In Defense of Uniform Convergence: Generalization via Derandomization with an Application to Interpolating Predictors

Jeffrey Negrea [1 2 3] **Gintare Karolina Dziugaite** [4 3] **Daniel M. Roy** [1 2 3]

Abstract

We propose to study the generalization error of a learned predictor $\hat{h}$ in terms of that of a surrogate (potentially randomized) predictor that is coupled to $\hat{h}$ and designed to trade empirical risk for control of generalization error. In the case where $\hat{h}$ interpolates the data, it is interesting to consider theoretical surrogate predictors that are partially derandomized or rerandomized, e.g., fit to the training data but with modified label noise. We also show that replacing $\hat{h}$ by its conditional distribution with respect to an arbitrary σ-field is a convenient way to derandomize. We study two examples, inspired by the work of Nagarajan and Kolter (2019) and Bartlett et al. (2020), where the learned predictor $\hat{h}$ interpolates the training data with high probability, has small risk, and, yet, does not belong to a nonrandom class with a tight uniform bound on two-sided generalization error. At the same time, we bound the risk of $\hat{h}$ in terms of surrogates constructed by conditioning and denoising, respectively, and shown to belong to nonrandom classes with uniformly small generalization error.

1. Introduction

One of the central problems in learning theory is to explain the statistical performance of deep learning algorithms. There is particular interest in explaining how overparameterized neural networks, trained by simple variants of stochastic gradient descent (SGD), simultaneously achieve low risk and zero empirical risk on benchmark datasets. While certain naive explanations have been ruled out (Zhang et al., 2017), progress has been slow.

[1]Stat. Sci., U. Toronto [2]Vector Inst. [3]Inst. Adv. Study [4]Element AI. Correspondence to: J. Negrea <negrea@utstat.toronto.edu>, G. K. Dziugaite <karolina.dziugaite@elementai.com>, D. M. Roy <droy@utstat.toronto.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

The bulk of recent work has approached this problem by arguing that the classifier learned by SGD belongs to a class for which there is a uniform and tight (two-sided) bound on the generalization error (Bartlett, Foster, and Telgarsky, 2017; Golowich, Rakhlin, and Shamir, 2018; Long and Sedghi, 2020; Neyshabur et al., 2017; Wei and Ma, 2019). After raising this observation, Nagarajan and Kolter (2019) argue that this approach may be unable to explain performance observed in overparameterized models. They argue this point by constructing a simple problem where an SGD-like algorithm learns a classifier that achieves low risk and zero empirical risk, yet the learned classifier does not belong (even with high probability) to a class whose generalization error is uniformly small. Nagarajan and Kolter conjecture that SGD finds a fit that is simple at a macroscopic level (leading to good generalization) but complex at the microscopic level due to data-dependent fluctuations (hurting "uniform convergence"). A similar observation for minimum-norm interpolating kernel methods is made by Belkin, Ma, and Mandal (2018, Sec. 4), where they find that norm-dependent bounds will be vacuously large.

In this work, we initiate a response to Nagarajan and Kolter (2019) in defense of the utility of uniform convergence for understanding learning algorithms that obtain zero empirical risk. We do so by introducing the technique of *derandomization*, whereby a risk bound is obtained by comparing the learned hypothesis with a surrogate that is less influenced by the data. This technique can be seen to be implicit in several classical analyses, including the bias–variance analysis of linear regression. We look at the high-dimensional case as our first example. Our second example mimics one by Nagarajan and Kolter and shows how *re*randomization can produce a derandomized surrogate without microscopic fluctuations. In each case, we show that the surrogate belongs to a class possessing an appropriate uniform convergence property and is similar enough to the original learned hypothesis to yield a tight bound on risk.

In order to formalize our results, we require a notion of uniform convergence that is applicable to learning algorithms that produce interpolating predictors.[1] In general,

[1]We use the term "interpolating predictors" to refer to predic-

to learn an interpolating predictor, the capacity of the hypothesis space must increase with the dataset size. This mirrors deep learning practice, where scientists will train larger, more complex models when presented with a larger dataset. Since the complexity of the learning problems in question—and possibly even the sample spaces generating the data—change with the sample size, the traditional notions of uniform convergence (Glivenko–Cantelli classes) are not applicable. Therefore we need to extend the concept of uniform convergence to the setting of sequences of learning problems of increasing complexity, which we do in Section 3 by defining the *structural Glivenko–Cantelli* property.

In Section 4, we introduce a general approach to relate sequences of learning problems which are not structural Glivenko–Cantelli to ones that may be. The basic idea is to introduce a surrogate hypothesis that is coupled to the output of the learning algorithm of interest, yet belongs to a class (the surrogate hypothesis class) for which a uniform and vanishing bound on two-sided generalization error holds. Surrogates arise implicitly in several existing analyses of interpolating hypotheses, including the risk analysis of 1-Nearest Neighbors (Devroye, Györfi, and Lugosi, 2013). Recent work can also be interpreted as employing surrogates, e.g., in high-dimensional halfspace learning (Kabán and Durrant, forthcoming) and deep learning (Mücke and Steinwart, 2019).

The observations of Nagarajan and Kolter (2019) relate to a number of other empirical learning phenomena that demand explanation. One example is the phenomenon of double descent, brought to light by Advani and Saxe (2017), Belkin et al. (2019), and Geiger et al. (2019). The difficulty of explaining these double descent curves using standard uniform convergence arguments is a central theme of recent talks by Belkin. In a line of work by Hastie et al. (2019) and Mei and Montanari (2019), double descent was observed in unregularized, overparameterized linear regression. Bartlett et al. (2020) show that, for sequences of overparameterized linear regression tasks, the minimum norm interpolating solution to least squares will achieve asymptotically optimal risk with high probability given constraints on the covariate (feature) distribution. In this setting, we show that no class containing the learned hypothesis with high probability can have a vanishing uniform bound on the absolute generalization error. In fact, such a bound cannot be representative of the risk of the learned hypothesis. In Section 5, we show that the analysis of Bartlett et al. (2020) may be viewed as introducing a surrogate classifier. The surrogate in this case is the minimum-norm interpolating solution on the training data *with label noise removed*. We use standard techniques from empirical process theory to demonstrate that the surrogate hypothesis class—the collection of all surrogate hypotheses that could have been learned from the training data—has the structural Glivenko–Cantelli property. We combine our uniform bound based on the structural Glivenko–Cantelli property with other components of the analysis of Bartlett et al. (2020) to obtain similar bounds on the expected risk of the minimum norm interpolating solution under weaker hypotheses.

