Journal of Machine Learning Research 23 (2022) 1-12 Submitted 10/21; Revised 3/22; Published 4/22

Foolish Crowds Support Benign Overfitting

Niladri S. Chatterji NILADRI@CS.STANFORD.EDU
Computer Science Department, Stanford University, 353 Jane Stanford Way, Stanford, CA 94305.

Philip M. Long PLONG@GOOGLE.COM
Google, 1600 Amphitheatre Parkway, Mountain View, CA, 94043.

Editor: Ohad Shamir

Abstract

We prove a lower bound on the excess risk of sparse interpolating procedures for linear regression with Gaussian data in the overparameterized regime. We apply this result to obtain a lower bound for basis pursuit (the minimum ℓ_1-norm interpolant) that implies that its excess risk can converge at an exponentially slower rate than OLS (the minimum ℓ_2-norm interpolant), even when the ground truth is sparse. Our analysis exposes the benefit of an effect analogous to the "wisdom of the crowd", except here the harm arising from fitting the *noise* is ameliorated by spreading it among many directions—the variance reduction arises from a *foolish* crowd.

Keywords: generalization, benign overfitting, interpolation, sparsity, lower bounds, regression

1. Introduction

Recently, there has been a surge of interest in benign overfitting, where a learning algorithm generalizes well despite interpolating noisy data (see, e.g., Zhang et al., 2021; Belkin et al., 2019; Bartlett et al., 2020; Belkin, 2021). Arguably the most basic setting in which this has been analyzed theoretically is linear regression with Gaussian data, where upper and nearly matching lower bounds have been obtained for the ordinary least squares (OLS) estimator, which chooses a parameter vector $\widehat{\theta}$ to minimize $\|\widehat{\theta}\|_2$ from among interpolating models (Bartlett et al., 2020; Negrea et al., 2020; Tsigler and Bartlett, 2020). Bounds have also been obtained for *basis pursuit* (Chen et al., 2001), which minimizes $\|\widehat{\theta}\|_1$ from among interpolating models (Muthukumar et al., 2020; Ju et al., 2020; Chinot et al., 2020; Koehler et al., 2021).

The upper bounds for the OLS estimator show that it rapidly approaches the Bayes risk when the structure of the covariance matrix Σ of the inputs is favorable. Informally, the following are necessary and sufficient:

- the sum of all of the eigenvalues of Σ is not too big;
- after excluding a few of the largest eigenvalues, there are many small eigenvalues of roughly equal magnitude.

A canonical example of such a benign covariance matrix is $\Sigma_{k,\varepsilon} := \mathrm{diag}(\overbrace{1,\ldots,1}^{k},\overbrace{\varepsilon,\ldots,\varepsilon}^{p-k})$. Consider a case where $k \ll p$, the rows of X are n i.i.d. draws from $\mathcal{N}(0,\Sigma_{k,\varepsilon})$, and, for

independent noise $\xi \sim \mathcal{N}(0, I)$ and a unit-length $\theta^\star$, $y = X\theta^\star + \xi$. Then a high probability upper bound on the excess risk of the OLS is proportional to

$$\frac{k}{n} + \frac{\varepsilon p}{n} + \frac{n}{p},$$

and a nearly matching lower bound is also known (Bartlett et al., 2020; Negrea et al., 2020; Tsigler and Bartlett, 2020). As an example, if $p = n^2$, $\varepsilon = 1/n^2$ and k is a constant, this is proportional to $1/n$.

If, in addition, $\theta^\star_i = 0$ for $i > k$, upper bounds are also known for basis pursuit in this setting (Koehler et al., 2021). On the other hand, they are much worse, scaling with p like $\frac{1}{\log p}$, and requiring that p be an exponentially large function of n to converge.

In this paper, we prove a lower bound for basis pursuit in this setting that scales with p like $\frac{1}{\log^2 p}$. Our lower bound requires that $\sigma \geq c\|\theta^*\|_2$ for an arbitrarily small positive constant c, along with a few mild technical conditions, including that $p > n$, so that interpolation is possible, and that $n \gg k$ (see Theorem 2 for the details). Note that OLS converges much faster than basis pursuit in this setting despite the fact that θ^* is sparse.

