### Title

Benign overfitting in linear regression

### Permalink

https://escholarship.org/uc/item/0v43g8r5

### Journal

Proceedings of the National Academy of Sciences of the United States of America, 117(48)

### ISSN

0027-8424

### Authors

Bartlett, Peter L
Long, Philip M
Lugosi, Gábor
et al.

### Publication Date

2020-12-01

### DOI

10.1073/pnas.1907378117

Peer reviewed

# Benign overfitting in linear regression

Peter L. Bartlett[a,b,1], Philip M. Long[c], Gábor Lugosi[d,e,f], and Alexander Tsigler[a]

[a]Department of Statistics, University of California, Berkeley, CA 94720-3860; [b]Computer Science Division, University of California, Berkeley, CA 94720-1776; [c]Google Brain, Mountain View, CA 94043; [d]Economics and Business, Pompeu Fabra University, 08005 Barcelona, Spain; [e]Institució Catalana de Recerca i Estudis Avançats, Passeig, Lluís Companys 23, 08010 Barcelona, Spain; and [f]Barcelona Graduate School of Economics, 08005 Barcelona, Spain

Edited by Richard Baraniuk, Rice University, Houston, TX, and accepted by Editorial Board Member David L. Donoho March 4, 2020 (received for review June 2, 2019)

**The phenomenon of benign overfitting is one of the key mysteries uncovered by deep learning methodology: deep neural networks seem to predict well, even with a perfect fit to noisy training data. Motivated by this phenomenon, we consider when a perfect fit to training data in linear regression is compatible with accurate prediction. We give a characterization of linear regression problems for which the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization is in terms of two notions of the effective rank of the data covariance. It shows that overparameterization is essential for benign overfitting in this setting: the number of directions in parameter space that are unimportant for prediction must significantly exceed the sample size. By studying examples of data covariance properties that this characterization shows are required for benign overfitting, we find an important role for finite-dimensional data: the accuracy of the minimum norm interpolating prediction rule approaches the best possible accuracy for a much narrower range of properties of the data distribution when the data lie in an infinite-dimensional space vs. when the data lie in a finite-dimensional space with dimension that grows faster than the sample size.**

statistical learning theory | overfitting | linear regression | interpolation

Deep learning methodology has revealed a surprising statistical phenomenon: overfitting can perform well. The classical perspective in statistical learning theory is that there should be a tradeoff between the fit to the training data and the complexity of the prediction rule. Whether complexity is measured in terms of the number of parameters, the number of nonzero parameters in a high-dimensional setting, the number of neighbors averaged in a nearest neighbor estimator, the scale of an estimate in a reproducing kernel Hilbert space, or the bandwidth of a kernel smoother, this tradeoff has been ubiquitous in statistical learning theory. Deep learning seems to operate outside the regime where results of this kind are informative since deep neural networks can perform well even with a perfect fit to the training data.

As one example of this phenomenon, consider the experiment illustrated in figure 1*C* in ref. 1: standard deep network architectures and stochastic gradient algorithms, run until they perfectly fit a standard image classification training set, give respectable prediction performance, even when significant levels of label noise are introduced. The deep networks in the experiments reported in ref. 1 achieved essentially zero cross-entropy loss on the training data. In statistics and machine learning textbooks, an estimate that fits every training example perfectly is often presented as an illustration of overfitting ["...interpolating fits...[are] unlikely to predict future data well at all" (ref. 2, p. 37)]. Thus, to arrive at a scientific understanding of the success of deep learning methods, it is a central challenge to understand the performance of prediction rules that fit the training data perfectly.

In this paper, we consider perhaps the simplest setting where we might hope to witness this phenomenon: linear regression. That is, we consider quadratic loss and linear prediction rules, and we assume that the dimension of the parameter space is large enough that a perfect fit is guaranteed. We consider data in an infinite-dimensional space (a separable Hilbert space), but our results apply to a finite-dimensional subspace as a special case. There is an ideal value of the parameters, $\theta^*$, corresponding to the linear prediction rule that minimizes the expected quadratic loss. We ask when it is possible to fit the data exactly and still compete with the prediction accuracy of $\theta^*$. Since we require more parameters than the sample size in order to fit exactly, the solution might be underdetermined, and therefore, there might be many interpolating solutions. We consider the most natural: choose the parameter vector $\hat{\theta}$ with the smallest norm among all vectors that gives perfect predictions on the training sample. (This corresponds to using the pseudoinverse to solve the normal equations; see below.) We ask when it is possible to overfit in this way—and embed all of the noise of the labels into the parameter estimate $\hat{\theta}$—without harming prediction accuracy.

Our main result is a finite sample characterization of when overfitting is benign in this setting. The linear regression problem depends on the optimal parameters $\theta^*$ and the covariance $\Sigma$ of the covariates $x$. The properties of $\Sigma$ turn out to be crucial since the magnitude of the variance in different directions determines both how the label noise gets distributed across the parameter space and how errors in parameter estimation in different directions in parameter space affect prediction accuracy. There is a classical decomposition of the excess prediction error into two terms. The first is rather standard: provided that the scale of the problem (that is, the sum of the eigenvalues of $\Sigma$) is small compared with the sample size $n$, the contribution to $\hat{\theta}$ that we can view as coming from $\theta^*$ is not too distorted. The second term is more interesting since it reflects the impact of the noise in the labels on prediction accuracy. We show that this part is small if and only if the effective rank of $\Sigma$ in the subspace corresponding to low-variance directions is large compared with $n$. This necessary and sufficient condition of a large effective rank can be viewed as a property of significant overparameterization: fitting the training data exactly but with near-optimal prediction accuracy occurs if and only if there are many low-variance (and

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "The Science of Deep Learning," held March 13–14, 2019, at the National Academy of Sciences in Washington, DC. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019 colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/science-of-deep-learning.

Author contributions: P.L.B., P.M.L., G.L., and A.T. designed research, performed research, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.B. is a guest editor invited by the Editorial Board.

[1] To whom correspondence may be addressed. Email: peter@berkeley.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1907378117/-/DCSupplemental.

First published April 24, 2020.

hence, unimportant) directions in parameter space where the label noise can be hidden.

The details are more complicated. The characterization depends in a specific way on two notions of effective rank, $r$ and $R$; the smaller one, $r$, determines a split of $\Sigma$ into large and small eigenvalues, and the excess prediction error depends on the effective rank as measured by the larger notion $R$ of the subspace corresponding to the smallest eigenvalues. For the excess prediction error to be small, the smallest eigenvalues of $\Sigma$ must decay slowly.

Studying the patterns of eigenvalues that allow benign overfitting reveals an interesting role for large but finite dimensions: in an infinite-dimensional setting, benign overfitting occurs only for a narrow range of decay rates of the eigenvalues. On the other hand, it occurs with any suitably slowly decaying eigenvalue sequence in a finite-dimensional space with dimension that grows faster than the sample size. Thus, for linear regression, data that lie in a large but finite-dimensional space exhibit the benign overfitting phenomenon with a much wider range of covariance properties than data that lie in an infinite-dimensional space.

