

On generalization bounds, projection profile, and margin distribution

Ashutosh Garg Sarel Har-Peled Dan Roth

Department of Computer Science and the Beckman Institute

University of Illinois, Urbana, IL. 61801, USA

`{ashutosh,sariel,dan}@uiuc.edu`

Abstract

We study generalization properties of linear learning algorithms and develop a data dependent approach that is used to derive generalization bounds that depend on the margin distribution. Our method makes use of random projection techniques to allow the use of existing VC dimension bounds in the effective, lower, dimension of the data. Comparisons with existing generalization bound show that our bounds are tighter and meaningful in cases existing bounds are not.

1 Introduction

The study of generalization abilities of learning algorithms and its dependence on sample complexity is one of the fundamental research efforts in learning theory. Understanding the inherent difficulty of learning problems allows one to evaluate whether learning is at all possible in certain situations, estimate the degree of confidence in the predictions made by learned classifiers, and is crucial in understanding and analyzing learning algorithms.

Understanding generalization is even more important when learning in very high dimensional spaces, as in many natural language and computer vision applications. Specifically, can one count on the behavior of a 10^6 dimensional classifier that is trained on a few examples, or even a few thousands examples? Existing bounds are loose and essentially meaningless in these (and even in simpler) cases¹.

This work develops a learning theory that is relevant for learning in very high dimensional spaces and uses it to establish informative generalization bounds, even for very high dimensional learning problems. The approach is motivated by recent works [RZ00, GR01, AV99] that argue that some high dimensional learning problems are naturally constrained in ways that make them, effectively, low dimensional problems. In these cases, although learning is done in a high dimension, generalization ought to depend on the true, lower dimensionality of the problem.

Technically, our approach builds on recent developments in the area of random projection of high dimensional data [JL84] which shows, informally, that it is possible to project high

¹Several statistical methods can be used to ensure the robustness of the empirical error estimate [KMNR97]. However, these typically require training on even less data, and do not contribute to understanding generalization and the domain.

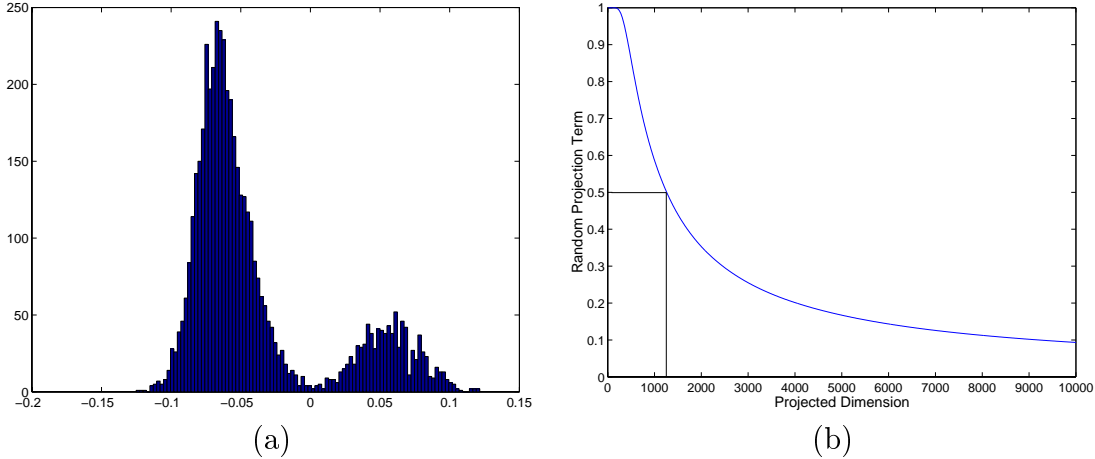


Figure 1: (a) Histogram of the distance of the points from the classifier for the context sensitive spelling correction. (b) The distortion error due to random projection as a function of the dimension of the projected space.

dimensional data randomly into a much smaller dimensional space, with a relatively small distortion of the distances between the projected points. This result is extended here to apply in the context of a sample of points along with a linear classifying hypothesis, and is used to develop generalization bounds for linear classifiers in very high dimensional spaces. The basic intuition underlying our results is as follows. If the effective dimensionality of the data is much smaller than the observed dimensionality, then it should be possible to randomly project the data into a lower dimensional space while, due to a small distortion in distances, incurring only a small amount of classification error, relative to what is possible in the original, high dimensional space. Since the projected space has low dimension, better “standard” generalization bounds hold there.

We introduce a new, data dependent, complexity measure for learning. The *projection profile* of data sampled according to a distribution \mathcal{D} , is the expected amount of error introduced when a classifier h is randomly projected, along with the data, into k -dimensions. Although this measure seems somewhat elusive, we show that it is captured by the following quantity: $a_k(\mathcal{D}, h) = \int_{x \in \mathcal{D}} u(x) d\mathcal{D}$, where

$$u(x) = \min \left(3 \exp \left(-\frac{(\nu(x))^2 k}{8(2 + |(\nu(x))|^2)} \right), 1 \right)$$

and $\nu(x)$ is the distance between x and the classifying hyperplane² defined by h , a linear classifier for \mathcal{D} . The sequence $\mathcal{P}(\mathcal{D}, h) = (a_1(\mathcal{D}, h), a_2(\mathcal{D}, h), \dots)$ is the *projection profile* of \mathcal{D} .

The projection profile turns out to be quite informative, both theoretically and in practice. In particular, it decreases monotonically (as a function of k), and provides a trade-off between dimension and accuracy. Namely, if the data is transformed from n to k dimensions then we expect the amount of error introduced to be $a_k(\mathcal{D}, h)$. This new complexity measure

²Our analysis does not assume the data to be linearly separable.

allows us to state a generalization bound in terms of the lower-dimensional projected space - the effective dimension of the data. We show that the overall performance will depend on an *estimation* of the projection profile when projecting to the effective dimension, with the addition of the standard, VC-dimension arguments, in the projected space.

