Open access • Journal Article • DOI:10.1017/S096354830000198X

Valid Generalisation from Approximate Interpolation — Source link

Martin Anthony, Peter L. Bartlett, Yuval Ishai, John Shawe-Taylor

Institutions: London School of Economics and Political Science

Published on: 01 Sep 1996 - Combinatorics, Probability & Computing (Cambridge University Press (CUP))

Valid Generalisation from Approximate Interpolation

Martin Anthony*
Department of Mathematics
The London School of Economics and Political Science
Houghton Street, London WC2A 2AE, United Kingdom
anthony@vax.lse.ac.uk

Peter Bartlett
Department of Systems Engineering
Research School of Information Sciences and Engineering
Australian National University, Canberra, Australia 0200
Peter.Bartlett@anu.edu.au

Yuval Ishai
Department of Computer Science
Technion
Haifa 32000, Israel

John Shawe-Taylor
Computer Science Department
Royal Holloway, University of London
Egham Hill, Egham, Surrey TW20 0EX, United Kingdom
john@dcs.rhbnc.ac.uk

*Part of this research was carried out while Martin Anthony was visiting the Department of Systems Engineering, Australian National University.

Abstract

Let $\mathcal{H}$ and $\mathcal{C}$ be sets of functions from domain X to $\Re$. We say that $\mathcal{H}$ validly generalises $\mathcal{C}$ from approximate interpolation if and only if for each $\eta > 0$ and $\epsilon, \delta \in (0,1)$ there is $m_0(\eta, \epsilon, \delta)$ such that for any function $t \in \mathcal{C}$ and any probability distribution $\mathcal{P}$ on X, if $m \geq m_0$ then with $\mathcal{P}^m$-probability at least $1-\delta$, a sample $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in X^m$ satisfies

$$\forall h \in \mathcal{H},\ |h(x_i) - t(x_i)| < \eta,\ (1 \leq i \leq m) \Longrightarrow \mathcal{P}(\{x : |h(x) - t(x)| \geq \eta\}) < \epsilon.$$

We find conditions that are necessary and sufficient for $\mathcal{H}$ to validly generalise $\mathcal{C}$ from approximate interpolation, and we obtain bounds on the sample length $m_0(\eta, \epsilon, \delta)$ in terms of various parameters describing the expressive power of $\mathcal{H}$.

1 Introduction and Definitions

Much work has recently been carried out on probabilistic models of machine learning such as the 'probably approximately correct' (or *pac*) model due to Valiant [26]. In particular, the pac learning of $\{0,1\}$-valued functions (equivalently, sets) has been studied in great depth; see [12, 5, 18], for example. More recently, attention has been focussed on the extension of the pac model to classes of real-valued functions; see, for example, [14, 1, 9]. The problem studied in this paper is a problem in probability theory which is motivated by, and has applications to, the learnability of real-valued function classes.

1.1 The problem

Suppose we have two sets of functions $\mathcal{H}$, the 'hypothesis space', and $\mathcal{C}$, the 'concept space', from a set X to $\Re$. Normally, we shall assume that $X \subseteq \Re^n$ for some n, but this is not necessary. Suppose also that there is a probability measure $\mathcal{P}$ defined on an appropriate σ-algebra of subsets of the domain X. In the case when $X \subseteq \Re^n$, this σ-algebra is taken to be the Borel σ-algebra. In a particular instance of the generalisation problem, $\mathcal{P}$ is fixed but is unknown to us — who may be thought of as 'the learner'— and there is some *target function* $t \in \mathcal{C}$. The aim is to guarantee that a function from $\mathcal{H}$ which approximates well to the target function on a sample of examples randomly drawn from X according to $\mathcal{P}$, is likely to be a good approximation of the target function on the whole of X. Less informally,

we would like to be sure that if a function from $\mathcal{H}$ closely approximates the target function on the points of the sample, then, with high probability, that function is, in some sense, a good approximation to the target function on the whole domain. Formally, let $\epsilon \in (0,1)$ be an *accuracy* parameter, $\delta \in (0,1)$ a *confidence* parameter, and $\eta \in \Re^+$ a *proximity* parameter. These are prescribed in advance and are part of the 'input' to a particular instance of the generalisation problem. Suppose we draw a sample $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ of length m, with the x_i being chosen independently according to $\mathcal{P}$. Let us say that $h \in \mathcal{H}$ is an *η-approximate interpolant of t on the sample* $\mathbf{x}$ if $|h(x_i) - t(x_i)| < \eta$ for each $1 \leq i \leq m$. (The idea of approximate interpolation occurs in other areas of learning theory; see Sontag [25], for example.) Let us also say that a sample $\mathbf{x} = (x_1, \ldots, x_m)$ is *$(\mathcal{P}, \mathcal{H}, \epsilon, \eta)$-reliable* for t if $h \in \mathcal{H}$ and $|h(x_i) - t(x_i)| < \eta$ for $1 \leq i \leq m$ imply that $\mathcal{P}(\{x \in X : |h(x) - t(x)| \geq \eta\}) < \epsilon$. We say that sample length m is *sufficient for valid (η, ϵ, δ)-generalisation of $\mathcal{C}$ (by $\mathcal{H}$) from η-approximate interpolation* if *for any target $t \in C$* and *for any distribution $\mathcal{P}$ on X* (by which we mean for any probability measure defined on the fixed σ-algebra), with $\mathcal{P}^m$-probability at least $1-\delta$, a sample $\mathbf{x} \in X^m$ is $(\mathcal{P}, \mathcal{H}, \epsilon, \eta)$-reliable for t.

In order to have the appropriate events measurable, some measurability constraints must be imposed on $\mathcal{H}$; we shall not discuss these here, but refer the reader to the appendix of [14] and to [20]. These constraints are mild, and are satisfied by all function classes discussed here. We arrive at the following formal definition.

