Dimension-Free Error Bounds from Random Projections

Ata Kabán

School of Computer Science
The University of Birmingham
Edgbaston, B15 2TT
Birmingham, UK
A.Kaban@cs.bham.ac.uk

Abstract

Learning from high dimensional data is challenging in general – however, often the data is not truly high dimensional in the sense that it may have some hidden low complexity geometry. We give new, user-friendly PAC-bounds that are able to take advantage of such benign geometry to reduce dimensional-dependence of error-guarantees in settings where such dependence is known to be essential in general. This is achieved by employing random projection as an analytic tool, and exploiting its structure-preserving compression ability. We introduce an auxiliary function class that operates on reduced dimensional inputs, and a new complexity term, as the distortion of the loss under random projections. The latter is a hypothesis-dependent data-complexity, whose analytic estimates turn out to recover various regularisation schemes in parametric models, and a notion of intrinsic dimension, as quantified by the Gaussian width of the input support in the case of the nearest neighbour rule. If there is benign geometry present, then the bounds become tighter, otherwise they recover the original dimension-dependent bounds.

Introduction

We consider learning settings where the generalisation error has a known essential dependence on the dimension of the input representation – examples include learning on unbounded input domains, learning with scale-insensitive loss functions, metric learning, and non-parametric methods.

Much of the statistical analysis of learning concerns how fast the empirical error converges to the true error as the sample size increases, and how large the sample must be to match the complexity of the model or hypothesis class of choice. However, in practice, especially in the above settings, the answers to these questions are often not sufficiently informative – for one cannot have access to unlimited sample sizes.

Instead, here we are mainly interested in the questions of what error-guarantee can be given for the available sample size in problems where the input dimension can be arbitrarily high, and what characteristics of the problem ensure good generalisation despite the sample size is small? A nice

example illustrating that analysis in small and large sample regimes may bring different insights is found in (Kontorovich and Weiss 2014).

We approach this problem from first principles. Although it is common practice, as well as technically convenient, to prime the analysis with a margin or some regularisation scheme in order to obtain dimension-free guarantees – where these geometric structures come from some form of prior knowledge, or experience – instead here we are interested in a generic principle that brings these in automatically when necessary.

To this end, we introduce a notion of compressibility that quantifies the distortion suffered by the loss when the inputs are subjected to a universal compression – that is a random projection. Random projection is often used for dimensionality reduction in algorithms, however in this context, the role of this compression will be purely analytic. In this role, it is somewhat analogous to the one-dimensional random projection of function outputs in Rademacher and Gaussian complexities, but here instead a random matrix acts on the inputs to the functions.

Based on these ideas we derive new generalisation bounds, which will depend on the complexity of an auxiliary low-dimensional hypothesis class instead the original one, plus a new complexity term that we call the data-complexity of the original function class. The latter captures and identifies benign geometric structures for the problem (of which 'margin' is a special case). We give the generic formalism first, which we then instantiate in concrete models.

Framework

Notations and Problem Setup

Let $\mathcal{X}_d \subseteq \mathbb{R}^d$ be an input domain, and $\mathcal{Y}$ the set of target values. In binary classification $\mathcal{Y} = \{-1, 1\}$ is the set of class labels, in regression $\mathcal{Y} \subseteq \mathbb{R}$. Let $\mathcal{H}_d$ be a function class (hypothesis class) with elements $h \in \mathcal{H}_d$ of the form $h : \mathcal{X}_d \to \mathcal{Y}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ be a bounded loss function. We are given a set of labelled examples $\mathcal{T}_N = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ drawn i.i.d. from some unknown distribution $\mathcal{D}_d$ over $\mathcal{X}_d \times \mathcal{Y}$. The learning problem is to use these to select a function from $\mathcal{H}_d$ with smallest generalisation error $E_{(x,y) \sim \mathcal{D}_d}[\ell(h(x), y)]$.

Let $\mathcal{G}_d = \ell \circ \mathcal{H}_d = \{g : (x,y) \in \mathcal{X}_d \times \mathcal{Y} \to \ell(h(x), y) : h \in \mathcal{H}_d\}$ denote the function class under study. Expectation w.r.t. the unknown data distribution $\mathcal{D}_d$ is denoted as $E[g] := E_{(x,y)\sim\mathcal{D}_d}[g(x,y)]$. Expectation w.r.t. the empirical measure defined by a sample $\mathcal{T}_N$ will is $\hat{E}_{\mathcal{T}_N}[g] = \frac{1}{N}\sum_{n=1}^N g(x_n, y_n)$ – where $\mathcal{D}_{\mathcal{T}_N} = \frac{1}{N}\sum_{n=1}^N \delta_{x_n}$, and δ_x is the probability distribution concentrated at the point x.

Definitions of Concepts

For the analysis that follows we make an auxiliary construction. Let $R \in \mathbb{R}^{k \times d}, k \leq d$ be a random matrix; a so-called random projection (RP) matrix. For instance a random matrix with i.i.d. Gaussian entries, or a Haar matrix is convenient, so R has full row rank almost surely (a.s.) and has low distortion property when used to reduce dimension (Dasgupta and Gupta 2003). As this is used in a purely analytic role, computationally fast variants are not required.

We apply R to the points in $\mathcal{X}_d$. For labelled points we use the convention $R(x,y) \equiv (Rx, y)$. This creates a random input space $R\mathcal{X}_d$ that is k-dimensional a.s., and which we denote as $\mathcal{X}_R$. When we refer to this as a k-dimensional input domain, then we use $\mathcal{X}_k$ instead.

On $\mathcal{X}_k$ (or $\mathcal{X}_R$), we define an auxiliary function class, denoted $\mathcal{G}_k = \ell \circ \mathcal{H}_k$ (or $\mathcal{G}_R = \ell \circ \mathcal{H}_R$) with elements $g_R = \ell \circ h_R$. This class may be chosen, and each choice gives rise to a different generalisation bound. Examples will be given later. A natural choice is for instance to have the same functional form as the elements of $\mathcal{G}_d$, but operating on k rather than d-dimensional inputs. An often more convenient choice is $\mathcal{G}_R := \mathcal{G}_d \circ R^T$.

Definition 1. *We define the following functional to represent the compressive distortion of a function $g \in \mathcal{G}_d$ relative to $g_R \in \mathcal{G}_R$:*

$$D_R(g, g_R) \equiv |g_R \circ R - g|$$

When $g_R = g \circ R^T$, we write instead $D_R(g) \equiv |g \circ R^T R - g|$.

We note that, conveniently, under suitable and fairly standard assumptions on the loss function, the compressive distortion can be bounded independently of the targets y. Examples in a later section will make this concrete.

