Sharp Generalization Error Bounds for Randomly-projected Classifiers

Robert J. Durrant R.J.DURRANT@CS.BHAM.AC.UK
School of Computer Science, University of Birmingham, Edgbaston, UK, B15 2TT

Ata Kabán A.KABAN@CS.BHAM.AC.UK
School of Computer Science, University of Birmingham, Edgbaston, UK, B15 2TT

Abstract

We derive sharp bounds on the generalization error of a generic linear classifier trained by empirical risk minimization on randomly-projected data. We make no restrictive assumptions (such as sparsity or separability) on the data: Instead we use the fact that, in a classification setting, the question of interest is really 'what is the effect of random projection on the predicted class labels?' and we therefore derive the exact probability of 'label flipping' under Gaussian random projection in order to quantify this effect precisely in our bounds.

1. Introduction

Random projection is fast becoming a workhorse in high dimensional learning (e.g. Boyali & Kavakli, 2012; Fard et al., 2012; Mahoney, 2011; Maillard & Munos, 2012; Paul et al., 2012; Pillai et al., 2011). However, except in a few specific settings, little is known about its effect on the generalization performance of a classifier.

Previous work quantifying the generalization error of a linear classifier trained on randomly projected data has, to the best of our knowledge, only considered specific families of classifiers and each approach previously employed has also assumed constraints of some form on the data. The earliest work is in a seminal paper by Arriaga & Vempala (1999), where the effect of randomly projecting well-separated data on the performance of the Perceptron is quantified. However the bounds in Arriaga & Vempala (1999) make use of high-probability geometry preservation guarantees via the Johnson-Lindenstrauss lemma (JLL) and therefore, contrary to expectation and experience, they become looser as the sample complexity increases. More recently, Calderbank et al. (2009) gave guarantees for SVM working with randomly projected *sparse* data using ideas from the field of compressed sensing (CS) - these however become looser as the number of non-zero features in the sparse representation of the data increases. Generative classifiers are considered in Davenport et al. (2010) where Neyman-Pearson detector was analyzed assuming spherical Gaussian classes, while Durrant & Kabán (2010); Durrant & Kabán (2011) considered Fisher's Linear Discriminant, assuming general sub-Gaussian classes. These bounds tighten with the sample complexity, but the assumptions on the class-conditional distributions may not hold in practice.

Along very different lines, Garg et al. (2002) use random projections to estimate the generalization error of a classifier learnt in the original data space, i.e. learning is not in the randomly projected domain, but random projections are used instead as a tool for deriving their generalization bounds. They have the nice idea, which we will also use, of quantifying the effect of random projection by how it changes class labels of projected points w.r.t. to the data space classifier. Their approach yields a data-dependent term that captures the margin distribution, and allows the use of existing VC-dimension bounds in the low dimensional space. However, although their result improves on previous margin bounds, it is still generally trivial (the probability of misclassification obtained is greater than 1). This is mainly because their estimate of how likely a class label is to be 'flipped' with respect to its label in the data space is extremely loose; in fact its contribution to the generalization error bound is typically greater than 1 and it never attains its true value. Here

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

we turn around the approach in Garg et al. (2002) in order to derive bounds for the generalization error of generic linear classifiers learnt by empirical risk minimization (ERM) from randomly-projected data. Moreover, instead of using bounds on the label-flipping probability (i.e. the margin distribution) as obtained in Garg et al. (2002) or Garg & Roth (2003), we derive the exact form of this quantity. Finally, we show that one can sometimes improve on their use of Markov inequality by Chernoff-bounding the dependent sum, and gain some additional improvement[1]. As a consequence we obtain non-trivial bounds on the generalization error of the randomly-projected classifier, which we note can also be extended to improve the results in Garg et al. (2002) in a straightforward way.

2. Preliminaries

2.1. The Classification Problem

We consider a 2-class classification problem where we observe N examples of labelled training data $\mathcal{T}^N = \{(x_i, y_i)\}_{i=1}^N$ where (x_i, y_i) drawn i.i.d from an unknown data distribution $\mathcal{D}$ over $\mathbb{R}^d \times \{0,1\}$. For a given class of functions $\mathcal{H}$, our goal is to learn from $\mathcal{T}^N$ the classification function $\hat{h} \in \mathcal{H}$ with the lowest possible *generalization error* in terms of some loss function $\mathcal{L}$. That is, find $\hat{h}$ such that $\mathcal{L}(\hat{h}(x_q), y_q) = \arg\min_{h\in\mathcal{H}} \mathrm{E}_{x_q,y_q}[\mathcal{L}(h(x_q), y_q)]$, where $(x_q, y_q) \sim \mathcal{D}$ is a query point with unknown label y_q.

Here we use the $(0,1)$-loss $\mathcal{L}_{(0,1)} : \{0,1\} \times \{0,1\} \to \{0,1\}$ which is the measure of performance of interest in classification, defined by:

$$\mathcal{L}_{(0,1)}(\hat{h}(x_q), y_q) = \begin{cases} 0 & \text{if } \hat{h}(x_q) = y_q \\ 1 & \text{otherwise.} \end{cases}$$

Working with the original data, the learned classifier $\hat{h}$ is a vector in $\mathbb{R}^d$ which, without loss of generality, we take to pass through the origin. For an unlabelled query point x_q the label returned by $\hat{h}$ is then:

$$\mathbf{1}\left\{\hat{h}^T x_q > 0\right\}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function which returns 1 if its argument is true and 0 otherwise. Since we are only interested in the sign of the dot product above, we may clearly assume without loss of generality that in the data space all data lie on the unit sphere $S^{d-1} \subseteq \mathbb{R}^d$ and that $\|\hat{h}\| = 1$, where $\|\cdot\|$ denotes the Euclidean norm.

[1] Our approach is numerically tighter when the confidence parameter δ in our bound is chosen to be small.