In Section 6 we provide a relatively flexible recipe for constructing surrogate classifiers via probabilistic conditioning. The approach produces a probability measure over hypotheses via retraining on data that is equal in distribution to the original training data but has been partially "rerandomized". The approach effectively trades empirical risk for generalization error. Lastly, in Section 7, we apply this recipe to an example, inspired by Nagarajan and Kolter (2019), where an interpolating learning algorithm is constructed for which there is no structural Glivenko–Cantelli class containing the learned hypothesis with high probability. In that example, we construct a surrogate by conditioning with respect to a specific σ-field. We show the corresponding surrogate class is structural Glivenko–Cantelli and can be used to derive risk bounds for the learned classifier, which exhibit a form of double descent.

1.1. Contributions

In this work, we extend our theoretical understanding of generalization, by way of the following contributions:

1. Defining the structural Glivenko–Cantelli property, a notion of uniform convergence for sequences of learning problems.

2. Proposing to study generalization error of learning algorithms—including interpolating ones—in terms of surrogate hypotheses that may belong to structural Glivenko–Cantelli classes, even when the original hypotheses do not.

3. Demonstrating that the hypothesis spaces corresponding to a sequence of unregularized, overparameterized linear regression tasks are not structural Glivenko–Cantelli, but that they can be analyzed by introducing a sequence of surrogates for which the surrogate hypothesis class is structural Glivenko–Cantelli. We further use this fact to provide bounds on the expected risk of the original sequence of tasks under weaker hypotheses than Bartlett et al. (2020).

4. Introducing a generic technique by which one may introduce surrogate learning algorithms via conditioning,

tors that achieve zero empirical risk, borrowing the terminology used for functions that achieve zero mean squared error.

which naturally trades empirical risk for generalization error relative to the original learning algorithm.

5. Analyzing an example that distills the key features of an example in Nagarajan and Kolter (2019), via a family of surrogates obtained from conditioning. We show that, while the original learning algorithm does not output hypotheses in a sequence of classes with the structural Glivenko–Cantelli property, discarding a few bits of information leads to one that does. We also show that bounds obtained via the surrogate learning algorithm exhibit a form of double descent.

2. Preliminaries

Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random elements in a space $\mathcal{S}$ with common distribution $\mathcal{D}$. Let $S = (Z_1, \ldots, Z_n)$ represent the training set. Fix a loss function $\ell : \mathcal{H} \times \mathcal{S} \to \mathbb{R}_+$ for a space $\mathcal{H}$ of hypotheses. Let $\mathcal{M}_1(\mathcal{H})$ be the space of distributions on $\mathcal{H}$. Note that $\mathcal{H}$ can be embedded into $\mathcal{M}_1(\mathcal{H})$ by the map $h \mapsto \delta_h$ taking a classifier to a Dirac measure degenerating on $\{h\}$. For $Q \in \mathcal{M}_1(\mathcal{H})$, the *(average) loss* and *risk* are defined to be

$$\ell(Q, z) = \int \ell(h, z) Q(\mathrm{d}h), \quad L_{\mathcal{D}}(Q) = \int \ell(Q, z) \mathcal{D}(\mathrm{d}z).$$

Let $L_S(Q) = L_{\widehat{\mathcal{D}}_n}(Q)$ denote the *empirical (average) risk*, where $\widehat{\mathcal{D}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ is the empirical distribution. For $h \in \mathcal{H}$, define $L_{\mathcal{D}}(h) = L_{\mathcal{D}}(\delta_h)$ and $L_S(h) = L_S(\delta_h)$. Let $\hat{h}$ or $\hat{h}(S)$ be a random element in $\mathcal{H}$, representing a learned classifier.

A hypothesis h *interpolates* a dataset S with respect to a non-negative loss ℓ when $L_S(h) = 0$. A learning algorithm $\hat{h}(S)$ is (almost surely) *interpolating* if $L_S(\hat{h}(S)) = 0$ a.s. (or equivalently $\mathbb{E}L_S(\hat{h}_S) = 0$). This extends our geometric intuition that a surface $h : \mathbb{R}^d \to \mathbb{R}$ interpolates points in $\{(x_i, y_i)\}_{i\in[n]} \mathbb{R}^d \times \mathbb{R}$ when $(h(x_i) - y_i)^2 = 0$ for all $i \in [n]$. The surprising properties of interpolating classifiers are explored in Belkin et al. (2019). See also Advani and Saxe (2017) and Geiger et al. (2019).

3. Structural Uniform Convergence

Nagarajan and Kolter (2019) argue that uniform convergence does not explain generalization in several examples that are emblematic of the modern interpolating regime. In those examples, however, the size of the learning problem varies with the cardinality of the training dataset. The standard notion of uniform convergence (i.e., of Glivenko–Cantelli classes, etc.) is not normally defined in this setting. In order to formalize the specific failure of "uniform convergence" in these sequences of learning problems, we introduce a *structural* version of the Glivenko–Cantelli property.

Definition 3.1. Let $\{(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})\}_{p\in\mathbb{N}}$ be a sequence of probability spaces where $\mathcal{S}^{(p)}$ denotes the sample space, $\mathcal{F}^{(p)}$ denotes the σ-field and $\mathcal{D}^{(p)}$ denotes the probability measure. Let $\mathcal{H}^{(p)}$ be a collection of measurable functions on $(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})$ and let $n_p \in \mathbb{N}$ for all $p \in \mathbb{N}$.

Then $\mathcal{H}^{(\cdot)}$ has the *structural $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-Glivenko–Cantelli* property, denoted $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-SGC, if

$$\lim_{p\to\infty} \mathbb{E}\left[\sup_{h\in\mathcal{H}^{(p)}} \left|\mathcal{D}^{(p)}h - \widehat{\mathcal{D}^{(p)}}_{n_p} h\right|\right] = 0,$$

where $Ph = \int h(x)P(dx)$ and $\widehat{\mathcal{D}^{(p)}}_{n_p}$ is the empirical distribution of an IID sample of size n_p from $\mathcal{D}^{(p)}$.

It is this property which is made to fail in the examples presented by Nagarajan and Kolter (2019). When $\{(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})\}_{p\in\mathbb{N}}$ and $\mathcal{H}^{(p)}$ are constant and $n_p = p$, this reduces to the classical notion of Glivenko–Cantelli.

Remark 3.2 (Relationship between PAC and nonuniform learning). PAC learnability and nonuniform learnability, as defined in (Shalev-Shwartz and Ben-David, 2014), can both be understood in terms of the structural Glivenko–Cantelli property. That PAC learnability implies structural Glivenko–Cantelli follows from the equivalence of PAC learnability with the uniform Glivenko–Cantelli property.