Definition 1 *Let $\mathcal{C}$ and $\mathcal{H}$ be sets of functions from X to $\Re$. We say that $\mathcal{H}$* validly generalises *$\mathcal{C}$* from approximate interpolation *if for all $\eta > 0$ and $\epsilon, \delta \in (0,1)$, there is $m_0(\eta, \epsilon, \delta)$ such that, for all probability distributions $\mathcal{P}$ on X and all $t \in \mathcal{C}$, if $m \geq m_0(\eta, \epsilon, \delta)$ then with $\mathcal{P}^m$-probability at least $1 - \delta$, a sample $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in X^m$ is $(\mathcal{P}, \mathcal{H}, \epsilon, \eta)$-reliable for t. In other words, with probability at least $1 - \delta$, $\mathbf{x}$ satisfies:*

$$\text{for all } h \in \mathcal{H},\ |h(x_i) - t(x_i)| < \eta,\ (1 \leq i \leq m) \Longrightarrow \mathcal{P}(\{x : |h(x) - t(x)| \geq \eta\}) < \epsilon.$$

Note that the sample length m_0 must be independent of t and $\mathcal{P}$, depending only on η, ϵ and δ; thus the requirement is similar to that of the standard 'probably approximately correct' (pac) learning model [12, 26, 5]. Another noticeable feature of this definition is the requirement that, with high probability, *any* η-approximate interpolant of t on the sample is required to be a good approximation to t. Thus, the notion of valid generalisation from approximate interpolation is a generalisation of what has been called 'solid learnability' by Ben-David *et al.* [10] and 'potential

learnability' by Anthony and Biggs [5] in the context of $\{0,1\}$-valued functions, where *every* consistent function from $\mathcal{H}$ is required to be close to the target function.

We shall assume throughout most of this paper that $\mathcal{H}$ is uniformly bounded, in that there is some bounded subset B of $\Re$ such that all functions in $\mathcal{H}$ map into B. Without loss, we may assume that $B = [0,1]$; the results may be modified easily if B is some other interval, by considering an equivalent problem in which the functions in $\mathcal{C}$ and $\mathcal{H}$ are composed with an affine transformation, and η is transformed appropriately.

Often, when $\mathcal{H}$ validly generalises from approximate interpolation the set $\Re^X$ of all functions from X to $\Re$, we shall say simply that $\mathcal{H}$ *validly generalises from approximate interpolation*. We shall be particularly interested in this case and in the case where $\mathcal{C} = \mathcal{H}$, which we shall refer to as the *restricted* problem. Occasionally, for convenience, we shall omit the words 'validly' and 'approximate'.

1.2 Relevance to function learning

We now briefly discuss the connection between a certain model of function learning and valid generalisation from approximate interpolation. For a function t from X to $\Re$, a positive integer m, and $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in X^m$, let

$$\mathbf{x}(t) = ((x_1, t(x_1)), \ldots, (x_m, t(x_m)))$$

be the *labelled training sample* arising from $\mathbf{x}$ and t. Suppose $\mathcal{C}$ is a set of functions from X to $\Re$ and, for a positive integer m, let

$$S_{\mathcal{C}}(m) = \{\mathbf{x}(t) : \mathbf{x} \in X^m,\ t \in \mathcal{C}\}$$

be the set of all labelled training samples of length m for functions in $\mathcal{C}$. For our purposes, a *learner* is a mapping[1]

$$L : \Re^+ \times \bigcup_{m=1}^{\infty} S_{\mathcal{C}}(m) \to \mathcal{H};$$

L receives as input a parameter η and a labelled training sample for some $t \in \mathcal{C}$, and L outputs some function $h \in \mathcal{H}$. We say that L is a *successful learner for* $\mathcal{C}$

[1] In considering a learner to be a function, we are unconcerned about questions of computability and computational complexity. In practical machine learning, one needs learners which arise from efficient algorithms. Our emphasis here, though, is on what might be termed the 'informational complexity' of learning.

by $\mathcal{H}$ (or that $\mathcal{H}$ learns $\mathcal{C}$ by L) if for all $\eta > 0$ and $\epsilon, \delta \in (0,1)$, there is $m_L(\eta, \epsilon, \delta)$ such that for any $m \geq m_L(\eta, \epsilon, \delta)$, any probability measure $\mathcal{P}$ on X and any $t \in \mathcal{C}$, the following holds: with $\mathcal{P}^m$-probability at least $1 - \delta$ a sample $\mathbf{x}$ is such that $\mathcal{P}(\{x : |h_L(x) - t(x)| \geq \eta\}) < \epsilon$, where $h_L = L(\eta, \mathbf{x}(t))$.

The criterion $\mathcal{P}(\{x \in X : |h(x) - t(x)| \geq \eta\}) < \epsilon$ is similar to the definition of a 'good model of probability' introduced by Kearns and Schapire [16] in their work on p-concepts, defined as functions from X to $[0,1]$. However, the problem they consider is quite different since, in learning a good model of probability of a p-concept as discussed in their work, one is given examples which are labelled 0 or 1 with certain probabilities, rather than examples of the form $(x, t(x))$ for the $[0,1]$-valued target p-concept t.

Let us say that $\mathcal{C}$ *is* $\mathcal{H}$*-approximable* if for any positive integer m, for any $\eta > 0$, for any $t \in \mathcal{C}$, if $\mathbf{x} \in X^m$ then there is $h \in \mathcal{H}$ such that $|h(x_i) - t(x_i)| < \eta$ for $1 \leq i \leq m$. (This is true, in particular, if $\mathcal{C} \subseteq \mathcal{H}$.) If $\mathcal{C}$ is $\mathcal{H}$-approximable, suppose that we have a learner $\mathcal{I}$ with the property that for $\epsilon, \delta \in (0,1)$, $\eta > 0$, and $t \in \mathcal{C}$, if m is a positive integer and $\mathbf{x} \in X^m$, then $\mathcal{I}(\eta, \mathbf{x}(t))$ is an η-approximate interpolant of t on $\mathbf{x}$. The above observations show that if $\mathcal{C}$ is $\mathcal{H}$-approximable *and* $\mathcal{H}$ validly generalises $\mathcal{C}$ from approximate interpolation, then $\mathcal{C}$ can be successfully learned by $\mathcal{H}$ and that any $\mathcal{I}$ as described above is a successful learner. (An important aspect of our definition of valid generalisation from approximate interpolation is that this 'learning result' holds regardless of how $\mathcal{I}$ produces its approximate interpolant.) Thus, in particular, if $\mathcal{C} \subseteq \mathcal{H}$ and $\mathcal{H}$ validly generalises $\mathcal{H}$ from approximate interpolation, then $\mathcal{C}$ can be successfully learned by $\mathcal{H}$. Note that, although the notion of $\mathcal{H}$ generalising from interpolation the set of *all* functions from X to $\Re$ may seem rather strong, it does not translate into a result concerning the learnability of all functions by $\mathcal{H}$, since one also needs approximability.

We remark that, although it is true that if $\mathcal{H}$ validly generalises $\mathcal{H}$ from approximate interpolation, then $\mathcal{H}$ can be successfully learned by $\mathcal{H}$, the converse is false. This is something we shall elaborate on later in the paper. This is in contrast to the corresponding situation in pac learning $\{0,1\}$-valued functions, where both 'solid learnability' and learnability are essentially equivalent (ignoring, as we have throughout, computational issues).