Now consider the case when d is very large and, for practical reasons, we would like to work with a lower dimensional representation of the data. There are many methods for carrying out such dimensionality reduction (see e.g. Fodor, 2002, for a survey) but here we focus on *random projection* which is a recent and very promising data-independent approach. Randomly projecting the data consists of simply left multiplying the data with a random matrix $R \in \mathcal{M}_{k\times d}$, $k \ll d$, where R has entries r_{ij} drawn i.i.d from a zero-mean subgaussian distribution. Again many matrices fit this bill – examples can be found in Achlioptas (2003); Dasgupta & Gupta (2002); Ailon & Chazelle (2006) and Matoušek (2008) – but for concreteness and analytical tractability we will focus here on matrices R where the entries $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

We are interested in quantifying the effect on the generalization error of randomly projecting the training set to a k-dimensional subspace, $k \ll d$, and learning the classifier there instead of in the original data space. In this setting, the training set now consists of instances of randomly-projected data $\mathcal{T}_R^N = \{(Rx_i, y_i)\}_{i=1}^N$, and the learned classifier is now a vector in $\mathbb{R}^k$ (possibly not through the origin - translation does not affect our proof technique) which we will denote by $\hat{h}_R$. The label returned by $\hat{h}_R$ is therefore:

$$\mathbf{1}\left\{\hat{h}_R^T R x_q + b > 0\right\}$$

where $b \in \mathbb{R}$. Denoting by $\hat{h}_R(Rx_q)$ the label returned by this classifier, we want to estimate:

$$\mathrm{E}_{x_q,y_q}\left[\mathcal{L}_{(0,1)}(\hat{h}_R(Rx_q), y_q)\right] = \Pr_{x_q,y_q}\left\{\hat{h}_R(Rx_q) \neq y_q\right\}$$

where $(x_q, y_q) \sim \mathcal{D}$ is a query point with unknown label y_q. To keep our results general we only assume that the data points are drawn i.i.d from $\mathcal{D}$, but we make no particular assumptions on the data distribution $\mathcal{D}$, in particular we make no assumption of a sparse data structure, nor do we assume that the classes are linearly separable.

3. Results

Our main result is the following bound on the generalization error of a classifier trained by ERM on the randomly projected data set:

Theorem 3.1 (Generalization Error). *Let $\mathcal{T}^N = \{(x_i, y_i) | x_i \in \mathbb{R}^d, y_i \in \{0,1\}\}_{i=1}^N$ be a set of d-dimensional labelled training examples of size N, and let $\hat{h}$ be the linear ERM classifier estimated from $\mathcal{T}^N$. Let $R \in \mathcal{M}_{k\times d}$, $k < d$ be a random projection matrix with entries $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. Denote by $\mathcal{T}_R^N =$*

$\{(Rx_i, y_i)\}_{i=1}^N$ the random projection of the training data $\mathcal{T}^N$, and let $\hat{h}_R$ be the linear classifier estimated from $\mathcal{T}_R^N$. Then for all $\delta \in (0,1]$, with probability at least $1-2\delta$ w.r.t. the random choice of $\mathcal{T}^N$ and R, the generalization error of $\hat{h}_R$ w.r.t the (0,1)-loss is bounded above by:

$$Pr_{x_q,y_q}\{\hat{h}_R(Rx_q) \neq y_q\} \leqslant \hat{E}(\mathcal{T}^N, \hat{h}) + \frac{1}{N}\sum_{i=1}^N f_k(\theta_i) + \min\left\{\sqrt{3\log\frac{1}{\delta}}\sqrt{\frac{1}{N}\sum_{i=1}^N f_k(\theta_i)}, \frac{1-\delta}{\delta}\cdot\frac{1}{N}\sum_{i=1}^N f_k(\theta_i)\right\} + 2\sqrt{\frac{(k+1)\log\frac{2eN}{k+1} + \log\frac{1}{\delta}}{N}} \tag{3.1}$$

where $f_k(\theta_i) := Pr_R\{sign(\hat{h}R^TRx_i) \neq sign(\hat{h}^Tx_i)\}$ is the flipping probability for the i-th training example with θ_i the principal angle between $\hat{h}$ and x_i, and $\hat{E}(\mathcal{T}^N, \hat{h}) = \frac{1}{N}\sum_{(x_i,y_i)\in\mathcal{T}^N}\mathcal{L}_{(0,1)}(\hat{h}(x_i), y_i)$ is the empirical risk of the data space classifier.

This theorem says that with high probability, the generalization error of any linear classifier trained on randomly projected data is upper bounded by the training error of the data space classifier plus the average flipping probabilities of the training points plus a 'projection-penalty' term, either $\sqrt{\frac{1}{N}\sum_{i=1}^N f_k(\theta_i)}\sqrt{3\log\frac{1}{\delta}}$ or $\frac{1-\delta}{\delta}\frac{1}{N}\sum_{i=1}^N f_k(\theta_i)$, plus the VC-complexity in the projection space. Notice that the terms involving flipping probabilities vanish when no flipping occurs and in particular, as $k \to d$, our bound recovers exactly the classical VC-bound for linear classifiers in $\mathbb{R}^d$. On the other hand, when $k < d$, these terms represent the bias of the classifier in the randomly projected domain and quantify the price paid for working there instead of in the data space.

Notice also that the average flipping probability term depends on the angles between the training points and the classifier; we therefore see from the geometry that when there is a large margin separating the classes this term will generally be small, and our bound captures well the effects of separated classes. On the other hand, a small average flipping probability is still possible even when the margin is small – for example provided that not too many points are close to the decision hyperplane (in other words, if the data are soft-separable with a large (soft) margin).

Finally we note that our theorem implies that we can get close to the best linear classifier in $\mathbb{R}^d$, but working in $\mathbb{R}^k$ and even with a relatively small sample complexity, provided that the data have some special structure which keeps this average flipping probability small (and we have already identified two such special structures). A key tool in obtaining theorem 3.1, which may also be of independent interest, is the following theorem 3.2:

Theorem 3.2 (Flipping Probability)**.** *Let $h, x \in \mathbb{R}^d$ and let the angle between them be $\theta \in [0, \pi/2]$. Without loss of generality take $\|h\| = \|x\| = 1$.*
Let $R \in \mathcal{M}_{k\times d}$, $k < d$, be a random projection matrix with entries $r_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ and let Rh, $Rx \in \mathbb{R}^k$ be the images of h, x under R with angular separation θ_R.

1. *Denote by $f_k(\theta)$ the 'flipping probability' $f_k(\theta) := Pr\{(Rh)^TRx < 0 | h^Tx > 0\}$. Then:*

$$f_k(\theta) = \frac{\Gamma(k)}{(\Gamma(k/2))^2}\int_0^\psi \frac{z^{(k-2)/2}}{(1+z)^k}dz \tag{3.2}$$

where $\psi = (1-\cos(\theta))/(1+\cos(\theta))$.

