CSC 2532: Statistical Learning Theory Winter 2020

Project Report: Overparameterizing Polynomial Interpolation

Ameer Dharamshi *ID: 1006378180*

1 Introduction

Conventional statistical wisdom tells us that overparameterizing a model results in high variance and poor generalization. In other words, the model overfits the training data. However, recent works have demonstrated empirically that as the number of parameters approaches the interpolation limit, the population risk diverges. Interestingly, the risk descends beyond the interpolation limit and in certain cases, the global minimum is achieved in the overparameterized region. Much effort has been spent trying to explain this behaviour. Hastie et al. in [2] demonstrate the benefits of overparameterization in minimum-norm least squares regression and Ba et al. in [1] derive exact population risk quantities in two-layer neural-networks. In this work, we turn to classic interpolating functions, polynomials, to observe whether they benefit from overparameterizing. Muthukumar et al. in [3] consider an empirical investigation of Vandermonde and Legendre features and offer bounds on the test mean-squared error though they assume that the data must be generated from a linear model.

2 Set-up

Consider a target function $G(x) : \mathbb{R} \to \mathbb{R}$ from which unique data pairs (x_i, y_i) are generated with additive Gaussian noise as $y_i = G(x_i) + \epsilon_i$, where $(x_i, \epsilon_i) \overset{iid}{\sim} P_x \times \mathcal{Z}$ and $\mathcal{Z} \sim N(0, \sigma^2)$. Before discussing the model class, let's first establish the relevant notation. Let $\mathcal{F}$ be the class of polynomial functions, $\mathcal{F} = \{f(x, \theta) = \theta_1 x + \theta_2 x^2 + \cdots + \theta_d x^d = \sum_{j=1}^{d} \theta_j x^j, d \in \mathbb{N}, \theta \in \mathbb{R}^d\}$. We have d denoting the degree of the polynomial and let n be the sample size. For convenience, define the feature map $\phi_d : \mathbb{R} \to \mathbb{R}^d$ as $\phi_d(x) = [x, x^2, ..., x^d]$. We can then represent $\mathcal{F}$ as $\mathcal{F} = \{f(x, \theta) = \langle \theta, \phi_d(x) \rangle, d \in \mathbb{N}, \theta \in \mathbb{R}^d\}$.

Our objective is to select a polynomial function $f \in \mathcal{F}$ that minimizes the mean squared error loss:

$$L((x, y), f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2$$

Once we have selected the parameter vector $\hat{\theta}$ based on the polynomial feature matrix constructed with the observed data, X_ϕ, we are then interested in the asymptotic behaviour of the prediction risk for a new point $x_0 \sim P_x$ as $n, d \to \infty$. This is defined as:

$$R(\hat{\theta}) = E\Big[(G(x_0) - \langle \hat{\theta}, \phi_d(x_0) \rangle)^2 | X_\phi\Big]$$

In order to decompose the prediction risk into the usual bias and variance terms, one typically uses the properties of the inner product. However, for a general non-linear G, this is not entirely straightforward. In much of the related work, a linear model for G is assumed or the model is linearized around a random initialization on the basis that the parameters will only change slightly. Here, we take a different approach. Assume that G is an analytic function in an open interval $\mathcal{D}$ where $\{0, x_1, \ldots, x_n\} \in \mathcal{D}$ and $G(0) = 0$. As we are considering polynomial functions, rewrite the target as $G(x) = \langle \theta^*, \phi_d(x) \rangle$ where θ^* is interpreted as the Taylor coefficients as $d \to \infty$. When $\hat{\theta}$ is an unbiased estimator, the risk can be decomposed as:

$$\begin{aligned} R(\hat{\theta}) &= E\Big[(G(x_0) - \langle \hat{\theta}, \phi_d(x_0) \rangle)^2 | X_\phi\Big] = E\Big[(\langle \theta^*, \phi_d(x_0) \rangle - \langle \hat{\theta}, \phi_d(x_0) \rangle)^2 | X_\phi\Big] \\ &= E\Big[(\phi_d(x_0)^T (\theta^* - \hat{\theta}))^2 | X_\phi\Big] = E\Big[(\theta^* - \hat{\theta})^T \phi_d(x_0) \phi_d(x_0)^T (\theta^* - \hat{\theta}) | X_\phi\Big] \end{aligned}$$

