Sparsity vs. Large Margins for Linear Classifiers

Ralf Herbrich
Statistics Research Group
Department of Computer Science
Technical University of Berlin
Berlin, Germany

Thore Graepel
Statistics Research Group
Department of Computer Science
Technical University of Berlin
Berlin, Germany

John Shawe-Taylor
Department of Computer Science
Royal Holloway
University of London
Egham, UK

Abstract

We provide small sample size bounds on the generalisation error of linear classifiers that take advantage of large observed margins on the training set *and* sparsity in the data dependent expansion coefficients. It is already known from results in the luckiness framework that both criteria independently have a large impact on the generalisation error. Our new results show that they can be combined which theoretically justifies learning algorithms like the Support Vector Machine [4] or the Relevance Vector Machine [12]. In contrast to previous studies we avoid using the classical technique of symmetrisation by a ghost sample but directly using the sparsity for the estimation of the generalisation error. We demonstrate that our result leads to practical useful results even in case of small sample size *if the training set witnesses our prior belief in sparsity and large margins.*

1 Introduction

In this paper we present a bound on the generalisation error of linear classifiers that takes advantage of the sparsity in terms of data dependent expansion coefficients *and* the margin attained at the given training set. It is already known that both criteria independently have an impact on the generalisation error of linear classifiers (see [13, 10]). We show that combining both criteria results in a bound that is tighter by orders of magnitudes and thus for the first time a practically useful bound for linear classifiers. Usually, bounds in the PAC framework are derived using a technique known as *symmetrisation by a ghost sample* [14], i.e. the probability over the random draw of the training set Z that there exists a classifier f with high generalisation error (larger than ε) but zero training error is upper bounded by twice the probability that there exists a classifier with zero training error on m iid examples but training error larger than $\frac{\varepsilon}{2}$ on a second ghost sample of size m drawn iid. This analysis then naturally leads to *covering numbers* for the function class because on a double sample of size $2m$ the number of different functions (in terms of training errors or attained margins) is finite (see [3, Lemma 4] or [11, Theorem 6.8]). For linear classifiers, a direct application of a lemma due to Alon et. al. [1] finally gives the margin bound in [10] having an additional $\log^2(m)$ factor. For comparison purposes we quote the bound here but using the slightly tighter bound on the fat shattering dimension contained in [2]. The result states that with probability at least $1-\delta$ over m randomly drawn samples Z , the generalisation error (see equation (2.4)) of a hyperplane with margin (see equation (2.2)) at least γ on the training set is bounded by

$$\frac{2}{m}\left(\left\lceil\frac{64\varsigma^2}{\gamma^2}\right\rceil\log\frac{8em}{\left\lceil\frac{64\varsigma^2}{\gamma^2}\right\rceil}\log(32m)+\log\left(\frac{8m}{\delta}\right)\right), \quad (1.1)$$

where ς is the radius of a ball containing the support of the distribution. Better results can be obtained by using tighter bounds on the covering numbers for linear function classes, in particular avoiding the double log factor, but even with these improvements the result will still give trivial bounds for most practical applications. We will demonstrate albeit artificial examples, where our new bound is non trivial with training set sizes as small as 300 .

Curiously, we shall totally avoid using the ghost sample technique. Conceptually, however, we *in fact* will make use of a ghost sample but from within the training set Z of size m . This can be accomplished by exploiting the sparseness of the classifier. If the classifier is determined by just d training points, we use the remaining $m-d$ points for testing the generalisation error of the classifier. This strategy was first proposed by Littlestone and Warmuth [6] for compression schemes. The novelty of the current paper is to combine their compression scheme argument with the large margin bounds on the growth function, resulting in bounds that are tighter than can be obtained by one or other approach on its own.

The paper is structured as follows: in the following section we will introduce the learning scenario we consider. In Section 3 we present our two main results together with some experiments. For the sake of readability the main proofs are delegated to Appendix A.

