Two models of double descent for weak features

Mikhail Belkin[1], Daniel Hsu[2], and Ji Xu[2]

[1]*The Ohio State University, Columbus, OH*
[2]*Columbia University, New York, NY*

March 19, 2019

Abstract

The "double descent" risk curve was recently proposed to qualitatively describe the out-of-sample prediction accuracy of variably-parameterized machine learning models. This article provides a precise mathematical analysis for the shape of this curve in two simple data models with the least squares/least norm predictor. Specifically, it is shown that the risk peaks when the number of features p is close to the sample size n, but also that the risk decreases towards its minimum as p increases beyond n. This behavior is contrasted with that of "prescient" models that select features in an *a priori* optimal order.

1 Introduction

The "double descent" risk curve was proposed by Belkin, Hsu, Ma, and Mandal [Bel+18] to qualitatively describe the out-of-sample prediction performance of several variably-parameterized machine learning models. This risk curve reconciles the classical bias-variance trade-off with the behavior of predictive models that interpolate training data, as observed for several model families (including neural networks) in a wide variety of applications [BO98; AS17; Spi+18; Bel+18]. In these studies, a predictive model with p parameters is fit to a training sample of size n, and the test risk (i.e., out-of-sample error) is examined as a function of p. When p is below the sample size n, the test risk is governed by the usual bias-variance decomposition. As p is increased towards n, the training risk (i.e., in-sample error) is driven to zero, but the test risk shoots up towards infinity. The classical bias-variance analysis identifies a "sweet spot" value of $p \in [0, n]$ at which the bias and variance are balanced to achieve low test risk. However, as p grows beyond n, the test risk again decreases, provided that the model is fit using a suitable inductive bias (e.g., least norm solution). In many (but not all) cases, the limiting risk as $p \to \infty$ is lower than what is achieved at the "sweet spot" value of p.

In this article, we study the key aspects of the "double descent" risk curve for the least squares/least norm predictor in two simple random features models. The first is a Gaussian model, which was studied by Breiman and Freedman [BF83] in the $p \leq n$ regime. The second is a Fourier series model for functions on the circle. In both cases, we prove that the risk is infinite around $p = n$, and decreases to towards its minimum as p increases beyond n. Our results provide a precise mathematical analysis of the mechanism described by Belkin et al. [Bel+18]. The transition from under- to over-parametrized regimes was also analyzed by Spigler, Geiger, d'Ascoli, Sagun, Biroli, and Wyart [Spi+18] by drawing a connection to the physical phenomenon of "jamming" in particle systems.

We note that in both of the models, the features are selected randomly, which makes them useful for studying scenarios where features are plentiful but individually too "weak" to be selected in an informed manner. Such scenarios are commonplace in machine learning practice, but they should be contrasted with scenarios where features are carefully designed or curated, as is often the case in scientific applications. For comparison, we give an example of "prescient" feature selection, where the p most useful features are included in the model. In this case, the optimal test risk is achieved at some $p \leq n$, which is consistent with the classical analysis of Breiman and Freedman [BF83].

E-mail: mbelkin@cse.ohio-state.edu, djhsu@cs.columbia.edu, jixu@cs.columbia.edu

2 Gaussian model

We consider a regression problem where the response y is equal to a linear function $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D) \in \mathbb{R}^D$ of D real-valued variables $\boldsymbol{x} = (x_1, \ldots, x_D)$ plus noise $\sigma\epsilon$:

$$y = \boldsymbol{x}^\intercal \boldsymbol{\beta} + \sigma\epsilon = \sum_{j=1}^{D} x_j \beta_j + \sigma\epsilon.$$

The learner observes n iid copies $((\boldsymbol{x}^{(i)}, y^{(i)}))_{i=1}^n$ of $(\boldsymbol{x}, y)$, but fits a linear model to the data only using a subset $T \subseteq [D] := \{1, \ldots, D\}$ of $p := |T|$ variables.

