

Maximal Width Learning of Binary Functions

Martin Anthony

*Department of Mathematics, London School of Economics, Houghton Street,
London WC2A2AE, U.K.*

Joel Ratsaby

*Electrical and Electronics Engineering Department, Ariel University Center of
Samaria, Ariel 40700, ISRAEL*

Abstract

This paper concerns learning binary-valued functions defined on \mathbb{R} , and investigates how a particular type of ‘regularity’ of hypotheses can be used to obtain better generalization error bounds. We derive error bounds that depend on the *sample width* (a notion analagous to that of sample margin for real-valued functions). This motivates learning algorithms that seek to maximize sample width.

Key words: Binary function classes, Learning algorithms

1 Introduction

1.1 *The idea of width*

It has proven useful, when using the sign of a real-valued function for binary classification, to use functions that achieve a ‘large margin’ on a labeled training sample (since better generalization error bounds are possible, and because such classifiers are also more robust). For general binary-valued functions, not arising in this way from real-valued functions, it is not immediately clear what one could use as an analogy to the margin. This paper investigates how an alternative notion of ‘regularity’ of binary-valued functions with respect to

Email addresses: m.anthony@lse.ac.uk (Martin Anthony),
ratsaby@ariel.ac.il (Joel Ratsaby).

a training sample can analogously be used to guide the selection of a ‘good’ classifier from the class.

The key concept is that of *sample width* of a function. Informally, a function $f : \mathbb{R} \rightarrow \{-1, 1\}$ has a sample width γ with respect to a sample of real numbers, each labeled with 1 or -1 , if γ is the largest number such that for each point x of the sample, we have not only that $f(x)$ matches the label associated with x , but, also, f is constant on an interval of length 2γ centered on each of the sample points. In a sense, then, the function f not only fits the data, but does so in a ‘simple’ or ‘robust’ (or perhaps even ‘convincing’) way. Here, we show how generalization error bounds on such hypotheses can be derived that depend explicitly on the sample width, improving (that is, decreasing) with the sample width.

1.2 Notation

Let the domain be $X = [0, B]$, for a finite $B > 0$. If A is a logical expression that can be evaluated to true or false, then we denote by $\mathbb{I}\{A\}$ the indicator function which takes the value 1 or 0 whenever the statement A is true or false, respectively. We denote by $\langle a, b \rangle$ a generalized interval set of the form $[a, b]$, (a, b) , $[a, b)$ or $(a, b]$. For an interval set R we write $\mathbb{I}_R(x)$ as the indicator function for the statement $x \in R$ or when the set is known explicitly to be $R = \langle a, b \rangle$ then we write $\mathbb{I}\langle a, b \rangle$. For any $a \in \mathbb{R}$, $\text{sgn}(a) = +1$ or -1 if $a > 0$ or $a \leq 0$, respectively. By a *binary function* h on X we mean a function which maps from X to $Y = \{-1, +1\}$. For simplicity, we allow functions h that have only simple discontinuities, i.e., at any point x the limits $h(x^+) \equiv \lim_{z \rightarrow x^+} h(z)$ from the right and similarly from the left $h(x^-)$ exist (but are not necessarily equal). We assume that the set of discontinuities is countable.

For $x \in X$, define the *width* of h on x by

$$\omega_h(x) = h(x) \sup\{a \geq 0 : h(z) = h(x), \text{ for all } z \text{ such that } x - a \leq z \leq x + a\}.$$

Let $Z = X \times Y$. A finite *sample* ζ is an element of Z^m (so it may include repetitions), and m is known as the length of the sample. For a sample $\zeta \in Z^m$, the *sample width* of h , denoted $\omega_\zeta(h)$, is defined as $\min_{(x,y) \in \zeta} y \omega_h(x)$. So, if $\omega_\zeta(h) = \gamma > 0$, then this implies that for each (x, y) in the sample, h is constant on an interval of the form $\langle x - \gamma, x + \gamma \rangle$. This definition of width resembles the notion of *sample margin* of a real-valued function f (see for instance [3]) which is defined as $m_\zeta(f) \equiv \min_{(x,y) \in \zeta} y f(x)$.

Following a form of the PAC model of computational learning theory [5, 8, 12],

we assume that some number, m , of labeled data points (x, b) (where $x \in X$ and $b \in Y$) are generated independently at random according to a fixed probability distribution P on $Z = X \times \{-1, 1\}$ and that we ‘learn’ about P from the sample. (Note that this model includes as a special case the situation in which x is drawn according to a fixed distribution μ on X and the label b is then given by $b = t(x)$ where t is some fixed function from X to Y .)

For a sample $\zeta \in Z^m$, we define the γ -width error (or, simply, γ -error) of a binary function h to be the following quantity:

$$L_\zeta^\gamma(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{y_i \omega_h(x_i) < \gamma\}$$

and we let

$$L(h) = P\{y h(x) < 0\} = P\{h(x) \neq y\}$$

be the probability that h misclassifies a randomly drawn pair $(x, y) \in X \times Y$. This is known as the *generalization error* of h . It is the probability of an error if we use the *hypothesis* h to predict the label y from x , for an element (x, y) of Z drawn according to P .

What we would like to be able to do is to infer that a hypothesis that fits a large randomly-drawn sample well (in the sense that it has small γ width error for a suitably large value of γ on a large P^m -random sample) will in fact have small generalization error (and will therefore have a high probability of correctly predicting the label y associated with x for a P -random $(x, y) \in Z$). The type of result we aim to derive, therefore is one of the following type: for any $\gamma, \delta > 0$ and any probability distribution P , with P^m probability at least $1 - \delta$, a random sample $\zeta \in Z^m$ will be such that for all $h \in H$,

$$L(h) < L_\zeta^\gamma(h) + \epsilon(m, \gamma, \delta),$$

where $\epsilon(m, \gamma, \delta) \rightarrow 0$ as $m \rightarrow \infty$ and where ϵ decreases as γ increases. (The product probability measure P^m is used because the m elements of the sample are generated independently and identically, according to P .)