We denote vectors by bold letters, whereas scalars are typeset in roman letters. Random variables are typeset in sans serif font; vector spaces are denoted by calligraphic capitalised letters. The symbols P, E and I denote a probability measure, the expectation of a random variable and the indicator function, respectively.

2 Preliminaries

Suppose we are given a fixed domain $\mathcal{X}$ of objects together with a fixed set $\mathcal{Y} = \{-1, +1\}$ of classes -1 and $+1$ abbreviated by $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Furthermore let us assume that there exists a stationary distribution P_{Z} from which we generate iid training sets $Z = (X, Y)$ of size m . Given a fixed mapping $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{K}$ we know that there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}) \rangle_{\mathcal{K}}$ where k is known as the *kernel* for the fixed *feature* space $\mathcal{K}$. Alternatively, we could choose a symmetric positive definite function k so as to assure that there exists a fixed space $\mathcal{K}$ by Mercer's theorem [8]. For a given training set Z we define the set of classifiers considered for learning as

$$\mathcal{H}(Z) = \{\operatorname{sign}(f) : f \in \mathcal{F}(Z)\} , \tag{2.1}$$
$$\mathcal{F}(Z) = \left\{ \mathbf{x} \mapsto \sum_{i=1}^{m} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b : \boldsymbol{\alpha} \in \mathcal{A}, b \in \mathbb{R} \right\} .$$

Though this is a data dependent set of classifiers we know by Mercer's theorem that each f is a linear classifier in the space $\mathcal{K}$. As we often bound the probability that a subset $Z' \subseteq Z$ of size exactly $d \in \{1, \ldots, m\}$ has a certain property we introduce the following notation: the symbol $\mathbf{i}$ denotes the index vector $\mathbf{i} = (i_1, \ldots, i_d) \in \{1, \ldots, m\}^d$ of d *distinct* indices $i_1 < i_2 < \cdots < i_d$ from the set $\{1, \ldots, m\}$. We use I_d to denote the set of all index vectors $\mathbf{i}$ of $\{1, \ldots, m\}$ of size d . Given a training set Z of size m we denote by $Z_{\mathbf{i}} = \{(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_d}, y_{i_d})\} \subseteq Z$ the subset of size d obtained by selecting the i_1 –th to i_d –th element from Z . A *learning algorithm* $L : \bigcup_{m=1}^{\infty} \mathcal{Z}^m \mapsto \mathcal{A}$ assigns a training set Z to a vector $\boldsymbol{\alpha}$ of coefficients $L(Z)$ in $\mathcal{A} \subseteq \mathbb{R}^m$ and is assumed to be invariant under permutation of the sample. We assume that the setting of the threshold b can then be inferred by a fixed rule from the examples. We denote by $\mathbf{i}_{L(Z)}$ the set of indices for which the coefficients are non zero, and by $f_{L(Z)} = f^b_{L(Z)}$ the corresponding function with appropriately chosen threshold b . If a learning algorithm L is applied to a subset $Z' \subset Z$ of the training set Z of size m we assume that L assigns all corresponding coefficients α_i not present in Z' to zero, i.e.

$$\forall i \in \{1, \ldots, m\} : (\mathbf{x}_i, y_i) \in (Z \setminus Z') \Rightarrow (L(Z'))_i = 0 .$$

Furthermore we assume that if the learning algorithm L is applied to $Z_{\mathbf{i}_{L(Z)}}$ the result obtained is $L(Z)$, that is the same function (and threshold) is reconstructed from the subsample. Note that this implies that the function $f_{L(Z)}$ is determined by the subsample $Z_{\mathbf{i}_{L(Z)}}$. Hence two distinct dichotomies of the same inputs must give rise to distinct sets of indices. Given a training set Z , the *margin* $\gamma_Z(\boldsymbol{\alpha}, b)$ is defined by

$$\gamma_Z(\boldsymbol{\alpha}, b) = \min_{(\mathbf{x}_i, y_i) \in Z} \left(\frac{y_i f^b_{\boldsymbol{\alpha}}(\mathbf{x}_i)}{\|f^b_{\boldsymbol{\alpha}}\|} \right) \tag{2.2}$$
$$\left\|f^b_{\boldsymbol{\alpha}}\right\|^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_i y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) .$$

