SIAM J. MATH. DATA SCI.
Vol. 2, No. 4, pp. 1167–1180



# Two Models of Double Descent for Weak Features*

Mikhail Belkin†, Daniel Hsu‡, and Ji Xu§

**Abstract.** The "double descent" risk curve was 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 sometimes decreases toward its minimum as $p$ increases beyond $n$. This behavior parallels some key patterns observed in large models, including modern neural networks, and is contrasted with that of "prescient" models that select features in an a priori optimal order.

**Key words.** overparameterization, least squares, inductive bias

**AMS subject classifications.** 62J05, 68T10

**DOI.** 10.1137/20M1336072

## 1. Introduction.

The "double descent" risk curve was proposed by Belkin et al. [4] as a general way to qualitatively describe the out-of-sample prediction performance of 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 (see section 1.1 for references). 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 $n$ (for regression or binary classification), the test risk is governed by the usual bias-variance decomposition. As $p$ is increased toward $n$, the training risk (i.e., in-sample error) is driven to zero, but the test risk shoots up, sometimes toward 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, in the "modern regime," as $p$ grows beyond $n$, the training risk remains zero, but the test risk decreases again, even when fitting noisy data, provided that the model is fit using a suitable inductive bias (e.g., least norm solution). In many (but not all) cases from [4], the limiting risk as $p \to \infty$ is lower than what is achieved at the "sweet spot" value of $p$.

*Received by the editors May 5, 2020; accepted for publication (in revised form) October 6, 2020; published electronically December 14, 2020.
https://doi.org/10.1137/20M1336072
**Funding:** This research was supported by NSF CCF-1740833 and IIS-1815697 awards, a Sloan Research Fellowship, a Google Faculty Award, and a Cheung-Kong Graduate School of Business Fellowship.

†Halıcıoğlu Data Science Institute, UC San Diego, La Jolla, CA 43221 USA (mbelkin@ucsd.edu).
‡Department of Computer Science and Data Science Institute, Columbia University, New York, NY 10027 USA (djhsu@cs.columbia.edu).
§Department of Computer Science, Columbia University, New York, NY 10027 USA (jixu@cs.columbia.edu).

In this article, we show that key aspects of the "double descent" risk curve can be observed with the least squares/least norm predictor in two simple random features models. The first is a Gaussian model studied by Breiman and Freedman [7] in the classical $p \leq n$ regime, while 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 again as $p$ increases beyond $n$. When the signal-to-noise ratio is high, the minimum risk is, in fact, achieved in the modern regime, when $p > n$. Our results provide a precise mathematical analysis in a simple and tractable setting of the mechanism that was qualitatively described by Belkin et al. [4]. In particular, it captures a key aspect of many practical overparameterized models: that increasing the number of parameters to the maximum can lead to better performance. We also establish some nonasymptotic concentration phenomena in the Gaussian model.

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 common in machine learning practice, and they should be contrasted with "scientific" 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 $p$ features a priori known to be most useful 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 [7].

**1.1. Related and concurrent works.** The "double descent" risk curve was posited by Belkin et al. [4] to connect the classical bias-variance trade-off to behaviors observed in overparameterized regimes for a variety of machine learning models. The shape and features of the risk curve itself appear throughout in the literature in a number of contexts, e.g., [21, 17, 13, 12, 6, 23, 1]; see also [14] for a "brief prehistory" that focuses on the curious peak in the curve. These prior works analyze the risk of linear classification and regression models and neural networks in high-dimensional asymptotic regimes. Our analysis in the Gaussian model gives an exact expression for the risk for any finite sample size and number of parameters.

More recently, Neal et al. [16] observed that similar phenomena in neural networks can be explained by a variance reduction effect of increasing network width. The transition from under- to overparametrized regimes was recently analyzed by Spigler et al. [20] by drawing a connection to the physical phenomenon of "jamming" in a class of glassy systems. Our analysis makes these ideas concrete and explicit in the context of simple regression models. For instance, our analysis captures the transition from under- to overparameterized regimes at a point where an inverse Wishart random matrix has no finite expectation. It also allows us to compare the risks at any points in the curve and explain how the risk in the overparameterized regime can be lower than any risk in the underparameterized regime.

The initial version of this article [5] appeared concurrently with works of Hastie et al. [11], Muthukumar et al. [15], and Bartlett et al. [3], all of which also study the behavior of the least squares/least norm predictor in overparameterized linear regression. Muthukumar et al. [15] focus on the well-specified scenario (essentially $p = D$) and provide upper bounds on the risk that go to zero as $p \to \infty$. (A related variance analysis was carried out by Neal et al. [16].) Hastie et al. [11] provide a much broader range of analyses in the high-dimensional

asymptotic regime, including a "misspecified" setup that is related to ours. Their analyses require weaker distributional assumptions than ours because of their reliance on asymptotic analysis. (A special case of the results in the follow-up work by Xu and Hsu [24] further broadens the range of analyses to allow highly nonisotropic designs, but again only in the high-dimensional asymptotic regime.) The analysis of Hastie et al. [11] also considers the effect of ridge regularization; in particular, they show that when the optimal level of regularization is used, the risk curve no longer shows the "double descent" shape. Finally, Bartlett et al. [3] study nonasymptotic upper and lower bounds on the risk in the overparameterized regime and provide a characterization in terms of certain "effective dimensions" based on the tail of the eigenvalue sequence of the covariance operator.