Let $\boldsymbol{X} := [\boldsymbol{x}^{(1)}| \cdots |\boldsymbol{x}^{(n)}]^\intercal$ be the $n \times D$ design matrix, and let $\boldsymbol{y} := (y^{(1)}, \ldots, y^{(n)})$ be the vector of responses. For a subset $A \subseteq [D]$ and a D-dimensional vector $\boldsymbol{v}$, we use $\boldsymbol{v}_A := (v_j : j \in A)$ to denote its $|A|$-dimensional subvector of entries from A; we also use $\boldsymbol{X}_A := [\boldsymbol{x}_A^{(1)}| \cdots |\boldsymbol{x}_A^{(n)}]^\intercal$ to denote the $n \times |A|$ design matrix with variables from A. For $A \subseteq [D]$, we denote its complement by $A^c := [D] \setminus A$. Finally, $\|\cdot\|$ denotes the Euclidean norm.

The learner fits regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_D)$ with

$$\hat{\boldsymbol{\beta}}_T := \boldsymbol{X}_T^\dagger \boldsymbol{y}, \quad \hat{\boldsymbol{\beta}}_{T^c} := \boldsymbol{0}.$$

Above, the symbol † denotes the Moore-Penrose pseudoinverse. In other words, the learner uses the solution to the normal equations $\boldsymbol{X}_T^\intercal \boldsymbol{X}_T \boldsymbol{v} = \boldsymbol{X}_T^\intercal \boldsymbol{y}$ of least norm for $\hat{\boldsymbol{\beta}}_T$ and forces $\hat{\boldsymbol{\beta}}_{T^c}$ to all-zeros.

In the remainder of this section, we analyze the risk of $\hat{\boldsymbol{\beta}}$ in the case where the distribution of $\boldsymbol{x}$ is the standard normal in $\mathbb{R}^D$, and then specialize the risk under particular selection models for T. The Gaussian model was also studied by Breiman and Freedman [BF83], although their analysis is restricted to the case where the number of variables used p is always at most n; our analysis will also consider the $p \geq n$ regime. We show that the "interpolating" $p \geq n$ regime is preferred to the "classical" $p \leq n$ regime in our model.

We focus on the noise-free setting in Section 2.1 and Section 2.2; we consider noise in Section 2.3.

2.1 Risk analysis

In this section, we derive an expression for the risk of $\hat{\boldsymbol{\beta}}$ for an arbitrary choice of p features $T \subseteq [D]$.

We assume $\sigma = 0$ (i.e., the noise-free setting), so $y = \boldsymbol{x}^\intercal \boldsymbol{\beta}$. Recall that we also assume $\boldsymbol{x}$ follows a standard normal distribution in $\mathbb{R}^D$; since $\boldsymbol{x}$ is isotropic (i.e., zero mean and identity covariance), the mean squared prediction error of any $\boldsymbol{\beta}' \in \mathbb{R}^D$ can be written as

$$\mathbb{E}[(y - \boldsymbol{x}^\intercal \boldsymbol{\beta}')^2] = \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|^2 = \|\boldsymbol{\beta}_{T^c} - \boldsymbol{\beta}'_{T^c}\|^2 + \|\boldsymbol{\beta}_T - \boldsymbol{\beta}'_T\|^2.$$

Since $\hat{\boldsymbol{\beta}}_{T^c} = \boldsymbol{0}$, it follows that the risk of $\hat{\boldsymbol{\beta}}$ is

$$\mathbb{E}[(y - \boldsymbol{x}^\intercal \hat{\boldsymbol{\beta}})^2] = \|\boldsymbol{\beta}_{T^c}\|^2 + \mathbb{E}[\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2].$$

Classical regime. The risk of $\hat{\boldsymbol{\beta}}$ was computed by Breiman and Freedman [BF83] in the regime where $p \leq n$:

$$\mathbb{E}[(y - \boldsymbol{x}^\intercal \hat{\boldsymbol{\beta}})^2] = \begin{cases} \|\boldsymbol{\beta}_{T^c}\|^2 \cdot \left(1 + \frac{p}{n-p-1}\right) & \text{if } p \leq n-2; \\ 0 & \text{if } \boldsymbol{\beta}_{T^c} = \boldsymbol{0}; \\ +\infty & \text{if } p \in \{n-1, n\} \text{ and } \boldsymbol{\beta}_{T^c} \neq \boldsymbol{0}. \end{cases}$$

