Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Failures of Model-dependent Generalization Bounds for Least-norm Interpolation

Peter L. Bartlett PETER@BERKELEY.EDU
University of California, Berkeley & Google, 367 Evans Hall #3860 Berkeley, CA 94720-3860.

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

Editor: Pierre Alquier

Abstract

We consider bounds on the generalization performance of the least-norm linear regressor, in the over-parameterized regime where it can interpolate the data. We describe a sense in which any generalization bound of a type that is commonly proved in statistical learning theory must sometimes be very loose when applied to analyze the least-norm interpolant. In particular, for a variety of natural joint distributions on training examples, any valid generalization bound that depends only on the output of the learning algorithm, the number of training examples, and the confidence parameter, and that satisfies a mild condition (substantially weaker than monotonicity in sample size), must sometimes be very loose—it can be bounded below by a constant when the true excess risk goes to zero.

Keywords: generalization bounds, benign overfitting, linear regression, statistical learning theory, lower bounds

1. Introduction

Deep learning methodology has revealed some striking deficiencies of classical statistical learning theory: large neural networks, trained to zero empirical risk on noisy training data, have good predictive accuracy on independent test data. These methods are overfitting (that is, fitting to the training data better than the noise should allow), but the overfitting is benign (that is, prediction performance is good). It is an important open problem to understand why this is possible.

The presence of noise is key to why the success of interpolating algorithms is mysterious. Generalization of algorithms that produce a perfect fit in the absence of noise has been studied for decades (see Haussler, 1992, and its references) A number of recent papers have provided generalization bounds for interpolating algorithms in the absence of noise, either for deep networks or in abstract frameworks motivated by deep networks (Li and Liang, 2018; Arora et al., 2019; Cao and Gu, 2019; Feldman, 2020). The generalization bounds in these papers either do not hold or become vacuous in the presence of noise: Li and Liang (2018) rule out noisy data in Assumption A1; the data-dependent bound of Arora et al. (2019, Theorem 5.1) becomes vacuous when independent noise is added to the y_i; adding a constant level of independent noise to the y_i in the setting of Cao and Gu (2019, Theorem 3.3) gives an upper bound on excess risk that is at least a constant; and the analysis of Feldman (2020) concerns the noise-free case.

There has also been progress on bounding the gap between average loss on the training set and expected loss on independent test data, based on uniform convergence arguments that bound the complexity of classes of real-valued functions computed by deep networks. For instance, the results of Bartlett (1998) for sigmoid nonlinearities rely on ℓ_1-norm bounds on parameters throughout the network, and those of Bartlett et al. (2017) for ReLUs rely on spectral norm bounds of the weight matrices throughout the network (see also Bartlett and Mendelson, 2002; Neyshabur et al., 2015; Bartlett et al., 2017; Golowich et al., 2018; Long and Sedghi, 2019). These bounds involve distribution-dependent function classes, since they depend on some measure of the complexity of the output model that may be expected to be small for natural training data. For instance, if some training method gives weight matrices that all have small spectral norm, the bound of Bartlett et al. (2017) will imply that the gap between empirical risk and predictive accuracy will be small. But while it is possible for these bounds to be small for networks that trade off fit to the data with complexity in some way, it is not clear that a network that interpolates noisy data could ever have small values of these complexity measures. This raises the question: are there any good data-dependent bounds for interpolating networks?

Zhang et al. (2017) claimed, based on empirical evidence, that conventional learning theoretic tools are useless for deep networks, but they considered the case of a fixed class of functions defined, for example, as the set of functions that can be computed by a neural network, or that can be reached by stochastic gradient descent with some training data, no matter how unlikely. These observations illustrate the need to consider distribution-dependent notions of complexity in understanding the generalization performance of deep networks. The study of such distribution-dependent complexity notions has a long history in nonparametric statistics, where it is central to the problem of model selection (see, for example, Bartlett et al., 2002, and its references); uniform convergence analysis over a level in a complexity hierarchy is part of a standard outline for analyzing model selection methods.

Nagarajan and Kolter (2019) provided an example of a scenario where, with high probability, an algorithm generalizes well, but two-sided uniform convergence fails for any hypothesis space that is likely to contain the algorithm's output. Their analysis takes an important step in allowing distribution-dependent notions of complexity, but only rules out the application of a specific set of tools: uniform convergence over a model class of the absolute differences between expectations and sample averages. Indeed, in their proof, the failure is an under-estimation of the accuracy of a model—a model has good predictive accuracy, but performs poorly on a sample (one obtained as a transformed but equally likely version of the sample that was used to train the model). However, in applying uniform convergence tools to show good performance of an algorithm, uniform bounds are only needed to show that bad models are unlikely to perform well on the training data. So if one wishes to provide bounds that guarantee that an algorithm has *poor* predictive accuracy, Nagarajan and Kolter (2019) provided an example where uniform convergence tools will not suffice. In contrast, we are concerned with understanding what tools can provide guarantees of *good* predictive accuracy of interpolating algorithms.

