Generalisation Error Bounds for Sparse Linear Classifiers

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

Ralf Herbrich
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 are sparse in their dual representation given by the expansion coefficients of the weight vector in terms of the training data. These results theoretically justify algorithms like the Support Vector Machine, the Relevance Vector Machine and K-nearest-neighbour. The bounds are a-posteriori bounds to be evaluated after learning when the attained level of sparsity is known. In a PAC-Bayesian style prior knowledge about the expected sparsity is incorporated into the bounds. The proofs avoid the use of double sample arguments by taking into account the sparsity that leaves unused training points for the evaluation of classifiers. We furthermore give a PAC-Bayesian bound on the average generalisation error over subsets of parameter space that may pave the way combining sparsity in the expansion coefficients and margin in a single bound. Finally, reinterpreting a mistake bound for the classical perceptron algorithm due to Novikoff we demonstrate that our new results put classifiers found by this algorithm on a firm theoretical basis.

1 Introduction

Sparseness in the representation of knowledge has long been considered advantageous. While sparsity in the original features is addressed in feature selection [11] we deal with a different kind of sparsity. Many learning algorithms are based on a dual representation of linear classifiers: The weight vector is represented as a linear combination of input vectors in a kernel-space whose existence can be ensured by the application of Mercer kernels. Examples are the Support Vector Machine (SVM) [2], the Relevance Vector Machine (RVM) [13] and the K-nearest-neighbour (KNN) classifier [3, 4], that can be viewed as a linear classifier in a collapsed kernel space in the limit of vanishing kernel bandwidth.

If a classifier is represented in terms of only a subset of the training sample and succeeds on the remaining training data it effectively compresses the sample [5]. We derive *a posteriori* results to be evaluated *after* learning by combining bounds under *prior* expectations, e.g. about the attained sparsity. In particular, we consider the complexity of *hypothesis classes* only w.r.t. particular *learning algorithms*. The sparsity allows us to avoid the double sample argument of the Basic Lemma [14]. In addition, we present a PAC-Bayesian theorem [9] about the average generalisation error over a subset of version space.

Finally, we reinterpret Novikoff's well known perceptron convergence theorem [12] as a *sparsity guarantee* for the classifier found by the well known perceptron learning algorithm: *the mere existence of large margin classifiers implies the existence of sparse consistent classifiers*. By combining the perceptron mistake bound with a compression bound that originated from the work of Littlestone and Warmuth [8] we are able to provide a PAC like generalisation error bound for the classifier found by the perceptron algorithm whose size is determined by the magnitude of the maximally achievable margin on the dataset.

The paper is structured as follows: In Section 2 we introduce the basic learning setting and provide the stratification lemma that will enable us later to combine bounds using prior knowledge. In Section 3 we give sparsity bounds for the zero-error case and the agnostic case. In Section 4 we give the PAC-Bayesian subset bound for sparse classifiers. Finally, we present an application of the sparsity result for the zero-error case for classifiers found by the perceptron learning algorithm. Most of the proofs have been delegated to the appendix.

2 Preliminaries

We assume a fixed domain $\mathcal{X}$ of objects together with a fixed set $\mathcal{Y} = \{-1, +1\}$ and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ from which we draw a training sample $Z = (X, Y)$ of size m iid according to $\mathsf{P}_Z \equiv \mathsf{P}_{XY}$. Given a fixed mapping $\phi : \mathcal{X} \to \mathcal{K}$ we know that there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_{\mathcal{K}}$ where k is known as the *kernel* for the fixed *feature* space $\mathcal{K}$ [10]. For a given

training set Z we define the set of classifiers considered

$$\begin{aligned}\mathcal{H}(Z) &= \{\operatorname{sign}(f) : f \in \mathcal{F}(Z)\}, \\ \mathcal{F}(Z) &= \left\{\mathbf{x} \mapsto \sum_{i=1}^{m} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) : \boldsymbol{\alpha} \in \mathcal{A}\right\}.\end{aligned}$$

A *learning algorithm* $L : \bigcup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{A}$ maps from a training set Z to a vector $\boldsymbol{\alpha}$ of coefficients in $\mathcal{A} \subseteq \mathbb{R}^m$, where the resulting hypothesis is assumed invariant under the permutation of the training set. 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 zero to all corresponding coefficients α_i not present in Z'. We define the *training error* $R_{\text{emp}}[f, Z]$ of a classifier f on a given training sample Z by