Remark 2. *(i) $\exists k \leq d$ s.t. $D_R(g) = 0$; (ii) If $g(x,y) \in [0, \bar{\ell}], \forall(x,y)$ then $D_R(g) \in [0, \bar{\ell}], \forall k$.*

By Remark 2(i), it is always possible to choose $\mathcal{G}_R$ and k to have zero compressive distortion. In particular, the choice $\mathcal{G}_R = \mathcal{G}_d \circ R^T$ with $k = d$ will recover the traditional error analysis. In turn, the use of the above quantity captures a condition that allows us to reduce dimensional dependence in error bounds. To this end we define the following new notion of complexity that will play a key role in the sequel.

Definition 3. *We define the data-complexity of a function class $\mathcal{G}_d$ as the following:*

$$\mathcal{C}_{2N,k}(\mathcal{G}_d) = E_{\mathcal{T}_{2N}\sim\mathcal{D}_d^{2N}} \sup_{g\in\mathcal{G}_d} E_R \inf_{g_R\in\mathcal{G}_k} \hat{E}_{\mathcal{T}_{2N}}[D_R(g, g_R)]$$

We may think of this as the largest (w.r.t. $g \in \mathcal{G}_d$) 'mimicking error' on average (over training sets) of an ensemble of learners that each receive a randomly compressed version of the inputs and train to behave as g.

We note that this complexity term is always non-negative and, as already mentioned, it can be made zero by our choices of $\mathcal{G}_R$ and k. The interesting cases are those in which this quantity is small despite $k < d$.

Generic Bound

This section gives a novel and generic PAC-style uniform generalisation bound. It bounds the error of any function in a given class in terms of the notion of data-complexity of the class introduced in the previous section, which allows the use of a low complexity auxiliary function class that operates on low dimensional random projections of the inputs. The former is based on the distortion of the loss incurred by the random projection of the inputs – therefore, due to the structure-preserving property of random projections, this term may be low when the data exhibits some benign geometry with respect to the function class under study. This allows us to quantify and exploit benign geometry that may be present – that is, naturally occurring structures that make a learning problem less complex than it appears to be. Examples will follow in the subsequent sections.

In generic terms, we prove the following Theorem 4. Recall the empirical Rademacher complexity of a function class $\mathcal{G}$ is defined as $\hat{\mathcal{R}}_N(G) = \frac{1}{N}E_\sigma \sup_{g\in\mathcal{G}} \sum_{n=1}^N \sigma_n g(x_n)$, where $\sigma = (\sigma_1, \ldots, \sigma_N) \overset{\text{i.i.d}}{\sim}$ Uniform(± 1).

Theorem 4. *Let $\mathcal{G}_d$ be the function class associated with the class of functions $\mathcal{H}_d$, with a bounded loss function ℓ taking values in $[0, \bar{\ell}]$: $\mathcal{G}_d = \{g : (x,y) \in \mathcal{X}_d \times \mathcal{Y} \to \ell(h(x), y), \text{ s.t. } h \in \mathcal{H}_d\}$. Let $\mathcal{T}_N = \{(x_n, y_n)_{n=1}^N \sim \mathcal{D}_d^N\}$ be a training set over $\mathcal{X}_d \times \mathcal{Y}$. Then, for any $\delta > 0$, w.p. at least $1 - 2\delta$, uniformly for all $g \in \mathcal{G}_d$, we have:*

$$
\begin{aligned}
E[g] \quad \leq \quad & \hat{E}_{\mathcal{T}_N}[g] + 2\mathcal{C}_{2N,k}(\mathcal{G}_d) \\
+ \quad & 2E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] + 3\bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}}
\end{aligned}
$$

where R is a $k \times d, k \leq d$ random matrix independent of the sample, and $\mathcal{G}_R = \{x' \to g_R(x') \in \mathcal{Y} : x' \in \mathcal{X}_R\}$ is an auxiliary function class chosen before seeing the sample.

By Remark 2(i), for any choice of $\zeta \geq 0$ we can have the data-complexity term $\mathcal{C}_{2N,k}(\mathcal{G}_d) \leq \zeta$ for a suitable choice of k. In particular, $\zeta = 0$ is always satisfied by $k = d$, in which case we recover the classical Rademacher complexity based bound. However when the Rademacher complexity term is dimension dependent, then the data-complexity term may reduce this dependence by reducing dimension and exploiting the presence of naturally occurring structures through the structure preserving properties of random projections. In addition, even in the absence of favourable geometry, k and ζ allow us to control the tradeoff between bias and variance by *choosing $k \leq d$* according to the available sample size N at the expense of a bias ζ.

As we shall see in concrete examples in the next section, the analytic estimate of the data-complexity term takes different forms, depending on the form of the function class of

interest – in some of the parametric models it brings in a constraint on the margin distribution, or a finiteness constraint on the norms of various parameters akin to regularisation. In the nonparametric case it takes the form of a notion of intrinsic dimension.

Proof of Theorem 4. Let

$$\phi(\mathcal{T}_N) := \sup_{g \in \mathcal{G}_d} |E[g] - \hat{E}_{\mathcal{T}_N}[g]|.$$

By McDiarmid inequality, w.p. $1 - \delta$,

$$\phi(\mathcal{T}_N) \le E_{\mathcal{T}_N \sim \mathcal{D}_d^N}[\phi(\mathcal{T}_N)] + \bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}}.$$

Now, bounding this expectation, we have:

$$E_{\mathcal{T}_N \sim \mathcal{D}_d^N}[\phi(\mathcal{T}_N)] = E_{\mathcal{T}_N \sim \mathcal{D}_d^N} \sup_{g \in \mathcal{G}_d} \left| E[g] - \hat{E}_{\mathcal{T}_N}[g] \right|$$

$$= E_{\mathcal{T}_N \sim \mathcal{D}_d^N} \sup_{g \in \mathcal{G}_d} \left| E_{\mathcal{T}'_N \sim \mathcal{D}_d^N}[\hat{E}_{\mathcal{T}'_N}[g]] - \hat{E}_{\mathcal{T}_N}[g] \right|$$

$$\le E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} \left| \hat{E}_{\mathcal{T}'_N}[g] - \hat{E}_{\mathcal{T}_N}[g] \right| \quad \text{by Jensen ineq.}$$

$$\le E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} E_R \inf_{g_R \in \mathcal{G}_R} \left\{ \left| \hat{E}_{\mathcal{T}'_N}[g] - \hat{E}_{R\mathcal{T}'_N}[g_R] \right| \right.$$

$$+ \left| \hat{E}_{R\mathcal{T}'_N}[g_R] - \hat{E}_{R\mathcal{T}_N}[g_R] \right|$$

$$\left. + \left| \hat{E}_{R\mathcal{T}_N}[g_R] - \hat{E}_{\mathcal{T}_N}[g] \right| \right\}$$

$$\le E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} E_R \inf_{g_R \in \mathcal{G}_R} \left\{ \left| \hat{E}_{\mathcal{T}'_N}[g] - \hat{E}_{R\mathcal{T}'_N}[g_R] \right| \right.$$

$$\left. + \left| \hat{E}_{R\mathcal{T}_N}[g_R] - \hat{E}_{\mathcal{T}_N}[g] \right| \right\}$$

$$+ E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} E_R \sup_{g_R \in \mathcal{G}_R} \left| \hat{E}_{R\mathcal{T}'_N}[g_R] - \hat{E}_{R\mathcal{T}_N}[g_R] \right|$$