The phenomenon of interpolating prediction rules has been an object of study by several authors over the last two years since it emerged as an intriguing mystery at the Simons Institute program on Foundations of Machine Learning in the spring of 2017. Belkin et al. (3) described an experimental study demonstrating that this phenomenon of accurate prediction for functions that interpolate noisy data also occurs for prediction rules chosen from reproducing kernel Hilbert spaces and explained the mismatch between this phenomenon and classical generalization bounds. Belkin et al. (4) gave an example of an interpolating decision rule—simplicial interpolation—with an asymptotic consistency property as the input dimension gets large. That work and subsequent work of Belkin et al. (5) studied kernel smoothing methods based on singular kernels that both interpolate and, with suitable bandwidth choice, give optimal rates for nonparametric estimation [building on earlier consistency results (6) for these unusual kernels]. Liang and Rakhlin (7) considered minimum norm interpolating kernel regression with kernels defined as nonlinear functions of the Euclidean inner product and showed that, with certain properties of the training sample (expressed in terms of the empirical kernel matrix), these methods can have good prediction accuracy. Belkin et al. (8) studied experimentally the excess risk as a function of the dimension of a sequence of parameter spaces for linear and nonlinear classes.

Subsequent to our work, ref. 9 considered the properties of the interpolating linear prediction rule with minimal expected squared error. After this work was presented at the NAS Colloquium on the Science of Deep Learning (10), we became aware of the concurrent work of Belkin et al. (11) and of Hastie et al. (12). Belkin et al. (11) calculated the excess risk for certain linear models (a regression problem with identity covariance and sparse optimal parameters, both with and without noise, and a problem with random Fourier features with no noise), and Hastie et al. (12) considered linear regression in an asymptotic regime, where sample size $n$ and input dimension $p$ go to infinity together with asymptotic ratio $p/n \to \gamma$. They assumed that, as $p$ gets large, the empirical spectral distribution of $\Sigma$ (the discrete measure on its set of eigenvalues) converges to a fixed measure, and they applied random matrix theory to explore the range of behaviors of the asymptotics of the excess prediction error as $\gamma$, the noise variance, and the eigenvalue distribution vary. They also studied the asymptotics of a model involving random nonlinear features. In contrast, we give upper and lower bounds on the excess prediction error for arbitrary finite sample size, for arbitrary covariance matrices, and for data of arbitrary dimension.

The next section introduces notation and definitions used throughout the paper, including definitions of the problem of linear regression and of various notions of effective rank of the covariance operator. The following section gives the characterization of benign overfitting, illustrates why the effective rank condition corresponds to significant overparameterization, and presents several examples of patterns of eigenvalues that allow benign overfitting, suggesting that slowly decaying covariance eigenvalues in input spaces of growing but finite dimension are the generic example of benign overfitting. Then we discuss the connections between these results and the benign overfitting phenomenon in deep neural networks and outline the proofs of the results.

## Definitions and Notation

We consider linear regression problems, where a linear function of covariates $x$ from a (potentially infinite-dimensional) Hilbert space $\mathbb{H}$ is used to predict a real-valued response variable $y$. We use vector notation, so that $x^\top \theta$ denotes the inner product between $x$ and $\theta$ and $xz^\top$ denotes the tensor product of $x, z \in \mathbb{H}$.

***Definition 1* (*Linear Regression*):** A linear regression problem in a separable Hilbert space $\mathbb{H}$ is defined by a random covariate vector $x \in \mathbb{H}$ and outcome $y \in \mathbb{R}$. We define

1) the covariance operator $\Sigma = \mathbb{E}[xx^\top]$ and
2) the optimal parameter vector $\theta^* \in \mathbb{H}$, satisfying $\mathbb{E}(y - x^\top\theta^*)^2 = \min_\theta \mathbb{E}(y - x^\top\theta)^2$.

We assume that

1) $x$ and $y$ are mean zero;
2) $x = V\Lambda^{1/2}z$, where $\Sigma = V\Lambda V^\top$ is the spectral decomposition of $\Sigma$ and $z$ has components that are independent $\sigma_x^2$ sub-Gaussian with $\sigma_x$ a positive constant: that is, for all $\lambda \in \mathbb{H}$,
$$\mathbb{E}[\exp(\lambda^\top z)] \le \exp(\sigma_x^2 \|\lambda\|^2/2),$$
where $\|\cdot\|$ is the norm in the Hilbert space $\mathbb{H}$;
3) the conditional noise variance is bounded below by some constant $\sigma^2$,
$$\mathbb{E}\left[(y - x^\top\theta^*)^2 \middle| x\right] \ge \sigma^2;$$
4) $y - x^\top\theta^*$ is $\sigma_y^2$ sub-Gaussian conditionally on $x$: that is, for all $\lambda \in \mathbb{R}$,
$$\mathbb{E}[\exp(\lambda(y - x^\top\theta^*))|x] \le \exp(\sigma_y^2\lambda^2/2)$$
(note that this implies $\mathbb{E}[y|x] = x^\top\theta^*$); and
5) almost surely, the projection of the data $X$ on the space orthogonal to any eigenvector of $\Sigma$ spans a space of dimension $n$.

Given a training sample $(x_1, y_1), \ldots, (x_n, y_n)$ of $n$ independent pairs with the same distribution as $(x, y)$, an estimator returns a parameter estimate $\theta \in \mathbb{H}$. The excess risk of the estimator is defined as
$$R(\theta) := \mathbb{E}_{x,y}\left[\left(y - x^\top\theta\right)^2 - \left(y - x^\top\theta^*\right)^2\right],$$
where $\mathbb{E}_{x,y}$ denotes the conditional expectation given all random quantities other than $x, y$ (in this case, given the estimate $\theta$). Define the vectors $\boldsymbol{y} \in \mathbb{R}^n$ with entries $y_i$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ with entries $\varepsilon_i = y_i - x_i^\top\theta^*$. We use infinite matrix notation: $X$ denotes the linear map from $\mathbb{H}$ to $\mathbb{R}^n$ corresponding to $x_1, \ldots, x_n \in \mathbb{H}$ so that $X\theta \in \mathbb{R}^n$ has $i$th component $x_i^\top\theta$. We use similar notation for the linear map $X^\top$ from $\mathbb{R}^n$ to $\mathbb{H}$.

Notice that Assumptions 1 to 5 are satisfied when $x$ and $y$ are jointly Gaussian with zero mean and $\operatorname{rank}(\Sigma) > n$.

We shall be concerned with situations where an estimator $\theta$ can fit the data perfectly: that is, $X\theta = \boldsymbol{y}$. Typically, this implies that there are many such vectors. We consider the interpolating estimator with minimal norm in $\mathbb{H}$. We use $\|\cdot\|$ to denote both the Euclidean norm of a vector in $\mathbb{R}^n$ and the Hilbert space norm.