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

and the *generalisation error* $R[f]$ of a classifier f by

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

Finally, the *version space* $V(Z)$ for a given training set Z is

$$V(Z) = \{\boldsymbol{\alpha} \in \mathcal{A} : R_{\text{emp}}[f_{\boldsymbol{\alpha}}, Z] = 0\}.$$

Our goal is a bound on $R[f]$ given only $R_{\text{emp}}[f, Z]$ and some easy-to-determine complexity measure of f. Assuming a *fixed* value of the complexity measure we prove that with high probability (at least $1 - \delta$) the generalisation error will be small (not more than $\epsilon(\delta)$). In order to plug in the *observed* value of the complexity measure we *stratify* over all r possible values of the complexity measure thereby encoding *prior belief* about which complexity value we expect to observe using probabilities p_i. Thus we combine a Bayesian prior (the numbers p_i) with PAC bounds leading to PAC-Bayesian theorems (see [9]).

Lemma 1 (Stratification Lemma). *Suppose we are given r logical formulas $\Upsilon_i : \mathcal{Z}^m \times \mathbb{R} \mapsto \{true, false\}$ such that*

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

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

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

3 Sparsity Bounds

3.1 The Zero Error Case

Let us start with a sparsity bound for classifiers f with $R_{\text{emp}}[f, Z] = 0$ using the following lemma [5].

Lemma 2 (Compression lemma). *Fix $d \in \{1, \ldots, m\}$ and a learning algorithm L. For any measure $\mathbf{P}_{\mathsf{Z}}$, the probability that m examples Z drawn iid according to $\mathbf{P}_{\mathsf{Z}}$ contain a subset $Z_d \subseteq Z$ of exactly d examples and the linear classifier $f_{L(Z_d)}$ is both consistent with Z and has generalisation error $R[f_{L(Z_d)}]$ larger than ε is at most*

$$\binom{m}{d} (1 - \varepsilon)^{m-d}.$$

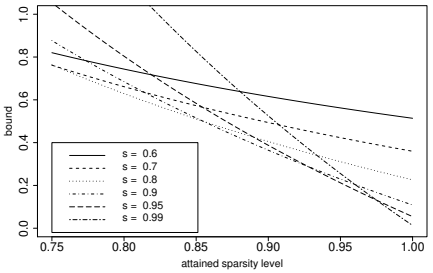

Figure 3.1: Bound values of Theorem 3 vs. attained sparsity level $\widehat{s}(Z)$ for $m = 1000$ and $\delta = 0.05$.

Using equation (7.1) the lemma implies that the following statement holds with probability at least $1 - \delta$ over the random draw of the training set Z:

$$\forall Z_d \subset Z : (|Z_d| \neq d) \vee \left(R_{\text{emp}}\left[f_{L(Z_d)}, Z\right] \neq 0\right) \vee \left(R\left[f_{L(Z_d)}\right] \leq \frac{\ln\left(\binom{m}{d}\right) + \ln\left(\frac{1}{\delta}\right)}{m - d}\right). \tag{3.1}$$

By a simple application of Lemma 1 we obtain our first sparsity bound.

Theorem 3 (Sparsity bound). *Fix a learning algorithm L and $s \in (0,1)$. For any measure $\mathbf{P}_{\mathsf{Z}}$, with probability $1 - \delta$ over the random draw of the training set Z of size m for all linear classifier $f_{L(Z)}$ that have zero training error $R_{\text{emp}}\left[f_{L(Z)}, Z\right] = 0$, the generalisation error $R\left[f_{L(Z)}\right]$ is bounded from above by*

$$\ln\left(\frac{1}{s}\right) + \frac{d \cdot \ln\left(\frac{1}{1-s}\right) + \ln(1 - s^m) + \ln\left(\frac{1}{\delta}\right)}{m - d}, \tag{3.2}$$

provided $d = \|L(Z)\|_0 > 0$.

Proof. Apply Lemma 1 to equation (3.1) using the sequence

$$p_d = \frac{\binom{m}{d} s^{m-d} (1 - s)^d}{1 - s^m}, \tag{3.3}$$

where s expresses our belief in sparsity. □

In terms of the *attained sparsity level* $\widehat{s}(Z) = 1 - \frac{\|L(Z)\|_0}{m}$ we have that with probability at least $1 - \delta$ over the random draw of the training set Z for all linear classifiers $f_{L(Z)}$ with zero training error $R_{\text{emp}}\left[f_{L(Z)}, Z\right] = 0$

$$R\left[f_{L(Z)}\right] \leq \ln\left(\frac{1-s}{s}\right) + \ln\left(\frac{1}{1-s}\right) \frac{1}{\widehat{s}(Z)} + \frac{\ln(1 - s^m) + \ln\left(\frac{1}{\delta}\right)}{m\widehat{s}(Z)}.$$

A good match of s and $\widehat{s}(Z)$ leads to low bound values as can be seen in Figure 3.1.