2 A related problem: learning with γ -regular functions

In this section, we look at a *different* problem which has some resemblance to the main one of this paper, as described above. We do so for two reasons: first, to see what sort of generalization error bound is obtained, so that the one we obtain for the main problem can be compared with it; and, secondly, because it draws on the ‘standard’ VC-theory of learning, which the reader can contrast with the rather different approach used to solve our main problem.

By considering sample width, we regard a binary function as being highly regular, or simple, with respect to a training sample, if it has long constant-value runs centered on the points of the sample. What would be an appropriate *sample-independent* counterpart to this? Perhaps the obvious approach is to regard a binary function as simple if it is piecewise constant, with the smallest ‘piece’ being of at least a certain length. Explicitly, let us say that $h : [0, B] \rightarrow \{-1, 1\}$ is γ -regular if for every $x \in [0, B]$, there is an interval $R = \langle a, a + 2\gamma \rangle$ such that $x \in R$ and h is constant on R (so that $h(z) = h(x)$ for all $z \in R$). (The fact that we take R to be of length 2γ rather than γ is so as to enable easier comparison with the sample-width based results we will obtain.)

A moment’s thought shows that this type of regularity does not imply large sample-width, because for the latter, we require the long constant-value segments to be centered on the sample points, which will fail to be the case if a sample point happens to be near the end-point of one of the intervals R of the type described above. Nonetheless, it does seem to be a comparable sample-independent version of the ‘width at least γ ’ property.

The following result bounds the generalization error of functions $h : X \rightarrow Y$ in terms of their regularity and their error on the sample, which is

$$\frac{1}{m} \sum_{i=1}^m \mathbb{I}\{h(x_i) \neq y_i\}.$$

What it shows, informally speaking, is that if we have a function that agrees well with the values on a random sample and which, moreover, is γ -regular for a large value of γ , then (with high probability) the function has small generalization error.

Theorem 1 *Let $B > 0$ and denote the domain by $X = [0, B]$ with range $Y = \{-1, +1\}$ and let $Z = X \times Y$. Let P be a probability distribution on Z and suppose that $\delta \in (0, 1)$. Then, with P^m -probability at least $1 - \delta$, $\zeta \in Z^m$ is such that for any function $h : X \rightarrow Y$ and for all $\gamma \in (0, B/2]$, if $h : X \rightarrow Y$ is γ -regular, then*

$$P\{h(x) \neq y\} < \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{h(x_i) \neq y_i\} + \epsilon(m, \gamma, \delta),$$

where, defining $k(\gamma)$ by

$$k(\gamma) = \left\lfloor \frac{B}{4\gamma} + \frac{1}{2} \right\rfloor,$$

$\epsilon(m, \gamma, \delta)$ denotes

$$\sqrt{\frac{8}{m} \left(2k(\gamma) \ln \left(\frac{em}{k(\gamma)} \right) + \ln \left(\frac{2^{k(\gamma)+2}}{\delta} \right) \right)}.$$

Proof. We sketch the proof. First, let us fix $\gamma \in (0, B/2]$. It can be seen that the class of γ -regular functions is contained in the set H_γ of all functions on $[0, B]$ that are indicator functions of unions of no more than $k(\gamma) = \lfloor B/(4\gamma) + 1/2 \rfloor$ intervals. Now we can apply some results from the standard theory of learning [3, 5, 8, 13]. Those results, together with bounds on the ‘growth function’ (see [3, 8, 13]) of H_γ , tell us that, for any probability distribution P on $Z = X \times Y$, and any $\delta \in (0, 1)$, we have the following: with P^m -probability at least $1 - \delta$, $\zeta \in Z^m$ is such that for any $h \in H_\gamma$ (and, therefore, for any γ -regular function $h : X \rightarrow Y$),

$$P\{h(x) \neq y\} < \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{h(x_i) \neq y_i\} + \epsilon_0(m, \gamma, \delta),$$

where

$$\epsilon_0(m, \gamma, \delta) = \sqrt{\frac{8}{m} \left(2k(\gamma) \ln \left(\frac{em}{k(\gamma)} \right) + \ln \left(\frac{4}{\delta} \right) \right)}.$$

So far, this requires γ to be fixed in advance. We can modify the result to obtain the required bound of Theorem 1 (that is, a bound that simultaneously applies for all γ) using a technique known as the method of sieves [3, 6, 9]. (See also the last part of Section 3.2 of this paper for more detail on this method.)

□

Given that $k(\gamma)$ is of order B/γ , if we suppress constants and focus on dependence on m , the bound of Theorem 1 states that with probability at least $1 - \delta$, we have

$$P\{h(x) \neq y\} < \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{h(x_i) \neq y_i\} + \epsilon(m, \gamma),$$

where $\epsilon(m, \gamma)$ is of order $\sqrt{\ln(\gamma m)/(\gamma m)}$. In fact, at the expense of larger constants, we can use a result of Talagrand [11] (see also [3]) to improve this to an ϵ that is of order $\sqrt{1/(\gamma m)}$.

3 Bounding generalization error in terms of width error

3.1 The main theorem

The following result bounds the generalization error of hypotheses in terms of their sample width error.

Theorem 2 Let $B > 0$ and denote the domain by $X = [0, B]$ with range $Y = \{-1, +1\}$ and let $Z = X \times Y$. Let P be a probability distribution on Z and suppose that $\delta \in (0, 1)$. Then, with P^m -probability at least $1 - \delta$, $\zeta \in Z^m$ is such that for any function $h : X \rightarrow Y$ and for all $\gamma > 0$,

$$L(h) < L_\zeta^\gamma(h) + \epsilon(m, \gamma, \delta),$$

where

$$\epsilon(m, \gamma, \delta) = \sqrt{\frac{8}{m} \left(\frac{2B}{\gamma} \ln 3 + \ln \left(\frac{32B}{\delta\gamma} \right) \right)}.$$

Note that the theorem makes no assumption on any class of hypotheses nor on its VC-dimension. (The error bound holds simultaneously for any $h : X \rightarrow Y$). Note also that γ is not prescribed in advance.

