Double Descent Notes*

Dominic Richards †

October 31, 2019

1 Motivation

In section we provide motivation behind the primary work reviewed in these notes [MM19].

1.1 Classical Understanding of Bias Variance Trade off

Given a class of models and a collection of training data, the objective in many machine learning settings is to find a model in that class that achieves low *Generalisation Error*[1], that is, error on predicting future observations. One common approach is *empirical risk minimisation*, which chooses the model within the class that achieves minimum error on the training data.

Given this approach, a standard technique in statistics is to decompose the *Generalisation Error* into two error terms which are controlled by the size of the model class. These two terms are commonly referred to as the *Bias* and *Variance*. The *Bias* is *typically decreasing* with the size of the model class, and represents the error from *under-fitting* the training data, specifically, the inability of models in the class to capture aspects the data that directly linked to its generative process. Meanwhile, the *Variance* is *typically increasing* with the model class size, and represents the error from models *over-fitting* the training data, namely, capturing spurious patterns (or noise) in the training data which are not inherently linked to predicting future observations. Having the Generalisation Error dominated by either one of these terms, *Bias* or *Variance*, is commonly referred to as respectively, *under-fitting* or *over-fitting*, the training data. The common wisdom is the existence of a *sweet spot* in model class size that balances the under-fitting and over-fitting, and minimises the generalisation error [FHT01]. Looking to Figure 1 a), we see a cartoon of this trade off.

*Some of these notes took inspiration from Andrea Montanari's presentation at the "Geometry of Deep Learning" work shop, a video of which can be found here `https://youtu.be/FCmUtpkDk-I?t=3727`

†Dominic.Richards@spc.ox.ac.uk

[1]We will commonly interchange between Generalisation Error, Test Error, Prediction Error, Test Risk and Prediction Risk. Whilst there is a distinct difference between some of the these quantities, namely, error is typically risk minus the minimum risk, these terms are meant to refer to performance of a model on unseen observations.

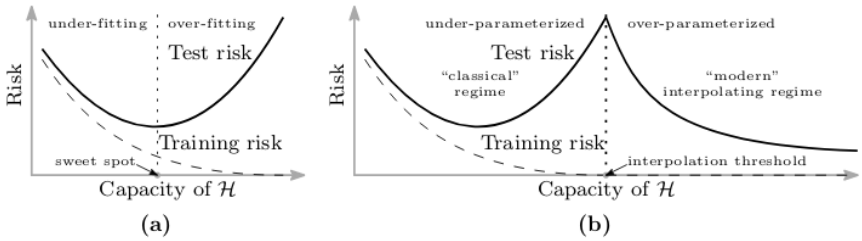

Figure 1: Picture from [BHMM18]. Test Risk (or Generalisation Error) plotted against Capacity of $\mathcal{H}$ (size of model class) (a) The classical trade off in under-fitting and over-fitting. (b) The *double descent risk*, which includes the classical regime to the left of the interpolation threshold. Beyond the *Interpolation threshold*, the Generalisation Error continues to decrease.

The size of the model class in this context can be controlled *explicitly* by, for instance, choosing the neural network architecture, or *implicitly* by regularising the problem through, for instance, early stopping when using gradient descent to the fit the model to training data. This goes alongside the conventional wisdom which states that achieving zero training error is a property of overfitting. For instance consider the following quote from [FHT01]:

> "... a model with zero training error is overfit to the training data and will typically generalize poorly" [FHT01, page 221].

The conventional wisdom alongside the Bias/Variance trade off just described, has recently been revisited in the context of overparameterised models like deep neural networks, which are able to achieve low generalisation error whilst simultaneously achieving very low or zero training error [ZBH$^+$16, BMM18]. Such methods often being referred to as "interpolators", since they achieve little to no training error, and thus, interpolate the training data. This has spurred a large number of recent works investigating the statistical guarantees of interpolation methods [MM19, BLLT19, HMRT19, BMM18, BHMM18, BRT18, RZ18, LR18, AS17]. The focus of this note will be on [MM19], as well as a few works surrounding it, which claims to give rigours proof, in a simplified setting, that the minimum generalisation error is achieved in the *overparameterised* regime. This has been observed previously and called the double descent curve, which we will now go on to describe.

1.2 Double Descent Curve

One motivation for the work [MM19] was a series of observations made by [BMM18, BHMM18], of a *Double Descent* phenomena in the Generalisation Error as the model class size is increased beyond *Interpolation Threshold* for neural networks and decision trees. The *Interpolation Threshold* being the point at which the complexity of the model class is sufficiently large to achieve zero training loss e.g. number of parameters

larger than number of data points. A cartoon of this is given Figure (1) (b), and a real example with a fully connected neural network can be seen in Figure 2.

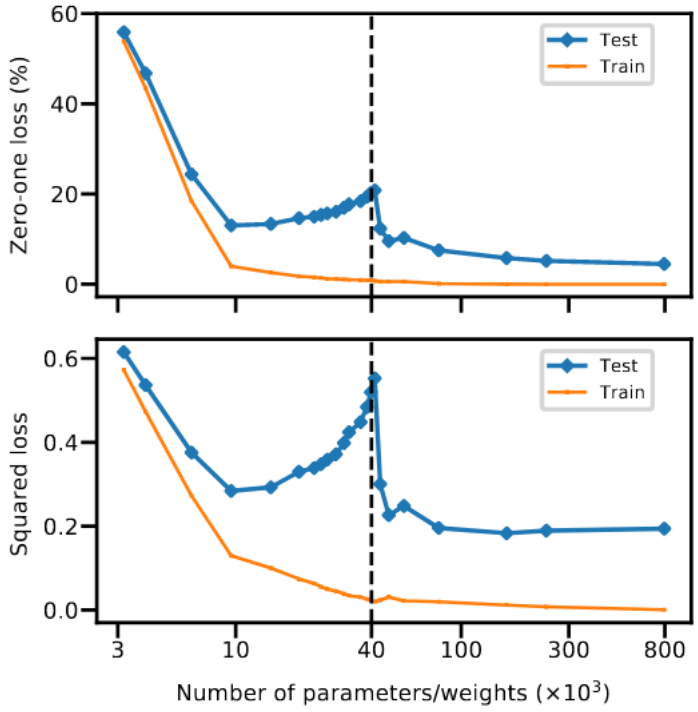

Figure 2: Taken from [BHMM18]. *Double Descent* curve for fully connect single layer neural network on MNIST. Plot of Risk (Generalisation Error) against number of parameters. Black dotted line indicates *Interpolation Threshold*.