## 2. Gaussian model.

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

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

Given $n$ independent and identically distributed copies $((\boldsymbol{x}^{(i)}, y^{(i)}))_{i=1}^n$ of $(\boldsymbol{x}, y)$, we fit a linear model to the data only using a subset $T \subseteq [D] := \{1, \dots, D\}$ of $p := |T|$ variables.

Let $\boldsymbol{X} := [\boldsymbol{x}^{(1)}|\cdots|\boldsymbol{x}^{(n)}]^*$ be the $n \times D$ design matrix, and let $\boldsymbol{y} := (y^{(1)}, \dots, y^{(n)})$ 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 $\boldsymbol{X}_A := [\boldsymbol{x}_A^{(1)}|\cdots|\boldsymbol{x}_A^{(n)}]^*$ 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.

We fit regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_D)$ with

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

Above, the symbol $^\dagger$ denotes the Moore–Penrose pseudoinverse. In other words, we use the solution to the normal equations $\boldsymbol{X}_T^*\boldsymbol{X}_T\boldsymbol{v} = \boldsymbol{X}_T^*\boldsymbol{y}$ of least norm for $\hat{\boldsymbol{\beta}}_T$ and force $\hat{\boldsymbol{\beta}}_{T^c}$ to all zeros.

In this section, our analysis assumes a model in which $(\boldsymbol{x}, \epsilon)$ follows a standard multivariate Gaussian distribution. This Gaussian model was also studied by Breiman and Freedman [7], 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.

### 2.1. Prediction risk.

We derive a formula for the (prediction) risk of $\hat{\boldsymbol{\beta}}$ for an arbitrary choice of $p$ features $T \subseteq [D]$ and then examine this risk under particular selection models for $T$.

Theorem 2.1. *Assume the distribution of $\boldsymbol{x}$ is the standard normal in $\mathbb{R}^D$, $\epsilon$ is a standard normal random variable independent of $\boldsymbol{x}$, and $y = \boldsymbol{x}^*\boldsymbol{\beta} + \sigma\epsilon$ 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*

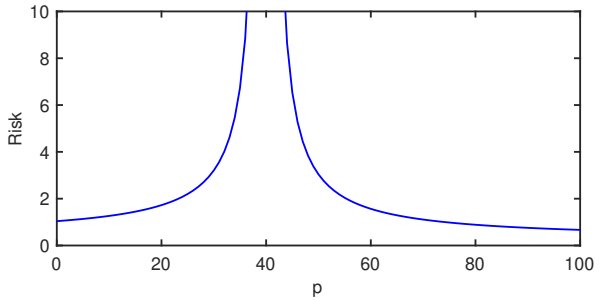


**Figure 1.** *Plot of risk* $\mathbb{E}[(y - \boldsymbol{x}^*\hat{\boldsymbol{\beta}})^2]$ *as a function of* $p$ *under the random selection model of* $T$*. Here,* $\|\boldsymbol{\beta}\|^2 = 1$*,* $\sigma^2 = 1/25$*,* $D = 100$*, and* $n = 40$*.*

$\hat{\boldsymbol{\beta}}_{T^c} = \mathbf{0}$*, is*

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

The proof of Theorem 2.1 is not hard; we give the details in subsection 2.2. We now turn to the risk of $\hat{\boldsymbol{\beta}}$ under a random selection model for $T$.

Corollary 2.2. *Let* $T$ *be a uniformly random subset of* $[D]$ *of cardinality* $p$*. In the setting of Theorem 2.1, the risk of* $\hat{\boldsymbol{\beta}}$ *(taking expectation with respect to the random choice of* $T$ *in addition to the random design matrix and response vector) satisfies*

$$\mathbb{E}[(y - \boldsymbol{x}^*\hat{\boldsymbol{\beta}})^2] = \begin{cases} \left((1 - \frac{p}{D}) \cdot \|\boldsymbol{\beta}\|^2 + \sigma^2\right) \cdot \left(1 + \frac{p}{n-p-1}\right) & \text{if } p \le n-2, \\ \|\boldsymbol{\beta}\|^2 \cdot \left(1 - \frac{n}{D} \cdot \left(2 - \frac{D-n-1}{p-n-1}\right)\right) + \sigma^2 \cdot \left(1 + \frac{n}{p-n-1}\right) & \text{if } p \ge n+2. \end{cases}$$

*Proof.* Since $T$ is a uniformly random subset of $[D]$ of cardinality $p$,

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

Plugging into Theorem 2.1 completes the proof. ∎

Thus, assuming $D > n + 1$, we observe that the risk first *increases* with $p$ up to the "interpolation threshold" ($p = n$), after which the risk *decreases* with $p$. Moreover, when the signal-to-noise ratio $\|\boldsymbol{\beta}\|^2/\sigma^2$ is larger than $D/(D - n - 1)$, the risk is smallest at $p = D$; in particular, it is smaller than the risk at any $p \le n$. This is the "double descent" risk curve, 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.