***Definition 2 (Minimum Norm Estimator)*:** Given data $X \in \mathbb{H}^n$ and $\boldsymbol{y} \in \mathbb{R}^n$, the minimum norm estimator $\hat{\theta}$ solves the optimization problem

$$\begin{aligned} &\min_{\theta\in\mathbb{H}} && \|\theta\|^2 \\ &\text{such that} && \|X\theta-\boldsymbol{y}\|^2 = \min_{\beta}\|X\beta-\boldsymbol{y}\|^2. \end{aligned}$$

By the projection theorem, parameter vectors that solve the least squares problem $\min_\beta \|X\beta-\boldsymbol{y}\|^2$ solve the normal equations, and therefore, we can equivalently write $\hat{\theta}$ as the minimum norm solution to the normal equations

$$\begin{aligned} \hat{\theta} &= \arg\min_{\theta}\left\{\|\theta\|^2 : X^\top X\theta = X^\top \boldsymbol{y}\right\} \\ &= \left(X^\top X\right)^\dagger X^\top \boldsymbol{y} \\ &= X^\top\left(XX^\top\right)^\dagger \boldsymbol{y}, \end{aligned}$$

where $\left(X^\top X\right)^\dagger$ denotes the pseudoinverse of the bounded linear operator $X^\top X$ (for infinite-dimensional $\mathbb{H}$, the existence of the pseudoinverse is guaranteed because $X^\top X$ is bounded and has a closed range) (13). When $\mathbb{H}$ has dimension $p$ with $p<n$ and $X$ has rank $p$, there is a unique solution to the normal equations. On the contrary, Assumption 5 in Definition 1 implies that we can find many solutions $\theta\in\mathbb{H}$ to the normal equations that achieve $X\theta = y$. The minimum norm solution is given by

$$\hat{\theta} = X^\top\left(XX^\top\right)^{-1}\boldsymbol{y}. \qquad \textbf{[1]}$$

Our main result gives tight bounds on the excess risk of this minimum norm estimator in terms of certain notions of effective rank of the covariance that are defined in terms of its eigenvalues.

We use $\mu_1(\Sigma) \geq \mu_2(\Sigma) \geq \cdots$ to denote the eigenvalues of $\Sigma$ in descending order, and we denote the operator norm of $\Sigma$ by $\|\Sigma\|$. We use $I$ to denote the identity operator on $\mathbb{H}$ and $I_n$ to denote the $n\times n$ identity matrix.

***Definition 3 (Effective Ranks)*:** For the covariance operator $\Sigma$, define $\lambda_i = \mu_i(\Sigma)$ for $i = 1, 2, \ldots$. If $\sum_{i=1}^\infty \lambda_i < \infty$ and $\lambda_{k+1} > 0$ for $k \geq 0$, define

$$r_k(\Sigma) = \frac{\sum_{i>k}\lambda_i}{\lambda_{k+1}}, \qquad R_k(\Sigma) = \frac{\left(\sum_{i>k}\lambda_i\right)^2}{\sum_{i>k}\lambda_i^2}.$$

## Main Results

The following theorem establishes nearly matching upper and lower bounds for the risk of the minimum norm interpolating estimator.

**Theorem 1.** *For any $\sigma_x$, there are $b, c, c_1 > 1$ for which the following holds. Consider a linear regression problem from Definition 1. Define*

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

*where the minimum of the empty set is defined as $\infty$. Suppose that $\delta < 1$ with $\log(1/\delta) < n/c$. If $k^* \geq n/c_1$, then $\mathbb{E}R(\hat{\theta}) \geq \sigma^2/c$. Otherwise,*

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

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

$$\mathbb{E}R(\hat{\theta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

*Moreover, there are universal constants $a_1, a_2, n_0$ such that, for all $n \geq n_0$, for all $\Sigma$, and for all $t \geq 0$, there is a $\theta^*$ with $\|\theta^*\| = t$ such that, for $x \sim \mathcal{N}(0, \Sigma)$ and $y|x \sim \mathcal{N}(x^\top\theta^*, \|\theta^*\|^2\|\Sigma\|)$ with probability at least $1/4$,*

$$R(\hat{\theta}) \geq \frac{1}{a_1}\|\theta^*\|^2\,\|\Sigma\|\,\mathbb{1}\left[\frac{r_0(\Sigma)}{n\log\left(1+r_0(\Sigma)\right)} \geq a_2\right].$$

**Effective Ranks and Overparameterization.** In order to understand the implications of Theorem 1, we now study relationships between the two notions of effective rank, $r_k$ and $R_k$, and establish sufficient and necessary conditions for the sequence $\{\lambda_i\}$ of eigenvalues to lead to small excess risk.

The following lemma shows that the two notions of effective rank are closely related. *SI Appendix*, section H has its proof and other properties of $r_k$ and $R_k$.

**Lemma 1.** *$r_k(\Sigma) \geq 1$, $r_k^2(\Sigma) = r_k(\Sigma^2)R_k(\Sigma)$, and*

$$r_k(\Sigma^2) \leq r_k(\Sigma) \leq R_k(\Sigma) \leq r_k^2(\Sigma).$$

Notice that $r_0(I_p) = R_0(I_p) = p$. More generally, if all of the nonzero eigenvalues of $\Sigma$ are identical, then $r_0(\Sigma) = R_0(\Sigma) = \operatorname{rank}(\Sigma)$. For $\Sigma$ with finite rank, we can express both $r_0(\Sigma)$ and $R_0(\Sigma)$ as a product of the rank and a notion of symmetry. In particular, for $\operatorname{rank}(\Sigma) = p$, we can write

$$r_0(\Sigma) = \operatorname{rank}(\Sigma)s(\Sigma), \qquad R_0(\Sigma) = \operatorname{rank}(\Sigma)S(\Sigma),$$

$$\text{with } s(\Sigma) = \frac{\frac{1}{p}\sum_{i=1}^p\lambda_i}{\lambda_1}, \qquad S(\Sigma) = \frac{\left(\frac{1}{p}\sum_{i=1}^p\lambda_i\right)^2}{\frac{1}{p}\sum_{i=1}^p\lambda_i^2}.$$

Both notions of symmetry $s$ and $S$ lie between $1/p$ (when $\lambda_2 \to 0$) and 1 (when the $\lambda_i$ are all equal).

Theorem 1 shows that, for the minimum norm estimator to have near-optimal prediction accuracy, $r_0(\Sigma)$ should be small compared with the sample size $n$ (from the first term) and $r_{k^*}(\Sigma)$ and $R_{k^*}(\Sigma)$ should be large compared with $n$. Together, these conditions imply that overparameterization is essential for benign overfitting in this setting: the number of nonzero eigenvalues should be large compared with $n$, they should have a small sum compared with $n$, and there should be many eigenvalues no larger than $\lambda_{k^*}$. If the number of these small eigenvalues is not much larger than $n$, then they should be roughly equal, but they can be more asymmetric if there are many more of them.

The following theorem shows that the kind of overparameterization that is essential for benign overfitting requires $\Sigma$ to have a heavy tail. (The proof—and some other examples illustrating the boundary of benign overfitting—are in *SI Appendix*, section I.) In particular, if we fix $\Sigma$ in an infinite-dimensional Hilbert space and ask when the excess risk of the minimum norm estimator approaches zero as $n \to \infty$, it imposes tight restrictions on the eigenvalues of $\Sigma$. However, there are many other possibilities for these asymptotics if $\Sigma$ can change with $n$. Since rescaling $X$ affects the accuracy of the least norm interpolant in an obvious way, we may assume without loss of generality that $\|\Sigma\| = 1$. If we restrict our attention to this case, then informally, Theorem 1 implies that, when the covariance operator for data with $n$ examples is $\Sigma_n$, the least norm interpolant converges if $\frac{r_0(\Sigma_n)}{n} \to 0$, $\frac{k_n^*}{n} \to 0$, and $\frac{n}{R_{k_n^*}(\Sigma_n)} \to 0$ and only if $\frac{r_0(\Sigma_n)}{n\log(1+r_0(\Sigma_n))} \to 0$, $\frac{k_n^*}{n} \to 0$, and $\frac{n}{R_{k_n^*}(\Sigma_n)} \to 0$, where $k_n^* = \min\{k \geq 0 : r_k(\Sigma_n) \geq bn\}$ for the universal constant $b$ in Theorem 1. This motivates the following definition.

