A PAC-Bayesian Margin Bound for Linear Classifiers

Ralf Herbrich and Thore Graepel

Abstract—**We present a bound on the generalisation error of linear classifiers in terms of a refined margin quantity on the training sample. The result is obtained in a PAC-Bayesian framework and is based on geometrical arguments in the space of linear classifiers. The new bound constitutes an exponential improvement of the so far tightest margin bound, which was developed in the luckiness framework, and scales logarithmically in the inverse margin. Even in the case of less training examples than input dimensions sufficiently large margins lead to non-trivial bound values and—for maximum margins—to a vanishing complexity term. In contrast to previous results, however, the new bound does depend on the dimensionality of feature space. The analysis shows that the classical margin is too coarse a measure for the essential quantity that controls the generalisation error: the fraction of hypothesis space consistent with the training sample. The practical relevance of the result lies in the fact that the well-known support vector machine is optimal with respect to the new bound only if the feature vectors in the training sample are all of the same length. As a consequence we recommend to use SVMs on normalised feature vectors only. Numerical simulations support this recommendation and demonstrate that the new error bound can be used for the purpose of model selection.**

Index Terms—**Bayes Classification Strategy, Computational Learning Theory, Generalisation Error Bound, Gibbs Classification Strategy, Linear Classifiers, Margin, Model Selection, PAC-Bayesian Framework, Support Vector Machine, Volume Ratios**

I. Introduction

Linear classifiers are popular in the machine learning and statistics communities due to their straightforward applicability and high flexibility that has been greatly improved by the so-called kernel method [1]. A natural and popular framework for the theoretical analysis of classifiers is the PAC (*probably approximately correct*) framework [2] which is closely related to Vapnik's ([3]) work on the generalisation error. For binary classifiers it turned out that the *growth function* is an appropriate measure of "complexity" and can tightly be upper bounded by the VC (Vapnik-Chervonenkis) dimension [4]. *Structural risk minimisation* [3] was suggested for directly minimising the VC dimension based on a training sample and an *a priori* structuring of the hypothesis space.

In practice, for example in the case of linear classifiers, often a thresholded *real-valued* function is used for classification. In 1993, Kearns and Schapire [5] demonstrated that considerably tighter bounds can be obtained by considering a scale-sensitive complexity measure known as the *fat shattering dimension*. Further results [6] provided bounds on the growth function similar to those proved by Vapnik and others [4], [7]. The popularity of the theory greatly increased by the invention of the *support vector machine* (SVM) [1] which aims at directly minimising the complexity as suggested by theory.

Until recently, however, the success of the SVM remained somewhat obscure because in PAC/VC theory the structuring of the hypothesis space must be *independent* of the training sample—in contrast to the data-dependence of the canonical hyper-plane. As a consequence Shawe-Taylor et al. [8] developed the *luckiness framework*, where luckiness refers to a complexity measure that is a function of both hypothesis *and* training sample.

First bounds on the generalisation error in a PAC-Bayesian spirit were obtained by Shawe-Taylor et al. [9] for single hypotheses. Recently, David McAllester presented some PAC-Bayesian theorems [10] that bound the generalisation error of randomised Bayesian classifiers *independently* of the correctness of the prior and regardless of the underlying data distribution—thus fulfilling the basic desiderata of PAC theory.

In this paper we extend McAllester's error bounds for the Gibbs classification strategy[1] (that draws classifiers randomly from the posterior distribution) to the Bayes (optimal) classification strategy (that weights the classification of each classifier by its posterior weight) and eventually to arbitrary consistent classifiers. Note, that the PAC-Bayesian framework provides *a posteriori* error bounds and is thus closely related in spirit to the luckiness framework[2].

The main contribution of this paper is a tight margin bound for linear classifiers in the PAC-Bayesian framework (see also [12]). The central idea is to identify the generalisation error of the classifier h of interest with that of the Bayes (optimal) classification strategy on a (point-symmetric) subset Q of hypothesis space that is *summarised* by h . For linear classifiers we show that for a uniform prior P_{W} over normalised weight vectors $\mathbf{w}$ the *normalised margin* $\Gamma_{\mathbf{z}}(\mathbf{w})$ of $h_{\mathbf{w}}$ is *directly* related to the volume of a large subset Q of hypothesis space summarised by $h_{\mathbf{w}}$. In particular, the result suggests that a learning algorithm for linear classifiers

Ralf Herbrich is with Microsoft Research Cambridge, 7 J J Thomson Avenue, Cambridge CB3 0FB, United Kingdom. Email: rherb@microsoft.com. Thore Graepel works at the Institute of Computational Science, Weinbergstrasse 43, WET, 2nd level, ETH Zentrum, 8006 Zurich, Switzerland. Email: graepel@inf.ethz.ch. This work was done while both authors were at Technical University of Berlin, Statistics and Business Mathematics, Sekr. FR 6-9, Franklinstr. 28/29, 10587 Berlin, Germany.

[1] The notion of the Gibbs classification strategy as used here should not be confused with the notion of Gibbs estimators as defined in [11].

[2] In fact, even Shawe-Taylor et al. concede that "... a Bayesian might say that luckiness is just a complicated way of encoding a prior. The sole justification for our particular way of encoding is that it allows us to get the PAC like results we sought..." [9, p. 4].

should aim at maximising the normalised margin instead of the classical margin. In Sections II and III we review the basic PAC-Bayesian theorem and show how it can be applied to single classifiers. In Section IV we give our main result and outline its proof. The necessary lemmata together with their proofs have been relegated to the appendix. In Section V we present an experimental study for 2 -dimensional data suggesting that the new bound can be successfully used for model selection in low dimensional feature spaces. Also, we discuss the consequences of the new result for the application of SVMs and demonstrate experimentally that, in fact, a normalisation of the feature vectors—as suggested by the new error bound—leads to considerably superior generalisation performance.

We denote m -tuples by italic bold letters (for example $\boldsymbol{x} = (x_1, \ldots, x_m)$), vectors (m -tuples of real numbers) by roman bold letters (for example $\mathbf{x}$), random variables by sans serif font (for example X), (vector) spaces by calligraphic capitalised letters (for example $\mathcal{X}$) and subsets of these vector spaces by capitalised roman letters (for example W). The symbols $\mathbf{P}, \mathbf{E}, \mathbf{I}$ and ℓ_2^n denote a probability measure, the expectation of a random variable, the indicator function and the normed space (2 -norm) of sequences of length n , respectively. The shorthand notation $z \in \boldsymbol{z}$ is formally defined by $z \in \boldsymbol{z} \Leftrightarrow \exists i \in \{1, \ldots, |\boldsymbol{z}|\} : z_i = z$.

II. A PAC Margin Bound

We consider learning in the PAC framework. Let $\mathcal{X}$ be the input space, and let $\mathcal{Y} := \{-1, +1\}$. Let a labelled training sample $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \mathcal{Y})^m =: \mathcal{Z}^m$ be drawn i.i.d. according to some unknown probability measure $\mathbf{P}_{\mathsf{Z}} = \mathbf{P}_{\mathsf{Y}|\mathsf{X}} \mathbf{P}_{\mathsf{X}}$. In the following we shall only study linear classifiers

$$\begin{aligned} \mathcal{H} &:= \{x \mapsto \operatorname{sign}(\langle \mathbf{w}, \boldsymbol{\phi}(x) \rangle) \mid \mathbf{w} \in \mathcal{W}\} , && (1) \\ \mathcal{W} &:= \{\mathbf{w} \in \mathcal{K} \mid \|\mathbf{w}\| = 1\} , && (2) \end{aligned}$$

where the mapping $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{K} = \ell_2^n$ maps[3] the input data to some feature space $\mathcal{K}$ and $\|\mathbf{w}\| = 1$ leads to a one-to-one correspondence of hypotheses[4] $h_{\mathbf{w}} \in \mathcal{H}$ to their parameters $\mathbf{w} \in \mathcal{W}$. Note that the unit length constraint does not restrict the classifiers considered because the classification $\operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$ of a point x is independent of the norm $\|\mathbf{w}\|$ of $\mathbf{w}$. Furthermore we assume the existence of a "true" hypothesis $\mathbf{w}^* \in \mathcal{W}$ that labelled the data[5], which leads to what we may refer to as a PAC-likelihood,

$$\mathbf{P}_{\mathsf{Y}|\mathsf{X}=x}(y) := \mathbf{I}_{y=\operatorname{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle)} . \qquad (3)$$

From the existence of $\mathbf{w}^*$ we know that there exists a version space $V(\boldsymbol{z}) \subseteq \mathcal{W}$,

$$V(\boldsymbol{z}) := \{\mathbf{w} \in \mathcal{W} \mid \forall (x, y) \in \boldsymbol{z} \ : \ \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) = y\} .$$

Our analysis aims at bounding the true risk $R[\mathbf{w}]$ of consistent hypotheses $\mathbf{w} \in V(\boldsymbol{z})$,

$$R[\mathbf{w}] := \mathbf{E}_{\mathsf{XY}}\left[\mathbf{I}_{\operatorname{sign}(\langle \mathbf{w}, \boldsymbol{\phi}(\mathsf{X}) \rangle) \neq \mathsf{Y}}\right] . \qquad (4)$$

Since all classifiers $\mathbf{w} \in V(\boldsymbol{z})$ are indistinguishable in terms of number of errors committed on the given training sample $\boldsymbol{z}$ we introduce the concept of the *margin* $\gamma_{\boldsymbol{z}}(\mathbf{w})$ of a classifier $\mathbf{w}$,

$$\gamma_{\boldsymbol{z}}(\mathbf{w}) := \min_{(x_i, y_i) \in \boldsymbol{z}} y_i \langle \mathbf{w}, \mathbf{x}_i \rangle . \qquad (5)$$

The following theorem due to Shawe-Taylor et al.[6] [8] bounds the generalisation errors $R[\mathbf{w}]$ of all classifiers $\mathbf{w} \in V(\boldsymbol{z})$ in terms of their margins $\gamma_{\boldsymbol{z}}(\mathbf{w})$.