2. *The expression above can be rewritten as the quotient of the surface area of a hyperspherical cap with an angle of 2θ by the surface area of the corresponding hypersphere, namely:*

$$f_k(\theta) = \frac{\int_0^\theta \sin^{k-1}(\phi)\ d\phi}{\int_0^\pi \sin^{k-1}(\phi)\ d\phi} \tag{3.3}$$

3. *The flipping probability is monotonic decreasing as a function of k: Fix $\theta \in [0, \pi/2]$, then $f_k(\theta) \geqslant f_{k+1}(\theta)$.*

4. Proofs

4.1. Proof of Flipping Probability - Theorem 3.2

Let $h, x \in \mathbb{R}^d$ be two unit vectors[2] with the angle between them $\theta \in [0, \pi/2]$ which we randomly project by premultiplying them with a random matrix $R \in \mathcal{M}_{k\times d}$ with entries drawn i.i.d from the Gaussian $\mathcal{N}(0, \sigma^2)$ to obtain $Rh, Rx \in \mathbb{R}^k$ with the angle between them θ_R. As a consequence of the Johnson-Lindenstrauss lemma, the angle between the projected vectors Rh, Rx is approximately θ with high probability (see e.g. Arriaga & Vempala (1999)) and the images of the vectors h, x under the same random projection are *not* independent.

We want to find the probability that following random projection the angle between these vectors becomes $\theta_R > \pi/2$, i.e. switches from being acute to being obtuse. We call this probability the 'flipping probability'

[2] In the proof of our generalization error bound, h will be instantiated as the data space ERM classifier $\hat{h}$ and x as a single training point x_i.

because its effect is to 'flip' the predicted class label in the projected space w.r.t the data space from the 1 class to the 0 class. It is easy to see that this probability is symmetric in the class labels, e.g. by considering the angle of x with $-h$, and so *mutatis mutandis* the probability of flipping from the 0 class to the 1 class has the same form.

We will prove parts 1 & 2 of theorem 3.2 here. Part 3 of our theorem is easy to believe using part 2 and the fact that the proportion of the surface of the k-dimensional unit sphere covered by a spherical cap with angle of 2θ is bounded above by $\exp\left(-\frac{1}{2}k\cos^2(\theta)\right)$ (Ball, 1997, Lemma 2.2, Pg 11); to save space we omit a rigorous proof of part 3 – this can be found in Durrant (2013).

Before proving theorem 3.2 we make some preliminary observations. First note, from the definition of the dot product, for $\theta \in [0, \pi/2]$ in the original d-dimensional space and θ_R in the k-dimensional randomly-projected space we have $\Pr_R\{\theta_R > \pi/2\} = \Pr_R\{(Rh)^T Rx < 0\}$, and this is the probability of our interest. In fact the arguments for the proof of parts 1 & 2 of our theorem will not rely on the condition $\theta \in [0, \pi/2]$ - this is only needed for part 3. Regarding random Gaussian matrices we note that, for any non-zero vector $x \in \mathbb{R}^d$, the event $Rx = 0$ has probability zero with respect to the random choices of R. This is because the null space of R, $\ker(R) = R(\mathbb{R}^d)^\perp$, is a linear subspace of $\mathbb{R}^d$ with dimension $d - k < d$, and therefore $\ker(R)$ has zero Gaussian measure in $\mathbb{R}^d$. Hence $\Pr_R\{x \in \ker(R)\} = \Pr_R\{Rx = 0\} = 0$. Likewise, R almost surely has rank k. In this setting we may therefore safely assume that $h, x \notin \ker(R)$ and that R has rank k. With these details out of the way, we begin:

4.1.1. Proof of part 1.

First we expand out the terms of $(Rh)^T Rx$ to obtain $\Pr_R\{(Rh)^T Rx < 0\}$:

$$= \Pr_R\left\{\sum_{i=1}^{k}\left(\sum_{j=1}^{d} r_{ij}h_j\right)\left(\sum_{j=1}^{d} r_{ij}x_j\right) < 0\right\} \tag{4.1}$$

Recall that the entries of R are independent and identically distributed with $r_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$ and make the change of variables $u_i = \sum_{j=1}^{d} r_{ij}h_j$ and $v_i = \sum_{j=1}^{d} r_{ij}x_j$. A linear combination of Gaussian variables is again Gaussian, however u_i and v_i are now no longer independent since they both depend on the same row of R. On the other hand, for $i \neq j$ the vectors (u_i, v_i) and (u_j, v_j) are independent of each other since the i-th row of R is independent of its j-th row. Moreover $(u_i, v_i) \sim (u_j, v_j)$, $\forall i, j$ so it is enough to consider a single term of the outer sum in (4.1). We have:

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim \mathcal{N}\left(\mathrm{E}_R\left[\begin{pmatrix} u_i \\ v_i \end{pmatrix}\right], \mathrm{Cov}_R\left[\begin{pmatrix} u_i \\ v_i \end{pmatrix}\right]\right)$$

Since u_i and v_i are zero mean, the expectation of this distribution is just $(0, 0)^T$, and its covariance is:

$$\Sigma_{u,v} = \begin{bmatrix} \mathrm{Var}(u_i) & \mathrm{Cov}(u_i, v_i) \\ \mathrm{Cov}(u_i, v_i) & \mathrm{Var}(v_i) \end{bmatrix} \tag{4.2}$$

Then:

$$\begin{aligned} \mathrm{Var}(u_i) &= \mathrm{E}[(u_i - \mathrm{E}(u_i))^2] \\ &= \mathrm{E}[(u_i)^2] \text{ since } \mathrm{E}(u_i) = 0 \\ &= \sum_{\substack{j=1 \\ j'=1}}^{d} h_j h_{j'} \mathrm{E}\left[r_{ij} r_{ij'}\right] \end{aligned}$$