The lower bound is a special case of a more general bound, Theorem 1, which can be paraphrased as follows. Under the same conditions as Theorem 2, including $\sigma \geq c\|\theta^*\|_2$, any interpolating procedure that, with high probability, outputs an s-sparse model $\widehat{\theta}$, must suffer excess loss proportional to $\frac{\sigma^2 n}{s \log^2 p}$. We then get Theorem 2 by showing that, in this setting, basis pursuit almost surely outputs an n-sparse model.

This analysis sheds light on why the overfitting of OLS is so benign. Because OLS interpolates, we can think of the parameters of its output as storing the noise—OLS benefits from spreading the noise evenly among many parameters, where each small fragment of noise has a tiny effect. This phenomenon is akin to the reduction in variance arising from prediction using a weighted average of many covariates that is commonly referred to as the "wisdom of the crowd" (Surowiecki, 2005). Here, by spreading the harm arising from fitting the *noise* among many parameters, the algorithm benefits from a *foolish* crowd.

Muthukumar et al. (2020) established the same lower bound for basis pursuit in a setting with isotropic covariates (see also, Ju et al., 2020) in the case that $\theta^* = 0$. Accommodating the possibility that $\theta^* \neq 0$ complicates the argument a bit, but our main contribution is to demonstrate slow convergence for basis pursuit in settings where OLS enjoys fast convergence. Chinot et al. (2020) proved an upper bound on the risk of basis pursuit, but, as pointed out by Koehler et al. (2021), it does not imply a bound on the excess risk. Limitations of algorithms that output sparse linear classifiers have also been studied previously (Helmbold and Long, 2012).

After a preliminary version of this work was posted on arXiv (Chatterji and Long, 2021), some related work was published (Wang et al., 2021; Li and Wei, 2021; Donhauser et al., 2022) that established upper bounds on the excess risk of the minimum ℓ_1-norm interpolator. In particular, Wang et al. (2021) showed an upper bound on the excess risk, that in the case with isotropic covariates scales as $\sigma^2/\log(p)$, almost matching our lower bound.

2. Preliminaries

For $p, n \in \mathbb{N}$, an *example* is a member of $\mathbb{R}^p \times \mathbb{R}$, and a *linear regression algorithm* takes as input n examples, and outputs $\widehat{\theta} \in \mathbb{R}^p$. For a joint probability distribution P over $\mathbb{R}^p \times \mathbb{R}$, the *excess risk* of $\widehat{\theta}$ with respect to P is

$$R(\widehat{\theta}) := \mathbb{E}_{(x,y)\sim P}[(\widehat{\theta} \cdot x - y)^2] - \inf_{\theta^\star} \mathbb{E}_{(x,y)\sim P}[(\theta^\star \cdot x - y)^2].$$

We refer to the following as the $(k, p, n, \varepsilon, \sigma)$-scenario:

- $X \in \mathbb{R}^{n\times p}$ is a matrix whose rows are i.i.d. draws from $\mathcal{N}(0, \mathrm{diag}(\overbrace{1, 1, \ldots, 1}^{k}, \overbrace{\varepsilon, \ldots, \varepsilon}^{p-k}))$, and
- for $\theta^\star \in \mathbb{R}^p$ with $(\theta^\star_{k+1}, \ldots, \theta^\star_p) = 0$, and $\xi \sim \mathcal{N}(0, \sigma^2 I)$, $y = X\theta^\star + \xi$.

For $\delta > 0$, $s, n \in \mathbb{N}$ and a joint probability distribution P over $\mathbb{R}^p \times \mathbb{R}$, we say that a regression algorithm A is an (s, δ)-sparse interpolator for P and $n \in \mathbb{N}$ if it satisfies the following: With probability $1 - \delta$ over the independent draw of the samples $(x_1, y_1), \ldots, (x_n, y_n) \sim P$, the output $\widehat{\theta}$ of A

- interpolates the data (that is, satisfies $\widehat{\theta} \cdot x_1 = y_1, \ldots, \widehat{\theta} \cdot x_n = y_n$), and
- has at most s non-zero components.

Given $X \in \mathbb{R}^{n\times p}$ and $y \in \mathbb{R}^n$, the *basis pursuit* algorithm (minimum ℓ_1-norm interpolant) A_{BP} outputs

$$\arg\min_{\theta} \|\theta\|_1, \quad \text{s.t. } X\theta = y,$$

if there is such a θ, and otherwise behaves arbitrarily (say outputting 0).