Theorem 1. *For all probability measures* $\mathbf{P}_{\mathsf{Z}}$ *such that* $\mathbf{P}_{\mathsf{X}}(\|\boldsymbol{\phi}(\mathsf{X})\| \leq \varsigma) = 1$ *, for any* $\delta \in (0, 1]$ *, with probability at least* $1 - \delta$ *over the random draw of the training sample* $\boldsymbol{z} \in \mathcal{Z}^m$ *, for any consistent classifier* $\mathbf{w} \in V(\boldsymbol{z})$ *with a positive margin* $\gamma_{\boldsymbol{z}}(\mathbf{w}) > \sqrt{32\varsigma^2/m}$ *the generalisation error* $R[\mathbf{w}]$ *is bounded from above by*

$$\frac{2}{m}\left(\kappa(\mathbf{w}) \log_2\left(\frac{8em}{\kappa(\mathbf{w})}\right) \log_2(32m) + \log_2\left(\frac{2m}{\delta}\right)\right) , \quad (6)$$

$$\kappa(\mathbf{w}) := \left\lceil \left(\frac{8\varsigma}{\gamma_{\boldsymbol{z}}(\mathbf{w})}\right)^2 \right\rceil .$$

As the bound on $R[\mathbf{w}]$ depends on $\gamma_{\boldsymbol{z}}^{-2}(\mathbf{w})$ we see that Theorem 1 provides a theoretical foundation of all algorithms that aim at maximising $\gamma_{\boldsymbol{z}}(\mathbf{w})$, for example, SVMs and Boosting [1], [14]. Nevertheless, its actual value is too large for any practically relevant training sample size m . For example, in order for (6) to be less than one (which is still trivially true) we need the astronomically large training sample size of $m_{\min} = 34\,816$ even in the luckiest case of $\gamma_{\boldsymbol{z}}(\mathbf{w}) = \varsigma$ and $\delta = 1$.

III. PAC-Bayesian Analysis

The PAC-Bayesian analysis requires the definition of a prior over hypothesis space. In the case of linear classifiers we assume a prior measure $\mathbf{P}_{\mathsf{W}}$ over normalised weight space $\mathcal{W}$ and assume independence of W and Z . We first present a result [10] that bounds the risk of the Gibbs classification strategy $\mathit{Gibbs}_{W(\boldsymbol{z})}$ by the measure $\mathbf{P}_{\mathsf{W}}(W(\boldsymbol{z}))$ of a consistent subset $W(\boldsymbol{z}) \subseteq V(\boldsymbol{z})$. This average risk is then related via the Bayes-Gibbs lemma to the risk of the Bayes classification strategy $\mathit{Bayes}_{W(\boldsymbol{z})}$ on $W(\boldsymbol{z})$. For a single consistent hypothesis $\mathbf{w} \in V(\boldsymbol{z})$ it is finally necessary to identify a consistent subset $Q(\mathbf{w}) \subseteq V(\boldsymbol{z})$ such that the Bayes classification strategy $\mathit{Bayes}_{Q(\mathbf{w})}$ on $Q(\mathbf{w})$ always agrees with $\mathbf{w}$.

Let us define the Gibbs classification strategy Gibbs_W with respect to the subset $W \subseteq \mathcal{W}$ by

$$\mathit{Gibbs}_W(x) := \operatorname{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) , \qquad \mathbf{w} \sim \mathbf{P}_{\mathsf{W}|\mathsf{W} \in W} , \qquad (7)$$

[3] For notational simplicity we will sometimes abbreviate $\boldsymbol{\phi}(x)$ by $\mathbf{x} \in \mathcal{K}$ which should not be confused with the sequence $\boldsymbol{x} \in \mathcal{X}^m$ of training examples.

[4] In the following, we synonymously refer to elements of $\mathcal{H}$ and $\mathcal{W}$ as hypotheses or classifiers always bearing in mind their relation given in (1) and (2).

[5] In fact, for the application of our main result, Theorem 6, the existence of $\mathbf{w}^*$ is not strictly necessary, but the existence of a version space $V(\boldsymbol{z})$ is sufficient.

[6] The present version of this theorem is taken from [13].

where for any two measurable subsets $W_1, W_2 \in \mathcal{W}$ we define $\mathsf{P}_{\mathsf{W}|\mathsf{W}\in W_2}(W_1) := \mathsf{P}_{\mathsf{W}}(W_1 \cap W_2) / \mathsf{P}_{\mathsf{W}}(W_2)$. Given a new test point $x \in \mathcal{X}$, the Gibbs classification strategy Gibbs_W draws a classifier $\mathbf{w} \in W$ with probability proportional to the prior and uses $\mathbf{w}$ for classification. As a consequence, this randomised decision strategy has a generalisation error given by

$$R[\mathit{Gibbs}_W] := \mathsf{E}_{\mathsf{XY}}\left[\mathsf{P}_{\mathsf{W}|\mathsf{W}\in W}\left(\mathrm{sign}\left(\langle \mathsf{W}, \boldsymbol{\phi}(\mathsf{X})\rangle\right) \neq \mathsf{Y}\right)\right]. \tag{8}$$

The following theorem [10] provides an upper bound on the generalisation error of the Gibbs classification strategy.

Theorem 2. *For any two independent measures P_{W} and P_{Z} , for any $\delta \in (0,1]$ with probability at least $1-\delta$ over the random draw of the training sample $\boldsymbol{z} \in \mathcal{Z}^m$ for all subsets $W(\boldsymbol{z}) \subseteq V(\boldsymbol{z})$ such that $\mathsf{P}_{\mathsf{W}}(W(\boldsymbol{z})) > 0$ the generalisation error $R\left[\mathrm{Gibbs}_{W(\boldsymbol{z})}\right]$ of the associated Gibbs classification strategy $\mathrm{Gibbs}_{W(\boldsymbol{z})}$ is bounded from above by*

$$\frac{1}{m}\left(\ln\left(\frac{1}{\mathsf{P}_{\mathsf{W}}(W(\boldsymbol{z}))}\right) + \ln\left(\frac{m^2}{\delta}\right) + 1\right). \tag{9}$$

Now consider the Bayes classification strategy Bayes_W which deterministically assigns a new test point $x \in \mathcal{X}$ to the class which achieves the highest vote from classifiers in W ,

$$\mathit{Bayes}_W(x) := \underset{y\in\mathcal{Y}}{\mathrm{argmax}}\ \mathsf{P}_{\mathsf{W}|\mathsf{W}\in W}\left(\mathrm{sign}\left(\langle \mathsf{W}, \mathbf{x}\rangle\right) = y\right). \tag{10}$$

There exists a simple relationship between the generalisation error of the Bayes classification strategy and the Gibbs classification strategy.

Lemma 3. *For any two independent measures P_{W} and P_{XY} and any set $W \subseteq \mathcal{W}$*

$$R[\mathrm{Bayes}_W] \leq 2 \cdot R[\mathrm{Gibbs}_W]. \tag{11}$$

Proof: It suffices to show that for all $(x,y) \in \mathcal{Z}$

$$\mathit{Bayes}_W(x) \neq y \ \Rightarrow\ \mathsf{P}_{\mathsf{W}|\mathsf{W}\in W}\left(\mathrm{sign}\left(\langle \mathsf{W}, \mathbf{x}\rangle\right) \neq y\right) \geq \frac{1}{2},$$

because combining (4) and (8) proves the lemma. However, since $|\mathcal{Y}| = 2$, by definition (11) the above statement always holds. Thus the lemma is proven. ∎

The combination of Lemma 3 with Theorem 2 yields a bound on the risk of the deterministic classification strategy $\mathit{Bayes}_{W(\boldsymbol{z})}$. In order to apply this bound to *single* classifiers in version space, we introduce the concept of *Bayes-admissibility*, i.e. for a single hypothesis $\mathbf{w} \in V(\boldsymbol{z})$ let us find a subset $Q(\mathbf{w})$ of version space $V(\boldsymbol{z})$ such that $\mathit{Bayes}_{Q(\mathbf{w})}$ on $Q(\mathbf{w})$ agrees with $\mathbf{w}$ on every point in $\mathcal{X}$.

Definition 4. Given the hypothesis space in (1) and a prior measure P_{W} over $\mathcal{W}$ we call a subset $Q(\mathbf{w}) \subseteq \mathcal{W}$ *Bayes-admissible* with respect to $\mathbf{w}$ *and* P_{W} if and only if

$$\forall x \in \mathcal{X}: \qquad \mathrm{sign}\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) = \mathit{Bayes}_{Q(\mathbf{w})}(x).$$

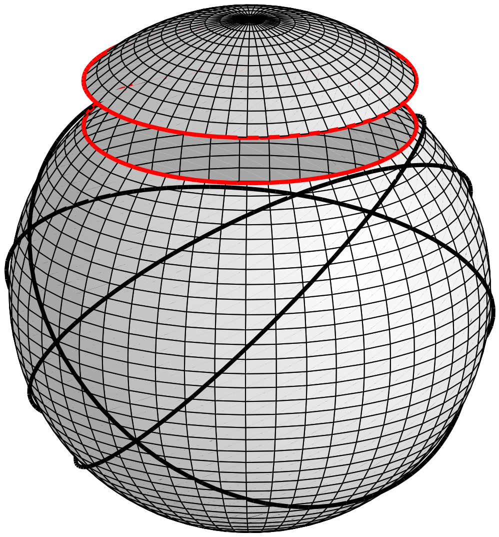


Figure 1. Illustration of the volume ratio for the classifier at the north pole. Four training points shown as grand circles make up version space—the polyhedron on top of the sphere. The radius of the “cap” of the sphere is proportional to the margin $\Gamma_{\boldsymbol{z}}$, which *only* for $\|\mathbf{x}_i\| = \mathrm{const.}$ is maximised by the SVM.