The ϵ of Theorem 2 is, if we suppress constants and focus on its dependence on m , of order $\sqrt{1/(\gamma m)}$. As we will explain at the end of Section 3.2, many analogous margin-based results for real-valued functions used in classification have an ϵ that includes also $\ln m$ factors and an additional factor related to the ‘fat-shattering dimension’ of the hypothesis space.

As noted, learning with γ -regular functions is a different problem, but it bears some analogy. In section 2 we obtained the high-probability bound

for all $\gamma \in (0, B]$, for all γ -regular h ,

$$P\{h(x) \neq y\} < \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{h(x_i) \neq y_i\} + O\left(\sqrt{1/(\gamma m)}\right), \quad (1)$$

where, here, the O -notation hides constants and δ -dependence. Theorem 2 gives the bound

for all $h : X \rightarrow Y$,

$$P\{h(x) \neq y\} < L_\zeta^\gamma(h) + O\left(\sqrt{1/(\gamma m)}\right). \quad (2)$$

These bounds look similar and, noting that $L_\zeta^\gamma(h) \geq \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{h(x_i) \neq y_i\}$, it might look as if (2) is weaker than (1). As we have noted, however, the two problems to which these bounds relate are different (though they are perhaps analogous). Importantly, it should be observed that (2) is a *sample-based bound* that applies to *any* $h : X \rightarrow Y$ (and not just those that are γ -regular). Even if a function h is not γ -regular, it might still have a large sample-width *on a given sample*, and it is this that potentially makes the sample-width approach useful in practice.

Overview

Any binary function h may be represented by thresholding a real-valued function f , i.e., $h(x) = \text{sgn}(f(x))$. The idea here is to choose a class F of real-valued functions f whose value $f(x)$ is equivalent to the width $\omega_h(x)$ of the corresponding binary functions. Then, the problem of bounding generalization error in terms of width error can be related to the previously-studied problem of bounding (classification) generalization error in terms of margin when real-valued functions are used, through thresholding, for classification. We can then use a margin-based ‘uniform convergence’ result (see [6] and Theorem 10.1 of [3]) to obtain generalization error bounds that depend on the *covering number* of the related class F . The covering numbers of the class F we construct are then bounded to provide a final error bound.

The related class of real functions

For a binary function h on X consider the corresponding set sequence $\{R_i\}_{i=1,2,\dots}$ which satisfies the following properties: (a) $[0, B] = \bigcup_{i=1,2,\dots} R_i$ and for any $i \neq j$, $R_i \cap R_j = \emptyset$, (b) h alternates in sign over consecutive sets R_i, R_{i+1} , (c) R_i is an interval set $\langle a, b \rangle$ with possibly $a = b$ (in which case $R_i = \{a\}$). Hence h has the following general form

$$h(x) = \pm \sum_{i=1,2,\dots} (-1)^i \mathbb{I}_{R_i}(x) \quad (3)$$

There are exactly two functions h corresponding to each sequence of sets R_i , $i = 1, 2, \dots$. Unless explicitly specified, the end points of $X = [0, B]$ are not considered roots of h , i.e., the default behavior is that outside X , the function ‘continues’ with the same value it takes at the endpoint $h(0)$ or $h(B)$, respectively. Now, associate with the set sequence R_1, R_2, \dots the unique non-decreasing sequence of right-endpoints a_1, a_2, \dots which define these sets (the sequence may have at most repetitions, or runs, of length two except for 0 and B) according to

$$R_i = \langle a_i, a_{i+1} \rangle, \quad i = 1, 2, \dots$$

Note that different choices for \langle and \rangle (see earlier definition of a generalized interval $\langle a, b \rangle$) give different sets R_i and hence different functions h . For instance, suppose $X = [0, 7]$ then the following set sequence $R_1 = [0, 2.4)$, $R_2 = [2.4, 3.6)$, $R_3 = [3.6, 3.6] = \{3.6\}$, $R_4 = (3.6, 7]$ has a corresponding end-point sequence $a_1 = 2.4, a_2 = 3.6, a_3 = 3.6, a_4 = 7$. Note that a singleton

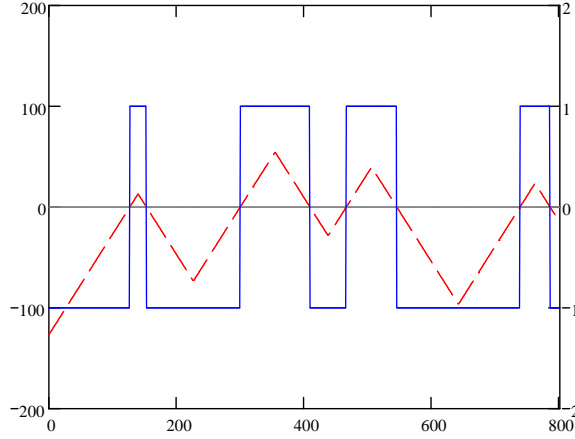


Fig. 1. h (solid with right vertical axis) and its corresponding f (dashed with left vertical axis) on $X = [0, B]$ with $B = 800$

set introduces a repeated value in this sequence. As another example consider $R_1 = [0, 0] = \{0\}$, $R_2 = (0, 4.1)$, $R_3 = [4.1, 7]$ with $a_1 = 0$, $a_2 = 4.1$, $a_3 = 7$.

Next, define the corresponding sequence of midpoints

$$\mu_i = \frac{a_i + a_{i+1}}{2}, \quad i = 1, 2, \dots$$

and the continuous real-valued function $f : X \rightarrow [-B, B]$ corresponding to h as:

$$f(x) = \pm \sum_{i=1,2,\dots} (-1)^{i+1} (x - a_i) \mathbb{I}[\mu_{i-1}, \mu_i] \quad (4)$$

where we take $\mu_0 = 0$.