Normally, the threshold b will be chosen to maximise $\gamma_Z(\boldsymbol{\alpha}, b)$ for the set of training examples. We define the *training error* $R_{\text{emp}}[f, Z]$ of a classifier f on a given training set Z by

$$R_{\text{emp}}[f, Z] = \frac{1}{m} \left| \{(\mathbf{x}_i, y_i) \in Z : y_i f(\mathbf{x}_i) \leq 0\} \right| . \tag{2.3}$$

Accordingly, the *generalisation error* $R[f]$ of a classifier f is defined by

$$R[f] = \mathsf{P}_{\mathsf{XY}}(\mathsf{Y} \cdot f(\mathsf{X}) \leq 0) . \tag{2.4}$$

We will also be interested in the following *conditional generalisation error* $R_\gamma[f]$ given that a margin γ is observed on a test example

$$R_\gamma[f] = \mathsf{P}_{\mathsf{XY}| \, |f(\mathsf{X})| \geq \gamma}(\mathsf{Y} \cdot f(\mathsf{X}) \leq 0) . \tag{2.5}$$

Our ultimate interest is to obtain bounds on $R[f]$ given only the observable training error $R_{\text{emp}}[f, Z]$ and some easy–to–determine complexity measure of f , e.g. the margin or the sparsity in terms of $\|L(Z)\|_0$. In course of derivation of such bounds we often proceed as follows: assuming a *fixed* value of the complexity measure shall allow us to determine that with high probability (at least $1 - \delta$) the generalisation error will be small (not more than $\varepsilon(\delta)$). In order to plug in the *observed* value of the complexity measure we *stratify* over all s possible values of the complexity measure using the following stratification lemma.

Lemma 2.1 (Stratification Lemma)**.** *Suppose we are given s logical formulas $\Upsilon_i : \mathcal{Z}^m \times \mathbb{R} \mapsto \{true, false\}$ such that*

$$\forall i \in \{1, \ldots, s\} \, \forall \delta \in [0, 1] : \mathsf{P}_{\mathsf{Z}^m}(\Upsilon_i(\mathsf{Z}, \delta)) \geq 1 - \delta .$$

Then for any set $p_1, \ldots, p_s$ of s positive numbers whose sum is upper bounded by one

$$\forall \delta \in [0, 1] : \mathsf{P}_{\mathsf{Z}^m}(\Upsilon_1(\mathsf{Z}, \delta p_1) \wedge \ldots \wedge \Upsilon_s(\mathsf{Z}, \delta p_s)) \geq 1 - \delta .$$

Note that for the stratification we can encode some *prior belief* which complexity value we expect to observe using positive real numbers p_i that sum up to at most one. This idea allows to combine Bayesian priors (the numbers p_i) with PAC bounds finally leading to PAC–Bayesian theorems (see [7] for details).

3 A Sparse Margin Bound

The core idea to obtain a generalisation error bound for a fixed learning algorithm is to exploit the (assumed) sparseness of a the returned linear classifier $f_{L(Z)}$ because if the learned classifier uses d training points,

i.e. $\|L(Z)\|_0 = d$, but has also large margins with correct classification on the remaining $m-d$ points, the latter can effectively be used as iid test points. The number of equivalence classes is then determined by the margin $\gamma_Z(L(Z),b) \geq \gamma$ attained on the whole training set Z . Note that the condition of a margin γ on the correctly classified $m-d$ points forces us to consider the conditional generalisation error $R_\gamma[f^b_{L(Z)}]$ rather than the more usual quantity — the generalisation error $R[f^b_{L(Z)}]$.