Although it may be difficult in general to find Bayes-admissible sets for a given classifier $\mathbf{w}$ under arbitrary priors P_{W} , the following geometrically plausible lemma establishes Bayes-admissibility for the case of interest.

Lemma 5. *For the uniform measure P_{W} over $\mathcal{W}$ each ball $Q(\mathbf{w}) = \{\mathbf{v} \in \mathcal{W} \mid \|\mathbf{w} - \mathbf{v}\| \leq \gamma\}$ is Bayes-admissible with respect to its centre $\mathbf{w} \in \mathcal{W}$.*

Proof: For any data point $x \in \mathcal{X}$ any subset $W \in \mathcal{W}$ can be written as the union $W = W_{+1} \cup W_{-1}$ of two disjunct sets W_{+1} and W_{-1} such that for $y \in \mathcal{Y}$ we have $\mathrm{sign}\left(\langle \mathbf{w}, \mathbf{x}\rangle\right) = y$ for all $\mathbf{w} \in W_y$. Then the Bayes classification strategy (10) can be written as

$$\mathit{Bayes}_W(x) = \underset{y\in\mathcal{Y}}{\mathrm{argmax}}\ \mathrm{vol}(W_{\mathrm{y}}).$$

If W is chosen to be the ball $Q(\mathbf{w})$, then we have by point-symmetry of $Q(\mathbf{w})$ that the weight vector $\mathbf{w}$ always lies in the half of greater volume and hence agrees with $\mathit{Bayes}_{Q(\mathbf{w})}$. Note that the frontier between W_{+1} and W_{-1} is a geodesic of $\mathcal{W}$. ∎

Please note that by considering a ball $Q(\mathbf{w})$ rather than just $\mathbf{w}$ we make use of the fact that $\mathbf{w}$ *summarises* all its neighbouring classifiers $\mathbf{v} \in Q(\mathbf{w})$. In order to minimise the bound given in Theorem 2 for any classifier $\mathbf{w} \in V(\boldsymbol{z})$ we would like to use the largest ball $Q^*(\mathbf{w})$ such that $Q^*(\mathbf{w}) \subset V(\boldsymbol{z})$. It turns out that for a uniform prior P_{W} the volume of $Q^*(\mathbf{w})$ can be expressed in terms of the *normalised* margin

$$\Gamma_{\boldsymbol{z}}(\mathbf{w}) := \min_{(x_i,y_i)\in\boldsymbol{z}} \frac{y_i \langle \mathbf{w}, \mathbf{x}_i\rangle}{\|\mathbf{x}_i\|}. \tag{12}$$

Note that in contrast to the classical margin $\gamma_{\boldsymbol{z}}(\mathbf{w})$ (see (5)) this *normalised* margin $\Gamma_{\boldsymbol{z}}(\mathbf{w})$ is a dimensionless quantity and constitutes a measure for the relative size of the version space invariant under rescaling of both weight vectors $\mathbf{w}$ and feature vectors $\mathbf{x}_i$.

IV. A PAC-Bayesian Margin Bound

Combining the ideas outlined in the previous section allows us to derive a generalisation error bound for linear classifiers $\mathbf{w} \in V(\boldsymbol{z})$ in terms of their *normalised* margin $\Gamma_{\boldsymbol{z}}(\mathbf{w})$.

Theorem 6. *Suppose $\mathcal{K} \subseteq \ell_2^n$ is a given feature space of dimensionality n . For all probability measures $\mathbf{P}_{\mathsf{Z}}$, for all $\delta \in (0,1]$ with probability at least $1-\delta$ over the random draw of the training sample $\boldsymbol{z} \in \mathcal{Z}^m$, for any consistent linear classifier $\mathbf{w} \in V(\boldsymbol{z})$ with positive margin $\Gamma_{\boldsymbol{z}}(\mathbf{w}) > 0$ the generalisation error $R[\mathbf{w}]$ of $\mathbf{w}$ is bounded from above by*

$$\frac{2}{m}\left(d \ln\left(\frac{1}{1-\sqrt{1-\Gamma_{\boldsymbol{z}}^2(\mathbf{w})}}\right) + \ln\left(\frac{(me)^2}{\delta}\right)\right) . \qquad (13)$$

where $d := \min(m,n) + 1$.

Proof: Geometrically the weight space $\mathcal{W}$ is the unit sphere in ℓ_2^n (see Figure 1). Let us assume that $\mathbf{P}_{\mathsf{W}}$ is uniform on the unit sphere as suggested by symmetry. Given the training sample $\boldsymbol{z}$ and a classifier $\mathbf{w}$, all classifiers $\mathbf{v} \in Q(\mathbf{w})$,

$$Q(\mathbf{w}) := \left\{ \mathbf{v} \in \mathcal{W} \,\middle|\, \langle \mathbf{w}, \mathbf{v} \rangle > \sqrt{1-\Gamma_{\boldsymbol{z}}^2(\mathbf{w})} \right\} , \qquad (14)$$

are within $V(\boldsymbol{z})$ (see Theorem 8). Such a set $Q(\mathbf{w})$ is Bayes-admissible by Lemma 5 and hence we can use $\mathbf{P}_{\mathsf{W}}(Q(\mathbf{w}))$ to bound the generalisation error of $\mathbf{w}$. Since $\mathbf{P}_{\mathsf{W}}$ is uniform, the value $-\ln(\mathbf{P}_{\mathsf{W}}(Q(\mathbf{w})))$ is simply the logarithm of the *volume ratio* between the surface of the unit sphere and $Q(\mathbf{w})$ (14). In Theorem 9 it is shown that the logarithm of this ratio is *exactly* given by

$$\ln\left(\frac{\int_0^{\pi} \sin^{n-2}(\theta)\, d\theta}{\int_0^{\arccos\left(\sqrt{1-\Gamma_{\boldsymbol{z}}^2(\mathbf{w})}\right)} \sin^{n-2}(\theta)\, d\theta}\right) .$$

For n odd, it can be shown that the logarithm of this ratio is bounded from above by (see Theorem 10)

$$n \cdot \ln\left(\frac{1}{1-\sqrt{1-\Gamma_{\boldsymbol{z}}^2(\mathbf{w})}}\right) + \ln(2) .$$

With $\ln(2) < 1$ we obtain the desired result. Note that m points maximally span an m -dimensional space and thus we can marginalise over the remaining $n-m$ dimensions of feature space[7]. This gives $d = \min(m,n) + 1$. ■

An appealing feature of (13) is that for $\Gamma_{\boldsymbol{z}}(\mathbf{w}) = 1$ the bound reduces to $\frac{2}{m}(2\ln(m) - \ln(\delta) + 2)$ with a rapid decay to zero as m increases. In the practice of kernel classification, one often encounters the case that the dimensionality n of the feature space exceeds the number m of training examples. This happens, for example, when the classification is carried out in feature spaces that are induced by mercer kernels with an infinite expansion in terms of eigenfunctions, as is, for example, the case for the so-called Gaussian RBF kernel that is frequently used in kernel methods such as the support vector machine [1]. Even in this case the bound may give non-trivial values for margins $\Gamma_{\boldsymbol{z}}(\mathbf{w}) > 0.91$. Furthermore, upper bounding $1/(1-\sqrt{1-\Gamma_{\boldsymbol{z}}^2(\mathbf{w})})$ by $2/\Gamma_{\boldsymbol{z}}^2(\mathbf{w})$ we see that Theorem 6 is an exponential improvement of Theorem 1 in terms of the attained margins. It should be noted, however, that in contrast to Theorem 1 the new bound depends on the dimensionality of the input space via $d = \min(m,n)$ and thus cannot serve as a direct motivation for margin maximisation in the case of $d = m$.

