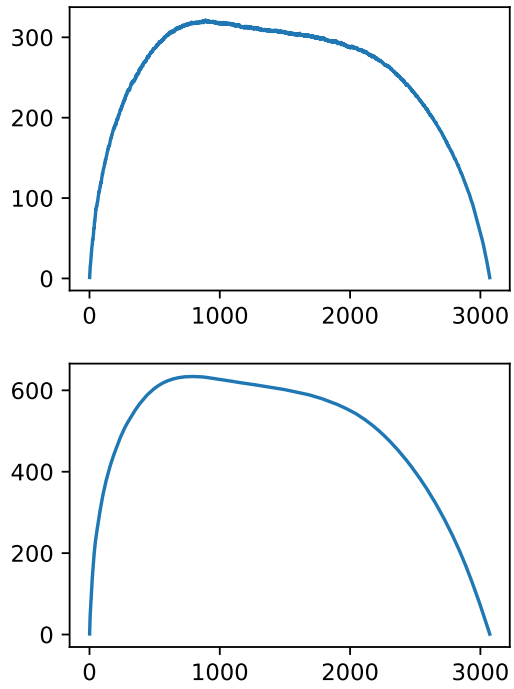


Figure 10: Plot of first and second effective rank for the cifar 10 dataset

5.2 Does double descent matter?

In the following figures (12-13), I have ran the basis of Hsu's idea but having $x \sim \mathcal{N}(0, 10I^{1000})$, n = 200, d = 1000 and $\sigma^2 = \frac{1}{25}$. The difference between each picture is only the value of β used to make y, which is that β becomes more and more independent of the first feature as the figure number increases. That is in the beginning I have $\beta = [10^6, 99.9, 99.8, ..., 1]$ but as we go down the pictures $\beta = [10^6, (4)^i * 99.9, ..., 1]$ (where the decay is uniform until 1 and it decays in 999 steps to it), for i = 1 through 6 . In doing so we see that as more weight gets distributed across the other features of β we start to see the double descent curve, which is an interesting phenomenon.

What is also interesting is to observe how lasso and ridge regression do for each of the figures (no need to look for double descent in those methods since they can be seen as feature selection models (especially lasso)). Therefore, in the figures (14-15) I have plotted the accuracy of the lasso and ridge regression where each point in the graph is correspondent