Lemma 3.1 (Margin Compression Lemma)**.** *Fix $\gamma \in (0,\varsigma)$, $d \in \{1,\ldots,m\}$ and a learning algorithm L . For any measure $\mathbf{P_Z}$ such that $\mathbf{P_X}(\{\mathbf{x} : \|\boldsymbol{\phi}(\mathbf{x})\|_{\mathcal{K}} \leq \varsigma\}) = 1$ the probability that m examples Z drawn iid according to $\mathbf{P_Z}$ contain a subset $Z_d \subseteq Z$ of exactly d examples and the linear classifier $f^b_{L(Z_d)}$ achieves a margin $\gamma_Z(L(Z_d),b)$ of at least γ and has conditional generalisation error $R_\gamma[f^b_{L(Z_d)}]$ larger than ε is less than*

$$\left(\frac{em}{\kappa}\right)^\kappa \exp\{-\varepsilon(m-d)\} ,$$

where $\kappa = \left\lceil \left(\frac{\varsigma}{\gamma}\right)^2 \right\rceil < m$.

The lemma implies the following statement that holds with probability at least $1-\delta$ over the random draw of the training set Z :

$$\forall Z_d \subset Z : (|Z_d| \neq d) \vee \left(\left\lceil \left(\frac{\varsigma}{\gamma}\right)^2 \right\rceil \neq \kappa \right) \vee (\gamma_Z(L(Z_d),b) < \gamma) \vee \left(R_\gamma\left[f^b_{L(Z_d)}\right] \leq \frac{\kappa \ln\left(\frac{em}{\kappa}\right) + \ln\left(\frac{1}{\delta}\right)}{m-d} \right) . \quad (3.1)$$

Using a double stratification over possible values of d and κ gives the following powerful theorem.

Theorem 3.2 (Sparse margin conditional bound)**.** *Fix a learning algorithm L . For any measure $\mathbf{P_Z}$ such that $\mathbf{P_X}(\{\mathbf{x} : \|\boldsymbol{\phi}(\mathbf{x})\|_{\mathcal{K}} \leq \varsigma\}) = 1$, with probability at least $1-\delta$ over the random draw of the training set Z of size m for all linear classifier $f^b_{L(Z)}$ that have margin $\gamma_Z(L(Z),b) = \gamma$ the conditional generalisation error $R_\gamma[f^b_{L(Z)}]$ is bounded from above by*

$$\frac{\left\lceil \left(\frac{\varsigma}{\gamma}\right)^2 \right\rceil \ln\left(\frac{em}{\left\lceil \left(\frac{\varsigma}{\gamma}\right)^2 \right\rceil}\right) + 2\ln(m) + \ln\left(\frac{1}{\delta}\right)}{m-d} , \quad (3.2)$$

provided $d = \|L(Z)\|_0 > 0$ *and* $\gamma > \frac{\varsigma}{\sqrt{m}}$.

Proof. The proof is obtained by an application of Lemma 2.1 to equation (3.1) using the sequence $p_d = \frac{1}{m}$ and $p_\kappa = \frac{1}{m}$. Note that κ is by definition always strictly positive. □

The two results given above only cover the conditional generalisation error. This may be useful if we are willing to discard test points falling within γ of the margin. We can, however, use these as intermediate results to obtain the following sparse margin bound on the generalisation error.

Theorem 3.3 (Sparse margin bound)**.** *Fix a learning algorithm L . For any measure $\mathbf{P_Z}$ such that all points are contained in a ball of radius ς in feature space $\mathcal{K}$, i.e. $\mathbf{P_X}(\{\mathbf{x} : \|\boldsymbol{\phi}(\mathbf{x})\|_{\mathcal{K}} \leq \varsigma\}) = 1$, with probability at least $1-\delta$ over the random draw of the training set Z of size m for all linear classifier $f^b_{L(Z)}$ that have margin $\gamma_Z(L(Z),b) = \gamma$ the generalisation error $R[f^b_{L(Z)}]$ is bounded from above by*

$$2\frac{\left\lceil \left(\frac{2\varsigma}{\gamma}\right)^2 \right\rceil \ln\left(\frac{em}{\left\lceil \left(\frac{2\varsigma}{\gamma}\right)^2 \right\rceil}\right) + 2\ln(m) + \ln\left(\frac{2}{\delta}\right)}{m-d} , \quad (3.3)$$

provided $d = \|L(Z)\|_0 > 0$ *and* $\gamma > \frac{\varsigma}{\sqrt{m}}$.