In this paper, motivated by the phenomenon of benign overfitting in deep networks, we consider a simpler setting where the phenomenon occurs, that of linear regression. Negrea et al. (2020, Lemma 5.2) adapt the construction of Nagarajan and Kolter (2019) to show

a similar failure of uniform convergence in this context, and similarly cannot shed light on tools that can or cannot guarantee good predictive accuracy. We study the minimum norm linear interpolant. Earlier work (Bartlett et al., 2020) provides tight upper and lower bounds on the excess risk of this interpolating prediction rule under suitable conditions on the probability distribution generating the data, showing that benign overfitting depends on the pattern of eigenvalues of the population covariance, and there is already a rich literature on related questions (Liang and Rakhlin, 2020; Belkin et al., 2019b,a; Hastie et al., 2019a,b; Negrea et al., 2020; Derezinski et al., 2020; Li et al., 2020; Tsigler and Bartlett, 2020). These risk bounds involve fine-grained properties of the distribution. Is this knowledge necessary? Is it instead possible to obtain data-dependent bounds for interpolating prediction rules? Already the proof of Bartlett et al. (2020) provides some clues that this might be difficult: when benign overfitting occurs, the eigenvalues of the empirical covariance are a very poor estimate of the true covariance eigenvalues—all but a small fraction (the largest eigenvalues) are within a constant factor of each other.

In this paper, we show that in these settings there cannot be good risk bounds based on data-dependent function classes in a strong sense: For linear regression with the minimum norm prediction rule, any "bounded-antimonotonic" model-dependent error bound that is valid for a sufficiently broad set of probability distributions must be loose—too large by an additive constant—for some (rather innocuous) probability distribution. Here the "model" in "model-dependent" refers to the output of the learning algorithm, which is an estimate of the regression function. The bounded-antimonotonic condition formalizes the mild requirement that the bound does not degrade very rapidly with additional data. Aside from this constraint, our result applies for any bound that is determined as a function of the output of the learning algorithm, the number of training examples, and the confidence parameter. This function could depend on the level in a hierarchy of models where the output of the algorithm lies. Our result applies whether the bound is obtained by uniform convergence over a level in the hierarchy, or in some other way.

The intuition behind our result is that benign overfitting can only occur when the test distribution has a vanishing overlap with the training data. Indeed, interpolating the data in the training sample guarantees that the conditional expectation of the prediction rule's loss on the training points that occur once must be at least twice the noise level. Using a Poissonization method, we show that a situation where the training sample forms a significant fraction of the support of the distribution is essentially indistinguishable from a benign overfitting situation where the training sample has measure zero. A data-dependent bound that is valid in both cases must be loose in the second case.

2. Preliminaries and Main Results

We consider prediction problems with patterns $x \in \ell_2$ and labels $y \in \mathbb{R}$, where ℓ_2 is the space of square summable sequences of real numbers. In fact, all probability distributions that we consider have support restricted to a finite dimensional subspace of ℓ_2, which we identify with $\mathbb{R}^d$ for an appropriate d. For a joint distribution[1] P over $\mathbb{R}^d \times \mathbb{R}$ and a (measurable)

1. Throughout the paper, whenever we refer to a probability distribution over $\mathbb{R}$, it is with respect to the Borel σ-field.

hypothesis $h : \mathbb{R}^d \to \mathbb{R}$, define the *risk* of h to be

$$R_P(h) = \mathbf{E}_{(x,y)\sim P}[(y - h(x))^2].$$

Let R_P^* be the minimum of R_P over measurable functions.

For any positive integer k, a distribution D over $\mathbb{R}^k$ is sub-Gaussian with parameter σ if, for any $u \in \mathbb{R}^k$, $\mathbf{E}_{x\sim D}[\exp(u \cdot (x - \mathbf{E}x))] \leq \exp\left(\frac{\|u\|^2\sigma^2}{2}\right)$. A joint distribution P over $\mathbb{R}^d \times \mathbb{R}$ has *unit scale* if $(X_1, ..., X_d, Y) \sim P$ is sub-Gaussian with parameter 1. It is *innocuous* if

- it is unit scale,
- the marginal on $(X_1, ..., X_d)$ is Gaussian, and
- the conditional of Y given $(X_1, ..., X_d)$ is continuous.

A *sample* is a finite multiset of elements of $\mathbb{R}^d \times \mathbb{R}$. A *least-norm interpolation algorithm* takes as input a sample $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, and outputs h_θ which maps x to $\theta \cdot x$ for $\theta \in \mathbb{R}^d$ that minimizes $\|\theta\|$ subject to

$$\sum_i (\theta \cdot \mathbf{x}_i - y_i)^2 = \min_{(\hat{y}_1,...,\hat{y}_n)\in\mathbb{R}^n} \sum_i (\hat{y}_i - y_i)^2.$$

We will refer both to the parameter vector θ output by the least-norm interpolation algorithm and the function $x \to \theta \cdot x$ parameterized by θ as the *least-norm interpolant*.

A function $\epsilon(h, n, \delta)$ mapping a hypothesis h, a sample size n and a confidence δ to a positive real number is a *uniform model-dependent bound for unit-scale distributions* if, for all unit-scale joint distributions P, all sample sizes n, and any least-norm interpolation algorithm A, with probability at least $1 - \delta$ over the random choice of $S \sim P^n$,

$$R_P(A(S)) - R_P^* \leq \epsilon(A(S), n, \delta).$$

Here h may be any measurable function from $\mathbb{R}^d$ to $\mathbb{R}$ for any d, and n and δ may be any positive integer and positive real number. The bound ϵ is *c-bounded antimonotonic* for $c \geq 1$ if for all h, δ, n_1 and n_2, if $n_2/2 \leq n_1 \leq n_2$ then $\epsilon(h, n_2, \delta) \leq c\epsilon(h, n_1, \delta)$. This formalizes the requirement that the bound cannot get too much worse too quickly with more data; doubling the sample size can degrade the bound by at most a factor of c. If $\epsilon(h, \cdot, \delta)$ is monotone-decreasing for all h and δ, then it is 1-bounded antimonotonic.