Interpolating regime. We consider the regime where $p \geq n$. Recall that the pseudoinverse of $\boldsymbol{X}_T$ can be written as $\boldsymbol{X}_T^\dagger = \boldsymbol{X}_T^\intercal (\boldsymbol{X}_T \boldsymbol{X}_T^\intercal)^\dagger$. Thus,

$$\begin{aligned} \boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T &= \boldsymbol{\beta}_T - \boldsymbol{X}_T^\intercal (\boldsymbol{X}_T \boldsymbol{X}_T^\intercal)^\dagger \boldsymbol{y} \\ &= \boldsymbol{\beta}_T - \boldsymbol{X}_T^\intercal (\boldsymbol{X}_T \boldsymbol{X}_T^\intercal)^\dagger (\boldsymbol{X}_{T^c} \boldsymbol{\beta}_{T^c} + \boldsymbol{X}_T \boldsymbol{\beta}_T) \\ &= (\boldsymbol{I} - \boldsymbol{X}_T^\intercal (\boldsymbol{X}_T \boldsymbol{X}_T^\intercal)^\dagger \boldsymbol{X}_T) \boldsymbol{\beta}_T - \boldsymbol{X}_T^\intercal (\boldsymbol{X}_T \boldsymbol{X}_T^\intercal)^\dagger \boldsymbol{X}_{T^c} \boldsymbol{\beta}_{T^c}. \end{aligned}$$

On the right hand side, the first term $(\boldsymbol{I} - \boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_T)\boldsymbol{\beta}_T$ is the orthogonal projection of $\boldsymbol{\beta}_T$ onto the null space of $\boldsymbol{X}_T$, while the second term $-\boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c}$ is a vector in the row space of $\boldsymbol{X}_T$. By the Pythagorean theorem, the squared norm of their sum is equal to the sum of their squared norms, so

$$\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2 = \|(\boldsymbol{I} - \boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_T)\boldsymbol{\beta}_T\|^2 + \|\boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c}\|^2.$$

We analyze the expected values of these two terms by exploiting properties of the standard normal distribution.

First term. Note that $\boldsymbol{\Pi}_T := \boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_T$ is the orthogonal projection matrix for the row space of $\boldsymbol{X}_T$. So, by the Pythagorean theorem, we have

$$\|(\boldsymbol{I} - \boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_T)\boldsymbol{\beta}_T\|^2 = \|\boldsymbol{\beta}_T\|^2 - \|\boldsymbol{\Pi}_T\boldsymbol{\beta}_T\|^2.$$

By rotational symmetry of the standard normal distribution, it follows that

$$\mathbb{E}[\|\boldsymbol{\Pi}_T\boldsymbol{\beta}_T\|^2] = \|\boldsymbol{\beta}_T\|^2 \cdot \frac{n}{p}.$$

Therefore

$$\mathbb{E}[\|(\boldsymbol{I} - \boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_T)\boldsymbol{\beta}_T\|^2] = \|\boldsymbol{\beta}_T\|^2 \cdot \left(1 - \frac{n}{p}\right).$$

Second term. We use the "trace trick" to write

$$\begin{aligned}\|\boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c}\|^2 &= \operatorname{tr}((\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})^\intercal) \\ &= \operatorname{tr}((\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})^\intercal)\end{aligned}$$

where the second equality holds almost surely because $\boldsymbol{X}_T\boldsymbol{X}_T^\intercal$ is almost surely invertible. Since $\boldsymbol{x}_T$ and $\boldsymbol{x}_{T^c}$ are uncorrelated, it follows that

$$\mathbb{E}[\|\boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c}\|^2] = \operatorname{tr}(\mathbb{E}[(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger]\mathbb{E}[(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})^\intercal]).$$

The distribution of $\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c}$ is normal with mean zero and covariance $\|\boldsymbol{\beta}_{T^c}\|^2 \cdot \boldsymbol{I} \in \mathbb{R}^{n\times n}$, so

$$\mathbb{E}[(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})(\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c})^\intercal] = \|\boldsymbol{\beta}_{T^c}\|^2 \cdot \boldsymbol{I}.$$

The distribution of $(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger$ is inverse-Wishart with identity scale matrix $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ and p degrees-of-freedom, so

$$\operatorname{tr}(\mathbb{E}[(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger]) = \begin{cases} \frac{n}{p-n-1} & \text{if } p \geq n+2; \\ +\infty & \text{if } p \in \{n, n+1\}. \end{cases}$$

Combining the last two displayed equations yields a simple expression for $\mathbb{E}[\|\boldsymbol{X}_T^\intercal(\boldsymbol{X}_T\boldsymbol{X}_T^\intercal)^\dagger\boldsymbol{X}_{T^c}\boldsymbol{\beta}_{T^c}\|^2]$.