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 $\beta_j^2$—a kind of "prescient" selection model studied by Breiman and Freedman [7]. 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.

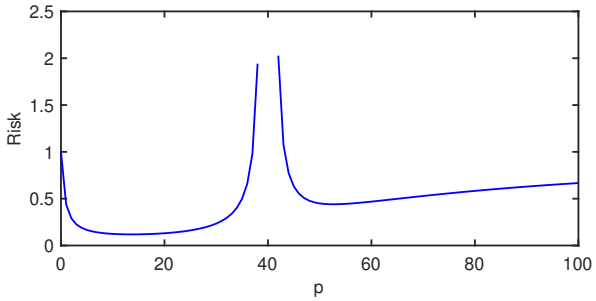


**Figure 2.** *Plot of risk* $\mathbb{E}[(y - \boldsymbol{x}^*\hat{\boldsymbol{\beta}})^2]$ *as a function of* $p$ *under the "prescient" selection model of* $T$*. Here,* $\|\boldsymbol{\beta}\|^2 = 1$*,* $\beta_j^2 \propto 1/j^2$*,* $\sigma^2 = 1/25$*,* $D = 100$*, and* $n = 40$*.*

**2.2. Proof of Theorem 2.1.** Recall that $\boldsymbol{x}$ is assumed to follow 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

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

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

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

*Classical regime.* The risk of $\hat{\boldsymbol{\beta}}$ was computed by Breiman and Freedman [7] in the regime where $p \le n$:

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

*Interpolating regime.* We consider the regime where $p \ge n$. Recall that the pseudoinverse of $\boldsymbol{X}_T$ can be written as $\boldsymbol{X}_T^\dagger = \boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger$. Thus, letting $\boldsymbol{\eta} := \boldsymbol{y} - \boldsymbol{X}_T\boldsymbol{\beta}_T$,

$$\begin{aligned} \boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T &= \boldsymbol{\beta}_T - \boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger \boldsymbol{y} \\ &= \boldsymbol{\beta}_T - \boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger(\boldsymbol{X}_T\boldsymbol{\beta}_T + \boldsymbol{\eta}) \\ &= (\boldsymbol{I} - \boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger \boldsymbol{X}_T)\boldsymbol{\beta}_T - \boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger \boldsymbol{\eta}. \end{aligned}$$

On the right-hand side, the first term $(\boldsymbol{I} - \boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger \boldsymbol{X}_T)\boldsymbol{\beta}_T$ is the orthogonal projection of $\boldsymbol{\beta}_T$ onto the null space of $\boldsymbol{X}_T$, while the second term $-\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger \boldsymbol{\eta}$ 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

$$\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2 = \|(\boldsymbol{I} - \boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger \boldsymbol{X}_T)\boldsymbol{\beta}_T\|^2 + \|\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger \boldsymbol{\eta}\|^2.$$

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

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

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

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,

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

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

$$\|\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\|^2 = \operatorname{tr}((\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger(\boldsymbol{X}_T\boldsymbol{X}_T^*)(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\boldsymbol{\eta}^*) = \operatorname{tr}((\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\boldsymbol{\eta}^*),$$

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

$$\mathbb{E}[\|\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\|^2] = \operatorname{tr}(\mathbb{E}[(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger]\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^*]).$$

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

$$\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^*] = (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \boldsymbol{I}.$$

The distribution of $\boldsymbol{P} := (\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger$ is inverse-Wishart with identity scale matrix $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ and $p$ degrees of freedom. Each diagonal entry $P_{i,i}$ of $\boldsymbol{P}$, for $i = 1, \ldots, n$, has a reciprocal that follows the $\chi^2$ distribution with $p - n + 1$ degrees of freedom. Hence, $\mathbb{E}[P_{i,i}] = 1/(p-n-1)$ if $p \geq n+2$ and $\mathbb{E}[P_{i,i}] = +\infty$ if $p \in \{n, n+1\}$. Therefore,

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

We conclude that

$$\mathbb{E}[\|\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\|^2] = \begin{cases} (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \frac{n}{p-n-1} & \text{if } p \geq n+2, \\ +\infty & \text{if } p \in \{n, n+1\}. \end{cases}$$

Combining the first and second terms gives the claimed expression for the risk. ∎

**2.3. Concentration.** We briefly consider the measure concentration of $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2$.