A set $\mathcal{B} \subseteq \mathbb{N}$ is *β-dense* if $\liminf_{N\to\infty} \frac{|\mathcal{B}\cap\{1,...,N\}|}{N} \geq \beta$. Say that $\mathcal{B} \subseteq \mathbb{N}$ is *strongly β-dense beyond n_0* if, for all $s \in \mathbb{N}$ such that $s^2 \geq n_0$,

$$\frac{|\mathcal{B} \cap \{s^2, ..., (s+1)^2 - 1\}|}{2s+1} \geq \beta.$$

The β-dense notion is standard; it roughly corresponds to the informal idea that at least a fraction β of the natural numbers are in $\mathcal{B}$. Notice that if a set is strongly β-dense beyond n_0, then it is β-dense, but also is, in a sense, "locally" β-dense as well.

The following is our main result.

Theorem 1 *If ϵ is a bounded-antimonotonic, uniform model-dependent bound for unit-scale distributions, then there are constants $c_0, c_1, c_2, c_3, c_4 > 0$ and innocuous distributions P_1 over $\mathbb{R}^{d_1} \times \mathbb{R}$, P_2 over $\mathbb{R}^{d_2} \times \mathbb{R}$, ... such that for, for any least-norm interpolation algorithm A, for all $0 < \delta < c_1$, for all large enough n,*

$$\mathbf{Pr}_{S \sim P_n^n}\left[R_{P_n}(A(S)) - R^*_{P_n} \leq c_0/\sqrt{n}\right] \geq 1 - \delta$$

but nonetheless, the set of n such that

$$\mathbf{Pr}_{S \sim P_n^n}\left[\epsilon(A(S), n, \delta) > c_2\right] \geq \frac{1}{2}$$

is strongly $(1 - \frac{c_3}{\log(1/\delta)})$-dense beyond $c_4 \log(1/\delta)$.

The distributions P_n in the theorem are slight variations on a joint gaussian distribution: the marginal distribution of x is gaussian with a covariance matrix chosen to satisfy the conditions in (Bartlett et al., 2020) that ensure the benign overfitting property, and the conditional distribution of y given x is a linear function of x plus noise, where the noise is a mixture of mean-zero gaussians with different variances.

3. Proof of Theorem 1

Our proof uses the following lemma (Birch, 1963; Feller, 1968; Batu et al., 2000, 2013), which has become known as the "Poissonization lemma". We use $\mathrm{Poi}(\lambda)$ to denote the Poisson distribution with mean λ: For $t \sim \mathrm{Poi}(\lambda)$ and $k \geq 0$,

$$\mathbf{Pr}[t = k] = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Lemma 2 *If, for $t \sim \mathrm{Poi}(n)$, you throw t balls independently uniformly at random into m bins,*

- *the numbers of balls falling into the bins are mutually independent, and*
- *the number of balls falling in each bin is distributed as $\mathrm{Poi}(n/m)$.*

For each n, our proof uses three distributions: D_n, Q_n and P_n. The first, D_n, is used to define Q_n and P_n; it is chosen so that the least-norm interpolant performs well on D_n. The distribution Q_n is defined so that the least-norm interpolant performs poorly on Q_n. The distribution P_n is defined so that, when the least-norm interpolant performs well on D_n, it also performs well on P_n. Crucially, the least norm interpolants that arise from Q_n and P_n are closely related.

For each n, the joint distribution D_n on (x, y)-pairs is defined as follows. Let $s = \lfloor\sqrt{n}\rfloor$, $N = s^2$, $d = N^2$. Let θ^* be an arbitrary unit-length vector. Let Σ_s be an arbitrary covariance matrix with eigenvalues $\lambda_1 = 1/81, \lambda_2 = \cdots = \lambda_d = 1/d^2$. The marginal of D_n on x is then $\mathcal{N}(0, \Sigma_s)$. For each $x \in \mathbb{R}^d$, the distribution of y given x is $\mathcal{N}\left(\theta^* \cdot x, \frac{1}{81}\right)$. For $d \geq 9$, since (x, y) is Gaussian, $\|\Sigma_s\| \leq 1/81$, and the variance of y is $1/81$, each D_n is innocuous.

For an absolute constant positive integer b, we get Q_n from D_n through the following steps.

1. Sample $(x_1, y_1), \ldots, (x_{bn}, y_{bn}) \sim D_n^{bn}$.

2. Define Q_n on $\mathbb{R}^d \times \mathbb{R}$ so that its marginal on $\mathbb{R}^d$ is uniform on $U = \{x_1, \ldots, x_{bn}\}$ and its conditional distribution of $Y|X$ is the same as D_n.

Definition 3 *For a sample S, the* compression *of S, denoted by $C(S)$, is defined to be*

$$C(S) = ((u_1, v_1), \ldots, (u_k, v_k)),$$

where $u_1, \ldots, u_k$ are the unique elements of $\{x_1, \ldots, x_n\}$, and, for each i, v_i is the average of $\{y_j : 1 \leq j \leq n, x_j = u_i\}$.