3.2 The Agnostic Case

In the case of non-zero training error we use the following analog of Lemma 2.

Lemma 4 (Agnostic compression lemma). *Fix $d \in \{1,\ldots,m\}$, $q \in \{1,\ldots,m-d\}$ and a learning algorithm L. For any measure P_Z, the probability that m examples Z drawn iid according to P_Z contain a subset $Z_d \subseteq Z$ of exactly d examples and the linear classifier $f_{L(Z_d)}$ has a training error $R_{\rm emp}\left[f_{L(Z_d)}, Z\right] \leq \frac{q}{m}$ on Z and has generalisation error $R\left[f_{L(Z_d)}\right]$ larger than ε is at most*

$$\binom{m}{d} \exp\left\{-2\left(m-d\right)\left(\varepsilon - \frac{q}{m-d}\right)^2\right\} .$$

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

$$\forall Z_d \subset Z : \left(|Z_d| \neq d\right) \vee \left(R_{\rm emp}\left[f_{L(Z_d)}, Z\right] > \frac{q}{m}\right) \vee \left(R\left[f_{L(Z_d)}\right] \leq \frac{q}{m-d} + \sqrt{\frac{\ln\left(\binom{m}{d}\right) + \ln\left(\frac{1}{\delta}\right)}{2\left(m-d\right)}}\right) \quad (3.4)$$

By a double application of Lemma 1 and 4 we obtain

Theorem 5 (Training error sparsity bound). *Fix a learning algorithm L and $s \in (0,1)$. For any measure P_Z, with probability $1-\delta$ over the random draw of the training set Z of size m for all linear classifier $f_{L(Z)}$, the generalisation error $R\left[f_{L(Z)}\right]$ is bounded from above by*

$$\frac{m}{m-d} R_{\rm emp}\left[f_{L(Z)}, Z\right] + \sqrt{\frac{1}{2}\ln\left(\frac{1}{s}\right) + \frac{d\ln\left(\frac{1}{1-s}\right) + \ln\left(1-s^m\right) + \ln\left(\frac{m}{\delta}\right)}{2\left(m-d\right)}},$$

provided $d \equiv \|L\left(Z\right)\|_0 > 0$.

Proof. Apply Lemma 1 to equation (3.4) using the sequences (3.3) and $p_q = \frac{1}{m}$. □

4 A PAC-Bayesian Analysis

In the PAC-Bayesian framework [9] we aim at incorporating Bayesian priors P_A over the data dependent expansion coefficients $\boldsymbol{\alpha}$ into PAC generalisation error bounds. Our first result bounds the generalisation error of a single classifiers $f_{\boldsymbol{\alpha}}$ whereas Theorem 7 is concerned with the average generalisation error over a subset of version space.

Theorem 6 (PAC–Bayesian folk theorem). *For any two measures P_A and P_Z with probability at least $1-\delta$ over the random draw of the training set Z for all classifiers $f_{\boldsymbol{\alpha}}$ that achieve zero training error $R_{\rm emp}\left[f_{\boldsymbol{\alpha}}, Z\right] = 0$ and $\mathsf{P}_\mathsf{A}\left(\boldsymbol{\alpha}\right) > 0$ the generalisation error is bounded from above by*

$$R\left[f_{\boldsymbol{\alpha}}\right] \leq \frac{1}{m - \|\boldsymbol{\alpha}\|_0}\left(\ln\left(\frac{1}{\mathsf{P}_\mathsf{A}\left(\boldsymbol{\alpha}\right)}\right) + \ln\left(\frac{1}{\delta}\right)\right) .$$