Our approach suggests a significant improvement over current approaches to generalization bounds, which are based either on VC theory [Vap82] and learning theory versions of Occam’s Razor [BEHW89] or, more recently, on the *margin* of a classifier with respect to a sample [ST98, STC99, STC00, HG01].

Although the development of margin-based bounds has been a significant improvement, they still are not meaningful. The main shortcoming is that the margin of the data might be defined by very few points of the distribution and thus might be very small. Our method can be viewed as allowing an explicit dependency on the distribution of the geometric distances of points from the classifier, rather than only the extreme points. We refer to this distance as the *margin distribution* of the data. In our method, the contribution of those nearby “problematic” points to the generalization bound is weighted together with their portion in the distribution. This is significant when most of the data is far from the optimal classifier - only very few points, those that determine the margin, are close to it. Our experiments reveal that this is indeed the case. The advantage of our method is exhibited in Figure 1, showing the margin distribution of data taken from a high dimensional natural language classification problem [GR99]. Despite the zero margin in this case, our method provides an informative bound.

This paper presents our method, analyzes it and uses it to develop new generalization bounds. We then evaluate the projection profile based method experimentally on real data and show its effectivity. Specifically, we show that meaningful generalization bounds are achieved for high dimensional problems in the natural language domain, and even in cases where learning is done using continuous kernels.

2 Preliminaries

We study a binary classification problem $f : \mathbb{R}^n \rightarrow \{-1, 1\}$. $S = \{(x_1, y_1), \dots, (x_m, y_m)\}$ denotes a sample set of m examples. The hypothesis $h \in \mathbb{R}^n$ is an n -dimensional linear classifier assumed, w.l.o.g, to pass through the origin. That is, for an example $x \in \mathbb{R}^n$ the hypothesis predicts $\hat{y}(x) = \text{sign}(h^T x)$. Throughout the paper, we use n to denote the original (high) dimensionality of the data, k to denote the (smaller) dimension into which projections are made and m to denote the sample size of the data.

Definition 2.1 The loss function considered is the 0-1 loss. Under this loss function, the *empirical error* \hat{E} of h over a sample set S and the *expected error* \bar{E} are given resp. by,

$$\hat{E}(h, S) = \frac{1}{m} \sum_i I(\hat{y}(x_i) \neq y_i) \quad \text{and} \quad \bar{E}(h, S) = E_x \left[\hat{y}(x) \neq f(x) \right],$$

where $I(\cdot)$ is the indicator function which is 1 when its argument is true and 0 otherwise. The expectation E_x is taken over the distribution of data.

We denote by $\|\cdot\|$ the L_2 norm of a vector. Clearly, we can assume w.l.o.g that all data points come from the surface of unit sphere (ie. $\forall x, \|x\| = 1$), and that $\|h\| = 1$ (as the classification just depends on the sign of the inner product.) Under these assumptions, the classification output $\widehat{y}(x)$ can be interpreted as the sign of the angle between the vectors x and h .

Let $\nu(x) = h^T x$ denote the signed distance of the sample point x from the classifier h . When x refers to the j^{th} sample from the sample set S , we denote it by $\nu_j = \nu(x_j) = h^T x_j$. With this notation (omitting the classifier h) the classification rule reduces simply to $\text{sign}(\nu(x))$.

Our method makes use of *random projections*, introduced in [JL84] and studied intensively since then [AV99, Ind01].

Definition 2.2 (Random Matrix) Let R be a $k \times n$ matrix where each entry $r_{ij} \sim N(0, 1/k)$. R is called a *random projection matrix*. For $x \in \mathbb{R}^n$, we denote by $x' = Rx \in \mathbb{R}^k$ the projection of x from an n to a k -dimensional space using projection matrix R .

In a similar fashion, for a classifier $h \in \mathbb{R}^n$, h' will denote it's projection to a k dimensional space via R (omitting k when clear from the context). Likewise, S' denotes the set of points which are the projections of the sample S , and $\nu'_j = (h')^T x'_j$, the signed distance of the projected point from the projected classifier.

Briefly, the method of random projection shows that with high probability, when n dimensional data is projected down to a lower dimensional space of dimension k , using a random $k \times n$ matrix, relative distances between points are *almost* preserved. Formally:

Theorem 2.3 ([AV99], Thm. 1) *Let $u, v \in \mathbb{R}^n$ and let u' and v' be the projections of u and v to \mathbb{R}^k via a random matrix R chosen as described above. Then for any constant c ,*

$$\Pr \left[(1 - c) \leq \frac{\|u' - v'\|^2}{\|u - v\|^2} \leq (1 + c) \right] \geq 1 - e^{-c^2 k/8}, \quad (1)$$

where the probability is over the selection of the random matrix R .

Note that if $\|u\| = \|v\| = 1$, then $\|u - v\| = 2 - 2u \cdot v$. Therefore the above theorem can also be viewed as stating that, with high probability, random projection preserves the angle between vectors which lie on the unit sphere.

3 A Margin Distribution based Bound

As mentioned earlier, the decision of the classifier h is based on the sign of $\nu(x) = h^T x$. Since both h and x are normalized, $|\nu(x)|$ can be thought of as the geometric distance between x and the hyperplane orthogonal to h that passes through the origin. Given a distribution on data points x , this induces a distribution on their distance from the hyperplane induced by h , which we refer to as the *margin distribution*.