V. Experimental Study

A. Model Selection

In order to investigate the tightness of the PAC-Bayesian margin bound we performed a controlled experiment where we considered varying distributions in feature space $\mathcal{K} = \ell_2^2$. Note that for a fixed input distribution $\mathbf{P}_{\mathsf{X}}$ every feature mapping (model) $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{K} = \ell_2^2$ into a two-dimensional feature space incurs a different distribution in $\mathcal{K}$. The points $\mathbf{x} \in \mathcal{K}$ were generated according to

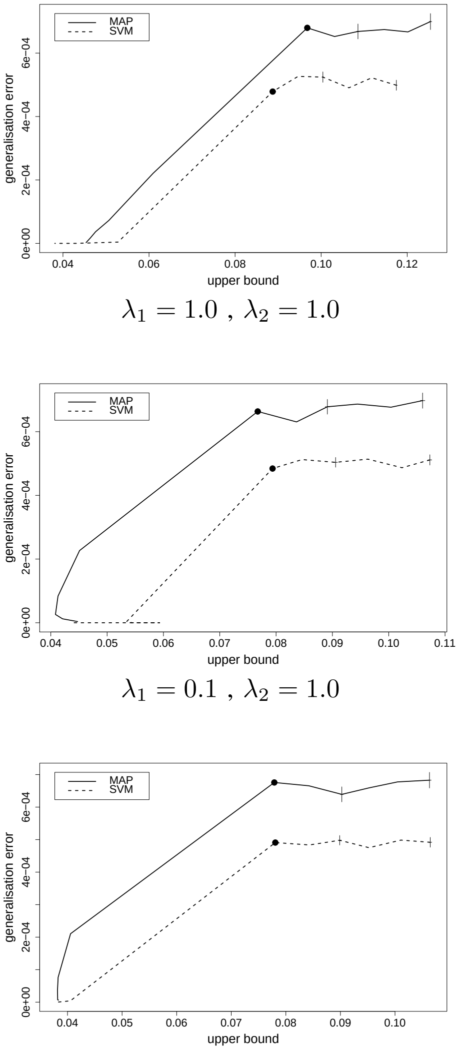
$$\mathbf{x} = (\lambda_1 \cos(\pi\beta), \lambda_2 \sin(\pi\beta))' . \qquad (15)$$

The incorporation of λ_1 and λ_2 allowed us to consider circles as well as ellipses. The hypothesis $\mathbf{w}^*$ labelling the data was given by $\mathbf{w}^* = (1,0)'$. By specifying a Beta(α ,α) distribution[8] over β we were able to vary from very good models far apart from the teacher (small α values) to very bad models concentrated at the teachers decision boundary (large α values). Note that the generalisation error $R[\mathbf{w}]$ of $\mathbf{w}$ is given by the probability of training points that fall into the "cone" $\{\mathbf{x} \mid \langle \mathbf{x}, \mathbf{w}^* \rangle \cdot \langle \mathbf{x}, \mathbf{w} \rangle < 0\}$.

In a set of experiments we checked the potential of the PAC-Bayesian margin bound for model selection. Using the shape parameter α we varied across different models. Over a random draw of 1000 training samples $\boldsymbol{z}$ for each α value we calculated the generalisation error of the solution as well as the bound valued based on the observed margin using (13) with $\delta = 0.05$. These two curves were linked via their common α -axis, that is, for the x -values we took the *mean upper bound value* and as the y -values the *mean generalisation error.* Horizontal bars (very short) are error bars for the upper bound, vertical bars are error bars for the generalisation error. In order to check the influence of ellipsoidal shapes we varied λ_1 and λ_2 . The curves named SVM were obtained by a modified SVM (maximising $\Gamma_{\boldsymbol{z}}$ instead of $\gamma_{\boldsymbol{z}}$ as in the the classical SVM); the curves named MAP (maximum a-posteriori estimator) were obtained by choosing $\mathbf{w} \in V(\boldsymbol{z})$ randomly from version space sampling from a uniform distribution, and using their margin $\Gamma_{\boldsymbol{z}}(\mathbf{w})$ for the calculation of the bound value. Note, that due to the uniform prior $\mathbf{P}_{\mathsf{W}}$ and the PAC-likelihood (3) the MAP estimator is not uniquely defined and any $\mathbf{w} \in V(\boldsymbol{z})$ maximises the posterior. The black dot indicates $\alpha = 1$. The

[7] More formally, the uniform distribution $\mathbf{P}_{\mathsf{W}}$ over the unit sphere in ℓ_2^n can be considered as a marginalised isotropic Gaussian in ℓ_2^n . Let $\mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_m) \subseteq \mathcal{K}$ be the m -dimensional subspace spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and let $\mathbf{P}_{\boldsymbol{x}}$ be the corresponding projection operator. Then, for any $\mathbf{x}_i$ in the training sample $\boldsymbol{z}$ and any $\mathbf{w} \in \mathcal{W}$, $\langle \mathbf{w}, \mathbf{x}_i \rangle = \langle \mathbf{P}_{\boldsymbol{x}} \mathbf{w}, \mathbf{x}_i \rangle$. Hence, we can further marginalise over $\mathcal{K} \cap \mathcal{L}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ resulting likewise in an isotropic Gaussian (see, e.g. [13]). As a consequence, it suffices to consider the uniform distribution over an m -dimensional unit sphere.

[8] We use the following parameterisation of the Beta density: $\mathbf{f}_{\mathrm{Beta}(\mu,\nu)}(x) := \frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)} x^{\mu-1} (1-x)^{\nu-1}$ with $\mu, \nu \in \mathbb{N}$ and $0 < x < 1$.



$\lambda_1 = 1.0$, $\lambda_2 = 1.0$

$\lambda_1 = 0.1$, $\lambda_2 = 1.0$

$\lambda_1 = 1.0$, $\lambda_2 = 0.1$

Figure 2. Plots of Upper bound versus generalisation error for a fixed training sample size of $m = 1000$. The black dot indicates $\alpha = 1$, that is, a uniform distribution over angles β (see (15)). See text for further explanation.

training sample size was fixed to $m = 1000$. The results are given in Figure 2. There are some interesting conclusions to be drawn:

- A straight line in the plot corresponds to equivalent shapes of the generalisation error and upper bound curve. We observe that *independently* of the shape of the feature space the bound resembles the shape of the generalisation error curve for all fortunate distributions up to the uniform distribution over angles $\pi\beta$. From this point on the bound increases while the generalisation error remains approximately constant.
- *Independently* of the shape of the feature space the bound values remain constant, that is, the x -axes in our plots always range from 0.04 to 0.12 . This is a *real* advantage over classical results which are sensitive to the length of training inputs $\mathbf{x}$ due to the unnormalised margin $\gamma_{\boldsymbol{z}}(\mathbf{w})$ (which is by construction always less than λ_1).
- For more ellipsoidal shapes of the feature space, margin maximisation techniques (SVM curves) are much more advantageous than "random guessing" in version space (MAP curves). This can be seen by observing the difference in the solid and dashed curves.
- There are *implicit* assumptions in bound-based model selection: It is assumed that a change of the model (feature space) leads to distributions which have large margins with high probability ($\alpha < 1$). Otherwise the bound value need not be related to the generalisation error. Nevertheless, *minimisation* of the bound would lead to such models.
- Although there is still a gap of approximately a factor of 100 between the estimated generalisation error (y -axis) and the upper bound value (x -axis) we would like to recall that the classical PAC margin bound given in Theorem 1, even in the best case of $\gamma_{\boldsymbol{z}}(\mathbf{w}) = 1$, needs the astronomically large training sample size $m = 1\,555\,878$ to achieve a value of 0.04 .

B. Normalising Data in Feature Space

Theorem 6 suggest the following learning algorithm: given a version space $V(\boldsymbol{z})$ (through a given training sample $\boldsymbol{z}$) find the classifier $\mathbf{w}$ that maximises $\Gamma_{\boldsymbol{z}}(\mathbf{w})$. This algorithm, however, is given by the SVM *only if* the training data in feature space $\mathcal{K}$ are normalised. We investigate the influence of such a normalisation on the generalisation error in the feature space $\mathcal{K}$ of all monomials up to the p th degree (well-known from handwritten digit recognition, see [1]). Since the SVM learning algorithm as well as the resulting classifier only refer to inner products in $\mathcal{K}$, it suffices to use an easy-to-calculate kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $x, \tilde{x} \in \mathcal{X}$, $k(x, \tilde{x}) = \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(\tilde{x}) \rangle$, given in our case by the polynomial kernel

$$\forall p \in \mathbb{N} \qquad k(x, \tilde{x}) := (\langle x, \tilde{x} \rangle + 1)^p \ .$$

Earlier experiments have shown [1] that without normalisation too large values of the exponent p may lead to "over-fitting". We used the UCI [15] data sets `thyroid` ($d = 5$, $m = 140$, $m_{\text{test}} = 75$) and `sonar` ($d = 60$, $m = 124$, $m_{\text{test}} = 60$) and plotted the generalisation error of SVM solutions (estimated over 100 different splits of the data set) as a function of p (see Figures 3 and 4). As suggested by Theorem 6 in almost all cases the normalisation improved the performance of the support vector machine solution at a statistically significant level. As a consequence, we recommend to always normalise data in feature space when training an SVM. Intuitively, it is only the *spatial direction* of both weight vector and feature vectors that determines the classification. Hence the different lengths of feature vectors in the training sample should not enter the SVM optimisation problem.

Note, that an alternative algorithm, the so-called Bayes point machine (BPM), also assumes a uniform prior P_{W} over weight space, but returns the centre of mass weight vector $\mathbf{w}_{\text{cm}} := \mathsf{E}_{\mathsf{W}|\mathsf{W}\in V(\boldsymbol{z})}[\mathsf{W}]$ obtained from averaging over samples from $\mathsf{P}_{\mathsf{W}|\mathsf{W}\in V(\boldsymbol{z})}$ (for example using the kernel Gibbs sampler [16]). The BPM has empirically been demonstrated

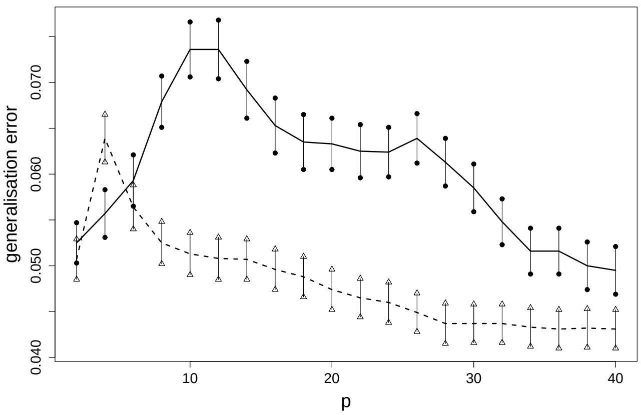


Figure 3. Estimated generalisation errors of classifiers learned by an SVM with (dashed line) and without (solid line) normalisation of the feature vectors $\mathbf{x}_i$. The error bars indicate one standard deviation over 100 random splits of the data sets. The plot is obtained on the `thyroid` dataset.