In order to proceed, we require $\text{Cov}(\phi_d(x))$. We specifically excluded a bias term x_0 in the feature matrix to avoid a 0 in the first element of this covariance matrix. For now, denote $\Sigma = \text{Cov}(\phi_d(x))$ and we will return to this point after imposing a specific distribution on $x \sim P_x$. Returning to the prediction risk decomposition, define $||z||_\Sigma^2 = z^T\Sigma z$ and note the following:

$$\begin{aligned}
R(\hat{\theta}) &= E\Big[(\theta^* - \hat{\theta})^T\phi_d(x_0)\phi_d(x_0)^T(\theta^* - \hat{\theta})|X_\phi\Big] = E\Big[||\theta^* - \hat{\theta}||_\Sigma^2|X_\phi\Big] \\
&= E\Big[||\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta^*||_\Sigma^2|X_\phi\Big] = E\Big[||\hat{\theta} - E[\hat{\theta}]||_\Sigma^2 + 2(\hat{\theta} - E[\hat{\theta}])\Sigma(E[\hat{\theta}] - \theta^*) + ||E[\hat{\theta}] - \theta^*||_\Sigma^2|X_\phi\Big] \\
&= E\Big[||\hat{\theta} - E[\hat{\theta}]||_\Sigma^2|\phi\Big] + E\Big[||E[\hat{\theta}] - \theta^*||_\Sigma^2|X_\phi\Big] = ||E[\hat{\theta}|X_\phi] - \theta^*||_\Sigma^2 + E\Big[||\hat{\theta} - E[\hat{\theta}]||_\Sigma^2|X_\phi\Big] \\
&= ||E[\hat{\theta}|X_\phi] - \theta^*||_\Sigma^2 + \text{Tr}(\text{Cov}(\hat{\theta}|X_\phi)\Sigma)
\end{aligned}$$

In the above, $B(\hat{\theta}) = ||E[\hat{\theta}|X_\phi] - \theta^*||_\Sigma^2$ represents the bias and $V(\hat{\theta}) = \text{Tr}(\text{Cov}(\hat{\theta}|X_\phi)\Sigma)$ represents the variance. To compute these values, we require an explicit form for $\hat{\theta}$.

3 Parameter Estimate

Recall the empirical loss function to be optimized,

$$\min_{f\in\mathcal{F}} L((x,y),f) = \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \implies \min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (y_i - \langle\theta, \phi_d(x_i)\rangle)^2 = \frac{1}{n}(y - X_\phi\theta)^2$$

The solution to this problem depends on n and d, the dimensions of the feature matrix $X_\phi \in \mathbb{R}^{n\times d}$ as they impact the rank of the matrix. The following are the possible solutions:

1. When $d < n$, X_ϕ is full column rank and X^TX is invertible. Thus, $\hat{\theta}$ can be obtained using the standard least squares solution, $\hat{\theta} = (X^TX)^{-1}X^Ty$.
2. When $d = n$, X_ϕ is a full rank square matrix so using least squares, $\hat{\theta} = (X^TX)^{-1}X^Ty = X^{-1}y$. Unless the error term is not truly random, there is a unique degree d interpolator of the data.
3. When $d > n$, there are infinite number of interpolating polynomials. In other words, the solution to the minimization problem is no longer unique and the objective function is equal to zero at each of these minimizers. We choose to compute the minimum norm solution. That is, solve:

$$\min_{\theta\in\mathbb{R}^d} \quad \theta^T\theta \quad \text{s.t.} \quad y = X_\phi\theta$$