The first term can be bounded by triangle inequality:

$$E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} E_R \inf_{g_R \in \mathcal{G}_R} \left\{ \left| \hat{E}_{\mathcal{T}'_N}[g] - \hat{E}_{R\mathcal{T}'_N}[g_R] \right| \right.$$

$$\left. + \left| \hat{E}_{R\mathcal{T}_N}[g_R] - \hat{E}_{\mathcal{T}_N}[g] \right| \right\} \le$$

$$E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} E_R \inf_{g_R \in \mathcal{G}_R} \left\{ \right.$$

$$\frac{1}{N} \left[\sum_{n=1}^{N} |g(x'_n, y'_n) - g_R(Rx'_n, y'_n)| \right.$$

$$\left. \left. + \sum_{n=1}^{N} |g(x_n, y_n) - g_R(Rx_n, y_n)| \right] \right\}$$

$$= E_{\mathcal{T}_{2N} \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} E_R \inf_{g_R \in \mathcal{G}_R} \left\{ \right.$$

$$\frac{2}{2N} \left[\sum_{n=1}^{2N} |g(x_n, y_n) - g_R(Rx_n, y_n)| \right] \right\}$$

$$= 2\mathcal{C}_{2N,k}(\mathcal{G}_d)$$

To estimate the second term we observe that it is the supremum of the empirical process indexed by the reduced class $\mathcal{G}_R$, hence we can apply the classical steps of symmetrization and McDiarmid inequality. Let $\sigma \sim \text{Uniform}(\pm 1)^N$ be Rademacher variables. We have:

$$E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} E_R \sup_{g_R \in \mathcal{G}_R} \left| \hat{E}_{R\mathcal{T}'_N}[g_R] - \hat{E}_{R\mathcal{T}_N}[g_R] \right| =$$

$$E_R E_{\mathcal{T}_N, \mathcal{T}'_N, \sigma} \sup_{g_R} \left| \frac{1}{N} \sum_{n=1}^{N} \sigma_n[g_R(Rx_n, y_n) - g_R(Rx'_n, y'_n)] \right|$$

$$= 2 E_R E_{\mathcal{T}_N \sim \mathcal{D}_d^N, \sigma} \sup_{g_R \in \mathcal{G}_R} \left| \frac{1}{N} \sum_{n=1}^{N} \sigma_n g_R(Rx_n, y_n) \right|$$

$$= 2 E_R[\mathcal{R}_N(\mathcal{G}_R \circ R)]$$

$$=_{1-\delta} 2 E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R \circ R)] + 2\bar{\ell}\sqrt{\frac{\log(1/\delta)}{2N}}$$

Plugging back and noticing that $\mathcal{R}_N(\mathcal{G}_R \circ R) = \mathcal{R}_N(\mathcal{G}_R)$ completes the proof. $\qquad\square$

Examples

To instantiate Theorem 4 in concrete learning settings, we will need to estimate our newly introduced data-complexity term for the specific function classes. The following properties will be useful for doing so.

Remark 5. *The following simpler expressions upper bound the data-complexity:*

$$(i) \ \mathcal{C}_{2N,k}(\mathcal{G}_d) \le E_{\mathcal{T}_N \sim \mathcal{D}_d^N} \sup_{g \in \mathcal{G}_d} \hat{E}_{\mathcal{T}_N} E_R[D_R(g)] \equiv \mathcal{C}_k(\mathcal{G}_d)$$

$$(ii) \ \mathcal{C}_k(\mathcal{G}_d) \le \sup_{g \in \mathcal{G}_d} \sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} E_R[D_R(g)].$$

Proof. Relaxing the infimum, and using Jensen's inequality,

$$\mathcal{C}_{2N,k}(\mathcal{G}_d) = ...$$

$$E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} E_R \inf_{g_R \in \mathcal{G}_R} \hat{E}_{\mathcal{T}_N \bigcup \mathcal{T}'_N} |g_R \circ R - g| \le$$

$$E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{g \in \mathcal{G}_d} E_R \hat{E}_{\mathcal{T}_N \bigcup \mathcal{T}'_N} |g_R \circ R - g|, \forall g_R \in \mathcal{G}_R$$

$$\le \frac{1}{2} E_{\mathcal{T}_N \sim \mathcal{D}_d^N} \sup_{g \in \mathcal{G}_d} E_R \hat{E}_{\mathcal{T}_N} |g_R \circ R - g|$$

$$+ \frac{1}{2} E_{\mathcal{T}'_N \sim \mathcal{D}_d^N} \sup_{g \in \mathcal{G}_d} E_R \hat{E}_{\mathcal{T}'_N} |g_R \circ R - g|, \ \forall g_R \in \mathcal{G}_R$$

$$\le E_{\mathcal{T}_N \sim \mathcal{D}_d^N} \sup_{g \in \mathcal{G}_d} E_R \hat{E}_{\mathcal{T}_N} |g_R \circ R - g|, \ \forall g_R \in \mathcal{G}_R$$

Now choose $\mathcal{G}_R = \mathcal{G}_d \circ R^T$ to complete the proof. $\qquad\square$

Thresholded Linear Model with the 0-1 Loss

We start with the classical example of learning a halfspace, which allows us to demonstrate the working of our generic Theorem 4 on a simple example, and at the same time it gives us chance to fix an error in the existing literature.

Consider the linear function class in $\mathbb{R}^d$, $\mathcal{H}_d = \{x \to h^T x : h, x \in \mathbb{R}^d\}$ with the 0-1 loss, $\ell_{01} : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$, $\ell(\hat{y}, y) = \mathbb{1}(\hat{y}y \le 0)$, where $\mathbb{1}(\cdot)$ takes the value 1 if its argument is true, and 0 otherwise. Let $\mathcal{G}_d = \ell_{01} \circ \mathcal{H}_d$.

By a slight abuse of notation we will identify the hypothesis h with its parameter vector.

It is known that the generalisation error of this class, when working with the 0-1 loss directly and allowing $\mathcal{X}$ to be unbounded, is of order $\Theta(\sqrt{d}/\sqrt{N})$, and this dependence on d cannot be removed in general. However, deploying our Theorem 4 yields a condition of geometric nature under which we can prove the following dimension-free bound.