For the least-norm interpolation algorithm A, for any pair S and S' of samples such that $C(S) = C(S')$, we have $A(S) = A(S')$. (This is true because the least-norm interpolant $A(S)$ is uniquely defined by the equality constraints specified by the compression $C(S)$.)

So that a generalization bound often must apply to Q_n, we need to show that it is likely to be unit scale. The proof of this lemma is in Appendix A.

Lemma 4 *There is a positive constant c_5 such that, for all large enough n, with probability at least $1 - \frac{c_5}{n}$, Q_n has unit scale.*

We can show that the least-norm interpolant is bad for Q_n by only considering the points in the support of Q_n that the algorithm sees exactly once.

Lemma 5 *For any constant $c > 0$, there are constants $c_6, c_7 > 0$ such that, for all sufficiently large n, almost surely for Q_n chosen randomly as described above, if t is chosen randomly according to* Poi(cn) *and S consists of t random draws from Q_n, then with probability at least $1 - e^{-c_6 n}$ over t and S,*

$$\mathbf{E}_{(x,y)\sim Q_n}\left[(A(C(S))(x) - y)^2\right] - \mathbf{E}_{(x,y)\sim Q_n}\left[(f^*(X) - Y)^2\right] \geq c_7$$

where f^* is the regression function for D_n (and hence also for Q_n).

Proof Recall that $U = \{x_1, \ldots, x_{bn}\}$ is the support of the marginal of Q_n on the independent variables. With probability 1, U has cardinality bn. Define $h = A(C(S))$. If some $x \in U$ appears exactly once in S, then $h(x)$ is a sample from the distribution of y given x under D_n. Thus, for such an x, the expected quadratic loss of $h(x)$ on a test point is the squared difference between two independent samples from this distribution. This is twice the variance of this distribution, that is, twice the expected loss of f^*, hence 2/81. On any $x \in U$, whether or not it was seen exactly once in S, by definition, $f^*(x)$ minimizes the expected loss given x.

Lemma 2 shows that, conditioned on the random choice of Q_n, the numbers of times the various x in U in S are mutually independent and, the probability that $x \in U$ is seen exactly once in S is $\frac{c}{b}\exp\left(-\frac{c}{b}\right) \geq \frac{ce^{-c}}{b}$. Applying a Chernoff bound (see, for example Mitzenmacher and Upfal, 2005, Theorem 4.5), the probability that fewer than $ce^{-c}n/2$ members of U

are seen exactly once in S is at most $e^{-c_6 n}$ for an absolute constant c_6. Thus if U_1 is the (random) subset of points in U that were seen exactly once, we have

$$\begin{aligned}
&Q_n\left[(h(X)-Y)^2\right] - Q_n\left[(f^*(X)-Y)^2\right] \\
&= \sum_{x\in U} Q_n\left[((h(X)-Y)^2 - (f^*(X)-Y)^2)1_{X=x}\right] \\
&\geq \sum_{x\in U_1} \mathbf{E}[(f^*(X)-Y)^2 1_{X=x}] \\
&= \frac{|U_1|}{81bn}.
\end{aligned}$$

Since, with probability $1-e^{-c_6 n}$, $|U_1| \geq ce^{-c}n/2$, this completes the proof. ■

Definition 6 *Define P_n as follows.*

1. *Set the marginal distribution of P_n on $\mathbb{R}^d$ the same as that of D_n.*
2. *To generate Y given $X = x$ for $(X,Y) \sim P_n$, first sample a random variable Z whose distribution is obtained by conditioning a draw from a Poisson with mean $\frac{c}{b}$ on the event that it is at least 1, then sample Z values $V_1, \ldots, V_Z$ from the conditional distribution $D_n(Y|X=x)$, and set $Y = \frac{1}{Z}\sum_{i=1}^{Z} V_j$.*

Note that, since D_n has a density, $x_1, ..., x_r$ are almost surely distinct and hence S drawn from P_n^r has $C(S) = S$ a.s.

The following lemma implies that the bounds for P_n tend to be as big as those for Q_n.

Lemma 7 *Define Q_n as above and let $\mathcal{Q}_n$ be the resulting distribution over the random choice of Q_n. Suppose P_n is defined as in Definition 6. Let $c > 0$ be an arbitrary constant. Choose S randomly by choosing $t \sim \mathrm{Poi}(cn)$, $Q_n \sim \mathcal{Q}_n$ and $S \sim Q_n^t$. Choose T by choosing $r \sim B\left(bn, 1-\exp\left(-\frac{c}{b}\right)\right)$ and $T \sim P_n^r$. Then $C(S)$ and T have the same distribution. In particular, for all $\delta > 0$, for any function ψ of the least norm interpolant h, a sample size r, and a confidence parameter δ, we have*

$$\mathbf{E}_{t\sim\mathrm{Poi}(cn), Q_n\sim\mathcal{Q}_n}[\mathbf{E}_{S\sim Q_n^t}[\psi(h(S), |C(S)|, \delta)]] = \mathbf{E}_{r\sim B\left(bn, 1-\exp\left(-\frac{c}{b}\right)\right)}[\mathbf{E}_{T\sim P_n^r}[\psi(h(T), r, \delta)].$$

Proof Let $\mathcal{C}$ be the probability distribution over training sets obtained by picking Q_n from $\mathcal{Q}_n$, picking t from $\mathrm{Poi}(cn)$, picking S from Q_n^t and compressing it. Let $C = C(S)$ be a random draw from $\mathcal{C}$. Let n_C be the number of examples in C.