To understand the nonuniform learnability of a class $\mathcal{H}$ in terms of the structural Glivenko–Cantelli property, recall the equivalence that $\mathcal{H}$ is non-uniformly learnable if and only if it is a countable union of VC classes—$\mathcal{H} = \bigcup_{j\in\mathbb{N}} \mathcal{H}_j$ with $\bigcup_{j\in[p]} \mathcal{H}_j$ of finite VC dimension d_p. Then, for any sequences $\delta_p \searrow 0$ and $\epsilon_p \searrow 0$, take $n_p \geq C_2 \frac{d_p + \log(1/\delta_p)}{\epsilon_p^2}$ where C_2 is the universal constant appearing in (Shalev-Shwartz and Ben-David, 2014, Theorem 6.8, Item 1.). Taking $\{(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})\}_{p\in\mathbb{N}}$ to be constant, and $\mathcal{H}^{(p)} = \bigcup_{j\in[p]} \mathcal{H}_j$, it follows immediately that

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}^{(p)}} \left|\mathcal{D}^{(p)}h - \widehat{\mathcal{D}^{(p)}}_{n_p} h\right|\right] \leq (1 - \delta_p)\epsilon_p + \delta_p \to 0,$$

and hence $\mathcal{H}^{(\cdot)}$ is $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-SGC. It would be reasonable in this case to say that $\mathcal{H}^{(\cdot)}$ is $(n_{(\cdot)})$-SGC uniformly over data generating distributions.

The partitioning of the hypothesis space in synchronization with increasing sample size in this derivation is similar to the partitioning of the hypothesis space by sample size occurring in the structural risk minimization algorithm for nonuniform learning. The analysis above tells us that any ERM algorithm restricted to $\mathcal{H}^{(p)}$ when the sample size is n_p will achieve low generalization error. ◁

4. Decompositions of Generalization Error using Surrogate Classifiers

We now describe how one may pass from bounding the generalization error of a learning algorithm to bounding the generalization error of a surrogate and controlling differences in the risk and empirical risk profiles between the original algorithm and the surrogate.

The following result is immediate from the linearity of expectation:

Lemma 4.1 (Surrogate decomposition)**.** *For every random element Q in $\mathcal{M}_1(\mathcal{H})$,*

$$\begin{aligned}\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h})] = \mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)] \\ + \mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)] \\ + \mathbb{E}[L_S(Q) - L_S(\hat{h})],\end{aligned}$$

provided the three expectations on the r.h.s. are finite.

This decomposition suggests that one can obtain a bound on the generalization error (and then the risk) of $\hat{h}$ by bounding the three terms individually. We interpret Q here as a (possibly randomized) surrogate hypothesis that is coupled with $\hat{h}$ via some information in the training algorithm and/or the training data. The choice of Q trades off one term for another. In the particular case of a.s. interpolating classifiers (i.e., $\mathbb{E}[L_S(\hat{h})] = 0$), one approach is to trade excess empirical risk, $\mathbb{E}[L_S(Q) - L_S(\hat{h})]$, for less generalization error, $\mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)]$.

One way to control generalization error is to show that Q belongs to a nonrandom class for which there holds a uniform and tight bound on generalization error.

Proposition 4.2 (Bounded loss, two-sided control)**.** *Assume ℓ takes values in an interval of length L. For every random element Q in $\mathcal{M}_1(\mathcal{H})$ and class $G \subseteq \mathcal{M}_1(\mathcal{H})$,*

$$\begin{aligned}&\mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)] \\ &\quad \le L\,\mathbb{P}[Q \notin G] + \mathbb{E}\Big[\sup_{P \in G} |L_{\mathcal{D}}(P) - L_S(P)|\Big].\end{aligned}$$

Just as we interpret Q as a surrogate hypothesis that depends on the dataset S, we view G in Proposition 4.2 as a surrogate hypothesis class that contains the surrogate hypothesis with high probability.

The surrogate decomposition may be viewed as similar to a one-step covering argument, where the cover is given by the class of surrogate hypotheses, and the approximation error is given by $\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)] + \mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)]$. In a typical one-step covering argument, the cover is chosen to be sufficiently fine as to have a uniformly small approximation error. The optimal cover density will vary with sample size so that approximation error vanishes as sample size increases. The key difference here is that we may not be able to control the approximation error uniformly or have any hope that it will vanish based on the covering induced by a surrogate. We will only attempt to uniformly control the cover given by the surrogate class. We then can rely on other techniques to handle the approximation error. This allows us to divide the objective of explaining generalization into a portion explained by uniform convergence and portion not explained by uniform convergence.

5. Overparameterized Linear Regression

Our first application of using surrogates and the structural Glivenko–Cantelli property to understand generalization error is inspired by the recent work of Bartlett et al. (2020). They determine necessary and sufficient conditions under which the *minimum norm interpolating linear predictor* generalizes well in mean-squared error for random design linear regression in the overparameterized regime (i.e., more features than observations) for sub-Gaussian random designs with conditionally sub-Gaussian residuals. We chose overparameterized linear regression as a first example to present because of 1) the failure of uniform convergence to directly explain performance of the learned hypothesis in the problem (in particular we show that the structural Glivenko–Cantelli property fails for any classes containing the learned hypotheses with high probability), and 2) recent work (e.g., by Hastie et al. (2019) and Mei and Montanari (2019) and others) showing that double descent occurs in variants of this problem such as random feature regression.

In overparameterized linear regression, our dataset is formed by n (feature, response) pairs, $((X_i, Y_i))_{i=1}^n$ with features as row vectors $X_i' \in \mathbb{R}^d$ and responses $Y_i \in \mathbb{R}$. The hypotheses are vectors in $\mathbb{R}^d$, and the loss function is squared error $\ell(\gamma, (x, y)) = (y - x\beta)^2$. Being *overparameterized* means that there are more features, d, than there are observations, n, i.e., $n < d$. For convenience we arrange the responses into a vector $Y = (Y_i)_{i=1}^n \in \mathbb{R}^n$, and the features into a *design matrix* $X = (X_i)_{i=1}^n \in \mathbb{R}^{n \times d}$ (so that the design matrix has one row per observation and one column per feature). Whenever X has full row rank and $n < d$, there is an affine space of dimension $d - n$ of interpolating hypotheses. The learned hypothesis which we will consider will be the minimum norm interpolating solution, $\hat{\beta} = (X'X)^+ X'Y$, where A^+ denotes the Moore–Penrose pseudoinverse of the matrix A. It is the risk properties of this learning algorithm which is analyzed by Bartlett et al. (2020). For simplicity, we will work with Gaussian design and Gaussian responses. Let $X_i \overset{\text{iid}}{\sim} \mathrm{N}_{1\times d}(0, \Sigma)$ be random row vectors with non-singular $d \times d$ feature covariance matrix Σ. Let $X = (X_1', \ldots X_n')'$ be the corresponding $n \times d$ random design matrix. Let $(Y_i \mid X) \overset{\text{ind}}{\sim} \mathrm{N}(X_i\beta, \sigma^2)$ and $Y = (Y_1, \ldots Y_n)'$ be the responses and response vector

respectively. Let $Z = Y - X\beta$ be the residual vector.