Proof. Apply Lemma 1 to the the statement

$$\Upsilon_{\boldsymbol{\alpha}}\left(Z\right) \equiv \left(R_{\rm emp}\left[f_{\boldsymbol{\alpha}}, Z\right] \neq 0\right) \vee \left(R\left[f_{\boldsymbol{\alpha}}\right] \leq \frac{\ln\left(\frac{1}{\delta}\right)}{m - \|\boldsymbol{\alpha}\|_0}\right) ,$$

that holds with probability at least $1-\delta$ over the random draw of Z for all $\boldsymbol{\alpha}$. For a fixed vector $\boldsymbol{\alpha}$ let $\mathbf{i}_{\boldsymbol{\alpha}}$ e the indices of non zero coefficients α_i and $d = \|\boldsymbol{\alpha}\|_0$. By noticing that $R_{\rm emp}\left[f_{\boldsymbol{\alpha}}, Z\right] = 0$ implies $R_{\rm emp}\left[f_{\boldsymbol{\alpha}}, Z \setminus Z_{\mathbf{i}_{\boldsymbol{\alpha}}}\right] = 0$ we have

$$\begin{aligned}
\forall \boldsymbol{\alpha} : \mathsf{P}_{\mathsf{Z}^m}&\left(\left(R_{\rm emp}\left[f_{\boldsymbol{\alpha}}, \mathsf{Z}\right] = 0\right) \wedge \left(R\left[f_{\boldsymbol{\alpha}}\right] > \varepsilon\right)\right) \\
&\leq \mathsf{P}_{\mathsf{Z}^m}\left(\left(R_{\rm emp}\left[f_{\boldsymbol{\alpha}}, \mathsf{Z} \setminus \mathsf{Z}_{\mathbf{i}_{\boldsymbol{\alpha}}}\right] = 0\right) \wedge \left(R\left[f_{\boldsymbol{\alpha}}\right] > \varepsilon\right)\right) \\
&= \mathsf{P}_{\mathsf{Z}^{m-d}}\left(\left(R_{\rm emp}\left[f_{\boldsymbol{\alpha}}, \mathsf{Z}\right] = 0\right) \wedge \left(R\left[f_{\boldsymbol{\alpha}}\right] > \varepsilon\right)\right) \\
&< \left(1-\varepsilon\right)^{m-d} \leq \exp\left\{-\varepsilon\left(m-d\right)\right\} ,
\end{aligned}$$

because for a fixed index vector $\mathbf{i}_{\boldsymbol{\alpha}}$ the classifier $f_{\boldsymbol{\alpha}}$ does not change over the random draw of the $m-d$ random variables Z_j with indices $j \in \{1,\ldots,m\} \setminus \mathbf{i}_{\boldsymbol{\alpha}}$. The result follows by solving for ε. □

Theorem 7 (PAC-Bayesian subset bound). *For any two measures P_A and P_Z with probability at least $1-\delta$ over the random draw of the training set Z for all subsets $A \in V\left(Z\right)$ of fixed sparsity d, i.e. $\forall \boldsymbol{\alpha} \in A : \|\boldsymbol{\alpha}\|_0 = d$, and $\mathsf{P}_\mathsf{A}\left(A\right) > 0$ the average generalisation error over A is bounded from above by*

$$\mathsf{E}_{\mathsf{A}\in A}\left[R\left[f_\mathsf{A}\right]\right] \leq \frac{\ln\left(\frac{1}{\mathsf{P}_\mathsf{A}(A)}\right) + 2\ln\left(m\right) + \ln\left(\frac{1}{\delta}\right) + 1}{m-d} .$$

As it stands this result bounds the generalisation error of the so-called Gibbs classifier that draws classifiers randomly from A according to the prior measure P_A. It thus justifies this simple Bayesian classification strategy. Furthermore, recent results [6] indicate that Theorem 7 may be useful for providing a PAC-Bayesian bound that benefits from both sparseness and margin of a classifier.

5 From Margin To Sparsity

In this section we present an application of Theorem 3 to the perceptron learning algorithm formulated in feature space $\mathcal{K}$ using kernels k. Given a fixed permutation $\Pi : \{1,\ldots,m\} \to \{1,\ldots,m\}$, the perceptron learning algorithm L^Π is as follows:

1. Start in step zero, i.e. $t=0$, with the vector $\boldsymbol{\alpha}_t = \mathbf{0}$.
2. For all $i \in \{1,\ldots,m\}$, if $y_{\Pi(i)} f_{\boldsymbol{\alpha}}\left(\mathbf{x}_{\Pi(i)}\right) \leq 0$ then
$$\left(\boldsymbol{\alpha}_{t+1}\right)_{\Pi(i)} = \left(\boldsymbol{\alpha}_t\right)_{\Pi(i)} + 1 . \quad (5.1)$$
and $t \leftarrow t+1$.
3. Stop, if there is no $i \in \{1,\ldots,m\}$ such that
$$y_{\Pi(i)} f_{\boldsymbol{\alpha}}\left(\mathbf{x}_{\Pi(i)}\right) \leq 0 .$$