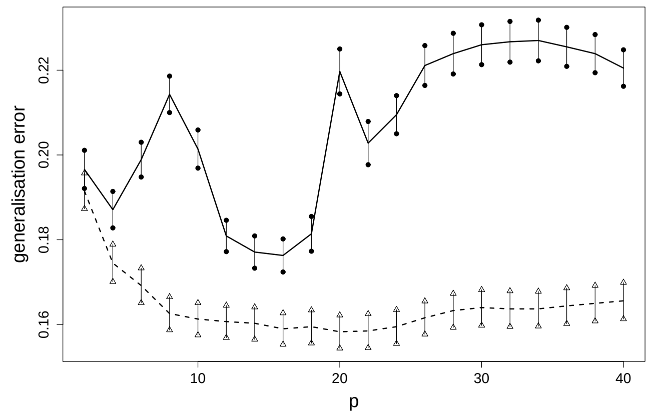


Figure 4. Estimated generalisation error on the `sonar` data set. For details see Figure 3.

to have generalisation properties superior to the support vector machine in the case of elongated version spaces [17]. However, it has not been justified in the sense that it minimises a PAC-style error bound as is the case for support vector machines.

VI. Discussion and Conclusion

The PAC-Bayesian framework together with simple geometrical arguments yields the so far tightest margin bound for linear classifiers. The novelty of the current approach to prove a PAC generalisation error bound in terms of a margin lies in the fundamentally different reasoning applied: classical PAC techniques use a ghost sample argument to consider equivalence classes of classifiers on a double sample and then aim at bounding the worst case number of equivalence classes in terms (or at the scale) of the margin observed (see, for example, [18], [19], [20], [3]). Interestingly, the *pure* covering number bound (see [21, Lemma 4] and [22, Theorem 6.8]) seems to be of the same order as our current result. The weakness of PAC margin bounds such as Theorem 1 mainly comes from the application of Alon's lemma (see [6]) when bounding the covering number at the observed margin scale by the fat shattering dimension.

In the current proof we avoided the usage of double samples and covering numbers at all. Instead we used pure volume ratio arguments and demonstrated that the margin $\Gamma_{\boldsymbol{z}}$ has a natural interpretation of characterising a subset of classifiers in version space summarised by the classifier under consideration. By instantiating a general result in the PAC-Bayesian framework with a special prior over classifiers we only needed to bound volume ratios in weight space. It is worthwhile mentioning that the quantity $\mathsf{P}_{\mathsf{W}}^{-1}\left(Q\left(\mathbf{w}\right)\right)$ can be considered as an upper bound on the packing number of classifiers with a margin at least $\Gamma_{\boldsymbol{z}}$ because if a classifier $\mathbf{w}_i$ has a consistent region $Q\left(\mathbf{w}_i\right)$ of classifiers around it then $\sum_{\mathbf{w}_i} \mathsf{P}_{\mathsf{W}}\left(Q\left(\mathbf{w}_i\right)\right) \leq 1$ by definition of probability and the disjointness of the $Q\left(\mathbf{w}_i\right)$ which implies that there are no more than $\max_i \left(\mathsf{P}_{\mathsf{W}}^{-1}\left(Q\left(\mathbf{w}_i\right)\right)\right)$ different weight vectors $\mathbf{w}_i$.

An interesting question is the role of the number d in Theorem 6, that is, the minimum of the number of dimensions of feature space and the training sample size. Although it appears that this number limits the applicability of the current results to low dimensional feature spaces it seems possible to reduce this number by considering different low dimensional projections of the training data *before* having seen the data and minimising the bound with respect to the increase in the margin $\Gamma_{\boldsymbol{z}}$ when projecting the training data to a low dimensional manifold. Our future work is focused on combining ideas from approximation theory and the theory of reproducing kernel Hilbert spaces to adopt our main theorem to take advantage of the eigenvalue spectrum of the observed inner product matrix $\mathbf{G}$, $\mathbf{G}_{ij} := \langle \mathbf{x}_i, \mathbf{x}_j \rangle = k\left(x_i, x_j\right)$ (see [23] for first results of this type).

The proof of the margin bound presented revealed an important point already stressed by Shawe-Taylor et al. in their seminal work on luckiness [8]: The margin as a characterisation of consistent classifiers is *one possible* prior belief in the data distribution underlying the training sample. Whenever this prior belief is not met by the observed training sample the margin bound will be trivial although still valid in the PAC sense. The arbitrariness of the prior P_{W} chosen in our proof should best be compared with the arbitrariness of a luckiness function when applying the main result of [8].

Acknowledgements

We would like to thank David McAllester, John Shawe-Taylor, Bob Williamson, Olivier Chapelle, John Langford, Alex Smola and Bernhard Schölkopf for interesting discussions and useful suggestions on earlier drafts. We are particularly indebted to Peter Bollmann-Sdorra for his invaluable help in proving Theorem 10. We also thank the two anonymous reviewers and Gabor Lugosi for their comments and suggestions that helped us to improve the manuscript.

Appendix

The appendix contains the lemmata and theorems necessary to prove our main theorem, Theorem 6. Due to the length of the single proofs we have split them into separate sections.

Balls in Version Space

In this section we prove that the ball

$$Q(\mathbf{w}) := \left\{ \mathbf{v} \in \mathcal{W} \;\middle|\; \langle \mathbf{w}, \mathbf{v} \rangle > \sqrt{1 - \Gamma_{\boldsymbol{z}}^2(\mathbf{w})} \right\}$$

around a linear classifier with normal $\mathbf{w}$ of unit length only contains classifiers within version space $V(\boldsymbol{z})$. Here, $\Gamma_{\boldsymbol{z}}(\mathbf{w})$ is the margin of the hyper-plane $\mathbf{w}$ on a set of points *normalised by the length* $\|\mathbf{x}_i\|$ *of the* $\mathbf{x}_i$ (see (12) for a formal definition). In order to prove this result we need the following lemma.

Lemma 7. *Suppose $\mathcal{K} \subseteq \ell_2^n$ is a fixed feature space. Assume we are given two points $\mathbf{w} \in \mathcal{W}$ and $\mathbf{x} \in \mathcal{K}$ such that $\langle \mathbf{w}, \mathbf{x} \rangle =: \gamma > 0$. Then for all $\mathbf{v} \in \mathcal{W}$ with*

$$\langle \mathbf{w}, \mathbf{v} \rangle > \sqrt{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}} \tag{16}$$

it follows that $\langle \mathbf{v}, \mathbf{x} \rangle > 0$.

Proof: Given $\mathbf{w} \in \mathcal{W}$ and $\mathbf{x} \in \mathcal{K}$ let $\mathcal{L}(\mathbf{x}, \mathbf{w}) \subset \mathcal{K}$ be the linear subspace of $\mathcal{K}$ spanned by $\mathbf{x}$ and $\mathbf{w}$. Then we can express any $\mathbf{v} \in \mathcal{K}$ as $\mathbf{v} = \mathbf{v}_{\parallel} + \mathbf{v}_{\perp}$ where $\mathbf{v}_{\parallel}$ is the projection of $\mathbf{v}$ onto $\mathcal{L}(\mathbf{x}, \mathbf{w})$ and $\mathbf{v}_{\perp}$ is its projection on the complement of $\mathcal{L}(\mathbf{x}, \mathbf{w})$. Since we have $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}_{\parallel}, \mathbf{w} \rangle$ and $\langle \mathbf{v}, \mathbf{x} \rangle = \langle \mathbf{v}_{\parallel}, \mathbf{x} \rangle$ we can make the following ansatz for $\mathbf{v}$ without loss of generality,

$$\mathbf{v} = \lambda \frac{\mathbf{x}}{\|\mathbf{x}\|} + \tau \left(\mathbf{w} - \gamma \frac{\mathbf{x}}{\|\mathbf{x}\|^2} \right) .$$

Note, that the vectors $\frac{\mathbf{x}}{\|\mathbf{x}\|}$ and $\mathbf{w} - \gamma \frac{\mathbf{x}}{\|\mathbf{x}\|^2}$ are orthogonal by construction. Furthermore, the squared length of $\mathbf{w} - \gamma \frac{\mathbf{x}}{\|\mathbf{x}\|^2}$ is given by $1 - \gamma^2 / \|\mathbf{x}\|^2$. Therefore, the unit norm constraint on $\mathbf{v}$ implies that

$$\tau^2 = \frac{1 - \lambda^2}{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}} .$$