***Definition 4*:** A sequence of covariance operators $\Sigma_n$ with $\|\Sigma_n\|=1$ is benign if

$$\lim_{n\to\infty}\frac{r_0(\Sigma_n)}{n}=\lim_{n\to\infty}\frac{k_n^*}{n}=\lim_{n\to\infty}\frac{n}{R_{k_n^*}(\Sigma_n)}=0.$$

We give some examples of benign and nonbenign settings.

**Theorem 2.**

1) *If* $\mu_k(\Sigma)=k^{-\alpha}\ln^{-\beta}((k+1)e/2)$, *then* $\Sigma$ *is benign if and only if* $\alpha=1$ *and* $\beta>1$.
2) *If*

$$\mu_k(\Sigma_n)=\begin{cases}\gamma_k+\epsilon_n & \textit{if } k\le p_n,\\ 0 & \textit{otherwise}\end{cases}$$

*and* $\gamma_k=\Theta(\exp(-k/\tau))$, *then* $\Sigma_n$ *with* $\|\Sigma_n\|=1$ *is benign if and only if* $p_n=\omega(n)$ *and* $ne^{-o(n)}=\epsilon_n p_n=o(n)$. *Furthermore, for* $p_n=\Omega(n)$ *and* $\epsilon_n p_n=ne^{-o(n)}$,

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

Compare the situations described by Theorem 2.1 and 2.2. Theorem 2.1 shows that, for infinite-dimensional data with a fixed covariance, benign overfitting occurs if and only if the eigenvalues of the covariance operator decay just slowly enough for their sum to remain finite. Theorem 2.2 shows that the situation is very different if the data have finite dimension and a small amount of isotropic noise is added to the covariates. In that case, even if the eigenvalues of the original covariance operator (before the addition of isotropic noise) decay very rapidly, benign overfitting occurs if and only if both the dimension is large compared with the sample size and the isotropic component of the covariance is sufficiently small—but not exponentially small—compared with the sample size.

These examples illustrate the tension between the slow decay of eigenvalues that is needed for $k/n+n/R_k$ to be small and the summability of eigenvalues that is needed for $r_0(\Sigma)/n$ to be small. There are two ways to resolve this tension. First, in the infinite-dimensional setting, slow decay of the eigenvalues suffices—decay just fast enough to ensure summability—as shown by Theorem 2.1. (*SI Appendix*, section I gives another example—Theorem S14.2—where the eigenvalue decay is allowed to vary with $n$; in that case, $\Sigma_n$ is benign iff the decay rate gets close—but not too close—to $1/k$ as $n$ increases.) Second, the other way to resolve the tension is to consider a finite-dimensional setting (which ensures that the eigenvalues are summable), and in this case, arbitrarily slow decay is possible. Theorem 2.2 gives an example of this: eigenvalues that are all at least as large as a small constant. *SI Appendix*, section I gives other examples with a similar flavor, including a truncated infinite series that decays sufficiently slowly that the sum does not converge (*SI Appendix*, section I, Theorem S14.3). Theorem 2.1 shows that a very specific decay rate is required in infinite dimensions, which suggests that this is an unusual phenomenon in that case. The more generic scenario where benign overfitting will occur is demonstrated by Theorem 2.2, with eigenvalues that are either at least a constant or slowly decaying in a very high—but finite-dimensional—space.

## Proof

Throughout the proofs, we treat $\sigma_x$ (the sub-Gaussian norm of the covariates) as a constant. Therefore, we use the symbols $b$, $c$, $c_1$, $c_2,\ldots$ to refer to constants that only depend on $\sigma_x$. Their values are suitably large (and always at least one) but do not depend on any parameters of the problems that we consider other than $\sigma_x$. For universal constants that do not depend on any parameters of the problem at all, we use the symbol $a$. Also, whenever we sum over eigenvectors of $\Sigma$, the sum is restricted to eigenvectors with nonzero eigenvalues.

**Outline.** The first step is a standard decomposition of the excess risk into two pieces, a term that corresponds to the distortion that is introduced by viewing $\theta^*$ through the lens of the finite sample and a term that corresponds to the distortion introduced by the noise $\boldsymbol{\varepsilon}=\boldsymbol{y}-X\theta$. The impact of both sources of error in $\hat\theta$ on the excess risk is modulated by the covariance $\Sigma$, which gives different weight to different directions in parameter space.

**Lemma 2.** *The excess risk of the minimum norm estimator satisfies* $R(\hat\theta)\le 2\theta^{*\top}B\theta^*+c\sigma^2\log(1/\delta)\operatorname{tr}(C)$ *with probability at least* $1-\delta$ *over* $\epsilon$, *and* $\mathbb{E}_{\boldsymbol{\varepsilon}}R(\hat\theta)\ge\theta^{*\top}B\theta^*+\sigma^2\operatorname{tr}(C)$, *where*

$$B=\left(I-X^\top\left(XX^\top\right)^{-1}X\right)\Sigma\left(I-X^\top\left(XX^\top\right)^{-1}X\right),$$
$$C=\left(XX^\top\right)^{-1}X\Sigma X^\top\left(XX^\top\right)^{-1}.$$

The proof of this lemma is in *SI Appendix*, section A. *SI Appendix*, sections J and K give bounds on the term $\theta^{*\top}B\theta^*$. The heart of the proof is controlling $\operatorname{tr}(C)$.

Before continuing with the proof, let us make a quick digression to note that Lemma 2 already begins to give an idea that many low-variance directions are necessary for the least norm interpolator to succeed. Let us consider the extreme case that $p=n$ and $\Sigma=I$. In this case, $C=\left(XX^\top\right)^{-1}$. For Gaussian data, for instance, standard random matrix theory implies that, with high probability, the eigenvalues of $XX^\top$ will all be within a constant factor of $n$, which implies that $\operatorname{tr}(C)$ is bounded below by a constant, and then, Lemma 2 implies that the least norm interpolant's excess risk is at least a constant.

To prove that $\operatorname{tr}(C)$ can be controlled for suitable $\Sigma$, the first step is to express it in terms of sums of outer products of unit-covariance, independent, sub-Gaussian random vectors. We show that, when there is a $k^*$ with $k^*/n$ small and $r_{k^*}(\Sigma)/n$ large, all of the smallest eigenvalues of these matrices are suitably concentrated, and this implies that $\operatorname{tr}(C)$ is bounded above by

$$\min_{l\le k^*}\left(\frac{l}{n}+n\frac{\sum_{i>l}\lambda_i^2}{(\lambda_{k^*+1}r_{k^*}(\Sigma))^2}\right).$$

(Later, we show that the minimizer is $l=k^*$.) Next, we show that this expression is also a lower bound on $\operatorname{tr}(C)$ provided that there is such a $k^*$. Conversely, we show that, for any $k$ for which $r_k(\Sigma)$ is not large compared with $n$, $\operatorname{tr}(C)$ is at least as big as a constant times $\min(1,k/n)$. Combining shows that, when $k^*/n$ is small, $\operatorname{tr}(C)$ is upper and lower bounded by constant factors times

$$\frac{k^*}{n}+\frac{n}{R_{k^*}(\Sigma)}.$$

**Unit Variance Sub-Gaussians.** Our assumptions allow the trace of $C$ to be expressed as a function of many independent sub-Gaussian vectors.