The connection between γ -width error of binary functions and the ‘margin error’ in the class F real-valued functions we have constructed is crucial. To help describe this link, some additional notation is useful. For a probability distribution P on $X \times Y$, as above, for $f : X \rightarrow \mathbb{R}$, and for $\zeta \in Z^m$ the error of f on ζ at margin γ is defined as

$$\text{er}_\zeta^\gamma(f) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{y_i f(x_i) < \gamma\}.$$

Note that if h has a width $\omega_h(x) = \gamma$ at x , then the corresponding function f satisfies $f(x) = \gamma$. That is, $f(x) = w_h(x)$. Also, for all x , $h(x) = \text{sgn}(f(x))$. It can be seen that, for any ζ ,

$$L_\zeta^\gamma(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{y_i \omega_h(x_i) < \gamma\} = \frac{1}{m} \sum_{i=1}^m \mathbb{I}\{y_i f(x_i) < \gamma\} = \text{er}_\zeta^\gamma(f)$$

and, for any P ,

$$L(h) = P\{yh(x) < 0\} = P\{\text{sgn}(f(x)) \neq y\}.$$

Note, in particular, that the problem of minimizing the γ -width error over all binary functions h on X is equivalent to minimizing the margin error (at margin γ) over this class F of piecewise-linear functions f .

Covering number bounds

We next need to consider *covering numbers*. For $S \subseteq X$, let the $l_\infty(S)$ -norm be defined as $\|f\|_{l_\infty(S)} = \max_{x \in S} |f(x)|$. For $\gamma > 0$, a γ -cover of F with respect to $l_\infty(S)$ is a subset \hat{F} of F with the property that for each $f \in F$ there exists $\hat{f} \in \hat{F}$ such that for all $x \in S$, $|f(x) - \hat{f}(x)| < \gamma$. The *covering number* $\mathcal{N}(F, \gamma, l_\infty(S))$ is the smallest cardinality of a covering for F with respect to $l_\infty(S)$ and the *uniform covering number* $\mathcal{N}_\infty(F, \gamma, m)$ is the maximum of $\mathcal{N}(F, \gamma, l_\infty(S))$, over all S with $S \subset X$ and $|S| = m$.

We shall use Theorem 10.1 of [3] (see also [6]), which is as follows:

Theorem 3 *Suppose that F is a set of real-valued functions defined on X and that P is any probability measure on $Z = X \times \{-1, 1\}$. Then, for any $\epsilon \in (0, 1)$, any $\gamma > 0$ and any positive integer m ,*

$$P^m \left(\left\{ L(\text{sgn}(f)) \geq \text{er}_\zeta^\gamma(f) + \epsilon \text{ for some } f \in F \right\} \right) \leq 2\mathcal{N}_\infty(F, \gamma/2, 2m)e^{-\epsilon^2 m/8}.$$

Given the connection between width error of functions $h : X \rightarrow \{-1, 1\}$ and margin error of corresponding functions in F , this means that, with probability at least $1 - 2\mathcal{N}_\infty(F, \gamma/2, 2m) \exp(-\epsilon^2 m/8)$, for all h , we have

$$L(h) < L_\zeta^\gamma(h) + \epsilon.$$

We now proceed to use this result to obtain useful generalization error bounds by bounding the covering numbers of F and relaxing the assumption that γ be prescribed in advance.

For a finite set $S \subset X$, let us compute the covering number of F with respect to the $l_\infty(S)$ -norm of f . Our approach is to construct and bound the size of a covering with respect to the sup-norm $\|f\|_\infty$ on X which clearly also serves as a covering with respect to $l_\infty(S)$. To do that we construct a finite class \hat{F} of functions as follows: fix γ and denote by $N = \lceil B/\gamma \rceil$. Let

$$\alpha_j = j\gamma, \quad 0 \leq j \leq N \tag{5}$$

and denote by $A = \{\alpha_j : 0 \leq j \leq N\}$. Then we define the finite class \hat{F} as consisting of all functions \hat{f} of the following general form

$$\hat{f}(x) = \pm \sum_{i=1,2,\dots} (-1)^{i+1} (x - \hat{a}_i) \mathbb{I}\langle \hat{\mu}_{i-1}, \hat{\mu}_i \rangle, \quad (6)$$

with

$$\hat{a}_i \in A, \hat{\mu}_0 = 0, \hat{\mu}_i = \frac{\hat{a}_i + \hat{a}_{i+1}}{2}, \quad i = 1, 2, \dots \quad (7)$$

where (similar to the end-point sequence a_i above) the sequence $\hat{a}_i, i = 1, 2, \dots$ is non-decreasing, may repeat up to two consecutive times (except for values of 0 and α_N) and its length does not exceed $2N$. As an extreme example consider the function

$$\hat{h}(x) = \begin{cases} -1 & \text{if } x \in A \\ +1 & \text{otherwise,} \end{cases}$$

whose corresponding \hat{f} has the sequence $\hat{a}_1 = 0, \hat{a}_2 = \alpha_1, \hat{a}_3 = \alpha_1, \hat{a}_4 = \alpha_2, \hat{a}_5 = \alpha_2, \hat{a}_6 = \alpha_3, \dots, \hat{a}_{2N-2} = \alpha_{N-1}, \hat{a}_{2N-1} = \alpha_{N-1}, \hat{a}_{2N} = \alpha_N$.

Next, we proceed to evaluate the approximation ability of \hat{F} . Given an $f \in F$ with its end-point sequence a_i let \hat{a}_i be any sequence (as in (7)) which also satisfies $|a_i - \hat{a}_i| \leq \gamma/2$. Note that while the sequence \hat{a}_i may have $r > 2$ repeated consecutive values $\{\hat{a}_{j+s}\}_{s=0}^{r-1}$ (for instance, due to a cluster of close points $\{a_{j+s}\}_{s=0}^{r-1}$) it is easy to see that the resulting function is equivalent to a function \hat{f} in \hat{F} whose sequence is obtained by replacing this long subsequence with a new subsequence a'_j of length equal to one (with $a'_j = \hat{a}_j$) or two (with $a'_j = a'_{j+1} = \hat{a}_j$) in case r is odd or even, respectively. For convenience, unless otherwise stated, we will use the original sequence \hat{a}_i (without such replacement) as the corresponding sequence of \hat{f} . We denote by μ_i and $\hat{\mu}_i$ the corresponding midpoint sequences, $i = 1, 2, \dots$, of f and \hat{f} .