For any $j \in \mathbb{N}$, we denote the set $\{1, \ldots, j\}$ by $[j]$. Given a vector v, let $\|v\|_2$ denote its Euclidean norm and $\|v\|_1$ denote its ℓ_1-norm. Given a matrix M, let $\|M\|_{op}$ denote its operator norm. For $z \in \mathbb{R}$, we denote $\max\{z, 0\}$ by $[z]_+$.

3. Main Results

We are ready to present our main result, a high probability lower bound on the excess risk for any (s, δ)-sparse interpolator.

Theorem 1 *For any $0 < c_1 \le 1$, there are absolute positive constants c_2, c_3 such that the following holds. For any $0 \le \delta \le c_2$, for any $(k, p, n, \varepsilon, \sigma)$ such that $\sigma \ge c_1\|\theta^\star\|_2$, $p \ge n + k$, $n \ge \log^2(1/\delta) + k^{1+c_1}$, for any $n \le s \le p - k$, and any regression algorithm A that is an (s, δ)-sparse interpolator for the $(k, p, n, \varepsilon, \sigma)$-scenario P, with probability $1 - 4\delta$ over n random draws from P, the output $\widehat{\theta}$ of A satisfies*

$$R(\widehat{\theta}) \ge \frac{c_3\sigma^2 n}{s \log^2(3p/s)}.$$

This theorem shows that a sparse interpolating predictor suffers large excess risk. Intuitively, the proof follows since an s-sparse interpolating predictor needs to hide the "energy" of the noise, which roughly scales like $\sigma^2 n$, in just s coordinates. If it attempts to hide it in the first k coordinates, then it suffers from large bias. If it hides it in the tail, then it suffers large variance.

Next, we state our result for basis pursuit, the minimum ℓ_1-norm interpolator.

Theorem 2 *For any $0 < c_1 \leq 1$, there are absolute positive constants c_2, c_3 such that the following holds. For any $0 \leq \delta \leq c_2$, for any $(k, p, n, \varepsilon, \sigma)$ such that $\sigma \geq c_1 \|\theta^\star\|_2$, $p \geq n + k$, $n \geq \log^2(1/\delta) + k^{1+c_1}$, with probability $1 - 4\delta$ over n random draws from P, the output $\widehat{\theta}$ of A_{BP} satisfies*

$$R(\widehat{\theta}) \geq \frac{c_3 \sigma^2}{\log^2(3p/n)}.$$

This theorem is proved by showing that the output of basis pursuit is always n-sparse and then by simply invoking the previous general result. Theorem 2 implies that the excess risk of A_{BP} can be much worse than OLS. For example, if $k = 5$, $p = n^2$, $\varepsilon = 1/n^2$, $\sigma^2 = \|\theta^\star\|_2^2 = 1$, then Theorem 2 implies an $\Omega\left(\frac{1}{\log^2 n}\right)$ lower bound for A_{BP} where a $O\left(\frac{1}{n}\right)$ upper bound holds for OLS (Bartlett et al., 2020; Negrea et al., 2020; Tsigler and Bartlett, 2020). If instead, $k = 5$, $p = n^2$, $\varepsilon = 1/n^2$, $\sigma^2 = \|\theta^\star\|_2^2 = \log^2 n$, then the excess risk of OLS goes to zero at a $O\left(\frac{\log^2 n}{n}\right)$ rate, but Theorem 2 implies that the excess risk of A_{BP} is bounded below by a constant.

When $\varepsilon = 1$, that is, when the covariates are isotropic, our lower bound coincides with the lower bound derived previously by Muthukumar et al. (2020).

4. Proof of Theorem 1

This section is devoting to proving Theorem 1, so the assumptions of Theorem 1 are in scope throughout this section. Our proof proceeds through a series of lemmas.

Definition 3 *For any $v \in \mathbb{R}^p$ and any $S \subseteq [p]$, let v_S be the vector obtained by selecting the components of S from v in order. For $X \in \mathbb{R}^{n \times p}$, define X_S similarly, except selecting columns from X. Let $H := \{1, \ldots, k\}$ and $T := \{k+1, \ldots, p\}$, so that $v_H = (v_1, \ldots, v_k)$ and $v_T = (v_{k+1}, \ldots, v_p)$.*

The first step is to break up the excess risk into contributions from the "head" H and the "tail" T.