In the early 60's Novikoff and Aizerman et al. [12, 1] were able to give an upper bound on the number t of mistakes made by this learning procedure. Given

a training set Z, the quantity determining the upper bound is the maximally achievable margin $\max_{\boldsymbol{\alpha}} \gamma_Z(\boldsymbol{\alpha})$ on the training sample $Z = (X, Y)$ normalised by the total extent ς of the data in feature space, i.e. $\varsigma = \max_{\mathbf{x}_i \in X} \|\boldsymbol{\phi}(\mathbf{x}_i)\|_{\mathcal{K}}$. This *margin* $\gamma_Z(\boldsymbol{\alpha})$ is given by

$$\begin{aligned} \gamma_Z(\boldsymbol{\alpha}) &= \frac{1}{\|f_{\boldsymbol{\alpha}}\|} \min_{(\mathbf{x}_i, y_i) \in Z} y_i f_{\boldsymbol{\alpha}}(\mathbf{x}_i) , \\ \|f_{\boldsymbol{\alpha}}\|^2 &= \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) . \end{aligned}$$

Theorem 8 (Mistake Bound for Perceptrons). *Fix a permutation $\Pi : \{1, \ldots, m\} \to \{1, \ldots, m\}$. Let $Z = (X, Y)$ be a training set of size m and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel. Suppose that there exists a vector $\boldsymbol{\alpha}^* \in \mathbb{R}^m$ such that $\gamma_Z(\boldsymbol{\alpha}^*) > 0$. Then the number of mistakes made by the perceptron learning algorithm L^{Π} on Z is at most*

$$\left(\frac{\varsigma}{\gamma_Z(\boldsymbol{\alpha}^*)} \right)^2 .$$

Considering the form of the update rule (5.1) we observe that this result not only bounds the number of mistakes made during learning but also the number $\|\boldsymbol{\alpha}\|_0$ of non-zero coefficients in the $\boldsymbol{\alpha}$ vector. To be precise, it bounds the ℓ_1 norm $\|\boldsymbol{\alpha}\|_1$ of the coefficient vector $\boldsymbol{\alpha}$ which, in turn, bounds the zero norm $\|\boldsymbol{\alpha}\|_0$ from above for all vectors with integer components. Theorem 8 thus establishes a relation between the *existence of a large margin classifier $\boldsymbol{\alpha}^*$ and the sparseness of any solution found by the perceptron algorithm.* Combining Theorem 8 and Lemma 2 and 1 with $p_d = \frac{1}{m}$ thus gives the following remarkable result.

Theorem 9 (Margin Bound). *Fix a permutation $\Pi : \{1, \ldots, m\} \to \{1, \ldots, m\}$. For any measure P_{Z} such that $\mathsf{P}_{\mathsf{X}}(\|\boldsymbol{\phi}(\mathsf{X})\|_{\mathcal{K}} \leq \varsigma) = 1$, with probability at least $1 - \delta$ over the random draw of the training set Z of size m, if there exists a vector $\boldsymbol{\alpha}^*$ such that $\gamma_Z(\boldsymbol{\alpha}^*) > \frac{\varsigma}{\sqrt{m}}$ then the generalisation error $R\left[f_{L^{\Pi}(Z)}\right]$ of the classifier $f_{L^{\Pi}(Z)}$ found by the perceptron learning algorithm L^{Π} is less than*

$$\frac{1}{m-d} \left(\ln \left(\binom{m}{d} \right) + \ln(m) + \ln \left(\frac{1}{\delta} \right) \right) , \qquad (5.2)$$

where $d = \left\lceil \varsigma^2 \gamma_Z^{-2}(\boldsymbol{\alpha}^*) \right\rceil$.