Theorem 2.3. *Consider the setting from Theorem 2.1, and fix any $\epsilon \in (0,1)$. If $\alpha := p/n < 1$, then*

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2 \in (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2)\left(1 + \left(\frac{1 \pm \epsilon}{1 \mp \epsilon}\right)\frac{p}{n-p+1}\right)$$

*with probability at least*

$$1 - 2\exp\left(-\frac{p\epsilon^4(\sqrt{\alpha^{-1}}-1)^2}{24((2-\epsilon)\sqrt{\alpha^{-1}}+\epsilon)^2}\right) - 2\exp\left(-\frac{p(1-\epsilon)^2(\sqrt{\alpha^{-1}}-1)^2}{2}\right) - 2p\exp\left(-\frac{p(\alpha^{-1}-1)\epsilon^2}{24}\right).$$

*If $\alpha > 1$, then*

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2 \in \|\boldsymbol{\beta}_T\|^2\left(1 - (1 \pm \epsilon)\frac{n}{p}\right) + (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2)\left(1 + \left(\frac{1 \pm \epsilon}{1 \mp \epsilon}\right)\frac{n}{p-n+1}\right)$$

*with probability at least*

$$1 - 2\exp\left(-\frac{n\epsilon^2}{12}\right) - 2\exp\left(-\frac{n\epsilon^4(\sqrt{\alpha}-1)^2}{24((2-\epsilon)\sqrt{\alpha}+\epsilon)^2}\right) - 2\exp\left(-\frac{n(1-\epsilon)^2(\sqrt{\alpha}-1)^2}{2}\right) - 2n\exp\left(-\frac{n(\alpha-1)\epsilon^2}{24}\right).$$

The proof is given in Appendix A. The main idea for the $p > n$ case is as follows. From the proof of Theorem 2.1, we have the decomposition

$$\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2 = \|(\boldsymbol{I} - \boldsymbol{\Pi}_T)\boldsymbol{\beta}_T\|^2 + \|\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\|^2.$$

The first term $\|(\boldsymbol{I} - \boldsymbol{\Pi}_T)\boldsymbol{\beta}_T\|^2$ is the squared distance from $\boldsymbol{\beta}_T$ to a uniformly random $n$-dimensional subspace of $\mathbb{R}^p$. This squared distance has the same distribution as the squared distance from a uniformly random vector of length $\|\boldsymbol{\beta}_T\|$ to a fixed $n$-dimensional subspace of $\mathbb{R}^p$. Thus, measure concentration on the unit sphere can be used here. The second term $\|\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\|^2$ is a (random) quadratic form in the Gaussian random vector $\boldsymbol{\eta}$. Gaussian concentration is readily applied after controlling the spectral properties of the Wishart random matrix $\boldsymbol{X}_T\boldsymbol{X}_T^*$. (The $p < n$ case is similar to the analysis of this second term.)

The same arguments can be used to give fixed-level confidence bounds; see Proposition B.1 in Appendix B.

Finally, it is also possible to compare $\|\boldsymbol{\beta}_T\|^2$ to $(p/D)\|\boldsymbol{\beta}\|^2$ (and $\|\boldsymbol{\beta}_{T^c}\|^2$ to $(1-p/D)\|\boldsymbol{\beta}\|^2$) under the random selection model of $T$ from Corollary 2.2 using concentration inequalities for sampling without replacement; see, e.g., [2], for a discussion. The following is a simple consequence of Proposition 1.4 of [2].

Proposition 2.4. *For any $t > 0$, with probability at least $1 - 2e^{-t}$,*

$$\left| \|\boldsymbol{\beta}_T\|^2 - \frac{p}{D}\|\boldsymbol{\beta}\|^2 \right| = \left| \|\boldsymbol{\beta}_{T^c}\|^2 - \left(1 - \frac{p}{D}\right)\|\boldsymbol{\beta}\|^2 \right| \leq \|\boldsymbol{\beta}\|^2 \left( \sqrt{2\left(\mu^2 - \frac{1}{D}\right)\min\left\{\frac{p}{D}, 1 - \frac{p}{D}\right\} t} + \frac{2\mu^2 t}{3} \right),$$

*where $\mu := \max_{i \in [D]} |\beta_i| / \|\boldsymbol{\beta}\|$.*

The proof is in Appendix C. The crucial parameter $\mu$ has range $[1/\sqrt{D}, 1]$. It is small when there are many relevant "weak" features, each with a relatively small coefficient in $\boldsymbol{\beta}$; conversely, it is large when $\boldsymbol{\beta}$ is concentrated on a sparse subset of features.

## 3. Fourier series model.

In this section, we consider a noise-free Fourier series model which can be regarded as a one-dimensional version of the random Fourier features model studied by Rahimi and Recht [18] for functions defined on the unit circle.