Now, when $j \neq j'$, r_{ij} and $r_{ij'}$ are independent, and so $\mathrm{E}[r_{ij}r_{ij'}] = \mathrm{E}[r_{ij}]\mathrm{E}[r_{ij'}] = 0$. On the other hand, when $j = j'$ we have $\mathrm{E}[r_{ij}r_{ij'}] = \mathrm{E}[r_{ij}^2] = \mathrm{Var}(r_{ij}) = \sigma^2$, since $r_{ij} \sim \mathcal{N}(0, \sigma^2)$. Hence:

$$\mathrm{Var}(u_i) = \sum_{j=1}^{d} \sigma^2 \hat{h}_j^2 = \sigma^2 \|h\|^2 = \sigma^2 \tag{4.3}$$

since $\|h\| = 1$. Likewise $\mathrm{Var}(v_i) = \sigma^2$.
Next the covariance $\mathrm{Cov}(u_i, v_i)$ is:

$$\begin{aligned} \mathrm{Cov}(u_i, v_i) &= \mathrm{E}\left[(u_i - \mathrm{E}[u_i])(v_i - \mathrm{E}[v_i])\right] = \mathrm{E}[u_i v_i] \\ &= \sum_{\substack{j=1 \\ j'=1}}^{d} h_j x_{j'} \mathrm{E}[r_{ij} r_{ij'}] \end{aligned} \tag{4.4}$$

Now, when $j \neq j'$ the expectation is zero, as before, and when $j = j'$ we have for (4.4):

$$= \sum_{j=1}^{d} h_j x_j \mathrm{E}[(r_{ij})^2] = \sum_{j=1}^{d} h_j x_j \mathrm{Var}(r_{ij}) = \sigma^2 h^T x \tag{4.5}$$

Hence for each $i \in \{1, \ldots, k\}$ the covariance matrix is:

$$\Sigma_{u,v} = \sigma^2 \begin{bmatrix} 1 & h^T x \\ h^T x & 1 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & \cos(\theta) \\ \cos(\theta) & 1 \end{bmatrix}$$

since $\|h\| = \|x\| = 1$, and we have $(u_i, v_i)^T \overset{\text{i.i.d}}{\sim} (0, \Sigma_{u,v})$. Now the probability in (4.1) can be written as:

$$\Pr\left\{\sum_{i=1}^{k} u_i v_i < 0\right\}$$

which it will be helpful to further rewrite as:

$$\Pr\left\{\sum_{i=1}^{k} (u_i, v_i) \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \begin{pmatrix} u_i \\ v_i \end{pmatrix} < 0\right\} \tag{4.6}$$

where the probability is now over the distribution of $(u_i, v_i)^T$. Making the final change of variables:

$$(y_i, z_i)^T = \Sigma_{u,v}^{-1/2}(u_i, v_i)^T \tag{4.7}$$

where the new variables y_i, z_i are independent unit variance spherical Gaussian variables, $(y_i, z_i)^T \overset{iid}{\sim} \mathcal{N}(0, I)$, we substitute into (4.6) to obtain the flip probability in the form:

$$\Pr\left\{\frac{1}{2}\sum_{i=1}^{k}(y_i, z_i)\,\Sigma_{u,v}^{1/2}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\Sigma_{u,v}^{1/2}\begin{pmatrix}y_i\\ z_i\end{pmatrix} < 0\right\} \tag{4.8}$$

where the probability now is w.r.t the standard Gaussian distribution. Now diagonalizing the symmetric matrix $\Sigma_{u,v}^{1/2}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\Sigma_{u,v}^{1/2}$ as $U\Lambda U^T$ with $UU^T = U^TU = I$ and Λ a diagonal matrix of its eigenvalues, we can rewrite (4.8) as:

$$\Pr\left\{\frac{1}{2}\sum_{i=1}^{k}(y_i, z_i)U\Lambda U^T\begin{pmatrix}y_i\\ z_i\end{pmatrix} < 0\right\} \tag{4.9}$$

The standard Gaussian distribution is invariant under orthogonal transformations, and so the form of U does not affect this probability. We can therefore take $U = I$ without loss of generality and rewrite (4.9) as:

$$\Pr\left\{\frac{1}{2}\sum_{i=1}^{k}(y_i, z_i)\Lambda\begin{pmatrix}y_i\\ z_i\end{pmatrix} < 0\right\}$$

Now we need the entries of Λ, which are the eigenvalues of:

$$\Sigma_{u,v}^{1/2}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\Sigma_{u,v}^{1/2}$$

Using the fact that the eigenvalues of AB are the same as the eigenvalues of BA these are the eigenvalues of

$$\sigma^2\begin{bmatrix}1 & \cos(\theta)\\ \cos(\theta) & 1\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} = \sigma^2\begin{bmatrix}\cos(\theta) & 1\\ 1 & \cos(\theta)\end{bmatrix}$$

which are $\lambda = \sigma^2(\cos(\theta) \pm 1)$. Substituting into the inequality (4.9) and dropping the positive scaling constant $\frac{1}{2}\sigma^2$ since it does not affect the sign of the left hand side, the probability we are after is:

$$\Pr\left\{\sum_{i=1}^{k}(y_i, z_i)^T\begin{bmatrix}\cos(\theta)+1 & 0\\ 0 & \cos(\theta)-1\end{bmatrix}\begin{pmatrix}y_i\\ z_i\end{pmatrix} < 0\right\}$$

$$= \Pr\left\{\sum_{i=1}^{k}((\cos(\theta)+1)y_i^2 + (\cos(\theta)-1)z_i^2) < 0\right\}$$

$$= \Pr\left\{(\cos(\theta)+1)\sum_{i=1}^{k}y_i^2 + (\cos(\theta)-1)\sum_{i=1}^{k}z_i^2 < 0\right\}$$

$$= \Pr\left\{\frac{\sum_{i=1}^{k}y_i^2}{\sum_{i=1}^{k}z_i^2} < \frac{1-\cos(\theta)}{1+\cos(\theta)}\right\} \tag{4.10}$$

Now, y_i and z_i are standard univariate Gaussian variables, hence $y_i^2, z_i^2 \overset{iid}{\sim} \chi^2$, and so the left hand side of (4.10) is F-distributed (Mardia et al., 1979, Appendix B.4, pg 487). Therefore:

$$\Pr_R\{(Rh)^TRx < 0\} = \frac{\Gamma(k)}{(\Gamma(k/2))^2}\int_0^{\psi}\frac{w^{(k-2)/2}}{(1+w)^k}\mathrm{d}w$$

where $\psi = (1-\cos(\theta))/(1+\cos(\theta))$ and $\Gamma(\cdot)$ is the gamma function. This proves the first part of Theorem 3.2. □