Furthermore, assumption (16) becomes

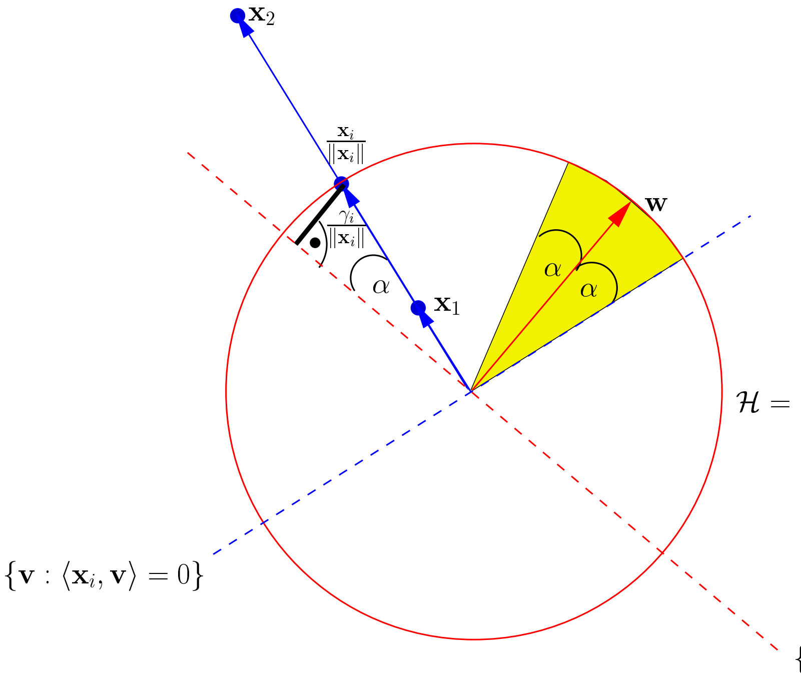
$$\begin{aligned} \left\langle \lambda \frac{\mathbf{x}}{\|\mathbf{x}\|} + \tau \left(\mathbf{w} - \gamma \frac{\mathbf{x}}{\|\mathbf{x}\|^2} \right), \mathbf{w} \right\rangle &> \sqrt{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}} \\ \lambda \frac{\gamma}{\|\mathbf{x}\|} \pm \sqrt{\frac{1 - \lambda^2}{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}}} \left(1 - \frac{\gamma^2}{\|\mathbf{x}\|^2} \right) &> \sqrt{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}} \\ \underbrace{\lambda \frac{\gamma}{\|\mathbf{x}\|} - \sqrt{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}} \left(1 \pm \sqrt{1 - \lambda^2} \right)}_{f(\lambda)} &> 0 . \end{aligned}$$

In order to solve for λ we consider the left-hand-side as a function of λ and determine the range of values where $f(\lambda)$ is positive. A straightforward calculation reveals that $[0, \lambda_{\max}]$ with

$$\lambda_{\max} = \frac{2\gamma}{\|\mathbf{x}\|} \sqrt{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}} ,$$

is the only range where $f(\lambda)$ is positive. Thus, the assumption $\langle \mathbf{w}, \mathbf{v} \rangle > \sqrt{1 - \gamma^2 / \|\mathbf{x}\|^2}$ implies

$$0 < \lambda \|\mathbf{x}\| < 2\gamma \sqrt{1 - \frac{\gamma^2}{\|\mathbf{x}\|^2}} .$$



$\mathbf{x}_2$
$\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$
$\frac{\gamma_i}{\|\mathbf{x}_i\|}$
α
$\mathbf{x}_1$
$\mathbf{w}$
α
α
$\mathcal{H} =$
$\{\mathbf{v} : \langle \mathbf{x}_i, \mathbf{v} \rangle = 0\}$
$\{$

Figure 5. Suppose the point $\mathbf{x}_1$ (or $\mathbf{x}_2$) is given. We have to show that all classifiers with normal $\tilde{\mathbf{w}}$ of unit length and $\langle \mathbf{w}, \tilde{\mathbf{w}} \rangle > \sqrt{1 - \gamma_i^2 / \|\mathbf{x}_i\|^2}$ are on the same side of the hyper-plane $\{\mathbf{v} \mid \langle \mathbf{x}_i, \mathbf{v} \rangle = 0\}$, that is $\langle \tilde{\mathbf{w}}, \mathbf{x}_i \rangle > 0$, where $\gamma_i = \langle \mathbf{x}_i, \mathbf{w} \rangle$. From the picture it is clear that regardless of $\|\mathbf{x}_i\|$, $\sin(\alpha) = (\gamma_i / \|\mathbf{x}_i\|)$ or equivalently $\cos(\alpha) = \sqrt{1 - \sin^2(\alpha)} = \sqrt{1 - \gamma_i^2 / \|\mathbf{x}_i\|^2}$. Obviously, all vector $\tilde{\mathbf{w}}$ of unit length which enclose an angle less than α with $\mathbf{w}$ are on the same side (the dark cone). As $\cos(\alpha)$ is monotonically decreasing for $\alpha \in \left(0, \frac{\pi}{2}\right)$, these classifiers have to fulfil $\langle \mathbf{w}, \tilde{\mathbf{w}} \rangle = \cos(\sphericalangle(\mathbf{w}, \tilde{\mathbf{w}})) > \sqrt{1 - \gamma_i^2 / \|\mathbf{x}_i\|^2}$.

Finally, the inner product of any $\mathbf{v}$ with $\mathbf{x}$ is given by

$$\begin{aligned} \langle \mathbf{v}, \mathbf{x} \rangle &= \left\langle \lambda \frac{\mathbf{x}}{\|\mathbf{x}\|} + \tau \left(\mathbf{w} - \gamma \frac{\mathbf{x}}{\|\mathbf{x}\|} \right), \mathbf{x} \right\rangle \\ &= \lambda \|\mathbf{x}\| + \tau (\gamma - \gamma) > 0 , \end{aligned}$$

where the last inequality follows from the previous consideration. The lemma is proven. For a geometrical reasoning see Figure 5. ■

Theorem 8. *Suppose $\mathcal{K} \subseteq \ell_2^n$ is a fixed feature space. Given a training sample $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in (\mathcal{X} \times \{-1, +1\})^m$ and $\mathbf{w} \in \mathcal{W}$ such that $\Gamma_{\boldsymbol{z}}(\mathbf{w}) > 0$, for all $\mathbf{v} \in \mathcal{W}$ such that $\langle \mathbf{w}, \mathbf{v} \rangle > \sqrt{1 - \Gamma_{\boldsymbol{z}}^2(\mathbf{w})}$ we have*

$$\forall i \in \{1, \ldots, m\} : \qquad y_i \langle \mathbf{v}, \mathbf{x}_i \rangle > 0 .$$

Proof: According to Lemma 7 we know that all $\mathbf{v} \in B_i$ with

$$B_i := \left\{ \mathbf{v} \in \mathcal{W} \;\middle|\; \langle \mathbf{w}, \mathbf{v} \rangle > \sqrt{1 - \frac{(y_i \langle \mathbf{x}_i, \mathbf{w} \rangle)^2}{\|\mathbf{x}_i\|^2}} \right\} ,$$

parameterise classifiers consistent with the i th point $\mathbf{x}_i$. Clearly, the intersection of all B_i gives the classifiers $\mathbf{w}$ which jointly fulfil the constraints $y_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0$. Noticing that the size of B_i depends inversely on $y_i \langle \mathbf{x}_i, \mathbf{w} \rangle$ we see that all $\mathbf{v}$ such that $\langle \mathbf{w}, \mathbf{v} \rangle > \Gamma_{\boldsymbol{z}}(\mathbf{w})$ jointly classify all points $\mathbf{x}_i$ correctly. The theorem is proven. ■

VOLUME RATIO THEOREM

In this section we explicitly derive the volume ratio between the largest inscribable ball in version space and the whole parameter space for the special case of linear classifiers in $\mathbb{R}^n$. Given a point $\mathbf{w} \in \mathcal{W}$ and a positive number $\gamma > 0$ we can characterise the ball of radius γ in the parameter space by

$$\begin{aligned} Q_\gamma(\mathbf{w}) &:= \left\{ \mathbf{v} \in \mathcal{W} \;\middle|\; \|\mathbf{w} - \mathbf{v}\|^2 < \gamma^2 \right\} \\ &= \left\{ \mathbf{v} \in \mathcal{W} \;\middle|\; \langle \mathbf{w}, \mathbf{v} \rangle > 1 - \gamma^2/2 \right\} . \end{aligned}$$

In the following we will calculate the exact value of the *volume ratio* $\mathrm{vol}(\mathcal{W}) / \mathrm{vol}(Q_\gamma(\mathbf{w}))$ where $\mathbf{w}$ can be chosen arbitrarily (due to the symmetry of the sphere).

Theorem 9. *Suppose we are given a fixed feature space $\mathcal{K} \subseteq \ell_2^n$. Then the fraction of the whole surface $\mathrm{vol}(\mathcal{W})$ of the unit sphere to the surface $\mathrm{vol}(Q_\gamma(\mathbf{w}))$ with Euclidean distance less than γ from any point $\mathbf{v} \in \mathcal{W}$ is given by*

$$\frac{vol(\mathcal{W})}{vol(Q_\gamma(\mathbf{w}))} = \frac{\int_0^\pi \sin^{n-2}(\theta)\, d\theta}{\int_0^{\arccos\left(1 - \frac{\gamma^2}{2}\right)} \sin^{n-2}(\theta)\, d\theta} . \qquad (17)$$

Proof: Consider spherical coordinates such that every $\mathbf{w} \in \mathcal{W}$ is expressed via $n-2$ angles $\theta_1, \ldots, \theta_{n-2}$ ranging from 0 to π , and one angle $0 \le \varphi \le 2\pi$. Choose $\mathbf{w} = (1, 0, \ldots, 0)'$ and the coordinate θ such that $w_1 = \cos(\theta)$. Then the intersection $\mathcal{W}_{\theta_0}^{n-1} := \mathcal{W} \cap \{\mathbf{w} \,| \theta = \theta_0\}$ is an $(n-1)$ -dimensional unit hyper-sphere of radius $r_{n-1} = \sin(\theta)$. The surface area $\mathrm{vol}\left(\mathcal{W}_{\theta_0}^{n-1}\right)$ of $\mathcal{W}_{\theta_0}^{n-1}$ is given by

$$\mathrm{vol}\left(\mathcal{W}_{\theta_0}^{n-1}\right) = c_{n-1} \sin^{n-2}(\theta) . \qquad (18)$$

The ratio can be expressed in terms of integrals over θ_0 as

$$\frac{\mathrm{vol}(\mathcal{W})}{\mathrm{vol}(Q_\gamma(\mathbf{w}))} = \frac{\int_0^\pi \mathrm{vol}\left(\mathcal{W}_{\theta_0}^{n-1}\right) d\theta_0}{\int_0^{\arccos\left(1 - \frac{\gamma^2}{2}\right)} \mathrm{vol}\left(\mathcal{W}_{\theta_0}^{n-1}\right) d\theta_0} ,$$

which together with (18) proves the theorem. ■

A VOLUME RATIO BOUND

In this section we present a practically useful upper bound for the logarithm of the expression given in (17). In order to check the usefulness of this expression we have compared the exact value with the upper bound and found that in the interesting regime of large margins the bound seems to be within a factor of 2 from the exact value (see Figure 6).