Consider μ_{i-1}, μ_i and $\hat{\mu}_{i-1}, \hat{\mu}_i$ which must satisfy $\mu_{i-1} \leq \mu_i, \hat{\mu}_{i-1} \leq \hat{\mu}_i$. Denote by $G_i \equiv \{x : \min\{\mu_i, \hat{\mu}_i\} \leq x \leq \max\{\mu_i, \hat{\mu}_i\}\}$. There are two cases: (I) the intervals G_{i-1} and G_i overlap (II) do not overlap. Suppose (II) then denote by $E_i = \{x : \max\{\mu_{i-1}, \hat{\mu}_{i-1}\} \leq x \leq \min\{\mu_i, \hat{\mu}_i\}\}$, $i = 1, 2, \dots$. Over E_i we have $f(x) = (-1)^{i+1}(x - a_i)$ and $\hat{f}(x) = (-1)^{i+1}(x - \hat{a}_i)$ hence

$$\sup_{x \in E_i} |f(x) - \hat{f}(x)| = |a_i - \hat{a}_i| \leq \gamma/2, \quad i = 1, 2, \dots$$

In either case (I) or (II), the worst-case deviation over the interval G_i occurs when either f increases and \hat{f} decreases (at a slope of absolute value 1) or vice

versa. Without loss of generality, suppose $\mu_i \leq \hat{\mu}_i$ so the latter is true. Then we have $\hat{f}(x) = x - \hat{a}_i$ and $f(x) = -(x - a_{i+1})$ so for $x \in G_i$,

$$|\hat{f}(x) - f(x)| = |(x - \hat{a}_i) - -(x - a_{i+1})| \leq |(\hat{\mu}_i - \hat{a}_i) + (\hat{\mu}_i - a_{i+1})|. \quad (8)$$

By (6) at $x = \hat{\mu}_i$ the function \hat{f} changes to $-(x - \hat{a}_{i+1})$ thus the right side of (8) equals

$$|-(\hat{\mu}_i - \hat{a}_{i+1}) + (\hat{\mu}_i - a_{i+1})| = |\hat{a}_{i+1} - a_{i+1}| \leq \gamma/2, \quad i = 1, 2, \dots$$

Combining the above, we have

$$\sup_{x \in X} |f(x) - \hat{f}(x)| = \max_{i=1,2,\dots} \max \left\{ \sup_{x \in E_i} |f(x) - \hat{f}(x)|, \sup_{x \in G_i} |f(x) - \hat{f}(x)| \right\} \leq \gamma/2.$$

Thus the class \hat{F} is a finite $\gamma/2$ -covering of the infinite class F . We proceed now to bound the cardinality of \hat{F} .

From (6), there is a two-to-one correspondence between an $\hat{f} \in \hat{F}$ (and its negation $-\hat{f}$) and the non-decreasing sequence \hat{a}_i , where $\hat{a}_i \in A$, $1 \leq i \leq n$, $1 \leq n \leq 2N$, which may have up to two consecutive repetitions (in case the original sequence \hat{a}_i has a repeated subsequence of length greater than two we henceforth replace it, as mentioned above, by a sequence with repeated runs of length no larger than two). Let b_i , $1 \leq i \leq m-1 \leq n$ be the sequence obtained from \hat{a}_i by removing all duplicates, 0 and α_N (if they appear). Define the sequence of differences as

$$c_i = \begin{cases} b_i/\gamma & i = 1 \\ (b_i - b_{i-1})/\gamma & i = 2, 3, \dots, m-1 \\ N - b_{i-1}/\gamma & i = m \end{cases}$$

which satisfies $\sum_{i=1}^m c_i = N$. For instance, for the sequence $\hat{a}_1 = 0$, $\hat{a}_2 = \hat{a}_3 = \alpha_4$, $\hat{a}_4 = \alpha_{N-3}$ we have $b_1 = \alpha_4$, $b_2 = \alpha_{N-3}$ and $c_1 = 4$, $c_2 = N - 7$, $c_3 = 3$. This sequence c_i , $i = 1, 2, \dots, m$ forms an ordered partition (or composition) of the integer N into m parts. By a classical result (see [2], p.54) the number of such compositions is exactly $\binom{N-1}{m-1}$. Clearly, given any such composition we may construct its corresponding b_i sequence and then have 2^{m-1} possible ways of duplicating any number b_i (this includes the choice of no duplication at all). The resulting sequence can then be modified by either preceding (or not) with a 0 or appending (or not) with an α_N (thus four possibilities) to obtain a valid \hat{a}_i sequence with a corresponding function $\hat{f} \in \hat{F}$. Negating to obtain $-\hat{f}$ also yields a possible function in \hat{F} . Hence there are exactly

$$4 \cdot 2 \cdot \sum_{m=1} \binom{N-1}{m-1} 2^{m-1} = 8 \sum_{k \geq 0} \binom{N-1}{k} 2^k = 8(1+2)^{N-1} = 8 \cdot 3^{N-1}$$

functions $\hat{f} \in \hat{F}$ and hence

$$|\hat{F}| = 8 \cdot 3^{N-1} = 8 \cdot 3^{\lceil B/\gamma \rceil - 1}.$$

To conclude, we therefore have shown that for any subset $S \subset X$ the class F has a covering number

$$\mathcal{N}(F, \gamma/2, l_\infty(S)) \leq \mathcal{N}(F, \gamma/2, l_\infty) \leq |\hat{F}| = 8 \cdot 3^{\lceil B/\gamma \rceil - 1}. \quad (9)$$

This bound therefore gives the following upper bound on the uniform $\gamma/2$ covering numbers (which is independent of m): for all m , for any $\gamma > 0$, $\mathcal{N}_\infty(F, \gamma/2, m) < 8 \cdot 3^{B/\gamma}$.