As expected, when the model class size is smaller than the *interpolation threshold*, we have the classical Bias/Variance trade off, with a peak in the error at the interpolation threshold. Meanwhile, most strikingly, as the model class size continues to increase beyond the interpolation threshold, the Generalisation/Prediction/Test error then decreases to a minimum. These observations are summarised in the following points:

O1): Peak in Generalisation error at the interpolation threshold

O2): Global minimum in generalisation error after interpolation threshold

O3): Monotone decreasing in the generalisation after interpolation threshold

O4): Vanishing explicit regularisation for Neural Network example [2]

One explanation for double descent phenomena, that is larger model classes generalise better, was put forward by [BHMM18] who stated that

> "Choosing the smoothest function that perfectly fits observed data is a form of Occam's razor: the simplest explanation compatible with the observations should be preferred (cf. [Vap13] [BEHW87]). By considering larger function classes, which contain more candidate predictors compatible with the data, we are able to find interpolating functions that have smaller norm and are thus 'simpler'. Thus increasing function class capacity improves performance of classifiers." - [BHMM18]

A point here is that, given two models that perfectly fit the data, we would always opt to choose the simpler one. As we will see later, the criteria of model simplicity in the context of squared loss with linear predictors is naturally the minimum norm solution.

The work [MM19] sets out to consider a *Neural Network model* that describes the double descent phenomna. That is, in the idealised setting, we would study directly what is done in practice: Stochastic Gradient Descent applied to a deep neural network. However, this is likely quite challenging, and therefore, we study a simplified setting (a *model*) which reproduces the observations **O 1-4)**. In this note we refer to the simplified settings as *Neural Network models*, since they only aim to model what occurs in practice.

2 Neural Network Models

In this section aim to give an introduction into Neural Network models, in particular those considered within [MM19].

2.1 Notation

Consider the canonical statistical learning problem, we are given independent and identically distributed (i.i.d) pairs $(y_i, \mathbf{x}_i)_{i \leq n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector and $y_i \in \mathbb{R}$ is a label or response variable. We aim to construct a function f which allows us to predict future responses. As eluded to previously, the performance of a predictor is measured in terms of its generalisation error or risk $R(f) := \mathbf{E}[(y - f(\mathbf{x}))^2]$ where the expectation is with respect to the observations $(y, \mathbf{x})$.

The models that we will be interested in understanding will be neural networks. These are define by composing linear functions as well as what are common called activation functions. The simplest model in this class is given by two-layer networks

[2] A form of implicit or algorithmic regularisation still occurs, i.e. least norm solution for squared loss or gradient descent. With this point, we mean there no explicit regularisation such as Tikhonov regularisation.

(NN):

$$\mathcal{F}_{NN} := \Big\{ f(\mathbf{x}; \mathbf{a}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle) : a_i \in \mathbb{R}, \boldsymbol{\theta}_i \in \mathbb{R}^d \forall i \leq N \Big\},$$

where the matrix of parameters $\boldsymbol{\Theta} \in \mathbb{R}^{N \times d}$ has the ith row $\boldsymbol{\Theta}_i = \boldsymbol{\theta}_i$, and the vector $\mathbf{a} = (a_1, \ldots, a_N) \in \mathbb{R}^N$. To be precise, the above defines a Neural Network with N neurons and $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function. We now go over two particular instances of Neural Network Models which simplify the above class by linearising it.

2.2 Linearised Neural Networks

In place of directly studying the class of functions produced by $\mathcal{F}_{NN}$ it has been popular recently to study two classes of models which can be regarded as linearizations of two-layer networks.

2.2.1 Random Features Model

The random features model of [RR08] only optimises over weights $\mathbf{a}$, while the second layer $\{\boldsymbol{\theta}_i\}_{i=1,\ldots,N}$ is chosen at random. This class of functions will therefore be parameterised the matrix $\boldsymbol{\Theta} \in \mathbb{R}^{N \times d}$ leading to

$$\mathcal{F}_{RF}(\boldsymbol{\Theta}) := \Big\{ f(\mathbf{x}; \mathbf{a}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle) : a_i \in \mathbb{R} \forall i \leq N \Big\}.$$

2.3 Neural Tangent Kernel

A more recent model is the Neural Tangent Kernel of [JGH18]. Similar to the Random Features model, the set of weights $\boldsymbol{\Theta}$ are chosen at random. Moreover, the model can be seen as a Taylor expansion of the Neural Network model around a random initalisation. With σ' denoting the derivative activation function, we have

$$\mathcal{F}_{NT}(\boldsymbol{\Theta}) := \Big\{ f(\mathbf{x}; (\mathbf{a}_i)_{i \leq N}, \boldsymbol{\Theta}) = \sum_{i=1}^{N} \langle \mathbf{a}_i, \mathbf{x} \rangle \sigma'(\langle \boldsymbol{\theta}_i, \mathbf{x}_i \rangle) : \mathbf{a}_i \in \mathbb{R}^d \forall i \leq N \Big\}.$$

3 Double Descent with Random Features

In this section we summarise one of the main results of [MM19], which is to show that the Random feature linearised neural network model class $\mathcal{F}_{RF}$ (appropiately rescaled) exhibits a double descent phenomena when the when the number of neurons N and number of samples n go to infinity alongside the dimension d, and the data is assumed to be generated from a particular parameteric model.

