Structure from Randomness in Halfspace Learning with the Zero-One Loss

Kaban, Ata; Durrant, Robert J.

Structure from Randomness in Halfspace Learning with the Zero-One Loss

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

Robert J. Durrant BobD@waikato.ac.nz
Department of Mathematics and Statistics, University of Waikato
Hamilton 3240, New Zealand

Abstract

We prove risk bounds for halfspace learning when the data dimensionality is allowed to be larger than the sample size, using a notion of compressibility by random projection. In particular, we give upper bounds for the empirical risk minimizer learned efficiently from randomly projected data, as well as uniform upper bounds in the full high-dimensional space. Our main findings are the following: i) In both settings, the obtained bounds are able to discover and take advantage of benign geometric structure, which turns out to depend on the cosine similarities between the classifier and points of the input space, and provide a new interpretation of margin distribution type arguments. ii) Furthermore our bounds allow us to draw new connections between several existing successful classification algorithms, and we also demonstrate that our theory is predictive of empirically observed performance in numerical simulations and experiments. iii) Taken together, these results suggest that the study of compressive learning can improve our understanding of which benign structural traits – if they are possessed by the data generator – make it easier to learn an effective classifier from a sample.

1. Introduction

Given a 'hypothesis class' of functions, $\mathcal{H}$, and a training set of N observations $\mathcal{T}^N = \{(x_n, y_n) : (x_n, y_n) \overset{\text{i.i.d}}{\sim} \mathcal{D}\}_{n=1}^N$, where $\mathcal{D}$ is an unknown distribution over $\mathcal{X} \times \mathcal{Y}, \mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \{-1, 1\}$, the goal in binary classification is to use the training data to learn a function (classifier) $\hat{h} \in \mathcal{H}$ such that, with respect to some specified loss function ℓ, its generalization error (or risk):

$$\mathrm{E}[\ell \circ \hat{h}] := \mathrm{E}_{(x,y)\sim\mathcal{D}}[\ell(\hat{h}(x), y)|\mathcal{T}^N] \tag{1}$$

is as small as possible. For binary classification the function $\ell : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}, \ell(\hat{h}(x), y) = \mathbf{1}(\hat{h}(x) \neq y)$, called the zero-one loss, is the main error measure of interest (Nguyen & Sanner, 2013). Here $\mathbf{1}(\cdot)$ denotes the indicator function which returns one if its argument is true and zero otherwise. The optimal classifier in $\mathcal{H}$ is denoted by $h^* := \arg\min_{h\in\mathcal{H}} \mathrm{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)]$.

In this work we consider functions of the form $\mathcal{H} := \{x \to \mathrm{sign}\,(h^T x) : h \in \mathbb{R}^d, x \in \mathcal{X}\}$, that is, $\mathcal{H}$ is identified with the set of normals to hyperplanes which, without loss of generality (since otherwise we can always concatenate a 1 to all inputs and work in $\mathbb{R}^{d+1}$ instead of $\mathbb{R}^d$), pass through the origin.

As $\mathcal{D}$ is unknown, one cannot minimize the generalization error directly. Instead, we have access to the empirical error over the training set – the minimizer of which is the Empirical Risk Minimizer or ERM classifier, $\hat{h} := \arg\min_{h \in \mathcal{H}} \frac{1}{N} \sum_{n=1}^{N} \ell(h(x_n), y_n)$.

Let $R \in \mathbb{R}^{k \times d}$, $k \leq d$ be an instance of a random matrix either (1) with i.i.d zero-mean Gaussian or subgaussian entries, or (2) such that, with probability 1, the columns of R^T form an orthonormal basis for some k-dimensional subspace of $\mathbb{R}^d$ selected uniformly from all such possible bases. Such matrices are commonly called Random Projection (RP) matrices – we will call matrices of type (1) either Gaussian or subgaussian RP matrices respectively, and those of type (2) Haar RP matrices[1]. A convenient way to think about R is as a means of compressing or 'sketching' the training data, and to carry out the compression we simply sample a single random RP matrix R of a particular kind and then left multiply the training observations with it. The celebrated Johnson-Lindenstrauss lemma and its variants guarantee that with high probability, such sketching has low distortion under mild conditions on the projection dimension, k (Johnson & Lindenstrauss, 1984; Achlioptas, 2003; Matoušek, 2008).

Now let $\mathcal{T}_R^N = \{(Rx_n, y_n)\}_{n=1}^N$ denote the RP of the training set, so the input points Rx_n are now k-dimensional. The hypothesis class defined on such k-dimensional inputs will be denoted by $\mathcal{H}_R := \{Rx \to \text{sign}\,(h_R^T Rx + b) : h_R \in \mathbb{R}^k, b \in \mathbb{R}, x \in \mathcal{X}\}$. Other analogous notations will be used in the k-dimensional space: $h_R^* = \arg\min_{h_R \in \mathcal{H}_R} \mathrm{E}[\ell \circ h_R]$ denotes the optimal classifier in $\mathcal{H}_R$, and $\hat{h}_R = \arg\min_{h_R \in \mathcal{H}_R} \frac{1}{N} \sum_{n=1}^{N} \ell(h_R(Rx_n), y_n)$ is the ERM in $\mathcal{H}_R$. For any particular instance of R the learned ERM classifier in the corresponding k-dimensional subspace $\hat{h}_R$ is possibly not through the origin, but any non-zero translation b will not affect our proof technique.

The generalization error of $\hat{h}_R$ is the following random variable that depends on both $\mathcal{T}^N$ and R:

$$\mathrm{E}[\ell \circ \hat{h}_R] := \mathrm{E}_{(x,y)\sim\mathcal{D}} \left[\ell(\hat{h}_R(Rx), y) | \mathcal{T}^N, R \right] \tag{2}$$

The remainder of this paper is concerned with the following two questions:

1. In terms of classification error, if we work with an RP sketch of the data how much, if anything, does it cost us? We made a start on this line of enquiry in preliminary work (Durrant & Kabán, 2013), which here we improve, simplify, and expand in Section 2.

2. What do the insights gained about learning from randomly projected data tell us about the original uncompressed learning problem, in terms of characteristics that make one instance of the problem easier than another? This is the subject of Section 3.

Although these two questions serve rather different goals, we propose to use similar techniques to approach both of them. The first of these questions resembles a model selection problem, namely the trade-off between loss minimization and model complexity. Furthermore it is also of more recent interest from the viewpoint of managing computational (time

1. We note that Haar RP matrices can be conveniently constructed by first generating an instance of a Gaussian RP matrix R and then either left multiplying it with $(RR^T)^{-1/2}$ so that $R' := (RR^T)^{-1/2}R$ is a Haar RP matrix, or else by carrying out Gram-Schmidt orthogonalization on the rows of R.

and space) complexity in high dimensional data analysis. Meanwhile the second of these questions targets another fundamental open problem concerning what properties make data easier or harder to learn from? Although both questions are difficult to approach in full generality, our findings here indicate that answers to the first question can facilitate answers to the second.

1.1 Related Work and Motivation

The use of RP in machine learning has a long history, most commonly as an efficient means to reduce computational and storage demand. Amongst the earliest works, Arriaga and Vempala (1999) gave generalization error bounds for the linear perceptron on randomly projected data, assuming that the class-conditional supports are separable by a hard margin. Their proof techniques combined the Johnson-Lindenstrauss lemma with union bounds and in order to control the generalization error following random projection, a target dimension that grows with the log of the sample size was required. In a different vein, Bălcan, Blum, and Vempala (2006) used RP to give an alternative interpretation for the kernel trick; they also assume *a priori* that the data classes are separated by a soft margin, i.e that the region between the classes has a low density with respect to both classes.

For learning in the original data space, Garg, Har-Peled, and Roth (2002) considered RP as an analytic tool, with similar aims to our Section 3. Unfortunately an error in the proof in that paper makes the findings incomplete and inconclusive. In Theorem 6 of Kabán (2019) we presented a corrected proof of the main result of Garg et al. (2002), however we were only able to reproduce its conclusion under an additional strong assumption that needs to hold with probability 1 for every training set of a given size. Besides which, even the originally stated generalization bound in (Garg et al., 2002) is generally trivial i.e. is greater than 1. In a follow-on experimental paper (Garg & Roth, 2003), the authors proposed an algorithm which learns a classifier by minimizing a heuristic simplification of the bound in (Garg et al., 2002) since the original was too loose to be of practical use. However, no theoretical guarantees have been given for the resulting algorithm.

Several years later, some spectacular advances in the area of compressed sensing (CS) revived interest in RP for machine learning, and the term 'compressive learning' was coined (Calderbank, Jafarpour, & Schapire, 2009). The works of Calderbank et al. (2009) and Fard, Grinberg, Pineau, and Precup (2012) gave guarantees for compressive learning of SVM and ordinary least square regression respectively, under the assumption that the data have a sparse representation. This assumption gave them access to tools from CS, which get around some of the problems seen in early works (Arriaga & Vempala, 2006). However, the guarantees obtained in this way cease to hold when the data do not admit a sparse representation in some basis. Indeed, we have no reason to believe that a sparse representation of the data is necessary for compressive learning to succeed. In fact, several works have studied compressive learning in specific generative parametric families that do not require a sparse representation of the data. In particular, the compressive Fisher's Linear Discriminant classifier has been analyzed (Durrant & Kabán, 2010), and more recently non-linear compressive classifiers have also been studied (Reboredo, Renna, Calderbank, & Rodrigues, 2016) under the assumption of Gaussian classes. Moreover, the sparse representation assumption also turned out to be unnecessary for compressive ordinary least squares

regression (Kabán, 2014), as it transpires that one can exploit the structure of the problem to obtain informative bounds without the need to retain assumptions from compressed sensing.

In 2013 we began a new line of inquiry on compressive classification (Durrant & Kabán, 2013), and derived the exact probability of label flipping under Gaussian RP, which we used to give tight upper bounds on the generalization error of a randomly-projected linear ERM classifier without any distributional assumptions on the data generator, and no margin or sparse representation requirements. Those results are the starting point of the present paper in the next section 2.

This paper develops a more complete analysis of linear classification in the high-dimensional, small sample size setting, with new results addressing both the compressive, and the original uncompressed problems, which we treat as two sides of the same coin. A large part of this paper is devoted to discussing the implications of this view. Our focus on linear models stems from two motivations: analytical tractability, and fundamental importance. Indeed, linear models have been a central object of study in statistical learning at all times (Vapnik, 1998; Kawaguchi, Kaelbling, & Bengio, 2019), as well as in compressed sensing (Donoho, 2006), and better understanding of such models has laid some of the main foundations for successful machine prediction.

It is well known from classical VC theory that, for linear classification in the agnostic setting, in the absence of any assumptions other than i.i.d. sampling of the data generator, the difference between the generalization error and the empirical error of a linear function class is of the order $\Theta(\sqrt{d/N})$ (Bartlett & Mendelson, 2002; Devroye & Lugosi, 1995), where d is the dimensionality of the feature representation and N is the sample size, so for meaningful guarantees – when we are agnostic about the properties of the data generator and its domain – we need the sample size N be of order d. However, often in practice we are faced with settings where $d > N$, and in this setting the VC bounds are clearly vacuous. Many approaches have been proposed to obtain non-vacuous bounds, including sparsity priors, low rank assumptions, margin maximization, and others – all of which assume, in some form, that the problem has a known simpler structure, which is specified beforehand. Our goal here is to obtain generalization bounds of a similar flavour, but without the need to pre-specify the form of structure expected to be present. We do this by exploiting the structure preserving ability of RPs, which is universal in the sense that it is oblivious to the data, so we can capture a range of structures without knowing what they are, by a notion of robustness to the perturbations created by RP.

Finally, we should mention the relevance of some very recent and ongoing debate about the puzzle of overparameterized models, in which linear models ($d > N$) are also being studied (Bartlett, Long, Lugosi, & Tsigler, 2019; Nagarajan & Kolter, 2019). A key result of Nagarajan and Kolter (2019) shows that in such settings, generalization bounds based on uniform convergence remain vacuous even if the biases of the optimizer are fully exploited. Other new findings (Negrea, Dziugaite, & Roy, 2019) indicate that it may still be possible to explain generalization via uniform convergence provided a suitable 'surrogate learning algorithm' may be constructed. In that context, our use of RP based compression may be viewed as a means by which to construct such a surrogate learning algorithm[2]. At a high

2. We thank an anonymous referee for pointing out this connection.

level the core idea here is that, despite $d \gg N$, a particular (representation of a) learning problem may have a simple, compressible, structure and for the compressed representation of the problem uniform convergence may indeed yield informative generalization error bounds for the original uncompressed problem.