The most intriguing feature of this result is that the *mere existence* of a large margin classifier $f_{\boldsymbol{\alpha}^*}$ is sufficient to guarantee a small generalisation error for the solution $f_{L^{\Pi}(Z)}$ of the perceptron learning algorithm although its attained margin $\gamma_Z(\boldsymbol{\alpha})$ is likely to be much smaller than $\gamma_Z(\boldsymbol{\alpha}^*)$. It has long been argued that the attained margin $\gamma_Z(\boldsymbol{\alpha})$ *itself* is the crucial quantity controlling the generalisation error of $\boldsymbol{\alpha}$. In light of our new result *if there exists a consistent classifier $f_{\boldsymbol{\alpha}^*}$ with large margin we know that there also exists at least one classifier $f_{\boldsymbol{\alpha}}$ with high sparsity* that can efficiently be found using the perceptron learning algorithm. In fact, whenever the SVM appears to be theoretically justified by a large observed margin, every solution found by the perceptron algorithm has a small guaranteed generalisation error — mostly even smaller than current bounds on the generalisation error of SVMs. Note that for a given training sample Z it is not unlikely that by permutation of Z via Π there exist exponentially many different consistent sparse classifiers $f_{\boldsymbol{\alpha}}$.

6 Conclusion

In this paper we proved a series of bounds for linear classifiers exploiting sparsity and prior knowledge thereabout. Double sample arguments could be avoided due to the sparseness that leaves iid samples from the training sample for witnessing the quality of the classifier. Thus we established proof for the common conception that sparse classifiers lead to good generalisation. Future work will be concerned with a thorough exploration of PAC-Bayesian results with priors on the data-dependent expansion coefficients in dual representations of linear classifiers.

Acknowledgements

We would like to thank Bob Williamson, Jon Baxter, Peter Bartlett, Alex Smola and Bernhard Schölkopf for many inspiring discussions. Parts of this work were done during a research stay of RH and TG at the Australian National University in Canberra.

7 Proofs

For subsets $Z' \subseteq Z$ of size exactly $d \in \{1, \ldots, m\}$ we denote by $\mathbf{i}$ the index vector $\mathbf{i} = (i_1, \ldots, i_d) \in \{1, \ldots, m\}^d$ of d *distinct* indices $i_1 < \ldots < i_d$ from the set $\{1, \ldots, m\}$. We use I_d to denote the set of all subsets $\mathbf{i}$ of $\{1, \ldots, m\}$ of size d. Given a training set Z of size m we write $Z_{\mathbf{i}} = \{(\mathbf{x}_{i_1}, y_{i_1}), \ldots, (\mathbf{x}_{i_d}, y_{i_d})\} \subseteq Z$. Finally, we use

$$\forall \varepsilon \in [0,1] : (1 - \varepsilon) \leq \exp\{-\varepsilon\} . \qquad (7.1)$$

7.1 Proof of Lemma 1

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

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

□

7.2 Proof of Lemma 2

Proof. We exploit the idea that for a fixed value of d there are still $m - d$ points drawn iid according to P_{Z} on which the classifier $f_{L(Z_d)}$ has succeeded. Thus, for a fixed set $Z_d \in \mathcal{Z}^d$ let us define the event

$$\begin{aligned} \Upsilon(Z_d, Z_{m-d}) \equiv & \left(R_{\text{emp}}\left[f_{L(Z_d)}, Z_{m-d}\right] = 0 \right) \wedge \\ & \left(R\left[f_{L(Z_d)}\right] > \varepsilon \right) , \end{aligned}$$

where $Z_{m-d} \in \mathcal{Z}^{m-d}$. By the binomial tail bound we know that

$$\forall Z_d \in \mathcal{Z}^d : \mathbf{P}_{\mathsf{Z}^{m-d}} \left(\Upsilon \left(Z_d, \mathsf{Z}^{m-d} \right) \right) < (1-\varepsilon)^{m-d} . \quad (7.2)$$

Thus we conclude

$$\begin{aligned} \mathbf{P}_{\mathsf{Z}^m} \left(\exists Z_d \subseteq \mathsf{Z} : \left(R_{\mathrm{emp}} \left[f_{L(Z_d)}, \mathsf{Z} \right] = 0 \right) \wedge \left(R \left[f_{L(Z_d)} \right] > \varepsilon \right) \right) \\ = \mathbf{P}_{\mathsf{Z}^m} \left(\exists \mathbf{i} : \Upsilon \left(\mathsf{Z}_\mathbf{i}, \mathsf{Z} \setminus \mathsf{Z}_\mathbf{i} \right) \right) \\ \leq \sum_\mathbf{i} \mathbf{P}_{\mathsf{Z}^m} \left(\Upsilon \left(\mathsf{Z}_\mathbf{i}, \mathsf{Z} \setminus \mathsf{Z}_\mathbf{i} \right) \right) , \end{aligned}$$