Thus, we obtain expressions for $\mathbb{E}[\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2]$ and hence also $\mathbb{E}[(y - \boldsymbol{x}^\intercal\hat{\boldsymbol{\beta}})^2]$.

We summarize the risk of $\hat{\boldsymbol{\beta}}$ in the following theorem.

Theorem 1. *Assume the distribution of $\boldsymbol{x}$ is the standard normal in $\mathbb{R}^D$ and $y = \boldsymbol{x}^\intercal\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^D$. Pick any $p \in \{0, \ldots, D\}$ and $T \subseteq [D]$ of cardinality p. The risk of $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}_T = \boldsymbol{X}_T^\dagger\boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}_{T^c} = \boldsymbol{0}$, is*

$$\mathbb{E}[(y - \boldsymbol{x}^\intercal\hat{\boldsymbol{\beta}})^2] = \begin{cases} \|\boldsymbol{\beta}_{T^c}\|^2 \cdot \left(1 + \frac{p}{n-p-1}\right) & \text{if } p \leq n-2; \\ +\infty & \text{if } n-1 \leq p \leq n+1 \text{ and } \boldsymbol{\beta}_{T^c} \neq \boldsymbol{0}; \\ \|\boldsymbol{\beta}_T\|^2 \cdot \left(1 - \frac{n}{p}\right) + \|\boldsymbol{\beta}_{T^c}\|^2 \cdot \left(1 + \frac{n}{p-n-1}\right) & \text{if } p \geq n+2; \\ \|\boldsymbol{\beta}_T\|^2 \cdot \max\left\{1 - \frac{n}{p}, 0\right\} & \text{if } \boldsymbol{\beta}_{T^c} = \boldsymbol{0}. \end{cases}$$

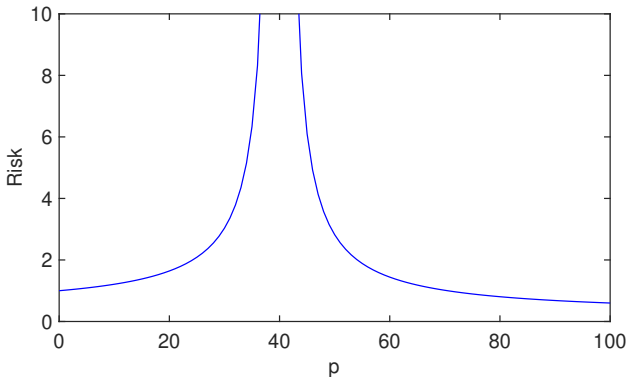


Figure 1: Plot of risk $\mathbb{E}[(y - \boldsymbol{x}^\intercal \hat{\boldsymbol{\beta}})^2]$ as a function of p, under the random selection model of T. Here, $\|\boldsymbol{\beta}\|^2 = 1$, $\sigma^2 = 0$, $D = 100$, and $n = 40$.

2.2 Feature selection model

We now study the risk of $\hat{\boldsymbol{\beta}}$ under a random selection model for T. Again, we assume $\sigma = 0$.

Let T be a uniformly random subset of $[D]$ of cardinality p, so

$$\mathbb{E}[\|\boldsymbol{\beta}_T\|^2] = \frac{p}{D} \cdot \|\boldsymbol{\beta}\|^2, \quad \mathbb{E}[\|\boldsymbol{\beta}_{T^c}\|^2] = \left(1 - \frac{p}{D}\right) \cdot \|\boldsymbol{\beta}\|^2.$$

We analyze the risk of $\hat{\boldsymbol{\beta}}$, taking expectation with respect to the random choice of T. First, consider $p \le n-2$. By Theorem 1, the risk of $\hat{\boldsymbol{\beta}}$ is

$$\mathbb{E}[(y - \boldsymbol{x}^\intercal \hat{\boldsymbol{\beta}})^2] = \|\boldsymbol{\beta}\|^2 \cdot \left(1 - \frac{p}{D}\right) \cdot \left(1 + \frac{p}{n-p-1}\right),$$

which increases with p as long as $D \ge n$. Now consider $p \ge n+2$. By Theorem 1, the risk of $\hat{\boldsymbol{\beta}}$ is

$$\mathbb{E}[(y - \boldsymbol{x}^\intercal \hat{\boldsymbol{\beta}})^2] = \|\boldsymbol{\beta}\|^2 \cdot \left(1 - \frac{n}{D} \cdot \left(2 - \frac{D-n-1}{p-n-1}\right)\right),$$

which decreases with p as long as $D > n+1$.