Theorem 10. *For all $j \in \mathbb{N}$ and all $0 < x \le \frac{1}{2}$*

$$\ln\left(\frac{\int_0^\pi \sin^{2j+1}(\theta)\, d\theta}{\int_0^{\Psi(x)} \sin^{2j+1}(\theta)\, d\theta} \right) \le \ln\left(\frac{1}{2x}\right)^{2j+1} + \ln(2) , \quad (19)$$

where $\Psi(x) := \arccos(1 - 2x)$.

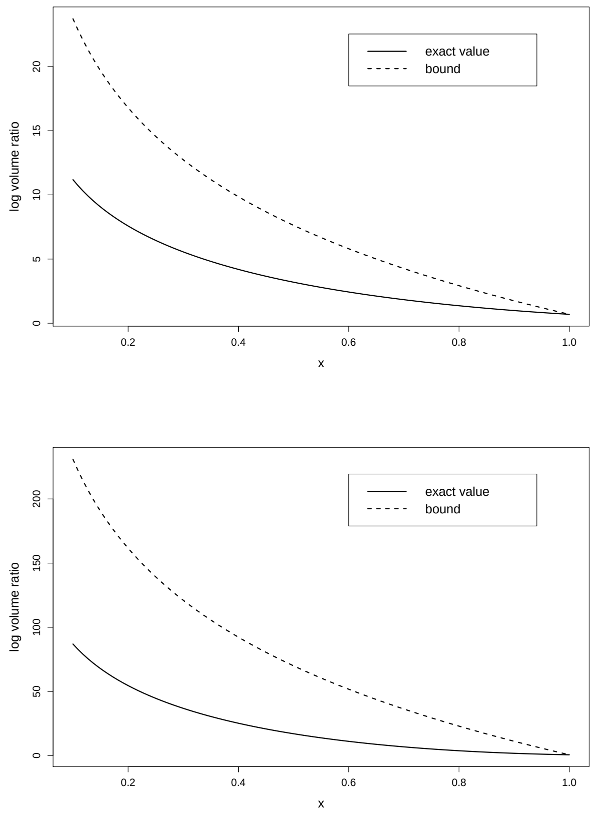


Figure 6. Comparison of the bound (r.h.s. of (19)) (dashed line) with the exact value (l.h.s. of (19)) (solid line) over the whole range of possible values of x for **(top)** $n = 10$ and **(bottom)** $n = 100$. Interestingly, in the relevant regime of large values of x the bound seems to be very tight regardless of the number n of dimensions.

Proof: From [24] we know that for all $j \in \mathbb{N}$

$$\int \sin^{2j+1}(\theta)\, d\theta = -\frac{\cos(\theta)}{2j+1} \sum_{i=0}^{j} \sin^{2i}(\theta)\, B_{j,i}, \qquad (20)$$

where

$$\begin{aligned} B_{j,i} &:= \frac{2(i+1) \cdot 2(i+2) \cdot \cdots \cdot 2j}{(2i+1) \cdot (2i+3) \cdot \cdots \cdot (2j-1)} \\ &= \frac{2 \cdot 4 \cdots 2j}{1 \cdot 3 \cdots (2j-1)} \cdot \frac{1 \cdot 3 \cdots (2i-1)}{2 \cdot 4 \cdots (2i)} \\ &= \frac{4^j (j!)^2 (2i)!}{(2j)! (i!)^2 4^i} = \frac{4^j}{4^i} \frac{\binom{2i}{i}}{\binom{2j}{j}} . \end{aligned} \qquad (21)$$

Let us introduce the abbreviation

$$S(j, x) := \int_0^{\arccos(1-2x)} \sin^{2j+1}(\theta)\, d\theta .$$

Then the numerator of (19) is given by $S(j, 1)$ whereas the denominator of (19) is simply $S(j, x)$. From (20) we see

$$\begin{aligned} S(j, x) &= -\frac{\cos(\theta)}{2j+1} \sum_{i=0}^{j} \sin^{2i}(\theta)\, B_{j,i} \Bigg|_0^{\arccos(1-2x)} \\ &= \frac{4^j}{(2j+1)\binom{2j}{j}} \left(1 + (2x-1) \sum_{i=0}^{j} \binom{2i}{i} x^i (1-x)^i \right) . \end{aligned}$$

where we have used (21) and

$$\begin{aligned}\sin^{2i}(\theta) &= \left(\sin^2(\theta)\right)^i = \left(1-\cos^2(\theta)\right)^i \\ &= \left(1-(1-2x)^2\right)^i = \left(4x-4x^2\right)^i .\end{aligned}$$

For the fraction we obtain

$$\ln\left(\frac{S(j,1)}{S(j,x)}\right) = -\ln\left(\frac{1}{2}\left(2x+(2x-1)\sum_{i=1}^{j}\binom{2i}{i}x^i(1-x)^i\right)\right).$$

In Lemma 15 we show that for any $j \in \mathbb{N}^+$ and $0 \le x < \frac{1}{2}$

$$\sum_{i=1}^{j}\binom{2i}{i}x^i(1-x)^i \le \frac{2x\left((2x)^{2j}-1\right)}{2x-1} .$$

Inserted into the last expression we obtain

$$\begin{aligned}\ln\left(\frac{S(j,1)}{S(j,x)}\right) &\le -\ln\left(\frac{1}{2}\left(2x+(2x-1)\frac{2x\left((2x)^{2j}-1\right)}{(2x-1)}\right)\right) \\ &= -\ln\left(\frac{(2x)^{2j+1}}{2}\right) \\ &= -(2j+1)\ln(2x)+\ln(2) .\end{aligned}$$

In the case of $x = \frac{1}{2}$ the problem reduces to showing that

$$\ln\left(\frac{S(j,1)}{S\left(j,\frac{1}{2}\right)}\right) = \underbrace{-(2j+1)\ln(2x)}_{0} + \ln(2) .$$

This is equal to the exact value of the volume ratio (19) with $x = \frac{1}{2}$ because we have for all $j \in \mathbb{N}$

$$\int_0^{\pi}\sin^{2j+1}(\theta)\,d\theta = 2\cdot\int_0^{\frac{\pi}{2}}\sin^{2j+1}(\theta)\,d\theta .$$

■

Bollmann's Lemma

In the course of the proof of Theorem 10 we need a tight upper bound on $\sum_{i=1}^{j}\binom{2i}{i}x^i(1-x)^i$ as a function of x . In the following we present a series of lemmata resulting in a reasonably accurate upper bound that we called *Bollmann's lemma*[9] (Lemma 15).

Lemma 11. *For all* $i \in \mathbb{N}^+$

$$\binom{2(i+1)}{i+1} = \binom{2i}{i}\left(4-\frac{2}{i+1}\right) .$$

Proof: A straightforward calculation shows that

$$\begin{aligned}\binom{2(i+1)}{i+1} &= \frac{2(i+1)(2i+1)}{(i+1)(i+1)}\binom{2i}{i} \\ &= \binom{2i}{i}\frac{4i+2}{i+1} \\ &= \binom{2i}{i}\left(4-\frac{2}{i+1}\right) .\end{aligned}$$

The lemma is proven. ■

Lemma 12. *For all* $i \in \mathbb{N}^+$ *and* $j \in \mathbb{N}^+$

$$\binom{2(j+1)}{j+1}\binom{2i}{i} \le 2\binom{2(i+j)}{i+j} .$$

Proof: We prove the lemma by induction over i . For $i = 1$ it follows that

$$\binom{2(j+1)}{j+1}\binom{2}{1} = 2\binom{2(j+1)}{j+1} .$$

Assume the assertion is true for $i \in \mathbb{N}^+$. Then

$$\begin{aligned}&\binom{2(j+1)}{j+1}\binom{2(i+1)}{i+1} \\ &\qquad = \binom{2(j+1)}{j+1}\binom{2i}{i}\left(4-\frac{2}{i+1}\right) \\ &\qquad \le 2\binom{2(i+j)}{i+j}\left(4-\frac{2}{i+1}\right) \\ &\qquad \le 2\binom{2(i+j)}{i+j}\left(4-\frac{2}{i+j+1}\right) \\ &\qquad = 2\binom{2(i+j+1)}{i+j+1},\end{aligned}$$

where we used Lemma 11 in the first and last line. ■

Lemma 13. *For all* $0 \le x < \frac{1}{2}$

$$\sum_{i=1}^{\infty}\binom{2i}{i}x^i(1-x)^i = \frac{2x}{1-2x} .$$

Proof: This can be seen by considering

$$\begin{aligned}\arcsin(u) &= u+\sum_{i=1}^{\infty}\binom{2i}{i}\frac{1}{4^i}\frac{u^{2i+1}}{2i+1} , \\ \frac{d\arcsin(u)}{du} &= 1+\sum_{i=1}^{\infty}\binom{2i}{i}\frac{1}{4^i}u^{2i} \\ &= \frac{1}{\sqrt{1-u^2}} .\end{aligned}$$