4.1.2. Proof of part 2.

Note that $\psi = \tan^2(\theta/2)$ and make the substitution $w = \tan^2(\theta/2)$. Then, from the trigonometric identity $\sin(\theta) = 2\tan(\theta)/(1+\tan^2(\theta))$ and $\frac{\mathrm{d}w}{\mathrm{d}\theta} = \tan(\theta/2)(1+\tan^2(\theta/2))$, we obtain:

$$f_k(\theta) = \frac{\Gamma(k)}{2^{k-1}(\Gamma(k/2))^2}\int_0^{\theta}\sin^{k-1}(\phi)\mathrm{d}\phi \tag{4.11}$$

To put the expression (4.11) in the form of the second part of the theorem, we need to show that the gamma term outside the integral is the reciprocal of $\int_0^{\pi}\sin^{k-1}(\phi)\mathrm{d}\phi$. This can be shown in a straightforward way using the beta function. Recall that the beta function is defined by (e.g. Abramowitz & Stegun, 1972, 6.2.2, pg 258):

$$\mathrm{B}(w, z) = \frac{\Gamma(w)\Gamma(z)}{\Gamma(w+z)}$$
$$= 2\int_0^{\pi/2}\sin^{2w-1}(\theta)\cos^{2z-1}(\theta)\mathrm{d}\theta,\ \mathrm{Re}(w), \mathrm{Re}(z) > 0 \tag{4.12}$$

and therefore from equation (4.12) we have:

$$\frac{1}{2}\mathrm{B}\left(\frac{k}{2}, \frac{1}{2}\right) = \int_0^{\pi/2}\sin^{k-1}(\theta)\mathrm{d}\theta$$

Next, from the symmetry of the sine function about $\pi/2$, equation (4.12), and using $\Gamma(1/2) = \sqrt{\pi}$ we have:

$$\int_0^{\pi}\sin^{k-1}(\theta)\mathrm{d}\theta = 2\int_0^{\pi/2}\sin^{k-1}(\theta)\mathrm{d}\theta$$
$$= \mathrm{B}\left(\frac{k}{2}, \frac{1}{2}\right) = \frac{\sqrt{\pi}\,\Gamma(k/2)}{\Gamma((k+1)/2)}$$

Now we just need to show that the leftmost factor on the right hand side of (4.11):

$$\frac{\Gamma(k)}{2^{k-1}(\Gamma(k/2))^2} = \frac{\Gamma((k+1)/2)}{\sqrt{\pi}\,\Gamma(k/2)} \tag{4.13}$$

To do this we use the duplication formula ((Abramowitz & Stegun, 1972), 6.1.18, pg 256):

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

with $z = k/2$. Then the left hand side of (4.13) is equal to:

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

as required. Putting everything together, we arrive at the alternative form for (4.11) given in equation (3.3), namely:

$$\Pr_R\{(R\hat{h})^T Rx < 0\} = \frac{\int_0^\theta \sin^{k-1}(\phi)\ \mathrm{d}\phi}{\int_0^\pi \sin^{k-1}(\phi)\ \mathrm{d}\phi} \tag{4.14}$$

This proves the second part of Theorem 3.2. □

4.1.3. Proof of part 3.

For reasons of space we omit the proof that the flipping probability is monotonic decreasing in the projection dimension k - this can be found in Ch. 6 of Durrant (2013). Note that although the value of the expressions in (3.3) and (3.2) can be calculated exactly for any given k and θ, e.g. using integration by parts, as k grows this becomes increasingly inconvenient. The final part of the theorem, bounding the flipping probability in the $(k+1)$-dimensional case above by the flipping probability in the k-dimensional case, is therefore useful in practice.

4.2. Proof of Generalization Error Bound - Theorem 3.1

We begin by considering the case when $R \in \mathcal{M}_{k\times d}$, $k < d$, is a fixed instance of a Gaussian random projection matrix. From classical VC theory (e.g. Vapnik, 1999; Herbrich, 2002) if $\hat{h}_R$ is the classifier with minimal empirical risk in the randomly projected space then we have, for any fixed R and any $\delta \in (0,1)$, with probability $1-\delta$ over the random draws of the training set $\mathcal{T}^N$ the following:

$$\Pr_{x_q,y_q}\{\hat{h}_R(Rx_q) \neq y_q\}$$
$$\leqslant \hat{E}(\mathcal{T}_R^N, \hat{h}_R) + 2\sqrt{\frac{VCdim\cdot\log(2eN/VCdim)+\log(1/\delta)}{N}}$$

where $\hat{E}(\mathcal{T}_R^N, \hat{h}_R)$ denotes the empirical risk $\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{\hat{h}_R(Rx_i) \neq y_i\}$. Further, since $\hat{h}_R$ is a linear classifier in k-dimensional space we also have $VCdim = k+1$ and we see immediately that random projection reduces the complexity term w.r.t the data space where $VCdim = d+1$. However, unless the data have some special structure, the empirical risk in the projected space will typically be greater than in the data space so we would especially like to quantify the effect of random projection on this term. With this goal in mind we first bound the empirical risk further by:

$$\hat{E}(\mathcal{T}_R^N, \hat{h}_R) \leqslant \hat{E}(\mathcal{T}_R^N, R\hat{h}) \tag{4.15}$$
$$= (\hat{E}(\mathcal{T}_R^N, R\hat{h}) - \hat{E}(\mathcal{T}^N, \hat{h})) + \hat{E}(\mathcal{T}^N, \hat{h}), \forall \hat{h} \in \mathbb{R}^d$$

where $\hat{E}(\mathcal{T}_R^N, R\hat{h})$ denotes the empirical error of a projected d-dimensional classifier evaluated on the projected training set, i.e. $\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{(R\hat{h})^T Rx_i \neq y_i\}$ for some $\hat{h} \in \mathbb{R}^d$.