This bound is almost tight since as we next show a lower bound on it grows at the same rate with respect to B/γ . To obtain the lower bound we use the *fat-shattering dimension*. This is a scale-sensitive version of the pseudo-dimension and was introduced by Kearns and Schapire [10]. Suppose that F is a set of functions from X to \mathbb{R} and that $\gamma \in (0, 1)$. We say that a finite subset $S = \{x_1, x_2, \dots, x_d\}$ of X is γ -shattered if there is $r = (r_1, r_2, \dots, r_d) \in \mathbb{R}^d$ such that for every $b = (b_1, b_2, \dots, b_d) \in \{0, 1\}^d$, there is a function $f_b \in F$ with $f_b(x_i) \geq r_i + \gamma$ if $b_i = 1$ and $f_b(x_i) \leq r_i - \gamma$ if $b_i = 0$. The *fat-shattering dimension*, $\text{fat}_F : \mathbb{R}^+ \rightarrow \mathbb{N} \cup \{0, \infty\}$, is

$$\text{fat}_F(\gamma) = \max \{|S| : S \subseteq X \text{ is } \gamma\text{-shattered by } F\},$$

or $\text{fat}_F(\gamma) = \infty$ if the maximum does not exist.

We can fairly easily lower bound the γ -fat-shattering dimension of our class F . Consider the sample $S_\gamma = \{x_i \equiv \alpha_{2i+1} : 0 \leq i \leq \lfloor N/2 \rfloor - 1\}$ where α_i are defined in (5). The function $f \in F$, whose corresponding sequence $a_{i+1} = \alpha_{2i}$, $0 \leq i \leq \lfloor N/2 \rfloor$, achieves the alternating dichotomy on S_γ , i.e., the corresponding binary function $h(x_i) = (-1)^i$, $0 \leq i \leq |S_\gamma|$ and its margin on S_γ equals γ . It is simple to see that for any other dichotomy $v \in \{-1, +1\}^{|S_\gamma|}$ there exists some $f \in F$ such that its corresponding h satisfies $h(x_i) = v_i$, $0 \leq i \leq |S_\gamma|$ with f having a margin at least γ on S_γ . Hence

$$\text{fat}_F(\gamma) \geq |S_\gamma| = \left\lfloor \frac{N}{2} \right\rfloor = \left\lfloor \frac{1}{2} \left\lceil \frac{B}{\gamma} \right\rceil \right\rfloor \geq \left\lfloor \frac{B}{2\gamma} \right\rfloor$$

It is known (see [7] and Theorem 12.10 of [3]) that for any $m \geq \text{fat}_F(16\epsilon)$, $\mathcal{N}_\infty(F, \gamma, m) \geq \exp(\text{fat}_F(16\gamma)/8)$. Hence we have

$$\mathcal{N}_\infty(F, \gamma/2, m) \geq e^{\text{fat}_F(8\gamma)/8} \geq e^{\lfloor B/16\gamma \rfloor / 8}. \quad (10)$$

From (9) and (10) we see that the log of the covering number is tightly estimated to within a constant multiple of B/γ .

Final steps

By Theorem 3, with probability at least $1 - 2\mathcal{N}_\infty(F, \gamma/2, 2m) \exp(-\epsilon^2 m/8)$, for all $h : X \rightarrow \{-1, 1\}$, we have $L(h) < L_\zeta^\gamma(h) + \epsilon$. Therefore, given the covering number bound we now have the following: for fixed $\gamma > 0$ and for $\delta \in (0, 1)$, with probability at least $1 - \delta$, for every function $h : X \rightarrow \{-1, 1\}$,

$$L(h) < L_\zeta^\gamma(h) + \sqrt{\frac{8}{m} \left(\frac{B}{\gamma} \ln 3 + \ln \left(\frac{16}{\delta} \right) \right)}. \quad (11)$$

The result obtained thus far requires γ to be fixed in advance. What we want instead is a bound that holds simultaneously for all γ . We can achieve this by using the ‘method of sieves’ (see [3, 6, 9]). Note that, since $X = [0, B]$ and by the way the functions in F are defined, we need never consider a width or margin greater than B . For $\gamma_1, \gamma_2 \in (0, B)$ and $\delta \in (0, 1)$, let $E(\gamma_1, \gamma_2, \delta)$ be the subset of Z^m consisting of $\zeta \in Z^m$ for which there exists some $h : X \rightarrow \{-1, 1\}$ with the property that $L(h) > L_\zeta^{\gamma_2}(h) + \epsilon(m, \gamma_1, \delta)$, where

$$\epsilon(m, \gamma, \delta) = \sqrt{\frac{8}{m} \left(\frac{B}{\gamma} \ln 3 + \ln \left(\frac{16}{\delta} \right) \right)}.$$

Then we have that for all γ , $P^m(E(\gamma, \gamma, \delta)) \leq \delta$, for this is simply the bound of (11) above, for fixed γ . Furthermore, if $0 < \gamma_1 \leq \gamma \leq \gamma_2 < 1$ and $0 < \delta_1 \leq \delta \leq 1$, then $E(\gamma_1, \gamma_2, \delta_1) \subseteq E(\gamma, \gamma, \delta)$. This observation enables us to argue, following [6], that

$$\begin{aligned} P^m \left(\bigcup_{\gamma \in (0, B]} E(\gamma/2, \gamma, \delta\gamma/(2B)) \right) &\leq P^m \left(\bigcup_{i=0}^{\infty} \bigcup_{\gamma \in (2^{-(i+1)}B, 2^{-i}B]} E(\gamma/2, \gamma, \delta\gamma/(2B)) \right) \\ &\leq P^m \left(\bigcup_{i=0}^{\infty} E(2^{-(i+1)}B, 2^{-(i+1)}B, \delta 2^{-(i+1)}) \right) \\ &\leq \sum_{i=0}^{\infty} \delta 2^{-(i+1)} = \delta. \end{aligned}$$

So, with probability at least $1 - \delta$, for all $h : X \rightarrow \{-1, 1\}$ and for all $\gamma \in (0, B)$,

$$L(h) < L_\zeta^\gamma(h) + \sqrt{\frac{8}{m} \left(\frac{2B}{\gamma} \ln 3 + \ln \left(\frac{32B}{\delta\gamma} \right) \right)},$$

which is exactly the statement of Theorem 2.