Theorem 6. *Let $\mathcal{G}_d = \ell_{01} \circ \mathcal{H}_d$ as above. Let $\mathcal{T}_N = \{(x_n, y_n)_{n=1}^N \sim \mathcal{D}_d^N\}$ be a training set over $\mathcal{X}_d \times \mathcal{Y}$. Let R be a $k \times d$ Gaussian random matrix, $k \leq d$. Suppose that for some $\zeta = \zeta(k) > 0$ we have for all $g \in \mathcal{G}_d$ and for all samples $\mathcal{T}_N \sim \mathcal{D}_d^N$ of size N that $\frac{1}{N} \sum_{n=1}^N \mathbb{1}\left\{ sign(h^T x_n) \neq sign(h^T R^T R x_n) \right\} \leq \zeta(k)$ w.p. $1 - \delta$ with respect to the random draw of R. Then, for any $\delta > 0$, w.p. $1 - 2\delta$ the following holds uniformly for all $g \in \mathcal{G}_d$:*

$$E[g] \leq \hat{E}[g] + 2\zeta(k)\mathbb{1}(k < d) + C\sqrt{\frac{k}{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}}$$

where $C > 0$ is an absolute constant.

The use of random projection (RP) to derive a dimension-free bound for this function class was previously attempted in (Garg, Har-Peled, and Roth 2002), but unfortunately an error in their proof[1] makes a meaningful comparison difficult. Nevertheless we believe the idea itself has potential. An alternative approach specialised to halfspace learning in pursued in (Kabán and Durrant 2017).

We note that both k and $\zeta(k)$ need to be chosen before seeing the sample. Interesting to notice that the role of k in the above bound replaces that of the VC dimension in classical bounds, which in this example would be d. A sensible choice is to set k proportional to N – which is typically known. The classical VC bound is recovered if we take the worst case complexity d to have $\zeta(k) = 0$, and demand sample size N to be proportional to it. However, one typically cannot have access to unlimited sample size N. Instead this new type of bound allows us to choose k proportional to the N we do have access to, and pay the price accordingly by $\zeta(k)$. Note however, that $\zeta(k)$ may still be small if there is benign geometry present – for instance if most points of the two classes are well separated.

The proof of Theorem 6 shows how the requirement for small data-complexity translates into an average margin distribution like condition in this case.

Proof of Theorem 6. We apply Theorem 4. Choosing $\mathcal{G}_R := \mathcal{G}_d \circ R^T$, by Remark 5 (i) we have:

$$\mathcal{C}_{2N,k}(\mathcal{G}_d) \leq E_{\mathcal{T}_N \sim \mathcal{D}_d^N} \sup_{g \in \mathcal{G}_d} E_R \hat{E}_{\mathcal{T}_N} |g \circ R^T R - g|$$

---

[1]In (Garg, Har-Peled, and Roth 2002), Lemma 3.5 correctly bounds the absolute difference of empirical errors from two independent samples S_1, S_2, in terms of their RP-ed counterpart, for a *fixed* classifier h. But the authors then attempt to apply Lemma 3.5, on page 8 after their eq.(3), with respect to a supremum over *all* hypotheses – unfortunately neglecting the effect of taking unions over events outside the subspace of the RP. The error carries over throughout the rest of the proof of their Theorem 3.1. and invalidates its main statement.

and plugging in the form of $\mathcal{G}_d$, we get:

$$\hat{E}_{\mathcal{T}_N} |g_R \circ R - g| \leq$$
$$\frac{1}{N} \sum_{n=1}^N \left| \mathbb{1}(h^T x_n \neq y_n) - \mathbb{1}(h^T R^T R x_n \neq y_n) \right|$$
$$= \frac{1}{N} \sum_{n=1}^N \mathbb{1}\left\{ \text{sign}(h^T x_n) \neq \text{sign}(h^T R^T R x_n) \right\}.$$

Hence, $\mathcal{C}_{2N,k}(\mathcal{G}_d) \leq \ldots$

$$E_{\mathcal{T}_N \sim \mathcal{D}_d^N} \sup_{h \in \mathcal{H}_d} \frac{1}{N} \sum_{n=1}^N \Pr_R \left\{ \text{sign}(h^T x_n) \neq \text{sign}(h^T R^T R x_n) \right\}$$
$$\leq \zeta(k)$$

It now remains to estimate the complexity of the function class in the reduced space, $\hat{\mathcal{R}}_N(\mathcal{H}_k)$. By Lemma 3.1 in (Mohri, Rostamizadeh, and Talwalkar 2012), and the known inequality between empirical Rademacher complexity and VC dimension for binary valued function classes (Bartlett and Mendelson 2002), we have $\hat{\mathcal{R}}_N(\mathcal{H}_k) \leq C\sqrt{\frac{V(\mathcal{H}_k)}{N}}$ where $C > 0$ is an absolute constant, and $V(\mathcal{H}_k) = k$ in this example. Hence the result follows from Theorem 4. $\square$

Generalised Linear Models

Consider the function class $\mathcal{G}_d = \ell \circ \mathcal{H}_d$ where $\mathcal{H}_d = \{x \rightarrow h^T x : x \in \mathcal{X}, h \in \mathbb{R}^d\}$ and $\ell : Y \times Y \rightarrow [0, \bar{\ell}]$ is a bounded loss function that is also L_ℓ-Lipschitz in its first argument. The input domain is not assumed to be bounded; it will be sufficient to require that $E_x[xx^T] < \infty$. Furthermore we do not impose any constraint a-priori on the parameter vector h – this will pop out of deploying our generic Theorem 4. Because of the unbounded input domain and the absence of a-priori constraints on h, the error is again known to be of order $\Theta(\sqrt{d}/\sqrt{N})$ in general.

By deploying our Theorem 4 we can prove following.

Theorem 7. *Let $\mathcal{G}_d$ be the class of generalised linear models of the form $\mathcal{G}_d = \ell \circ \mathcal{H}_d$, where $\mathcal{H}_d = \{x \rightarrow h^T x : h, x \in \mathbb{R}^d\}$, and the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, \bar{\ell}]$ is L_ℓ-Lipschitz in its first argument. Let $\mathcal{T}_N = \{(x_n, y_n)_{n=1}^N \sim \mathcal{D}_d^N\}$ be a training set over $\mathcal{X}_d \times \mathcal{Y}$, where $\mathcal{D}_d$ satisfies that $\Sigma \equiv E_{x \sim \mathcal{D}_d}[xx^T]$ is finite. Then, for any $k \leq d$ positive integer, and any $\delta > 0$, w.p. $1 - 2\delta$ we have uniformly for all $g \in \mathcal{G}_d$:*

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g]...$$
$$+ 2L_\ell \sqrt{\frac{2}{k}} E_x[\|x\|_2] \mathbb{1}(k < rank(\Sigma)) \cdot \sup_{h \in \mathcal{H}_d} \|h\|_2$$
$$+ 2\bar{\ell} C \sqrt{\frac{k}{N}} + + 3\bar{\ell} \sqrt{\frac{\log(2/\delta)}{2N}}$$

where C is an absolute constant.