2. Risk Bounds for the Compressive ERM Classifier

In this section we consider learning by Empirical Risk Minimization (ERM) from data that is available only in randomly projected form, such as that captured by compressive sensing devices, or because e.g. practical considerations make the original high-dimensional data too onerous to work with. This section also serves to introduce the necessary foundations that will be built upon in the subsequent section 3. Furthermore, Section 2.1.1 will shed some light on the effect of sparse representation of the data and other structures that potentially make learning 'easy'.

We start by extending our previous result (Durrant & Kabán, 2013) – which originally concerned only Gaussian random projection – to allow sub-Gaussian families of RP matrices, while we also simplify our arguments.

Let R be a RP matrix of size $k \times d$, with $k \leq d$. We shall denote by θ_u^h the angle between two vectors $h, u \in \mathbb{R}^d$, and let $f_k^+(\theta_u^h) := f_k(\theta_u^h) \cdot \mathbf{1}(h^T u > 0)$, where $f_k(\theta_u^h) = \Pr_R\left\{h^T R^T R u \leq 0\right\}$. That is, if we now identify h with the normal to some classifying hyperplane in $\mathbb{R}^d$ and likewise identify $u := xy$ for some observation $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ then $f_k^+(\theta_{xy}^h)$ signifies how likely it is that a point x flips from the correct side of the decision boundary of classifier h to the wrong side after RP (or vice-versa). This quantity will be our basic tool for telling us what happens to correct label predictions after a RP.

With the notations and definitions just introduced, we have the following result which holds with any zero-mean subgaussian matrix R:

Theorem 2.1. *For any $\delta \in (0, 1)$, the following holds for the compressive linear ERM classifier $\hat{h}_R$ with probability $1 - 2\delta$:*

$$\begin{aligned} Pr_{x,y}[(\hat{h}_R^T R x + b)y \leq 0] \quad &\leq \quad Pr_{x,y}[h^{*T} xy \leq 0] + c\sqrt{\frac{k + 1 + \log(1/\delta)}{N}} ... \\ &+ \quad E_{x,y}[f_k^+(\theta_{xy}^{h^*})] + \min\left\{\frac{1-\delta}{\delta} \cdot E_{x,y}[f_k^+(\theta_{xy}^{h^*})], \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right\} \end{aligned}$$

where $c > 0$ is a universal constant, independent of d and N.

We should observe the main features of the bound. The VC complexity term of $\mathcal{O}(\sqrt{d/N})$ is reduced to $\mathcal{O}(\sqrt{k/N})$ at the price of terms that represent the error incurred by working in a random subspace projection of the data.

We have so far not specified the distribution of R, and we shall see that the proof does not require this specification as long as $f_k^+(\theta)$ can be controlled. Conceptually it is useful to think about R as a matrix that orthogonally projects the data into a lower dimensional, uniformly randomly oriented subspace $\mathbb{R}^k$. As k grows towards d, then the probability of flipping across the boundary decreases to zero. Hence $f_k^+(\theta)$ decreases towards zero so the last two terms on the r.h.s. vanish – consequently, Theorem 2.1 approaches the classical VC bound for the original d-dimensional ERM classifier. Moreover, suppose the

data distribution is supported in a fixed but arbitrary subspace $\mathbb{R}^m \subseteq \mathbb{R}^d$, $m < d$ – then the probability of flipping will be zero provided $\mathbb{R}^m \subseteq \mathbb{R}^k$, so again increasing k to m will reduce the last two terms on r.h.s towards zero.

Proof of Theorem 2.1. For a fixed instance of R, we apply a classical VC bound (Bartlett & Mendelson, 2002, Theorem 1): For any $\delta \in (0,1)$ w.p. $1-\delta$ over the random draws of the training set, the following holds uniformly for all $h_R \in \mathbb{R}^k$:

$$\Pr_{x,y}[(h_R^T Rx + b)y \leq 0] \leq \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}((h_R^T Rx_n + b)y_n \leq 0) + c\sqrt{\frac{k+1+\log(1/\delta)}{N}} \tag{3}$$

for some absolute constant $c > 0$.

This implies that the ERM classifier $\hat{h}_R$ satisfies the following:

$$\Pr_{x,y}[(\hat{h}_R^T Rx + b)y \leq 0] \leq \Pr_{x,y}[(h_R^{*T} Rx + b^*)y \leq 0] + 2c\sqrt{\frac{k+1+\log(1/\delta)}{N}} \tag{4}$$

and one can absorb the factor 2 into c.

Since (h_R^*, b^*) is optimal in $\mathbb{R}^{k+1}$, its error is upper bounded by the error of the optimal homogeneous classifier, which we denote by $h_R^{*'} \in \mathbb{R}^k$:

$$\begin{aligned}
\Pr_{x,y}((h_R^{*T} Rx + b^*)y \leq 0) &\leq \Pr_{x,y}(h_R^{*'T} Rxy \leq 0) \\
&\leq \Pr_{x,y}(h^{*T} R^T Rxy \leq 0) \\
&\leq \underbrace{\mathrm{E}_{x,y}[\mathbf{1}(h^{*T} R^T Rxy \leq 0) \cdot \mathbf{1}(h^{*T} xy > 0)]}_{T} + \Pr_{x,y}(h^{*T} xy \leq 0)
\end{aligned}$$

where we used the fact that both $h_R^{*'} \in \mathbb{R}^k$ and $Rh^* \in \mathbb{R}^k$ are elements of the same k-dimensional linear subspace $R(\mathbb{R}^d) \equiv \mathbb{R}^k$.

We then bound T from its expectation using a combination of Höffding and Markov inequalities. Indeed, since $T \in [0,1]$, by the Höffding bound it holds for any $\epsilon > 0$ that:

$$\Pr\{T \geqslant \mathrm{E}_R[T] + \epsilon\} \leqslant \exp\left(-2\epsilon^2\right) \tag{5}$$

which implies w.p at least $1-\delta$:

$$T \leqslant \mathrm{E}_R[T] + \sqrt{\frac{1}{2}\log\frac{1}{\delta}} \tag{6}$$

Noting that T is a positive random variable, smaller values should imply smaller deviations. This is not reflected by eq. (6), hence we combine this with a Markov inequality, which is tighter for small $\mathrm{E}_R[T]$ – while the Höffding bound is tighter for small values of δ. From Markov's inequality we have:

$$T \leqslant \mathrm{E}_R[T] \cdot \frac{1-\delta}{\delta} \tag{7}$$

Taking the minimum of these two bounds in eqs. (6) and (7) completes the proof. □

We now discuss the choice of the distribution of R. If R has i.i.d. Gaussian entries, then it is rotationally invariant, and based on this property the exact expression of $f_k^+(\theta)$ has been computed (Durrant & Kabán, 2013). When $d \gg k$, such Gaussian RP is also known to be a good approximation of the Haar RP, i.e. a uniformly randomly oriented orthogonal projection, thanks to the concentration of measure. However, for implementing compressive classifiers, certain sub-Gaussian random projection matrices R can be computationally cheaper (Achlioptas, 2003) and hence may be preferable in practice. Rotational invariance then no longer holds exactly, but it still holds approximately (Vershynin, 2018), and Lemma 2.2 below will exploit this to establish the applicability of Theorem 2.1 to the whole family of random matrices with i.i.d. sub-Gaussian entries by controlling $f_k^+(\theta)$.

Lemma 2.2 gives the relevant expressions needed for computing or upper bounding the term $f_k^+(\theta)$ needed in the bound of Theorem 2.1. Parts (a) and (b) both follow from our earlier work (Durrant & Kabán, 2013; Kabán, 2015, Corollary 3.1.) however, for completeness we include new proofs in Appendix A, as we found simple ways to show how these expressions arise.

Lemma 2.2 (Flipping probability)**.** *Let $h, x \in \mathbb{R}^d$, and let $\theta = \theta_x^h \in [0, \pi)$ be the angle between them. Let R be a RP matrix, and let $Rh, Rx \in \mathbb{R}^k$ be the images of h, x under R. Then the following hold:*

(a) If R has entries $r_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$, we have for $h^T x \neq 0$,

$$f_k(\theta) := Pr\big\{(Rh)^T Rx \leq 0\big\} \quad = \quad \frac{\Gamma(k)}{(\Gamma(k/2))^2} \int_0^{\frac{1-\cos(\theta)}{1+\cos(\theta)}} \frac{z^{(k-2)/2}}{(1+z)^k} dz \tag{8}$$

$$Pr\left\{\frac{(Rh)^T Rx}{h^T x} \leq 0\right\} \quad = \quad f_k(\theta) \cdot \mathbf{1}(h^T x > 0) + (1 - f_k(\theta)) \cdot \mathbf{1}(h^T x < 0) \tag{9}$$

$$\leq \quad \exp(-k \cos^2(\theta)/2) \tag{10}$$

(b) If R has i.i.d. sub-Gaussian entries, and $h^T x \neq 0$, we have:

$$Pr\left\{\frac{(Rh)^T Rx}{h^T x} \leq 0\right\} \leq \exp(-k \cos^2(\theta)/8) \tag{11}$$

To develop more intuition, Figure 1 provides a visualization of the numerical evaluations of $f_k(\theta)$ from eq. (8). Note that it depends only on the angle between a pair of vectors and the projection dimension k, it is independent of the dimensionality of the data, d.

The inequalities of Lemma 2.2 reveal, for instance, that the probability that a point flips to the wrong side of the decision hyperplane decreases exponentially not only with increasing k but also with the square of the cosine similarity of the point with h^*. Therefore, taken together Theorem 2.1 and Lemma 2.2 discover that data distributions that have more probability mass at regions where points have large cosine (i.e. large normalized margin) w.r.t. h^* are more benign for compressive classification. This is consistent with margin distribution bounds in learning theory (Shawe-Taylor & Cristianini, 1999), and we should note that no sparse representation is required for the compressive learning guarantee provided by Theorem 2.1.

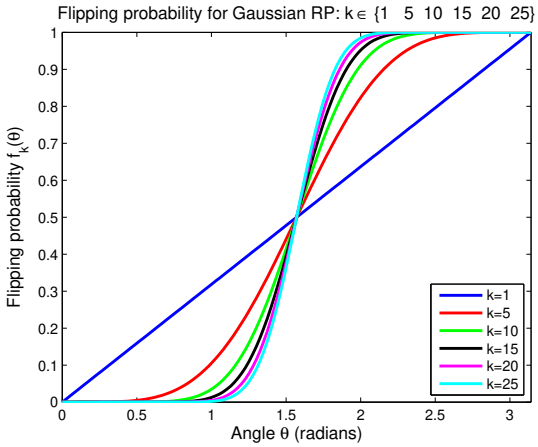

Figure 1: Illustration of the function $f_k(\theta)$ from eq. (8) in Lemma 2.2

.

In fact, an appealing advantage of RP-based dimension reduction is its universality: The structure preserving ability of RP makes the subsequent classifier adapt to all structures that ensure low distortion, without needing prior knowledge of what these structures might be, without needing a separate treatment for each. Indeed, we already mentioned two different structures that reduce flip rates – the case when the data lives in a low-dimensional subspace, and the case when the class-conditional supports of the distribution are separated by a margin – both of which tighten our generalization guarantee, essentially for free.

For learning from RP-ed data, we may also use the bounds from Lemma 2.2 to deduce the required value of k that ensures that the contribution to the risk from flip probabilities remains (with high probability) below some user-specified threshold ϵ. From Theorem 2.1 combined with Part (b) of Lemma 2.2, if the classes are separable with a margin $\inf_{x\in\mathcal{X}} \cos(\theta^{h^*}_{xy}) > 0$, then for any small $\epsilon > 0$, the last two terms on the r.h.s. in Theorem 2.1 will be below ϵ for:

$$k \geq \frac{8\log(1/(\epsilon\delta))}{\inf_{x\in\mathcal{X}} \cos^2(\theta^{h^*}_{xy})}. \tag{12}$$

The case of zero margin can also be accommodated, as it is easy to modify the proof of Theorem 2.1 by artificially introducing a margin parameter $\gamma > 0$, which yields the following corollary.

Corollary 2.2.1. *Fix some $\gamma > 0$. Let R be a $k \times d$ subgaussian random matrix with i.i.d. entries, $k \leq d$, and denote $f^\gamma_k(\theta^h_u) := f_k(\theta^h_u) \cdot \mathbf{1}(\cos(\theta^h_u) > \gamma)$. For any $\delta \in (0,1)$, the following holds for the compressive linear ERM classifier $\hat{h}_R$ with probability $1-2\delta$:*

$$\begin{aligned} Pr_{x,y}\{(\hat{h}_R^T Rx + b)y \leq 0\} \quad &\leq \quad E_{x,y}[\mathbf{1}\{\cos(\theta^{h^*}_{xy}) \leq \gamma\}] + c\sqrt{\frac{k+1+\log(1/\delta)}{N}}... \\ &+ \quad E_{x,y}[f^\gamma_k(\theta^{h^*}_{xy})] + \min\left\{\frac{1-\delta}{\delta}\cdot E[f^\gamma_k(\theta^{h^*}_{xy})], \sqrt{\frac{1}{2}\log\frac{1}{\delta}}\right\} \end{aligned}$$

where $c > 0$ is a universal constant, independent of d and N.