3.1 Setup

Begin by denoting the sphere of radius r in d dimensions by $\mathbb{S}^{d-1}(r)$. The objective is to learn a function $f_d \in L^2(\mathbb{S}^{d-1}(\sqrt{d}))$ where L^2 denotes the standard space of square integrable functions. Assume the data $(\mathbf{x}_i, y_i)_{i\leq n}$ is generated i.i.d from a joint distribution such that the features are sampled uniformly from the sphere $\mathbf{x}_i \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$ and the responses are set $y_i = f_d(\mathbf{x}_i) + \epsilon_i$, where the noise ϵ_i is independent of $\mathbf{x}_i$. The noise distribution satisfies $\mathbf{E}_\epsilon[\epsilon_i] = 0$, and $\mathbf{E}_\epsilon[\epsilon_i^2] = \tau^2$. We now overload notation and introduce rescaled variant of the Random Features Neural Network Model class from previously

$$\mathcal{F}_{RF}(\mathbf{\Theta}) := \Big\{ f(\mathbf{x}; \mathbf{a}, \mathbf{\Theta}) = \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle / \sqrt{d}) : a_i \in \mathbb{R} \forall i \leq N \Big\}.$$

where we have now divided the inner product within the activation function by $\sqrt{d}$. We recall that the parameters $\mathbf{\Theta}$ will be chosen randomly and independently of the data. It will be assumed that $\|\boldsymbol{\theta}_i\|_2 = \sqrt{d}$, which justifies the factor $\frac{1}{\sqrt{d}}$ in the above expression, yielding $\langle \boldsymbol{\theta}_i, \mathbf{x}_j \rangle / \sqrt{d}$ of order one.

The coefficients left to choose in the Random Features Neural Network Model given above are $\mathbf{a} \in \mathbb{R}^d$. These will be learned by performing ridge regression with penalisation $\lambda \geq 0$ on the data points $(\mathbf{x}_i, y_i)_{i\leq n}$ to yield

$$\hat{\mathbf{a}}(\lambda) = \text{argmin}_{\mathbf{a}\in\mathbb{R}^N} \Big\{ \frac{1}{n} \sum_{j=1}^{n} \big(y_j - \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{\theta}_i, \mathbf{x}_j \rangle) / \sqrt{d} \big)^2 + \frac{N\lambda}{d} \|\mathbf{a}\|_2^2 \Big\}.$$

Going back to the key points, in particular **O4)**, we will be interested in the ridgeless limit $\lambda \to 0$. Finally, we recall we are interested in the generalisation/test/prediction error. We denote this quantity for a particular function f_d, training data $\mathbf{X} = (\mathbf{x}_i)_{i\leq n}$, noise $\epsilon = (\epsilon_i)_{i\leq n}$, and random features $\mathbf{\Theta} = (\boldsymbol{\theta}_i)_{i\leq N}$:

$$R_{RF}(f_d, \mathbf{X}, \mathbf{\Theta}, \lambda) = \mathbf{E}_{\mathbf{x}} \Big[\Big(f_d(\mathbf{x}) - f(\mathbf{x}; \hat{\mathbf{a}(\lambda)}, \mathbf{\Theta}) \Big)^2 \Big],$$

where expectation is over a new fresh feature $\mathbf{x} \sim \mathbb{S}^{d-1}(\sqrt{d})$. Note the above is a *random* quantity depending on the data, noise and random features. The error above is then studied in the following setting

- The random features are uniformly and independently distributed on the sphere: $\boldsymbol{\theta}_i \sim \text{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$.
- N, n, d lie in a proportional asymptotic regime. Namely, $N, n, d \to \infty$ with $N/d \to \psi_1$ and $n/d \to \psi_2$ for some $\psi_1, \psi_2 \in (0, \infty)$
- Assume a linear model $f_d(\mathbf{x}) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \mathbf{x} \rangle$, where $\boldsymbol{\beta}_{d,1} \in \mathbb{R}^d$ is arbitrary with $\|\boldsymbol{\beta}_{d,1}\|_2^2 = F_1^2$

We see that $\psi_1/\psi_2 = N/n$ characterises the degree of over-parameterisation, and thus, the size of the complexity of the model class as described at the beginning.

Remark 1 (Other models and Limitations of Linearised Neural Networks) *We note that the work [MM19] also consider a more general non-linear model* $f_d(\mathbf{x}) = \beta_{d,0} + \langle \boldsymbol{\beta}_{d,1}, \mathbf{x} \rangle + f_d^{NL}(\mathbf{x})$ *where the nonlinear component* $f_d^{NL}(\mathbf{x})$ *is a centered isotropic Gaussian process indexed by* $\mathbf{x} \in \mathbb{S}^{d-1}(\sqrt{d})$. *The linear model is motivated by [GMMM19], which shows when* $N = O(d^{\ell-1})$ *for some* $\ell > 1$ *then the Random Feature model does not out perform linear regression over all monomials of degree at most* ℓ *in* $\mathbf{x}$. *As such, when* f_d *is non-linear as previously, the RF model only fits the linear component, while the non-linear component* f_d^{NL} *gets interpreted as noise (see Remark 5 in [MM19] for a precise statement of this). We note it is a similar case for the Neural Tangent Kernel. For details see [GMMM19].*

3.2 Main Result

In this section one of the main results of [MM19] is presented. That is to give precise asymptotic bounds for $R_{RF}(f_d, \mathbf{X}, \boldsymbol{\Theta}, \lambda)$ as $d \to \infty$ in terms of the norm of the signal $\|\boldsymbol{\beta}_{d,1}\|_2^2 = F_1^2$, the ratio of network size to dimension $N/d \to \psi_1$ and ratio of the sample size to dimension $n/d \to \psi_2$ as well as the noise τ^2. This is summarised within the following Theorem. It is as follows.