Note that this is different from the margin of the sample set S with respect to a classifier h . Traditionally in the learning community, margin of a sample set S (referred to simply as

“the margin”) is the distance of the point which is closest to the hyperplane. Formally,

$$\text{margin of } S \text{ w.r.t. } h \text{ is } \gamma(S, h) = \min_{i=1}^m |h^T x_i|.$$

Consider a scenario in which one learns a classifier in a very high dimensional space (typical in image processing, language processing and data mining application). According to existing theory, in order to learn a classifier which, with high confidence, performs well on previously unseen data, one needs to train it on a large amount of data. In what follows, we develop alternative bounds that show that if “many” of the high dimensional points are classified with high confidence, that is, $|h^T x|$ is large for these, then one doesn’t need as many points as predicted by VC-theory or margin based theory. The main result of the paper formalizes this insight and is given in the following theorem.

Theorem 3.1 *Let $S = \{(x_1, y_1), \dots, (x_{2m}, y_{2m})\}$ be a set of n -dimensional labeled examples and h a linear classifier. Then, for all constants $0 < \delta < 1; 0 < k$, with probability at least $1 - 4\delta$, the expected error of h is bounded by*

$$\overline{E} \leq \widehat{E}(S, h) + \min_k \left\{ \mu_k + 2\sqrt{\frac{(k+1) \ln \frac{me}{k+1} + \ln \frac{1}{\delta}}{2m}} \right\} \quad (2)$$

where $\mu_k = \frac{6}{m\delta} \sum_{j=1}^{2m} \exp\left(-\frac{\nu_j^2 k}{8(2+|\nu_j|)^2}\right)$, $\nu_j = \nu(x_j) = h^T x_j$.

3.1 Proving the Margin Distribution Bound

The bound given in Eqn. 2 has two main components. The first component, μ_k , is the distortion incurred by the random projection to dimension k , and the second follows directly from VC theory for this dimension.

Recall that the random projection theorem states that relative distances are (almost) preserved when projecting to lower dimensional space. Therefore, we first argue that the image, under projection, of data points that are far from h in the original space, will still be far from its image in the projected (k dimensional) space. The first term quantifies the penalty incurred due to data points whose images will not be consistent with the image of h . That is, this term bounds the *empirical* error in the projected space. Once the data lies in the lower dimensional space, we can bound the expected error of the classifier on the data as a function of the dimension of the space, number of samples and the empirical error there (that is, the first component).

Decreasing the dimension of the projected space implies increasing the contribution of the first term, while the VC-dimension based term decreases. To get the optimal bound, one has to balance these two quantities and choose the dimension k of the projected space so that the generalization error is minimized. We will use the following lemmas to compute the penalty incurred while projecting the data down to k dimensional space.

Lemma 3.2 *Let h be an n -dimensional classifier, $x \in \mathbb{R}^n$ a sample point, such that $\|h\| = \|x\| = 1$, and $\nu = h^T x$. Let $R \in \mathbb{R}^{k \times n}$ a random projection matrix (Def. 2.2), with $h' =$*

$Rh, x' = Rx$. Then the probability of misclassifying x , relative to its classification in the original space, due to the random projection, is

$$P \left[\text{sign}(h^T x) \neq \text{sign}(h'^T x') \right] \leq 3 \exp \left(-\frac{\nu^2 k}{8(2 + |\nu|)^2} \right).$$

Proof: From Theorem 2.3 we know that with probability at least $Z(c) = 1 - 3 \exp(-c^2 k/8)$, we have

$$\begin{aligned} (1-c)\|h\|^2 &\leq \|h'\|^2 \leq (1+c)\|h\|^2, \\ (1-c)\|x\|^2 &\leq \|x'\|^2 \leq (1+c)\|x\|^2, \\ (1-c)\|h-x\|^2 &\leq \|h'-x'\|^2 \leq (1+c)\|h-x\|^2. \end{aligned}$$

Since $\|h\| = \|x\| = 1$, and setting $\nu = h^T x$, $\nu' = h'^T x'$, we have $\|h-x\|^2 = \|h\|^2 + \|x\|^2 - 2h^T x = 2 - 2\nu$ and $\|h'-x'\|^2 = \|h'\|^2 + \|x'\|^2 - 2\nu'$. In particular $\|h'\|^2 + \|x'\|^2 - 2\nu' \leq (1+c)(2-2\nu)$. Namely, $\nu' \geq (\|h'\|^2 + \|x'\|^2)/2 - (1+c)(1-\nu) \geq (1-c) - (1+c)(1-\nu) = c(\nu-2) + \nu$. Thus, when $\nu > 0$ we have $\nu' > 0$ if $c(\nu-2) + \nu > 0$. This implies that we need $\nu/(2-\nu) > c$.

Similarly, $\|h'\|^2 + \|x'\|^2 - 2\nu' \geq (1-c)(2-2\nu)$. Namely, $\nu' \leq (\|h'\|^2 + \|x'\|^2)/2 - (1-c)(1-\nu) \leq 1+c-1+c+\nu(1-c) = 2c+\nu(1-c) = c(2-\nu) + \nu$. In particular, if $\nu < 0$ then $\nu' < 0$ if $c(2-\nu) + \nu < 0$, which implies $c < -\nu/(2-\nu)$.

Combining the above two inequalities, we conclude that ν and ν' have the same sign if $c < |\nu|/(2+|\nu|)$. Namely, the required probability is

$$\Pr \left[\text{sign}(h^T x) \neq \text{sign}(h'^T x') \right] \leq 1 - Z \left(\frac{|\nu|}{2+|\nu|} \right) = 3 \exp \left(-\nu^2 k / (8(2+|\nu|)^2) \right),$$

which is obtained by picking $c = |\nu|/(2+|\nu|)$. ■

Next we define the projection error for a sample – this is essentially the projection profile introduced in Section 1 for a finite sample. A natural interpretation of the projection error is that it is an estimate of the projection profile by sampling.