Advantage of directly bounding covering numbers

As noted after the statement of Theorem 2, our $\epsilon(m, \gamma, \delta)$ is of order $\sqrt{1/(\gamma m)}$. We commented earlier that analogous margin-based results for real-valued functions [3] have an ϵ that is larger. Common margin-based approaches to learning make use of Theorem 3 together with the fact that the covering numbers can, by [1], be bounded by the fat-shattering dimension of the class. Explicitly, the following bound (from [3]) is a straightforward corollary of the main result of Alon *et al.*.

Theorem 4 *Suppose that F is a set of functions from X to $[0, B]$ and that F has finite fat-shattering function. Let $m \in \mathbb{N}$ and $\alpha > 0$. Let $d = \text{fat}_F(\alpha/4)$. Then, for all $m \geq d$,*

$$\mathcal{N}_\infty(F, \alpha, m) < 2 \left(\frac{4mB^2}{\alpha^2} \right)^{d \log_2(4eBm/(\alpha\epsilon))}.$$

The following results from now using Theorem 3.

Theorem 5 *Suppose that F is a set of functions from X to $[0, B]$ and that F has finite fat-shattering function. Let $m \in \mathbb{N}$ and $\gamma > 0$. Let $d = \text{fat}_F(\gamma/8)$. Let $\delta \in (0, 1)$. Then, for any probability measure P on $Z = X \times \{-1, 1\}$, with P^m -probability at least $1 - \delta$,*

$$L(\text{sgn}(f)) \geq \text{er}_\zeta^\gamma(f) + \epsilon$$

where

$$\epsilon = \sqrt{\frac{8}{m} \left(d \log_2 \left(\frac{8eBm}{d\gamma} \right) \ln \left(\frac{32mB^2}{\gamma^2} \right) + \ln \left(\frac{4}{\delta} \right) \right)}.$$

In the present context, where F is obtained from H in the way described in Section 3.2, we have seen that d is of order at least $1/\gamma$. This means that the margin-based bounds involve a term of order $\sqrt{(\ln m)^2/(\gamma m)}$. So the bounds obtained here through bounding the covering number directly yield better results than those based on using the fat-shattering dimension.

3.3 A special case

The next result applies to the more specific case where we use a hypothesis that has $L_\zeta^\gamma(h) = 0$. (This is sometimes termed the *restricted model* of learning

[3].)

Theorem 6 *Let $B > 0$ and denote the domain by $X = [0, B]$ with range $Y = \{-1, +1\}$ and let $Z = X \times Y$. Let P be a probability distribution on Z and suppose that $\delta \in (0, 1)$. Then, with P^m -probability at least $1 - \delta$, $\zeta \in Z^m$ is such that for all $\gamma > 0$, for any function $h : X \rightarrow Y$ with the property that $L_\zeta^\gamma(h) = 0$, we have*

$$L(h) < \frac{2}{m} \left(\frac{2B}{\gamma} \ln 3 + \ln \left(\frac{32B}{\delta\gamma} \right) \right).$$

Proof: It follows from a result in [4] (see also [3, 6] for similar results) that, for fixed γ , the P^m -probability that there is $f \in F$ with $\text{er}_\zeta^\gamma(f) = 0$ and $L(\text{sgn}(f)) \geq \epsilon$ is no larger than

$$2\mathcal{N}_\infty(F, \gamma/2, 2m)2^{-\epsilon m/2}.$$

This means that, with probability at least $1 - 2\mathcal{N}_\infty(F, \gamma/2, 2m)2^{-\epsilon m/2}$, for all $h : X \rightarrow Y$ such that $L_\zeta^\gamma(h) = 0$, we have $L(h) < \epsilon$. Given the covering number bound, this means that for fixed $\gamma > 0$ and for $\delta \in (0, 1)$, with probability at least $1 - \delta$, for every function $h : X \rightarrow Y$ which satisfies $L_\zeta^\gamma(h) = 0$, we have

$$L(h) < \frac{2}{m} \left(\frac{B}{\gamma} \ln 3 + \ln \left(\frac{16}{\delta} \right) \right).$$

We turn this into a result that holds simultaneously for all $\gamma \in (0, B]$ using the same technique as in the proof of Theorem 2. The resulting bound is that stated in Theorem 6. \square

Theorem 2 gives the bound

$$\text{for all } h : X \rightarrow Y, \quad P\{h(x) \neq y\} < L_\zeta^\gamma(h) + O\left(\sqrt{1/(\gamma m)}\right).$$

If we simply apply this in the case where $L_\zeta^\gamma(h) = 0$, we obtain a (high-probability) generalization bound of order $\sqrt{1/(\gamma m)}$. Theorem 6 improves this to one of order only $1/(\gamma m)$.

4 Implications for learning algorithms

4.1 Sample width maximization algorithms

The generalization error bound results, Theorems 2 and 6, have some fairly practical implications. Consider, in particular, Theorem 6. The error bound decreases as γ increases; however, as γ increases, the condition that $L_\zeta^\gamma(h) = 0$ becomes more demanding. This suggests using a learning algorithm which will maximize the sample width.

Definition 1 *Given a hypothesis space H (a set of functions from X to Y), we say that a learning algorithm $\mathcal{A} : \bigcup_{m=1}^\infty Z^m \rightarrow H$ is a sample-width maximization algorithm for H if for all m and all $\zeta \in Z^m$, \mathcal{A} returns a hypothesis in H which has zero $\gamma(\zeta)$ -width error on ζ , where*

$$\gamma(\zeta) = \max\{\gamma : \exists h \in H, L_\zeta^\gamma(h) = 0\}.$$

So, a sample-width maximization algorithm for H will give an output hypothesis that agrees with the classifications of the sample points and achieves maximum possible width on the sample of all such functions. (There may be many such hypotheses.) The generalization performance of such an algorithm can be bounded directly by Theorem 6.