Theorem 1 ([MM19]) *Let* $\sigma : \mathbb{R} \to \mathbb{R}$ *be weakly differentiable, with* σ' *be a weak derivative of* σ. *Assume* $|\sigma(u)|, |\sigma'(u)| \leq c_0 e^{c_1|u|}$ *for some constants* $c_0, c_1 < \infty$. *Define the parameters for a standard normal random variable* $G \sim \mathcal{N}(0,1)$

$$b_1 = \mathbf{E}[G\sigma(G)], \quad b_\star^2 = \mathbf{E}[\sigma(G)(G)] - \mathbf{E}[\sigma(G)]^2 - b_1^2, \quad \zeta = \frac{b_1^2}{b_\star^2}$$

Then, for any $\lambda > 0$ *we have*

$$R_{RF}(f_d, \mathbf{X}, \boldsymbol{\Theta}, \lambda) = \underbrace{F_1^2 \mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/b_\star^2)}_{\textit{Bias}} + \underbrace{\tau^2 \mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/b_\star^2)}_{\textit{Variance}} + o_{d,\mathbf{P}}(1)$$

where $\mathscr{B}$ *and* $\mathscr{V}$ *are defined in Definition 1.*

We now highlight some observations about the above result.

- The prediction error R_{RF} has been written in terms of two functions which align with the bias $\mathscr{B}(\zeta, \psi_1, \psi_2, \lambda/b_\star^2)$ and variance $\mathscr{V}(\zeta, \psi_1, \psi_2, \lambda/b_\star^2)$ as well as high order terms than go to zero as $d \to \infty$ in high probability[3] i.e. $o_{d,\mathbf{P}}(1)$. The above Theorem can therefore be seen as a asymptotic bound in the high-dimensional limit.
- The functions $\mathscr{B}$ and $\mathscr{V}$ depend on: properties of the activation function σ (through ζ), the number of parameters Neural network normalised by the dimension $\psi_1 = N/d$, the number of samples normalised by the dimension $\psi_2 = n/d$ and the regularisation $\lambda/b_\star^2$.
- The functions $\mathscr{B}$ and $\mathscr{V}$ take a complicated form, the explicit expression of which is given later in Definition 1. They can be evaluated and plotted using a computer.

[3] We have $h_1(d) = o_{d,\mathbf{P}}(h_2(d))$, if $h_1(d)/h_2(d)$ converges to 0 in probability.

As Theorem 1 gives an asymptotic bound for the Generalisation Error/ prediction risk, we can now investigate it to gain insights into how error depends on the overparameterisation and number of samples i.e. ψ_1, ψ_2. As eluded to previously, looking to Figure 3 we see an alignment between the theoretical predictions from Theorem 1 and what is observed by Random feature regression. That is the double descent curve as well as points **O1-4)** are predicted by Theorem 1. Given the theoretical predictions we can

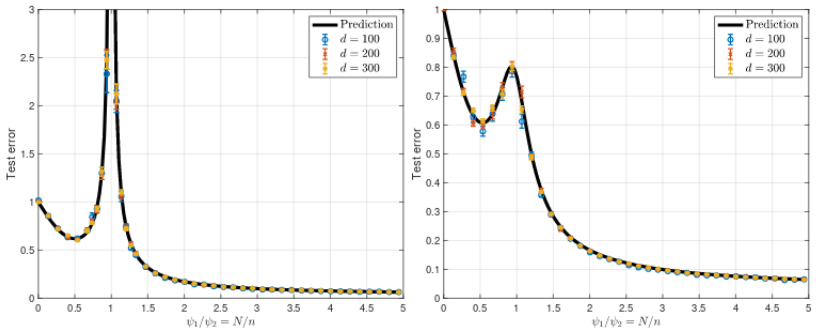

Figure 3: Plot from [MM19]. Random features ridge regression with ReLU activation ($\sigma = \max\{x, 0\}$). Data are generated via $y_i = \langle \boldsymbol{\beta}_1, \mathbf{x}_1 \rangle$ (zero noise) with $\|\boldsymbol{\beta}\|_2^2 = 1$, $\psi_2 = n/d = 3$. Left frame: $\lambda = 10^{-8}$. Right frame: $\lambda = 10^{-3}$. Black line is theoretical prediction, colored symbols are numerical results for several dimensions d.

now search for other insights about these models. These will be summarised within the following subsection.

3.3 Insights

In this section we summarise a number of additional insights given by Theorem 1. This will include some comparisons to prior literature, which help show what the primary contributions of [MM19].

3.3.1 Double Descent in Bias

The double descent curve has been theoretical shown to appear in the linear regression model of [AS17, HMRT19, BHX19]. Together these works show a number of the key features of the double descent curve, namely those summarised by observations **O 1, 3, 4)**. For some details on these references see Section 4. Although, it is claimed that these models did not show observation **O 2)**, that is, *the minimum in Generalisation error was not achieved in the highly overparameterised regime*, that is, without ad-hoc model mispecification. This is in contrast to [MM19] which show that the minimum is achieved in the overparametertised regime.

The authors of [MM19] state the reason that their model has minimum generalisation error in the highly overparameterised regime is due to Bias term $\mathscr{B}$ also monotonically decreasing after the interpolation threshold. The Variance term was already

characterised in [HMRT19] and shown to be monotonically decreasing after the interpolation threshold.

3.4 Regularisation for low Signal to Noise

Define the Signal to Noise ratio (SNR) as F_1^2/τ^2, that is, the ratio of predictor norm $\|\boldsymbol{\beta}_{d,1}\|_2^2 = F_1^2$ and the noise covariance $\mathbf{E}_{\boldsymbol{\epsilon}}[\epsilon^2] = \tau^2$. Then looking to Figure 4 the Test/prediction/generalisation Error is plotted as a function of the regularisation parameter λ for a high (left) and low (right) signal to noise levels. Observe that in both cases of high and low signal to noise the minimum in Test Error is achieved in the overparameterised regime $\psi_1/\psi_2 = N/n \to \infty$. In the case of high signal to noise the error is minimised at vanishing regularisation $\lambda \to 0$, meanwhile for low signal to noise the minimum is achieved at a non-zero regularisation λ. The existence of such a cutoff in the signal to noise is given in Proposition 5.4 of [MM19]. Intuitively, theory states that if the signal to noise is too low then you should include explicit regularisation.