We can interpret the various terms in the bound as follows. The term after the empirical error on the r.h.s. is an upper bound on the data-complexity term. It provides, as in the previous section, a condition under which a dimension-free bound holds. Specifically, in this example it tells us that

finiteness of $\|h\|_2$ is such a condition, and that small values of this norm represent a benign geometric structure that reduces generalisation error when N is limited. While this is no news, finding it from our generic theorem validates the principle behind it, namely that the distortion of the loss under a random projection of the inputs is able to capture a meaningful condition that explains what makes the learning problem easier despite the limited sample size. It will be interesting to follow this principle through in several other models too in the next couple of subsections.

It is also interesting to note similarities and differences of the obtained Theorem 7 with traditional Rademacher bounds. It is well known that, if $\mathcal{X}$ is bounded, then together with the norm constraint on $\|h\|_2$ derived above (or pre-imposed, as usual in the literature), one can have a dimension-free estimate of the empirical Rademacher complexity term. If we were to assume that $\mathcal{X}$ is bounded then our bound recovers exactly the Rademacher bound with the choice $k \geq \text{rank}(\Sigma)$ – that is, when the data-complexity term vanishes – and other choices of k have a disadvantage. On the other hand, boundedness of $\mathcal{X}$ is not required for Theorem 7 to hold, and as already discussed, the constraint on $\|h\|_2$ pops out from a generic procedure without the need for any prior knowledge.

Proof of Theorem 7. We can work with Gaussian R with i.i.d. 0-mean entries and variance $1/k$ so that on average any projected vector has the same squared length as the original. Choose the auxiliary function class be $\mathcal{G}_R = \mathcal{G}_d \circ R^T$, and we have for the data-complexity term: $\mathcal{C}_{2N,k}(\mathcal{G}_d) \leq \ldots$

$$L_\ell E_{\mathcal{T}_N, \mathcal{T}'_N \sim \mathcal{D}_d^{2N}} \sup_{h \in \mathcal{H}_d} E_R \frac{1}{2N} \sum_{n=1}^{2N} \left| h^T x_n - h^T R^T R x_n \right| \tag{1}$$

By Jensen inequality, and Lemma 2 from (Kabán 2014) we get (after some algebra):

$$E_R \left| h^T x - h^T R^T R x \right| \leq \left\{ E_R \left[\|h^T x - h^T R^T R x\|^2 \right] \right\}^{1/2}$$
$$\leq \sqrt{\frac{2}{k}} \|h\|_2 \|x\|_2$$

Having decoupled h and the data, we plug this back into eq. (1) to get the following when $k < \text{rank}(\mathcal{X}_d)$ (and 0 otherwise, by construction):

$$\mathcal{C}_{2N,k}(\mathcal{G}_d) \leq L_\ell \sqrt{\frac{2}{k}} E\|x\|_2 \sup_{h \in \mathcal{H}_d} \|h\|_2 \tag{2}$$

We note that the factor of $\sqrt{2}$ can be improved to 1 at the expense of a lengthier derivation if we chose to work with a scaled Haar distributed R.

Next, we estimate the Rademacher complexity term in the reduced space. Since there is no constraint on the parameters or the input domain, we instead exploit that the loss function is bounded through the following lemma (proof omitted).

Lemma 8. *Let $\mathcal{F}_k = \{x \to f(w^T x) \in [0,1] : x \in \mathbb{R}^k\}$. Then $\exists C > 0$ s.t. $\hat{\mathcal{R}}_N(\mathcal{F}_k) \leq C\sqrt{\frac{k}{N}}$.*

Using this, let $\mathcal{F}_R := \mathcal{G}_R/\bar{\ell}$, and we have:

$$\hat{\mathcal{R}}_N(\mathcal{G}_R) \leq \bar{\ell}\hat{\mathcal{R}}_N(\mathcal{G}_R/\bar{\ell}) \leq \bar{\ell}C\sqrt{\frac{k}{N}} \tag{3}$$

Putting together eqs. (3) and (2) completes the proof. $\quad\square$

Mahalanobis Metric Learning Classifiers

In this section we consider the problem of learning a generic classifier simultaneously with a linear transformation of the inputs – equivalent to learning a Mahalanobis metric that enhances classification performance. Here we will assume a bounded input space living in a ball of radius B for convenience, i.e. $\mathcal{X}_d \subseteq \mathcal{B}(0, B)$ as previous work by (Verma and Branson 2015) that studied the exact same problem.

The work of (Verma and Branson 2015) gave an error bound with dimensional dependence of order $\mathcal{O}(d\sqrt{\log(d)})$, and proved that this is not improvable in general. The authors then proposed a restriction on the Mahalanobis metrics, under which they show for the specific class of 2-layer perceptrons that the dimensional dependence of the risk upper bound reduces to $\mathcal{O}(\sqrt{\log(d)})$. Our main purpose here is to demonstrate how this regulariser comes out automatically from the data-complexity term when applying our generic Theorem 4, without the need to specify it by a-priori knowledge, and we may take advantage if it more widely to tighten the bound whenever there are weakly informative features.

Let $\mathcal{H}_d = \{x \to h(x) : x \in \mathcal{X}_d\}$ be some parametric L_h-Lipschitz function class with uniformly bounded fat shattering dimension over all scales[2]. Define $\mathcal{M}_d = \{M \in \mathbb{R}^{d \times d}, \sigma_{\max}(M) \leq 1\}$, where the condition $\sigma_{\max}(M) \leq 1$ (also assumed in (Verma and Branson 2015)) is not a restriction but serves to remove arbitrary scaling. The matrix $M^T M$ may be thought of as a metric tensor in the high dimensional space, although we work with M directly. As in the previous section, the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ is assumed to be bounded and L_ℓ-Lipschitz in its first argument. The function class of our interest is then $\mathcal{G}_d = \ell \circ \mathcal{H}_d \circ \mathcal{M}_d$.

Again we deploy Theorem 4, and now obtain the following.

Theorem 9. *Consider the class of functions of the form $\mathcal{G}_d = \ell \circ \mathcal{H}_d \circ \mathcal{M}_d$, where $\mathcal{M}_d = \{M \in \mathbb{R}^{d \times d}, \sigma_{\max}(M) \leq 1\}$, $\mathcal{H}_d = \{x \to h(x) : x \in \mathcal{X}_d, h$ is L_h-Lipschitz, $\text{fat}_\alpha(\mathcal{H}) \leq f_d\}$ for some $f_d > 0$, the loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \bar{\ell}]$ is L_ℓ-Lipschitz in its first argument, and $\mathcal{X} \subseteq \mathcal{B}(0, B)$. Let $\mathcal{T}_N = \{(x_n, y_n)_{n=1}^N \sim \mathcal{D}_d^N\}$ be a training set of size N taking values in $\mathcal{X}_d \times \mathcal{Y}$. Then, for any $k \leq d$ positive integer, and any $\eta > 0, \delta > 0$, w.p.*

---

[2]This condition can be dropped at the expense of a $\sqrt{\log(N)}$ factor in the error bound (by using a fixed scale α instead of Dudley's integral inequality in the proof, as it was done in Lemma 3 of (Verma and Branson 2015)). However, for function classes that are closed under scalar multiplication it is simply equivalent to having a bounded pseudo-dimension, cf. Theorem 11.14. in (Anthony and Bartlett 1999).