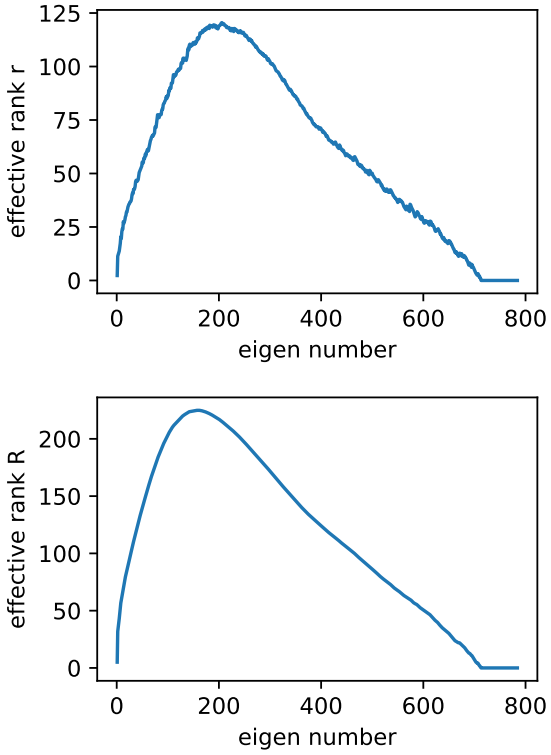
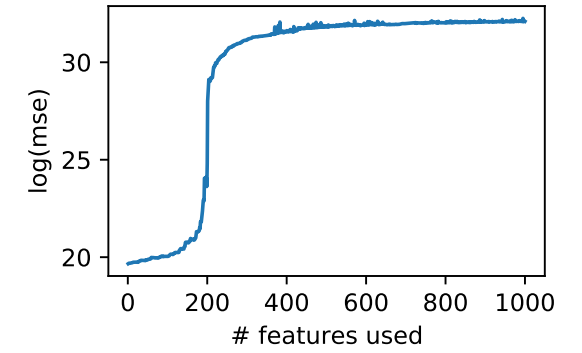
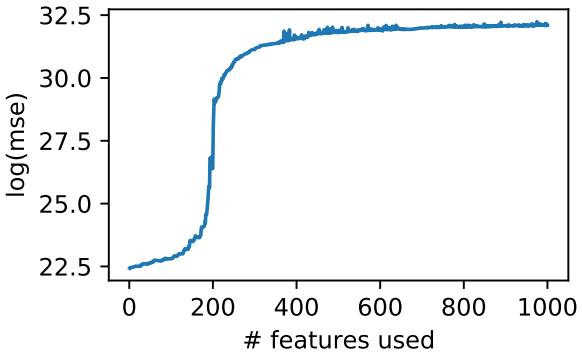
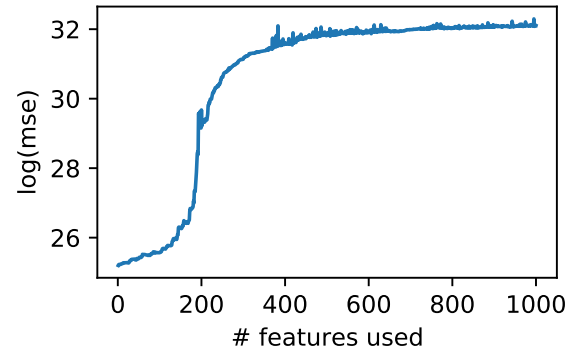


Figure 11: Plot of first and second effective rank for mnist dataset







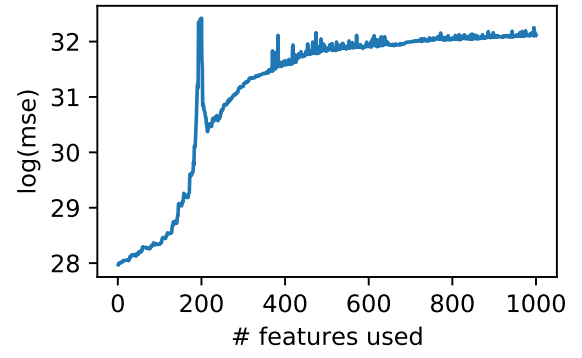
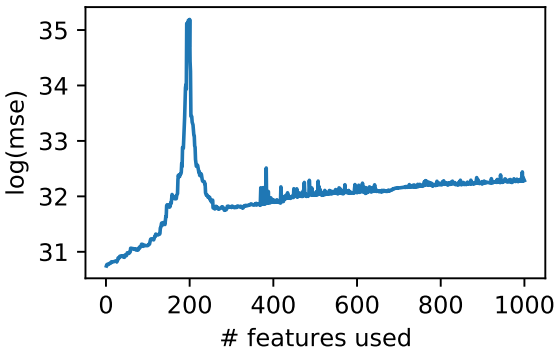
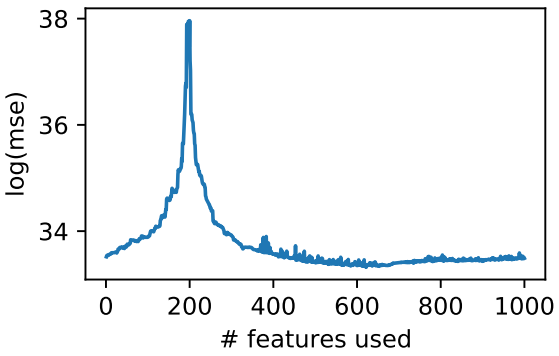