Using $u = 2\sqrt{x(1-x)}$ we obtain the result, i.e.

$$\begin{aligned}\sum_{i=1}^{\infty}\binom{2i}{i}\frac{1}{4^i}\left(2\sqrt{x(1-x)}\right)^{2i} &= \frac{1}{\sqrt{1-4x(1-x)}}-1 \\ \sum_{i=1}^{\infty}\binom{2i}{i}x^i(1-x)^i &= \frac{1-\sqrt{1-4x(1-x)}}{\sqrt{1-4x(1-x)}} \\ &= \frac{1-\sqrt{(1-2x)^2}}{\sqrt{(1-2x)^2}} .\end{aligned}$$

The lemma is proven. ■

Lemma 14. *For all* $0 \le x < \frac{1}{2}$ *and* $j \in \mathbb{N}^+$

$$\begin{aligned}&4x^2\sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j} \le \\ &\qquad \sum_{i=1}^{\infty}\binom{2(i+j+1)}{i+j+1}x^{i+j+1}(1-x)^{i+j+1} .\end{aligned}$$

[9] We do not know of any prior appearance of this inequality in the literature. The name "Bollmann's lemma" was chosen in honour of our colleague Peter Bollmann-Sdorra, who proved this result based on what he calls "high school algebra". He thought it too minor to justify an authorship.

Proof: By Lemma 12 we have

$$\sum_{i=1}^{\infty}\binom{2i}{i}\binom{2(j+1)}{j+1}x^{i+j}(1-x)^{i+j} \le \sum_{i=1}^{\infty}2\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j} .$$

As $0 < 1-x \le 1+2x$ we have that

$$(1-x)\binom{2(j+1)}{j+1}x^{j}(1-x)^{j}\sum_{i=1}^{\infty}\binom{2i}{i}x^{i}(1-x)^{i} \quad \le \quad 2(1+2x)\sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j}$$

implies

$$(1-x)\binom{2(j+1)}{j+1}x^{j}(1-x)^{j}\frac{2x}{1-2x} \quad \le \quad 2(1+2x)\sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j} .$$

Multiplying both sides by $\frac{1-2x}{2}$ (which is by assumption positive) yields

$$\binom{2(j+1)}{j+1}x^{j+1}(1-x)^{j+1} \le \left(1-4x^2\right)\sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j} .$$

Rearranging terms gives

$$\begin{aligned}
&4x^2\sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j} \\
&\le \sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j} \\
&\qquad -\binom{2(j+1)}{j+1}x^{j+1}(1-x)^{j+1} \\
&= \sum_{i=1}^{\infty}\binom{2(i+j+1)}{i+j+1}x^{i+j+1}(1-x)^{i+j+1} .
\end{aligned}$$

The lemma is proven. ■

Lemma 15. *For any* $j \in \mathbb{N}^{+}$ *and* $0 < x < \frac{1}{2}$

$$\sum_{i=1}^{j}\binom{2i}{i}x^{i}(1-x)^{i} \le \frac{2x\left((2x)^{2j}-1\right)}{2x-1} .$$

Proof: The assertion can be transformed into

$$\begin{aligned}
\sum_{i=1}^{j}\binom{2i}{i}x^{i}(1-x)^{i} \quad &\le \quad \frac{2x\left((2x)^{2j}-1\right)}{2x-1} \\
&= \quad \frac{2x\left(1-(2x)^{2j}\right)}{1-2x} \\
&\le \quad \frac{2x}{1-2x}-\frac{(2x)^{2j+1}}{1-2x} \\
&\le \quad \sum_{i=1}^{\infty}\binom{2i}{i}x^{i}(1-x)^{i}-\frac{(2x)^{2j+1}}{1-2x} ,
\end{aligned}$$

which is equivalent to

$$(2x)^{2j+1} \le (1-2x)\sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j} .$$

We prove this by induction over j . For $j=1$ we have

$$\begin{aligned}
\binom{2}{1}x(1-x) \quad &= \quad 2x-2x^2 \le 2x+4x^2 = \frac{8x^3-2x}{2x-1} \\
&= \quad \frac{2x\left((2x)^2-1\right)}{2x-1} .
\end{aligned}$$

Assume the assertion is true for j . Then

$$\begin{aligned}
&(2x)^{2(j+1)+1} = 4x^2(2x)^{2j+1} \\
&\le 4x^2\left((1-2x)\sum_{i=1}^{\infty}\binom{2(i+j)}{i+j}x^{i+j}(1-x)^{i+j}\right) \\
&\le (1-2x)\sum_{i=1}^{\infty}\binom{2(i+j+1)}{i+j+1}x^{i+j+1}(1-x)^{i+j+1} ,
\end{aligned}$$

where the second line was assumed to be true and the third line follows from Lemma 14. The lemma is proven. ■

References

[1] V. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[2] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[3] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*. Berlin: Springer, 1982.

[4] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–281, 1971.

[5] M. J. Kearns and R. E. Schapire, "Efficient distribution-free learning of probabilistic concepts," *Journal of Computer and System Sciences*, vol. 48, no. 3, pp. 464–497, 1994.

[6] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *Journal of the ACM*, vol. 44, no. 4, pp. 615–631, 1997.

[7] N. Sauer, "On the density of families of sets," *Journal of Combinatorial Theory*, vol. 13, pp. 145–147, 1972.

[8] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data-dependent hierarchies," *IEEE Transactions on Information Theory*, vol. 44, no. 5, pp. 1926–1940, 1998.

[9] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," tech. rep., Royal Holloway, University of London, 1997. NC2-TR-1997-013.

[10] D. McAllester, "Some pac-bayesian theorems," *Machine Learning*, vol. 37, pp. 355–363, 1999.

[11] O. Catoni, "Gibbs estimators," Tech. Rep. LMENS-98-21, École normale supérieure, Département de mathématiques et applications (DMA), 1998.

[12] R. Herbrich and T. Graepel, "A PAC-Bayesian margin bound for linear classifiers: Why SVMs work," in *Advances in Neural Information Processing Systems 13* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), (Cambridge, MA), pp. 224–230, MIT Press, 2001.

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

[14] R. Schapire, Y. Freund, P. L. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Annals of Statistics*, vol. 26, pp. 1651–1686, 1998.

[15] UCI, "University of California Irvine: Machine Learning Repository," 1990.

[16] T. Graepel and R. Herbrich, "The kernel Gibbs sampler," in *Advances in Neural Information Processing Systems 13* (T. K. Leen, T. G. Dietterich, and V. Tresp, eds.), (Cambridge, MA), pp. 514–520, MIT Press, 2001.

[17] R. Herbrich, T. Graepel, and C. Campbell, “Bayes point machines,” *Journal of Machine Learning Research*, vol. 1, pp. 245–279, 2001.

[18] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. No. 31 in Applications of mathematics, New York: Springer, 1996.

[19] M. Vidyasagar, *A Theory of Learning and Generalization*. New York: Springer, 1997.

[20] M. Anthony and P. Bartlett, *A Theory of Learning in Artificial Neural Networks*. Cambridge University Press, 1999.

[21] 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*, vol. 44, no. 2, pp. 525–536, 1998.

[22] J. Shawe-Taylor and N. Cristianini, “Robust bounds on generalization from the margin distribution,” NeuroCOLT Technical Report NC-TR-1998-029, ESPRIT NeuroCOLT2 Working Group, http://www.neurocolt.com, 1998.

[23] B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Kernel-dependent support vector error bounds,” in *Ninth International Conference on Artificial Neural Networks*, Conference Publications No. 470, (London), pp. 103–108, IEE, 1999.

[24] G. P. Bois, *Tables of Indefinite Integrals*. Dover Publications, 1961.

PLACE
PHOTO
HERE

Ralf Herbrich studied computer science at the Technical University of Berlin where he also completed his PhD thesis on learning theory and kernel methods in 2000. He had research stays at University of Bristol, UK, and at the Australian National University in Canberra. Currently, he is holding a research position at Microsoft Research Cambridge, UK, as well as a Research Fellowship at the Darwin College Cambridge. Ralf Herbrich has published in a number of key international journals and conferences. His research interests revolve around the topics of machine learning, theory of generalisation and inference. He is an enthusiastic player of computer games, which he views as a perfect playground for machine learning algorithms.

PLACE
PHOTO
HERE

Thore Graepel studied physics at the University of Hamburg, at Imperial College London, and at the Technical University of Berlin where he also completed his PhD thesis on theory and algorithms for pattern classification with an emphasis on kernel methods and statistical learning theory in 2002. In the course of his doctoral work he enjoyed research stays at RIKEN, Japan, and at the Australian National University in Canberra. Currently, he is holding postdoctoral position at the Institute of Computational Science at the Swiss Federal Institute of Technology (ETH) Zurich. Thore Graepel has published in a number of key international journals and conferences, and is a member of the editorial board of Kluwer’s Machine Learning journal. His research interests revolve around the topics of inference, generalisation, and learning. They include multivariate statistics, statistical learning theory, machine learning algorithms, and optimisation. He is an enthusiastic player of the Japanese board game Go, which he views as a fascinating playground for cognitive models and machine learning algorithms.