Lemma 4 *The excess risk of any parameter vector $\widehat{\theta}$ satisfies*

$$R(\widehat{\theta}) = \|\theta^*_H - \widehat{\theta}_H\|_2^2 + \varepsilon \|\widehat{\theta}_T\|_2^2.$$

Proof By the projection lemma,

$$\begin{aligned} R(\widehat{\theta}) &= \mathbb{E}_{(x,y)\sim P}[(\widehat{\theta} \cdot x - y)^2] - \mathbb{E}_{(x,y)\sim P}[(\theta^\star \cdot x - y)^2] \\ &= \mathbb{E}_{(x,y)\sim P}[(\widehat{\theta} \cdot x - \theta^\star \cdot x)^2] + \sigma^2 - \sigma^2 \\ &= \|\theta^*_H - \widehat{\theta}_H\|_2^2 + \varepsilon \|\widehat{\theta}_T\|_2^2, \end{aligned}$$

since $\theta^*_T = 0$ and the distribution of x has covariance $\mathrm{diag}(\overbrace{1, 1, \ldots, 1}^{k}, \overbrace{\varepsilon, \ldots, \varepsilon}^{p-k})$. ■

Lemma 4 leads to the subproblem of establishing a lower bound on $\|\widehat{\theta}_T\|_2^2$. The following lemma is an easy step in this direction.

Lemma 5 *Given any estimator $\widehat{\theta}$ such that $X\widehat{\theta} = y$ we have*

$$\|X_T\widehat{\theta}_T\|_2 = \|y - X_H\widehat{\theta}_H\|_2.$$

Proof The lemma follows from the fact that $y = X\widehat{\theta} = X_H\widehat{\theta}_H + X_T\widehat{\theta}_T$. ■

Lemma 5 provides a means to establish a lower bound on $\|X_T\widehat{\theta}_T\|_2$. This in turn can lead to a lower bound on $\|\widehat{\theta}_T\|_2$ if we can show that the linear operator associated with X_T does not "blow up" $\widehat{\theta}_T$. It turns out, when $\widehat{\theta}$ (and thus $\widehat{\theta}_T$) is sparse, a random X_T is especially unlikely to "blow up" $\widehat{\theta}_T$, as reflected in the following lemma. It is an immediate consequence of (Adamczak et al., 2012, Theorem 4.2).

Lemma 6 *There exists a constant c such that for any $t \geq 1$, we have*

$$\mathbb{P}\left[\max_{S \subseteq T : |S| \leq s} \|X_S\|_{op} \geq c\sqrt{\varepsilon}\left(\sqrt{s}\log\left(\frac{3(p-k)}{s}\right) + \sqrt{n} + t\right)\right] \leq e^{-t}.$$

Lemma 6 implies a lower bound on $\|\theta_T\|_2$ for any s-sparse θ.

Lemma 7 *There exists a constant c such that, with probability at least $1-\delta$, any s-sparse θ has*

$$\|\theta_T\|_2 \geq \frac{\|X_T\theta_T\|_2}{c\sqrt{\varepsilon}\left(\sqrt{s}\log\left(\frac{3(p-k)}{s}\right) + \sqrt{n} + \log(1/\delta)\right)}. \tag{1}$$

Proof For any s-sparse θ, if $S = T \cap \{i : \theta_i \neq 0\}$, we have $|S| \leq s$, so for any X, we have $\|X_T\theta_T\|_2 = \|X_S\theta_S\|_2$. Applying Lemma 6 with $t = \log(1/\delta)$, with probability at least $1-\delta$, (1) holds for all such θ. ■

Since $\widehat{\theta}$ is likely to be s-sparse, Lemma 7 implies a high-probability lower bound on $\|\widehat{\theta}_T\|_2$, the contribution of the tail to the excess risk. This bound is in terms of $\|X_T\widehat{\theta}_T\|_2 = \|y - X_H\widehat{\theta}_H\|_2$. We will bound this by proving a high-probability lower bound on $\|y\|_2$, and a high-probability upper bound on $\|X_H\widehat{\theta}_H\|_2$. We start with a lower bound on $\|y\|_2$.