Hence, by a similar argument, setting $k \geq \frac{8\log(1/(\epsilon\delta))}{\gamma^2}$ is sufficient to guarantee that the last two terms in Corollary 2.2.1 are below a threshold of ϵ.

2.1 What If the Data Has a Sparse Representation?

Recall that compressive learning guarantees, inspired by the compressed sensing (CS) literature, relied heavily on techniques that assume that the input data has a sparse representation. Yet, we have just demonstrated that a sparse representation is not a necessary condition for good generalization from compressive data. One may wonder then, where do such sparse representation conditions fit into this picture? In which ways does sparsity facilitate learning?

To address this question, we look at what properties of the data could ensure that with high probability no flipping events occur? This is rather a strong requirement: we are asking for conditions on the data generator such that (all other things being equal) with high probability learning from RP-ed data incurs *no reduction in classification performance.*

Now, since we work with the zero-one loss directly, we can take all $h \in \mathcal{H}$ and $x \in \mathcal{X}$ to have unit norm without any loss of generality. Therefore let:

$$U := \left\{ \frac{xy}{\|x\|} : x \in \mathcal{X}, y \in \{-1, 1\} \right\}. \tag{13}$$

Then for any $h \in \mathcal{H}$, we define the margin of h to be $\gamma_h := \inf_{u \in U} \cos(\theta_u^h)$. We further define, for any fixed γ and h, the following set:

$$T_{h,\gamma}^{+} \quad := \quad \left\{ u \in U : \cos(\theta_u^h) \geq \gamma \right\} \subset S^{d-1}; \text{ where } \gamma > 0 \tag{14}$$

Thus, the set $T_{h,\gamma}^{+}$ defined above is the set of all points in the support $\mathcal{X}$ of the underlying (unknown) input data distribution that the high dimensional vector $h \in \mathcal{H}$ classifies correctly with a pre-specified margin of at least γ. Note, however, that we are *not* requiring the data support of the classes to have a margin, nor to be linearly separable.

With these definitions in place, the following theorem gives a condition on the compressed dimension k that ensures a risk guarantee for the compressive linear ERM classifier with a similar flavour to either margin bounds or VC bounds for a k (rather than d) dimensional dataspace classifier. As we shall see in Theorem 2.3 below, this condition depends on the geometric structure of the problem, which will be reflected by the so-called Gaussian width of the set $T_{h^*,\gamma}^{+}$.

Definition 1. *(Vershynin, 2018) The Gaussian width of a set T is defined as:*

$$w(T) := E_{g \sim \mathcal{N}(0,I)}[\sup_{x \in T} g^T x]. \tag{15}$$

Theorem 2.3. *Let R be a $k \times d, k \leq d$ (isotropic) subgaussian random matrix with independent rows each having subgaussian norm bounded as $\|R_i\|_{\psi_2} \leq K$. Fix some $\gamma > 0$ large enough that $\gamma \geq \gamma_{h^*}$. Then, for any $\delta > 0$ there are absolute constants $C, c > 0$ such that with probability $1 - 2\delta$ the generalization error of the compressive linear ERM classifier, $\hat{h}_R$, is bounded as the following:*

$$Pr_{x,y}[(\hat{h}_R^T R x + b)y \leq 0] \leq E_{x,y}\left[\mathbf{1}\left(\cos(\theta_{xy}^{h^*}) < \gamma\right)\right] + c\sqrt{\frac{k + 1 + \log(1/\delta)}{N}} \tag{16}$$

provided that $k \geq CK^4 \left(w(T^+_{h^*,\gamma}) + \sqrt{\log(1/\delta)}\right)^2 \gamma^{-1}$.

In the special case when h^* has a margin and $T^+_{h^*,\gamma} = T^+_{h^*,\gamma_{h^*}}$ then the empirical risk on the compressive classifier for the required k incurs no increase in error from the compression, and has a guarantee as strong as that for a k-dimensional dataspace classifier, regardless of the dimensionality of the original data.

On the other hand, if h^* has no positive margin, then $\gamma_{h^*} = 0$, but $\gamma > 0$ ensures that the bound still holds, in the same way as in Corollary 2.2.1. However, the last two terms from Corollary 2.2.1 have now disappeared due to the more stringent requirement of no flipping, which is ensured (with high probability) by the target dimension k in the statement of Theorem 2.3. Observe furthermore that, despite this stringent requirement, the bound of Theorem 2.3 does still not explicitly depend on the dimension d of the original data, but only through k. This is where the effect of sparse representation of the input data will become apparent – indeed, as we shall see shortly, one example where $\mathcal{O}(k)$ is less than d is when the data has a sparse representation.

2.1.1 Examples when $\mathcal{O}(k) < d$

As $T^+_{h^*,\gamma} \subseteq S^{d-1}$ we always have $w(T^+_{h^*,\gamma})^2 \leq d$. This follows e.g. from (Vershynin, 2018, Prop. 7.7.2). Equality holds if $T^+_{h^*,\gamma}$ is the whole d-dimensional hypersphere, and it can be much less than d when the data support has a simpler structure, such as the following:

- If h^* has a large margin for the points of the data support that don't contribute to the empirical error term, then $w(T^+_{h^*,\gamma})$ reduces. To see this, note that, if the correctly classified datapoints are concentrated around the antipodes of the unit sphere and h^* is roughly in the direction of the north pole (say) – so that it has a large margin – then $T^+_{h^*,\gamma}$ is contained in a spherical cap making a small angle $\phi := \arccos(\gamma_{h^*})$ at the origin, while if the margin of h^* is small then ϕ for the spherical cap containing $T^+_{h^*,\gamma}$ is larger. As shown by Amelunxen et al. (2014) and Bandeira et al. (2017), the squared Gaussian width of a spherical cap is $d(1-\cos^2(\phi)) + O(1)$. Observe also that the vector h^* that reduces this quantity also increases the margin, $\cos(\theta^{h^*}_{xy})$.

- If the data lives in an s-dimensional subspace, then $[w(T^+_{h^*,\gamma})]^2$ is of order $\mathcal{O}(s)$ rather than $\mathcal{O}(d)$.

- If the data has a sparse representation, i.e. its support lives in a union of disjoint s-dimensional subspaces, then the squared Gaussian width of the support of the classes is of order $\Theta(s\log(2d/s))$ (Plan & Vershynin, 2013, Lemma 3.5). Here we use the facts that the intersection of the length-normalized data support with the unit sphere is then a union of disjoint $(s-1)$-spheres, the Gaussian width of a sphere is the same as that of a ball, and $T^+_{h^*,\gamma}$ is a subset of the projection of the support onto the unit sphere, so its squared Gaussian width $[w(T^+_{h^*,\gamma})]^2$ is no larger.

- If the optimal classifier vector h^* is sparse, this will reduce the contribution of the Gaussian width term to the error bound. However, interesting to note that in the case when irrelevant noise features exist, then even a sparse h^* cannot completely circumvent their bad effect on compressive classification. Indeed, even though $[w(T^+_{h^*,\gamma})]^2$ is

reduced in this way, the noise components increase $\|x\|$ in the denominator of the cosine, so the empirical error term will still tend to increase with the extent of irrelevant features.

More examples of structured sets exist of course, as there are many ways in which the data may not 'fill' the full d-dimensional space. Our generic bound captures the effect of these structures on the compressive classifier performance without the need of prior knowledge of the particular structure that might be present. This explains, for instance, why a drastic random compression still works surprisingly well in practice for face recognition (Goel, Bebis, & Nefian, 2005) but not so well on difficult data sets that contain large amounts of unstructured noise (Xie, Li, Zhang, & Wang, 2016).

To prove Theorem 2.3, we need the following lemma.

Lemma 2.4. *(Uniform bound on sign flipping) Fix $h \in \mathbb{R}^d, \|h\| = 1$ w.l.o.g. Let R be an isotropic subgaussian random matrix with independent rows having subgaussian norm bounded as $\|R_i\|_{\psi_2} \leq K$. Let $T^+_{h,\gamma}$ be as defined in eq. (14). Then, for any fixed h, $\forall \delta \in (0,1)$, w.p. $1-\delta$ w.r.t. draws of R,*

$$P_R\left\{\exists u \in T^+_{h,\gamma} : h^T R^T R u \leq 0\right\} < \delta \tag{17}$$

provided that $k \geq CK^4 \left(w(T^+_{h,\gamma}) + \sqrt{\log(2/\delta)}\right)^2 / \gamma$ for some absolute constant C.

Proof of Lemma 2.4. By the parallelogram law,

$$\begin{aligned} -\frac{h^T R^T R u}{k} &= \frac{1}{4k}\left(\|R(h-u)\|^2 - \|R(h+u)\|^2\right) \\ &= \frac{1}{4}\left(\frac{\|R(h-u)\|^2}{k} - \|h-u\|^2\right) - \frac{1}{4}\left(\frac{\|R(h+u)\|^2}{k} - \|h+u\|^2\right) - h^T u \end{aligned}$$

Now each of the brackets is an empirical process, and we can make use a the following result (Liaw, Mehrabian, Plan, & Vershynin, 2017, Theorem 1.4) that bounds the suprema of such processes[3].

Theorem (Liaw et al., 2017). Let R be an isotropic subgaussian random matrix with independent rows having subgaussian norm bounded as $\|R_i\|_{\psi_2} \leq K$. Let T be a bounded subset of R^d, and denote its radius by $\mathrm{rad}(T) = \sup_{u \in T}\|u\|$. Then there is an absolute constant C s.t. with probability $1-\delta$ w.r.t. the random draws of R,

$$\begin{aligned} \sup_{u \in T}\left|\frac{\|Ru\|_2^2}{k} - \|u\|_2^2\right| &\leq \frac{1}{k}\left[C^2K^4\left(w(T) + \mathrm{rad(T)}\sqrt{\log(1/\delta)}\right)^2\right. \\ &+ \left. 2CK^2\mathrm{rad}(T)\left(w(T) + \mathrm{rad}(T)\sqrt{\log(1/\delta)}\right)\sqrt{k}\right] \end{aligned} \tag{18}$$

3. This result is an improvement on seminal work by Klartag and Mendelson (2005) that first connected random projections with empirical processes.

Denoting by $V(T',\delta)/k$ the r.h.s. of eq. (18), where T' is the set whose Gaussian width appears in it, we get w.p. $1-2\delta$:

$$\sup_{u\in T^+_{h,\gamma}}\left[-\frac{h^TR^TRu}{k}\right] \leq \frac{V(T^+_{1,h,\gamma},\delta)}{4k}+\frac{V(T^+_{2,h,\gamma},\delta)}{4k}-\gamma \tag{19}$$

where

$$T_1=\{h-u : u\in T^+_{h,\gamma}\};\quad T_2=\{h+u : u\in T^+_{h,\gamma}\} \tag{20}$$

Now, since h is a fixed vector, and the Gaussian width is invariant to translation, $w(T_1)=w(T_2)=w(T^+_{h,\gamma})$.