**Lemma 3.** *Consider a covariance operator* $\Sigma$ *with* $\lambda_i=\mu_i(\Sigma)$ *and* $\lambda_n>0$. *Write its spectral decomposition* $\Sigma=\sum_j\lambda_j v_j v_j^\top$, *where the orthonormal* $v_j\in\mathbb{H}$ *are the eigenvectors corresponding to the* $\lambda_j$. *For* $i$ *with* $\lambda_i>0$, *define* $z_i=Xv_i/\sqrt{\lambda_i}$. *Then,*

$$\operatorname{tr}(C)=\sum_i\left[\lambda_i^2 z_i^\top\left(\sum_j\lambda_j z_j z_j^\top\right)^{-2}z_i\right],$$

*and these $z_i \in \mathbb{R}^n$ are independent $\sigma_x^2$ sub-Gaussian. Furthermore, for any $i$ with $\lambda_i > 0$, we have*

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top\right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

*where $A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top$.*

***Proof*:** By Assumption 2 in Definition 1, the random variables $x^\top v_i/\sqrt{\lambda_i}$ are independent $\sigma_x^2$ sub-Gaussian. We consider $X$ in the basis of eigenvectors of $\Sigma$, $Xv_i = \sqrt{\lambda_i} z_i$, to see that

$$XX^\top = \sum_i \lambda_i z_i z_i^\top, \qquad X\Sigma X^\top = \sum_i \lambda_i^2 z_i z_i^\top,$$

and therefore, we can write

$$\begin{aligned} \operatorname{tr}(C) &= \operatorname{tr}\left(\left(XX^\top\right)^{-1} X\Sigma X^\top \left(XX^\top\right)^{-1}\right) \\ &= \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top\right)^{-2} z_i\right]. \end{aligned}$$

For the second part, we use *SI Appendix*, section B, Lemma S3, which is a consequence of the Sherman–Woodbury–Morrison formula

$$\begin{aligned} \lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top\right)^{-2} z_i &= \lambda_i^2 z_i^\top \left(\lambda_i z_i z_i^\top + A_{-i}\right)^{-2} z_i \\ &= \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1+\lambda_i z_i^\top A_{-i}^{-1} z_i)^2}, \end{aligned}$$

by *SI Appendix*, section B, Lemma S3 for the case $k = 1$ and $Z = \sqrt{\lambda_i} z_i$. Note that $A_{-i}$ is invertible by Assumption 5 in Definition 1. □

The weighted sum of outer products of these sub-Gaussian vectors plays a central role in the rest of the proof. Define

$$A = \sum_i \lambda_i z_i z_i^\top, \quad A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top, \quad A_k = \sum_{i>k} \lambda_i z_i z_i^\top,$$

where the $z_i \in \mathbb{R}^n$ are independent vectors with independent $\sigma_x^2$ sub-Gaussian coordinates with unit variance defined in Lemma 3. Note that the vector $z_i$ is independent of the matrix $A_{-i}$, and therefore, in the last part of Lemma 3, all of the random quadratic forms are independent of the points where those forms are evaluated.

**Concentration of *A*.** The next step is to show that eigenvalues of $A$, $A_{-i}$, and $A_k$ are concentrated. The proof of the following inequality is in *SI Appendix*, section C. Recall that $\mu_1(A)$ and $\mu_n(A)$ denote the largest and the smallest eigenvalues of the $n \times n$ matrix $A$.

**Lemma 4.** *There is a constant $c$ such that, for any $k \geq 0$ with probability at least $1 - 2e^{-n/c}$,*

$$\frac{1}{c}\sum_{i>k} \lambda_i - c\lambda_{k+1} n \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c\left(\sum_{i>k} \lambda_i + \lambda_{k+1} n\right).$$

The following lemma uses this result to give bounds on the eigenvalues of $A_k$, which in turn, give bounds on some eigenvalues of $A_{-i}$ and $A$. For these upper and lower bounds to match up to a constant factor, the sum of the eigenvalues of $A_k$ should dominate the term involving its leading eigenvalue, which is a condition on the effective rank $r_k(\Sigma)$. The lemma shows that, after $r_k(\Sigma)$ is sufficiently large, all of the eigenvalues of $A_k$ are identical up to a constant factor.

**Lemma 5.** *There are constants $b, c \geq 1$ such that, for any $k \geq 0$, with probability at least $1 - 2e^{-n/c}$,*

1) *for all $i \geq 1$,*

$$\mu_{k+1}(A_{-i}) \leq \mu_{k+1}(A) \leq \mu_1(A_k) \leq c\left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right);$$

2) *for all $1 \leq i \leq k$,*

$$\mu_n(A) \geq \mu_n(A_{-i}) \geq \mu_n(A_k) \geq \frac{1}{c}\sum_{j>k} \lambda_j - c\lambda_{k+1} n;$$

*and*

3) *if $r_k(\Sigma) \geq bn$, then*

$$\frac{1}{c}\lambda_{k+1} r_k(\Sigma) \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c\lambda_{k+1} r_k(\Sigma).$$

***Proof*:** By Lemma 4, we know that, with probability at least $1 - 2e^{-n/c_1}$,

$$\begin{aligned} \frac{1}{c_1}\sum_{j>k} \lambda_j - c_1\lambda_{k+1} n &\leq \mu_n(A_k) \\ &\leq \mu_1(A_k) \leq c_1\left(\sum_{j>k} \lambda_j + \lambda_{k+1} n\right). \end{aligned}$$

First, the matrix $A - A_k$ has rank at most $k$ (as a sum of $k$ matrices of rank 1). Thus, there is a linear space $\mathscr{L}$ of dimension $n - k$ such that, for all $v \in \mathscr{L}$, $v^\top A v = v^\top A_k v \leq \mu_1(A_k)\|v\|^2$ and therefore, $\mu_{k+1}(A) \leq \mu_1(A_k)$.

Second, by the Courant–Fischer–Weyl Theorem, for all $i$ and $j$, $\mu_j(A_{-i}) \leq \mu_j(A)$ (*SI Appendix*, section G, Lemma S11). On the other hand, for $i \leq k$, $A_k \preceq A_{-i}$, and therefore, all of the eigenvalues of $A_{-i}$ are lower bounded by $\mu_n(A_k)$.

Finally, if $r_k(\Sigma) \geq bn$,

$$\begin{aligned} \sum_{j>k} \lambda_j + \lambda_{k+1} n &= \lambda_{k+1} r_k(\Sigma) + \lambda_{k+1} n \\ &\leq \left(1 + \frac{1}{b}\right)\lambda_{k+1} r_k(\Sigma), \\ \frac{1}{c_1}\sum_{j>k} \lambda_j - c_1\lambda_{k+1} n &= \frac{1}{c_1}\lambda_{k+1} r_k(\Sigma) - c_1\lambda_{k+1} n \\ &\geq \left(\frac{1}{c_1} - \frac{c_1}{b}\right)\lambda_{k+1} r_k(\Sigma). \end{aligned}$$

Choosing $b > c_1^2$ and $c > \max\left\{c_1 + 1/c_1, (1/c_1 - c_1/b)^{-1}\right\}$ gives the third claim of the lemma. □

**Upper Bound on the Trace Term.** **Lemma 6.** *There are constants $b, c \geq 1$ such that, if $0 \leq k \leq n/c$, $r_k(\Sigma) \geq bn$, and $l \leq k$, then with probability at least $1 - 7e^{-n/c}$,*

$$\operatorname{tr}(C) \leq c\left(\frac{l}{n} + n\frac{\sum_{i>l} \lambda_i^2}{(\sum_{i>k} \lambda_i)^2}\right).$$

The proof uses the following lemma and its corollary. Their proofs are in *SI Appendix*, section C.