We claim that n_C is distributed as $B\left(bn, 1-\exp\left(-\frac{c}{b}\right)\right)$. Conditioned on Q_n, and recalling that U is the support of Q_n, for any $x \in U$, Lemma 2 implies that for each $x \in U$, the probability that x is not seen is the probability, under a Poisson with mean $\frac{c}{b}$, of drawing a 0. Thus, the probability that x is seen is $1-\exp\left(-\frac{c}{b}\right)$. Since the numbers of times different x are seen in S are independent, the number seen is distributed as $B\left(bn, 1-\exp\left(-\frac{c}{b}\right)\right)$.

Now, for each $x \in U$, the event that it is in $C(S)$ is the same as the event that at appears at least once in S. Thus, conditioned on the event that x appears in S, the number of y

values that are used to compute the y value in $C(S)$ is distributed as a Poisson with mean $\frac{c}{b}$, conditioned on having a value at least 1.

Let $D_{n,X}$ be the marginal distribution of D_n on the x's. If we make n independent draws from $D_{n,X}$, and then independently reject some of these examples, to get n_C draws, the resulting n_C examples are independent. (We could first randomly decide the number n_C of examples to keep, and then draw those independently from D_n, and we would have the same distribution.)

The last two paragraphs together, along with the definition of Q_n, imply that the distribution over T obtained by sampling r from $B(bn, 1-e^{-c/b})$ and T from P_n^r is the same as the distribution over C obtained by sampling Q_n from $\mathcal{Q}_n$, t from $\text{Poi}(cn)$, then sampling S from Q_n^t and compressing it. Thus, the distributions of T and $C(S)$ are the same, and hence the distributions of $(h(T), |T|)$ and $(h(S), |C(S)|)$ are the same, because $h(S) = h(C(S))$. ■

We will use the following bound on a tail of the Poisson distribution, due to Canonne (2017).

Lemma 8 *For any $\lambda, \alpha > 0$, $\mathbf{Pr}_{r\sim\text{Poi}(\lambda)}(r \geq (1+\alpha)\lambda) \leq \exp\left(-\frac{\alpha^2}{2(1+\alpha)}\lambda\right)$.*

Armed with these tools, we now show that ϵ must often have a large value.

Lemma 9 *Then there are positive constants c_1, c_2, c_3, c_4 such that, for all $0 < \delta < c_1$, the set of n such that*

$$\mathbf{Pr}_{S\sim P_n^n}\left[\epsilon(h, n, \delta) > c_2\right] \geq \frac{1}{2}$$

is strongly $(1 - \frac{c_3}{\log(1/\delta)})$-dense beyond $c_4 \log(1/\delta)$.

Proof We will think of the natural numbers as being divided into bins $[1,2), [2,4), [4,7), ...$ Let us focus our attention on one bin: $\{s^2, ..., (s+1)^2 - 1\}$. Let n denote the center of the bin, $n = s^2 + s$, so that $s \sim \sqrt{n}$.

For a constant $c_8 > 0$ and any $\delta > 0$, Lemma 7 implies

$$\mathbf{E}_{r\sim B(bn,1-e^{-c/b})}[\mathbf{Pr}_{T\sim P_n^r}[\epsilon(h(T), r, \delta) \leq c_8]] = \mathbf{E}_{t\sim\text{Poi}(cn), Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[\epsilon(h(S), |C(S)|, \delta) \leq c_8]]. \tag{1}$$

Suppose that ϵ is B'-bounded-antimonotonic. Fix $B > 0$ such that $B > B'$. Then

$$\begin{aligned}
&\mathbf{E}_{t\sim\text{Poi}(cn), Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[\epsilon(h, |C(S)|, \delta) \leq c_8]] \\
&= \mathbf{E}_{t\sim\text{Poi}(cn), Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[B\epsilon(h, |C(S)|, \delta) \leq c_8 B]] \\
&\leq \mathbf{E}_{t\sim\text{Poi}(cn), Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[R_{Q_n}(h) - R^*_{Q_n} > B\epsilon(h, |C(S)|, \delta)]] \\
&\qquad + \mathbf{E}_{t\sim\text{Poi}(cn), Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[R_{Q_n}(h) - R^*_{Q_n} \leq c_8 B]].
\end{aligned} \tag{2}$$

For each sample size t and any Q_n that has unit scale

$$\begin{aligned}
&\mathbf{Pr}_{S\sim Q_n^t}[R_{Q_n}(h) - R^*_{Q_n} > B\epsilon(h, |C(S)|, \delta)] \\
&\leq \mathbf{Pr}_{S\sim Q_n^t}\left[R_{Q_n}(h) - R^*_{Q_n} > \epsilon(h, t, \delta)\right] + \mathbf{Pr}_{S\sim Q_n^t}\left[B\epsilon(h, |C(S)|, \delta) \leq \epsilon(h, t, \delta)\right] \\
&\leq \delta + \mathbf{Pr}_{S\sim Q_n^t}[|C(S)| < t/2]
\end{aligned}$$

where the second inequality follows from the fact that ϵ is a valid B'-bounded-antimonotonic uniform model-dependent bound for unit-scale distributions and $B > B'$. Combining this with Lemma 4, we have

$$\mathbf{E}_{t\sim\mathrm{Poi}(cn),Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[R_{Q_n}(h)-R^*_{Q_n} > B\epsilon(h,|C(S)|,\delta)]] \le \delta+\frac{c_5}{n}+\mathbf{Pr}_{S\sim Q_n^t}[|C(S)|<t/2].$$