Proof. The result is obtained by an application of Theorem 3.2 with two different settings of the value for the threshold b . Since $\gamma_Z(L(Z),b) = \gamma$, we can take a threshold of

$$b' = b + \frac{\gamma}{2\left\|f^b_{L(Z)}\right\|} ,$$

and still obtain a margin of $\frac{\gamma}{2}$. Since we are assuming the threshold b was chosen by a fixed rule, we can adapt the rule to choose b' . The same applies for the threshold

$$b'' = b - \frac{\gamma}{2\left\|f^b_{L(Z)}\right\|} .$$

We now apply the theorem for these two threshold values using $\frac{\delta}{2}$ in place of δ . In both cases we have a margin of at least $\frac{\gamma}{2}$, hence obtaining a conditional generalisation error of

$$R_{\frac{\gamma}{2}}\left[f^{b'}_{L(Z)}\right] \leq \frac{\left\lceil \left(\frac{2\varsigma}{\gamma}\right)^2 \right\rceil \ln\left(\frac{em}{\left\lceil \left(\frac{2\varsigma}{\gamma}\right)^2 \right\rceil}\right) + 2\ln(m) + \ln\left(\frac{2}{\delta}\right)}{m-d} ,$$

Now consider the errors made by the classifier $f^b_{L(Z)}$. Since any test point must have margin $\frac{\gamma}{2}$ for either $f^{b'}_{L(Z)}$ or $f^{b''}_{L(Z)}$, a misclassified point must be conditionally misclassified by one of these two functions. Hence the probability of a randomly drawn test point being misclassified by $f^b_{L(Z)}$ is bounded by the sum of the $\frac{\gamma}{2}$ conditional error bounds for the functions $f^{b'}_{L(Z)}$ and $f^{b''}_{L(Z)}$. The result follows. □

In order to check the practical usefulness of the bounds (3.2) and (3.3) we generated training sets on the unit sphere in $\mathbb{R}^{50}$ using normalised points $\mathbf{x}$ drawn according to two multidimensional Gaussian with mean vectors $\boldsymbol{\mu}_{+1} = (1,\ldots,1)'$ and $\boldsymbol{\mu}_{-1} = (-1,\ldots,-1)'$ and the same covariance matrix $\sigma^2\mathbf{I}$. For the determination

σ^2	eqn. (3.2)	$\widehat{R}_\gamma[f^b_{L(Z)}]$	eqn. (3.3)	$\widehat{R}[f^b_{L(Z)}]$
0.1	0.29	0.00	0.69	0.00
1	0.34	0.00	0.83	0.00
10	1.00	0.07	1.00	0.07

Table 1: Bound values of Theorem 3.2 ($m = 100$) and 3.3 ($m = 300$) over 100 random draws of the training set with $\delta = 0.05$. Small values of σ^2 lead to training sets that can be separated with a large margin.

of the classes we applied the fixed rule $y = \text{sign}\left(\sum_{i=1}^{50} x_i\right)$. Our learning algorithm maximises the margin using only $0.25 \cdot m$ training points. In Table 1 we see that even for very small training set sizes, e.g. $m = 100$, our bound provides non trivial values *if σ^2 was such that with high probability the margin as well as the sparsity is large.* Note that the sparsity of 75 % alone does not suffice to give non trivial generalisation error bounds (see [2]). Similarly, in order to achieve a bound value of 0.29 for maximally large margins $\gamma_Z(L(Z), b) = \varsigma$ with the classical margin bound given by equation (1.1) we would need the astronomical number of $m = 153\,892$ as the minimal training set size.