Definition 3.3 (projection error) Given a classifier h , a sample S , and a random projection matrix R , let $\text{Err}_{\text{proj}}(h, R, S)$ be the classification error caused by the projection matrix R . Namely,

$$\text{Err}_{\text{proj}}(h, R, S) = \frac{1}{|S|} \sum_{x \in S} I(\text{sign}(h^T x) \neq \text{sign}(h'^T x')).$$

Lemma 3.4 Let h be an n -dimensional classifier, R a random projection matrix, and a sample of m points

$$S = \left\{ (x_1, y_1), \dots, (x_m, y_m) \mid x_i \in \mathbb{R}^n, y_i \in \{0, 1\} \right\}.$$

Then, with probability $\geq 1 - \delta$ (over the choice of the random projection matrix R), the projection error satisfies $\text{Err}_{\text{proj}}(h, R, S) \leq \varepsilon_1(S, \delta)$, where

$$\varepsilon_1(S, \delta) = \frac{1}{m} \frac{1}{\delta} \sum_{i=1}^m 3 \exp \left(-\nu_i^2 k / (8(2+|\nu_i|)^2) \right),$$

$\nu_i = h^T x_i$, for $i = 1, \dots, m$.

Proof: Let Z be the expected projection error of a sample where the expectation is taken with respect to the choice of the projection matrix. That is,

$$\begin{aligned}
Z &= E \left[\text{Err}_{proj}(h, R, S) \right] = E \left[\frac{1}{|S|} \sum_{x \in S} I(\text{sign}(h^T x) \neq \text{sign}(h'^T x')) \right] \\
&= \frac{1}{m} \sum_{x \in S} E \left[I(\text{sign}(h^T x) \neq \text{sign}(h'^T x')) \right] = \frac{1}{m} \sum_{x \in S} \Pr \left[\text{sign}(h^T x) \neq \text{sign}(h'^T x') \right] \\
&\leq \frac{1}{m} \sum_{x \in S} 3 \exp \left(-\nu_i^2 k / (8(2 + |\nu_i|)^2) \right) = \delta \varepsilon_1(S, \delta),
\end{aligned}$$

which follows by linearity of expectation and Lemma 3.2. Now, using Markov inequality,

$$\Pr \left[\text{Err}_{proj}(h, R, S) \geq \frac{Z}{\delta} \right] \leq \delta,$$

which establishes the lemma, as $Z/\delta \leq \varepsilon_1(S, \delta)$. ■

Lemma 3.5 *Let S_1, S_2 be two samples of size m from \mathbb{R}^n , R a random projection matrix, and S'_1, S'_2 the projected sets. Then, with probability $\geq 1 - 2\delta$,*

$$\Pr \left[\left| \widehat{E}(h, S_1) - \widehat{E}(h, S_2) \right| > \varepsilon \right] < \Pr \left[\left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| > \rho \right],$$

where $\rho = \varepsilon - \varepsilon_1(S_1, \delta) - \varepsilon_1(S_2, \delta)$.

Proof: Applying the result of the Lemma 3.4 on sample sets S_1, S_2 , we obtain that with probability at least $1 - 2\delta$, we have that

$$\left| \widehat{E}(h, S_1) - \widehat{E}(h', S'_1) \right| < \varepsilon_1(S_1, \delta), \quad \text{and} \quad \left| \widehat{E}(h, S_2) - \widehat{E}(h', S'_2) \right| < \varepsilon_1(S_2, \delta).$$

Using simple algebra, we have,

$$\begin{aligned}
\left| \widehat{E}(h, S_1) - \widehat{E}(h, S_2) \right| &\leq \left| \widehat{E}(h, S_1) - \widehat{E}(h', S'_1) \right| + \left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| \\
&\quad + \left| \widehat{E}(h', S'_2) - \widehat{E}(h, S_2) \right| \\
&\leq \varepsilon_1(S_1, \delta) + \varepsilon_1(S_2, \delta) + \left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right|.
\end{aligned}$$

In particular, with probability $\geq 1 - 2\delta$,

$$\begin{aligned}
\Pr \left[\left| \widehat{E}(h, S_1) - \widehat{E}(h, S_2) \right| > \varepsilon \right] &\leq \Pr \left[\varepsilon_1(S_1, \delta) + \varepsilon_1(S_2, \delta) + \left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| > \varepsilon \right] \\
&= \Pr \left[\left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| > \varepsilon - \varepsilon_1(S_1, \delta) - \varepsilon_1(S_2, \delta) \right]
\end{aligned}$$

We have obtained a bound on the additional classification error that is incurred when projecting the sample down from some n dimensional space to a k dimensional space. As a ■

result, we have established the fact that the difference between the classification performance on two samples, in high dimension, is very similar to the difference in low dimension. This is now used to prove the main result of the paper.

Proof of Theorem 3.1: Let S_1 denote a sample of size m from \mathbb{R}^n . Let \mathcal{H} denotes the space of all linear classifiers in n dimensional space and let $h \in \mathcal{H}$. Also, let $\overline{E}(h)$ denote the expected error of a classifier h , on the data sampled according to distribution \mathcal{D} and $\widehat{E}(h, S_1)$, the empirical (observed) error of the same classifier, h , on the sample set S_1 when the data was sampled according to the same distribution \mathcal{D} .