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 random subsets of $[D]$. For any $i \in [D]$, the membership of $i$ in $S$ (respectively, $T$) is determined by an independent Bernoulli variable with mean $\rho_n := n/D$ (respectively, $\rho_p := p/D$).
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 $S$.

We fit regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_D)$ with

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

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, we have

$$\boldsymbol{F}_{S,T}^{\dagger} = \begin{cases} \boldsymbol{F}_{S,T}^T (\boldsymbol{F}_{S,T} \boldsymbol{F}_{S,T}^T)^{-1}, & |T| \geq |S|, \\ (\boldsymbol{F}_{S,T}^T \boldsymbol{F}_{S,T})^{-1} \boldsymbol{F}_{S,T}^T, & |T| \leq |S|. \end{cases}$$

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

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

(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 $\rho_p > \rho_n$.

Following the arguments from subsection 2.1, we have

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

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

(3.1)
$$\mathbb{E}[\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|^2 \mid S,T] = 1-\frac{1}{D}\cdot\operatorname{tr}((\boldsymbol{F}_{S,T}^{\dagger}\boldsymbol{F}_{S,T})^T(\boldsymbol{F}_{S,T}^{\dagger}\boldsymbol{F}_{S,T})) + \frac{1}{D}\cdot\operatorname{tr}((\boldsymbol{F}_{S,T}^{\dagger}\boldsymbol{F}_{S,T^c})^T(\boldsymbol{F}_{S,T}^{\dagger}\boldsymbol{F}_{S,T^c})).$$

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

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

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

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

where the $\lambda_i \in [0,1]$ are the eigenvalues of $\boldsymbol{F}_{S,T^c}\boldsymbol{F}_{S,T^c}^T$. Therefore, from (3.1), we have

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

To determine the asymptotic behavior of $(*)$, we use a recent result of Farrell [10],

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

as $D,n,p\to\infty$ with $\rho_n = n/D$ and $\rho_p = p/D$ held fixed. Further, under this limit, we have

$$\mathbb{E}\min\left\{\frac{|S|}{D},\frac{|T|}{D}\right\} \to \rho_n$$

since $\rho_p \geq \rho_n$. Hence, we have the following.

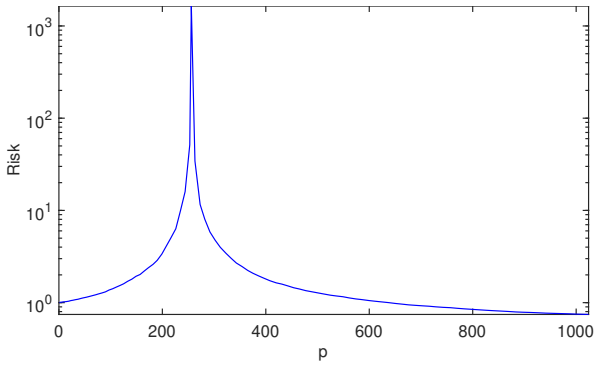


**Figure 3.** *Plot of risk as a function of $p$ in the Fourier series model. Here, $\boldsymbol{\beta}$ was chosen uniformly at random (once) from the unit sphere in $\mathbb{R}^D$ for $D = 1024$. We then computed $\hat{\boldsymbol{\beta}}$ from 10 independent random choices of $S$ (with $n = 256$)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

Theorem 3.1. *Assume the setting as above, with $D, n, p \to \infty$ and $\rho_n = n/D$ and $\rho_p = p/D$ held fixed. Then*

$$\lim \mathbb{E}\left[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2\right] = 1 - \frac{n}{D}\left(2 - \frac{p(1-n/D)}{p-n}\right).$$

Note that the right-hand side in the equation from Theorem 3.1 is well defined in the limit because the ratios $\rho_n, \rho_p$ are fixed. It diverges to $+\infty$ when $\rho_p$ is close to $\rho_n$ and decreases as $\rho_p$ approaches 1. This is the same behavior as in the Gaussian model from section 2 with random feature selection; we depict a nonasymptotic instantiation of it in Figure 3.

**4. Discussion.** Our analysis shows 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 as suggested by classical analyses. This choice is simple, both conceptually and algorithmically (although it may incur a computational penalty for processing large numbers of parameters), and avoids the need for precise control of regularization parameters. It is reflective of the practice in modern 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, notably neural networks, are highly successful. This stands in contrast to the "scientific" scenarios with informed selection of features; for example, in many science and medical applications, features are purposefully chosen based on the detailed understanding of the underlying phenomena. As illustrated by the "prescient" model that selects the best features, in that case 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.

Finally, we remark that there appears to be a sharp divide between the classical analyses of statistics and machine learning in $p < n$ regimes and the modern "weak but plentiful features" interpolating settings. While the former are deeply explored, an understanding of the latter is only starting to emerge. It is clear that the best practices for model and feature selection depend crucially on the regime of the application.