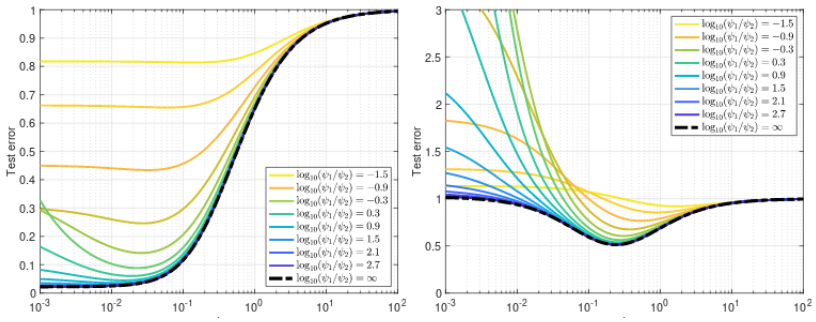

Figure 4: Plot from [MM19]. Analytical predictions for the test error of learning a linear function with $\|\boldsymbol{\beta}_1\|_2^2 = 1$ using random features with ReLU activation function $\sigma(x) = \max\{x, 0\}$. Rescaled sample size is fixed to $\psi_2 = n/d = 10$. Different curves are for different values of the number of neurons $\psi_1 = N/d$. On the left: high SNR $1/\tau^2 = 5$, on the right low SNR $1/\tau^2 = 1/10$. Each line aligns with different level of overparameterisation $\psi_1/\psi_2 = N/n$.

3.5 Detailed Theoretical Results

In this section we present the more detail decomposition of the Bias and Variance terms given in Theorem 1. The analysis in [MM19] heavily utilises random matrix theory, in particular fixed point equations for the Stieltjes transform. We begin with the following Definition. Within the following for any complex number $z \in \mathbb{C}_+$ we denote the imaginary part by $\Im(z)$ and the real part by $\Re(z)$.

Definition 1 *Let the functions $\nu_1, \nu_2 : \mathbb{C}_+ \to \mathbb{C}_+$ be uniquely defined by the following conditions: (i) ν_1, ν_2 are analytic on $\mathbb{C}_+$; (ii) For $\Im(\xi) > 0$, $\nu_1(\xi), \nu_2(\xi)$ satisfy the*

following equations

$$\nu_1 = \psi_1\Big(-\xi - \nu_2 - \frac{\zeta^2\nu_2}{1-\zeta^2\nu_1\nu_2}\Big)$$
$$\nu_2 = \psi_2\Big(-\xi - \nu_1 - \frac{\zeta^2\nu_1}{1-\zeta^2\nu_1\nu_2}\Big)$$

(iii) $(\nu_1(\xi), \nu_2(\xi))$ *is the unique solution of these equations with* $|\nu_1(\xi)| \leq \psi_1/\Im J(\xi)$, $|\nu_2(\xi)| \leq \psi_2/\Im J(\xi)$ *for* $\Im J(\xi) > C$, *with* C *a sufficiently large constant. Let*

$$\mathcal{X} \equiv \nu_1(\mathbf{i}(\psi_1\psi_2\overline{\lambda})^{1/2}) \cdot \nu_2(\mathbf{i}(\psi_1\psi_2\overline{\lambda})^{1/2})$$

and

$$\begin{aligned}
\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda}) &\equiv -\mathcal{X}^5\zeta^6 + 3\mathcal{X}^4\zeta^4 + (\psi_1\psi_2 - \psi_2 - \psi_1 + 1)\mathcal{X}^3\zeta^6 - 2\mathcal{X}^3\zeta^4 - 3\mathcal{X}^3\zeta^2 \\
&\quad (\psi_1 + \psi_2 - 3\psi_1\psi_2 + 1)\mathcal{X}^2\zeta^4 + 2\mathcal{X}^2\zeta^2 + \mathcal{X}^2 + 3\psi_1\psi_2\mathcal{X}\zeta^2 - \psi_1\psi_2 \\
\mathcal{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda}) &\equiv \psi_2\mathcal{X}^3\zeta^4 - \psi_2\mathcal{X}^2\zeta^2 + \psi_1\psi_2\mathcal{X}\zeta^2 - \psi_1\psi_2 \\
\mathcal{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda}) &\equiv \mathcal{X}^5\zeta^6 - 3\mathcal{X}^4\zeta^4 + (\psi_1 - 1)\mathcal{X}^3\zeta^6 + 2\mathcal{X}^3\zeta^4 + 3\mathcal{X}^3\zeta^2 + (-\psi_1 - 1)\mathcal{X}^2\zeta^4 - 2\mathcal{X}^2\zeta^2 - \mathcal{X}^2.
\end{aligned}$$

We then define

$$\underbrace{\mathscr{B}(\zeta,\psi_1,\psi_2,\lambda/b_\star^2)}_{\textit{Bias}} \equiv \frac{\mathcal{E}_1(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}$$
$$\underbrace{\mathscr{V}(\zeta,\psi_1,\psi_2,\lambda/b_\star^2)}_{\textit{Variance}} \equiv \frac{\mathcal{E}_2(\zeta,\psi_1,\psi_2,\overline{\lambda})}{\mathscr{E}_0(\zeta,\psi_1,\psi_2,\overline{\lambda})}$$

Without delving into complex analysis, the above gives a formula for the Bias and Variance in terms of two fixed point equation. The fact that we go to complex analysis is as a result of using Stieltjes transform to investigate limiting empirical spectral distribution of the random matrices involved. An investigation into the analytical techniques involved would likely require its own set of notes and reading group. For the brave soul who wishes to go deeper, a pair of useful references on random matrix theory are: Terrance Tao's blog post on the semi-circular law `https://terrytao.wordpress.com/2010/02/02/254a-notes-4-the-semi-circular-law/`; as well as the following monograph [TV$^+$04].

4 Related Literature

In this section we go back and survey the literature surrounding [MM19]. Each subsection will be dedicated to summarising the results of a single work.

4.1 "Two models of Double Descent for Weak Features" [BHX19]

The work [BHX19] gives two simple settings in which some of the features **O 1-4)** of the double descent curve arise. In this note we will focus on the simple noiseless linear regression example.