where the last inequality follows from the union bound. Since the number of summands is $\binom{m}{d}$ and the summands are bounded by equation (7.2) we finally have that the probability under consideration is at most

$$\binom{m}{d} (1-\varepsilon)^{m-d} .$$

□

7.3 Proof of Lemma 4

Proof. For a fixed set $Z_d \in \mathcal{Z}^d$ let us define the

$$\begin{aligned} \Upsilon \left(Z_d, Z_{m-d} \right) \quad \equiv \quad & \left(R_{\mathrm{emp}} \left[f_{L(Z_d)}, Z_{m-d} \right] \leq \frac{q}{m-d} \right) \wedge \\ & \left(R \left[f_{L(Z_d)} \right] > \varepsilon \right) , \end{aligned}$$

where $Z_{m-d} \in \mathcal{Z}^{m-d}$. We have chosen $\frac{q}{m-d}$ because if $f_{L(Z_d)}$ commits not more than q errors on Z, the *number* of errors on the subset Z_{m-d} is upper bounded by q. By Hoeffding's inequality [7] we know that

$$\begin{aligned} & \forall Z_d \in \mathcal{Z}^d : \mathbf{P}_{\mathsf{Z}^{m-d}} \left(\Upsilon \left(Z_d, \mathsf{Z}_{m-d} \right) \right) \\ & < \exp \left\{ -2 (m-d) \left(\varepsilon - \frac{q}{m-d} \right)^2 \right\} . \quad (7.3) \end{aligned}$$

Thus we conclude

$$\begin{aligned} \mathbf{P}_{\mathsf{Z}^m} \left(\exists Z_d \subseteq \mathsf{Z} : \left(R_{\mathrm{emp}} \left[f_{L(Z_d)}, \mathsf{Z} \right] \leq \frac{q}{m} \right) \wedge \left(R \left[f_{L(Z_d)} \right] > \varepsilon \right) \right) \\ = \mathbf{P}_{\mathsf{Z}^m} \left(\exists \mathbf{i} : \Upsilon \left(\mathsf{Z}_\mathbf{i}, \mathsf{Z} \setminus \mathsf{Z}_\mathbf{i} \right) \right) \\ \leq \sum_\mathbf{i} \mathbf{P}_{\mathsf{Z}^m} \left(\Upsilon \left(\mathsf{Z}_\mathbf{i}, \mathsf{Z} \setminus \mathsf{Z}_\mathbf{i} \right) \right) , \end{aligned}$$

where the last inequality follows from the union bound. Since the number of summands is $\binom{m}{d}$ and the summands are bounded by equation (7.3) we finally have that the probability under consideration is at most

$$\binom{m}{d} \exp \left\{ -2 (m-d) \left(\varepsilon - \frac{q}{m-d} \right)^2 \right\} .$$

□

7.4 Proof of Theorem 7

Let us recall the *quantifier reversal lemma* [9].

Lemma 10 (Quantifier Reversal Lemma). *Let* X *and* Y *be random variables and let* δ *range over* $(0,1]$. *Let* $\Upsilon : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \mapsto \{\mathrm{true}, \mathrm{false}\}$ *be any measurable formula on the product space* $\mathcal{X} \times \mathcal{Y}$ *such that for any* $x \in \mathcal{X}$ *and* $y \in \mathcal{Y}$ *we have*

$$\{ \delta \in (0,1] : \Upsilon (x, y, \delta) \} = (0, \delta_{\max}]$$

for some $\delta_{\max}$. *If*

$$\forall x \forall \delta \in (0,1] : \quad \mathbf{P}_{\mathsf{Y}|\mathsf{X}=x} \left(\Upsilon \left(x, \mathsf{Y}, \delta \right) \right) \geq 1 - \delta$$

then we have for all $\beta \in (0,1)$ *and* $\delta \in (0,1]$

$$\mathbf{P}_{\mathsf{Y}} \left(\forall \gamma \in (0,1] : \mathbf{P}_{\mathsf{X}|\mathsf{Y}=y} \left(\Upsilon \left(\mathsf{X}, y, (\gamma \beta \delta)^{\frac{1}{1-\beta}} \right) \right) \geq 1 - \gamma \right) \geq 1 - \delta .$$