4 Conclusion

In this paper we have proven a conditional generalisation and a generalisation error bound for linear classifiers both in terms of margins and sparsity. The novelty with the approach is the avoidance of moving to a ghost sample by using the points not appearing in the sparse representation as test points. By using this technique we are able to avoid using covering number bounds working instead with the VC dimension of large margin hyperplanes together with Sauer's Lemma. The result is a bound which is significantly tighter than previous large margin bounds and indeed many standard PAC results.

A Proofs

A.1 Proof of Lemma 2.1

Proof. The proof is a simple union bound argument. By definition

$$\begin{aligned}
&\forall \delta \in [0,1] : \mathbf{P}_{\mathsf{Z}^m}\left(\Upsilon_1(\mathsf{Z}, \delta p_1) \wedge \ldots \wedge \Upsilon_s(\mathsf{Z}, \delta p_s)\right) \\
&\quad = 1 - \mathbf{P}_{\mathsf{Z}^m}\left(\neg\Upsilon_1(\mathsf{Z}, \delta p_1) \vee \ldots \vee \neg\Upsilon_s(\mathsf{Z}, \delta p_s)\right) \\
&\quad \geq 1 - \sum_{i=1}^{s} \mathbf{P}_{\mathsf{Z}^m}\left(\neg\Upsilon_i(\mathsf{Z}, \delta p_i)\right) > 1 - \sum_{i=1}^{s} \delta p_i \\
&\quad \geq 1 - \delta\,.
\end{aligned}$$

□

A.2 Proof of Lemma 3.1

Before proving the lemma we recall the following bound on the number of dichotomies realisable with hyperplanes having margin γ (see [13, p. 128] and [2, 5, 9] for details).

Lemma A.1 (VC dimension of hyperplanes)**.** *For any measure $\mathbf{P}_{\mathsf{Z}}$ such that $\mathbf{P}_{\mathsf{X}}(\{\mathbf{x} : \|\boldsymbol{\phi}(\mathbf{x})\|_{\mathcal{K}} \leq \varsigma\}) = 1$ the number of different classifications Y realisable on m randomly drawn points X by a linear classifier $f^b_{\boldsymbol{\alpha}}$ of the form (2.1) with $\gamma_{(X,Y)}(\boldsymbol{\alpha}, b) \geq \gamma$ is bounded from above by*

$$\left(\frac{em}{\kappa}\right)^{\kappa},$$

where $\kappa = \left\lceil \left(\frac{\varsigma}{\gamma}\right)^2 \right\rceil < m$.

Proof of Lemma 3.1. We exploit the idea that for a fixed value of d there are still $m - d$ points drawn iid according to $\mathbf{P}_{\mathsf{Z}}$ on which the classifier $f^b_{L(Z_d)}$ has to succeed. For a fixed index set $\mathbf{i} \in I_d$ and margin γ we define the propositions

$$\begin{aligned}
A^c_{\mathbf{i}}(Z) &\equiv \left(\mathbf{i}_{L(Z)} = \mathbf{i}\right) \wedge \left(\gamma_Z\left(L(Z_{\mathbf{i}}), b\right) \geq \gamma\right) \wedge \\
&\qquad \left(R_\gamma\left[f^b_{L(Z_{\mathbf{i}})}\right] > \varepsilon\right), \\
A_{\mathbf{i}}(Z) &\equiv \left(\mathbf{i}_{L(Z')} = \mathbf{i}\right) \wedge \left(\gamma_{Z'}\left(L(Z_{\mathbf{i}}), b\right) \geq \gamma\right) \wedge \\
&\qquad \left(R_\gamma\left[f^b_{L(Z_{\mathbf{i}})}\right] > \varepsilon\right),
\end{aligned}$$

where with examples $Z' = Z[f_{L(Z_{\mathbf{i}})}]$ we denote the set of training examples Z relabelled using the function $f_{L(Z_{\mathbf{i}})}$. The idea behind these definition is that we use the examples indexed by $\mathbf{i}$ to find a hypothesis. For the first proposition this is the hypothesis for the whole training set (the exponent c indicates consistency). For the second proposition it is the hypothesis for the whole training set when appropriately relabelled. We wish to bound the probability