We need to require that the r.h.s. of eq. (19) is non-positive. Since $\|h\|=\|u\|=1$, and $h^Tu=\cos(\theta^h_u)\geq\gamma$, this is equivalent to requiring that:

$$\frac{V(T^+_{h,\gamma},\delta)}{2k}\leq\gamma \tag{21}$$

Hence, for

$$k\geq\frac{V(T^+_{h,\gamma},2\delta)}{2\gamma}=\frac{\Omega(w(T^+_{h,\gamma})^2)}{\gamma} \tag{22}$$

we have the statement of Lemma 2.4, noting again that by definition $T^+_{h,\gamma}$ is a set of unit vectors, so its radius is 1. □

Proof of Theorem 2.3. For a fixed instance of R, we have the uniform VC bound of eq. (4). We upper bound the error of the optimal classifier in $\mathbb{R}^{k+1}$ as the following:
$\Pr_{x,y}((h^{*T}_R Rx+b^*)y\leq 0)\leq ...$

$$\begin{aligned}
... &\leq \Pr_{x,y}(h^{*T}R^TRxy\leq 0)\\
&= \mathrm{E}_{x,y}[\mathbf{1}\left(h^{*T}R^TRxy\leq 0\right)-\mathbf{1}\left(\cos(\theta^{h^*}_{x_ny_n})\leq\gamma\right)]+\Pr_{x,y}\left(\cos(\theta^{h^*}_{xy})\leq\gamma\right)\\
&= \mathrm{E}_{x,y}[\mathbf{1}\left(h^{*T}R^TRxy\leq 0\right)\mathbf{1}\left(\cos(\theta^{h^*}_{xy})>\gamma\right)\\
&- \mathbf{1}\left(h^{*T}R^TRxy>0\right)\mathbf{1}\left(\cos(\theta^{h^*}_{xy})\leq\gamma\right)]+\Pr_{x,y}\left(\cos(\theta^{h^*}_{xy})\leq\gamma\right)\\
&\leq \mathrm{E}_{x,y}[\mathbf{1}\left(h^{*T}R^TRxy\leq 0\right)\mathbf{1}\left(\cos(\theta^{h^*}_{xy})>\gamma\right)]+\Pr_{x,y}\left(\cos(\theta^{h^*}_{xy})\leq\gamma\right)]
\end{aligned}$$

Now, by Lemma 2.4, for the stated k the the first term on the r.h.s. is zero w.p. $1-\delta$. This completes the proof. □

3. From Random Projections Back to the Dataspace: Geometry-Aware Error Bounds with the Zero-One Loss

With the insights gained regarding the ability of RP to exploit benign geometry for compressed classification, it is natural to wonder if a similar approach is possible to better understand the original high dimensional classification problem. If so, this would yield more

informative bounds that can predict generalization performance from quantities computed on the training set.

We will not assume *a priori* knowledge of properties of the data generator to prime the analysis, but instead we will let RP, as an analytical tool, capture and discover any beneficial structure automatically. We will evaluate the success of this strategy in terms of how informative or predictive of generalization performance are the bounds so obtained. We note that any positive outcomes could also be indicative of some potentially wider applicability of such an RP-based approach.

Throughout this section, let R be a $k \times d, k \leq d$, rotation-invariant random matrix, such as a random matrix having i.i.d. zero-mean Gaussian entries. That is, R is a random projection (RP) just as in the previous Section 2, but instead here it will serve as an analytical tool rather than as a method for dimension reduction. Due to this change of scope, we no longer need to take computational efficiency into consideration, and we will consider ourselves free to make the choice concerning the distribution of the entries of R for analytical convenience.

Again, as in Section 2, θ_u^h will denote the angle between the d-dimensional vectors u and h, and $f_k(\theta_u^h) := \Pr_R((Rh)^T Ru \leq 0)$.

3.1 Risk Bounds for Dataspace Classification

We start by decomposing the risk into a distortion term that consists of the expected absolute difference between the value of the zero-one loss before and after random projection, plus the error on the compressed classification:

$$\Pr_{x,y}[h^T xy \leq 0] \leq \mathrm{E}_{x,y}\left|\mathbf{1}(h^T xy \leq 0) - \mathrm{E}_R[\mathbf{1}(h^T R^T Rxy \leq 0)]\right| + \mathrm{E}_{x,y,R}[\mathbf{1}(h^T R^T Rxy \leq 0)] \tag{23}$$

The distortion term will measure the compressibility of the problem, i.e. to what extent the data geometry would allow the high dimensional problem to be solvable in lower dimension. We can further re-write this latter term as the following:

$$\begin{aligned} &\mathrm{E}_{x,y}\left|\mathbf{1}(h^T xy \leq 0) - \mathrm{E}_R[\mathbf{1}(h^T R^T Rxy \leq 0)]\right| \\ &= \mathrm{E}_{x,y}\left[(1 - f_k(\theta_{xy}^h)) \cdot 1(h^T xy \leq 0) + f_k(\theta_{xy}^h) \cdot 1(h^T xy > 0)\right] \end{aligned}$$

Plugging back, for Gaussian R the r.h.s. of eq. (23) becomes:

$$\mathrm{E}_{x,y}[\mathbf{1}(h^T xy \leq 0) + 2f_k(\theta_{xy}^h)\mathbf{1}(h^T xy > 0)] = \mathrm{E}_{x,y}[\min(1, 2f_k(\theta_{xy}^h))] \tag{24}$$

Eq. (24) holds for any k and any $h, x \in \mathbb{R}^d, y \in \{-1, 1\}$, since R was chosen to be rotationally invariant. Indeed, for the i.i.d. Gaussian R, we can read this off directly from the integral representation of f_k in eq. (8). There are three cases: First, notice that $f_k(\pi/2) = 1/2$ for any k – that is, a point on the decision boundary has probability $1/2$ to flip across. Furthermore, $f_k(\theta_{xy}^h)\mathbf{1}(h^T xy > 0) \leq 1/2$ – that is because a point on the correct side corresponds to $\theta_{xy}^h < \pi/2$, and has no more than $1/2$ probability to flip across to the wrong side. Finally, by symmetry, we have $f_k(\theta_{xy}^h)\mathbf{1}(h^T xy < 0) \geq 1/2$ – a point on the wrong side of the boundary has no less than $1/2$ probability to remain on the wrong side. In this latter case, both sides of eq. (24) evaluate to 1; in the former case both sides evaluate to $2f_k(\theta_{xy}^h)$, and the two cases coincide when x is on the boundary.

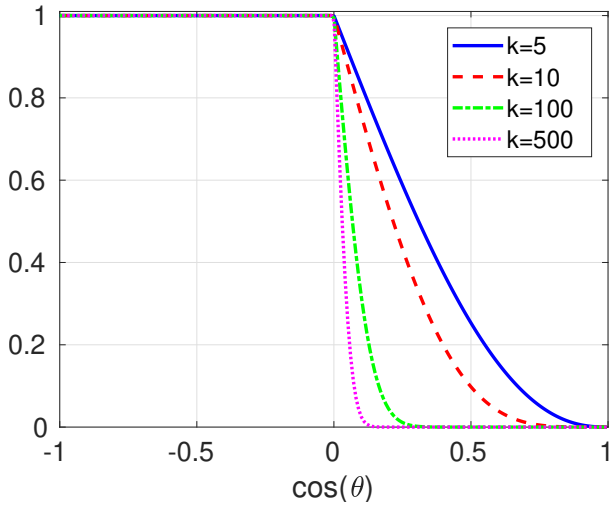

Figure 2: Loss functions from eq. (24).

Now it may be interesting to regard eq. (24) as a family of new loss functions. We illustrate these, as functions of $\cos(\theta)$, in Figure 2, for a few different values of k.

Many surrogate loss functions are in use already. As our discussion unfolds we shall notice some similarities with the margin loss. However, while the margin loss pre-specifies that margin is a benign structure, our loss function discovers it from the requirement of low distortion under RP. Moreover, contrary to the margin loss and other surrogate losses, it retains the scale-invariance of the zero-one loss. Indeed, if we scale the data or/and the classifier, the new loss function value remains the same.

However, whether this will be useful, will depend on the complexity of the function class defined by the new loss function, which we need to evaluate. To make this tractable, we will work with the Gaussian R, as this allows us to exploit the availability of the exact expression of the loss function, using Lemma 2.2. We will prove the following result:

Theorem 3.1. *Fix any positive integer $k \leq d$. For any $\delta > 0$, with probability at least $1-\delta$ with respect to the random draws of $\mathcal{T}^N$ of size N, $\forall h \in \mathcal{H}$ the generalization error of h is upper bounded as the following:*

$$Pr_{x,y}[h^T xy \leq 0] \quad \leq \quad \frac{1}{N}\sum_{n=1}^{N} \min(1, 2f_k(\theta^h_{x_n y_n})) + \frac{2\sqrt{2}}{\sqrt{\pi}}\sqrt{\frac{k}{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}}$$

where $f_k(\theta)$ is the expression given in eq. (8).

The first term is the empirical error computed from the sample through our new loss function. The second term is analogous to the complexity term in VC bounds – however, now the target dimension k of the RP takes the place of the VC dimension, reducing this term from $\sqrt{d/N}$ to $\sqrt{k/N}$ at the expense of a higher empirical average, unless there is enough benign structure to support a smaller value of k. If k grows to d and there is no structure at all, we recover the classical VC bound.

The analogy with margin bounds becomes apparent if we interpret k as an inverse margin, and look at the tradeoff this manages. The usual norm constraints in margin bounds are now taken care of automatically by the cosine similarities, and in this view the bound may be interpreted as an alternative derivation of margin distribution arguments. In later subsections we shall see how this is useful for establishing new connections between existing methods and algorithms.

Before delving into the proof, we should mention that, in the form stated, k needs to be specified before seeing the sample. We can think of it as an 'affordable complexity' which we can conveniently choose relative to the available sample size. The distortion term measured on the sample will then reflect the extent of error incurred. Alternatively we may apply structural risk minimization (SRM) (Vapnik, 1998) to make the bound hold uniformly for all values of k. Take the sequence $(k_i)_{i\geq 1}$ with $k_i = i$, and let μ_i – chosen before seeing the sample – be our prior belief in the value k_i s.t. $\sum_{i\geq 1}\mu_i = 1$. Then applying Theorem 3.1 with $\delta_i := \delta\mu_i$, and applying union bound over the sequence of values k_i, we get the conclusion of Theorem 3.1 simultaneously for all k at an expense of a small additive error. For instance, if we choose an exponential prior probability sequence, $\mu_i = 2^{-k}$, then this additional error term evaluates to $3\sqrt{\frac{\log(2)}{2}}\sqrt{\frac{k}{N}}$, yielding the following corollary.

Corollary 3.1.1. *For any $\delta > 0$, with probability at least $1-\delta$ with respect to the random draws of $\mathcal{T}^N$ of size N, $\forall h \in \mathcal{H}, \forall k \leq d$ the generalization error of h is upper bounded as the following:*

$$Pr_{x,y}[h^T xy \leq 0] \leq \frac{1}{N}\sum_{n=1}^{N}\min(1, 2f_k(\theta^h_{x_n y_n})) + \left(\frac{2\sqrt{2}}{\sqrt{\pi}} + \frac{3\sqrt{\log(2)}}{\sqrt{2}}\right)\sqrt{\frac{k}{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}}$$

where $f_k(\theta)$ is the expression given in eq. (8).

We remark that the constant in the brackets above is $\left(\frac{2\sqrt{2}}{\sqrt{\pi}} + \frac{3\sqrt{\log(2)}}{\sqrt{2}}\right) \approxeq 3.362$.

Figure 3 demonstrates the bound of Theorem 3.1 for a synthetic example, with $\delta = 0.05$. The sample size was $N = 5000$, and the cosine values were generated from a 0-mean Gaussian with variance $1/9$; values outside of $[-1, 1]$ were replaced with samples from the uniform distribution on $[-1, 1]$. The first term on the r.h.s. of the bound ('Flip p') is plotted against the sum of the last two terms ('Complexity'), along with their sum ('Bound'). We see the trade-off between the average flip probability under RP, and the complexity of the function class in the RP space, as the RP dimension k varies.

To prove Theorem 3.1, we need the following lemma.

Lemma 3.2. *The function $f_k(\theta)$ satisfies the following properties:*

(a) Lipschitz continuity as a function of $\cos(\theta)$, with constant $\sqrt{\frac{k}{2\pi}}$:
For any k and any $\theta_1, \theta_2 \in [0, 2\pi]$,

$$|f_k(\theta_1) - f_k(\theta_2)| \leq \sqrt{\frac{k}{2\pi}}|\cos(\theta_1) - \cos(\theta_2)|. \tag{25}$$

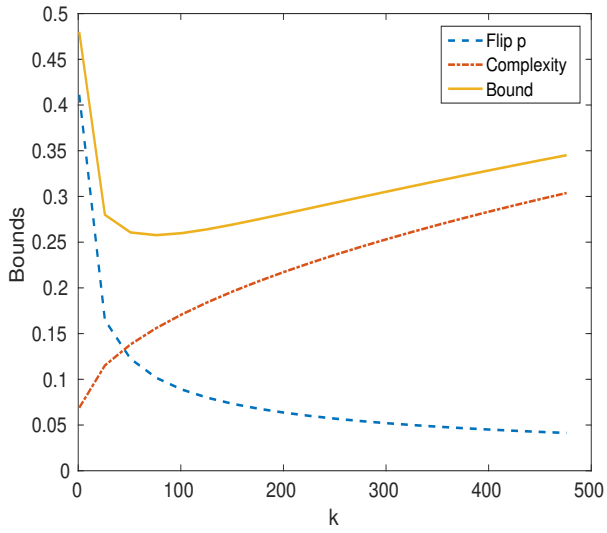

Figure 3: Illustration of the bound of Theorem 3.1.