Figure 12: $x \sim \mathcal{N}(0, 10 * I^{1000})$, $\sigma^2 = \frac{1}{25}$, n = 200, β gets more independent of the first feature after each figure, meaning more signal is distributed across the other features





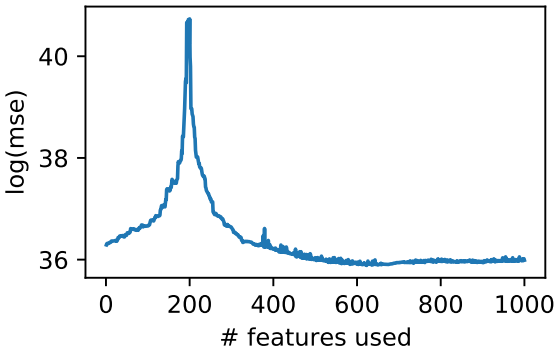


Figure 13: $x \sim \mathcal{N}(0, 10 * I^{1000})$, $\sigma^2 = \frac{1}{25}$, n = 200, β gets more independent of the first feature after each figure, meaning more signal is distributed across the other features

to the accuracy of the methods in the distribution of figures 12-13 (I have labeled the x-axis of these figures as β signal to indicate that as we move toward the right the sparsity of β decreases). Hence we can observe that although Ridge regression doesn't consistently outperform the best min norm interpolator for each of the setups, lasso regularization certainly does almost as good as every model that is explored in the double descent curve for each β. Specifically the log of the mse for lasso for each of the pictures in figures 12-13 is [19, 22, 25, 28, 31, 33 36] respectively. This shows that even though double descent is something that can happen, overparameterization in linear regression should not necessarily be the preferred method.

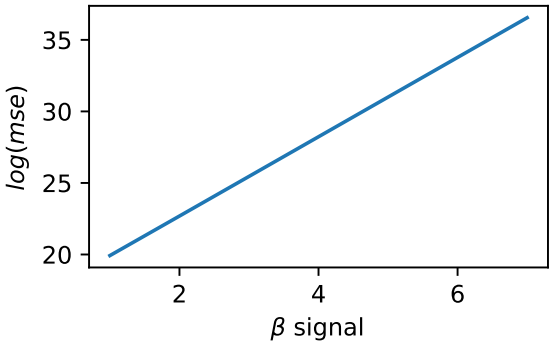


Figure 14: Lasso accuracy for the setup described in figures 12-13

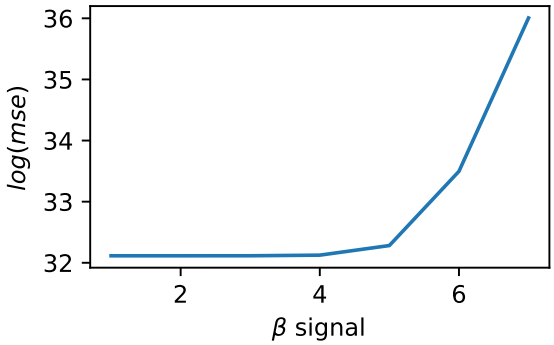


Figure 15: Ridge Regression accuracy for the setup described in figures 12-13

6 Conclusion

In this project we have covered some of the intuition and main results on why over-parameterization can be good for linear regression. The main argument being, if you have a signal that lies mostly in a low dimensional subspace and has some noise, then you can use the over-parameterization to interpolate the noise while also preserving the low dimensional characteristics. Hence why over-fitting can lead to good generalization. This view is clearly seen from Sahai et al. Fourier Perspective on linear regression and how overparameterization can lead to noise absorption, but also if you are not careful, signal bleeding. Ultimately this resulted in the need to introduce some priors in order to assert that while interpolation can happen, the low frequency features should be weighted higher than the high frequency ones. Bartlet et al. in a similar topic categorizes how this prior can be implicit by the covariance matrix. His results show that the min interpolator is dependent on two notions of effective rank of the covariance matrix and that in order for the interpolator to do well these should compete with the number of samples in some sense. Furthermore, Bartlet et al. results show that in order to achieve good mse the data should have a small number of heavy directions and a large number of insignificant directions which can be used to interpolate noise. We have also seen that under certain assumptions of gaussianity Hsu et al. can provide explicit curves of what the accuracy of an estimator $\hat{\beta}$ is with respect to the number of dimensions, number of points and true parameter β. These expressions described by Hsu can indeed show a double descent behavior under some suitable choices of β, meaning we can have good generalizability in the overfitting regime. However, this is not true for all choices of β. In fact, we can construct sparse β's where the double descent curve won't happen and in fact we will only see an ascending mse as the number of dimensions used to predict β increases. Lastly, we have also