2 Measures of Dimension and the Main Result

In this paper, we derive necessary and sufficient condition for $\mathcal{H}$ to validly generalise $\mathcal{C}$ from approximate interpolation. The cases $\mathcal{C} = \Re^X$ and $\mathcal{C} = \mathcal{H}$ are of particular interest. When $\mathcal{C} = \Re^X$, the class of all functions from X to $\Re$, a particularly succinct characterisation theorem can be given. This is the main result of this paper, which we state in this section. Before doing so, a number of definitions are required.

Although not explicit in the statement of our results, one definition worth giving at this stage is that of the *Vapnik-Chervonenkis dimension* [28, 12]. This combinatorial parameter is central in the pac learning theory of $\{0,1\}$-valued functions and is used in the proofs of the results here. Suppose that $\mathcal{G}$ is a set of $\{0,1\}$-valued functions on X. We say that the finite subset $S = \{x_1, x_2, \ldots, x_d\}$ of X is *shattered* by $\mathcal{G}$ if for every $\mathbf{b} = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $g_{\mathbf{b}} \in \mathcal{G}$ such that

$$g_{\mathbf{b}}(x_i) = b_i.$$

The *VC-dimension* of $\mathcal{G}$, denoted $\mathrm{VCdim}(\mathcal{G})$, is then (infinity, or) the largest cardinality of a shattered set.

The main results of this paper involve two generalisations of the VC-dimension for classes of real-valued functions. One of these—the *pseudo-dimension*—is fairly standard, but the other, the *band dimension*, has only been used rarely [19].

The pseudo-dimension [14, 20]—sometimes called the *combinatorial dimension* or *Pollard dimension*—of a set $\mathcal{H}$ of real-valued functions arises from generalising to real-valued functions the notion of shattering as follows. We say that the finite subset $S = \{x_1, x_2, \ldots, x_d\}$ of X is *shattered* if there is $\mathbf{r} = (r_1, r_2, \ldots, r_d) \in \Re^d$ such that for every $\mathbf{b} = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $h_{\mathbf{b}} \in \mathcal{H}$ with

$$h_{\mathbf{b}}(x_i) \begin{cases} > r_i & \text{if } b_i = 1 \\ < r_i & \text{if } b_i = 0. \end{cases}$$

The *pseudo-dimension* of $\mathcal{H}$, denoted $\mathrm{Pdim}(\mathcal{H})$, is the largest cardinality of a shattered set, or infinity if there is no bound on the cardinalities of the shattered sets. It is clear that if $\mathcal{G}$ is a class of $\{0,1\}$-valued functions, then its pseudo-dimension equals its VC-dimension. The pseudo-dimension is a well-understood and useful measure of expressive power. One attractive feature of this dimension is that if the set of functions is a vector space then its pseudo-dimension coincides with its linear dimension, as shown in [14]. We are mainly concerned in this paper with sets of

functions mapping into a bounded range and hence not with vector spaces of functions, as such, but, in view of this result, if such a function class is a subset of a vector space of dimension d, then its pseudo-dimension is at most d.

The *band-dimension* of a class $\mathcal{H}$ of real-valued functions is a 'scale-sensitive' extension of the VC-dimension. This means that the band-dimension is not simply one number depending on $\mathcal{H}$, but is, rather, a *function* depending on $\mathcal{H}$. (A number of such scale-sensitive dimensions have proven to be useful in learning theory [16, 1, 9, 23, 24].) Let $\mathcal{H}$ be a set of real-valued functions. Given any $\gamma \in \Re^+$, let us say that the finite subset $T = \{(x_1, y_1), (x_2, y_2) \ldots, (x_d, y_d)\}$ of $X \times \Re$ is *γ-band-shattered* by $\mathcal{H}$ if for every $\mathbf{b} = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $h_{\mathbf{b}} \in \mathcal{H}$ with

$$|h_{\mathbf{b}}(x_i) - y_i| \begin{cases} < \gamma & \text{if } b_i = 1 \\ \geq \gamma & \text{if } b_i = 0. \end{cases}$$

The *γ-band-dimension* of $\mathcal{H}$, denoted $\mathrm{Bdim}_{\mathcal{H}}(\gamma)$, is the largest cardinality of a γ-band-shattered set, or infinity if there is no bound on the cardinalities of these sets. The *band-dimension* is the function $\mathrm{Bdim}_{\mathcal{H}}(\gamma)$ of γ, from $\Re^+$ to $\mathbb{N}_0 \cup \{\infty\}$. (Here, $\mathbb{N}_0$ denotes the set of non-negative integers.) If $\mathrm{Bdim}_{\mathcal{H}}(\gamma)$ is finite for all $\gamma > 0$, we say that $\mathrm{Bdim}_{\mathcal{H}}$ *is finite*. It is easy to see that if $\mathcal{G}$ is a class of $\{0,1\}$-valued functions, then for all $\gamma > 0$, $\mathrm{Bdim}_{\mathcal{G}}(\gamma) = \mathrm{VCdim}(\mathcal{G})$. The band-dimension was used in [19, 27, 8].

We prove the following result.

Theorem 2 *Suppose that $\mathcal{H}$ is a set of functions from a set X into $[0,1]$ and that $\Re^X$ is the set of all functions from X into $\Re$. Then, the following are equivalent.*

- *$\mathcal{H}$ validly generalises $\Re^X$ from approximate interpolation.*
- *$\mathrm{Bdim}_H(\gamma)$ is finite, for all $\gamma \in (0,1)$.*
- *$\mathrm{Pdim}(\mathcal{H})$ is finite.*

In proving this theorem, we shall derive a result relating the band-dimension and the pseudo-dimension. Although the pseudo-dimension has been more widely studied, we shall show later in the paper, when providing bounds on the 'sample complexity' function $m_0(\eta, \epsilon, \delta)$, that the band-dimension characterises the sample complexity more precisely than does the pseudo-dimension.

3 Characterising with the Band Dimension

In this section, we derive a necessary and sufficient condition for $\mathcal{H}$ to validly generalise $\mathcal{C}$ from approximate interpolation. We then concentrate attention on the case in which $\mathcal{C}$ is $\Re^X$, the set of all real functions on X. In this case, we obtain bounds on the sample complexity $m_0(\eta, \epsilon, \delta)$ in terms of the band-dimension of $\mathcal{H}$.