Lemma 8 *With probability $1-\delta$,*

$$\|y\|_2^2 \geq (\sigma^2 + \|\theta^\star\|_2^2)n\left(1 - 2\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

Proof We have $y = X\theta^\star + \xi$. That is, for each sample $i \in [n]$, $y_i = x_i \cdot \theta^\star + \xi_i$.

Observe that $x_i \cdot \theta^\star \sim \mathcal{N}(0, \|\theta^\star\|_2^2)$, $\xi_i \sim \mathcal{N}(0, \sigma^2)$, and x_i and ξ_i are independent. Therefore, we have that $y_i \sim \mathcal{N}(0, \sigma^2 + \|\theta^\star\|_2^2)$. Thus

$$\|y\|_2^2 = \sum_{i=1}^{n} |y_i|^2 = \left(\sigma^2 + \|\theta^\star\|_2^2\right) q, \tag{2}$$

where q is a random variable with a $\chi^2(n)$ distribution. Applying Lemma 1 from (Laurent and Massart, 2000), we have

$$\mathbb{P}\left(q \geq n - 2\sqrt{tn}\right) \geq 1 - \exp(-t).$$

If we set $t = \log(1/\delta)$ then with probability at least $1 - \delta$,

$$q \geq n\left(1 - 2\sqrt{\frac{\log(1/\delta)}{n}}\right).$$

This combined with (2) completes the proof. ■

Recall that we also want an upper bound on $\|X_H\widehat{\theta}_H\|_2$; we will use a bound that is an immediate consequence of (Vershynin, 2010, Corollary 5.35).

Lemma 9 *With probability* $1 - \delta$*,*

$$\|X_H\|_{op} \leq \sqrt{n} + \sqrt{k} + \sqrt{2\log(2/\delta)}.$$

Armed with these lemmas, we can now prove Theorem 1.

Proof of Theorem 1 With foresight, set $\zeta = \sqrt{\frac{c_4\sigma^2 n}{s\log^2(3p/s)}}$ for a constant c_4 that will be determined by the analysis.

Case 1 ($\|\widehat{\theta}_H\|_2 \geq \|\theta^\star\|_2 + \zeta$). Recall that $\|\theta^\star\|_2 = \|\theta^\star_H\|_2$, since it is zero for all entries after the kth coordinate. By Lemma 4 we have

$$R(\widehat{\theta}) \geq \|\widehat{\theta}_H - \theta^\star_H\|_2^2 \geq \left(\|\widehat{\theta}_H\|_2 - \|\theta^\star\|_2\right)_+^2 \geq \zeta^2 = \frac{c_4\sigma^2 n}{s\log^2(3p/s)}.$$

Case 2 ($\|\widehat{\theta}_H\|_2 \leq \|\theta^\star\|_2 + \zeta$). By Lemma 4, we have

$$R(\widehat{\theta}) \geq \varepsilon\|\widehat{\theta}_T\|_2^2.$$

The estimator $\widehat{\theta}$ is s-sparse with probability at least $1 - \delta$. Hence, combining Lemmas 5 and 7, and taking a union bound we get that, for an absolute positive constant c, with probability $1 - 2\delta$,

$$R(\widehat{\theta}) \geq c\frac{\|y - X_H\widehat{\theta}_H\|_2^2}{s\log^2\left(\frac{3(p-k)}{s}\right) + n + \log^2(1/\delta)} \geq c\frac{\left(\|y\|_2 - \|X_H\widehat{\theta}_H\|_2\right)_+^2}{s\log^2\left(\frac{3(p-k)}{s}\right) + n + \log^2(1/\delta)}.$$

Applying Lemma 8, we find that with probability at least $1-3\delta$,

$$R(\widehat{\theta}) \geq c\frac{\left[\sqrt{(\sigma^2+\|\theta^\star\|_2^2)n\left(1-2\sqrt{\frac{\log(1/\delta)}{n}}\right)}-\|X_H\widehat{\theta}_H\|_2\right]_+^2}{s\log^2\left(\frac{3(p-k)}{s}\right)+n+\log^2(1/\delta)}.$$