Thus, we observe that the risk first *increases* with p up to the "interpolation threshold" ($p = n$), after which the risk *decreases* with p. Moreover, the risk is smallest at $p = D$. This is the "double descent" risk curve observed by Belkin et al. [Bel+18] where the first "descent" is degenerate (i.e., the "sweet spot" that balances bias and variance is at $p = 0$). See Figure 1 for an illustration. For a scenario where the first "descent" is non-degenerate, see Appendix A.

It is worth pointing out that the behavior under the random selection model of T can be very different from that under a deterministic model of T. Consider including variables in T by decreasing order of β_j^2—a kind of "prescient" selection model studied by Breiman and Freedman [BF83]. For simplicity, assume $\beta_j^2 = 1/j^2$ and $D = \infty$, so

$$\|\boldsymbol{\beta}_T\|^2 = \sum_{j=1}^{p} \frac{1}{j^2}, \quad \|\boldsymbol{\beta}_{T^c}\|^2 = \frac{\pi^2}{6} - \sum_{j=1}^{p} \frac{1}{j^2}.$$

The behavior of the risk as a function of p, illustrated in Figure 2, reveals a striking difference between the random selection model and the "prescient" selection model.

2.3 Noise

We now briefly discuss the effect of additive noise in the responses. The following is a straightforward generalization of Theorem 1.

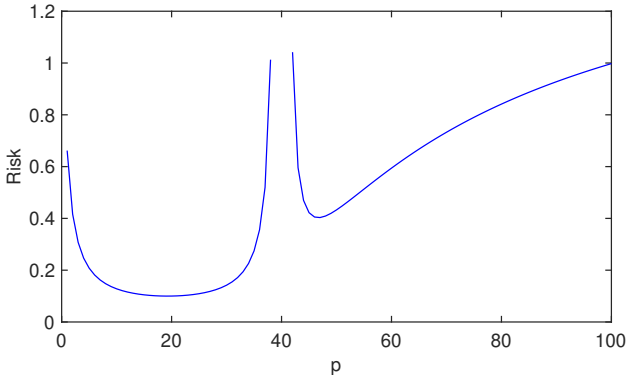


Figure 2: Plot of risk $\mathbb{E}[(y - \boldsymbol{x}^\intercal \hat{\boldsymbol{\beta}})^2]$ as a function of p, under the "prescient" selection model of T. Here, $\|\boldsymbol{\beta}\|^2 = \pi^2/6$, $\sigma^2 = 0$, $D = \infty$, and $n = 40$. As $p \to \infty$, the risk approaches $\|\boldsymbol{\beta}\|^2$ from below.

Theorem 2. *Assume the distribution of $\boldsymbol{x}$ is the standard normal in $\mathbb{R}^D$, ϵ is a standard normal random variable independent of $\boldsymbol{x}$, and $y = \boldsymbol{x}^\intercal \boldsymbol{\beta} + \sigma\epsilon$ for some $\boldsymbol{\beta} \in \mathbb{R}^D$ and $\sigma > 0$. Pick any $p \in \{0, \dots, D\}$ and $T \subseteq [D]$ of cardinality p. The risk of $\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}_T = \boldsymbol{X}_T^\dagger \boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}_{T^c} = \boldsymbol{0}$, is*

$$\mathbb{E}[(y - \boldsymbol{x}^\intercal \hat{\boldsymbol{\beta}})^2] = \begin{cases} (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \left(1 + \frac{p}{n-1-p}\right) & \text{if } p \le n-2; \\ +\infty & \text{if } n-1 \le p \le n+1; \\ \|\boldsymbol{\beta}_T\|^2 \cdot \left(1 - \frac{n}{p}\right) + (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \left(1 + \frac{n}{p-n-1}\right) & \text{if } p \ge n+2. \end{cases}$$

A similar analysis with a random selection model for T can be obtained from Theorem 2.