We first require some standard results concerning the 'probably approximately correct' (pac) model of generalisation. Suppose that $\mathcal{G}$ is a set of functions with range $\{0,1\}$, defined on a domain X. We say that $\mathcal{G}$ *validly pac-generalises* if for any $\epsilon, \delta \in (0,1)$, there is $m_0 = m_0(\epsilon, \delta)$ such that for any function $t : X \to \{0,1\}$ and any probability measure $\mathcal{P}$ on X, with $\mathcal{P}^m$ probability at least $1-\delta$, a sample $\mathbf{x} \in X^m$ is such that

$$\text{for all } g \in \mathcal{G},\ g(x_i) = t(x_i)\,(1 \le i \le m) \Longrightarrow \mathcal{P}(\{x : g(x) \neq t(x)\}) < \epsilon$$

for $m \ge m_0$. Blumer *et al.* [12], following work of Vapnik and Chervonenkis [28], proved that if $\mathcal{G}$ has finite VC dimension then $\mathcal{G}$ validly pac-generalises. They obtained a bound on a suitable value of $m_0(\epsilon, \delta)$. This was subsequently improved in [6, 22] to show that a suitable value of m_0 is

$$m_0(\epsilon, \delta) = \frac{1}{\epsilon(1-\sqrt{\epsilon})}\left(2\mathrm{VCdim}(\mathcal{H}) \ln\left(\frac{6}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right).$$

The problem of valid generalisation from approximate interpolation can be reduced to the problem of valid pac generalisation of a set of $\{0,1\}$-valued functions, as we now describe. Let $\mathcal{H}$ be a class of functions from X to $[0,1]$ and $\mathcal{C}$ a class of real-valued functions on X. Fix $\eta > 0$ and $t \in \mathcal{C}$ throughout the following discussion. For $h : X \to [0,1]$, define the function $h_{[\eta,t]}$ from X to $\{0,1\}$ by

$$h_{[\eta,t]}(x) = 1 \Longleftrightarrow |h(x) - t(x)| \ge \eta$$

and let $\mathcal{H}_{[\eta,t]} = \left\{h_{[\eta,t]} : h \in \mathcal{H}\right\}$. Note that $t_{[\eta,t]}$ is the identically-0 function. (These definitions implicity use the *loss functions* approach discussed by Haussler [14], where we take the loss function to be $l^\eta : [0,1] \times [0,1] \to \{0,1\}$ defined by $l^\eta(y, y') = 1$ if and only if $|y - y'| \ge \eta$.) Then the *error* of $h_{[\eta,t]}$ with respect to $t_{[\eta,t]}$ is

$$\mathrm{er}_{\mathcal{P}}(h_{[\eta,t]}) = \mathcal{P}\left(\{x \in X : h_{[\eta,t]}(x) \neq t_{[\eta,t]}(x)\}\right) = \mathcal{P}\left(\{x \in X : |h(x) - t(x)| \ge \eta\}\right).$$

Furthermore, the hypothesis $h_{[\eta,t]}$ is consistent with $t_{[\eta,t]}$ on a sample $(x_1, x_2, \ldots, x_m)$ if and only if $h_{[\eta,t]}(x_i) = 0$ for $1 \le i \le m$; that is, if and only if $|h(x_i) - t(x_i)| < \eta$ for $1 \le i \le m$.

Theorem 3 *Let $\mathcal{H}$ be a set of functions from X to $[0,1]$ and $\mathcal{C}$ a set of real functions on X. Then $\mathcal{H}$ validly generalises $\mathcal{C}$ from approximate interpolation if and only if, for all $\eta > 0$, the set*

$$\left\{\mathrm{VCdim}(\mathcal{H}_{[\eta,t]}) : t \in \mathcal{C}\right\}$$

is a bounded set of integers. When this is so, then, with

$$\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta) = \max_{t\in\mathcal{C}} \mathrm{VCdim}(\mathcal{H}_{[\eta,t]}),$$

a suitable sample length function $m_0(\eta,\epsilon,\delta)$ is

$$m_0(\eta,\epsilon,\delta) = \frac{1}{\epsilon(1-\sqrt{\epsilon})}\left(2\,\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta)\ln\left(\frac{6}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right).$$

Furthermore, for $0 < \delta \leq 1/6$ and η satisfying $\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta) \geq 2$, any sample length function must satisfy

$$m_0(\eta,\epsilon,\delta) > \max\left(\frac{1-\epsilon}{\epsilon}\log\frac{1}{\delta}, \frac{\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta)-1}{12\epsilon}\right).$$

Proof: Suppose first that the set of VC-dimensions described is a bounded set of integers, and let $\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta)$ be as in the statement of the theorem. Then, for each $t \in \mathcal{C}$, by the standard results on the basic pac-generalisation model, provided

$$m \geq \frac{1}{\epsilon(1-\sqrt{\epsilon})}\left[2\,\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta)\ln(6/\epsilon) + \ln(2/\delta)\right],$$

for any distribution $\mathcal{P}$ on X, and any $t \in \mathcal{C}$,

$$\mathcal{P}^m\Big(\Big\{\mathbf{x}\in X^m : \exists h\in\mathcal{H} \text{ with } h_{[\eta,t]}(x_i) = t_{[\eta,t]}(x_i)\,(1\leq i\leq m) \text{ and } \mathrm{er}_{\mathcal{P}}(h_{[\eta,t]}) \geq \epsilon\Big\}\Big) < \delta.$$

But this means that for $m \geq m_0(\eta,\epsilon,\delta)$, for any probability distribution $\mathcal{P}$ on X and any $t \in \mathcal{C}$, with $\mathcal{P}^m$-probability at least $1-\delta$, a sample $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in X^m$ satisfies:

$$\text{for all } h\in\mathcal{H},\ |h(x_i)-t(x_i)| < \eta,\ (1\leq i\leq m) \Longrightarrow \mathcal{P}(\{x : |h(x)-t(x)| \geq \eta\}) < \epsilon.$$

In other words, $\mathcal{H}$ validly generalises $\mathcal{C}$ from approximate interpolation, with m_0 as a suitable sample length function.