The inequality (4.15) holds because $\hat{h}_R$ and $R\hat{h}$ lie in the same k-dimensional subspace of $\mathbb{R}^d$, and $\hat{h}_R$ is the ERM classifier in that subspace. Now $\hat{h} \in \mathbb{R}^d$ is an arbitrary vector that we can choose to minimize this bound, but we will take it to be the ERM classifier in $\mathbb{R}^d$ in order to keep the link between the randomly projected classifier $\hat{h}_R$ and its high-dimensional counterpart $\hat{h}$. Now observe that:

$$\hat{E}(\mathcal{T}_R^N, R\hat{h}) - \hat{E}(\mathcal{T}^N, \hat{h}) \ldots$$
$$\leqslant \frac{1}{N}\sum_{i=1:N} \mathbf{1}\{sign((R\hat{h})^T Rx_i) \neq sign(\hat{h}^T x_i)\}$$

and so, for any fixed R, w.p. $1-\delta$ w.r.t. random draws of $\mathcal{T}^N$ we have:

$$Pr_{x_q,y_q}\{\hat{h}_R^T Rx_q \neq y_q\} \ldots$$
$$\leqslant \frac{1}{N}\sum_{i=1}^N \mathbf{1}\{sign((R\hat{h})^T Rx_i) \neq sign(\hat{h}^T x_i)\} + \hat{E}(\mathcal{T}^N, \hat{h})$$
$$+2\sqrt{\frac{(k+1)\log(2eN/(k+1))+\log(1/\delta)}{N}} \tag{4.16}$$

Denote $S := \frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{\mathrm{sign}\,(\hat{h}^T x_i) \neq \mathrm{sign}\,(\hat{h}^T R^T Rx_i)\right\}$ in the above bound. This is an empirical estimate of the average flipping probability on this data from a single random projection. Our next step is to show that this estimate is not far from its expectation (with respect to random matrices R), that is S is close to $\mathrm{E}_R[S] = \frac{1}{N}\sum_{i=1}^N f_k(\theta_i)$, where $f_k(\theta_i)$ is the flipping probability of theorem 3.2. The main technical issue is the dependency between the $\hat{h}^T R^T Rx_i$ due to the common random matrix instance R and hence we cannot obtain decay with N since the random variable of interest is the projection matrix R and it is independent of N. To make the best of the situation, we derive two large deviation bounds for S: The first is a straightforward application of Markov inequality to give w.p. at least $1-\delta$:

$$S \leqslant \left(1 + \frac{1-\delta}{\delta}\right)\mathrm{E}_R[S] \tag{4.17}$$

where we recall that $\mathrm{E}_R[S] = \frac{1}{N}\sum_{i=1}^N f_k(\theta_i)$. Replacing the empirical estimate of the average flipping probability in (4.16) by RHS of (4.17) yields one high probability upper bound on the generalization error.

The upper bound on S given in (4.17) can be improved somewhat for small values of δ by using the following lemma, which is Corollary 3 on page 24 of Siegel (1995).

Lemma 4.1 (Chernoff bound for dependent variables)**.** *Let $X = \sum_{i=1}^N X_i$, where the X_i may be dependent. Let $Y = \sum_{i=1}^N Y_i$ where the Y_i are independent and $Y_i \sim X_i$ (i.e. $Pr\{Y_i \leqslant a\} = Pr\{X_i \leqslant a\}$, $\forall i$). Let B be a Chernoff bound on $Pr\{Y - E[Y] \geqslant \epsilon\}$ then:*

$$Pr\{X - E[X] \geqslant \epsilon\} \leqslant B^{1/N}$$

Now let R_i, $i \in \{1, 2, \ldots, N\}$ be a collection of N i.i.d draws of random matrices with i.i.d zero-mean Gaussian entries and define $S_N := \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{\text{sign}\left(\hat{h}^T x_i\right) \neq \text{sign}\left(\hat{h}^T R_i^T R_i x_i\right)\right\}$. The sum S_N differs from S in that S has the same random matrix in each summand while S_N has independent random matrices in each summand. However, for any $i \in \{1, ..., N\}$, the i-th term of S has the same distribution as the i-th term of S_N and a Chernoff bound for S_N can therefore be used to bound the deviation of S from its expectation, via lemma 4.1.

Now, by construction S_N is a sum of independent Bernoulli variables. Using a standard Chernoff bound for sums of Bernoulli random variables (e.g. Anthony & Bartlett, 1999, Pg 360) we obtain $\forall \epsilon \in (0, 1)$:

$$\Pr\left\{S_N \geqslant (1+\epsilon)\mathrm{E}_R[S_N]\right\} \leqslant \exp(-N\mathrm{E}_R[S_N]\epsilon^2/3) \quad (4.18)$$

and applying lemma 4.1 then yields:

$$\begin{aligned}\Pr_R\left\{S - \mathrm{E}_R[S] \geqslant \epsilon \mathrm{E}_R[S]\right\} &\leqslant \exp(-N\mathrm{E}_R[S_N]\epsilon^2/3)^{1/N} \\ &= \exp(-\mathrm{E}_R[S_N]\epsilon^2/3) \quad (4.19)\end{aligned}$$

This bound is 'Chernoff tight', i.e. tight w.r.t the Chernoff bound (4.18), when no assumptions are made on the set of points $\mathcal{T}^N$ and, in particular, it gives the appropriate Chernoff bound when all points of $\mathcal{T}^N$ are identical.

Now, specifying $\delta \in (0, 1)$, setting δ to the LHS of eq. (4.19), and using the fact that $\mathrm{E}_R[S_N] = \mathrm{E}_R[S]$, we obtain $\epsilon\sqrt{\mathrm{E}_R[S]} = \sqrt{3\log(1/\delta)}$. Rearranging we obtain, w.p. at least $1 - \delta$:

$$S \leqslant \mathrm{E}_R[S] + \sqrt{\mathrm{E}_R[S]}\sqrt{3\log(1/\delta)} \quad (4.20)$$

Replacing the empirical estimate of the average flipping probability in (4.16) with RHS of (4.20) yields a further high probability upper bound on generalization error. Taking the minimum over these two bounds, and finally applying union bound delivers the theorem. □

5. Geometric Interpretation of Flipping Probability

It is easy to verify that (4.14) recovers the known result for $k = 1$, namely θ/π, as given in Goemans & Williamson (1995, Lemma 3.2). Geometrically, when $k = 1$ the flipping probability is the quotient of the length of the arc with angle 2θ by the circumference of the unit circle which is 2π. In the form of (4.14) our result gives a natural generalization of this result, as follows: Recall that the surface area of the unit hypersphere in $\mathbb{R}^{k+1}$ is given by (Kendall, 2004):

$$2\pi \cdot \prod_{i=1}^{k-1} \int_0^{\pi} \sin^i(\phi)\mathrm{d}\phi$$

while the surface area of the hyperspherical cap with angle 2θ is given by:

$$2\pi \cdot \prod_{i=1}^{k-2} \int_0^{\pi} \sin^i(\phi)\mathrm{d}\phi \cdot \int_0^{\theta} \sin^{k-1}(\phi)\mathrm{d}\phi$$

Now taking the quotient of these two areas all but the last factors cancel and so we obtain our flipping probability as given in (4.14). Therefore, the probability that the sign of a dot product flips from being positive to being negative (equivalently the angle flips from acute to obtuse) after Gaussian random projection is given by the ratio of the surface area in $\mathbb{R}^{k+1}$ of a hyperspherical cap with angle 2θ to the surface area of the unit hypersphere.