$1 - 2\delta$ *we have uniformly for all* $g \in \mathcal{G}_d$ *that:*

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + 2L_\ell L_h \mathbb{1}(k < rank(\mathcal{X}_d))$$

$$\cdot \left\{ \frac{\eta + \sqrt{2}}{\sqrt{k}} E\|x\|_2 \sup_{M \in \mathcal{M}} \|M\|_{Fro} + \exp\left(-\frac{\eta^2}{2}\right) \right\} +$$

$$\frac{12L_\ell}{\sqrt{N}} \left(\sqrt{kd \ln\left(1 + 2L_h B\sqrt{d}\right) + f_k \ln(4)} + 2\sqrt{\pi(kd + f_k)} \right)$$

$$+ 3\bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}}$$

The proof is rather lengthy, and deferred to the full version. The main steps are to choose $\mathcal{G}_R$ in a way to make sure that the scaling indeterminacy of the metric is taken care of in the auxiliary function class – if we work with Gaussian R, this involves a truncation step – and to estimate the Rademacher complexity in the reduced class. The latter involves Dudley inequality and covering number estimation (Mendelson 2003) on the matrix class that represents the metric in the reduced space.

The bound of Theorem 9 has the same high-level structure to what we have seen in the previous sections: On the r.h.s. we have the empirical error, the data-complexity term – which can only be nonzero for $k < d$ – and the complexity of the auxiliary function class – whose dependence on d is reduced to $\mathcal{O}(\sqrt{kd \log(\sqrt{d})})$ whenever $k < d$ is chosen. The choice $k = d$ (assuming $\mathcal{X}_d$ is not degenerate) recovers the bound of (Verma and Branson 2015) with two small improvements: in the log factor d is improved to $\sqrt{d}$, and a $\log(N)$ factor is eliminated. Note that the complexity of the high dimensional class $\mathcal{H}_d$ is not present in this bound, instead we have the complexity of its k-variate version, f_k, which is the fat shattering dimension of $\mathcal{H}_k$ instead. Note also that we did not specify the class $\mathcal{H}_d$ so we can plug in any function class if we have an estimate of its fat shattering dimension. Upon more specification the dependence on d may be further reducible.

In general, coming up with the extra constraints that reduce dimensional dependence is not a trivial task, and typically relies on some prior knowledge about the problem. By contrast, in our framework we do not require any a-priori knowledge of the specific problem, instead require a robustness to perturbations created by the random projection. This automatically yields an appropriate constraint – in the present example this is the magnitude of $\|M\|_{Fro}$ – that reflects robustness to distortions from random projection that our generic Theorem 4 rests on. Since $\sigma_{\max}(M) \leq 1$, this is at most $\sqrt{d}$ – but it can be much smaller if there are weakly informative features that are then down-weighted by the linear mapping being learned.

An equivalent formulation of Theorem 9 is to say that for some k and $\zeta(k)$ chosen before seeing the data, we require the precondition that, $\forall M \in \mathcal{M}_d, \|M\|_{Fro} \leq \zeta(k)$. On the r.h.s. then $\zeta(k)$ takes the place of $\sup_{M \in \mathcal{M}_d} \|M\|_{Fro}$. Of course, the supremum of the Frobenius norms of all matrices in $\mathcal{M}_d$ is not likely to be known, but we can use the technique of Structural Risk Minimisation (Vapnik 1998) to covert the bound into an algorithm that uses an estimate from the data. In this example, this would yield Frobenius-norm regularisation of the metric.

Finally, for the specific class of 2-layer perceptrons, it is natural to wonder whether we can recover a bound in $\mathcal{O}(\sqrt{\log(d)})$ as in (Verma and Branson 2015). It turns out that, by using from our main Theorem 4 with a choice of fixed metric in the auxiliary class $\mathcal{G}_R$ we can actually obtain a dimension-free bound:

Corollary 10. *Consider the class of functions of the form* $\mathcal{G}_d = \ell \circ \mathcal{H}_d \circ \mathcal{M}_d$, *where* $\mathcal{H}_d$ *has the following form. Let* $\phi : \mathbb{R} \to [-b, b]$ *be* L_ϕ-*Lipschitz, and* $\mathcal{H}_d = \{x \to \sum_{i=1}^m v_i \phi(w_i^T x) : x \in \mathcal{X}_d, \|w_i\|_1 \leq 1, \|v_i\|_1 \leq 1\}$. *Let* $\mathcal{T}_N$ *be a training set, as before. Then, for any* $k \leq d$ *positive integer, and any* $\delta > 0$, *w.p.* $1 - \delta$ *we have uniformly for all* $g \in \mathcal{G}_d$ *that:*

$$E[g] \leq \hat{E}_{\mathcal{T}_N}[g] + CL_\ell b \sqrt{\frac{k}{N}} + 3\bar{\ell}\sqrt{\frac{\log(2/\delta)}{2N}} +$$

$$\frac{2\sqrt{2}L_\ell L_\phi}{\sqrt{k}} E\|x\|_2 \mathbb{1}(k < rk(\mathcal{X}_d)) \sup_{v, W, M} \|v\|_2 \|W^T M\|_{Fro}$$

Nearest Neighbour

The previous sections concerned various linear and nonlinear parametric models. Here we take a simple representative of nonparametric models – a nearest neighbour classifier.

The nearest neighbour rule can be expressed as the following (Kontorovich and Weiss 2015). Denote by $\mathcal{T}_N^+, \mathcal{T}_N^- \subset \mathcal{T}_N, \mathcal{T}_N^+ \cup \mathcal{T}_N^- = \mathcal{T}_N$ the positively and negatively labelled training points respectively. Define the distance of a point $x \in \mathcal{X}$ to a set S as $d(x, S) = \inf_{z \in S}\{\|x - z\|\}$. Then $N^+(x) \equiv d(x, \mathcal{T}_N^+)$ and $N^-(x) \equiv d(x, \mathcal{T}_N^-)$ are the nearest positive and nearest negative neighbours of x respectively, and the label prediction for $x \in \mathcal{X}$ is given by the sign of the following 1-Lipschitz function:

$$h(x : \mathcal{T}_N^+, \mathcal{T}_N^-) := \frac{1}{2}(d(x, \mathcal{T}_N^-) - d(x, \mathcal{T}_N^+)) \quad (4)$$

$$= \frac{1}{2}\left(\|x - N^-(x)\| - \|x - N^+(x)\|\right)$$

We use Euclidean norms throughout.