Conversely, fix η and suppose there is a function $t \in \mathcal{C}$ such that $\mathrm{VCdim}(\mathcal{H}_{[\eta,t]}) \geq d$ for some $d \geq 2$. Fix ϵ and δ. We shall use an argument similar to Blumer,

Ehrenfeucht, Haussler, and Warmuth's proof of Theorem 2.1 in [12] to prove the first term in the maximum. Because $d \geq 2$, there is a set $\{a, b\} \subseteq X$ and a function $f \in \mathcal{H}$ such that $f_{[\eta,t]}(a) = 1$ but $f_{[\eta,t]}(b) = 0$. Let $\mathcal{P}$ be the probability distribution on $\{a, b\}$ with $\mathcal{P}(\{a\}) = \epsilon$, $\mathcal{P}(\{b\}) = 1-\epsilon$. Suppose the sample $\mathbf{x} = (b, \ldots, b) \in X^m$ is drawn. Clearly, f η-approximately interpolates t on this sample, since $|f(b) - t(b)| < \eta$, but $\mathcal{P}(\{x : |f(x) - t(x)| \geq \eta\}) = \mathcal{P}(\{a\}) = \epsilon$. So with $\mathcal{P}^m$-probability at least $(1-\epsilon)^m$, a sample $x \in X^m$ is not $(\mathcal{P}, \mathcal{H}, \epsilon, \eta)$-reliable for t. This probability is at least δ for

$$m \leq \frac{1-\epsilon}{\epsilon} \log \frac{1}{\delta}.$$

To prove the second term in the maximum, we use an argument similar to one used in [13]. Let $X_0 = \{y_0, y_1, \ldots, y_k\} \subseteq X$ be shattered by $\mathcal{H}_{[\eta,t]}$, where $k = d - 1$. Choose a set $F \subseteq \mathcal{H}_{[\eta,t]}$ such that $|F| = 2^d$ and F shatters X_0. Let $\mathcal{P}$ be the probability distribution on X defined by

$$\mathcal{P}(\{x\}) = \begin{cases} 1 - 2\epsilon & \text{if } x = y_0 \\ 2\epsilon/k & \text{if } x \in \{y_1, \ldots, y_k\} \\ 0 & \text{otherwise.} \end{cases}$$

Let $Q \subseteq X_0^m$ consist of those sequences of length m which contain no more than $k/2$ elements of the set $\{y_1, \ldots, y_k\}$. Then for any sample $\mathbf{x} = (x_1, \ldots, x_m)$ in Q there is a function $h_{[\eta,t]}$ in F such that $|h(x_i) - t(x_i)| < \eta$ for $i = 1, \ldots, m$, but h satisfies

$$|\{i \in \{1, \ldots, m\} : |h(y_i) - t(y_i)| \geq \eta\}| \geq \frac{k}{2},$$

so $\mathcal{P}(\{x : |h(x) - t(x)| \geq \eta\}) \geq \epsilon$. That is, with probability at least $\mathcal{P}^m(Q)$, a sample is not $(\mathcal{P}, \mathcal{H}, \epsilon, \eta)$-reliable for t.

Now, the probability of drawing a sample of length m that is not in Q is no more than the probability of $k/2$ successes in a sequence of m Bernoulli trials, where the probability of success at each trial is 2ϵ. From standard Chernoff bounds on the tails of the binomial distribution (see [3]), this probability is no more than

$$\exp\left(-\frac{2m\epsilon}{3}\left(\frac{k}{4m\epsilon} - 1\right)^2\right)$$

and for $0 < \delta \leq 1/6$ and $k \geq 1$, this is no more than $1 - \delta$ when $m \leq k/(12\epsilon)$. □

In what follows, it will be convenient to define $\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta)$ to be infinite if the set $\{\mathrm{VCdim}(\mathcal{H}_{[\eta,t]}) : t \in \mathcal{C}\}$ is unbounded or if one of these VC-dimensions is infinite. Then Theorem 3 provides a necessary and sufficient condition for the general

problem of $\mathcal{H}$ validly generalising a class $\mathcal{C}$ from approximate interpolation, namely $\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta) < \infty$ for all $\eta > 0$. This is, however, a rather cumbersome condition. We now show that if $\mathcal{C} = \Re^X$, then $\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta)$ is closely related to the 'simpler' band-dimension. Indeed, we have the following result.

Proposition 4 *Suppose that $\mathcal{H}$ is a set of $[0,1]$-valued functions and that $\eta > 0$. Then, if $\mathcal{C} = \Re^X$,*

$$\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta) \leq \mathrm{Bdim}_{\mathcal{H}}(\eta) \leq 2\mathrm{d}^*_{\mathcal{H},\mathcal{C}}(\eta).$$

Proof: Assume that both dimensions are finite and let $\eta > 0$. Suppose first that $t : X \to \Re$ and that the set $S = \{x_1, x_2, \ldots, x_d\}$ is shattered by $\mathcal{H}_{[\eta,t]}$. Then, if $t_i = t(x_i)$, it is clear that the subset $\{(x_1, t_1), (x_2, t_2), \ldots, (x_d, t_d)\}$ is η-band-shattered by $\mathcal{H}$. This proves the first inequality. Now suppose that the subset

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_d, y_d)\}$$

of $X \times \Re$ is η-band-shattered by $\mathcal{H}$. It is possible to have $x_i = x_j$ if $i \neq j$. However, it is easy to see that no X-coordinate may be repeated three times in T. It follows that there is a subset T' of T, of cardinality at least $d/2$ such that the X-coordinates of the points in T' are distinct. The set of X-coordinates of the points in T' is then shattered by $\mathcal{H}_{[\eta,t]}$, where $t : X \to \Re$ is any function such that if $(x_i, y_i) \in T'$, then $t(x_i) = y_i$. This proves the second inequality. □

Combining this proposition and Theorem 3 gives the following result, which provides a simpler characterisation of valid generalisation from approximate interpolation.

Theorem 5 *Let $\mathcal{H}$ be a set of functions from X to $[0,1]$. Then $\mathcal{H}$ validly generalises from approximate interpolation if and only if $\mathrm{Bdim}_{\mathcal{H}}(\eta)$ is finite for all $\eta > 0$. When $\mathrm{Bdim}_{\mathcal{H}}$ is finite, a suitable sample length function $m_0(\eta, \epsilon, \delta)$ is*

$$\frac{1}{\epsilon(1-\sqrt{\epsilon})}\left(2\,\mathrm{Bdim}_{\mathcal{H}}(\eta)\ln\left(\frac{6}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right).$$

Furthermore, any sample length function must satisfy

$$m_0(\eta, \epsilon, \delta) \geq \max\left(\frac{1-\epsilon}{\epsilon}\log\frac{1}{\delta}, \frac{\mathrm{Bdim}_{\mathcal{H}}(\eta) - 2}{24\epsilon}\right)$$

when $\delta \leq 1/6$ and $\mathrm{Bdim}_{\mathcal{H}}(\eta) \geq 4$.