To obtain the generalization error of a classifier which is learned in a high dimensional space, we want to compute the bound on the following quantity as a function of the error ε :

$$\Pr \left[\sup_{h \in \mathcal{H}} \left| \overline{E}(h) - \widehat{E}(h, S_1) \right| > \varepsilon \right]. \quad (3)$$

We are interested in the probability that uniformly all classifiers from the hypothesis space \mathcal{H} will do well on future data. To compute the bound, we use the technique of double samples which has been used in proving bounds of this kind. Assume we observe a sample of size $2m$, where S_1, S_2 denote the first and second half of sample resp. From [Vap98, 131–133]:

$$\Pr \left[\sup_{h \in \mathcal{H}} \left| \overline{E}(h) - \widehat{E}(h, S_1) \right| > \varepsilon \right] \leq 2 \Pr \left[\sup_{h \in \mathcal{H}} \left| \widehat{E}(h, S_1) - \widehat{E}(h, S_2) \right| > \frac{\varepsilon}{2} \right]$$

Now suppose we project the data along with the hyperplane down to k dimensional space, using a random projection matrix. Then according to Lemma 3.5, with high probability most of the data stays on the correct side of the hyperplane. Formally, with probability $\geq 1 - 2\delta$,

$$\Pr \left[\sup_{h \in \mathcal{H}} \left| \widehat{E}(h, S_1) - \widehat{E}(h, S_2) \right| > \frac{\varepsilon}{2} \right] \leq \Pr \left[\sup_{h \in \mathcal{H}} \left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| > \rho \right],$$

where $\rho = \frac{\varepsilon}{2} - \varepsilon_1(S_1, \delta) - \varepsilon_1(S_2, \delta)$.

Since the sample sets S_1, S_2 contains independent samples so do S'_1, S'_2 , and using results from [Vap98, p. 134] we write

$$\begin{aligned} & \Pr \left[\sup_{h \in \mathcal{H}} \left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| > \rho \right] \\ &= \Pr \left[\sup_{h' \in \mathcal{H}'} \left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| > \rho \right] \leq \sum_{h' \in \mathcal{H}'} \Pr \left[\left| \widehat{E}(h', S'_1) - \widehat{E}(h', S'_2) \right| > \rho \right] \\ &\leq N^k(2m) \exp(-2\rho^2 m) \leq \left(\frac{2em}{k+1} \right)^{k+1} \exp(-2\rho^2 m), \end{aligned}$$

where $N^k(2m)$ is the maximum number of different partitions of $2m$ samples of k dimensional data that can be realized by a k dimensional linear classifier. The last inequality uses the Sauer's lemma to bound this quantity as a function of the VC dimension of a classifier which in this case happens to be $(k+1)$. To bound this probability by δ , the confidence parameter, we isolate ρ from the inequality, which gives $\left(\frac{2em}{k+1} \right)^{k+1} \exp(-2\rho^2 m) \leq \delta$. Simplification reveals that this inequality holds for

$$\rho = \sqrt{\frac{(k+1) \ln(2em/(k+1)) + \ln(1/\delta)}{2m}}.$$

Solving for ε , we have $\varepsilon = 2\rho + 2\varepsilon_1(S_1, \delta) + 2\varepsilon_1(S_2, \delta)$. Putting it together, we get

$$\Pr \left[\sup_{h \in \mathcal{H}} \left| \overline{E}(h) - \widehat{E}(h, S_1) \right| > \varepsilon \right] \leq 2\delta.$$

Combining this, together with the bound on the confidence that Equation (4) holds, we get that with probability $\geq (1 - 2\delta)(1 - 2\delta) \geq 1 - 4\delta$, we have

$$\begin{aligned} \overline{E}(h, S_1) &\leq \widehat{E}(h, S_1) + \varepsilon \\ &\leq \widehat{E}(h, S_1) + 2\varepsilon_1(S_1, \delta) + 2\varepsilon_1(S_2, \delta) + 2\sqrt{\frac{(k+1) \ln(2em/(k+1)) + \ln(1/\delta)}{2m}}. \end{aligned}$$

Plugging in the value of ρ along with the value of $\varepsilon_1(\cdot, \cdot)$, gives the final bound. \blacksquare

In fact, it is possible to improve the bound for some ranges of k , using the fact that we only care about the distortion in the distance of points from the classifier, rather than the distortion in the size of the projected vectors themselves, as in Lemma 3.2.

Lemma 3.6 *Let $x' = Rx$ and $h' = Rh$. Let $\nu = h^T x$ and $\nu' = h'^T x'$, where R is $n \times k$ random projection matrix. Then, we have*

$$E[\nu'] = \nu, \quad \text{var}[\nu'] = \frac{1 + \nu^2}{k}.$$

Proof: Let $R = \{r_1; r_2; \dots; r_k\}$, where r_i are the row vectors, see Def. 2.2. Then, $\nu' = h'^T x' = \sum_{i=1}^k h^T r_i r_i^T x$. Let $\nu'_i = h^T r_i r_i^T x$. Clearly, the ν'_i 's are independent random variables, and we can express $\nu' = \sum_{i=1}^k \nu'_i$ as the sum of independent random variables. Furthermore,

$$E[\nu'_i] = E[h^T r_i r_i^T x] = h^T E[r_i r_i^T] x = h^T (I/k) x = h^T x = \nu/k,$$

where I is the identity matrix. This hold as each entry of the random projection matrix is $\sim N(0, 1/k)$, and thus the matrix $E[r_i r_i^T]$ has zero in each non-diagonal entry, as it is the expectation of the product of two independent variables, each of expectation zero, and the value of a diagonal entry is the second moment of $N(0, 1/k)$, which is $1/k$. In particular, $E[\nu'] = E[\sum_i \nu'_i] = \nu$.