Proof of Theorem 7. We decompose the expectation at some point $\varepsilon \in \mathbb{R}$ by

$$\mathbf{E}_{\mathsf{A} \in A} \left[R \left[f_\mathsf{A} \right] \right] \leq \varepsilon \cdot \mathbf{P}_{\mathsf{A} \in A} \left(R \left[f_\mathsf{A} \right] \leq \varepsilon \right) + 1 \cdot \mathbf{P}_{\mathsf{A} \in A} \left(R \left[f_\mathsf{A} \right] > \varepsilon \right) ,$$

using $R \left[f_\mathsf{A} \right] < 1$ by definition. In the proof of Theorem 6 we have already shown that for all $\boldsymbol{\alpha} \in A$ and for all $\delta \in (0,1]$

$$\mathbf{P}_{\mathsf{Z}^m | \mathsf{A} = \boldsymbol{\alpha}} \left(\left(\boldsymbol{\alpha} \notin V \left(\mathsf{Z} \right) \right) \vee \left(R \left[f_{\boldsymbol{\alpha}} \right] \leq \frac{\ln \left(\frac{1}{\delta} \right)}{m - \left\| \boldsymbol{\alpha} \right\|_0} \right) \right) \geq 1 - \delta .$$

By Lemma 10 this implies that for all $\beta \in (0,1)$ with probability at least $1 - \delta$ over the random draw of the training set Z for all $\gamma \in (0,1]$

$$\mathbf{P}_{\mathsf{A} | \mathsf{Z}^m = Z} \left(\left(\mathsf{A} \in V \left(Z \right) \right) \wedge \left(R \left[f_\mathsf{A} \right] > \underbrace{\frac{1}{1-\beta} \frac{\ln \left(\frac{1}{\delta \gamma \beta} \right)}{m - \left\| \mathsf{A} \right\|_0}}_{\varepsilon(\gamma, \beta)} \right) \right) < \gamma .$$

By assumption of the theorem $\mathbf{P}_{\mathsf{A} | \mathsf{Z}^m = Z} = \mathbf{P}_\mathsf{A}$ from which it follows that

$$\begin{aligned} \mathbf{P}_{\mathsf{A} \in A} \left(R \left[f_\mathsf{A} \right] > \varepsilon \left(\gamma, \beta \right) \right) \quad &= \quad \frac{\mathbf{P}_\mathsf{A} \left(\left(\mathsf{A} \in A \right) \wedge \left(R \left[f_\mathsf{A} \right] > \varepsilon \left(\gamma, \beta \right) \right) \right)}{\mathbf{P}_\mathsf{A} \left(A \right)} \\ &\leq \quad \frac{\gamma}{\mathbf{P}_\mathsf{A} \left(A \right)} , \end{aligned}$$

because $A \in V \left(Z \right)$. If we set $\gamma = \frac{\mathbf{P}_\mathsf{A}(A)}{m}$ and $\beta = \frac{1}{m}$ we finally obtain that with probability at least $1 - \delta$ over the random draw of the training set Z

$$\mathbf{E}_{\mathsf{A} \in A} \left[R \left[f_\mathsf{A} \right] \right] \leq \frac{\ln \left(\frac{1}{\mathbf{P}_\mathsf{A}(A)} \right) + 2 \ln \left(m \right) + \ln \left(\frac{1}{\delta} \right)}{m - d} + \frac{1}{m} ,$$

because by assumption for all $\boldsymbol{\alpha} \in A : \left\| \boldsymbol{\alpha} \right\|_0 = d$. Bounding $\frac{1}{m}$ by $\frac{1}{m-d}$ from above completes the proof. □

References

[1] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

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

[3] E. Fix and J. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, U.S., 1951.

[4] E. Fix and J. Hodges. Discriminatory analysis: small sample performance. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, TX, U.S., 1951.

[5] S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, (21):269–304, 1995.

[6] R. Herbrich, T. Graepel, and C. Campbell. Bayesian learning in reproducing kernel Hilbert spaces. Technical report, Technical University Berlin, 1999. TR 99-11.

[7] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

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

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

[10] 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.

[11] A. Y. Ng. On feature selection: Learning with exponentially many irrelevant features as training examples. In *Proceedings of the 15th International Conference on Machine Learning*, pages 404–412. Morgan Kaufmann, 1998.

[12] A. Novikoff. On convergence proofs for perceptrons. In *Report at the Symposium on Mathematical Theory of Automata*, pages 24–26, Politechnical Institute Brooklyn, 1962.

[13] M. E. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems*, San Mateo, CA, 2000. Morgan Kaufmann. in press.

[14] V. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, New York, 1998.