Next, by applying Lemma 9, with probability at least $1-4\delta$,

$$\begin{aligned}
&R(\widehat{\theta})\\
&\geq c\frac{\left[\sqrt{(\sigma^2+\|\theta^\star\|_2^2)n\left(1-2\sqrt{\frac{\log(1/\delta)}{n}}\right)}-(\sqrt{n}+\sqrt{k}+\sqrt{2\log(2/\delta)})\|\widehat{\theta}_H\|_2\right]_+^2}{s\log^2\left(\frac{3(p-k)}{s}\right)+n+\log^2(1/\delta)}\\
&\geq c\frac{\left[\sqrt{(\sigma^2+\|\theta^\star\|_2^2)n\left(1-2\sqrt{\frac{\log(1/\delta)}{n}}\right)}-(\sqrt{n}+\sqrt{k}+\sqrt{2\log(2/\delta)})\left(\|\theta^\star\|_2+\zeta\right)\right]_+^2}{s\log^2\left(\frac{3(p-k)}{s}\right)+n+\log^2(1/\delta)}\\
&= c\frac{\left[\sqrt{(\sigma^2+\|\theta^\star\|_2^2)n\left(1-2\sqrt{\frac{\log(1/\delta)}{n}}\right)}-(\sqrt{n}+\sqrt{k}+\sqrt{2\log(2/\delta)})\left(\|\theta^\star\|_2+\sqrt{\frac{c_4\sigma^2 n}{s\log^2(3p/s)}}\right)\right]_+^2}{s\log^2\left(\frac{3(p-k)}{s}\right)+n+\log^2(1/\delta)}.
\end{aligned}$$

By choosing $c_2>0$ to be small enough, we can choose n to be as large as desired. Recall that, as $n\to\infty$, both $k=o(n)$ and $\log(2/\delta)=o(n)$. Thus with probability at least $1-4\delta$, we have that

$$R(\widehat{\theta}) \geq (c/2)\frac{\left[\sqrt{\sigma^2+\|\theta^\star\|_2^2}-\left(1+\frac{c_1^2}{8}\right)\left(1+\sqrt{\frac{c_4\sigma^2 n}{\|\theta^\star\|_2^2 s\log^2(3p/s)}}\right)\|\theta^\star\|_2\right]_+^2 n}{s\log^2\left(\frac{3(p-k)}{s}\right)+n+\log^2(1/\delta)}.$$

Since $s\geq n$, we have

$$\begin{aligned}
R(\widehat{\theta}) &\geq c'\frac{\left[\sqrt{\sigma^2+\|\theta^\star\|_2^2}-\left(1+\frac{c_1^2}{8}\right)\left(1+\sqrt{\frac{c_4\sigma^2}{\|\theta^\star\|_2^2\log^2(3p/s)}}\right)\|\theta^\star\|_2\right]_+^2 n}{s\log^2(3p/s)}\\
&= c'\frac{\left[\sqrt{1+\frac{\|\theta^\star\|_2^2}{\sigma^2}}-\left(1+\frac{c_1^2}{8}\right)\left(1+\sqrt{\frac{c_4\sigma^2}{\|\theta^\star\|_2^2\log^2(3p/s)}}\right)\frac{\|\theta^\star\|_2}{\sigma}\right]_+^2\sigma^2 n}{s\log^2(3p/s)}.
\end{aligned}$$

Defining $r:=\frac{\|\theta^\star\|_2}{\sigma}$ and simplifying, we have

$$R(\widehat{\theta}) \geq c'\frac{\left[\sqrt{1+r^2}-\left(1+\frac{c_1^2}{8}\right)r-\left(1+\frac{c_1^2}{8}\right)\sqrt{\frac{c_4}{\log^2(3p/s)}}\right]_+^2\sigma^2 n}{s\log^2(3p/s)}.$$

Since $\sqrt{1+r^2}-\left(1+\frac{c_1^2}{8}\right)r$ is a decreasing function of r, and, by assumption, $r=\frac{\|\theta^\star\|}{\sigma}\leq 1/c_1$, we have

$$R(\widehat{\theta})\geq c'\frac{\left[\sqrt{1+\frac{1}{c_1^2}}-\frac{1}{c_1}-\frac{c_1}{8}-\left(1+\frac{c_1^2}{8}\right)\sqrt{\frac{c_4}{\log^2(3p/s)}}\right]_+^2\sigma^2 n}{s\log^2(3p/s)}.$$