We next compute $\text{var}[\nu'_i]$. Let $r_i = [\mu_1, \dots, \mu_n]$, where $\mu_i \sim N(0, 1/k)$. Then:

$$E[\nu_i'^2] = E[(h^T r_i r_i^T x)^2] = E[(h^T r_i r_i^T x)(h^T r_i r_i^T x)] = \sum_{a=1}^n \sum_{b=1}^n \sum_{c=1}^n \sum_{d=1}^n x_a x_b h_c h_d E[\mu_a \mu_b \mu_c \mu_d].$$

Note, that if a is not equal to either b, c or d , then $E[\mu_a \mu_b \mu_c \mu_d] = E[\mu_a] E[\mu_b \mu_c \mu_d] = 0 \cdot E[\mu_b \mu_c \mu_d]$, as μ_a is independent of μ_b, μ_c, μ_d . We can apply the same argumentation to b, c , and d . It follows, thus, that the only non-zero terms in the above summation are when: (i) $a = b$ and $c = d$, or (ii) $a = c$ and $b = d$, or (iii) $a = d$ and $b = c$ (note that the case $a = b = c = d$ is counted three times in those cases). Thus

$$\begin{aligned} E[\nu_i'^2] &= \sum_{a=1}^n \sum_{c=1}^n x_a x_a h_c h_c E[\mu_a \mu_a \mu_c \mu_c] + \sum_{a=1}^n \sum_{b=1}^n x_a x_b h_a h_b E[\mu_a \mu_b \mu_a \mu_b] \\ &\quad + \sum_{a=1}^n \sum_{b=1}^n x_a x_b h_b h_a E[\mu_a \mu_b \mu_b \mu_a] - 2 \sum_{a=1}^n x_a x_a h_a h_a E[\mu_a \mu_a \mu_a \mu_a]. \end{aligned}$$

We observe that $E[\mu_u^2] = 1/k$ and $E[\mu_u^4] = 3/k^2$. Thus, $E[\mu_a^2\mu_b^2] = 1/k^2$ if $a \neq b$ and $E[\mu_a^2\mu_b^2] = 3/k^2$ if $a = b$. Thus,

$$\begin{aligned} E[\nu_i'^2] &\leq \frac{1}{k^2} \sum_{a=1}^n \sum_{c=1}^n x_a x_a h_c h_c + \frac{1}{k^2} \sum_{a=1}^n \sum_{b=1}^n x_a x_b h_a h_b + \frac{1}{k^2} \sum_{a=1}^n \sum_{b=1}^n x_a x_b h_b h_a \\ &\quad + 3 \cdot \frac{2}{k^2} \sum_{a=1}^n x_a x_a h_a h_a - 2 \frac{3}{k^2} \sum_{a=1}^n x_a x_a h_a h_a \\ &= \frac{1}{k^2} (||x||^2 ||h||^2 + 2||xh||^2) = \frac{1 + 2\nu^2}{k^2}, \end{aligned}$$

as $||x|| = ||h|| = 1$. Finally,

$$\text{var}[\nu_i'] = E[\nu_i'^2] - (E[\nu_i'])^2 = \frac{1 + 2\nu^2}{k^2} - \frac{\nu^2}{k^2} = \frac{1 + \nu^2}{k^2}.$$

We conclude that $\text{var}[\nu'] = k \cdot \text{var}[\nu_i'] = \frac{1 + \nu^2}{k}$. ■

Using Chebyshev bound (details omitted) we get:

Lemma 3.7 *Let R be a random projection matrix as in Def. 2.2, $x' = Rx, h' = Rh, \nu = h^T x, \nu' = h'^T x'$. Then $\Pr[\text{sign}(\nu) \neq \text{sign}(\nu')] \leq \frac{2}{k\nu^2}$.*

Proof: Using the Chebyshev bound, we know that

$$\Pr\left[|\nu' - E[\nu']| \geq \frac{\varepsilon}{\sigma(\nu')} \sigma(\nu')\right] \leq \frac{(\sigma(\nu'))^2}{\varepsilon^2},$$

where $\sigma(\nu')$ is the standard deviation of ν' . Plugging in the bounds of Lemma 3.6, we have

$$\Pr[|\nu' - \nu| \geq \varepsilon] \leq \frac{2}{k\varepsilon^2}.$$

Now, the $\text{sign}(\nu) \neq \text{sign}(\nu')$ only if $|\nu' - \nu| \geq |\nu|$. Which implies that

$$P(\text{sign}(\nu) \neq \text{sign}(\nu')) \leq P(|\nu' - \nu| > |\nu|) \leq \frac{2}{k\nu^2}. \quad \blacksquare$$

Note that the difference between this and the result in Lemma 3.2 is that there we used the Chernoff bound, which is tighter for large values of k . For smaller values of k the above result will provide a better bound. This result can be further improved if the random projection matrix used has entries in $\{-1, +1\}$, using a variation of Lemma 3.6 along with a recent result [Ach01].

4 Analysis

Based on Lemma 3.2 and Lemma 3.7, the expected probability of error for a k -dimensional image of x , given that the point x is at distance $\nu(x)$ from the n -dimensional hyperplane (where the expectation is with respect to selecting a random projection) is given by

$$\min\left(3 \exp\left(-\frac{(\nu(x))^2 k}{8(2 + |(\nu(x))|)^2}\right), \frac{2}{kl^2}, 1\right). \quad (4)$$

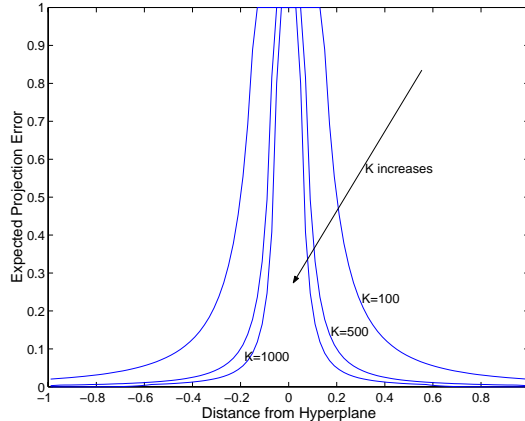


Figure 2: The contribution of data points to the generalization error as a function of their distance from the hyperplane.