To facilitate the analysis, we assume a bounded input domain, $\mathcal{X}_d \subseteq \mathcal{B}(0, B)$, same as in (Kontorovich and Weiss 2015) and the truncated γ-margin loss, which is $[0, 1]$-valued and $1/\gamma$-Lipschitz, where $\gamma \in (0, 1]$ – so that in the reduced space we can make use of existing estimates. The function class of our interest is therefore the composition of the $1/\gamma$-Lipschitz loss and the 1-Lipschitz classifier of the form given in eq. (4) – that is, $\mathcal{G}_d \subseteq \{x \to g(x) : x \in \mathcal{X}_d, g \text{ is } 1/\gamma\text{-Lipschitz}\}$.

By applying again our generic Theorem 4, we will obtain a bound where the data-complexity term turns out to be bounded by the Gaussian width of $\mathcal{X}_d$. For a set T the Gaussian width (Vershynin 2018; Liaw et al. 2017) is defined as:

$$w(T) = E_g \left[\sup_{x \in T} \{\langle g, x \rangle\} \right],$$

where $g \sim N(0, I_d)$. It is a measure of complexity of a set (see e.g. (Vershynin 2018), sec. 7.5 and references therein).

Hence we shall see from the below example that the name 'data-complexity' that we introduced in an early section is quite appropriate.

In this setting, Theorem 4 yields the following:

Theorem 11. *Let* $\mathcal{X}_d \subseteq \mathcal{B}(0, B)$, $\mathcal{Y} = \{-1, 1\}$, *and* $\mathcal{T}_N \sim \mathcal{D}^N$. *For any* $k \leq d$ *positive integer, for any* $\gamma > 0, \delta \in (0, 1)$, *w.p.* $1 - \delta$, *uniformly for all* $g \in \mathcal{G}_d$ *we have:*

$$
\begin{aligned}
E[g] &\leq \hat{E}_{\mathcal{T}_N}[g] + \frac{4c}{\gamma\sqrt{k}} w(\mathcal{X}_d) \mathbb{1}(k < d) \\
&+ C\frac{1}{\gamma} B N^{-\frac{1}{k+1}} + 3\sqrt{\frac{\log(2/\delta)}{2N}}
\end{aligned} \tag{5}
$$

where c and C are constants.

Eq. (5) holds for any positive integer k chosen before seeing the data. For instance, if we set it to make the data-complexity term below some $\eta \in (0, 1)$, this is:

$$
k \geq \frac{16cw^2(\mathcal{X}_d)}{\eta^2\gamma^2} \tag{6}
$$

Then replacing this choice of k into the bound of Theorem 11 resembles the flavour of the bounds obtained previously in doubling metric spaces in (Gottlieb, Kontorovich, and Krauthgamer 2016), with the squared Gaussian width taking the place of the doubling dimension. This is an interesting connection since there is a known link between the doubling dimension and the squared Gaussian width (Indyk 2007) – in an Euclidean metric space with algebraic dimension d they are both of order $\Theta(d)$, but are otherwise more general and can take fractional values. The Gaussian width is sensitive to structure embedded in Euclidean spaces, such as the existence of a sparse representation, smooth manifold structure, and so on.

Despite the above connection, there are differences. If the sample size is too small compared to intrinsic dimension of the input space, then we might opt for a lower value for k in the bound of Theorem 11 than that of eq.(6) at the expense of a larger bias. This is also practical since N is known while the Gaussian width may be unknown. As we see from the function class complexity term, the sample size needs to be exponential in k. Conversely, if N increases then k should be increased as well, in order to reduce or to eliminate (as k reaches d) the bias.

Another difference is in the methodological focus: In (Kontorovich and Weiss 2015; Gottlieb, Kontorovich, and Krauthgamer 2016), bounding the error in terms of a notion of intrinsic dimension was made possible due to a property of the Lipschitz class, by which the covering numbers of the function class are upper bounded in terms of the covering numbers of the input space. By contrast, in our strategy the starting point was to exploit random projection to obtain an auxiliary class with lower complexity, and as such, the Lipschitz property of the classifier functions is not in general required for our strategy to yield bounds in terms of the complexity of the input space. Indeed, we have seen throughout the various examples in this section that the same starting point has drawn together margin distribution and some widely used regularisation schemes in the case of parametric models, as well as the Gaussian width in the nearest neighbour example.

We note that in Theorem 11, the parameter γ needs to be chosen before seeing the data. Alternatively, if the Gaussiuan width is known, and noting that the constant C is specified in (Kontorovich and Weiss 2015), one can pursue SRM to tune the value of γ on the training set by minimising the bound.

Proof of Theorem 11. We take $R \in \mathbb{R}^{k \times d}$ a random projection matrix with 0-mean $1/k$-variance i.i.d. Gaussian entries and will use the notations $N_R^+(x)$ and $N_R^-(x)$ for the points whose image under a random projection is the nearest positive or nearest negative to Rx.

For $\mathcal{G}_R$ we choose a function class of the same form as $\mathcal{G}_d$, but acting on the compressed k-dimensional inputs instead.

The data-complexity term can be bounded as follows:

$$
C_{2N,k}(\mathcal{G}_d) \leq C_k(\mathcal{G}_d) \tag{7}
$$

$$
= E_{\mathcal{T}_N} \sup_{g \in \mathcal{G}_d} E_R \frac{1}{N} \sum_{n=1}^{N} |g_R(Rx_n, y_n) - g(x_n, y_n)|
$$

where

$$
\begin{aligned}
|g_R(Rx, y) - g(x, y)| &\leq \frac{1}{2\gamma} \Big| \|Rx - RN_R^-(x)\| - \|Rx - RN_R^+(x)\| \\
&\quad - \|x - N^-(x)\| + \|x - N^+(x)\| \Big| \\
&\leq \frac{1}{2\gamma} \Big(\big| \|Rx - RN_R^-(x)\| - \|x - N^-(x)\| \big| \\
&\quad + \big| \|Rx - RN_R^+(x)\| - \|x - N^+(x)\| \big| \Big)
\end{aligned} \tag{8}
$$

Note that $\|Rx - RN_R^{\pm}(x)\| \leq \|Rx - RN^{\pm}(x)\|$, and $\|x - N^{\pm}(x)\| \leq \|x - N_R^{\pm}(x)\|$, hence

$$
\begin{aligned}
\text{eq. (8)} \leq \frac{1}{2\gamma} \Big(&\max \big\{ \big| \|Rx - RN^-(x)\| - \|x - N^-(x)\| \big|, \\
&\big| \|Rx - RN_R^-(x)\| - \|x - N_R^-(x)\| \big| \big\} \\
&+ \max \big\{ \big| \|Rx - RN^+(x)\| - \|x - N^+(x)\| \big|, \\
&\big| \|Rx - RN_R^+(x)\| - \|x - N_R^+(x)\| \big| \big\} \Big)
\end{aligned}
$$