Now by a union bound, for some constant $c_9 > 0$,

$$\begin{aligned}
&\mathbf{E}_{t\sim\mathrm{Poi}(cn),Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[|C(S)|<t/2]] \\
&\le \mathbf{E}_{t\sim\mathrm{Poi}(cn),Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[|C(S)|<c_9 n]] + \mathbf{Pr}_{t\sim\mathrm{Poi}(cn)}[t/2\ge c_9 n] \\
&= \mathbf{Pr}_{Z\sim B(bn,1-e^{-c/b})}[Z<c_9 n] + \mathbf{Pr}_{t\sim\mathrm{Poi}(cn)}[t/2\ge c_9 n] \\
&\le \delta, \qquad (3)
\end{aligned}$$

where the last inequality follows from a Chernoff bound and from Lemma 8 with $\alpha = 1-2c_9/c$, provided $n=\Omega(\log(1/\delta))$ and provided we can choose c_9 to satisfy $c/2 < c_9 < b(1-e^{-c/b})$. Our choice of b and c, specified below, will ensure this. In that case, we have that

$$\begin{aligned}
&\mathbf{E}_{t\sim\mathrm{Poi}(cn),Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[\epsilon(h,|C(S)|,\delta)\le c_8]] \\
&\le 2\delta + \frac{c_5}{n} + \mathbf{E}_{t\sim\mathrm{Poi}(cn),Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[R_{Q_n}(h)-R^*_{Q_n}\le c_8 B]].
\end{aligned}$$

Applying Lemma 5 to bound the RHS, if n is large enough and $c_8 B < c_7$, then

$$\mathbf{E}_{t\sim\mathrm{Poi}(cn),Q_n\sim\mathcal{Q}_n}[\mathbf{Pr}_{S\sim Q_n^t}[\epsilon(h,|C(S)|,\delta)\le c_8]] \le 3\delta+\frac{c_5}{n}.$$

Returning to (1), we get

$$\mathbf{E}_{r\sim B(bn,1-e^{-c/b})}[\mathbf{Pr}_{T\sim P_n^r}[\epsilon(h,r,\delta))\le c_8]] \le 3\delta+\frac{c_5}{n}. \qquad (4)$$

Let us now focus on the case that $b=2$ and $c=2\ln 2$, so that

$$\mathbf{E}_{r\sim B(bn,1-e^{-c/b})}[r] = (1-e^{-c/b})bn = n.$$

(And note that $c/2 = \ln 2 < 1 = b(1-e^{-c/b})$, as required for (3).) Chebyshev's inequality implies

$$\mathbf{Pr}_{r\sim B(bn,1-e^{-c/b})}[r\notin[n-s,n+s]] \le c_{10}$$

for an absolute positive constant c_{10}. Returning now to (4), Markov's inequality implies

$$\mathbf{Pr}_{r\sim B(bn,1-e^{-c/b})}[\mathbf{Pr}_{T\sim P_n^r}[\epsilon(h,r,\delta))\le c_8] > 1/2] \le c_{11}\left(\delta+\frac{1}{n}\right). \qquad (5)$$

Further, it is known (Slud, 1977; Box et al., 1978) that there is an absolute constant c_{12} such that, for all large enough n and all $r_0\in[n-s,n+s]$,

$$\mathbf{Pr}_{r\sim B(bn,1-e^{-c/b})}[r=r_0] \ge \frac{c_{12}}{\sqrt{n}}.$$

Combining this with (5) and recalling that s and s' are $\Theta(\sqrt{n})$, we get

$$\frac{|\{r \in [n-s, n+s] : \mathbf{Pr}_{T\sim P_n^r}[\epsilon(h, r, \delta)) \leq c_8] > 1/2\}|}{2s+1} \leq c_{13}\left(\delta + \frac{1}{n}\right) \leq \frac{c_{14}}{\log(1/\delta)},$$

for $n \geq c_4 \log(1/\delta)$ and small enough δ. Since, for all $r \in [n-s, n+s]$, we have $P_r = P_n$, this completes the proof. ■

The following bound can be obtained through direction application of the results of Bartlett et al. (2020). The details are given in Appendix B.

Lemma 10 *There is a constant c such that, for all large enough n, with probability at least $1-\delta$, for $S \sim P_n^n$, the least-norm interpolant h satisfies $R_{P_n}(h) - R^*_{P_n} \leq c\sqrt{\frac{\log(1/\delta)}{n}}$.*

Combining this with Lemma 9 proves Theorem 1.

4. Conclusion

We have shown that valid bounds that depend only on the output of the learning algorithm must be too loose to establish benign overfitting of linear regressors in the sense of Bartlett et al. (2020); Tsigler and Bartlett (2020); additional information, such as properties of the covariance, is necessary.

There are several interesting directions for further work. The construction presented in this paper uses a different probability distribution for each sample size. It is not clear whether this is necessary. It seems likely that a wider variety of generalization bounds than the class analyzed in this paper must sometimes be loose. Our theoretical understanding of benign overfitting in the context of deep learning, where it was originally observed, is much less developed than in the linear regression setting considered here.

Acknowledgments

We thank Andrea Montanari for alerting us to a flaw in an earlier version of this paper, and the JMLR reviewers for their helpful comments and suggestions. We also thank Vaishnavh Nagarajan and Zico Kolter for helpful comments on an earlier draft of this paper, and Dan Roy for calling our attention to (Negrea et al., 2020, Lemma 5.2).