In order to remain overparameterized as the number of observations increases, it is necessary to consider a sequence of distinct learning problems so that the number of features grow with the number of observations. Thus, whether or not the learned hypothesis belongs to a classical Glivenko–Cantelli class with high probability is irrelevant to the determinaiton of the risk properties of the learned hypothesis when we stay overparameterized. We will see that the structural Glivenko–Cantelli property is relevant to the problem. However, there is no SGC class containing the learned hypothesis with high probability.

Lemma 5.1 (Failure of uniform convergence for overparameterized linear regression)**.** *There is no sequence of measurable sets* $\{A_n\}_{n\in\mathbb{N}}$ *such that* $\mathbb{P}((X,Y) \in A_n) > 2/3$ *for all* $n \in \mathbb{N}$ *and for which*

$$\limsup_{n\to\infty} \mathbb{E} \sup_{(\tilde{X},\tilde{Y})\in A_n} \left| L_D(\hat{\beta}(\tilde{X},\tilde{Y})) - L_S(\hat{\beta}(\tilde{X},\tilde{Y})) \right| \leq \frac{3}{2} L_D(\beta).$$

The proof of this result is found in Appendix B.

Even though there is no SGC class containing the learned hypothesis with high probability, Bartlett et al. (2020) are able to show that, under some conditions, the minimum norm solution is consistent, i.e., its risk converges to the Bayes risk. In the remainder of this section we discuss the conditions required by those authors to ensure consistency, and weaker conditions under which we prove similar results. Having formalied our hypotheses, we show that the bias–variance decomposition used by Bartlett et al. (2020) can be seen to be a natural surrogate decomposition, and that the sequence of surrogate classes is structural Glivenko–Cantelli. We then use a uniform convergence argument to prove the risk of the learned hypothesis converges to the Bayes risk under weaker conditions than those of Bartlett et al. (2020), closing the gap between their necessary and sufficient conditions for consistency.

We handle only the Gaussian design / Gaussian response case, but note that the results can be extended to the sub-Gaussian with minor modifications.[2] We provide bounds on the expected generalization error only, as the purpose of this example is to illustrate how uniform convergence of a surrogate may be used. However, we would like the reader to be aware that high-probability bounds may be obtained using the high-probability versions of the chaining arguments that appear in the proofs, as is standard in empirical process theory.

Bartlett et al. (2020) define a sequence of covariance matrices $\Sigma_n \in \mathbb{R}^{d_n \times d_n}$ to be *benign* when $\|\Sigma_n\| = 1$ and

$$\lim_{n\to\infty} \left(\sqrt{\frac{r_0(\Sigma_n)}{n}} + \frac{k_n^*(\Sigma_n)}{n} + \frac{n}{R_{k_n^*(\Sigma_n)}(\Sigma_n)} \right) = 0,$$

where, writing $\{\lambda_i(\Sigma)\}_{i\in[d]}$ for the eigenvalues of Σ in decreasing order (with multiplicity),

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

and $k_n^*(\Sigma) = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, for some universal constant $b > 0$ defined by Bartlett et al. Their work shows that it is sufficient that Σ_n be *benign* and $\|\beta_n\|^2 \sqrt{r_0(\Sigma_n)/n} \to 0$ in order that $L_D(\hat{\beta}) \to \sigma^2$ in probability. (In fact, $\mathbb{E}L_D(\hat{\beta}) \to \sigma^2$ also.) They then show that it is necessary that

$$\lim_{n\to\infty} \left(\frac{k_n^*(\Sigma_n)}{n} + \frac{n}{R_{k_n^*(\Sigma_n)}(\Sigma_n)} \right) = 0$$

in order that $\mathbb{E}L_D(\hat{\beta}) \to \sigma^2$. Note that the risk of the true coefficient vector (and global optimizer of the risk), β, is σ^2, and so we expect the risk of any estimator to be at least that large.

Bartlett et al. (2020) describe $r_k(\Sigma)$ and $R_k(\Sigma)$ as measures of *effective rank*. The condition, $\sqrt{r_0(\Sigma_n)/n} \to 0$, i.e., the effective rank grows slower than the sample size, parallels the classical fixed dimension setting where we need fewer features (measured by the rank of Σ) than observations, n. To achieve this, the spectrum of Σ must decay sufficiently quickly. In contrast, for $k_n^*(\Sigma)/n$ and $n/R_{k_n^*(\Sigma_n)}(\Sigma_n)$ to vanish, Bartlett et al. (2020) argue that we need the spectrum of Σ to decay sufficiently slowly. The requirement that the spectrum decay quickly enough to ensure $\sqrt{r_0(\Sigma_n)/n} \to 0$, but not too quickly to prevent $k_n^*(\Sigma)/n + n/R_{k_n^*(\Sigma_n)}(\Sigma_n) \to 0$, suggests there is a balance to strike.

In fact, for convergence in mean, we show that fast decay is unnecessary by establishing convergence in mean under strictly weaker hypotheses:

Definition 5.2. A sequence $\{\Sigma_n\}_{n\in\mathbb{N}}$ is *weakly benign* if

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

Clearly if $\{\Sigma_n\}_{n\in\mathbb{N}}$ is benign then it is weakly benign since $r_0(\Sigma) \geq 1$. The converse does not hold in general. The condition that $\{\Sigma_n\}_{n\in\mathbb{N}}$ forms a weakly benign sequence is, therefore, weaker than the sufficient condition of Bartlett et al. (2020) and equivalent to the necessary condition. Our weaker conditions remove the tension between rapid and slow decay of the spectrum required for the sequence to be benign. Instead, we need only favor slow decay of the spectrum.

[2] Bartlett et al. (2020) handled the Gaussian design / Gaussian response case in the first version of their work, and gave extensions to the sub-Gaussian case in a recent update.

5.1. Introducing a surrogate

Consider the surrogate predictor $\hat{\beta}_0$ corresponding to the minimum norm interpolating predictor for the training data *without* label noise. Mathematically, this surrogate is defined by $\hat{\beta}_0 = (X'X)^+X'X\beta$. Notice that $\hat{\beta}_0 = P(X)\beta$ where $P(X)$ is the projection onto the row-space of X.