**Lemma 7.** *Suppose that $\{\lambda_i\}_i^\infty$ is a nonincreasing sequence of nonnegative numbers such that $\sum_{i=1}^\infty \lambda_i < \infty$ and that $\{\xi_i\}_{i=1}^\infty$ are independent centered $\sigma$-subexponential random variables. Then, for* *some universal constant $a$ for any $t > 0$, with probability at least $1 - 2e^{-t}$,*

$$\left|\sum_i \lambda_i \xi_i\right| \le a\sigma \max\left(t\lambda_1, \sqrt{t \sum_i \lambda_i^2}\right).$$

**Corollary 1.** *Suppose that $z \in \mathbb{R}^n$ is a centered random vector with independent $\sigma^2$ sub-Gaussian coordinates with unit variances, $\mathscr{L}$ is a random subspace of $\mathbb{R}^n$ of codimension $k$, and $\mathscr{L}$ is independent of $z$. Then, for some universal constant $a$ and any $t > 0$, with probability at least $1 - 3e^{-t}$,*

$$\|z\|^2 \le n + a\sigma^2(t + \sqrt{nt}),$$
$$\|\Pi_{\mathscr{L}} z\|^2 \ge n - a\sigma^2(k + t + \sqrt{nt}),$$

*where $\Pi_{\mathscr{L}}$ is the orthogonal projection on $\mathscr{L}$.*

***Proof of Lemma 6*:** Fix $b$ to its value in Lemma 5. By Lemma 3,

$$\begin{aligned}\mathrm{tr}(C) &= \sum_i \lambda_i^2 z_i^\top A^{-2} z_i \\ &= \sum_{i=1}^{l} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} + \sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i. \qquad [2]\end{aligned}$$

First, consider the sum up to $l$. If $r_k(\Sigma) \ge bn$, Lemma 5 shows that, with probability at least $1 - 2e^{-n/c_1}$ for all $i \le k$, $\mu_n(A_{-i}) \ge \lambda_{k+1} r_k(\Sigma)/c_1$ and for all $i$, $\mu_{k+1}(A_{-i}) \le c_1 \lambda_{k+1} r_k(\Sigma)$. The lower bounds on the $\mu_n(A_{-i})$ imply that, for all $z \in \mathbb{R}^n$ and $1 \le i \le l$,

$$z^\top A_{-i}^{-2} z \le \frac{c_1^2 \|z\|^2}{(\lambda_{k+1} r_k(\Sigma))^2},$$

and the upper bounds on the $\mu_{k+1}(A_{-i})$ give

$$z^\top A_{-i}^{-1} z \ge (\Pi_{\mathscr{L}_i} z)^\top A_{-i}^{-1} \Pi_{\mathscr{L}_i} z \ge \frac{\|\Pi_{\mathscr{L}_i} z\|^2}{c_1 \lambda_{k+1} r_k(\Sigma)},$$

where $\mathscr{L}_i$ is the span of the $n - k$ eigenvectors of $A_{-i}$ corresponding to its smallest $n - k$ eigenvalues. Therefore, for $i \le l$,

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \le \frac{z_i^\top A_{-i}^{-2} z_i}{(z_i^\top A_{-i}^{-1} z_i)^2} \le c_1^4 \frac{\|z_i\|^2}{\|\Pi_{\mathscr{L}_i} z_i\|^4}. \qquad [3]$$

Next, we apply Corollary 1 $l$ times together with a union bound to show that, with probability at least $1 - 3e^{-t}$ for all $1 \le i \le l$,

$$\|z_i\|^2 \le n + a\sigma_x^2(t + \ln k + \sqrt{n(t + \ln k)}) \le c_2 n, \qquad [4]$$
$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \ge n - a\sigma_x^2(k + t + \ln k + \sqrt{n(t + \ln k)}) \ge n/c_3, \qquad [5]$$

provided that $t < n/c_0$ and $c > c_0$ for some sufficiently large $c_0$ (note that $c_2$ and $c_3$ only depend on $c_0$, $a$, and $\sigma_x$, and we can still take $c$ large enough in the end without changing $c_2$ and $c_3$). Combining Eqs. **3–5**, with probability at least $1 - 5e^{-n/c_0}$,

$$\sum_{i=1}^{l} \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \le c_4 \frac{l}{n}.$$

Second, consider the second sum in Eq. **2**. Lemma 5 shows that, on the same high-probability event that we considered in bounding the first half of the sum, $\mu_n(A) \ge \lambda_{k+1} r_k(\Sigma)/c_1$. Hence,

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \le \frac{c_1^2 \sum_{i>l} \lambda_i^2 \|z_i\|^2}{(\lambda_{k+1} r_k(\Sigma))^2}.$$

Notice that $\sum_{i>l} \lambda_i^2 \|z_i\|^2$ is a weighted sum of $\sigma_x^2$-subexponential random variables, with the weights given by the $\lambda_i^2$ in blocks of size $n$. Lemma 7 implies that, with probability at least $1 - 2e^{-t}$,

$$\begin{aligned}\sum_{i>l} \lambda_i^2 \|z_i\|^2 &\le n \sum_{i>l} \lambda_i^2 + a\sigma_x^2 \max\left(\lambda_{l+1}^2 t, \sqrt{tn \sum_{i>l} \lambda_i^4}\right) \\ &\le n \sum_{i>l} \lambda_i^2 + a\sigma_x^2 \max\left(t \sum_{i>l} \lambda_i^2, \sqrt{tn} \sum_{i>l} \lambda_i^2\right) \\ &\le c_5 n \sum_{i>l} \lambda_i^2\end{aligned}$$

because $t < n/c_0$. Combining the above gives

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \le c_6 n \frac{\sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}.$$

Finally, putting both parts together and taking $c > \max\{c_0, c_4, c_6\}$ gives the lemma. □

**Lower Bound on the Trace Term.** We first give a bound on a single term in the expression for $\mathrm{tr}(C)$ in Lemma 3 that holds regardless of $r_k(\Sigma)$. The proof is in *SI Appendix*, section D.

**Lemma 8.** *There is a constant $c$ such that, for any $i \ge 1$ with $\lambda_i > 0$ and any $0 \le k \le n/c$, with probability at least $1 - 5e^{-n/c}$,*

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \ge \frac{1}{cn}\left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2}.$$

We can extend these bounds to a lower bound on $\mathrm{tr}(C)$ using the following lemma. The proof is in *SI Appendix*, section E.

**Lemma 9.** *Suppose that $n \le \infty$, $\{\eta_i\}_{i=1}^n$ is a sequence of nonnegative random variables, and that $\{t_i\}_{i=1}^n$ is a sequence of nonnegative real numbers (at least one of which is strictly positive) such that, for some $\delta \in (0, 1)$ and any $i \le n$, $\Pr(\eta_i > t_i) \ge 1 - \delta$. Then,*

$$\Pr\left(\sum_{i=1}^{n} \eta_i \ge \frac{1}{2} \sum_{i=1}^{n} t_i\right) \ge 1 - 2\delta.$$

These two lemmas imply the following lower bound.