Appendix A. Proof of Lemma 4

To prove Lemma 4, we will need some lemmas. The first is due to Buldygin and Kozachenko (1980) (see Rivasplata, 2012).

Lemma 11 *If X_1 is a sub-Gaussian random variable with parameter σ_1, and X_2 is a (not necessarily independent) sub-Gaussian random variable with parameter σ_2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sigma_1 + \sigma_2$.*

This immediately implies the following.

Lemma 12 *If X_1 is a sub-Gaussian random vector with parameter σ_1, and X_2 is a (not necessarily independent) sub-Gaussian random vector with parameter σ_2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sigma_1 + \sigma_2$.*

Proof Any projection of $X_1 + X_2$ is the sum of the projections of X_1 and X_2, so this follows from Lemma 11. ■

Lemma 13 *For a random vector $X = (X_1, ..., X_k)$, if X_1 is sub-Gaussian with parameter $1/3$, X_2 is sub-Gaussian with parameter $1/3$, and $(X_3, ..., X_k)$ is sub-Gaussian with parameter $1/3$, then X is sub-Gaussian with parameter 1.*

Proof Embedding a random vector into a higher-dimensional space by adding components that always evaluate to zero does not affect whether it is sub-Gaussian, or its sub-Gaussian parameter. Since

$$X = (X_1, 0, ..., 0) + (0, X_2, 0, ..., 0) + (0, 0, X_3, ..., X_k),$$

applying Lemma 12 above completes the proof. ■

Now, for $U \sim D_n^m$, the uniform distribution Q over U, and $(X_1, ..., X_d, Y) \sim Q$, we now would like to show that X_1 is sub-Gaussian with parameter $1/3$. We will use the following known sufficient condition, which can be recovered by tracing through the constants in the proof of (Vershynin, 2018, Proposition 2.5.2).

Lemma 14 *If a random variable X satisfies $\mathbf{E}\left[\exp\left(\frac{18X^2}{e}\right)\right] \leq 2$, then X is sub-Gaussian with parameter $1/3$.*

Now we are ready to analyze the marginal distribution of the first component.

Lemma 15 *For U obtained from m independent samples from $\mathcal{N}(0, \sigma^2)$ for $\sigma \leq 1/9$ if Q is the uniform distribution over U, then, with probability at least $1 - \frac{3}{m}$, Q is sub-Gaussian with parameter $1/3$.*

Proof Define $a = 18/e$ and let $Z = \mathbf{E}_{x\sim Q}[\exp(ax^2)]$.

We have

$$\begin{aligned}
\mathbf{E}_{S\sim\mathcal{N}(0,\sigma)^m}[Z] &= \mathbf{E}_{S\sim\mathcal{N}(0,\sigma)^m}[\mathbf{E}_{x\sim Q}[\exp(ax^2)]] \\
&= \mathbf{E}_{x\sim\mathcal{N}(0,\sigma)}[\exp(ax^2)] \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{ax^2}\exp\left(-\frac{x^2}{2\sigma^2}\right)\,dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} \exp\left(-\left(\frac{1}{2\sigma^2} - a\right)x^2\right)\,dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \times \frac{\sqrt{\pi}}{\sqrt{\frac{1}{2\sigma^2} - a}} \\
&= \frac{1}{\sqrt{1 - 2a\sigma^2}}.
\end{aligned}$$

Similarly

$$\begin{aligned}\mathbf{Var}_S[Z] &= \mathbf{Var}_S[\mathbf{E}_{x\sim Q}[\exp(ax^2)]] \\ &= \frac{1}{m}\mathbf{Var}_{x\sim\mathcal{N}(0,\sigma)}[\exp(ax^2)] \\ &\leq \frac{1}{m}\mathbf{E}_{x\sim\mathcal{N}(0,\sigma)}[\exp(2ax^2)] \\ &= \frac{1}{m\sqrt{1-4a\sigma^2}}.\end{aligned}$$

By Chebyshev's inequality,

$$\mathbf{Pr}\left[Z \geq \frac{1}{\sqrt{1-2a\sigma^2}} + \frac{1}{\sqrt{3}(1-4a\sigma^2)^{1/4}}\right] \leq \frac{3}{m}.$$

For $\sigma \leq 1/9$, recalling that $a = 18/e$ shows that $\mathbf{Pr}[Z \geq 2] \leq \frac{3}{m}$ and applying Lemma 14 completes the proof. ■

Armed with these lemmas, we are now ready to prove Lemma 4. For $S \sim D_n^m$, let Q be the uniform distribution over S. For $(X_1, ..., X_d, Y) \sim Q$, Lemma 15 implies that, with probability $1 - 6/m$, X_1 and Y are both sub-Gaussian with parameter $1/3$. It remains to analyze $(X_2, ..., X_d)$. Let S' be the projections of the elements of S onto these coordinates. With probability at least $1 - 3/m$, for all $s' \in S'$, $||s'|| \leq \log(em^2/3)/\sqrt{d}$; see (Lovász and Vempala, 2007, Lemma 5.17). Recalling that $d = \Theta(n^2)$, if $m = bn$, then, for all large enough n, with probability $1 - 3/m$, $\max_{s'\in S'} ||s|| \leq 1/6$, which implies that $(X_2, ..., X_d)$ is sub-Gaussian with parameter $1/3$. Putting this together with the analysis of X_1 and Y, and applying Lemma 13, completes the proof.