The surrogate decomposition of the generalization error for $\hat{\beta}$ is given in the following lemma.

Lemma 5.3 (Surrogate decomposition of $\hat{\beta}$)**.**

$$\begin{aligned} &L_D(\hat{\beta}) - L_S(\hat{\beta}) \\ &\quad = (L_S(\hat{\beta}_0) - L_S(\hat{\beta})) + (L_D(\hat{\beta}) - L_D(\hat{\beta}_0)) \\ &\qquad + (L_D(\hat{\beta}_0) - L_S(\hat{\beta}_0)), \end{aligned}$$

with

$$\begin{aligned} L_S(\hat{\beta}_0) - L_S(\hat{\beta}) &= \frac{1}{n}\|Z\|^2, \\ L_D(\hat{\beta}) - L_D(\hat{\beta}_0) &= \mathrm{Tr}(X(X'X)^+\Sigma(X'X)^+X'ZZ'), \\ L_D(\hat{\beta}_0) - L_S(\hat{\beta}_0) &= \sigma^2 - \frac{\|Z\|^2}{n} + \beta' P(X)^\perp \Sigma P(X)^\perp \beta. \end{aligned}$$

The proof appears in Appendix B. Note that (in expectation) $(L_D(\hat{\beta}) - L_D(\hat{\beta}_0))$ and $(L_S(\hat{\beta}_0) - L_S(\hat{\beta})) + (L_D(\hat{\beta}_0) - L_S(\hat{\beta}_0))$ are exactly the terms which Bartlett et al. (2020) arrive at via the bias–variance decomposition and bound separately. They, however, made this decision using a problem-specific decomposition of the generalization error rather than arriving at the decomposition because it naturally arose from a choice of surrogate.

Lemma 5.4. *The sequence of implied surrogate hypothesis classes, $\{\hat{\beta}_0(S) : S \in \mathcal{S}^{(n)}\}_{n\in\mathbb{N}}$ is $(\mathcal{D}^{(n)}, n)$-SGC as long as $(\beta_n'\Sigma_n\beta_n)/\sqrt{n} \to 0$. Quantitatively, for a universal constant $C > 0$,*

$$\begin{aligned} &\mathbb{E} \sup_{(X_0,Y_0)\in\mathbb{R}^{n\times d}\times\mathbb{R}^n} \left| L_D(\hat{\beta}_0(X_0,Y_0)) - L_S(\hat{\beta}_0(X_0,Y_0)) \right| \\ &\qquad \le C\frac{\sigma^2 + \beta'\Sigma\beta}{\sqrt{n}}. \end{aligned}$$

The proof appears in Appendix B. Combining Lemma 5.4 with (Bartlett et al., 2020, Lemma 11) (which controls $L_D(\hat{\beta}) - L_D(\hat{\beta}_0)$), we get the following bound on the expected generalization error.

Theorem 5.5 (Expected risk bound for overparameterized linear regression)**.** *For some universal constant $C, c, b > 0$,*

$$\begin{aligned} \mathbb{E}L_D(\hat{\beta}) \le \sigma^2 &+ C\frac{\sigma^2 + \beta'\Sigma\beta}{\sqrt{n}} \\ &+ c\sigma^2\left(\frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)}\right) \end{aligned}$$

where b is the same constant appearing in the definition of k^* above. In particular, if $\{\Sigma_n\}_{n\in\mathbb{N}}$ is weakly benign and $(\beta_n'\Sigma_n\beta_n)/\sqrt{n} \to 0$ then $\mathbb{E}L_D(\hat{\beta}) \to \sigma^2$.

This result is similar to what one would obtain by converting the high-probability bound of Bartlett et al. (2020) into a bound in expectation. However, this result removes the dependence on $r_0(\Sigma_n)$ and replaces the dependence on $\|\beta_n\|^2\|\Sigma_n\|$ with $\beta_n'\Sigma_n\beta_n/\sqrt{n}$, hence weakening the conditions sufficient to ensure generalization. Indeed, the gap between the necessary and sufficient conditions is reduced to $\beta_n'\Sigma_n\beta_n/\sqrt{n}$. In addition, the apparent need to balance slow and fast decay of the spectrum of Σ_n highlighted by Bartlett et al. (2020) is not necessary.

More importantly, our approach highlights the role of uniform convergence in this problem. It is noteworthy that, when viewing the surrogate class as a type of one-step covering as discussed in the comments after Proposition 4.2, the approximation error component of the surrogate decomposition does not vanish in this case. Instead, it tends to σ^2 when the covariance matrices form a weakly benign sequence and $\beta_n'\Sigma_n\beta_n/\sqrt{n} \to 0$, but may have more erratic behavior otherwise.

6. Constructing Surrogates by Conditioning

In the overparameterized linear regression example, the surrogate obtained by training on "de-label-noised" data allowed us to construct meaningful generalization bounds for our learning problem via uniform bounds on the generalization error of the surrogate. For a generic learning problem, however, there may be no notion of label noise or such an approach may not prove useful. This leads us to seek natural constructions of surrogates in less structured problems.

One generic way to introduce such a surrogate is by conditioning. Let $\mathbb{P}^{\mathcal{F}}$ denote the conditional probability operator given a σ-field $\mathcal{F}$ (or a random variable), taking an event to its conditional probability. For a random variable ψ, let $\mathbb{P}^{\mathcal{F}}[\psi]$ denote the conditional distribution of ψ given $\mathcal{F}$.

Lemma 6.1 (Derandomization via conditioning)**.** *Let $\mathcal{F}$ be a σ-field on (some possible extension of) the underlying probability space upon which S and $\hat{h}$ are defined. Let $Q = \mathbb{P}^{\mathcal{F}}[\hat{h}]$. Then $\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)] = 0$.*

The following is then immediate by Lemmas 4.1 and 6.1.

Lemma 6.2 (Surrogate decomposition by conditioning)**.** *Let $\mathcal{F}$ and Q be as in Lemma 6.1. Then*

$$\begin{aligned} \mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h})] &= \mathbb{E}[L_S(Q) - L_S(\hat{h})] \\ &\quad + \mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)]. \end{aligned}$$

If $\hat{h}$ is a.s. interpolating (i.e., $\mathbb{E}[L_S(\hat{h})] = 0$), then

$$\mathbb{E}[L_{\mathcal{D}}(\hat{h})] = \mathbb{E}[L_S(Q)] + \mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)].$$

Every conditional distribution $Q = \mathbb{P}^{\mathcal{F}}[\hat{h}]$ represents a *derandomization* of $\hat{h}$: i.e., by the definition of conditioning, $\hat{h}$ has equal or greater dependence on the data S

than Q. There are other ways to achieve derandomization rather than conditioning. However, they may require one to obtain some explicit control on the risk difference, $\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)]$.