$$\begin{aligned}
\mathbf{P}_{\mathsf{Z}^m}\left(\exists \mathbf{i} \in I_d : A^c_{\mathbf{i}}(\mathsf{Z})\right) &\leq \sum_{\mathbf{i} \in I_d} \mathbf{P}_{\mathsf{Z}^m}\left(A^c_{\mathbf{i}}(\mathsf{Z})\right) \\
&= \binom{m}{d} \mathbf{P}_{\mathsf{Z}^m}\left(A^c_{\mathbf{i}_0}(\mathsf{Z})\right),
\end{aligned}$$

where $\mathbf{i}_0 = \{1, \ldots, d\}$ which follows from the union bound. The event $A^c_{\mathbf{i}_0}(Z)$ can be decomposed as follows

$$A^c_{\mathbf{i}_0}(Z) \equiv A_{\mathbf{i}_0}(Z) \bigwedge_{j=d+1}^{m} \left(y_j f^b_{L(Z_{\mathbf{i}_0})}(\mathbf{x}_j) \geq \gamma\right).$$

Hence, we can now write $\mathbf{P}_{\mathsf{Z}^m}\left(A^c_{\mathbf{i}_0}(\mathsf{Z})\right)$ as

$$\begin{aligned}
&\mathbf{P}_{\mathsf{Z}^m}\left(A_{\mathbf{i}_0}(\mathsf{Z})\right) \mathbf{P}_{\mathsf{Z}^m | A_{\mathbf{i}_0}(\mathsf{Z})}\left(\bigwedge_{j=d+1}^{m} \mathsf{Y}_j f^b_{L(Z_{\mathbf{i}_0})}(\mathsf{X}_j) \geq \gamma\right) \\
&= \mathbf{P}_{\mathsf{Z}^m}\left(A_{\mathbf{i}_0}(\mathsf{Z})\right) \prod_{j=d+1}^{m} \mathbf{P}_{\mathsf{Z}^m | A_{\mathbf{i}_0}(\mathsf{Z})}\left(\mathsf{Y}_j f^b_{L(\mathsf{Z}_{\mathbf{i}_0})}(\mathsf{X}_j) \geq \gamma\right).
\end{aligned}$$

By the independence assumption and the fact that the effect of the conditional probability is the same as the conditional generalisation error $R_\gamma[f^b_{L(Z_{\mathbf{i}_0})}] > \varepsilon$, each

factor in the product is less than $(1-\varepsilon)$ so that we obtain

$$\begin{aligned}\mathbf{P}_{\mathbf{Z}^m}\left(A^c_{\mathbf{i}_0}(\mathbf{Z})\right) &< \mathbf{P}_{\mathbf{Z}^m}\left(A_{\mathbf{i}_0}(\mathbf{Z})\right)(1-\varepsilon)^{m-d} \\ &\leq \mathbf{P}_{\mathbf{Z}^m}\left(A_{\mathbf{i}_0}(\mathbf{Z})\right)\exp\left\{-\varepsilon(m-d)\right\},\end{aligned}$$

where we have used $\forall\varepsilon \in [0,1] : (1-\varepsilon) \leq \exp\{-\varepsilon\}$. Let Σ be the set of permutations U of the m examples. By the invariance of the probability under permutations of the sample, we can now write the probability $\mathbf{P}_{\mathbf{Z}^m}(A_{\mathbf{i}_0})$ as follows