3 Fourier series model

Let $\boldsymbol{F} \in \mathbb{C}^{D \times D}$ denote the $D \times D$ discrete Fourier transform matrix: its (i,j)-th entry is

$$F_{i,j} = \frac{1}{\sqrt{D}} \omega^{(i-1)(j-1)},$$

where $\omega := \exp(-2\pi \mathrm{i}/D)$ is a primitive root of unity. Let $\boldsymbol{\mu} := \boldsymbol{F}\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{C}^D$. Consider the following observation model:

1. S and T are independent and uniformly random subsets of $[D]$ of cardinalities n and p, respectively.
2. We observe the $n \times p$ design matrix $\boldsymbol{F}_{S,T}$ and n-dimensional vector of responses $\boldsymbol{\mu}_S$. Here, $\boldsymbol{F}_{S,T}$ is the submatrix of $\boldsymbol{F}$ with rows from S and columns from T, and $\boldsymbol{\mu}_S$ is the subvector of $\boldsymbol{\mu}$ of entries from S.

The learner fits regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_D)$ with

$$\hat{\boldsymbol{\beta}}_T := \boldsymbol{F}_{S,T}^\dagger \boldsymbol{\mu}_S, \quad \hat{\boldsymbol{\beta}}_{T^c} := \boldsymbol{0}.$$

This can be regarded as a one-dimensional version of the random Fourier features model studied by Rahimi and Recht [RR08] for functions defined on the unit circle.

One important property of the discrete Fourier transform matrix that we use is that the matrix $\boldsymbol{F}_{A,B}$ has rank $\min\{|A|, |B|\}$ for any $A, B \subseteq [D]$. This is a consequence of the fact that $\boldsymbol{F}$ is Vandermonde. Thus, for $p \ge n$, we have

$$\boldsymbol{F}_{S,T}^\dagger = \boldsymbol{F}_{S,T}^* (\boldsymbol{F}_{S,T} \boldsymbol{F}_{S,T}^*)^{-1}.$$

In the remainder of this section, we analyze the risk of $\hat{\boldsymbol{\beta}}$ under a random model for $\boldsymbol{\beta}$, where

$$\mathbb{E}[\boldsymbol{\beta}\boldsymbol{\beta}^*] = \frac{1}{D} \cdot \boldsymbol{I}$$

(which implies $\mathbb{E}[\|\boldsymbol{\beta}\|^2] = 1$). The random choice of $\boldsymbol{\beta}$ is independent of S and T. Considering the risk under this random model for $\boldsymbol{\beta}$ is a form of average-case analysis. For simplicity, we only consider the regime where $p \geq n$, as it suffices to reveal some key aspects of the risk of $\hat{\boldsymbol{\beta}}$.

Following the arguments from Section 2.1, we have

$$\begin{aligned}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2 &= \|\boldsymbol{\beta}_{T^c}\|^2 + \|(\boldsymbol{I} - \boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T})\boldsymbol{\beta}_T\|^2 + \|\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T^c}\boldsymbol{\beta}_{T^c}\|^2 \\ &= \|\boldsymbol{\beta}\|^2 - \|\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T}\boldsymbol{\beta}_T\|^2 + \|\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T^c}\boldsymbol{\beta}_{T^c}\|^2.\end{aligned}$$

Now we take (conditional) expectations with respect to $\boldsymbol{\beta}$, given S and T:

$$\mathbb{E}[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2 \mid S, T] = 1 - \frac{1}{D} \cdot \operatorname{tr}((\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T})^*(\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T})) + \frac{1}{D} \cdot \operatorname{tr}((\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T^c})^*(\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T^c})). \quad (1)$$

Since $\boldsymbol{F}_{S,T}$ has rank n, the first trace expression is equal to

$$\operatorname{tr}((\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T})^*(\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T})) = n.$$

For the second trace expression, we use the explicit formula for $\boldsymbol{F}_{S,T}^\dagger$ and the fact that $\boldsymbol{F}_{S,T}\boldsymbol{F}_{S,T}^* + \boldsymbol{F}_{S,T^c}\boldsymbol{F}_{S,T^c}^* = \boldsymbol{I}$ to obtain