In some work on function learning, such as [17, 7, 11], a dimension known as the *graph dimension* has proven to be useful. The graph dimension of a class $\mathcal{H}$ of functions that map from X to a set Y is the VC-dimension of the class

$$\left\{(x,y) \mapsto \begin{cases} 1 & \text{if } y = h(x) \\ 0 & \text{otherwise} \end{cases} : h \in \mathcal{H}\right\}.$$

It appears that this dimension is more useful for functions taking values in a finite set, rather than in the reals, and there is some further evidence of this here. For, although it might seem that the band-dimension is a 'scale-sensitive' version of the graph dimension, the two are in fact unrelated, as the following example shows. For each positive integer i, define a function h_i from $\mathbb{N}$ to $[0,1]$ by

$$h_i(n) = \begin{cases} 1/2 + 1/(i+1) & \text{if } \mathrm{bit}_n(i) = 1; \\ 1/2 - 1/(i+1) & \text{otherwise,} \end{cases}$$

where $\mathrm{bit}_n(i)$ is the nth bit from the right in the binary representation of i. The class $\mathcal{H} = \{h_i : i \in \mathbb{N}\}$ has graph dimension 1 since no two functions of $\mathcal{H}$ agree at any point of $\mathbb{N}$. However, for any $\eta > 0$, $\mathcal{H}$ has infinite η-band-dimension.

4 Relationships Between Dimensions

In this section, we show that the pseudo-dimension $\mathrm{Pdim}(H)$ and the band dimension $\mathrm{Bdim}_H(\eta)$ are within a factor of $\log \frac{1}{\eta}$ of each other. The proofs involve several notions of dimension of discretised versions of the function class $\mathcal{H}$, and provide a characterisation of those dimensions whose finiteness is necessary and sufficient for generalisation from approximate interpolation.

The following definitions are from [11]. Let $\mathcal{F}$ be a class of functions defined on X that take values in a finite set S with $|S| = n$. Let Ψ be a class of $\{0,1,*\}$-valued functions defined on S. We say that F Ψ-shatters a sequence $\mathbf{x} = (x_1, \ldots, x_d) \in X^d$ if there is a sequence $\psi = (\psi_1, \ldots, \psi_d) \in \Psi^d$ satisfying

$$\{0,1\}^d \subseteq \{(\psi_1(f(x_1)), \ldots, \psi_d(f(x_d))) : f \in \mathcal{F}\}.$$

The Ψ-dimension of $\mathcal{F}$ is

$$\Psi\text{-dim}(F) = \max\left\{d : \exists \mathbf{x} \in X^d,\ \mathcal{F} \text{ } \Psi\text{-shatters } \mathbf{x}\right\}.$$

Two important examples of dimensions defined in this way are the Ψ_B-dimension and the Ψ_{Nat}-dimension, where $\Psi_B = \{0,1\}^S$ and $\Psi_{\mathrm{Nat}} = \{\psi_{a,b} : a, b \in S, a < b\}$

with

$$\psi_{a,b}(y) = \begin{cases} 0 & \text{if } y = a \\ 1 & \text{if } y = b \\ * & \text{otherwise.} \end{cases}$$

We say that a class Ψ is a distinguisher if, for all distinct $y_1, y_2 \in S$, there is a ψ in Ψ and a $b \in \{0,1\}$ for which $\psi(y_1) = b$ and $\psi(y_2) = 1 - b$. Notice that $\Psi_{\rm Nat}$ and Ψ_B are distinguishers. Ben-David, Cesa-Bianchi, Haussler, and Long show in [11] that if Ψ is a distinguisher then the Ψ-dimension is closely related to the $\Psi_{\rm Nat}$- and Ψ_B-dimensions.

Theorem 6 ([11]) *Suppose S is a set of cardinality $n \in \mathbb{N}$, $\mathcal{F}$ is a class of S-valued functions defined on X, and Ψ is a class of $\{0,1,*\}$-valued functions defined on S. If Ψ is a distinguisher, we have*

$$\Psi_{\rm Nat}\text{-dim}(\mathcal{F}) \leq \Psi\text{-dim}(\mathcal{F}) \leq \Psi_B\text{-dim}(\mathcal{F}) \leq 4.67 \log_2 n \;\; \Psi_{\rm Nat}\text{-dim}(\mathcal{F}).$$

We use this result to prove the following theorem. Here, as elsewhere in the paper, no serious attempt has been made to optimise the constants.

Theorem 7 *If $\mathcal{H}$ is a set of functions that map from a set X to $[0,1]$, then*

$$\mathrm{d}^*_{\mathcal{H},\Re^X}(\eta) < 7.5 \mathrm{Pdim}(\mathcal{H}),$$

for all $\eta > 0$.

Notice that if $\mathcal{H}$ is a set of $\{0,1\}$-valued functions, then $\mathrm{Bdim}_{\mathcal{H}}(\eta) = \mathrm{d}^*_{\mathcal{H},\Re^X}(\eta) = \mathrm{Pdim}(\mathcal{H}) = \mathrm{VCdim}(\mathcal{H})$ for all $\eta > 0$. It follows that Theorem 7 cannot be improved by more than a constant factor.

Proof: If $t \in \Re^X$ and $\eta > 0$, let $h'_{[\eta,t]} : X \to \{0,1,2\}$ be defined by

$$h'_{[\eta,t]}(x) = \begin{cases} 0 & \text{if } \frac{h(x)-t(x)}{2\eta} \leq 0 \\ 1 & \text{if } 0 < \frac{h(x)-t(x)}{2\eta} < 1 \\ 2 & \text{if } \frac{h(x)-t(x)}{2\eta} \geq 1 \end{cases}$$

Let $\mathcal{H}'_{[\eta,t]} = \left\{h'_{[\eta,t]} : h \in \mathcal{H}\right\}$. Define $\Psi_{\rm Nat}, \Psi_B : \{0,1,2\} \to \{0,1,*\}$ as above. Let $\Psi_G = \{\psi_G\}$, where

$$\psi_G(z) = \begin{cases} 1 & \text{if } z = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Clearly,

$$\mathrm{d}^*_{\mathcal{H},\Re^X}(\eta) = \max_{t\in\Re^X} \Psi_G\text{-dim}\left(\mathcal{H}'_{[\eta,t]}\right).$$