(b) Scale invariance:
For any $c_1, c_2 > 0$,

$$f_k(\theta^{c_2 h}_{c_1 x}) = f_k(\theta^h_x). \tag{26}$$

Proof of Lemma 3.2. Part (a). Let $a := \cos(\theta)$, and define

$$\varphi_k(a) := \frac{\Gamma(k)}{(\Gamma(k/2))^2} \int_0^{\frac{1-a}{1+a}} \frac{z^{(k-2)/2}}{(1+z)^k} dz \tag{27}$$

so we have

$$f_k(\theta^h_x) = \varphi_k(\cos(\theta)). \tag{28}$$

By the Leibniz integration rule we have that:

$$|\varphi_k(a)| = \left| -\frac{\Gamma(k)}{2^{k-1}(\Gamma(k/2))^2}(1-a^2)^{\frac{k-2}{2}} \right| \tag{29}$$

$$\leq \frac{\Gamma(k)}{(\Gamma(k/2))^2 2^{k-1}} =: L \tag{30}$$

Hence, $\varphi_k(\cdot)$ is L-Lipschitz.

We can further simplify the expression of L by rewriting it into a Gamma function ratio. Using the duplication formula (Abramowitz & Stegun, 1965, 6.1.18, pg. 256):

$$\Gamma(2z) = (2\pi)^{-\frac{1}{2}} 2^{2z-\frac{1}{2}} \Gamma(z)\Gamma((2z+1)/2) \tag{31}$$

with $z = k/2$, the expression of L is equal to:

$$\frac{2^{k-\frac{1}{2}}\Gamma(k/2)\Gamma((k+1)/2)}{\sqrt{2\pi}2^{k-1}(\Gamma(k/2))^2} = \frac{\Gamma(k/2)\Gamma((k+1)/2)}{\sqrt{\pi}\ (\Gamma(k/2))^2} = \frac{\Gamma((k+1)/2)}{\sqrt{\pi}\ \Gamma(k/2)} \tag{32}$$

Now we use an inequality for Gamma function ratios (Wendel, 1948):

$$x(x+y)^{y-1} \leq \frac{\Gamma(x+y)}{\Gamma(x)} \leq x^y, \forall y \in [0,1], \tag{33}$$

which yields:

$$L \;=\; \frac{\Gamma(k/2+1/2)}{\sqrt{\pi}\;\Gamma(k/2)} \leq \sqrt{\frac{k}{2\pi}} \tag{34}$$

This completes the proof of part (a).

Part (b) is immediate, as the expression of $f_k(\theta_x^h)$ depends on x and h only through their angle. □

With all ingredients in place, the proof of Theorem 3.1 is now straightforward.

Proof of Theorem 3.1. By the classical Rademacher complexity based risk bound (Mohri, Rostamizadeh, & Talwalkar, 2012; Koltchinskii & Panchenko, 2002, Theorem 3.1.), for any positive integer k, the following holds uniformly for all halfspace classifiers $h \in \mathcal{H}$:

$$\Pr_{x,y}[h^T xy \leq 0] \;\leq\; \frac{1}{N}\sum_{n=1}^{N} \min(1, 2f_k(\theta^h_{x_n y_n})) + 2\hat{\mathcal{R}}_N(G_k) + 3\sqrt{\frac{\log(2/\delta)}{2N}} \tag{35}$$

where $\hat{\mathcal{R}}_N(\cdot)$ denotes the empirical Rademacher complexity of the function class in its argument, and we defined the function class G_k:

$$G_k = \{u \to \min(1, 2f_k(\theta_u^h)) : h \in \mathbb{R}^d\} \tag{36}$$

By Lemma 3.2 (a), the functions in this class are Lipschitz functions of $\cos(\theta^h_{xy})$. Hence, to compute the Rademacher complexity, we interpret G_k as a composition:

$$G_k = \ell_k \circ \mathcal{F}, \tag{37}$$

where

$$\ell_k : [-1,1] \to [0,1], \quad \ell_k(a) \;=\; \min\left(1,\; 2\frac{\Gamma(k)}{(\Gamma(k/2))^2}\int_0^{\frac{1-a}{1+a}} \frac{z^{(k-2)/2}}{(1+z)^k} dz\right)$$

$$\mathcal{F} \;=\; \left\{u \to \frac{h^T}{\|h\|}\frac{u}{\|u\|} : h \in \mathbb{R}^d\right\}$$

By Lemma 3.2 part (a), ℓ_k has Lipschitz constant $\sqrt{\frac{2k}{\pi}}$. In consequence, by Talagrand's contraction lemma we have:

$$\hat{\mathcal{R}}_N(G_k) \;\leq\; \sqrt{\frac{2k}{\pi}} \cdot \hat{\mathcal{R}}_N(\mathcal{F}) \tag{38}$$

Finally, since $\mathcal{F}$ is a linear function class of $h/\|h\|$, and the vectors $h/\|h\|$ and $xy/\|x\|$ have unit norms, we have (Mohri et al., 2012, Theorem 4.3) that $\hat{\mathcal{R}}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}}$ Combining this with eqs (35) and (38) completes the proof. □

3.1.1 A Variant with Local Choices of k

The idea in this section is to allow the values of k to differ for each input point. This will give us choices to pre-define more compressible and less compressible regions of the input space when appropriate. This leads to the following result.

Theorem 3.3. *Let $k : \mathbb{R}^d \times \mathcal{H} \to \mathbb{N}$ a deterministic function specified independently of $\mathcal{T}^N$. Then $\forall h \in \mathcal{H}$, with probability $1-\delta$ with respect to the random draw of a training set of size N, the generalization error of h is upper bounded as the following:*

$$\begin{aligned} Pr_{x,y}[h^T xy \leq 0] &\leq \frac{1}{N}\sum_{n=1}^{N} \min(1, 2f_{k(x_n y_n,h)}(\theta^h_{x_n y_n})) + 2\sqrt{\frac{2}{\pi}}\sqrt{\frac{1}{N}\max_{n=1}^{N} k(x_n y_n, h)} \\ &+ 3\sqrt{\frac{\log(2/\delta)}{2N}} + 3\sqrt{\frac{\log(2)}{2}}\sqrt{\frac{1}{N}\max_{n=1}^{N} k(x_n y_n, h)} \end{aligned}$$

Some comments are now in order, as it may seem impractical to compress each point to a different dimension. However, recall that our use of random projections only serves an analytic purpose throughout this section; we are interested in what can be said about the original learning problem in the original space, through the lens of random projections.

The implications of Theorem 3.3 in theory and practice will be discussed further in Section 3.2 and Section 3.3.1, where this result will allow us to make connections between various existing results and successful algorithms. In particular, the bounds in Corollary 3.4.1 and Corollary 3.5.2 will both be implied by Theorem 3.3, and the r.h.s. of those bounds are minimized by the Large Margin Distribution Machine (Zhang & Zhou, 2014) and a regularized boosting (Schapire, 2013) respectively.

Before turning to the proof, it should be observed that, at a first sight, the goal of the bound in Theorem 3.3 might seem tricky technically, since each choice of k is a model order selection problem, and we now try to have as many as the number of data points. The trick here is to notice and exploit a specific property of the function $f_k(\theta)$, given in Lemma 3.4 below, namely its monotonicity with respect to k on $\theta \in [0, \pi/2]$. Using this, the proof of Theorem 3.3 becomes straightforward.

Lemma 3.4. *For any integer $k \geq 1$ and any $h, x \in \mathbb{R}^d$, we have:*

$$f_k(\theta^h_x)\mathbf{1}(h^T x > 0) \geq f_{k+1}(\theta^h_x)\mathbf{1}(h^T x > 0). \tag{39}$$

Proof of Lemma 3.4. We will use an equivalent rewriting of $f_k(\phi)$ as the ratio of the area of a hyperspherical cap to the area of the corresponding hypersphere (Durrant & Kabán, 2013, part 2 of Theorem 3.2), that is:

$$f_k(\phi) = \frac{\int_0^\phi \sin^{k-1}(\theta)\ \mathrm{d}\theta}{\int_0^\pi \sin^{k-1}(\theta)\ \mathrm{d}\theta}. \tag{40}$$

It is sufficient to show that for all $\phi \in [0, \pi/2]$, the ratio of successive flip probabilities:

$$\frac{f_{k+1}(\phi)}{f_k(\phi)} = \left(\frac{\int_0^\phi \sin^{k}(\theta)\ \mathrm{d}\theta}{\int_0^\pi \sin^{k}(\theta)\ \mathrm{d}\theta}\right) / \left(\frac{\int_0^\phi \sin^{k-1}(\theta)\ \mathrm{d}\theta}{\int_0^\pi \sin^{k-1}(\theta)\ \mathrm{d}\theta}\right) \leqslant 1 \tag{41}$$

Let us rewrite the ratio eq. (41) above as:

$$\left(\frac{\int_0^\phi \sin^k(\theta)\ \mathrm{d}\theta}{\int_0^\phi \sin^{k-1}(\theta)\ \mathrm{d}\theta}\right) / \left(\frac{\int_0^\pi \sin^k(\theta)\ \mathrm{d}\theta}{\int_0^\pi \sin^{k-1}(\theta)\ \mathrm{d}\theta}\right) \tag{42}$$

Denote the numerator of eq. (42) by $g_k(\phi)$, and notice that the denominator is nothing but $g_k(\pi)$.

Now observe that the denominator, $g_k(\pi) = g_k(\pi/2)$. Indeed,

$$\frac{\int_0^\pi \sin^k(\theta)\ \mathrm{d}\theta}{\int_0^\pi \sin^{k-1}(\theta)\ \mathrm{d}\theta} = \frac{2\int_0^{\pi/2} \sin^k(\theta)\ \mathrm{d}\theta}{2\int_0^{\pi/2} \sin^{k-1}(\theta)\ \mathrm{d}\theta} = \frac{\int_0^{\pi/2} \sin^k(\theta)\ \mathrm{d}\theta}{\int_0^{\pi/2} \sin^{k-1}(\theta)\ \mathrm{d}\theta} \tag{43}$$

where the first equality follows from the symmetry of the sine function about $\pi/2$. Hence, we see that the whole expression (42) is equal to 1 when $\phi = \pi/2$.

Now it remains to show that $g_k(\phi) \leq g_k(\pi/2), \forall \phi \in [0, \pi/2], \forall k$. In fact more is true: We show that $\forall k$, the function $g_k(\phi)$ is monotonic increasing on the interval $[0, \pi/2]$. From this the required inequality follows, and hence the expression (42) has its maximum value of 1 on this domain, from which the result follows.

To show monotonicity, we differentiate the function $g_k(\phi)$ w.r.t ϕ, and obtain:

$$\frac{\mathrm{d}}{\mathrm{d}\phi} g_k(\phi) = \frac{\sin^k(\phi)\int_0^\phi \sin^{k-1}(\theta)\mathrm{d}\theta - \sin^{k-1}(\phi)\int_0^\phi \sin^k(\theta)\mathrm{d}\theta}{\left(\int_0^\theta \sin^{k-1}(\theta)\mathrm{d}\theta\right)^2} \tag{44}$$

Then (44) is greater than zero when its numerator is, and:

$$\begin{aligned} &\sin^k(\phi)\int_0^\phi \sin^{k-1}(\theta)\mathrm{d}\theta - \sin^{k-1}(\phi)\int_0^\phi \sin^k(\theta)\mathrm{d}\theta \\ &= \sin^{k-1}(\phi)\left[\sin(\phi)\int_0^\phi \sin^{k-1}(\theta)\mathrm{d}\theta - \int_0^\phi \sin^k(\theta)\mathrm{d}\theta\right] \qquad (45) \\ &= \sin^{k-1}(\phi)\left[\int_0^\phi \sin(\phi)\sin^{k-1}(\theta)\mathrm{d}\theta - \int_0^\phi \sin(\theta)\sin^{k-1}(\theta)\mathrm{d}\theta\right] \geqslant 0 \qquad (46) \end{aligned}$$

The last step follows from the monotonicity and non-negativity of the sine function on $[0, \pi/2]$ and so $\sin(\phi) \geqslant \sin(\theta)$ for $\phi \geqslant \theta \geqslant 0$, $\theta \in [0, \pi/2]$. It follows now that the numerator of (42) is monotonic increasing with $\phi \in [0, \pi/2]$ and so the whole expression (41) takes its maximum value of 1 when $\phi = \pi/2$. This completes the proof. □

Proof of Theorem 3.3. We use Theorem 3.1 with SRM on k. This allows us to select the value for k after seeing the sample. Let this value be $k_{\max} = \max_{n=1}^N k(x_n y_n, h)$. Hence it holds w.p. $1-\delta$, $\forall h \in \mathcal{H}$ that:

$$\begin{aligned} \Pr_{x,y}[h^T xy \leq 0] \leq \frac{1}{N}\sum_{n=1}^N \min(1, 2f_{k_{\max}}(\theta^h_{x_n y_n})) + \frac{2\sqrt{2}}{\sqrt{\pi}}\sqrt{\frac{k_{\max}}{N}} \\ +3\sqrt{\frac{\log(2/\delta)}{2N}} + 3\sqrt{\frac{\log(2)}{2}}\sqrt{\frac{k_{\max}}{N}} \end{aligned} \tag{47}$$

Due to Lemma 3.4, for any u, h, the function $\min(1, 2f_k(\theta_u^h))$ is non-increasing with k. Therefore the first term on the r.h.s. in Theorem 3.3 is a deterministic upper bound on the first term of eq. (47), and all the other terms are identical. □

3.2 Application to Uncovering New Connections Between Existing Algorithms

A desirable property for new theoretical results is an ability to provide a unifying context that can connect earlier work. In this section we show that the theorems presented in this section facilitate this.