Since $s\leq p$, we have

$$R(\widehat{\theta})\geq c'\frac{\left[\sqrt{1+\frac{1}{c_1^2}}-\frac{1}{c_1}-\frac{c_1}{8}-\left(1+\frac{c_1^2}{8}\right)\sqrt{c_4}\right]_+^2\sigma^2 n}{s\log^2(3p/s)}.$$

Recall that $0<c_1\leq 1$, and choose

$$c_3=\min\left\{c'\left(\sqrt{1+\frac{1}{c_1^2}}-\frac{1}{c_1}-\frac{c_1}{8}-\left(1+\frac{c_1^2}{8}\right)\sqrt{c_4}\right)_+^2,c_4\right\}.$$

Thus, if c_4 is chosen to be a sufficiently small positive constant then

$$c_3\geq\min\left\{c'\left(\sqrt{1+\frac{1}{c_1^2}}-\frac{1}{c_1}-\frac{c_1}{4}\right)_+^2,c_4\right\}>0$$

completing the proof. ■

5. Proof of Theorem 2

This section is devoted to proving Theorem 2, so the assumptions of Theorem 2 are in scope throughout this section. As in Section 4, our proof proceeds through a series of lemmas.

The first lemma is an immediate consequence of (Schneider and Tardivel, 2020, Proposition 1).

Lemma 10 *Almost surely, there is a unique minimizer of $\|\theta\|_1$ subject to $X\theta=y$.*

The following lemma appears to be known (Chen et al., 2001); we included a proof in Appendix A because we did not find one that applies in our setting.

Lemma 11 *Almost surely, the output $\widehat{\theta}$ of A_{BP} is n-sparse.*

Proof of Theorem 2 By Lemma 11, for any $\delta>0$, A_{BP} is a (n,δ)-sparse interpolator for the $(k,p,n,\varepsilon,\sigma)$-scenario P. Applying Theorem 1 with $s=n$ completes the proof. ■

6. Discussion

We have demonstrated that for interpolating linear regression with Gaussian data, outputting a sparse parameter vector can be harmful, even when learning a sparse target.

Our proofs only use a few of the properties of Gaussian distributions, so in the case that the covariance is $\Sigma_{k,\varepsilon}$, our results should generalize to sub-Gaussian and log-concave distributions. We chose to analyze $\Sigma_{k,\varepsilon}$ because it is arguably the canonical case where OLS enjoys benign overfitting, and it leads to clean and interpretable bounds. Handling a wider variety of covariance matrices is another very natural future direction. A starting point would be to generalize (Adamczak et al., 2012, Theorem 4.2).

Recent research has shown that linear models parameterized by simple two-layer linear neural networks with diagonal weight matrices leads to implicit regularization that interpolates between the ℓ_1-norm used by basis pursuit and the ℓ_2-norm used by OLS (Woodworth et al., 2020; Azulay et al., 2021). The connection of this work to neural networks, together with the stark difference between ℓ_1 and ℓ_2 regularization in the context of benign overfitting highlighted in this paper, motivates the study of benign overfitting with these models. (We thank Olivier Bousquet for suggesting this last problem.)

Appendix A. Proof of Lemma 11

By Lemma 10, we may assume without loss of generality that there is a unique minimizer of $\|\theta\|_1$ subject to $X\theta = y$. Assume for the sake of contradiction that $\widehat{\theta}$ has $\|\widehat{\theta}\|_0 = s > n$.

Let

$$I := \{i \in [p] : \widehat{\theta}_i \neq 0\}.$$

Since $|I| > n$, the columns in $\{X_i : i \in I\}$ are linearly dependent. That is, there exists a set of weights $\{\lambda_i : i \in I\}$, at least one of which is nonzero, such that

$$\sum_{i \in I} \lambda_i X_i = 0. \tag{3}$$

Let the vector $\lambda \in \mathbb{R}^p$ be obtained by filling in $\lambda_i = 0$ for $i \notin I$.

From here, we will divide our analysis into cases.