Furthermore, since all functions in Ψ_G are $\{0,1\}$-valued, Ψ_G is a subset of Ψ_B so $\Psi_G\text{-dim}(\mathcal{H}'_{[\eta,t]}) \leq \Psi_B\text{-dim}(\mathcal{H}'_{[\eta,t]})$. Let $\Psi_P = \{\psi_1, \psi_2\}$, where

$$\psi_1(z) = \begin{cases} 0 & \text{if } z = 0 \\ 1 & \text{otherwise} \end{cases}$$

and

$$\psi_2(z) = \begin{cases} 0 & \text{if } z = 2 \\ 1 & \text{otherwise.} \end{cases}$$

Clearly, for all $\eta > 0$

$$\mathrm{Pdim}(\mathcal{H}) = \max_{t\in\Re^X} \Psi_P\text{-dim}\left(\mathcal{H}'_{[\eta,t]}\right),$$

and Ψ_P is a distinguisher. So Theorem 6 implies $\Psi_P\text{-dim}(\mathcal{H}'_{[\eta,t]}) \geq \Psi_{\mathrm{Nat}}\text{-dim}(\mathcal{H}'_{[\eta,t]})$ and

$$\Psi_G\text{-dim}\left(\mathcal{H}'_{[\eta,t]}\right) < (4.67\log_2 3)\,\Psi_P\text{-dim}\left(\mathcal{H}'_{[\eta,t]}\right).$$

□

To show a converse relationship between $\mathrm{d}^*_{\mathcal{H},\Re^X}(\eta)$ and $\mathrm{Pdim}(\mathcal{H})$, we consider a more general discretisation.

Definition 8 *Suppose $t \in \Re^X$, $h \in \mathcal{H}$, and $\eta > 0$. Let $S = \{i/2 : i \in \mathbf{Z}\}$. Let the function $h''_{[\eta,t]} : X \to S$ be defined by*

$$h''_{[\eta,t]}(x) = \phi\left(\frac{h(x) - t(x)}{2\eta}\right),$$

where

$$\phi(\alpha) = \begin{cases} \alpha + \frac{1}{2} & \text{if } \alpha \in \mathbf{Z} \\ \lceil\alpha\rceil & \text{otherwise.} \end{cases}$$

Let $\mathcal{H}''_{[\eta,t]} = \{h''_{[\eta,t]} : h \in \mathcal{H}\}$.

The graph of the function ϕ is illustrated in Figure 1. As above, we can define various dimensions of $\mathcal{H}''_{[\eta,t]}$ using classes of $\{0,1,*\}$-valued functions. Because, for any fixed t and $\mathbf{x}$, the functions in $\mathcal{H}_{[\eta,t]}$ map $\mathbf{x}$ to a bounded subset of S, we need consider only certain $\{0,1,*\}$-valued function classes.

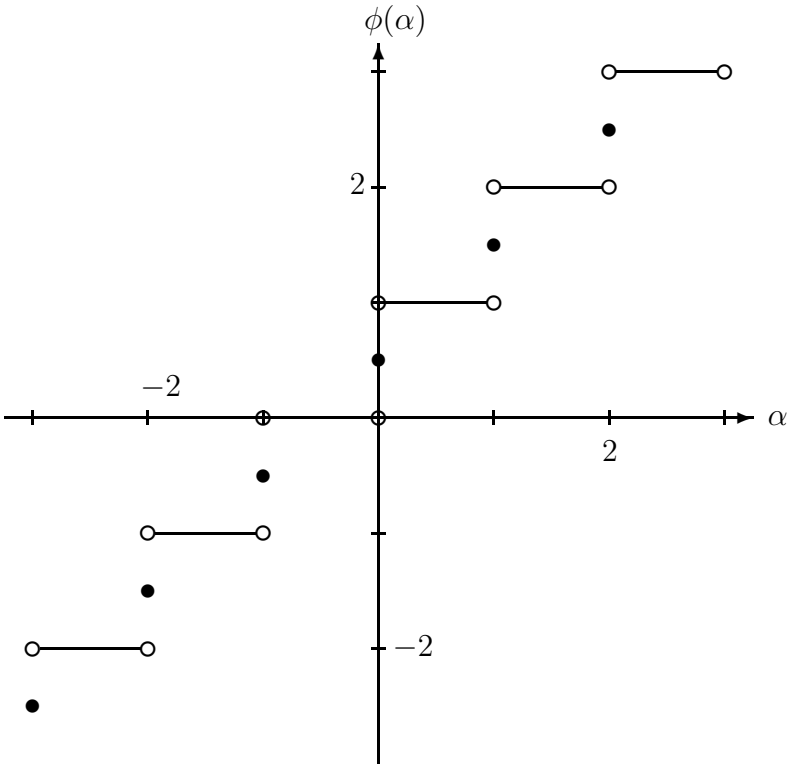

Figure 1: The definition of the discretisation $\mathcal{H}''_{[\eta,t]}$ uses the function ϕ.

Definition 9 *Let $S = \{i/2 : i \in \mathbf{Z}\}$ and suppose that Ψ is a sequence of sets $\Psi = \langle \Psi_1, \Psi_2, \Psi_3, \ldots \rangle$, where each Ψ_i is a set of functions from S to $\{0, 1, *\}$. For such a sequence, let*

$$\Psi\text{-dim}_{\mathcal{H}}(\eta) = \max_{t \in \Re^X} \Psi_n\text{-dim}(\mathcal{H}''_{[\eta,t]}),$$

where $n = \left\lceil \frac{1}{2\eta} \right\rceil$.

For $i \in \mathbb{N}$, let $S_i = \{-1/2, 0, 1/2, 1, \ldots, i - 1/2, i\}$. We say that the sequence Ψ is admissible if, for all $i \in \mathbb{N}$, for all $\psi \in \Psi_i$, and for all $y \in S - S_i$, $\Psi(y) = *$.

The following result shows that we can assume a sequence Ψ is admissible without loss of generality.

Proposition 10 *Let $S = \{i/2 : i \in \mathbf{Z}\}$. Let Ψ be a sequence of sets of functions from S to $\{0, 1, *\}$. Then there is an admissible sequence $\tilde{\Psi}$ satisfying*

$$\tilde{\Psi}\text{-dim}_{\mathcal{H}}(\eta) = \Psi\text{-dim}_{\mathcal{H}}(\eta)$$

for all $\eta > 0$ and all classes $\mathcal{H}$ of $[0,1]$-valued functions.

Proof: Let $\Psi = \langle \Psi_1, \Psi_2, \ldots \rangle$. We will show that the sequence $\tilde{\Psi} = \langle \tilde{\Psi}_1, \tilde{\Psi}_2, \ldots \rangle$ will suffice, where

$$\tilde{\Psi}_n = \left\{ y \mapsto \begin{cases} \psi(y-m) & \text{if } y - m \in S_n \\ * & \text{otherwise} \end{cases} : m \in \mathbf{Z}, \psi \in \Psi_n \right\}$$

(Recall that $S_n = \{-1/2, 0, 1/2, 1, \ldots, n-1/2, n\}$.)