This expression measures the contribution of the point to the generalization error, as a function of its distance from the hyperplane (and for a fixed k). Figure 2 shows this term (Eqn. 4) for different values of k . This bell shaped curve, depicting the results, exhibits that all points have some contribution to the error, and the relative contribution of a point decays exponentially fast as a function of its distance from the hyperplane. Given the probability distribution over the instance space and a fixed classifier, one can compute the distribution over the margin which is then used to compute the projection profile of the data as

$$\int_{x \in \mathcal{D}} \min \left(3 \exp \left(-\frac{(\nu(x))^2 k}{8(2 + |\nu(x)|)^2} \right), \frac{2}{kl^2}, 1 \right) d\mathcal{D}, \quad (5)$$

Consider, for example, the case when the distribution over the distance of the points from the hyperplane, D_l is normal with mean μ_l and variance σ_l . (Since the distance is bounded in $[0, 1]$, for the analysis we need to make sure that means and variances are such that no points lies outside this region.) In this case, the projection profile can be compute analytically. Figure 3 shows the bound for this case (mean = 0.3, variance = 0.1) as a function of k (the projected dimension of the data) and compares the VC-dimension term with the random projection term. It is evident that when the dimension of the data is very small, it is better to consider the VC-dimension based bounds but as soon as the dimension of the data increases, the VC-dimension term is much larger. Our bound can be thought of as doing a tradeoff between the two terms.

4.1 Comparison with Existing Bounds

The most basic generalization bound is the one derived from VC-theory [BEHW89]. The true error of an n -dimensional linear classifier whose empirical error on a sample of size m is $\hat{\epsilon}$ is bounded, with probability at least $1 - \delta$ by,

$$\epsilon \leq \hat{\epsilon} + \sqrt{\frac{(n+1)(\ln(\frac{2m}{n+1}) + 1) - \ln \delta/4}{m}}, \quad (6)$$

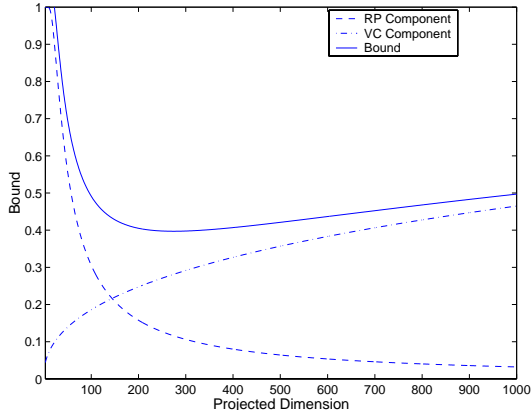


Figure 3: The bound obtained when the margin has a normal distribution. It shows the tradeoff between the VC-dimension and the random projection terms in the bound

where we use the fact that the VC-dimension of a linear classifier is $n + 1$. This bound is very general and gives the worst case generalization performance of the classifier. It depends only on the number of samples and the dimensionality of the data.

[ST98, STC99] have explored a new direction in which the margin of the data is used to derive bounds on the generalization error. Their bound depends on the fat-shattering function, $afat$, a generalization of the VC dimension, introduced in [KS94]. For sample size m the bound is given by:

$$\varepsilon \leq \frac{2}{m} \left(f \log_2(32m) \log_2 \frac{8em}{f} + \log_2 \frac{8m}{\delta} \right), \quad (7)$$

where $f = afat(\delta/8)$ and δ – the minimum margin of data points in the sample. For linear functions this is bounded by $(BR/\delta)^2$, where B is the norm of the classifier and R is the maximal norm of the data.

It is useful to observe the key difference between these two bounds; while the former depends only on the space in which the data lies and is totally independent of the actual data, the latter is independent of this space and depends on the performance (margin) of the classifier on the given data.

The new bound proposed here can be thought of as providing a link between the two existing bounds described above. The first component is a function of the data and independent of the true dimension whereas the second component is a function of the projected dimension of the data.

In most cases, the bounds in Eqn. 6, 7 are weak in the sense that the amount of data required before the bound is meaningful (< 0.5) is huge. For the VC-dimension based bounds, for example, the amount of data required for a meaningful bound is at least 17 times the dimension of the data. This is not feasible in many natural language processing and computer vision tasks where data dimensionality may be very high [Rot98]. Figure 4 provides some quantitative assessment of these claims. It gives the number of samples required before the fat-shattering bound is meaningful (< 0.5). This shows that even for the case when the margin is 0.9, almost a hundred thousand points need to be observed before the bounds are useful.

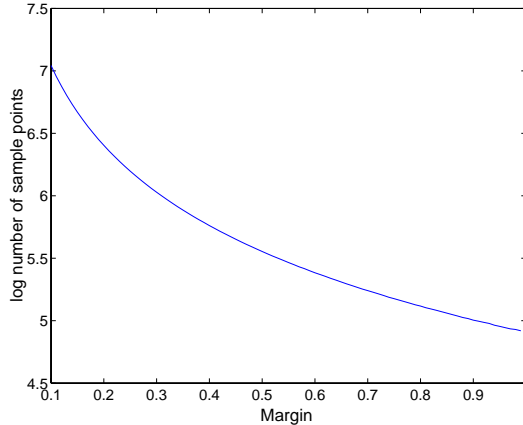


Figure 4: The number of sample points required before the fat-shattering bound in Eqn. 7 is meaningful, as a function of margin. Here the logarithm is in base 10.