Case 1 ($\sum_{i \in I} \lambda_i \text{sign}(\widehat{\theta}_i) > 0$). We will prove by contradiction that this case cannot happen. For an $\eta > 0$ to be set later, consider

$$v = \widehat{\theta} - \eta \sum_{i \in I} \lambda_i e_i = \widehat{\theta} - \eta\lambda.$$

First, note that

$$Xv = X\widehat{\theta} - \eta \sum_{i \in I} \lambda_i X_i = y - 0 = y.$$

We will now prove the claim that, for a small enough η, $\|v\|_1 < \|\widehat{\theta}\|_1$. This will lead to the desired contradiction. To establish this claim, it suffices to prove that $\frac{-\lambda}{\|\lambda\|_2}$ is a descent direction for $\|\cdot\|_1$ at $\widehat{\theta}$. Toward this end, consider an arbitrary member z of the subgradient of $\|\cdot\|_1$ at $\widehat{\theta}$. Recalling that $\lambda_i = 0$ when $\widehat{\theta}_i = 0$, we have

$$\lambda \cdot z = \sum_{i \in I} \lambda_i \text{sign}(\widehat{\theta}_i) > 0,$$

by the assumption of this case. This implies that $\frac{-\lambda}{\|\lambda\|_2}$ is indeed a descent direction, so that, for a small enough η, $\|v\|_1 < \|\widehat{\theta}\|_1$. Recalling that $Xv = y$ then yields a contradiction.

Case 2 ($\sum_{i\in I} \lambda_i \text{sign}(\widehat{\theta}_i) < 0$). This case leads to a contradiction symmetrically to the proof in Case 1, using

$$v = \widehat{\theta} + \eta\lambda.$$

Case 3 ($\sum_{i\in I} \lambda_i \text{sign}(\widehat{\theta}_i) = 0$). Choose i_0 arbitrarily from among those $i \in I$ such that $\lambda_i \neq 0$ with the minimum values of $|\lambda_i|$, that is, $i_0 \in \arg\min_{i\in[p]} \{|\lambda_i| : \lambda_i > 0\}$. As in the first case, suppose that $\lambda_{i_0} > 0$ (the other case can be handled symmetrically). Set $\eta = \theta_{i_0}/\lambda_{i_0}$, and once again consider the vector

$$v = \widehat{\theta} - \eta\lambda.$$

As before, for all $\eta' \in [0, \eta]$, $X(\widehat{\theta} - \eta'\lambda) = y$. Furthermore, since

$$i_0 \in \arg\min_{i\in[p]} \{|\lambda_i| : \lambda_i > 0\},$$

for each such η', for all i,

$$\text{sign}((\widehat{\theta} - \eta'\lambda)_i) = \text{sign}(\widehat{\theta}_i).$$

This means that along the path from $\widehat{\theta}$ to v, any subgradient z of the ℓ_1 norm satisfies

$$z_i = \text{sign}(\widehat{\theta}_i), \quad \text{for all } i \in I.$$

But this means, throughout this path, λ is orthogonal to any subgradient, which in turn means that the ℓ_1-norm is unchanged. When $\eta' = \eta$, we have an interpolator with the same ℓ_1-norm as $\widehat{\theta}$ but one fewer nonzero component, a contradiction.

References

Radosław Adamczak, Rafał Latała, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Chevet type inequality and norms of submatrices. *Studia Mathematica*, 210 (1):35–56, 2012.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pages 468–477, 2021.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *arXiv preprint arXiv:2105.14368*, 2021.

Mikhail Belkin, Daniel J Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Niladri S Chatterji and Philip M Long. Foolish crowds support benign overfitting. *arXiv preprint arXiv:2110.02914*, 2021.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. On the robustness of minimum-norm interpolators. *arXiv preprint arXiv:2012.00807*, 2020.

Konstantin Donhauser, Nicolo Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effects of inductive bias. *arXiv preprint arXiv:2203.03597*, 2022.

David P Helmbold and Philip M Long. On the necessity of irrelevant variables. *Journal of Machine Learning Research*, 13(1):2145–2170, 2012.

Peizhong Ju, Xiaojun Lin, and Jia Liu. Overfitting can be harmless for basis pursuit, but only to a degree. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. In *Advances in Neural Information Processing Systems*, 2021.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, pages 1302–1338, 2000.

Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*, pages 7263–7272, 2020.

Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness, sparsity and clustering in penalized estimation. *arXiv preprint arXiv:2004.09106*, 2020.

James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum l1-norm interpolation of noisy data. *arXiv preprint arXiv:2111.05987*, 2021.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.