Constructing the Lagrangian $L(\theta,\lambda) = \theta^T\theta + \lambda^T(X_\phi\theta - y)$, taking derivatives and setting to 0 will yield $\hat{\theta} = X_\phi^T(X_\phi X_\phi^T)^{-1}y$. Notice that $X_\phi^T(X_\phi X_\phi^T)^{-1}$ is the Moore-Penrose pseudoinverse of X_ϕ. Alternatively, we can represent this by letting $(X_\phi^TX_\phi)^+ = \lim_{\lambda\to 0^+}(X_\phi^TX_\phi + \lambda I)^{-1}$ and recalling that $(X_\phi^TX_\phi)^+X^T \to X_\phi^T(X_\phi X_\phi^T)^{-1}$.

4 Gaussian Data - Vandermonde Features

Assuming that $x_i \sim P_x = N(0,1)$, notice that $\Sigma = \text{Cov}(\phi_d(x))$ is:

$$\Sigma = \begin{bmatrix} E[X^2] & E[X^3] & E[X^4] & \dots & E[X^{d+1}] \\ E[X^3] & E[X^4] & E[X^5] & \dots & E[X^{d+2}] \\ E[X^4] & E[X^5] & E[X^6] & \dots & E[X^{d+3}] \\ \vdots & & & \ddots & \vdots \\ E[X^{d+1}] & E[X^{d+2}] & E[X^{d+3}] & \dots & E[X^{2d}] \end{bmatrix} = \begin{bmatrix} 1 & 0 & 3 & \dots & E[X^{d+1}] \\ 0 & 3 & 0 & \dots & E[X^{d+2}] \\ 3 & 0 & 15 & \dots & E[X^{d+3}] \\ \vdots & & & \ddots & \vdots \\ E[X^{d+1}] & \dots & \dots & \dots & (2d-1)!! \end{bmatrix}$$

Equipped with Σ, let's first consider the underparameterized case. The bias is:

$$B(\hat{\theta}) = ||E[\hat{\theta}|X_\phi] - \theta^*||_\Sigma^2 = ||(X_\phi^T X_\phi)^{-1} X_\phi^T E[y|X_\phi] - \theta^*||_\Sigma^2 = ||(X_\phi^T X_\phi)^{-1} X_\phi^T X_\phi \theta^* - \theta^*||_\Sigma^2 = 0$$

and the variance is:

$$\begin{aligned} V(\hat{\theta}) &= \operatorname{Tr}(\operatorname{Cov}(\hat{\theta}|X_\phi)\Sigma) = \operatorname{Tr}(\operatorname{Cov}((X_\phi^T X_\phi)^{-1} X_\phi^T y|X_\phi)\Sigma) \\ &= \operatorname{Tr}((X_\phi^T X_\phi)^{-1} X_\phi^T \operatorname{Cov}(\epsilon) X_\phi (X_\phi^T X_\phi)^{-1}\Sigma) = \sigma^2 \operatorname{Tr}((X_\phi^T X_\phi)^{-1}\Sigma) \end{aligned}$$

There are two challenges in computing an exact expression for the variance. First, Σ is not an easy to handle matrix. While all elements are guaranteed to be finite, it is far from the identity matrix. Second, the entries of X_ϕ are not iid. While each $x_i \overset{iid}{\sim} N(0,1)$, the features generated by $\phi_d(x_i)$ are not independent. In addition to the above challenges, we will observe in Section 6 that we do not observe benefits of overparameterization on the plain Vandermonde polynomial features. However, in the next section we will reconfigure the feature map to generate a better conditioned Σ.