**Appendix A. Proof of Theorem 2.3.** We first consider $p > n$ (i.e., $\alpha > 1$). From the proof of Theorem 2.1, we have the decomposition

$$\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2 = \|(\boldsymbol{I} - \boldsymbol{\Pi}_T)\boldsymbol{\beta}_T\|^2 + \|\boldsymbol{X}_T^*(\boldsymbol{X}_T\boldsymbol{X}_T^*)^\dagger\boldsymbol{\eta}\|^2,$$

where $\boldsymbol{\Pi}_T$ is the orthogonal projection matrix for the row space of $\boldsymbol{X}_T$ and $\boldsymbol{\eta}$ is normal with mean zero and covariance $(\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2)\boldsymbol{I}$ and independent of $\boldsymbol{X}_T$. By symmetry of the standard normal distribution, the first term $\|(\boldsymbol{I} - \boldsymbol{\Pi}_T)\boldsymbol{\beta}_T\|^2$ is the squared distance from $\boldsymbol{\beta}_T$ to a uniformly random $n$-dimensional subspace of $\mathbb{R}^p$. This squared distance has the same distribution as the squared distance from a uniformly random vector of length $\|\boldsymbol{\beta}_T\|$ to a fixed $n$-dimensional subspace of $\mathbb{R}^p$. This argument was also used by Dasgupta and Gupta [9] in their proof of the Johnson–Lindenstrauss lemma. By Lemma 2.2 from [9], we have, for any $\epsilon \in (0, 1)$,

$$\Pr\left[\|(\boldsymbol{I} - \boldsymbol{\Pi}_T)\boldsymbol{\beta}_T\|^2 \notin \left(1 - (1 \pm \epsilon)\frac{n}{p}\right)\|\boldsymbol{\beta}_T\|^2\right] \le 2\exp\left(-\frac{n\epsilon^2}{12}\right).$$

The second term $\|\boldsymbol{X}_T^T(\boldsymbol{X}_T\boldsymbol{X}_T^T)^\dagger\boldsymbol{\eta}\|^2$ is a (random) quadratic form in $\boldsymbol{\eta}$. Let $\boldsymbol{K}_T := \boldsymbol{X}_T\boldsymbol{X}_T^T$, which is nonsingular almost surely. By Lemma 4 from [8], we have, for any $\epsilon \in (0, 1)$,

$$\Pr\left[\|\boldsymbol{X}_T^T(\boldsymbol{X}_T\boldsymbol{X}_T^T)^\dagger\boldsymbol{\eta}\|^2 \notin (1 \pm \epsilon)(\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2)\mathrm{tr}(\boldsymbol{K}_T^{-1}) \mid \boldsymbol{K}_T \text{ nonsingular}\right] \le 2\exp\left(-\frac{n\epsilon^2}{24\kappa(\boldsymbol{X}_T)^2}\right),$$

where $\kappa(\boldsymbol{X}_T) = \sigma_{\max}(\boldsymbol{X}_T)/\sigma_{\min}(\boldsymbol{X}_T)$ is the ratio of the largest singular value of $\boldsymbol{X}_T$ to the smallest singular value of $\boldsymbol{X}_T$. For any $t > 0$,

$$\Pr\left[\sigma_{\max}(\boldsymbol{X}_T) \ge \sqrt{p} + (1 + t)\sqrt{n}\right] \le \exp(-nt^2/2),$$
$$\Pr\left[\sigma_{\min}(\boldsymbol{X}_T) \le \sqrt{p} - (1 + t)\sqrt{n}\right] \le \exp(-nt^2/2).$$

These inequalities follow from Gaussian comparison inequalities and concentration of measure on the sphere and in Gaussian space; see, e.g., [19, 22]. Therefore, for $p > (1 + t)^2 n$,

$$\Pr\left[\kappa(\boldsymbol{X}_T)^2 \ge \left(\frac{\sqrt{p} + (1 + t)\sqrt{n}}{\sqrt{p} - (1 + t)\sqrt{n}}\right)^2\right] \le 2\exp\left(-\frac{nt^2}{2}\right).$$

Finally, observe that $1/(\boldsymbol{K}_T^{-1})_{i,i}$ has a $\chi^2$-distribution with $p - n + 1$ degrees of freedom. Therefore, again using Lemma 4 from [8] and a union bound, we have, for any $\epsilon \in (0, 1)$,

$$\Pr\left[\mathrm{tr}(\boldsymbol{K}_T^{-1}) \notin \frac{n}{p - n + 1} \cdot \frac{1}{1 \mp \epsilon}\right] \le 2n\exp\left(-\frac{(p - n + 1)\epsilon^2}{24}\right).$$

Putting these probability inequalities together (with $t = (1 - \epsilon)(\sqrt{\alpha} - 1)$) completes the proof for $p > n$.