Informally, if $\hat{h}$ interpolates (or more generally overfits), we would expect a derandomized classifier to have excess empirical risk, yet lower generalization error.

Finally, it is important to understand how tautologies can arise from this perspective. If Q is a.s. nonrandom (corresponding, e.g., to conditioning on the trivial σ-algebra), then $Q = \mathbb{P}[\hat{h}]$ a.s., i.e., Q is the distribution of $\hat{h}$. In this case, $\mathbb{E}L_S(Q) = \mathbb{E}L_{\mathcal{D}}(Q) = \mathbb{E}L_{\mathcal{D}}(\hat{h})$, and we obtain the tautology

$$\begin{aligned}\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h})] &= \mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)] \\ &\quad + \mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)] \\ &\quad + \mathbb{E}[L_S(Q) - L_S(\hat{h})] \\ &= 0 + \mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h})] + 0.\end{aligned}$$

For this extreme example, Q belongs to the singleton class $\{\mathbb{P}[\hat{h}]\}$, which exhibits "uniform convergence" trivially. On the other end of the spectrum, if $Q = \delta_{\hat{h}}$, i.e., we condition on $\mathcal{F} = \sigma(S)$, then we get an equally tautological statement from the decomposition. The idea behind introducing the surrogate classifier Q is that it allows one to conceptually interpolate between these two tautological end points in order to find a (non-tautological) bound on the generalization error of a learning algorithm.

7. Hypercube Classifier

The following example is inspired by theoretical and empirical work by Nagarajan and Kolter (2019). Like in their work modeling SGD, we describe an example of a low-risk learned classifier, $\hat{h}$, such that there is no nonrandom class containing $\hat{h}$ almost surely for which one may establish a uniform and non-vacuous bound on generalization error. Using Lemma 6.2 and Proposition 4.2, we show that a derandomization of $\hat{h}$, obtained by conditioning on an explicit σ-field $\mathcal{F}$, yields a tight generalization bound based on uniform convergence of the surrogate.

In this section, we first construct the learning problem we will address. Second, we show that the structural Glivenko–Cantelli property fails on this example even though is has low generalization error. Lastly, we introduce our surrogate learning algorithm, show that it has similar empirical and test performance to the original algorithm, verify that the surrogate has the structural Glivenko–Cantelli property, and finally use this to establish a generalization bound for the original learning algorithms.

7.1. Construction

Let $d \gg 1$ index the dimension of the feature space and our sequence of learning problems, and let n_d be the sample size for the problem with index d. Let $\mathcal{X}^{(d)} = \{0,1\}^{2d}$ be the feature space and $\mathcal{Y} = \{0,1\}$ be the label space, and let $\mathcal{S}^{(d)} = \mathcal{X}^{(d)} \times \mathcal{Y}$. Let $f_d^* : \mathcal{X} \to \mathcal{Y}$ be given by

$$f_d^*(x) = \begin{cases} 1, & \|X\|_1 \le d, \\ 0, & \text{otherwise.} \end{cases}$$

For $x \in \mathcal{X}^{(d)}$, note that $f_d^*(x) = 1 - f_d^*(1-x)$. Let $\mathcal{D}^{(d)}$ be the distribution of $(X, f_d^*(X))$, where $X \sim \text{Unif}\left(\mathcal{X}^{(d)}\right)$. Let $S = (Z_1, \ldots, Z_{n_d}) \sim (\mathcal{D}^{(d)})^{n_d}$ where $Z_i = (X_i, f_d^*(X_i))$. Let $\hat{f}_d$ be the random element in $\{0,1\}^{2d} \to \{0,1\}$ given by

$$\hat{f}_d^{(S)}(x) = \begin{cases} 1 - f_d^*(x), & x \notin S \text{ and } 1-x \in S \\ f_d^*(x), & \text{otherwise} \end{cases}.$$

Let $\bar{Z}_i = (1 - X_i, 1 - Y_i)$ and $\bar{S} = (\bar{Z}_1, \ldots, \bar{Z}_n)$. We refer to pairs $(Z_i, \bar{Z}_i)$ as antipodes. Our learning algorithm, $\mathcal{A}_d : S \to \hat{f}_d^{(S)}$, only makes a classification error when a test point was not in the training set, but its antipode was. The loss function will be the $0-1$ loss, $\ell(f, (x,y)) = \mathbf{1}_{f(x) \neq y}$.

7.2. Failure of Uniform Convergence for this Problem

First, we note that at every problem size, d, the VC dimension of the collection of accessible decision rules is at least as large as the training dataset. We do not use this fact again, but it does highlight the apparent complexity of the learning problem.

Proposition 7.1 (VC theory not applicable)**.** *For $n_d \le 2^{2d-1}$, $\mathcal{H}^{(d)} = \{\hat{f}_d^{(S)} : S \in (\mathcal{S}^{(d)})^{n_d}\}$ has VC dimension at least n_d.*

Proof. For $n_d \le 2^{2d-1}$, any set of features $(X_1, \ldots X_{n_d})$ of size n_d with no antipodal points and no repeated points can be shattered by the subcollection of $\mathcal{H}^{(d)}$ given by $\left\{\hat{f}_d^{(S)} : S \in \prod_{i \in [n_d]} \{Z_i, \bar{Z}_i\}\right\}$. □

Notice this algorithm never makes an error on training data.

Lemma 7.2 ($\hat{f}_d$ is interpolating.)**.** $L_S(\hat{f}_d) = 0$ *a.s.*

Furthermore, by construction, the learning algorithm cannot return a classifier with high risk, no matter the training data observed, as long as $n_d \in o(2^{2d})$.

Lemma 7.3 ($\hat{f}_d$ has small risk)**.** $L_{\mathcal{D}^{(d)}}(h) \le n_d 2^{-2d}$ *for all $h \in \mathcal{H}^{(d)} = \{\hat{f}_d^{(S)} : S \in (\mathcal{S}^{(d)})^{n_d}\}$, and hence*

$$L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d) \le n_d 2^{-2d} \quad a.s.$$

The proof appears in Appendix A. The following result demonstrates that uniform convergence (of a class containing $\hat{h}$) *does not* explain the risk. The argument mirrors that of Nagarajan and Kolter (2019).