5 Uniform Data - Legendre Features

In this section, we will assume that $x_i \sim P_x = U(-1,1)$. We will also make a change to the feature map. Instead of the ill conditioned pure Vandermonde features, we will define a new map $\nu_d(x) : \mathbb{R} \to \mathbb{R}^D$ as $\nu_d(x) = \left[\sqrt{\frac{3}{2}}p_1(x), \sqrt{\frac{5}{2}}p_2(x), \ldots, \sqrt{d+\frac{1}{2}}p_d(x)\right]^T$ where $p_j(x)$ is the jth Legendre polynomial evaluated at $x \in [-1,1]$. The reasoning behind this choice of polynomials is that the Legendre polynomials form an orthogonal system and have 0 mean. That is, when $x_i \sim P_x = U(-1,1)$, $p(x) = \frac{1}{2}$ and:

$$\begin{aligned} \int_{-1}^{1} p_j(x)p(x)dx &= \frac{1}{2}\int_{-1}^{1} p_j(x)dx = 0 \\ \int_{-1}^{1} p_i(x)p_j(x)p(x)dx &= \frac{1}{2}\int_{-1}^{1} p_i(x)p_j(x)dx = 0 \quad \text{for } i \neq j \\ \int_{-1}^{1} p_j(x)p_j(x)p(x)dx &= \frac{1}{2}\int_{-1}^{1} p_j(x)p_j(x)dx = \frac{1}{j+\frac{1}{2}} \end{aligned}$$

For our choice of constants multiplied to the Legendre polynomials, the final integral reduces to $\frac{1}{2}$. Now, putting these pieces together, we see that $\Sigma = \operatorname{Cov}(\nu_d(x)) = \frac{1}{2}I_d$. To further understand the risk decomposition, we must first replace all instances of X_ϕ with X_ν.

Now, in the underparameterized regime, once again the bias is:

$$B(\hat{\theta}) = ||E[\hat{\theta}|X_\nu] - \theta^*||_\Sigma^2 = ||(X_\nu^T X_\nu)^{-1} X_\nu^T E[y|X_\nu] - \theta^*||_\Sigma^2 = ||(X_\nu^T X_\nu)^{-1} X_\nu^T X_\nu \theta^* - \theta^*||_\Sigma^2 = 0$$

and the variance is:

$$\begin{aligned} V(\hat{\theta}) &= \operatorname{Tr}(\operatorname{Cov}(\hat{\theta}|X_\nu)\Sigma) = \operatorname{Tr}(\operatorname{Cov}((X_\nu^T X_\nu)^{-1} X_\nu^T y|X_\nu)\Sigma) \\ &= \frac{1}{2}\operatorname{Tr}((X_\nu^T X_\nu)^{-1} X_\nu^T \operatorname{Cov}(\epsilon) X_\nu (X_\nu^T X_\nu)^{-1} I_d) = \frac{\sigma^2}{2}\operatorname{Tr}((X_\nu^T X_\nu)^{-1} X_\nu^T X_\nu (X_\nu^T X_\nu)^{-1}) \\ &= \frac{\sigma^2}{2}\operatorname{Tr}((X_\nu^T X_\nu)^{-1}) = \frac{\sigma^2}{2n}\operatorname{Tr}(\hat{\Sigma}^{-1}) = \frac{\sigma^2}{2n}\sum_{i=1}^{d}\frac{1}{s_i} = \frac{\sigma^2 d}{2n}\sum_{i=1}^{d}\frac{1}{ds_i} \end{aligned}$$

where $n\hat{\Sigma} = X_\nu^T X_\nu$ is the empirical covariance matrix and s_i are the eigenvalues of $\hat{\Sigma}$. Given that the rows of X_ν are iid, isotropic, the Uniform distribution is log-concave, and that the use of polynomial features

ensures that $X_\nu^T X_\nu$ is invertible when $d \leq n$, we have the Marchenko-Pastur theorem for isotropic x_i from [6] and [5]. If we can assume that $s_{\min} \geq \frac{(1-\sqrt{d/n})^2}{2}$, this is the same as the underparameterized setting in Hastie et. al [2] aside from the factor of $\frac{1}{2}$. Thus we find that $R(\hat{\theta}) = V(\hat{\theta}) = \frac{\sigma^2\gamma}{2-2\gamma}$ where $\gamma = d/n$.