Now we consider $p < n$ (i.e., $\alpha < 1$). We have

$$\hat{\boldsymbol{\beta}}_T = (\boldsymbol{X}_T^* \boldsymbol{X}_T)^\dagger \boldsymbol{X}_T^* (\boldsymbol{X}_T \boldsymbol{\beta}_T + \boldsymbol{\eta}).$$

The matrix $\boldsymbol{X}_T^* \boldsymbol{X}_T$ is nonsingular almost surely, so $\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}\|^2 = \boldsymbol{\eta}^* (\boldsymbol{X}_T \boldsymbol{X}_T^*)^\dagger \boldsymbol{\eta} = \boldsymbol{\eta}^* \boldsymbol{K}_T^\dagger \boldsymbol{\eta}$ also holds almost surely. Note that $\boldsymbol{K}_T$ has the same eigenvalues as $\boldsymbol{X}_T^* \boldsymbol{X}_T$, and hence $\boldsymbol{K}_T^\dagger$ has the same eigenvalues as $(\boldsymbol{X}_T^* \boldsymbol{X}_T)^{-1}$. Therefore, following essentially the same arguments as above for handling $\|\boldsymbol{X}_T^* (\boldsymbol{X}_T \boldsymbol{X}_T^*)^\dagger \boldsymbol{\eta}\|^2$ (but switching the roles of $p$ and $n$ and hence replacing $\alpha$ with $\alpha^{-1}$) completes the proof for $p < n$. ∎

## Appendix B. Confidence bounds.

Fixed-level confidence bounds can be immediately derived from the probability inequalities in Appendix A.

Proposition B.1. *Consider the setting from Theorem 2.1, and fix any $\delta \in (0,1)$. If $p < n$, then with probability at least $1 - \delta$,*

$$\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2 \in \left(1 \pm \frac{1 + \sqrt{\frac{p}{n}} + \sqrt{\frac{2\ln(8/\delta)}{n}}}{1 - \sqrt{\frac{p}{n}} - \sqrt{\frac{2\ln(8/\delta)}{n}}} \cdot \sqrt{\frac{48\ln(256/\delta)}{p}}\right) \cdot (\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \frac{p}{n-p+1} \cdot \frac{1}{1 \mp \sqrt{\frac{24\ln(8p/\delta)}{n-p+1}}}.$$

*If $p > n$, then with probability at least $1 - \delta$,*

$$\|\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\|^2 \in \left(1 - \left(1 \pm \sqrt{\frac{12\ln(8/\delta)}{n}}\right)\frac{n}{p}\right)\|\boldsymbol{\beta}_T\|^2 + \left(1 \pm \frac{1 + \sqrt{\frac{n}{p}} + \sqrt{\frac{2\ln(8/\delta)}{p}}}{1 - \sqrt{\frac{n}{p}} - \sqrt{\frac{2\ln(8/\delta)}{p}}} \cdot \sqrt{\frac{48\ln(256/\delta)}{n}}\right)(\|\boldsymbol{\beta}_{T^c}\|^2 + \sigma^2) \cdot \frac{n}{p-n+1} \cdot \frac{1}{1 \mp \sqrt{\frac{24\ln(8n/\delta)}{p-n+1}}}.$$

*In the expressions above, we assume $n$ and $p$ are large enough (perhaps in relation to each other) so that all denominators are positive.*

## Appendix C. Proof of Proposition 2.4.

Let $X_1, \ldots, X_p$ be a random sample of cardinality $p$ from the finite population $(\beta_1^2, \ldots, \beta_D^2)$, drawn without replacement, so that $\|\boldsymbol{\beta}_T\|^2 = \sum_{j=1}^p X_j$. Since $\|\boldsymbol{\beta}_{T^c}\|^2 = \|\boldsymbol{\beta}\|^2 - \|\boldsymbol{\beta}_T\|^2$, we have

$$\left|\|\boldsymbol{\beta}_T\|^2 - \frac{p}{D}\|\boldsymbol{\beta}\|^2\right| = \left|\|\boldsymbol{\beta}_{T^c}\|^2 - \left(1 - \frac{p}{D}\right)\|\boldsymbol{\beta}\|^2\right|.$$

Observe that the finite population $(\beta_1^2, \ldots, \beta_D^2)$ has mean $\frac{1}{D}\|\boldsymbol{\beta}\|^2$, variance $\frac{1}{D}\sum_{j=1}^D \beta_j^4 - (\frac{1}{D}\sum_{j=1}^D \beta_j^2)^2 \le \frac{1}{D}\|\boldsymbol{\beta}\|^4\mu^2 - (\frac{1}{D}\|\boldsymbol{\beta}\|^2)^2 = \frac{1}{D}\|\boldsymbol{\beta}\|^4(\mu^2 - \frac{1}{D})$, and range $\max_{j\in[D]} \beta_j^2 = \|\boldsymbol{\beta}\|^2\mu^2$.