Theorem 7.4 ($\mathcal{H}^{(\cdot)}$ is not SGC)**.** *If* $n_d \in o(2^d)$ *then* $\left\{\ell \circ \mathcal{H}^{(d)}\right\}_{d\in\mathbb{N}}$ *is* not $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$*-SGC, in fact*

$$\mathbb{E} \sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| = 1 - O(n^2 2^{-2d}).$$

The proof appears in Appendix A. In this example, a generalization error bound was tractable because $L_{\mathcal{D}}(h)$ was readily bounded for all h, despite the fact that uniform convergence failed for $\mathcal{H}^{(\cdot)}$. One may then ask "how many bits of information do we need to forget about our training data in order for the sequence surrogate hypothesis class obtained by conditioning is structural Glivenko–Cantelli?"

7.3. Introducing a Surrogate Classifier

Let $k_d \leq 2d$. Let $\pi_{k_d} : \{0,1\}^{2d} \to \{0,1\}^{2d}$ satisfy

$$\pi_{k_d}(x_1, \ldots, x_{2d})_j = \begin{cases} 0, & j \leq k_d, \\ x_j, & \text{otherwise.} \end{cases}$$

That is, $\pi_{k_d}(x)$ zeros out the first k_d entries of x. Now, let $\pi_{k_d}(S) = (\pi_{k_d}(X_1), \ldots, \pi_{k_d}(X_{n_d}))$, $\mathcal{G}^{(d)} = \sigma(\pi_{k_d}(S))$, and put $Q(S) = \mathbb{P}^{\mathcal{G}^{(d)}}[\hat{f}_d^{(S)}]$. $Q(S)$ is a Gibbs classifier that is learned from the data, but is less coupled with the data than $\hat{f}_d^{(S)}$. Intuitively, conditioning our learned classifier on $\mathcal{G}^{(d)}$ can be interpreted as redrawing the first k features and the labels associated with the training data independently for each new test point, holding the last $2d - k$ features of each training point fixed. Since $Q(S)$ is $\sigma(\pi_{k_d}(S))$-measurable, then for distinct datasets S and S' with $\pi_{k_d}(S) = \pi_{k_d}(S')$, we have $Q(S) = Q(S')$.

The map $S \mapsto Q(S)$ is the *surrogate learning algorithm* and $Q(S)$ is the *surrogate classifier*. Note that in this example, our chosen surrogate learning algorithm returns a Gibbs classifier, while the original algorithm returns a deterministic classification rule. When the argument S of Q is omitted then it is assumed to be the training dataset, S.

We first evaluate the risk properties of our surrogate.

Lemma 7.5 (Risk and empirical risk of the surrogate Q)**.** *The following all hold almost surely*

$$L_S(Q) \leq \frac{2^{-k_d}(1 - 2^{-k_d})}{n_d} \times \left|\left\{(i,j) : X_i[k+1:2d] = \overline{X_j[k+1:2d]}\right\}\right|$$

$$L_{\overline{S}}(Q) \leq \frac{2^{-k}}{n_d} \left|\{(i,j) : X_i[k+1:2d] = X_j[k+1:2d]\}\right| \text{ and}$$

$$L_{\mathcal{D}}(Q) \leq n_d 2^{-2d}.$$

Furthermore,

$$\mathbb{E}L_S(Q) \leq (n_d - 1)2^{-2d}(1 - 2^{-k_d}),$$
$$\mathbb{E}L_{\overline{S}}(Q) \leq 2^{-k_d} + (n_d - 1)2^{-2d}, \text{ and}$$
$$\mathbb{E}L_{\mathcal{D}}(Q) \leq n_d 2^{-2d}.$$

The proof appears in Appendix A. As was foreshadowed in Section 4, we have increased the empirical risk by replacing $\hat{f}_d$ with Q. However, at the same time, we have dramatically lowered the empirical risk on the adversarial (antipodal) dataset, and not affected the true risk at all. In fact we are able to trade off empirical risk on the training data with worst case risk on an adversarial dataset explicitly by varying the parameter k_d. Even a small amount of re-randomization in the surrogate (k_d small) can yield very tight control on the adversarial empirical risk. In this example, the adversarial empirical risk decreases exponentially fast in the number of bits of information lost per example. This allows us to demonstrate that the sequence of surrogate hypothesis classes is structural Glivenko–Cantelli.

Lemma 7.6 (The surrogate is SGC)**.** *Consider the sequence of surrogate hypothesis classes given by* $\mathcal{I}^{(d)} = \{Q(S) : S \in \mathcal{S}^{(d)}\}$*. If* $n_d \in o(2^{k_d}/\sqrt{d})$ *then we have* $\{\ell \circ \mathcal{I}^{(d)}\}_{d\in\mathbb{N}}$ *is* $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$*-SGC. In particular, when* $n_d \in o(2^{2d})$ *and* $k_d \geq \lceil (1+\epsilon)\log_2(n_d) + \log_2(d)/2 \rceil$ *then we have* $\{\ell \circ \mathcal{I}^{(d)}\}_{d\in\mathbb{N}}$ *is* $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$*-SGC. If* $\log(n_d) \in \Omega(d)$ *this is only possible if* $k_d \in \Omega(d)$*. If* $\log(n_d) \in o(d)$ *then this is possible even when* $k_d \in o(d)$

The proof appears in Appendix A. Note that the restrictions upon k_d we provide may be a product of our particular approach to bounding the Rademacher complexity (Massart's Lemma). A more refined approach may yield looser restrictions on k_d. Since the surrogate behaves similarly on training and test data to the original learning algorithm, and since the sequence of classes of achievable surrogates is SGC, we can establish a generalization error bound for the original learning algorithm using the uniform convergence of the surrogate.

Theorem 7.7 (Bounding generalization error via an SGC surrogate)**.** *We have the following bound on the generalization error:*

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d)] \leq (n_d - 1)2^{-2d} + 2\sqrt{\log(2)}n_d^{1/2}((2d - k_d)n_d + 1)^{1/2}2^{-k_d}.$$

If $n_d \in o\left(2^{2d}\right)$*, then for any choice of* $\{k_d\}_{d\in\mathbb{N}}$ *with* $\lim_{d\to\infty}(k_d - \log_2(n_d) - \log_2(d)/2) = \infty$*, our surrogate witnesses that the generalization error vanishes*

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d)] \to 0.$$

The proof appears in Appendix A.