**Lemma 10.** *There are constants $c$ such that, for any $0 \le k \le n/c$ and any $b > 1$ with probability at least $1 - 10e^{-n/c}$,*

1) *if $r_k(\Sigma) < bn$, then $\mathrm{tr}(C) \ge \frac{k+1}{cb^2 n}$*; and
2) *if $r_k(\Sigma) \ge bn$, then*

$$\mathrm{tr}(C) \ge \frac{1}{cb^2} \min_{l \le k}\left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}\right).$$

*In particular, if all choices of $k \le n/c$ give $r_k(\Sigma) < bn$, then $r_{n/c}(\Sigma) < bn$ implies that, with probability at least $1 - 10e^{-n/c}$, $\mathrm{tr}(C) = \Omega_{\sigma_x}(1)$.*

***Proof*:** From Lemmas 3, 8, and 9, with probability at least $1 - 10e^{-n/c_1}$,

$$\begin{aligned}\mathrm{tr}(C) &\ge \frac{1}{c_1 n} \sum_i \left(1 + \frac{\sum_{j>k} \lambda_j + n\lambda_{k+1}}{n\lambda_i}\right)^{-2} \\ &\ge \frac{1}{c_2 n} \sum_i \min\left\{1, \frac{n^2 \lambda_i^2}{\left(\sum_{j>k} \lambda_j\right)^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\} \\ &\ge \frac{1}{c_2 b^2 n} \sum_i \min\left\{1, \left(\frac{bn}{r_k(\Sigma)}\right)^2 \frac{\lambda_i^2}{\lambda_{k+1}^2}, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\}.\end{aligned}$$

Now, if $r_k(\Sigma) < bn$, then the second term in the minimum is always bigger than the third term, and in that case,

$$\operatorname{tr}(C) \ge \frac{1}{c_2 b^2 n} \sum_i \min\left\{1, \frac{\lambda_i^2}{\lambda_{k+1}^2}\right\} \ge \frac{k+1}{c_2 b^2 n}.$$

On the other hand, if $r_k(\lambda) \ge bn$,

$$\begin{aligned}\operatorname{tr}(C) &\ge \frac{1}{c_2 b^2} \sum_i \min\left\{\frac{1}{n}, \frac{b^2 n \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}\right\} \\ &= \frac{1}{c_2 b^2} \min_{l \le k}\left(\frac{l}{n} + \frac{b^2 n \sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}\right),\end{aligned}$$

where the equality follows from the fact that the $\lambda_i$ are nonincreasing. □

**A Simple Choice of *l*.** Recall that $\sigma_x$ is a constant. If no $k \le n/c$ has $r_k(\Sigma) \ge bn$, then Lemmas 2 and 10 imply that the expected excess risk is $\Omega(\sigma^2)$, which proves the first paragraph of Theorem 1 for large $k^*$. If some $k \le n/c$ does have $r_k(\Sigma) \ge bn$, then the upper and lower bounds of Lemmas 6 and 10 are constant multiples of

$$\min_{l \le k}\left(\frac{l}{n} + n \frac{\sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}\right).$$

It might seem surprising that any suitable choice of $k$ suffices to give upper and lower bounds: what prevents one choice of $k$ from giving an upper bound that falls below the lower bound that arises from another choice of $k$? However, the freedom to choose $k$ is somewhat illusory: Lemma 5 shows that, for any qualifying value of $k$, the smallest eigenvalue of $A$ is within a constant factor of $\lambda_{k+1} r_k(\Sigma)$. Thus, any two choices of $k$ satisfying $k \le n/c$ and $r_k(\Sigma) \ge bn$ must have values of $\lambda_{k+1} r_k(\Sigma)$ within constant factors. The smallest such $k$ simplifies the bound on $\operatorname{tr}(C)$ as the following lemma shows. The proof is in *SI Appendix*, section F.

**Lemma 11.** *For any* $b \ge 1$ *and* $k^* := \min\{k : r_k(\Sigma) \ge bn\}$, *if* $k^* < \infty$, *we have*

$$\begin{aligned}&\min_{l \le k^*}\left(\frac{l}{bn} + \frac{bn \sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2}\right) \\ &= \frac{k^*}{bn} + \frac{bn \sum_{i>k^*} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}.\end{aligned}$$

Finally, we can finish the proof of Theorem 1. Set $b$ in Lemma 10 and Theorem 1 to the constant $b$ from Lemma 6. Take $c_1$ to be the maximum of the constants $c$ from Lemmas 6 and 10.

By Lemma 10, if $k^* \ge n/c_1$, then with high probability $\operatorname{tr}(C) \ge 1/c_2$. However, by Lemma 10.2 and by Lemma 6, if $k^* < n/c_1$, then with high probability $\operatorname{tr}(C)$ is within a constant factor of $\min_{l \le k^*}\left(\frac{l}{n} + n \frac{\sum_{i>l} \lambda_i^2}{(\lambda_{k^*+1} r_{k^*}(\Sigma))^2}\right)$, which by Lemma 11, is within a constant factor of $\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}$. Taking $c$ sufficiently large and combining these results with Lemma 2 and with the upper bound on the term $\theta^{*\top} B \theta^*$ in *SI Appendix*, section J completes the proof of the first paragraph of Theorem 1.

The proof of the second paragraph is in *SI Appendix*, section K.

## Deep Neural Networks

How relevant are Theorems 1 and 2 to the phenomenon of benign overfitting in deep neural networks? One connection appears by considering regimes where deep neural networks are well approximated by linear functions of their parameters. This so-called neural tangent kernel (NTK) viewpoint has been vigorously pursued recently in an attempt to understand the optimization properties of deep learning methods. Very wide neural networks, trained with gradient descent from a suitable random initialization, can be accurately approximated by linear functions in an appropriate Hilbert space, and in this case, gradient descent finds an interpolating solution quickly (14–19). (Note that these papers do not consider prediction accuracy, except when there is no noise; for example, ref. 14, Assumption A1 implies that the network can compute a suitable real-valued response exactly, and the data-dependent bound of ref. 19, Theorem 5.1 becomes vacuous when independent noise is added to the $y_i$.) The eigenvalues of the covariance operator in this case can have a heavy tail under reasonable assumptions on the data distribution (20, 21), and the dimension is very large but finite as required for benign overfitting. However, the assumptions of Theorem 1 do not apply in this case. In particular, the assumption that the random elements of the Hilbert space are a linearly transformed vector with independent components is not satisfied. Thus, our results are not directly applicable in this—somewhat unrealistic—setting. Note that the slow decay of the eigenvalues of the NTK is in contrast to the case of the Gaussian and other smooth kernels, where the eigenvalues decay nearly exponentially quickly (22).

The phenomenon of benign overfitting was first observed in deep neural networks. Theorems 1 and 2 are steps toward understanding this phenomenon by characterizing when it occurs in the simple setting of linear regression. Those results suggest that covariance eigenvalues that are constant or slowly decaying in a high (but finite)-dimensional space might be important in the deep network setting also. Some authors have suggested viewing neural networks as finite-dimensional approximations to infinite-dimensional objects (23–25), and there are generalization bounds—although not for the overfitting regime—that are applicable to infinite-width deep networks with parameter norm constraints (26–30). However, the intuition from the linear setting suggests that truncating to a finite-dimensional space might be important for good statistical performance in the overfitting regime. Confirming this conjecture by extending our results to the setting of prediction in deep neural networks is an important open problem.