conducted a set of experiments demonstrating how as you decrease the sparsity of a vector β the mse curve with respect to the number of features used to make a predictor $\hat{\beta}$ can go from simply ascending to a double descent curve. This is a significant argument against general over-parameterization in linear regression which is only furthered if we consider that lasso outperformed all of the over-parameterized mse regimes.

References

[1] Peter L. Bartlett et al. *Benign Overfitting in Linear Regression*. 2020. arXiv: 1906.11300 [stat.ML].

[2] Mikhail Belkin, Daniel Hsu, and Partha Mitra. *Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate*. 2018. arXiv: 1806.05161 [stat.ML].

[3] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two Models of Double Descent for Weak Features". In: *SIAM Journal on Mathematics of Data Science* 2.4 (Jan. 2020), pp. 1167–1180. ISSN: 2577-0187. DOI: 10.1137/20m1336072. URL: http://dx.doi.org/10.1137/20M1336072.

[4] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. *To understand deep learning we need to understand kernel learning*. 2018. arXiv: 1802.01396 [stat.ML].

[5] Mikhail Belkin et al. *Reconciling modern machine learning practice and the bias-variance trade-off*. 2019. arXiv: 1812.11118 [stat.ML].

[6] Koby Bibas, Yaniv Fogel, and Meir Feder. "A New Look at an Old Problem: A Universal Learning Approach to Linear Regression". In: *2019 IEEE International Symposium on Information Theory (ISIT)* (July 2019). DOI: 10.1109/isit.2019.8849398. URL: http://dx.doi.org/10.1109/ISIT.2019.8849398.

[7] Trevor Hastie et al. *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*. 2020. arXiv: 1903.08560 [math.ST].

[8] Partha P Mitra. *Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation*. 2019. arXiv: 1906.03667 [cs.LG].

[9] Vidya Muthukumar et al. *Harmless interpolation of noisy data in regression*. 2019. arXiv: 1903.09139 [cs.LG].

[10] Abraham J. Wyner et al. *Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers*. 2017. arXiv: 1504.07676 [stat.ML].

[11] Chiyuan Zhang et al. *Understanding deep learning requires rethinking generalization*. 2017. arXiv: 1611.03530 [cs.LG].

7 Appendix

7.1 Key Ideas for Hsu's Main Result on Over-parametrized Regimes

If the subset A is picked under the prescient model, and then $\hat{\beta}$ is learned, then the:

$$\mathbb{E}_{(x,y),S_n}[(y-\hat{\beta}^T x)^2] = \begin{cases} (\|\beta_{T^c}\|^2+\sigma^2)(1+\frac{p}{n-p-1}) \text{ if } p \leq n-2 \\ +\infty \text{ if } n-1 \leq p \leq n-2 \\ \|\beta_A\|^2(1-\frac{n}{p})+ \\ (\|\beta_{A^c}\|^2+\sigma^2)(1+\frac{n}{p-n-1}) \text{o.w} \end{cases} \tag{15}$$