$$\begin{aligned}\mathbf{P}_{\mathbf{Z}^m}\left(A_{\mathbf{i}_0}(\mathbf{Z})\right) &= \mathbf{E}_{\mathbf{U}}\left[\mathbf{E}_{\mathbf{Z}^m|\mathbf{U}=U}\left[\mathbf{I}_{A_{\mathbf{i}_0}(U(\mathbf{Z}))}\right]\right] \\ &= \mathbf{E}_{\mathbf{Z}^m}\left[\mathbf{E}_{\mathbf{U}|\mathbf{Z}^m=Z}\left[\mathbf{I}_{A_{\mathbf{i}_0}(\mathbf{U}(Z))}\right]\right] \\ &= \mathbf{E}_{\mathbf{Z}^m}\left[\frac{1}{m!}\sum_{U\in\Sigma}\mathbf{I}_{A_{\mathbf{i}_0}(U(Z))}\right],\end{aligned}$$

where we have used the uniform measure $\mathbf{P}_{\mathbf{U}}$ over the $m!$ possible permutations. We now bound the number of non zero summands for a fixed set Z . First observe that if the summand is non zero for some $U \in \Sigma$, then all the $d!(m-d)!$ permutations U' realising the same split, that is with $U'(Z)_{\mathbf{i}_0} = U(Z)_{\mathbf{i}_0}$, have non zero summands. We must therefore bound the number of index vectors $\mathbf{i} \in I_d$ of d examples that when placed in the first d positions give a non zero summand. For each such set $\mathbf{i}$ we have $\gamma_{Z'}(L(Z_{\mathbf{i}}), b) \geq \gamma$ and $\mathbf{i}_{L(Z')} = \mathbf{i}$, where $Z' = Z[f^b_{L(Z_{\mathbf{i}})}]$. Since the set $\mathbf{i}_{L(Z')}$ is uniquely determined by Z' , distinct sets must correspond to distinct dichotomies of the examples in Z , each of which is realised with margin at least γ . Thus, by Lemma A.1 the number of non zero summands cannot exceed

$$\left(\frac{em}{\kappa}\right)^{\kappa} d!\,(m-d)!\,.$$

Putting together the partial results we obtain

$$\begin{aligned}\mathbf{P}_{\mathbf{Z}^m}\left(\exists\mathbf{i}\in I_d : A^c_{\mathbf{i}}(\mathbf{Z})\right) &\leq \binom{m}{d}\mathbf{P}_{\mathbf{Z}^m}\left(A^c_{\mathbf{i}_0}(\mathbf{Z})\right) \\ &< \binom{m}{d}\mathbf{P}_{\mathbf{Z}^m}\left(A_{\mathbf{i}_0}(\mathbf{Z})\right)\exp\left\{-\varepsilon(m-d)\right\} \\ &\leq \binom{m}{d}\left(\frac{em}{\kappa}\right)^{\kappa}\frac{d!\,(m-d)!}{m!}\exp\left\{-\varepsilon(m-d)\right\} \\ &= \left(\frac{em}{\kappa}\right)^{\kappa}\exp\left\{-\varepsilon(m-d)\right\}.\end{aligned}$$

□

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale sensitive dimensions, uniform convergence and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[2] P. Bartlett and J. Shawe-Taylor. Generalization performance of Support Vector Machines and other pattern classifiers. In *Advances in Kernel Methods — Support Vector Learning*, pages 43–54. MIT Press, 1998.

[3] P. L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[4] C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, 20:273–297, 1995.

[5] L. Gurvits. A note on a scale–sensitive dimension of linear bounded functionals in banach spaces. In *Proceedings of Algorithm Learning Theory, ALT-97*, 1997.

[6] N. Littlestone and M. Warmuth. Relating data compression and learnability. Technical report, University of California Santa Cruz, 1986.

[7] D. A. McAllester. Some PAC Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, Madison, Wisconsin, 1998.

[8] T. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Transaction of London Philosophy Society (A)*, 209:415–446, 1909.

[9] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972.

[10] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data–dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[11] J. Shawe-Taylor and N. Cristianini. Robust bounds on generalization from the margin distribution. Technical report, Royal Holloway, University of London, 1998. NC2–TR–1998–029.

[12] M. E. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems 12*, pages 652–658, San Mateo, CA, 2000. Morgan Kaufmann.

[13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[14] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Application*, 16(2):264–281, 1971.