## Conclusions and Further Work

Our results characterize when the phenomenon of benign overfitting occurs in high-dimensional linear regression with Gaussian data and more generally. We give finite sample excess risk bounds that reveal the covariance structure that ensures that the minimum norm interpolating prediction rule has near-optimal prediction accuracy. The characterization depends on two notions of the effective rank of the data covariance operator. It shows that overparameterization (that is, the existence of many low-variance and hence, unimportant directions in parameter space) is essential for benign overfitting and that data that lie in a large but finite-dimensional space exhibit the benign overfitting phenomenon with a much wider range of covariance properties than data that lie in an infinite-dimensional space.

There are several natural future directions. Our main theorem requires the conditional expectation $\mathbb{E}[y|x]$ to be a linear function of $x$, and it is important to understand whether the results are also true in the misspecified setting, where this assumption is not true. Our main result also assumes that the covariates are distributed as a linear function of a vector of independent random variables. We would like to understand the extent to which this assumption can be relaxed since it rules out some important examples, such as infinite-dimensional reproducing kernel Hilbert spaces with continuous kernels defined on finite-dimensional spaces. We would also like to understand how our results extend to other loss functions other than squared error and what we can say about overfitting estimators beyond the minimum norm interpolating estimator. The most interesting future direction is understanding how these ideas could apply to nonlinearly parameterized function classes, such as neural networks, the methodology that uncovered the phenomenon of benign overfitting.

**Data Availability.** There are no data associated with this manuscript.

**ACKNOWLEDGMENTS.** We acknowledge the support of NSF Grant IIS-1619362 and of a Google research award. G.L. was supported by the Spanish Ministry of Economy and Competitiveness, Grant PGC2018-101643-B-I00; "High-dimensional problems in structured probabilistic models - Ayudas Fundación BBVA a Equipos de Investigación Cientifica 2017"; and Google Focused Award "Algorithms and Learning for AI." Part of this work was done as part of the fall 2018 program on Foundations of Data Science at the Simons Institute for the Theory of Computing.

1. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, "Understanding deep learning requires rethinking generalization" in *5th International Conference on Learning Representations*. https://openreview.net/forum?id=Sy8gdB9xx. Accessed 30 March 2020.
2. T. Hastie, R. Tibshirani, J. H. Friedman, *Elements of Statistical Learning* (Springer, 2001).
3. M. Belkin, S. Ma, S. Mandal, "To understand deep learning we need to understand kernel learning" in *Proceedings of the 35th International Conference on Machine Learning* (Proceedings of Machine Learning Research, 2018), vol. 80, pp. 540–548.
4. M. Belkin, D. Hsu, P. Mitra, "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate" in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, S. Bengio et al., Eds. (NIPS, 2018), pp. 2306–2317.
5. M. Belkin, A. Rakhlin, A. B. Tsybakov, Does data interpolation contradict statistical optimality? arXiv:1806.09471 (25 June 2018).
6. L. Devroye, L. Györfi, A. Krzyżak, The Hilbert kernel regression estimate. *J. Multivariate Anal.* **65**, 209–227 (1998).
7. T. Liang, A. Rakhlin, Just interpolate: Kernel "ridgeless" regression can generalize. arXiv:1808.00387 (1 August 2018).
8. M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine learning and the bias-variance trade-off. arXiv:1812.11118 (28 December 2018).
9. V. Muthukumar, K. Vodrahalli, A. Sahai, Harmless interpolation of noisy data in regression. arXiv:1903.09139 (21 March 2019).
10. P. L. Bartlett, "Accurate prediction from interpolation: A new challenge for statistical learning theory (presentation at the National Academy of Sciences workshop, The Science of Deep Learning)" (video recording, 2019). https://www.youtube.com/watch?v=1y2sB38T6FU&feature=youtu.be. Accessed 14 March 2019.
11. M. Belkin, D. Hsu, J. Xu, Two models of double descent for weak features. arXiv:1903.07571 (18 March 2019).
12. T. Hastie, A. Montanari, S. Rosset, R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation. arXiv:1903.08560 (19 March 2019).
13. C. A. Desoer, B. H. Whalen, A note on pseudoinverses. *J. Soc. Ind. Appl. Math.* **11**, 442–446 (1963).
14. Y. Li, Y. Liang, Learning overparameterized neural networks via stochastic gradient descent on structured data. arXiv:1808.01204 (3 August 2018).
15. S. S. Du, B. Poczós, X. Zhai, A. Singh, Gradient descent provably optimizes over-parameterized neural networks. arXiv:1810.02054 (4 October 2018).
16. S. S. Du, J. D. Lee, H. Li, L. Wang, X. Zhai, Gradient descent finds global minima of deep neural networks. arXiv:1811.03804 (9 November 2018).
17. D. Zou, Y. Cao, D. Zhou, Q. Gu, Stochastic gradient descent optimizes over-parameterized deep relu networks. arXiv:1811.08888 (21 November 2018).
18. A Jacot, F Gabriel, C Hongler, "Neural tangent kernel: Convergence and generalization in neural networks" in *32nd Conference on Neural Information Processing Systems*, S. Bengio *et al.*, Eds. (NeurIPS, 2018), pp. 8580–8589.
19. S. Arora, S. S. Du, W. Hu, Z. Li, R. Wang, Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. arXiv:1901.08584 (24 January 2019).
20. B. Xie, Y. Liang, L. Song, Diverse neural network learns true target functions. arXiv:1611.03131 (9 November 2016).
21. Y. Cao, Z. Fang, Y. Wu, D. X. Zhou, Q. Gu, Towards understanding the spectral bias of deep learning. arXiv:1912.01198 (3 December 2019).
22. M. Belkin, "Approximation beats concentration? An approximation view on inference with smooth radial kernels" in *Conference On Learning Theory, 2018, Stockholm, Sweden, 6-9 July 2018*, S. Bubeck, V. Perchet, P. Rigollet, Eds. (PMLR, 2018), vol. 75, pp. 1348–1361.
23. W. S. Lee, P. L. Bartlett, R. C. Williamson, Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inf. Theor.* **42**, 2118–2132 (1996).
24. Y. Bengio, N. L. Roux, P. Vincent, O. Delalleau, P. Marcotte, "Convex neural networks" in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, J. C. Platt, Eds. (MIT Press, Cambridge, MA, 2006), pp. 123–130.
25. F. Bach, Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.* **18**, 1–53 (2017).
26. P. L. Bartlett, The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Trans. Inf. Theor.* **44**, 525–536 (1998).
27. P. L. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3**, 463–482 (2002).
28. B. Neyshabur, R. Tomioka, N. Srebro, "Norm-based capacity control in neural networks" in *Proceedings of the 28th Conference on Learning Theory, Proceedings of Machine Learning Research*, P. Grünwald, E. Hazan, S. Kale, Eds. (PMLR, Paris, France, 2015), vol. 40, pp. 1376–1401.
29. P. Bartlett, D. Foster, M. Telgarsky, "Spectrally-normalized margin bounds for neural networks" in *Advances in Neural Information Processing Systems 30*, I. Guyon *et al.*, Eds. (Curran Associates, Inc., 2017), pp. 6240–6249.
30. N. Golowich, A. Rakhlin, O. Shamir, "Size-independent sample complexity of neural networks" in *Proceedings of the 31st Conference on Learning Theory, Proceedings of Machine Learning Research*, S. Bubeck, V. Perchet, P. Rigollet, Eds. (PMLR, 2018), vol. 75, pp. 297–299.