$$\begin{aligned}\operatorname{tr}((\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T^c})^*(\boldsymbol{F}_{S,T}^\dagger \boldsymbol{F}_{S,T^c})) &= \operatorname{tr}(\boldsymbol{F}_{S,T^c}^*(\boldsymbol{F}_{S,T}\boldsymbol{F}_{S,T}^*)^{-1}\boldsymbol{F}_{S,T^c}) \\ &= \operatorname{tr}(\boldsymbol{F}_{S,T^c}^*(\boldsymbol{I} - \boldsymbol{F}_{S,T^c}\boldsymbol{F}_{S,T^c}^*)^{-1}\boldsymbol{F}_{S,T^c}) \\ &= \operatorname{tr}((\boldsymbol{I} - \boldsymbol{F}_{S,T^c}\boldsymbol{F}_{S,T^c}^*)^{-1}\boldsymbol{F}_{S,T^c}\boldsymbol{F}_{S,T^c}^*) \\ &= \sum_{i=1}^n \frac{\lambda_i}{1-\lambda_i} \\ &= -n + \sum_{i=1}^n \frac{1}{1-\lambda_i},\end{aligned}$$

where $\lambda_1, \ldots, \lambda_n \in [0,1]$ are the eigenvalues of $\boldsymbol{F}_{S,T^c}\boldsymbol{F}_{S,T^c}^*$. Therefore, from Equation (1), we have

$$\mathbb{E}[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2] = 1 - \frac{2n}{D} + \frac{n}{D} \cdot \underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{1-\lambda_i}\right]}_{(*)}.$$

A precise characterization of $(*)$ is difficult to obtain. Under a slightly different model, in which membership in S (respectively, T) is determined by independent Bernoulli variables with mean n/D (respectively, p/D), we can use asymptotic arguments to characterize the empirical eigenvalue distribution for $\boldsymbol{F}_{S,T}\boldsymbol{F}_{S,T}^*$.[1] Assuming the asymptotic equivalence of these random models for S and T, we find that the quantity $(*)$ approaches

$$\frac{\rho_p \cdot (1-\rho_n)}{\rho_p - \rho_n}$$

as $D, n, p \to \infty$, where $\rho_n := n/D$ and $\rho_p := p/D$ are held fixed and $\rho_p > \rho_n$ [Far11]. So, in this limit, we have

$$\mathbb{E}[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2] \to 1 - \rho_n \cdot \left(2 - \frac{\rho_p \cdot (1-\rho_n)}{\rho_p - \rho_n}\right).$$

This quantity diverges to $+\infty$ as $\rho_p \to \rho_n$, and decreases as $\rho_p \to 1$. This is the same behavior as in the Gaussian model from Section 2 with random selection; we depict it empirically in Figure 3.

[1] We can also derive essentially the same formula for the risk under this Bernoulli model, although the derivation is somewhat more cumbersome since we do not always have $|T| \geq |S|$ even if $\mathbb{E}|T| \geq \mathbb{E}|S|$. Hence, we have opted to present the derivation only for the simpler model.

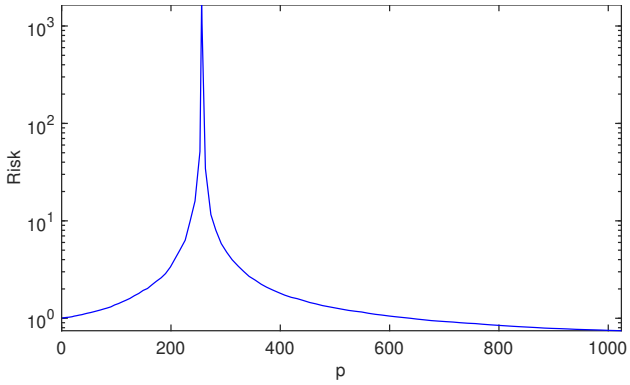


Figure 3: Plot of risk as a function of p in the Fourier series model. Here, $\boldsymbol{\beta}$ was chosen uniformly at random (once) from the unit sphere in $\mathbb{R}^D$ for $D = 1024$. We then computed $\hat{\boldsymbol{\beta}}$ from 10 independent random choices of S (with $n = 256$) and T and plotted the average value of $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2$.

4 Discussion

Our analyses show that when features are chosen in an uninformed manner, it may be optimal to choose as many as possible—even more than the number of data—rather than limit the number to that which balances bias and variance. This choice is conceptually and algorithmically simple and avoids the need for precise control of regularization parameters. It is reminiscent of the practice in machine learning applications like image and speech recognition, where signal processing-based features are individually weak but in great abundance, and models that use all of the features are highly successful. This stands in contrast to scenarios where informed selection of features is possible; for example, in many science and medical applications, features are hand-crafted and purposefully selected. As illustrated by the "prescient" selection model, choosing the number of features to balance bias and variance can be better than incurring the costs that come with using all of the features. The best practices for model and feature selection thus crucially depend on which regime the application is operating under.