Appendix B. Proof of Lemma 10

The lemma follows from (Bartlett et al., 2020, Theorem 1); before showing how to apply it, let us first restate a special case of the theorem for easy reference.

B.1 A useful upper bound

The special case concerns the least-norm interpolant applied to training data $(x_1, y_1), ..., (x_n, y_n)$ drawn from a joint distribution P over (x, y) pairs. The marginal distribution of x is Gaussian with covariance Σ. There is a unit length θ^* such that, for all x, the conditional distribution of y given x has mean $\theta^* \cdot x$ is sub-gaussian with parameter 1 and variance at most 1.

We will apply an upper bound in terms of the eigenvalues $\lambda_1 \geq \lambda_2 \geq ...$ of Σ. The bound is in terms of two notions of the effective rank of the tail of this spectrum:

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

The rank of Σ is assumed to be greater than n.

Lemma 16 () *There are $b, c, c_1 > 1$ for which the following holds. For all n, P and Σ defined as above, write $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$. Suppose that $\delta < 1$ with $\log(1/\delta) < n/c$. If $k^* < n/c_1$, then, with probability at least $1-\delta$, the least-norm interpolant h satisfies*

$$R_P(h) - R_P^* \leq c\left(\max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}.\frac{r_0(\Sigma)}{n}.\sqrt{\frac{\log(1/\delta)}{n}}\right\} + \log(1/\delta)\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\right).$$

Lemma 16 was sharpened and generalized by Tsigler and Bartlett (2020). Both Bartlett et al. (2020) and Tsigler and Bartlett (2020) provide examples of concrete instantiations of their bounds.

B.2 The proof

To prove Lemma 10, we need to show that P_n satisfies the requirements on P in Lemma 16, and evaluate the effective ranks r_k and R_k of P_n's covariance Σ_s. Define $\alpha = 1/d^2$. We have

$$r_0 = \frac{1/81 + (d-1)\alpha}{1/81} = 1 + 81(d-1)\alpha$$

(which is bounded by a constant) and

$$R_0 = \frac{(1/81 + (d-1)\alpha)^2}{1/81^2 + (d-1)\alpha^2}.$$

For $k > 0$,

$$r_k = R_k = d - k.$$

Since d grows faster than n, for large enough n, $k^* := \min\{k : r_k \geq bn\} = 1$. So

$$R_{k^*} = d - 1 = \Omega(n^2).$$

Each sample from the distribution of Y given $X = x$ has a mean of $\theta^* \cdot x$, and is sub-Gaussian with parameter at most $\frac{1}{9}$, and with variance at most $1/81$ (because increasing Z only decreases the variance of Y).

Evaluating Lemma 16 on P_n then gives Lemma 10.

References

S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, volume 97, pages 322–332, 2019.

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 Transactions on Information Theory*, 44(2):525–536, 1998.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

P. L. Bartlett, D. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NIPS*, pages 6240–6249, 2017.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *PNAS*, 2020.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 259–269. IEEE, 2000.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4:1–4:25, 2013. doi: 10.1145/2432622.2432626. URL `https://doi.org/10.1145/2432622.2432626`.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical biasvariance trade-off. *PNAS*, 116(32):15849–15854, 2019a.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? *AISTATS*, pages 1611–1619, 2019b.

MW Birch. Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 25(1):220–233, 1963.

GEP Box, WH Hunter, S Hunter, et al. *Statistics for experimenters*, volume 664. John Wiley and sons New York, 1978.

V. V. Buldygin and Y. V. Kozachenko. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32(6):483–489, 1980.

C. L. Canonne. A short note on Poisson tail bounds, 2017.

Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10835–10845, 2019.

M. Derezinski, F. T. Liang, and M. W. Mahoney. Exact expressions for double descent and implicit regularization via surrogate random design. In *NeurIPS*, 2020.

V. Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

W. Feller. *An introduction to probability theory and its applications*, volume 1. John Wiley & Sons, 1968.

N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *COLT*, volume 75, pages 297–299, 2018.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019a.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. Technical Report 1903.08560 [math.ST], arXiv, 2019b. URL `https://arxiv.org/abs/1903.08560`.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

Y. Li and Y. Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *NeurIPS*, pages 8157–8166. 2018.

Z. Li, W. Su, and D. Sejdinovic. Benign overfitting and noisy features. *arXiv preprint arXiv:2008.02901*, 2020.

T. Liang and A. Rakhlin. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

P. M. Long and H. Sedghi. Generalization bounds for deep convolutional neural networks. *ICLR*, 2019.

L. Lovász and S. S. Vempala. The geometry of logconcave functions and sampling algorithms. *Random Struct. Algorithms*, 30(3):307–358, 2007.

M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

V. Nagarajan and J. Z. Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *NeurIPS*, pages 11611–11622, 2019.

J. Negrea, G. K. Dziugaite, and D. M. Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. *ICML*, 2020.

B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *COLT*, pages 1376–1401, 2015.

O. Rivasplata. Subgaussian random variables: An expository note. `https://sites.ualberta.ca/~omarr/publications/subgaussians.pdf`, 2012. Downloaded 1/12/2021.

E. V. Slud. Distribution inequalities for the binomial law. *The Annals of Probability*, pages 404–412, 1977.

A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression, 2020. ArXiv, 2009.14286.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.