So we are trying to evaluate:

$$\begin{aligned} \mathbb{E}_{(x,y),S_n}[(y-x^T\hat{\beta})^2] = \mathbb{E}_{(x,y),S_n}[(x^T\beta+\sigma-x^T\hat{\beta})^2] = \\ \mathbb{E}_{(x,y),S_n}[(x^T(\beta-\hat{\beta}))^2]+\sigma^2 = \\ \mathbb{E}_{(x,y),S_n}[(\beta-\hat{\beta})^T xx^T(\beta-\hat{\beta})]+\sigma^2 = \\ \mathbb{E}_{S_n}[(\beta-\hat{\beta})^T\mathbb{E}_{(x,y)}[xx^T](\beta-\hat{\beta})]+\sigma^2 = \\ \mathbb{E}_{S_n}[(\beta-\hat{\beta})^T(\beta-\hat{\beta})]+\sigma^2 = \\ \mathbb{E}_{S_n}[\|\beta-\hat{\beta}\|^2]+\sigma^2 = \\ \mathbb{E}_{S_n}[\|\beta_A-\hat{\beta}_A\|^2]+\mathbb{E}_{S_n}[\|\beta_{A^c}-\hat{\beta}_{A^c}]\|^2+\sigma^2 = \\ \mathbb{E}_{S_n}[\|\beta_A-\hat{\beta}_A\|^2]+\sigma^2+\|\beta_{A^c}\|^2 \end{aligned}$$

Now:

$$\beta_A - \hat{\beta}_A = \beta_A - (X_A^T)(X_A X_A^T)^{-1} y =$$
$$\beta_A - (X_A^T)(X_A X_A^T)^{-1}(X_A\beta_A + \eta) =$$
$$\beta_A - (X_A^T)(X_A X_A^T)^{-1}(X_A\beta_A) - (X_A^T)(X_A X_A^T)^{-1}(\eta) =$$
$$(I - X_A^T(X_A X_A^T)^{-1} X_A)\beta_A - (X_A^T)(X_A X_A^T)^{-1}(\eta)$$

where $\eta = y - X_A\beta_A$. Now observe that the first term is the projection of β_A onto the nullspace of X_A while the second term is the projection of η onto the rowspace of X_A, hence we can apply the pythagorean theorem to get:

$$\|\beta_A - \hat{\beta}_A\|^2 = \|(I - X_A^T(X_A X_A^T)^{-1} X_A)\beta_A\|^2 + \|(X_A^T)(X_A X_A^T)^{-1}(\eta)\|^2$$

Now the first term of the prior can be evaluated in expectation to be $\|\beta_A(1 - \frac{n}{p})\|^2$ due to the rotational symmetry of the Standard Normal distribution (to be honest Professor I don't really follow this step). The second term $(X_A^T)(X_A X_A^T)^{-1}(\eta)$ can be evaluated using the trace trick to show that it is almost surely $tr(\mathbb{E}[\|(X_A X_A^T)^{-1}\|^2]\mathbb{E}[\eta\eta^T])$. Since $\eta = y - X_A\beta_A = X_{A^c}\beta_{A^c} + \epsilon$. Therefore η has expected mean zero and expected covariate $\mathbb{E}[\eta\eta^T] = (\|\beta_{A^c}\|^2 + \sigma^2)I$. Now the term $tr(\mathbb{E}[\|(X_A X_A^T)^{-1}\|^2])$ has the form of the inverse Wishart Matrix where each diagonal entry follows a χ^2 distribution with p-n+1 degrees of freedom. Hence the diagonal entries of this matrix are in expectation $\frac{1}{p-n-1}$ if $p \geq n+2$ and infinity if $p \in \{n, n+1\}$. Which means:

$$\mathbb{E}[\|X_A^T(X_A X_A^T)^{-1}\eta\|^2] = \begin{cases} (\|\beta_A^c\|^2 + \sigma^2)\frac{n}{n-p-1} \text{ if} p \geq n+2 \\ \infty \text{ if } p \in \{n, n+1\} \end{cases}$$

Now combining it with β_{A^c} gives the desired expression for the overparameterized regime and where $p \in \{n-1, n+1\}$.