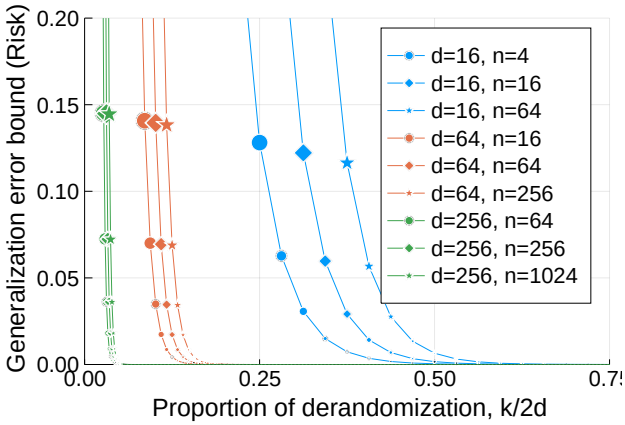

Figure 1: Visualizing the bound of Theorem 7.7.

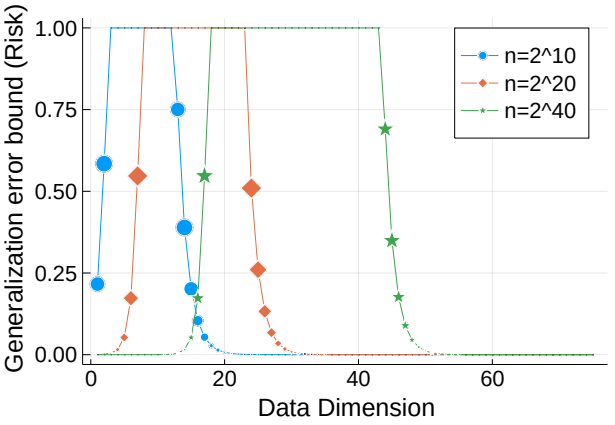

Figure 2: Double descent in the bound of Theorem 7.7.

7.4. Interpreting the Derandomization Bound

We visualize the dependence of our generalization error bound on k, n and d in Fig. 1. Notice that the bound decays very rapidly in the proportion of randomness removed by conditioning, $k/2d$. When the learning problem is sufficiently complex, even for larger sample size, a small proportion of derandomization leads to a strong control of the generalization error.

7.5. Relationship with Double Descent

The bound produced by derandomization also exhibits a form of double descent. The rising left half of each curve in Fig. 2 is a bound on the generalization error based on uniform convergence of a VC class containing the learned predictor, where the VC dimension is bounded by $\min(n, 2d-1)$. For $d \ll n$ this gives non-vacuous bounds for the low-dimensional setting. (Since the classifier is interpolating almost surely regardless of dimension or sample size, only the ascent of the first "descent" is present. A more complex example that models a learning phase would produce a first descent.) The second half of each curve in Fig. 2—based on the bound in Theorem 7.7—shows the second descent in the high dimensional setting. We see that, in sufficiently high dimensions, one may bound the generalization error of the learned classifier via uniform convergence of a suitable derandomized classifier. *The bound obtained via derandomization is non-vacuous in the region of the second descent, exactly where standard uniform convergence techniques based on VC theory would give vacuous bounds.*

Recalling that the weakly benign condition for overparameterized linear regression essentially requires that the spectrum of the feature covariance decays slowly enough, we can interpret that condition as requiring a sufficient degree of overparametrization for the risk to converge to the Bayes risk, showing that weakly benign overparameterized linear regression exhibits a form of double descent as well. As the surrogate for overparameterized linear regression and the surrogate constructed for this example both yield useful generalization bounds which exhibit double descent, we expect that there are useful connections between derandomization, surrogates and double descent to be explored in future research.

Acknowledgments The authors would like to thank Blair Bilodeau, Yasaman Mahdaviyeh, and Mufan Li for feedback on drafts of this work. JN is supported by a NSERC Vanier Canada Graduate Scholarship, and by the Vector Institute. DMR is supported in part by an NSERC Discovery Grant, an Ontario Early Researcher Award, and a stipend provided by the Charles Simonyi Endowment. This research was partially carried out while DMR and GKD were attending the Foundations of Deep Learning program at the Simons Institute for the Theory of Computing, and partially while JN, DR, and GKD were visiting the Institute for Advanced Study during the Special Year on Optimization, Statistics, and Theoretical Machine Learning. JN's travel to the Institute for Advanced Study was funded by a NSERC Michael Smith Foreign Study Supplement.

References

Advani, M. S. and Saxe, A. M. (2017). *High-dimensional dynamics of generalization error in neural networks*. arXiv: 1710.03667.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). "Spectrally-normalized margin bounds for neural networks". In: *Advances in Neural Information Processing Systems*, pp. 6241–6250.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). "Benign overfitting in linear regression". *Proc. National Academy of Sciences*.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). "Reconciling modern machine-learning practice and the clas-

sical bias–variance trade-off". *Proceedings of the National Academy of Sciences* 116.32, pp. 15849–15854.

Belkin, M., Ma, S., and Mandal, S. (2018). "To Understand Deep Learning We Need to Understand Kernel Learning". In: *International Conference on Machine Learning*.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media.

Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., and Wyart, M. (2019). "Jamming transition as a paradigm to understand the loss landscape of deep neural networks". *Physical Review E* 100.1, p. 012115.

Golowich, N., Rakhlin, A., and Shamir, O. (2018). "Size-Independent Sample Complexity of Neural Networks". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 297–299.

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

Kabán, A. and Durrant, R. J. (forthcoming). *J. Artificial Intelligence Research*.

Kuriki, S. and Numata, Y. (2010). "On formulas for moments of the Wishart distributions as weighted generating functions of matchings". *Discrete Mathematics & Theoretical Computer Science*.

Long, P. M. and Sedghi, H. (2020). "Size-free generalization bounds for convolutional neural networks". In: *International Conference on Learning Representations*.

Mei, S. and Montanari, A. (2019). *The generalization error of random features regression: Precise asymptotics and double descent curve*. arXiv: 1908.05355.

Mücke, N. and Steinwart, I. (2019). *Global Minima of DNNs: The Plenty Pantry*. arXiv: 1905.10686.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. Vol. 197. John Wiley & Sons.

Nagarajan, V. and Kolter, J. Z. (2019). "Uniform convergence may be unable to explain generalization in deep learning". In: *Advances in Neural Information Processing Systems (33)*. arXiv: 1902.04742.

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. (2017). *A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks*. arXiv: 1707.09564.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Wei, C. and Ma, T. (2019). "Data-dependent sample complexity of deep neural networks via lipschitz augmentation". In: *Advances in Neural Information Processing Systems*, pp. 9722–9733.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). "Understanding deep learning requires rethinking generalization". In: *International Conference on Representation Learning (ICLR)*. arXiv: 1611.03530v2 [cs.LG].