Therefore the supremum over $g \in \mathcal{G}_d$ amounts to a supremum over $N^+(x), N^-(x) \in \mathcal{X}_d$. So, eq. (7) $\leq \dots$

$$
\frac{1}{2\gamma} E_{\mathcal{T}_N} \sup_{N^-(x)} E_R \frac{1}{N} \sum_{n=1}^{N} \big| \|Rx_n - RN^-(x_n)\| - \|x_n - N^-(x_n)\| \big|
$$

$$
+ \frac{1}{2\gamma} E_{\mathcal{T}_N} \sup_{N^+(x)} E_R \frac{1}{N} \sum_{n=1}^{N} \big| \|Rx_n - RN^+(x_n)\| - \|x_n - N^+(x_n)\| \big|
$$

$$
\leq \frac{1}{2\gamma} \sup_{x, N^-(x) \in \mathcal{X}_d} E_R \big| \|Rx - RN^-(x)\| - \|x - N^-(x)\| \big|
$$

$$
+ \frac{1}{2\gamma} \sup_{x, N^+(x) \in \mathcal{X}_d} E_R \big| \|Rx - RN^+(x)\| - \|x - N^+(x)\| \big|
$$

$$
\leq \frac{1}{2\gamma} E_R \sup_{x, N^-(x) \in \mathcal{X}_d} \big| \|Rx - RN^-(x)\| - \|x - N^-(x)\| \big|
$$

$$
+ \frac{1}{2\gamma} E_R \sup_{x, N^+(x) \in \mathcal{X}_d} \big| \|Rx - RN^+(x)\| - \|x - N^+(x)\| \big|
$$

$$
\leq \frac{1}{\gamma} c E_R \sup_{x, x' \in \mathcal{X}_d} \big| \|Rx - Rx'\| - \|x - x'\| \big| \tag{9}
$$

$$
\lesssim \frac{1}{\gamma} w(\mathcal{X}_d - \mathcal{X}_d)/\sqrt{k} \tag{10}
$$

$$
= \frac{2c}{\gamma\sqrt{k}} w(\mathcal{X}_d) \tag{11}
$$

where $w(\mathcal{X}_d)$ is the Gaussian width of $\mathcal{X}_d$, and the last two steps follow from (Liaw et al. 2017).

For the function class complexity term we can use existing estimates. Note that $\mathcal{G}_R$ is a class of $1/\gamma$-Lipschitz functions, on k-dimensional inputs. The empirical Rademacher complexity[3] of the class of Lipschitz functions with a fixed Lipschitz constant has been derived in (Gottlieb, Kontorovich, and Krauthgamer 2016), which in our case takes the following form:

$$
\begin{aligned}
\hat{\mathcal{R}}_N(\mathcal{G}_R) &\leq \left[\frac{34(4\frac{1}{\gamma}\mathrm{diam}(R\mathcal{X}_d))^{k/2}}{N} \left(\frac{k-1}{2} \right) \right]^{\frac{2}{k+1}} \\
&\leq C\frac{1}{\gamma}\mathrm{diam}(R\mathcal{X}_d)N^{-\frac{1}{k+1}} \qquad (12)
\end{aligned}
$$

where C is an absolute constant.

To assemble the generalisation bound, we require the expected value $E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)]$. Since $E_R\|Rx\|^2 = \|x\|^2$, and by convexity of the supremum and the square root functions, Jensen's inequality yields that:

$$
E_R[\mathrm{diam}(R\mathcal{X}_d)] \leq 2B, \qquad (13)
$$

where the factor of 2 can be absorbed into the constant C above, and so

$$
E_R[\hat{\mathcal{R}}_N(\mathcal{G}_R)] \leq C\frac{1}{\gamma}BN^{-\frac{1}{k+1}} \qquad (14)
$$

Putting the pieces together completes the proof. $\qquad\square$

Conclusions

We presented an approach to reduce dimensional dependence of error bounds for learning settings where such dependence is known to be essential in general. This is achieved by the ability of random projections to take advantage of benign low complexity geometry – leveraged here in a purely analytic (rather than algorithmic) role. First we gave a generic uniform upper bound on the generalisation error in terms of the complexity of an auxiliary function class, and a new complexity term that quantifies benign geometry in a problem-dependent manner. We then instantiated this to parametric linear and nonlinear models, as well as a simple non-parametric model. If there is benign geometry present, then the bounds become tighter, otherwise they recover existing bounds. It is also possible in principle to use these results in conjunction with the classical technique of structural risk minimisation to convert them into regularised estimators. Future work will extend this framework to other learning settings, to find out what other geometric structures are benign for high dimensional learning.

---

[3]An alternative approach, pursued in (Gilad-Bachrach, Navot, and Tishby 2004; Gottlieb and Kontorovich 2014), would be to use fat-shattering dimension of the Lipschitz class, which gives a better convergence rate once the sample size exceeds a large threshold, but it is less tight in the small sample regime.

References

Anthony, M., and Bartlett, P. L. 1999. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

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

Dasgupta, S., and Gupta, A. 2003. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms* 22(1):60–65.

Garg, A.; Har-Peled, S.; and Roth, D. 2002. On generalization bounds, projection profile, and margin distribution. In *International Conference on Machine Learning (ICML)*, 171–178.

Gilad-Bachrach, R.; Navot, A.; and Tishby, N. 2004. Margin based feature selection: Theory and algorithms. In *International Conference on Machine Learning (ICML)*.

Gottlieb, L.-A., and Kontorovich, A. 2014. Efficient classification for metric data. *IEEE Trans. Inform. Theory* 60(9):5750–5759.

Gottlieb, L.-A.; Kontorovich, A.; and Krauthgamer, R. 2016. Adaptive metric dimensionality reduction. *Theoretical Computer Science* 620(21):105–118.

Indyk, P. 2007. Nearest-neighbor-preserving embeddings. *ACM Transactions on Algorithms* 3:3.

Kabán, A., and Durrant, R. J. 2017. Structure-aware error bounds for linear classification with the zero-one loss. arXiv preprint arXiv:1709.09782.

Kabán, A. 2014. New bounds on compressed linear least squares regression. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 448–456.

Kontorovich, A., and Weiss, R. 2014. Maximum margin multiclass nearest neighbors. In *Proceedings of the 31th International Conference on Machine Learning, ICML*, 892–900.

Kontorovich, A., and Weiss, R. 2015. *A Bayes consistent 1-NN classifier*. AISTATS.

Liaw, C.; Mehrabian, A.; Plan, Y.; and Vershynin, R. 2017. A simple tool for bounding the deviation of random matrices on geometric sets. *Geometric Aspects of Functional Analysis* 277–299.

Mendelson, S. 2003. A few notes on statistical learning theory. In Mendelson, S., and Smola, A. J., eds., *Advanced Lectures in Machine Learning, vol. 2600 of Lecture Notes in Computer Science*. Springer-Verlag, Berlin. 1–40.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of machine learning*. MIT Press.

Vapnik, V. N. 1998. *Statistical Learning Theory*. New York: Wiley-Interscience.

Verma, N., and Branson, K. 2015. Sample complexity of learning mahalanobis distance metrics. *Advances in Neural Information Processing Systems (NIPS)* 28:2584–2592.

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