Therefore, Proposition 1.4 of [2] and a union bound implies, with probability at least $1-2e^{-t}$,

$$\left|\|\boldsymbol{\beta}_T\|^2 - \frac{p}{D}\|\boldsymbol{\beta}\|^2\right| = \left|\|\boldsymbol{\beta}_{T^c}\|^2 - \left(1-\frac{p}{D}\right)\|\boldsymbol{\beta}\|^2\right| \le \|\boldsymbol{\beta}\|^2\left(\sqrt{2\left(\mu^2-\frac{1}{D}\right)\frac{pt}{D}}+\frac{2\mu^2 t}{3}\right).$$

If $p/D$ is more than $1/2$, then we can replace $p/D$ by $1-p/D$ on the right-hand side by analogously applying the previous argument to the random sample of cardinality $D-p$ that determines $\boldsymbol{\beta}_{T^c}$. ∎

**Acknowledgments.** We thank the anonymous referees for their remarks and suggestions (which, in particular, led to the inclusion of subsection 2.3). This work was carried out in part while MB was at Ohio State University.

## REFERENCES

[1] M. S. Advani and A. M. Saxe, *High-Dimensional Dynamics of Generalization Error in Neural Networks*, preprint, arXiv:1710.03667, 2017.
[2] R. Bardenet and O.-A. Maillard, *Concentration inequalities for sampling without replacement*, Bernoulli, 21 (2015), pp. 1361–1385.
[3] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, *Benign overfitting in linear regression*, Proc. Natl. Acad. Sci. USA, (2020), doi:10.1073/pnas.1907378117.
[4] M. Belkin, D. Hsu, S. Ma, and S. Mandal, *Reconciling modern machine learning practice and the bias-variance trade-off*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 15849–15854.
[5] M. Belkin, D. Hsu, and J. Xu, *Two Models of Double Descent for Weak Features*, preprint, arXiv:1903.07571v1, 2019.
[6] S. Bös and M. Opper, *Dynamics of batch training in a perceptron*, J. Phys. A, 31 (1998), 4835.
[7] L. Breiman and D. Freedman, *How many variables should be entered in a regression equation?*, J. Amer. Statist. Assoc., 78 (1983), pp. 131–136.
[8] S. Dasgupta, *Learning Probability Distributions*, Ph.D. thesis, University of California, Berkeley, 2000.
[9] S. Dasgupta and A. Gupta, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures Algorithms, 22 (2003), pp. 60–65.
[10] B. Farrell, *Limiting empirical singular value distribution of restrictions of discrete Fourier transform matrices*, J. Fourier Anal. Appl., 17 (2011), pp. 733–753.
[11] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, *Surprises in High-Dimensional Ridgeless Least Squares Interpolation*, preprint, arXiv:1903.08560, 2019.
[12] A. Krogh and J. A. Hertz, *Generalization in a linear perceptron in the presence of noise*, J. Phys. A, 25 (1992), 1135.
[13] Y. Le Cun, I. Kanter, and S. A. Solla, *Eigenvalues of covariance matrices: Application to neural-network learning*, Phys. Rev. Lett., 66 (1991), 2396.
[14] M. Loog, T. Viering, A. Mey, J. H. Krijthe, and D. M. Tax, *A brief prehistory of double descent*, Proc. Natl. Acad. Sci. USA, 117 (2020), pp. 10625–10626.
[15] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, *Harmless interpolation of noisy data in regression*, IEEE J. Sel. Areas Inform. Theory, (2020), pp. 67–83.
[16] B. Neal, S. Mittal, A. Baratin, V. Tantia, M. Scicluna, S. Lacoste-Julien, and I. Mitliagkas, *A Modern Take on the Bias-Variance Tradeoff in Neural Networks*, preprint, arXiv:1810.08591, 2018.
[17] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, *On the ability of the optimal perceptron to generalise*, J. Phys. A, 23 (1990), L581.
[18] A. Rahimi and B. Recht, *Random features for large-scale kernel machines*, in Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press, 2008, pp. 1177–1184.
[19] M. Rudelson and R. Vershynin, *Smallest singular value of a random rectangular matrix*, Comm. Pure Appl. Math., 62 (2009), pp. 1707–1739.

[20] S. Spigler, M. Geiger, S. d'Ascoli, L. Sagun, G. Biroli, and M. Wyart, *A jamming transition from under- to over-parametrization affects generalization in deep learning*, J. Phys. A, 52 (2019), 474001.

[21] F. Vallet, J.-G. Cailton, and P. Refregier, *Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions*, Europhys. Lett., 9 (1989), 315.

[22] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Vol. 47, Cambridge University Press, Cambridge, 2018.

[23] T. L. Watkin, A. Rau, and M. Biehl, *The statistical mechanics of learning a rule*, Rev. Mod. Phys., 65 (1993), 499.

[24] J. Xu and D. Hsu, *On the number of variables to use in principal component regression*, in Advances in Neural Information Processing Systems 32, Cambridge, MA, MIT Press, 2019, pp. 5095–5104.