Theorem 7 *Suppose that H is the set of binary functions mapping $X = [0, B]$ to $\{-1, 1\}$. Suppose that \mathcal{A} is a sample-width maximization algorithm for H . Given a sample $\zeta \in Z^m$, let $\mathcal{A}(\zeta)$ denote the output hypothesis. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$L(\mathcal{A}(\zeta)) < \frac{2}{m} \left(\frac{2B}{\gamma} \ln 3 + \ln \left(\frac{32B}{\delta\gamma} \right) \right).$$

Note how important it is that, in Theorem 6, the parameter γ is *not* prescribed in advance, because $\gamma(\zeta)$ cannot be known *a priori*.

If there is no particular fixed hypothesis space from which we must choose our hypothesis, then it seems natural, given a labeled sample, to take as hypothesis the simplest $\{-1, 1\}$ -valued function that achieves maximum sample width. That is, we have the following algorithm for learning binary functions on $X = [0, B]$.

Algorithm MW:

Input: A sample $\zeta = \{(x_i, y_i)\}_{i=1}^m$, $x_i \in X$, $y_i \in Y$, $1 \leq i \leq m$, ordered according to $x_1 \leq x_2 \leq \dots \leq x_m$,

- (1) Locate all set-pairs of consecutive points $\{\{x_{i_j}, x_{i_{j+1}}\}\}_{j=1}^\ell$ such that $y_{i_j} \neq y_{i_{j+1}}$, $1 \leq j \leq \ell$. (These set-pairs can have a non-empty intersection).
- (2) Define the corresponding ℓ midpoints as follows:

$$\nu_j = \frac{x_{i_j} + x_{i_{j+1}}}{2}, \quad 1 \leq j \leq \ell$$

- (3) Let h' be defined as follows:

$$h'(x) = \begin{cases} y_{i_1} & \text{if } x \leq \nu_1 \\ y_{i_{j+1}} & \text{if } \nu_j < x \leq \nu_{j+1}, \quad 1 \leq j \leq \ell - 1 \\ y_{i_{\ell+1}} & \text{if } x \geq \nu_\ell \end{cases}$$

Output: h'

It is clear that this is a sample width maximization algorithm. The width $\gamma(\zeta)$ will depend, of course, on the x_i in the sample and on their classifications, but, certainly, we have $\gamma(\zeta) \geq \min_{1 \leq i \neq j \leq m} |x_i - x_j|/2$, the minimum distance between two points in the sample.

4.2 Model selection

A range of ‘model selection’ results for learning with real-valued functions have been obtained, a number of which involve the margin. (See, for instance [3].) In a similar way, the error bounds obtained here can lead to analogous results. The bound of Theorem 2 takes the form

$$L(h) < E(m, \gamma, \delta, h) = L_\zeta^\gamma(h) + \epsilon(m, \gamma, \delta), \quad (12)$$

where, for fixed m and δ , $\epsilon(m, \gamma, \delta)$ decreases as γ increases. A sample-width maximization algorithm will find h such that $L_\zeta^\gamma(h) = 0$ and γ is as large as possible. In general, for any h , and any sample, $L_\zeta^\gamma(h)$ increases as γ increases. Therefore $E(m, \gamma, \delta, h)$ is the sum of two quantities, one of which increases and one of which decreases as γ increases and there is hence a trade-off between the two quantities. This motivates the use of a learning algorithm \mathcal{A} that returns a hypothesis h which minimizes the combination $E(m, \gamma, \delta, h)$. The (high-probability) generalization error bound for such an algorithm take the

form

$$L(\mathcal{A}(\zeta)) \leq \inf_{\gamma} \left(L_{\zeta}^{\gamma}(h) + \sqrt{\frac{8}{m} \left(\frac{2B}{\gamma} \ln 3 + \ln \left(\frac{32B}{\delta\gamma} \right) \right)} \right).$$

5 Conclusions and further work

For learning with binary-valued functions, it is not immediately clear how to use the notion of ‘margin’, which has proven useful in considering learning with real-valued functions. This paper has studied how fairly natural notions of ‘regularity’ of binary-valued functions can be used to bound generalization error, and, in particular, it has shown that a *sample-based* measure of regularity known as the *sample width* can be useful. These results suggest ways in which to guide the selection of a ‘good’ classifier, by selecting those that have high sample width.

This paper only concerns the case in which the domain is an interval on the real line. Clearly, for other domains, there may be other ways of defining notions corresponding to sample ‘width’, and we are currently considering approaches to this.

Acknowledgement

This work was supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

References

- [1] N. Alon, S. Ben-David, N. Cesa-Bianchi and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM* **44**(4): 616–631, 1997.
- [2] G. E. Andrews, *The Theory of Partitions*. Cambridge University Press, 1988.
- [3] Martin Anthony and Peter L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge UK, 1999.
- [4] Martin Anthony and Peter L. Bartlett, Function learning from interpolation. *Combinatorics, Probability and Computing*, **9**, 213–225, 2000.
- [5] Martin Anthony and Norman L. Biggs, *Computational Learning Theory*:

- An Introduction*. Cambridge Tracts in Theoretical Computer Science, 30, 1992. Cambridge University Press, Cambridge, UK.
- [6] Peter Bartlett, The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory* **44**(2): 525–536, 1998.
 - [7] P.L. Bartlett, S.R. Kulkarni and S.E. Posner, Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory* **43**(5): 1721–1724, 1997.
 - [8] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth, Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, **36**(4): 929–965, 1989.
 - [9] U. Grenander, *Abstract Inference*. Wiley, 1981.
 - [10] Michael J. Kearns and Robert E. Schapire, Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, **48**(3): 464–497, 1994.
 - [11] M. Talagrand, Sharper bounds for Gaussian and empirical processes. *Annals of Probability* 22: 28–76, 1994.
 - [12] V. N. Vapnik (1998). *Statistical Learning Theory*, Wiley.
 - [13] 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*, **16**(2): 264–280, 1971.