We also compared the performance of the new bound with respect to the existing bounds on some real problems. In the first experiment we considered 17000 dimensional data taken from the problem of context sensitive spelling correction [GR99]. A winnow based algorithm, which was shown very successful on this problem, was used to learn a linear classifier. Figure 1 shows the histogram of the distance of the data with respect to the learned classifier. It is evident that a large number of data points are very close to the classifier and therefore the fat-shattering bounds are not useful. At the same time, to gain confidence from the VC-dimension based bounds, we need over 120,000 data points. Figure 1 shows the random projection term for this case; this term is below 0.5 already after 2000 samples and thus for the overall bound to be small, we need much less examples as compared to the VC-dimension case. The second experiment considered the problem of face detection. RBF³ kernel was used to learn the classifier. Figure 5(a) shows the histogram of the margin of the learned classifier and (b) gives the random projection term as a function of the dimension of the data.

5 Conclusions

We have presented a new analysis method for linear learning algorithms that uses random projections and margin distribution analysis. The main contribution of the work is using this method to develop a new data dependent complexity measure for learning and a bound on the true error of a learning algorithm, as a function of the margin distribution of the data relative to the learned classifier. While the random projection method was used in this paper as an analysis tool, one of the main directions of future research is to investigate algorithmic implications of the ideas presented here. In addition we plan to study the bounds on real data sets and develop a better understanding of the projection profile introduced here.

³Details and the extension of our theory to support the case of infinite dimensional space are omitted.

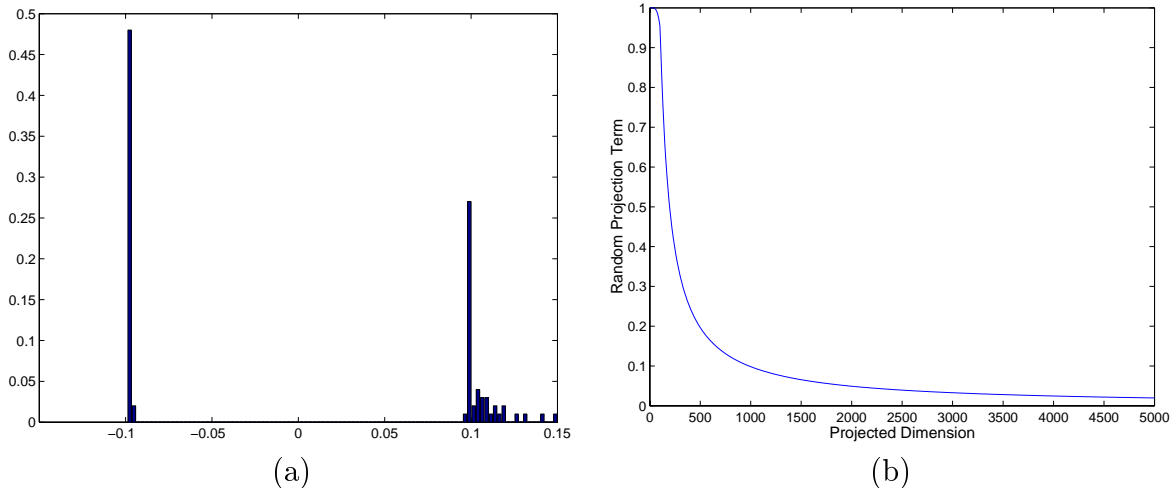


Figure 5: (a) Histogram of the distance of the points from the classifier for the face detection experiment. (b) The distortion error due to random projection as a function of the dimension of the projected space.

References

- [Ach01] D. Achlioptas. Database-friendly random projections. In *Symposium on Principles of Database Systems*, pages 274–281, 2001.
- [AV99] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proc. of the 40th Foundations of Computer Science*, pages 616–623, 1999.
- [BEHW89] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–865, 1989.
- [GR99] A. R. Golding and D. Roth. A Winnow based approach to context-sensitive spelling correction. *Machine Learning*, 34(1-3):107–130, 1999.
- [GR01] A. Garg and D. Roth. Learning coherent concepts. In *Proc. 12th Int. Workshop on Algorithmic Learning Theory*. Springer-Verlag, 2001. To Appear.
- [HG01] R. Herbrich and T. Graepel. A pac-bayesian margin bound for linear classifiers: Why svms work. In *Advances in Neural Information Processing Systems*, pages 224–230, 2001.
- [Ind01] P. Indyk. Algorithmic applications of low-distortion geometric embeddings. In *Proc. 42nd Annu. IEEE Sympos. Found. Comput. Sci.*, pages 10–31, 2001. Tutorial.
- [JL84] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In *Conference in modern analysis and probability*, pages 189–206, 1984.

- [KMNR97] M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27:7–50, 1997.
- [KS94] M. Kearns and R. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48:464–497, 1994.
- [Rot98] D. Roth. Learning to resolve natural language ambiguities: A unified approach. In *Proc. of the American Association of Artificial Intelligence*, pages 806–813, 1998.
- [RZ00] D. Roth and D. Zelenko. Towards a theory of coherent concepts. In *Proc. of the American Association of Artificial Intelligence*, pages 639–644, 2000.
- [ST98] J. Shawe-Taylor. Classification accuracy based on observed margin. *Algorithmica*, 22(1/2):157–172, 1998.
- [STC99] J. Shawe-Taylor and N. Cristianini. Further results on the margin distribution. In *Proc. 12th Annu. Conf. on Comput. Learning Theory*, pages 278–285. ACM Press, New York, NY, 1999.
- [STC00] J. Shawe-Taylor and N. Christianini. *An Introduction to Support Vector Machines: And Other kernel based methods*. Cambridge University Press, 2000.
- [Vap82] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.
- [Vap98] V. N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons Inc., New York, 1998.