4.1.1 Setup

They consider a simple regression model where a response y is equal to a linear function $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ of a d real-valued variable $\mathbf{x} = (x_1, \ldots, x_d)$:

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \sigma\epsilon = \sum_{j=1}^{d} x_j \beta_j.$$

Once again following the standard leaning setting, we are given n i.i.d copies $(\mathbf{x}_i, y_i)_{i \leq n}$. Although we only fit a linear model using a subset $P \subseteq [d] := \{1, \ldots, d\}$ of $p := |P|$ variables.

Let the matrix of features be denoted $\mathbf{X} \in \mathbb{R}^{n \times d}$ so that the ith row aligns with the covariates for the ith sample $\mathbf{X}_i = \mathbf{x}_i$. Similarly let the vector of responses be denoted $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$. For a subset $A \subset [d]$ and a $d-$dimensional vector $\mathbf{v}$, we use $\mathbf{v}^A := (v_j : j \in A)$to denote the $|A|-$dimensional subvector of entries from A. Similarly for a matrix we have $\mathbf{X}^A \in \mathbb{R}^{n \times |A|}$ whose ith row are the covariates for sample i associated to A i.e. $\mathbf{X}_i^A = \mathbf{x}_i^A$. For $A \subseteq [d]$, we denote its complement by $A^c := [d] \backslash A$. Finally, $\| \cdot \|_2$ denotes the Euclidean norm.

We will assume that the features $\mathbf{x}$ follow a standard normal distribution in $\mathbb{R}^d$. The objective is to produce an estimate $\boldsymbol{\beta}'$ such that $\boldsymbol{\beta}'_{P^c} = 0$ that minimises the out of sample prediction error

$$R(\boldsymbol{\beta}') = \mathbf{E}[(y - \mathbf{x}^\top \boldsymbol{\beta}')^2] = \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_2^2 = \|\boldsymbol{\beta}_{P^c}\|_2^2 + \|\boldsymbol{\beta}_P - \boldsymbol{\beta}'_P\|_2^2$$

where we used $\mathbf{E}[(y - \mathbf{x}^\top \boldsymbol{\beta}')^2] = \mathbf{E}[(\mathbf{x}^\top(\boldsymbol{\beta} - \boldsymbol{\beta}'))^2] = \mathbf{E}\big[\operatorname{Tr}\big((\boldsymbol{\beta} - \boldsymbol{\beta}')^\top \mathbf{x}\mathbf{x}^\top(\boldsymbol{\beta} - \boldsymbol{\beta}')\big)\big]$, passing expectation into the trace and the feature covariance is the identity $\mathbf{E}[\mathbf{x}\mathbf{x}] = I$. The second equality comes from splitting the euclidean norm across the co-ordinates that we have access to P and those that are hidden P^c.

The estimator considered will be the *simplest* (in the ℓ_2 sense) vector minimising the squared loss on the training data over the P covariates given. This estimator can succinctly written as

$$\widehat{\boldsymbol{\beta}}^P := (\mathbf{X}^P)^\dagger \mathbf{y}, \quad \widehat{\boldsymbol{\beta}}^{P^c} := 0 \tag{1}$$

The symbol † denotes the Moore-Penrose pseudoinverse. In the classical setting where $p \leq n$ and $(\mathbf{X}^P)^\top(\mathbf{X}^P)$ is invertible, the minimiser of the empirical loss i.e. $\beta \to \|\mathbf{X}^P\beta - \mathbf{y}\|_2^2$ is unique and found by the derivative to zero and setting $\widehat{\boldsymbol{\beta}}_P = ((\mathbf{X}^P)^\top \mathbf{X}^P)^{-1}(\mathbf{X}^P)^\top \mathbf{y}$. We arrive at (1) above by then using the identity involving the Moore-Penrose pseudoinverse.[4] In the case $p \geq n$, there are multiple vectors β that minimise the empirical loss, specifically achieve zero training loss. As such we pick the simplest, that is, the one with the smallest euclidean norm $\|\beta\|_2^2$. The vector in this case is exactly aligns with $(\mathbf{X}^P)^\dagger \mathbf{y}$.[5] We now consider the risk $\mathbf{E}[R(\widehat{\boldsymbol{\beta}})]$, under the assumption that $\mathbf{x}$ is sampled from a standard Guassian distribution.

[4]For a matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ its Moore-Penrose pseudoinverse can be written as $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^\dagger \mathbf{A}^\top$ So when $\mathbf{A}^\top \mathbf{A}$ is invertible we have $\mathbf{A}^\dagger = (\mathbf{A}^\top \mathbf{A})^{-1}\mathbf{A}^\top$, where we note when a matrix is invertible the Moore-Penrose inverse aligns with the standard inverse.

[5]For under deter minded linear system $z \to \mathbf{A}z = \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{n \times p}, \mathbf{b} \in \mathbb{R}^n$, the minimum norm solution i.e. $\mathbf{z}^\star = \arg\min_{\mathbf{z}:\mathbf{A}\mathbf{z}=\mathbf{b}} \|z\|_2^2$ is unique and defined by $\mathbf{z}^\star = \mathbf{A}^\dagger \mathbf{b}$.

4.1.2 Risk Analysis

We begin with the following Theorem that gives the expected risk for different regimes of p.