3.2.1 Connection with the Large Margin Distribution Machine

We shall instantiate the bound of Theorem 3.3. The idea is to define the function $k = k(xy, h)$ in a way to lead to an analytic convex expression of $h/\|h\|$.

To do this, first we bound the $f_k(\theta)$ term using the analytic upper bound given in eq. (10), which is tight on the interval $\theta \in [-\pi/2, \pi/2)$, and is still a bound on $\theta \in [\pi/2, 3\pi/2]$.

$$\min(1, 2f_k(\theta)) \leq 2\exp\left\{-\frac{k\cos^2(\theta)\cdot \mathrm{sgn}(\cos(\theta))}{2}\right\} \tag{48}$$

For $k(\cdot)$ we choose the following:

$$k(xy, h) := \frac{2}{\left|\cos(\theta_{xy}^h)\right|} \tag{49}$$

Plugging this into Theorem 3.3 we obtain:

Corollary 3.4.1. *With probability $1-\delta$ w.r.t. the training set of size N, $\forall h \in \mathcal{H}$,*

$$\begin{aligned} Pr_{x,y}[h^T xy \leq 0] \quad &\leq \quad \frac{1}{N}\sum_{n=1}^{N} 2\exp\left(-\frac{h^T x_n y_n}{\|h\|\cdot\|x_n\|}\right) + \frac{4}{\sqrt{\pi}}\frac{1}{\sqrt{N}}\cdot\max_n \sqrt{\frac{\|h\|\cdot\|x_n\|}{|h^T x_n|}} \\ &+ \quad 3\sqrt{\frac{\log(2/\delta)}{2N}} + 3\sqrt{\frac{\log(2)}{N}}\cdot\max_n \sqrt{\frac{\|h\|\cdot\|x_n\|}{|h^T x_n|}} \end{aligned} \tag{50}$$

Observe that if we were to turn this bound into a minimization objective, we would minimize an exponential loss and maximize the minimum margin. That is, minimizing the exponential loss in fact minimizes a tight upper bound on the flipping probability.

Denote by $\gamma_n^h = \frac{h^T x_n y_n}{\|h\|\cdot\|x_n\|}$ the margin of the point x_n with respect to the hyperplane defined by h. Using the Taylor expansion for $\exp(\cdot)$ we can write:

$$\frac{1}{N}\sum_{n=1}^{N}\exp\left(-\frac{h^T x_n y_n}{\|h\|\cdot\|x_n\|}\right) = \frac{1}{N}\sum_{n=1}^{N}\exp\left(-\gamma_n^h\right) = 1 - \frac{1}{N}\sum_{n=1}^{N}\gamma_n^h + \frac{1}{N}\sum_{n=1}^{N}(\gamma_n^h)^2 - ...$$

Now, observe that the minimizer of this term in our generalization bound implies that the average of the empirical margin distribution is maximized and its second moment (so also the variance) is minimized. Hence, replacing the exponential term with its second order Taylor approximation in the bound of eq. (50) we obtain an objective that recovers the

recently proposed and successful method of Large Margin Distribution Machine (LDM) (Zhang & Zhou, 2014). Of course, the new loss function from our bound contains also the higher order moments, capturing the entire empirical distribution of margins.

Indeed, the LDM (Zhang & Zhou, 2014) was formulated as a quadratic objective, implemented in an efficient algorithm that maximizes the sample mean and minimizes the sample variance of the observed margin distribution, in addition to maximizing the minimum margin. Its original motivation was a boosting bound of Gao and Zhou (2013), given as a function of the average and 'some notion of' variance of the empirical margin distribution, derived by entirely different means. Via a completely different route, here we obtained a new explanation of LDM – namely as implementing an *approximate* minimizer of the bound in eq. (50).

3.3 An Empirical Assessment of Theorem 3.1

To assess the informativeness of our basic bound from Theorem 3.1, we create a classifier by minimizing the bound, without the ambition of a computationally efficient approach.

To be precise, our rationale is the following. A good theory should be capable of explaining essential characteristics of learning. The VC theory, despite its completeness (matching upper and lower bounds) is known to fall short of this desideratum, and it is therefore not suited to generate new algorithms. Margin theory is one remedy, which created the SVM. Since a RP-based analysis has an ability to capture low dimensional structures, and we have just seen how this can help explain new connections between different existing approaches, our aim in this section is to assess to what extent this theoretical finding is in tune with empirically observed behavior of classifiers.

We shall compare the bound-minimizing classifier with the most related existing algorithms: i) the gold-standard SVM – which maximizes the minimal margin and disregards other geometry and is therefore not robust against perturbations or outliers; ii) the direct zero-one loss minimizer (Nguyen & Sanner, 2013) – which in our framework corresponds to choosing a very large value for k; iii) the LDM (Zhang & Zhou, 2014) – which in our theory may be viewed as approximately minimizing the local version of our bound.

To this end, first we observe that so far it was sufficient for our purposes to consider h that goes through the origin, however in this context this is worth fine-tuning. The reason is that adding an intercept in the usual manner (by padding the inputs with a dummy feature of ones) when we work with the exact form of $f_k(\cdot)$, may not achieve the best achievable cosine values. Instead, we incorporate a new parameter vector z that allows us to slightly shift the data simultaneously with learning h to minimize the objective formed by the h-dependent terms of the bound. Thus, we minimize the following:

$$Obj(h,z) = \sum_{n=1}^{N} \frac{\Gamma(k)}{\left(\Gamma(k/2)\right)^2} \int_0^{\frac{1-a_n(h,z)}{1+a_n(h,z)}} \frac{v^{(k-2)/2}}{(1+v)^k} dv \tag{51}$$

where

$$a_n(h,z) := \cos(\theta^h_{(x_n-z)y_n}) = \frac{h^T}{\|h\|} \frac{(x_n-z)y_n}{\|x_n-z\|} \tag{52}$$

We initialize z in a low density region (e.g. the mid-point between data centres, z_0) and fine-tune it from the data in a small neighbourhood of z_0. Before doing so, let us

theoretically show that z is indeed learnable in this way. Replace x by $x - z$ throughout, and define the modified function class

$$\mathcal{F}_k^{shift}(h,z) = \left\{ x \to \frac{h^T}{\|h\|} \cdot \frac{(x-z)}{\|x-z\|} : h, z \in \mathbb{R}^d \right\} \tag{53}$$

Lemma 3.5. *Let $\mathcal{B}(z_0, r)$ be the Euclidean ball of radius r centred at z_0. Fix $\epsilon > 0$, and suppose we can choose $r > 0$ small enough so that $\sup_{z \in \mathcal{B}(z_0,r)} \frac{r}{\sqrt{N}} \sum_{n=1}^N \frac{1}{\|x_n - z\|} \leq \epsilon$. Then:*

$$\hat{\mathcal{R}}(\mathcal{F}_k^{shift}(h,z)) \leq \frac{1+\epsilon}{\sqrt{N}} \tag{54}$$

Note the technical condition on Lemma 3.5 does not depend on the class labels, and only requires a low-density region around the center of the input set.

Proof of Lemma 3.5. First, observe that for any fixed z_0, the empirical Rademacher complexity remains unchanged.

$$\hat{\mathcal{R}}(\mathcal{F}_k^{shift})(h, z = z_0) = \mathrm{E}_\sigma \sup_{h \in \mathbb{R}^d} \frac{1}{N} \left[\sum_{n=1}^N \sigma_n \frac{h^T}{\|h\|} \cdot \frac{(x_n - z_0)}{\|x_n - z_0\|} \right] = \frac{1}{\sqrt{N}}$$

Now, for any tolerance $\epsilon > 0$, by the mean value theorem and Cauchy-Schwartz, we have:

$$\begin{aligned} \hat{\mathcal{R}}(\mathcal{F}_k^{shift}(h,z)) &= \mathrm{E}_\sigma \sup_{h \in \mathbb{R}^d, z \in \mathcal{B}} \frac{1}{N} \left[\sum_{n=1}^N \sigma_n \frac{h^T}{\|h\|} \cdot \frac{(x_n - z)}{\|x_n - z\|} \right] \\ &\leq \mathrm{E}_\sigma \sup_{h,z} \|\Delta_z\| \cdot \|z - z_0\| + \hat{\mathcal{R}}(\mathcal{F}_k^{shift}(h, z_0)) \end{aligned} \tag{55}$$

where Δ_z is the gradient w.r.t. z of the function in the argument of the supremum, so:

$$\|\Delta_z\| = \left\| \frac{h^T}{\|h\|} \cdot \frac{1}{N} \sum_{n=1}^N \left(I_d - \frac{(x_n - z)(x_n - z)^T}{\|x_n - z\|^2} \right) \cdot \frac{\sigma_n}{\|x_n - z\|} \right\| \tag{56}$$

$$\leq \sup_{z \in \mathcal{B}(z_0,r)} \frac{1}{N} \sum_{n=1}^N \frac{1}{\|x_n - z\|} \tag{57}$$

$$\leq \frac{\epsilon}{r\sqrt{N}} \tag{58}$$

The inequality in eq. (57) is because

$$\lambda_{\max} \left(I_d - \frac{(x_n - z)(x_n - z)^T}{\|x_n - z\|^2} \right) = 1. \tag{59}$$

Plugging this back into eq. (55) and noting that $\|z - z_0\| \leq r$ gives:

$$\hat{\mathcal{R}}(\mathcal{F}_k^{shift}(h,z)) \leq \frac{\epsilon}{\sqrt{N}} + \hat{\mathcal{R}}(\mathcal{F}_k^{shift}(h, z_0)) = \frac{1+\epsilon}{\sqrt{N}}$$

□

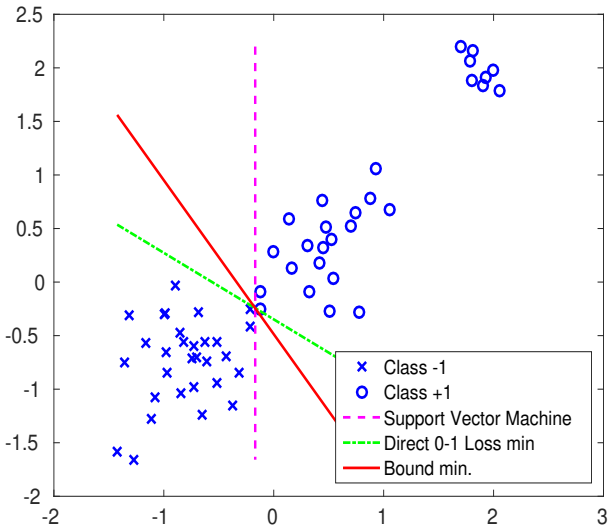

Figure 4: Illustration of our bound-optimizing classifier against the SVM and the zero-one loss minimizer. Note the very different separating planes obtained by each approach. In particular, the orientation of the SVM boundary is exclusively determined by the four support vectors.

Now we are ready to turn Theorem 3.1 into an algorithm. The gradients w.r.t. h and z are as follows:

$$\begin{aligned}
\Delta_h &= \frac{1}{N}\sum_{n=1}^{N} \ell_k'(a_n(h,z)) \cdot \frac{(x_n - z)y_n}{\|x_n - z\|} \cdot \left(I_d - \frac{hh^T}{\|h\|^2}\right) \frac{1}{\|h\|} \\
\Delta_z &= -\frac{1}{N}\sum_{n=1}^{N} \ell_k'(a_n(h,z)) \cdot \frac{hy_n}{\|h\|} \cdot \left(I_d - \frac{(x_n - z)(x_n - z)^T}{\|x_n - z\|^2}\right) \cdot \frac{1}{\|x_n - z\|}
\end{aligned}$$

where

$$\ell_k'(a_n(h,z)) = -\frac{\Gamma(k)}{2^{k-1}(\Gamma(k/2))^2}(1 - a_n(h,z)^2)^{\frac{k-2}{2}}. \tag{60}$$

We used numerical integration (Simpson quadrature, cf. MatLab's built-in function 'quad') to evaluate the objective, and a freely available generic nonlinear optimizer[4] that employs a combination of conjugate gradient and line search methods.