References

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

[Bel+18] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine learning and the bias-variance trade-off". In: *arXiv preprint arXiv:1812.11118* (2018).

[BF83] Leo Breiman and David Freedman. "How many variables should be entered in a regression equation?" In: *Journal of the American Statistical Association* 78.381 (1983), pp. 131–136.

[BO98] Siegfried Bös and Manfred Opper. "Dynamics of batch training in a perceptron". In: *Journal of Physics A: Mathematical and General* 31.21 (1998), p. 4835.

[Far11] Brendan Farrell. "Limiting empirical singular value distribution of restrictions of discrete Fourier transform matrices". In: *Journal of Fourier Analysis and Applications* 17.4 (2011), pp. 733–753.

[RR08] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in Neural Information Processing Systems*. 2008, pp. 1177–1184.

[Spi+18] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. "A jamming transition from under-to over-parametrization affects loss landscape and generalization". In: *arXiv preprint arXiv:1810.09665* (2018).

A Non-degenerate double descent in a Fourier series model

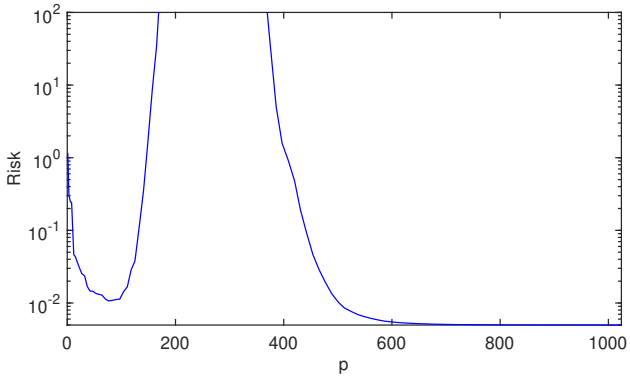
To observe the "double descent" risk curve where the first "descent" is non-degenerate, we consider a model in which the distribution of the feature vector $\boldsymbol{x}$ is non-isotropic; instead, $\boldsymbol{x}$ has a diagonal covariance matrix $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^*] = \mathrm{diag}(t_1^2, \dots, t_D^2)$ with decaying eigenvalues (e.g., $t_i^2 \propto i^{-2}$). In such a scenario, it is natural to select features in decreasing order of the eigenvalues, as is done in principal components regression. Like random feature selection, this form of feature selection is also generally uninformed by the responses.

Formally, we let $\boldsymbol{\mu} := \boldsymbol{F}\mathrm{diag}(t_1, \dots, t_D)\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{C}^D$ and positive sequence $t_1 > t_2 > \cdots$, and use the same observation model as in Section 3, except that T is deterministically set to $T := \{1, \dots, p\}$. We fit $\hat{\boldsymbol{\beta}}$ in the same manner as in Section 3. We are interested in the risk

$$\sum_{j=1}^{D} t_j^2(\beta_j - \hat{\beta}_j)^2,$$

which can be regarded as the mean squared error when $\boldsymbol{x}$ is drawn uniformly at random from the rows of the design matrix $\boldsymbol{F}\mathrm{diag}(t_1, \dots, t_D)$.

We carried out the same simulation as from Figure 3 under the modified model, with $t_i^2 \propto i^{-2}$ and $\sum_{i=1}^{D} t_i^2 = 1$. We chose $\boldsymbol{\beta}$ uniformly at random (once) from the unit sphere in $\mathbb{R}^D$ for $D = 1024$. Then, for each p, we computed $\hat{\boldsymbol{\beta}}$ from 10 independent random choices of S (with $n = 256$), and plotted the average value of $\sum_{j=1}^{D} t_j^2(\beta_j - \hat{\beta}_j)^2$. The plot is shown below. (The vertical axis is truncated for clarity; the curve peaks around $p = n$ with value on the order of 10^{16}.)



The plot shows the usual "U"-shaped curve arising from the bias-variance trade-off when $p < n$, and a second "descent" towards the overall minimum for $p > n$. This risk curve is qualitatively the same as those observed by Belkin et al. [Bel+18] for neural networks and other predictive models.