Note the rather useful fact that the flipping probability depends only on the angular separation of the two vectors and on the projection dimensionality k: It is independent of the embedding dimensionality d which can therefore be arbitrarily large without affecting this quantity. Moreover this geometric interpretation shows that equation (4.14) decays exponentially with increasing k, since the proportion of the surface of the k-dimensional unit sphere covered by a spherical cap with angle of 2θ is bounded above by $\exp\left(-\frac{1}{2}k\cos^2(\theta)\right)$ (Ball, 1997, Lemma 2.2, Pg 11). Therefore we see that the additional loss arising from random projection (that is, the cost of working with randomly-projected data rather than working with the original high-dimensional data) is both independent of the original data dimensionality and will decay approximately exponentially as a function of k – approximately because there is some trade-off with the complexity term and the upper bound $\exp\left(-\frac{1}{2}k\cos^2(\theta)\right)$ is not tight. Our results therefore generalize the findings of Davenport et al. (2010); Durrant & Kabán (2010) where the same exponential decay was observed, but for specific choices of classifier. It is perhaps also worth noting that this loose upper bound on the flipping probability already improves considerably on the estimate of Garg et al. (2002), which is seen by substituting $\cos(\theta)$ for ν in their error bounds.

6. Two Straightforward Corollaries

6.1. Upper Bound on Generalization Error for Data Separable with a Margin

In proving theorem 3.1 we made no assumption that the data classes were linearly separable. However, if the classes are separable with a margin, m, in the data space (Cristianini & Shawe-Taylor, 2000) then straightforward geometry combined with Ball (1997, Lemma 2.2, Pg 11) yields an upper bound on our

flipping probability of $\exp\left(-\frac{1}{2}km^2\right)$. This bound holds deterministically, and so we then have the following high probability guarantee for separable data:

Corollary 6.1 (Generalization Error - Separable Classes)**.** *If the conditions of Theorem 3.1 hold and the data classes are also separable with a margin, m, in the data space then for all $\delta \in (0,1)$ with probability at least $1-2\delta$ we have:*

$$Pr_{x_q,y_q}\{\hat{h}_R(Rx_q) \neq y_q\} \leqslant \hat{E}(\mathcal{T}^N, \hat{h}) + \exp\left(-\frac{1}{2}km^2\right) + \min\left\{\sqrt{3\log\frac{1}{\delta}}\cdot\exp\left(-\frac{1}{4}km^2\right), \frac{1-\delta}{\delta}\cdot\exp\left(-\frac{1}{2}km^2\right)\right\} + 2\sqrt{\frac{(k+1)\log\frac{2eN}{k+1} + \log\frac{1}{\delta}}{N}}$$

Here we see that the bias introduced to the classifier by random projection decays exponentially with the square of the margin. Note that if we consider the margin at each training point individually then we have a setting analogous to the *margin distribution* considered in Shawe-Taylor (1998); Shawe-Taylor & Cristianini (1999). Using the margin distribution a tighter upper bound on the flipping probability is straightforward to derive - for reasons of space we do not do so here.

6.2. Upper Bound on Generalization Error in Data Space

An upper bound on the average label flipping probability whose exact form we derived here is key in the bounds of Garg et al. (2002), where it serves as a data-dependent complexity measure (termed the 'projection profile') to characterize the generalization error of data space linear classifiers. It is now straightforward to use our exact form in place of their projection profile term to give the following bound on the generalization error of *data space* classifiers as a corollary, which is an improvement on the main result in (Garg et al., 2002):

Corollary 6.2 (Data Space Generalization Error)**.** *Let $\mathcal{T}^{2N} = \{(x_i, y_i)\}_{i=1}^{2N}$ be a set of d-dimensional labelled training examples drawn i.i.d. from some data distribution $\mathcal{D}$, and let $\hat{h}$ be a linear classifier estimated from $\mathcal{T}^{2N}$ by ERM. Let $k \in \{1, 2, \ldots, d\}$ be an integer and let $R \in \mathcal{M}_{k\times d}$ be a random projection matrix, with entries $r_{ij} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$. Then for all $\delta \in (0, 1]$, with probability at least $1-4\delta$ w.r.t. the random draws of $\mathcal{T}^{2N}$ and R the generalization error of $\hat{h}$ w.r.t the (0,1)-loss is bounded above by:*

$$Pr_{x_q,y_q}\{\hat{h}^T x_q \neq y_q\} \leqslant \hat{E}(\mathcal{T}^{2N}, \hat{h}) + 2\cdot\min_k\left\{\frac{1}{N}\sum_{i=1}^{2N} f_k(\theta_i) + \min\left\{\sqrt{3\log\frac{1}{\delta}}\sqrt{\frac{1}{N}\sum_{i=1}^{2N} f_k(\theta_i)}, \frac{1-\delta}{\delta}\cdot\frac{1}{N}\sum_{i=1}^{2N} f_k(\theta_i)\right\} + \sqrt{\frac{(k+1)\log\frac{2eN}{k+1} + \log\frac{1}{\delta}}{2N}}\right\} \quad (6.1)$$

Proof Sketch: Follow the two-part proof in Garg et al. (2002): One part bounds the generalization error using classical tools of the double sample trick and Sauer lemma after making a move into the random projection space; while the other is an estimate of our flipping probability obtained using the JLL. To obtain the result in (6.1) plug in our exact form for the flipping probability for their estimate and use lemma 4.1 as well as Markov inequality in their lemma 3.4.