Turning to the overparameterized setting where $d > n$ and $\hat{\theta} = (X_\phi^T X_\phi)^+ X^T y$, notice that the Σ-norm is the l_2-norm as $\Sigma = I_d$. Denote $r^2 = ||\theta^*||_2^2$. Then, the bias is:

$$\begin{aligned} B(\hat{\theta}) &= ||E[\hat{\theta}|X_\nu] - \theta^*||_2^2 = ||X_\nu^T(X_\nu X_\nu^T)^{-1}X_\nu\theta^* - \theta^*||_2^2 = ||\theta^*(I_d - X_\nu^T(X_\nu X_\nu^T)^{-1}X_\nu)||_2^2 \\ &= \theta^{*T}(I_d - X_\nu^T(X_\nu X_\nu^T)^{-1}X_\nu)(I_d - X_\nu^T(X_\nu X_\nu^T)^{-1}X_\nu)\theta^* = \theta^{*T}(I_d - X_\nu^T(X_\nu X_\nu^T)^{-1}X_\nu)\theta^* \\ &= ||\theta^*||_2^2 - (X_\nu\theta^*)^T(X_\nu X_\nu^T)^{-1}(X_\nu\theta^*) \leq ||\theta^*||_2^2 = r^2 \end{aligned}$$

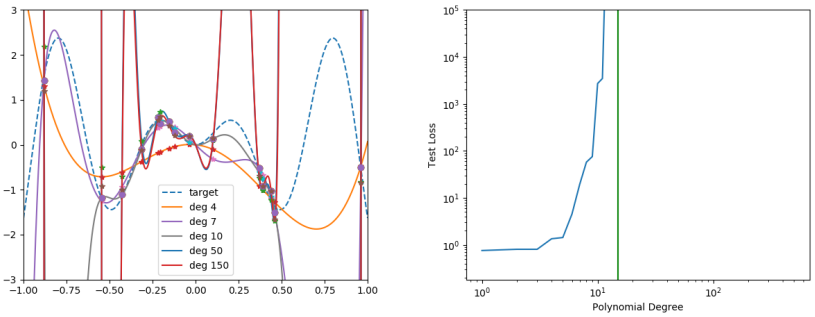
where the inequality follows from the fact that $X_\nu X_\nu^T$ is positive definite. In Hastie et. al [2], the authors demonstrate that r^2 also exhibits behaviour following a double descent pattern. and that we expect the upper bound on $B(\hat{\theta}_)$ to decay as γ increases. Finally, the overparamaterized variance is:

$$\begin{aligned} V(\hat{\theta}) &= \mathrm{Tr}(\mathrm{Cov}(\hat{\theta}|X_\nu)\Sigma) = \mathrm{Tr}(\mathrm{Cov}(X_\nu^T(X_\nu X_\nu^T)^{-1}y|X_\nu)\Sigma) \\ &= \frac{1}{2}\mathrm{Tr}(X_\nu^T(X_\nu X_\nu^T)^{-1}\mathrm{Cov}(\epsilon)(X_\nu X_\nu^T)^{-1}X_\nu I_d) = \frac{\sigma^2}{2}\mathrm{Tr}(X_\nu^T(X_\nu X_\nu^T)^{-1}(X_\nu X_\nu^T)^{-1}X_\nu) \\ &= \frac{\sigma^2}{2}\mathrm{Tr}((X_\nu X_\nu^T)^{-1}) = \frac{\sigma^2}{2d}\sum_{i=1}^{n}\frac{1}{t_i} = \frac{\sigma^2 n}{2d}\sum_{i=1}^{n}\frac{1}{nt_i} \end{aligned}$$

where t_i are the eigenvalues of of $\frac{1}{d}X_\nu X_\nu^T$. Under the same assumptions as the $d < n$ case, this is the isotropic features overparameterized case of Hastie et. al [2] and so $V(\hat{\theta}) = \frac{\sigma^2}{2\gamma-2}$ and $R(\hat{\theta}) \leq r^2 + \frac{\sigma^2}{2\gamma-2}$. In summary, when $d < n$, the risk diverges as $d \to n$ and after, we see the risk descend as d/n increases.