Figure 4 illustrates the classifier that results from optimizing our bound on toy data created to showcase the difference from SVM and the zero-one loss minimizer. It is most apparent that the bound-minimizing classifier is robust against unessential detail in the data, and captures the essential geometric structure. The difference from LDM was nearly indistinguishable on this data set, which agrees with the interpretation of LDM that we derived from our RP-based analyses in the previous subsection.

4. http://learning.eng.cam.ac.uk/carl/code/minimize/

Table 1: Mean of test error rates ± one standard error for our bound optimizer and some comparisons. We marked in bold font all statistically significant out-performances in test error over SVM at the 0.05 level using a paired t-test. Italic font in last two columns indicates performance was statistically significantly worse than SVM for the competing methods on the corresponding dataset.

Data set	N	d	Bound min.	SVM	Zero-one Loss	LDM
Australian	690	14	**0.137± 0.015**	0.148± 0.013	0.156±0.077	*0.149± 0.014*
German	1000	24	**0.260± 0.018**	0.280± 0.016	0.264±0.021	*0.315±0.015*
Haberman	306	3	**0.265± 0.025**	0.285± 0.050	0.268±0.024	0.276±0.030
Parkinsons	195	22	**0.141± 0.032**	0.221± 0.049	0.141± 0.036	0.135±0.034
PlRelax	182	12	**0.285± 0.029**	0.361± 0.166	*0.299±0.035*	0.290±0.051
Sonar	208	60	0.256± 0.045	0.271± 0.036	0.245±0.044	0.264±0.044

Further experimental tests on UCI data sets (Dua & Graff, 2017) are presented in Table 1. For each data set we performed 50 independent splits into two halves. Our parameter k, and SVM's C parameter were set by 5-fold cross-validation on the training half. The error rates on the held-out testing half of the data are reported in Table 1 in comparison with those of SVM (linear kernel). Despite the non-convex optimization involved in optimizing our bound, we observe improved generalization performance in 5 out of 6 data sets tested, which were statistically significant at the 0.05 level using a paired t-test. On the Sonar data set no statistically significant differences have been observed at the 0.05 level.

The statistically significant improvements achieved may or may not be practically significant – this would of course depend on the application domain, and the computing resources, as a direct minimization of our bound is far less efficient than the fast QP solver based LDM. However, what these results demonstrate is that the RP-based theory presented in the earlier sections does capture some essential characteristics that govern generalization, the obtained generalization bound is structure-aware and informative, its predictions are in agreement with practical experience.

3.3.1 Beyond Linear Classification: Connecting Two Views of Boosting

In this section we demonstrate the applicability of the ideas and techniques presented so far to a nonlinear situation: a linearly weighted ensemble of binary valued base learners from the function class $B = \{b : \mathcal{X} \to \{-1, 1\}\}$, with weights $\alpha = (\alpha_1, \alpha_2, ..., \alpha_T)$:

$$F_{ens} = \left\{ x \to \sum_{t=1}^{T} \alpha_t b_t(x) : b_t \in B, \sum_{i=1}^{T} |\alpha_i| \leq 1 \right\} \tag{61}$$

such as a boosting ensemble, which is known to be a very successful ensemble technique in practice (Opitz & Maclin, 1999). The bounds derived so far can be adapted to this class simply by replacing the empirical Rademacher complexity term of the unit-norm linear function class with that of the base-learner.

By adapting our Theorem 3.1, we obtain the following as a corollary.

Corollary 3.5.1. *Fix any $k(\leq T)$ positive integer, and $\delta > 0$. With probability $1-\delta$ w.r.t. the training set of size N, uniformly for all $\alpha_t, \sum_{t=1}^{T} |\alpha_t| \leq 1$ and all $b_t \in B, t = 1, ..., T$,*

$$Pr_{x,y}[\sum_{t=1}^{T} \alpha_t b_t(x) y \leq 0] \leq \frac{1}{N} \sum_{n=1}^{N} \min\left(1, 2 f_k(\theta^{\alpha}_{b(x_n) y_n})\right) + c\sqrt{\frac{k \cdot V(B)}{N}} + 3\sqrt{\frac{\log(2/\delta)}{2N}} \tag{62}$$

where V denotes VC-dimension, and $c > 0$ is a universal constant.

If we regard k as the inverse of a margin parameter, then Corollary 3.5.1 is analogous to the empirical margin distribution bound on boosting (Schapire, Freund, Bartlett, & Lee, 1998), derived by very different means, which gave rise to the classic margin-based explanation for the performance of boosting.

Before giving the proof, let us obtain another classical view of boosting, namely that of loss minimization (Schapire, 2013), from the same principle, by adapting our Theorem 3.3. While both views of boosting have coexisted for a long time, to our knowledge there has not previously been a single generic principle to connect them.

Indeed, applying Theorem 3.3 with the choice

$$k(h, b(x)y) := \frac{2\|b(x)\|_2}{|\cos(\theta^{\alpha}_{b(x)y})|} \cdot \frac{\|\alpha\|_2}{\|\alpha\|_1}, \tag{63}$$

where b is the vector of binary predictions $(b_t)_{t=1,...,T}$ we get the following:

Corollary 3.5.2. *With probability $1-\delta$ w.r.t. the training set of size N, uniformly for all $\alpha_t, \sum_{t=1}^{T} |\alpha_t| \leq 1$ and all $b_t \in B, t = 1, ..., T$,*

$$\begin{aligned} Pr_{x,y}[\alpha^T b(x) y \leq 0] \quad &\leq \quad \frac{1}{N} \sum_{n=1}^{N} 2 \exp\left(-\frac{\alpha^T b(x_n) y_n}{\|\alpha\|_1}\right) + 3\sqrt{\frac{\log(2/\delta)}{2N}} \\ &+ \quad \left(c\sqrt{\frac{V(B)}{N}} + 3\sqrt{\frac{\log(2)}{2N}}\right) \sqrt{2T} \max_n \sqrt{\frac{\|\alpha\|_1}{|\alpha^T b(x_n)|}} \end{aligned} \tag{64}$$

The dependence on T comes from $\|b(x)\|_2 = \sqrt{T}$ that enters in cosine-margins.

The point here is that, looking at the r.h.s. of the bound of Corollary 3.5.2 *as a minimization objective*, we recognize the first term is the well-known exponential loss of adaboost, and the last term contains the inverse of the minimum margin – together these recover a regularized adaboost (Schapire, 2013). Thus, these two different views of boosting can now be understood as manifestations of the same principle, more precisely a requirement of robustness to perturbations created by random projections.

Proof of Corollary 3.5.1. We apply Theorem 3.1 to the linear-convex aggregation in F_{ens}, but since here the inputs into this aggregation are the outputs of the base classifiers learned from the data, we need to replace the empirical Rademacher complexity contained in that bound with the following:

$$\begin{aligned} \hat{\mathcal{R}}_N(F_{ens}) \quad &= \quad \frac{1}{N} \mathrm{E}_\sigma \sup_{\alpha, b} \sum_{n=1}^{N} \sigma_n \frac{\alpha^T b(x_n)}{\|\alpha\| \cdot \|b(x_n)\|} \\ &= \quad \sup_{\alpha} \frac{\|\alpha\|_1}{\|\alpha\|_2} \frac{1}{N} \mathrm{E}_\sigma \sup_{\alpha, b} \sum_{n=1}^{N} \sigma_n \frac{\alpha^T}{\|\alpha\|_1} \cdot \frac{1}{\|b(x_n)\|_2} b(x_n) \end{aligned} \tag{65}$$

Since $b(x_n) \in \{-1,1\}^T$, we have $\|b(x_n)\|_2 = \sqrt{T}, \forall x_n$. We also have $\frac{\|\alpha\|_1}{\|\alpha\|_2} \leq \sqrt{T}$. Therefore, the elements of F_{ens} belong to the absolute convex hull of B, so by a classical result (Boucheron, Bousquet, & Lugosi, 2005, Theorem 3.3.) we have the r.h.s. of eq. (65) upper bounded by:

$$\frac{1}{N}\mathrm{E}_\sigma \sup_b \sum_{n=1}^{N} \sigma_n b(x_n) = \hat{\mathcal{R}}_N(B) \leq c\sqrt{\frac{V(B)}{N}} \tag{66}$$

for some absolute constant c. The last inequality is a known link between the Rademacher complexities and VC dimension (Koltchinskii & Panchenko, 2002).

□

Proof of Corollary 3.5.2. We apply Theorem 3.3, replacing the empirical Rademacher complexity, and we note that:

$$\sqrt{k(x_n y_n, h)} \quad = \quad \sqrt{\frac{2\|b(x_n)\|}{|\cos(\theta^{\alpha}_{b(x_n)y_n})|} \cdot \frac{\|\alpha\|_2}{\|\alpha\|_1}} \leq \sqrt{\frac{2\|\alpha\|_1}{|\alpha^T b(x_n)|}} \cdot \sqrt{T} \tag{67}$$

since $\frac{\|\alpha\|_2}{\|\alpha\|_1} \leq 1$, and $\|b(x_n)\|_2 = \sqrt{T}$. Multiplying together eqs. (67) and (65) completes the proof.

□

4. Conclusions

We proved risk bounds for halfspace learning, which automatically tighten in the presence of compressible structure. We addressed both learning from compressive data as well as learning from the original high-dimensional data, with generalization error guarantees. In particular we demonstrated that insights gained from studying the ERM classifier on randomly projected data can be employed to develop novel uniform bounds in the full high-dimensional space. In this context, the probability of a point flipping across the boundary can also be viewed as a Lipschitz approximation of the zero-one loss, which keeps the scale invariance of the zero-one loss, while at the same time it gains us access to the effects of small perturbations on low complexity sets. We extensively discussed several implications of our results, to draw connections between some existing successful classification approaches, including the computationally efficient LDM, and two different explanations of boosting. We also demonstrated the informativeness of our bounds empirically in simulation experiments.

Our main focus was linear classification, although we have also shown an extension to boosting ensembles. For nonlinear problems one could also, in principle, replace x with its feature space representation $\phi(x)$ induced by a fixed choice of kernel s.t. $K(x_1, x_2) = \phi(x_1)^T\phi(x_2)$, and the bounds we presented for the linear models here would then hold in unchanged form for the kernel-based function class. Of course various quantities will change under the mapping $\phi(\cdot)$ relative to the original data, and such an approach sacrifices some practical advantages of our bounds, and addressing this remains for future work. Another interesting avenue is to extend the approach of this analysis to other function classes and more complex models.

Acknowledgements

AK is funded by EPSRC Fellowship EP/P004245/1 "FORGING: Fortuitous Geometries and Compressive Learning". Part of this paper was written while RJD was visiting the University of Birmingham on sabbatical – RJD acknowledges financial support from the University of Waikato that made this visit possible. We thank Dehua Xu (MSc student at Birmingham) for early work on the implementation of the algorithm in Section 3.3.

Appendix A. Proof of Lemma 2.2

Without loss of generality (w.l.o.g.) we can take $\|h\| = \|x\| = 1$.

Part (a). Rewrite the probability of interest as the following:

$$\Pr_R\left\{h^T R^T R x \leq 0\right\} = \Pr_R\left\{\|R(h+x)\|^2 - \|R(h-x)\|^2 \leq 0\right\} \tag{68}$$

Now, observe that the two terms $\|R(h+x)\|^2$ and $\|R(h-x)\|^2$ are statistically independent. Indeed, each component of the random vectors $R(h+x)$ and $R(h-x)$ are Gaussian distributed, and denoting by R_i the i-th row of R, it is easy to verify that

$$\mathrm{Cov}(R_i(h+x), R_i(h-x)) = \|h\|^2 - \|x\|^2 = 0$$

since w.l.o.g. we assumed $\|h\| = \|x\| = 1$. Likewise,

$$\mathrm{Cov}(R_i(h+x), R_j(h-x))_{i\neq j} = 0$$

by the independence of the rows of R.