7. Summary and Discussion

We derived the exact probability of 'label flipping' as a result of Gaussian random projection, and used it to derive sharp upper bounds on the generalization error of a randomly-projected classifier. Unlike earlier results of Arriaga & Vempala (1999) and Calderbank et al. (2009), we require neither a large margin nor data sparsity for our bounds to hold, while unlike Davenport et al. (2010) and Durrant & Kabán (2010); Durrant & Kabán (2011) our guarantees hold for an arbitrary data distribution.

Our proof makes use of the orthogonal invariance of the standard Gaussian distribution, which cannot be applied for other random matrices with entries whose distribution is not orthogonally invariant: It would be interesting to extend these results to more general random projection matrices, and we are working on ways to do this. Furthermore we note that the form of VC complexity term in our bounds is not optimal, for example better guarantees (albeit without explicit constants) are given in Bartlett & Mendelson (2002) and these could be used in place of the bounds we adopted to sharpen our results further.

Our findings show that good generalization performance can be obtained from a classifier trained on randomly projected data, provided that the data have some structure which keeps the probability of label flipping low – we saw that two such structures are when data classes are separable or soft-separable with a margin. Identifying other structural properties of data which also imply a low flipping probability remains for future work.

References

Abramowitz, M. and Stegun, I.A. *Handbook of Mathematical Functions*. Dover, New York, 10th edition, 1972.

Achlioptas, D. Database-friendly Random Projections: Johnson-Lindenstrauss with Binary Coins. *J. Computer and System Sciences*, 66(4):671–687, 2003.

Ailon, N. and Chazelle, B. Approximate Nearest Neighbors and the Fast Johnson–Lindenstrauss Transform. In *Proc. 38th Annual ACM Symposium on Theory of Computing (STOC 2006)*, pp. 557–563. ACM, 2006.

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

Arriaga, R.I. and Vempala, S. An Algorithmic Theory of Learning: Robust Concepts and Random Projection. In *40th Annual Symposium on Foundations of Computer Science (FOCS 1999).* , pp. 616–623. IEEE, 1999.

Ball, K. An Elementary Introduction to Modern Convex Geometry. *Flavors of Geometry*, 31:1–58, 1997.

Bartlett, P.L. and Mendelson, S. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *J. Machine Learning Research*, 3:463–482, 2002.

Boyali, A. and Kavakli, M. A Robust Gesture Recognition Algorithm based on Sparse Representation, Random Projections and Compressed Sensing. In *7th IEEE Conference on Industrial Electronics and Applications (ICIEA 2012)*, pp. 243–249, july 2012.

Calderbank, R., Jafarpour, S., and Schapire, R. Compressed Learning: Universal Sparse Dimensionality Reduction and Learning in the Measurement Domain. Technical Report, Rice University, 2009.

Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

Dasgupta, S. and Gupta, A. An Elementary Proof of the Johnson–Lindenstrauss Lemma. *Random Structures & Algorithms*, 22:60–65, 2002.

Davenport, M.A., Boufounos, P.T., Wakin, M.B., and Baraniuk, R.G. Signal Processing with Compressive Measurements. *IEEE J. Selected Topics in Signal Processing*, 4(2):445–460, April 2010.

Durrant, R.J. *Learning in High Dimensions with Projected Linear Discriminants*. PhD thesis, School of Computer Science, University of Birmingham, January 2013.

Durrant, R.J. and Kabán, A. Compressed Fisher Linear Discriminant Analysis: Classification of Randomly Projected Data. In *Proc. 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010)*, 2010.

Durrant, R.J. and Kabán, A. A Tight Bound on the Performance of Fishers Linear Discriminant in Randomly Projected Data Spaces. *Pattern Recognition Letters*, 33 (7):911–919, 2011.

Fard, M., Grinberg, Y., Pineau, J., and Precup, D. Compressed Least-squares Regression on Sparse Spaces. In *Proc. 26th AAAI Conference on Artificial Intelligence (AAAI 2012)*, 2012.

Fodor, I.K. A Survey of Dimension Reduction Techniques. Technical Report UCRL-ID-148494, US Dept. of Energy, Lawrence Livermore National Laboratory, 2002.

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

Garg, A., Har-Peled, S., and Roth, D. On Generalization Bounds, Projection Profile, and Margin Distribution. In *Proc. 19th International Conference on Machine Learning (ICML 2002)*, pp. 171–178, 2002.

Goemans, M.X. and Williamson, D.P. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems using Semidefinite Programming. *Journal of the ACM*, 42(6):1145, 1995.

Herbrich, R. *Learning Kernel Classifiers: Theory and Algorithms*. The MIT Press, 2002.

Kendall, MG. *A Course in the Geometry of n Dimensions*. Dover, New York, 2004.

Mahoney, M.W. Randomized Algorithms for Matrices and Data. *arXiv preprint arXiv:1104.5557*, 2011.

Maillard, O. and Munos, R. Linear Regression with Random Projections. *J. Machine Learning Research*, 13: 2735–2772, 2012.

Mardia, K.V., Kent, J.T., and Bibby, J.M. *Multivariate Analysis*. Academic Press, London, 1979.

Matoušek, J. On Variants of the Johnson–Lindenstrauss Lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

Paul, S., Boutsidis, C., Magdon-Ismail, M., and Drineas, P. Random Projections for Support Vector Machines. *arXiv preprint arXiv:1211.6085*, 2012.

Pillai, J.K., Patel, V.M., Chellappa, R., and Ratha, N.K. Secure and Robust Iris Recognition using Random Projections and Sparse Representations. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(9):1877–1893, 2011.

Shawe-Taylor, J. Classification Accuracy based on Observed Margin. *Algorithmica*, 22(1-2):157–172, 1998.

Shawe-Taylor, J. and Cristianini, N. Further Results on the Margin Distribution. In *Proc. 12th Annual Conference on Computational Learning Theory (COLT 1999)*, pp. 278–285. ACM, 1999.

Siegel, A. Toward a Usable Theory of Chernoff bounds for Heterogeneous and Partially Dependent Random Variables. Technical Report, New York University, 1995.

Vapnik, V.N. An Overview of Statistical Learning Theory. *IEEE Trans. Neural Networks*, 10(5):988–999, 1999.