Theorem 2 ([BHX19]) *Assume the distribution of* $\mathbf{x}$ *is a standard normal in* $\mathbb{R}^d$ *and* $y = \mathbf{x}^\top \boldsymbol{\beta}$ *for some* $\boldsymbol{\beta} \in \mathbb{R}^d$*. Pick any* $p \in \{0, \ldots, d\}$ *and* $P \subset [d]$ *of cardinality* p*. The risk of* $\widehat{\boldsymbol{\beta}}$ *as defined in* (1) *is*

$$\mathbf{E}[R(\widehat{\boldsymbol{\beta}})] = \begin{cases} \|\boldsymbol{\beta}^{P^c}\|_2^2\Big(1 + \frac{p}{n-p-1}\Big) & \text{if } p \leq n-2 \\ +\infty & \text{if } n-1 \leq p \leq n+1 \text{ and } \boldsymbol{\beta}^{P^c} \neq 0 \\ \|\boldsymbol{\beta}^{P}\|_2^2\Big(1 - \frac{n}{p}\Big) + \|\boldsymbol{\beta}^{P^c}\|_2^2\Big(1 + \frac{n}{p-n-1}\Big) & \text{if } p \geq n+2 \\ \|\boldsymbol{\beta}^{P}\|_2^2 \max\big\{1 - \frac{n}{p}, 0\big\} & \text{if } \boldsymbol{\beta}^{P^c} = 0 \end{cases}$$

Specifically, the above Theorem characterises how the error depends upon the true predictor $\boldsymbol{\beta}$ as well as the subset of predictors P given. How the above changes with respect to the number of parameters p, naturally depends on how P is assumed to be chosen in conjunction with the true predictor $\boldsymbol{\beta}$. The work [BHX19] considers two models for how P is chosen: Uniformly at random from $[d]$; and a "prescient" selection model where by the largest coefficients of $\boldsymbol{\beta}$ are chosen first.

Random Subset: Suppose that P is a subset of $[d]$ of size p chosen uniformly at random. In this case we have

$$\mathbf{E}[\|\boldsymbol{\beta}^{P}\|_2^2] = \frac{p}{d}\|\boldsymbol{\beta}\|_2^2, \quad \mathbf{E}[\|\boldsymbol{\beta}^{P^c}\|_2^2] = \Big(1 - \frac{p}{d}\Big)\|\boldsymbol{\beta}\|_2^2.$$

Plugging these into Theorem 1 and assuming no co-ordinates of $\boldsymbol{\beta}$ are zero (so event $\boldsymbol{\beta}^{P^c} = 0$ has zero probability) we arrive at the risk

$$\mathbf{E}[R(\widehat{\boldsymbol{\beta}})] = \|\boldsymbol{\beta}\|_2^2 \times \begin{cases} \Big(1 - \frac{p}{d}\Big)\Big(1 + \frac{p}{n-p-1}\Big) & \text{if } p \leq n-2 \\ +\infty & \text{if } n-1 \leq p \leq n+1 \text{ and } \boldsymbol{\beta}^{P^c} \neq 0 \\ \Big(1 - \frac{n}{d}\Big(2 - \frac{d-n-1}{p-n-1}\Big)\Big) & \text{if } p \geq n+2 \end{cases}$$

Now when $p \leq n-2$ we see that the risk bound above is increasing upto the *interpolation threshold* (p=n), after which the risk decreases with p. The risk is smallest at $p = d$. Figure 5 plots the above risk for a range of p. As we see the double descent phenomna is observed.

Prescient choice: The authors also consider the infinite dimensional case $d = \infty$ where by $\beta_j = \frac{1}{j}$ for $j \geq 1$ and $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$. In this setting they suppose that P is the largest p co-ordinates of $\boldsymbol{\beta}$ i.e. the first p. The risk curve as a function of p in this case is very much different case, as can be seen in Figure 6. In this case the number of features that minimises the risk is within the classic $p \leq n$ regime.

4.1.3 Discussion

The work considers two simplified settings (only one focused on here) which show the double descent phenomena. The authors state that when features are chosen in

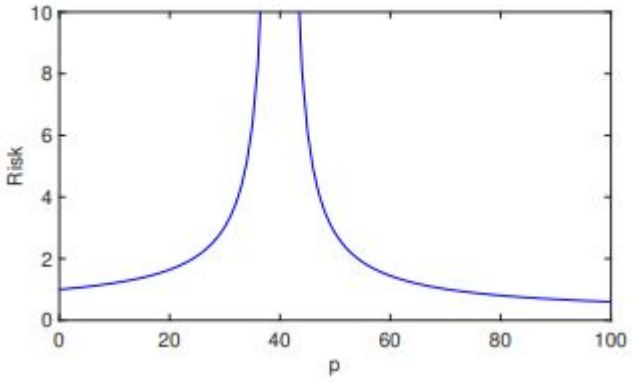

Figure 5: Plot from [BHX19]. Plot of risk $\mathbf{E}[R(\widehat{\boldsymbol{\beta}})]$ as a function of p, under the random selection model of P. Here $\|\boldsymbol{\beta}\|_2^2 = 1$, $d = 100$ and $n = 40$

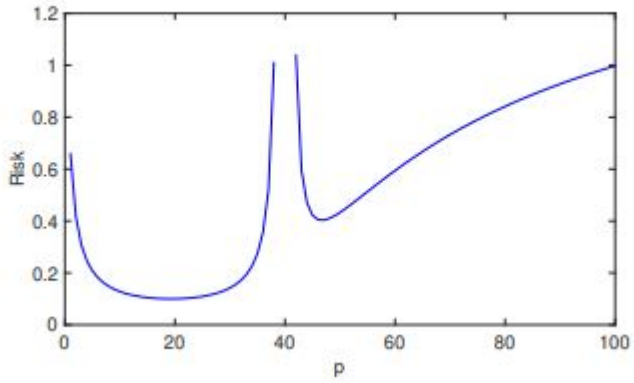

Figure 6: Plot from [BHX19]. Plot of risk $\mathbf{E}[R(\widehat{\boldsymbol{\beta}})]$ as a function of p, under the "prescient" model of P. Here $\|\boldsymbol{\beta}\|_2^2 = \pi^2/6$, $d = \infty$ and $n = 40$. As $p \to \infty$ the risk approaches $\|\boldsymbol{\beta}\|_2^2$ from below

an uninformed manner, it maybe optimal to choose as many as possible. In contrast, in scenarios where informed selection of features is possible, choosing the number of features that balance the bias and variance can be better then incurring the costs that come with using all of the features.