The variances are $\mathrm{Var}(R(h+x)) = \|h+x\|^2$ and $\mathrm{Var}(R(h-x)) = \|h-x\|^2$. Hence, denoting

$$U^2 := \left\| R\frac{h+x}{\|h+x\|} \right\|^2; \quad V^2 := \left\| R\frac{h-x}{\|h-x\|} \right\|^2 \tag{69}$$

these are independent standard χ^2 variables. Therefore, we may rewrite eq. (68) as the following:

$$\begin{aligned} \Pr_R\left\{h^T R^T R x \leq 0\right\} &= \Pr_{U^2,V^2}\left\{U^2\|h+x\|^2 < V^2\|h-x\|^2\right\} \\ &= \Pr_{U^2,V^2}\left\{\frac{U^2}{V^2} < \frac{\|h-x\|^2}{\|h+x\|^2}\right\} \end{aligned} \tag{70}$$

where the fraction on the l.h.s. is F-distributed. Further, observe that the r.h.s. is

$$\frac{\|h-x\|^2}{\|h+x\|^2} = \frac{2-2h^T x}{2+2h^T x} = \frac{1-\cos(\theta)}{1+\cos(\theta)} \tag{71}$$

since $\|h\| = \|x\| = 1$. Denoting this value by ψ, the integral of the cumulative density function of the F-distribution at ψ gives us the result stated in eq. (8).

The upper bound in eq. (10) follows from convex geometry (Ball, 1997) after noticing the geometric interpretation of eq. (8) as the ratio between the area of a hyperspherical cap

with angle 2θ and the surface area of the corresponding sphere when $\cos(\theta) > 0$ (Durrant & Kabán, 2013). Instead, below we give an elementary proof.

It is sufficient to consider the case $\cos(\theta) > 0$, and the case $\cos(\theta) \leq 0$ follows in the same way by symmetry. When $\cos(\theta) > 0$, we rewrite the r.h.s. of eq. (70) as the following:

$$\begin{aligned}\Pr\{(Rh)^T Rx \leq 0\} &= \Pr_{U^2,V^2}\left\{-(\cos(\theta)+1)U^2 - (\cos(\theta)-1)V^2 > 0\right\} \\ &\leq \mathrm{E}[\exp\left\{-\lambda(\cos(\theta)+1)U^2 - \lambda(\cos(\theta)-1)V^2)\right\}] \\ &= (1+2\lambda(\cos(\theta)+1))^{-k/2}(1+2\lambda(\cos(\theta)-1))^{-k/2} \qquad (72)\end{aligned}$$

for all $\lambda > 0$ such that $1 + 2\lambda(\cos(\theta) - 1) > 0$. In the last line we used that U^2 and V^2 are independent χ^2 variables.

After straightforward algebra, the r.h.s. of eq. (72) equals:

$$(1 + 4\lambda\cos(\theta) - 4\lambda^2\sin^2(\theta))^{-k/2}$$

minimizing this w.r.t. λ gives that the optimal λ satisfies $2\sin^2(\theta)\lambda = \cos(\theta)$.

So, if $\theta \neq 0$ then $\lambda = \frac{\cos(\theta)}{2\sin^2(\theta)}$ – which satisfies the condition required above. In turn, if $\theta = 0$ then the probability of interest is trivially 0 so the upper-bound we derive holds in both cases.

Plugging back, after cancellations we get for the case $h^T x > 0$ that:

$$\begin{aligned}\Pr\{(Rh)^T Rx \leq 0\} &\leq \left(1 + \frac{\cos^2(\theta)}{\sin^2(\theta)}\right)^{-k/2} \\ &= (\sin^2(\theta))^{k/2} \\ &= (1-\cos^2(\theta))^{k/2} \\ &\leq \exp\left\{-\frac{k}{2}\cos^2(\theta)\right\} \qquad \square\end{aligned}$$

Part (b). We start by rewriting as in eq. (68). But now the two quadratic terms are not independent in general (albeit they are uncorrelated), since these are now non-Gaussian – hence we need to purse a different strategy.

Rewrite eq. (68) by inserting the following expression, which evaluates to zero:

$$k\sigma^2\|h+x\|^2(1-\cos(\theta)) - k\sigma^2\|h-x\|^2(1+\cos(\theta)) \qquad (73)$$

Indeed, it is easy to check that this expression equals $k\sigma^2(4h^T x - 4\cos(\theta)) = 0$ because $\|h\| = \|x\| = 1$.

We insert eq. (73) into eq. (68):

$$\begin{aligned}\Pr\big\{&-[\|R(h+x)\|^2 - k\sigma^2\|h+x\|^2(1-\cos(\theta))] + ... \\ &[\|R(h-x)\|^2 - k\sigma^2\|h-x\|^2(1+\cos(\theta))] > 0 | \cos(\theta) > 0\big\} \qquad (74)\end{aligned}$$

Exponentiating both sides, and employing Markov inequality, for any $\lambda > 0$ the r.h.s. of eq. (74) is upper bounded by:

$$\begin{aligned}\mathrm{E}\big[\exp\big\{&-\lambda\left(\|R(h+x)\|^2 - k\sigma^2\|h+x\|^2(1-\cos(\theta))\right) + ... \\ &\lambda\left(\|R(h-x)\|^2 - k\sigma^2\|h-x\|^2(1+\cos(\theta))\right)\big\} \,|\, \cos(\theta) > 0\big] \qquad (75)\end{aligned}$$

Next, we introduce a convex combination which will serve us to exploit the convexity of the exponential function. For any $\alpha \in (0,1)$, eq. (75) equals:

$$\begin{aligned} &= \mathrm{E}\left[\exp\left\{-\alpha\frac{\lambda}{\alpha}\left(\|R(h+x)\|^2 - k\sigma^2\|h+x\|^2(1-\cos(\theta))\right) + ...\right.\right. \\ &\quad \left.\left.(1-\alpha)\frac{\lambda}{1-\alpha}\left(\|R(h-x)\|^2 - k\sigma^2\|h-x\|^2(1+\cos(\theta))\right)\right\} \mid \cos(\theta) > 0\right] \\ &\leq \alpha\mathrm{E}\left[\exp\left\{-\frac{\lambda}{\alpha}[\|R(h+x)\|^2 - k\sigma^2\|h+x\|^2(1-\cos(\theta))]\right\} \mid \cos(\theta) > 0\right] + ... \\ &\quad (1-\alpha)\mathrm{E}\left[\exp\left\{\frac{\lambda}{1-\alpha}[\|R(h-x)\|^2 - k\sigma^2\|h-x\|^2(1+\cos(\theta))]\right\} \mid \cos(\theta) > 0\right] \end{aligned}$$

where we used Jensen's inequality in the last line.

Now, $\lambda_1 := \frac{\lambda}{\alpha}$ and $\lambda_2 := \frac{\lambda}{1-\alpha}$ are two free parameters each of which may be optimized independently because we can take $\lambda = 1/(1/\lambda_1 + 1/\lambda_2)$. Further, since R was subgaussian, we have two sub-exponential moment generating functions in eq. (75) that are identical to those that appear in the proof of the two sides of the Johnson-Lindenstrauss lemma (JLL) by Dasgupta and Gupta (2002), but now with $\epsilon := \cos(\theta) \in (0,1)$ playing the role of the distortion parameter. Hence, by the same arguments as in JLL, both expectations above are upper-bounded by $\exp(-k\epsilon^2/8) = \exp(-k\cos^2(\theta)/8)$. So we obtain the upper bound:

$$\alpha\exp(-k\cos^2(\theta)/8) + (1-\alpha)\exp(-k\cos^2(\theta)/8) = \exp(-k\cos^2(\theta)/8) \quad \square$$

References

Abramowitz, M., & Stegun, I. (1965). *Handbook of Mathematical Functions*. Dover Publications.

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, *66*(4), 671–687.

Amelunxen, D., Lotz, M., McCoy, M., & Tropp, J. (2014). Living on the edge: phase transitions in convex programs with random data. *Inf. Inference*, *3*(3), 224–294.

Arriaga, R., & Vempala, S. (2006). An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, *63*(2), 161–182.

Arriaga, R., & Vempala, S. (1999). An algorithmic theory of learning: Robust concepts and random projection. In *40th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 616–623.

Ball, K. (1997). An elementary introduction to modern convex geometry. *Flavors of Geometry*, *31*, 1–58.

Bandeira, A. S., Mixon, D. G., & Recht, B. (2017). *Compressive classification and the rare eclipse problem*, pp. 197–220. Compressed Sensing and its Applications. Applied and Numerical Harmonic Analysis. Birkhäuser, Cham.

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

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2019). Benign overfitting in linear regression. *CoRR*, *abs/1906.11300*.

Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: Probability & Statistics*, *9*, 323–375.

Bălcan, M. F., Blum, A., & Vempala, S. (2006). Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, *65*, 79–94.

Calderbank, R., Jafarpour, S., & Schapire, R. (2009). Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical Report, Rice University.

Dasgupta, A., & Gupta, A. (2002). An elementary proof of the Johnson-Lindenstrauss lemma. *Random Structures & Algorithms*, *22*, 60–65.

Devroye, L., & Lugosi, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognition*, *28*, 1011–1018.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, *52*(4), 1289–1306.

Dua, D., & Graff, C. (2017). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Durrant, R. J., & Kabán, A. (2010). Compressed fisher linear discriminant analysis: Classification of randomly projected data. In *Proc. 16th ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1119–1128.

Durrant, R. J., & Kabán, A. (2013). Sharp generalization error bounds for randomly-projected classifiers. In *Proc. 30-th International Conference on Machine Learning (ICML)*, pp. 693–701, Journal of Machine Learning Research W&CP 28(3).

Fard, M. M., Grinberg, Y., Pineau, J., & Precup, D. (2012). Compressed least-squares regression on sparse spaces. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, AAAI'12, pp. 1054–1060. AAAI Press.

Gao, W., & Zhou, Z.-H. (2013). On the doubt about margin explanation of boosting. *Artificial Intelligence*, *203*, 1–18.

Garg, A., Har-Peled, S., & Roth, D. (2002). On generalization bounds, projection profile and margin distribution. In *Proc. 19th International Conference on Machine Learning (ICML)*, pp. 171–178.

Garg, A., & Roth, D. (2003). Margin distribution and learning algorithms. In *Proc. 20th International Conference on Machine Learning (ICML)*, pp. 210–217.

Goel, N., Bebis, G., & Nefian, A. (2005). Face recognition experiments with random projection. In Jain, A. K., & Ratha, N. K. (Eds.), *Biometric Technology for Human Identification II*, Vol. 5779, pp. 426–437.

Johnson, W., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, *26*, 189–206.

Kabán, A. (2015). Improved bounds on the dot product under random projection and random sign projection. In *Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 487–496.

Kabán, A. (2019). Dimension-free error bounds from random projections.. In *33rd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4049–4056.

Kabán, A. (2014). New bounds on compressive linear least squares regression. In Kaski, S., & Corander, J. (Eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, Vol. 33 of *Proceedings of Machine Learning Research*, pp. 448–456, Reykjavik, Iceland. PMLR.

Kawaguchi, K., Kaelbling, L. P., & Bengio, Y. (2019). Generalization in deep learning..

Klartag, B., & Mendelson, S. (2005). Empirical processes and random projections. *Journal of Functional Analysis*, *225*(1), 229–245.

Koltchinskii, V., & Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, *30*, 1–50.

Liaw, C., Mehrabian, A., Plan, Y., & Vershynin, R. (2017). *A simple tool for bounding the deviation of random matrices on geometric sets*, pp. 277–299. Springer, Berlin.

Matoušek, J. (2008). On variants of the Johnson-Lindenstrauss lemma. *Random Structures & Algorithms*, *33*(2), 142–156.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Nagarajan, V., & Kolter, J. Z. (2019). Uniform convergence may be unable to explain generalization in deep learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32*, pp. 11615–11626. Curran Associates, Inc.

Negrea, J., Dziugaite, G., & Roy, D. M. (2019). In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. *ArXiv*, *abs/1912.04265*.

Nguyen, T., & Sanner, S. (2013). Algorithms for direct 0–1 loss optimization in binary classification. *Proc. 30-th International Conference on Machine Learning (ICML)*, *28*(3), 1085–1093.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*(1), 169–198.

Plan, Y., & Vershynin, R. (2013). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, *59*, 482–494.

Reboredo, H., Renna, F., Calderbank, R., & Rodrigues, M. R. D. (2016). Bounds on the number of measurements for reliable compressive classification. *IEEE Transactions on Signal Processing*, *64*(22), 5778–5793.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference*, pp. 37–52. Springer, Berlin.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W.-S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*(5), 1651–1686.

Shawe-Taylor, J., & Cristianini, N. (1999). Margin distribution bounds on generalization. In *Proc. European Conference on Computational Learning Theory (EuroCOLT)*, pp. 263–273.

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

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

Wendel, J. G. (1948). Note on the gamma function. *American Math Monthly*, *55*, 563–564.

Xie, H., Li, J., Zhang, Q., & Wang, Y. (2016). Comparison among dimensionality reduction techniques based on random projection for cancer classification. *Computational Biology and Chemistry*, *65*, 165–172.

Zhang, T., & Zhou, Z. (2014). Large margin distribution machine. In *Proc. 20-th SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 313–322.