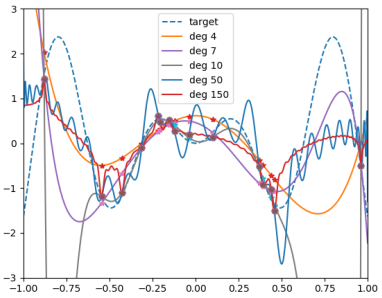
6 Empirical Evidence

Consider the data generating target function $G(x) = 3x\sin(10x)$. We generate 15 points from $G(x)$ and fit polynomials of varying degrees using each feature map. Then we examine the average test risk as an estimate of the prediction risk. Figure 1 demonstrates that pure polynomial features fail to generalize. We attribute this to the inability to control the covariance matrix of the Vandermonde features. The test loss plot shows that prediction risk diverges before reaching the interpolation limit and does not decline when the polynomial degree exceeds sample size.

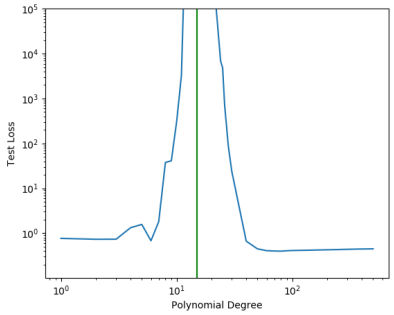


(a) Sample Models (b) Test Loss

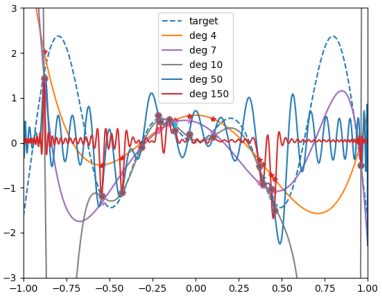
Figure 1: Vandermonde Features



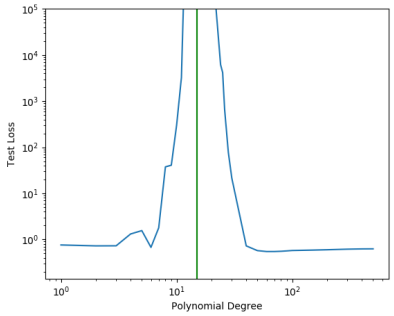
(a) Sample Models - Legendre



(b) Test Loss - Legendre



(c) Sample Models - Adjusted Legendre



(d) Test Loss - Adjusted Legendre

Figure 2: Legendre Features Example

On the other hand, in Figure 2, we see the desired behaviour. The high degree polynomials constructed with Legendre features do tend to generalize well. As $d \to n$, the test loss diverges before descending as the polynomial degree greatly exceeds the sample size. Notice that this is present in the unadjusted and adjusted Legendre features. In the adjusted Legendre features, the interpolating polynomials tend to concentrate around 0 and deviate close to the observations with the amount of fluctuation increasing as the degree decreases. In summary, our empirical results match the theoretical analysis in that we observe the double descent curve when using Legendre polynomial features but not with Vandermonde features.

References

[1] Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. "Generalization of Two-layer Neural Networks: An Asymptotic Viewpoint". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1gBsgBYwH.

[2] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*. 2019. arXiv: 1903.08560 [math.ST].

[3] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. *Harmless interpolation of noisy data in regression*. 2019. arXiv: 1903.09139 [cs.LG].

[4] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. *Deep Double Descent: Where Bigger Models and More Data Hurt*. 2019. arXiv: 1912.02292 [cs.LG].

[5] Alain Pajor and Leonid Pastur. *On the Limiting Empirical Measure of the sum of rank one matrices with log-concave distribution*. 2007. arXiv: 0710.1346 [math.PR].

[6] Pavel Yaskov. *The necessary and sufficient conditions in the Marchenko-Pastur theorem*. 2015. arXiv: 1511.02711 [math.PR].