4.2 "Surprises in High-Dimensional Ridgeless Least Squares Interpolation" [HMRT19]

The work [HMRT19] also considers a simple linear regression setting, but with the dimension and number of samples going to infinity in a proportional regime $d/n \to \gamma \in (0, \infty)$. In this setting the degree of over parameterisation is characterised by γ, and they are able to describe a number of the double descent curve properties.

4.2.1 Setup

Once again suppose we have n independently and identically distributed samples $(\mathbf{x}_i, y_i)$ that are generated with $\boldsymbol{\beta} \in \mathbb{R}^d$ for $i = 1, \dots, n$

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$$

where ϵ_i is independent noise such that $\mathbf{E}[\epsilon_i] = 0$ and $\mathbf{E}[\epsilon_i^2] = \sigma^2$. Suppose that the feature vectors are generated according to a standard d-dimensional Gaussian distribution $\mathbf{x}_i \sim \mathcal{N}(0, I)$ for $i = 1, \dots, n$. The matrix of feature vectors will be denoted $\mathbf{X} \in \mathbb{R}^{n \times d}$ whose ith row aligns with the sample i i.e. $\mathbf{X}_i = \mathbf{x}_i$, and the vector of responses as $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$ The quantity of interest in this case is once again the sample prediction which will be defined for a fresh sample $\mathbf{x} \sim \mathcal{N}(0, I)$

$$R_X(\boldsymbol{\beta}') = \mathbf{E}[(\mathbf{x}^\top \boldsymbol{\beta}' - \mathbf{x}^\top \boldsymbol{\beta})^2 | \mathbf{X}]$$

where we have conditioned the data matrix $\mathbf{X}$, therefore the above is a random quantity depending upon it.

The estimator considered will once again be the (least norm in the case it is not unique) parameter minimising the squared loss on the training data. That is

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}'} \|\mathbf{X}\boldsymbol{\beta}' - \mathbf{y}\|_2^2 = (\mathbf{X}^\top \mathbf{X})^\dagger \mathbf{X}^\top \mathbf{y},$$

where $\dagger$ again denotes the Moore-Penrose psudoinverse.

4.2.2 Double Descent in Linear Model

We now go on to summarise the results from [HMRT19], which describe many aspects of the double descent curve.

Theorem 3 ([HMRT19]) *Consider the setting as described in Section 4.2.1 with $\|\boldsymbol{\beta}\|_2^2 = r^2$. Then as $d, n \to \infty$ such that $d/n \to \gamma$, then almost surely*

$$R_X(\widehat{\boldsymbol{\beta}}) \to \begin{cases} \sigma^2 \frac{\gamma}{1-\gamma} & \text{if } \gamma < 1 \\ \sigma^2 \frac{1}{\gamma - 1} + r^2\left(1 - \frac{1}{\gamma}\right) & \text{if } \gamma > 1 \end{cases}$$

We now make a number of observations about the above Theorem.

- The two cases of Theorem 3 align with the under parameterised regime $\gamma < 1$ as well as the over parameterised regime $\gamma > 1$.

- In the under parameterised regime the error is all variance, as it scales with the noise variance σ^2, and diverges at $\gamma = 1$
- In the over parameterised regime the error composed a variance term, scaling with the noise variance σ^2, as well as a bias term which scales with the norm of the underlying parameter r. The variance is now decreasing in the degree of overparameterisation γ, meanwhile the bias is increasing in the amount of over parameterisation.

A plot the above for the risk in Theorem 3 has then been given in Figure 7. Looking to the properties of the double descent curve highlighted at the start of this note **O 1-4)**, we see a number of them are present. Namely, **O 1)**: a peak is seen at the interpolation threshold ($\gamma = 1$), and **O 4)**: we are in a regime with no explicit regularisation. Although, there is only a monotone decrease in the error after the interpolation threshold when there is a large Signal to Noise ratio, and as such, does not satisfy observation **O 3)**. Moreover, the minimum in generalisation error is not achieved in the over parameterised regime, thus not satisfying observation **O 2)**.

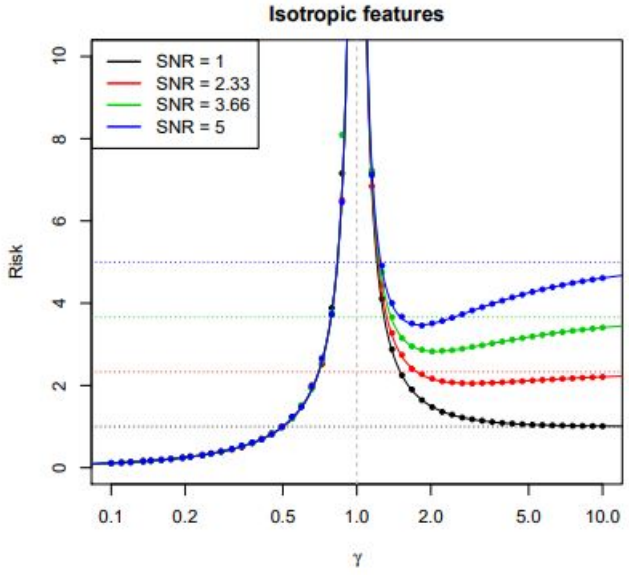

Figure 7: Plot from [HMRT19]. Plot of risk $\mathbf{E}[R_X(\widehat{\boldsymbol{\beta}})]$ as a function of γ. Signal to Noise (SNR) defined as the ratio of true norm to noise $\|\beta\|_2^2/\sigma^2 = r^2/\sigma^2$. In plots $r^2 = 1$

References

[AS17] Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

[BEHW87] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

[BHMM18] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[BHX19] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[BLLT19] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *arXiv preprint arXiv:1802.01396*, 2018.

[BRT18] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[GMMM19] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

[HMRT19] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[LR18] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

[MM19] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[RR08] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[RZ18] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. *arXiv preprint arXiv:1812.11167*, 2018.

[TV+04] Antonia M Tulino, Sergio Verdú, et al. Random matrix theory and wireless communications. *Foundations and Trends® in Communications and Information Theory*, 1(1):1–182, 2004.

[Vap13] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[ZBH+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.