Fix η, $n = \lceil 1/(2\eta) \rceil$, and $\mathcal{H}$. Suppose there is a function $t : X \to \Re$ and $\mathbf{x} = (x_1, \ldots, x_d) \in X^d$ such that $\mathcal{H}''_{[\eta,t]}$ Ψ_n-shatters $\mathbf{x}$. Then there are functions $\psi_1, \ldots, \psi_d$ in Ψ_n such that

$$\left\{ (\psi_1(f(x_1)), \ldots, \psi_d(f(x_d))) : f \in \mathcal{H}''_{[\eta,t]} \right\}$$

contains $\{0,1\}^d$. Now define

$$\tilde{\psi}_i : y \mapsto \psi_i(y + \lceil t(x_i)/(2\eta) \rceil - 1).$$

Let $\tilde{t}(x_i) = t(x_i) - 2\eta \lceil t(x_i)/(2\eta) \rceil + 2\eta$. Then

$$\psi_i \left(\phi \left(\frac{h(x_i) - t(x_i)}{2\eta} \right) \right) = \tilde{\psi}_i \left(\phi \left(\frac{h(x_i) - \tilde{t}(x_i)}{2\eta} \right) \right).$$

Furthermore, the argument α of $\tilde{\psi}_i$ satisfies

$$\begin{aligned} & \phi\left(\frac{-t(x_i)}{2\eta}\right) \leq \ \alpha \ \leq \phi\left(\frac{1-t(x_i)}{2\eta}\right) \\ \iff \quad & \phi\left(\left\lceil \frac{t(x_i)}{2\eta} \right\rceil - \frac{t(x_i)}{2\eta} - 1\right) \leq \ \alpha \ \leq \phi\left(\frac{1}{2\eta} + \left\lceil \frac{t(x_i)}{2\eta} \right\rceil - \frac{t(x_i)}{2\eta} - 1\right). \end{aligned}$$

If $t(x_i)/(2\eta) \in \mathbf{Z}$, then $-1/2 \leq \alpha \leq \lceil 1/(2\eta) \rceil - 1$. Otherwise $0 \leq \alpha \leq \lceil 1/(2\eta) \rceil$. In either case, $\alpha \in S_n$. It follows that $\mathcal{H}''_{[\eta,\tilde{t}]}$ $\tilde{\Psi}_n$-shatters $\mathbf{x}$, so

$$\max_{t \in \Re^X} \tilde{\Psi}_n\text{-dim}\left(\mathcal{H}''_{[\eta,t]}\right) \geq \max_{t \in \Re^X} \Psi_n\text{-dim}\left(\mathcal{H}''_{[\eta,t]}\right).$$

A similar argument gives the reverse inequality. □

We are interested here in sequences of admissible $\{0,1,*\}$-valued function classes that can distinguish intervals in the following sense.

Definition 11 *Let $\Psi = \langle \Psi_i : i \in \mathbb{N} \rangle$ be a sequence of $\{0, 1, *\}$-valued function classes defined on the set $S = \{i/2 : i \in \mathbf{Z}\}$. We say that Ψ is an* interval distinguisher *if it is admissible and, for all $n \in \mathbb{N}$ and all Δ in $\{1, \ldots, n\}$, there is an m in $\{0, 1, \ldots, n - \Delta\}$ such that, for some $\psi \in \Psi_n$ and $b \in \{0, 1\}$, $\psi(m) = b$ and $\psi(m + \Delta) = 1 - b$.*

We can define two admissible sequences based on the function classes Ψ_{Nat} and Ψ_B defined above. Let the sequence $\Psi_{\text{Nat}} = \langle \Psi_{\text{Nat},n} : n \in \mathbb{N} \rangle$ be defined by $\Psi_{\text{Nat},n} = \{\psi_{a,b} : a, b \in S_n, a < b\}$ with

$$\psi_{a,b}(y) = \begin{cases} 0 & \text{if } y = a \\ 1 & \text{if } y = b \\ * & \text{otherwise.} \end{cases}$$

Let the sequence $\Psi_B = \langle \Psi_{B,n} : n \in \mathbb{N} \rangle$ be defined by

$$\Psi_{B,n} = \left\{ \psi \in \{0, 1, *\}^S : \forall a \in S_n,\ \psi(a) \in \{0, 1\} \text{ and } \forall a \in S - S_n,\ \psi(a) = * \right\}.$$

Obviously, Ψ_{Nat} and Ψ_B are interval distinguishers.

The following theorem relates Ψ_{Nat}-dim$_{\mathcal{H}}$, Ψ_B-dim$_{\mathcal{H}}$, and Ψ-dim$_{\mathcal{H}}$, for any interval distinguisher Ψ. It is analogous to Theorem 6.

Theorem 12 *Suppose Ψ is an interval distinguisher, $\mathcal{H}$ is a set of functions from some set X to $[0, 1]$, and $\eta > 0$. Then*

$$\Psi_{\text{Nat}}\text{-dim}_{\mathcal{H}}(\eta) \leq \Psi\text{-dim}_{\mathcal{H}}(\eta) \leq \Psi_B\text{-dim}_{\mathcal{H}}(\eta) \leq 4.67 \log_2 \left(2 \left\lceil \frac{1}{2\eta} \right\rceil + 2 \right) \Psi_{\text{Nat}}\text{-dim}_{\mathcal{H}}(\eta).$$

Proof: Fix η and let $n = \left\lceil \frac{1}{2\eta} \right\rceil$. To prove the first inequality, assume Ψ_{Nat}-dim$_{\mathcal{H}}(\eta) \geq d$ for some $d \in \mathbb{N}$. Then there is a function $t : X \to \Re$, sequences $\mathbf{x} = (x_1, \ldots, x_d) \in X^d$ and $\psi = (\psi_1, \ldots, \psi_d) \in \Psi^d_{\text{Nat},n}$, and a subset $\mathcal{H}_0 \subseteq \mathcal{H}$ of cardinality 2^d such that

$$\left\{ \left(\psi_1(h''_{[\eta,t]}(x_1)), \ldots, \psi_d(h''_{[\eta,t]}(x_d)) \right) : h \in \mathcal{H}_0 \right\} = \{0, 1\}^d.$$

By definition, $\psi_i = \psi_{a_i,b_i}$ for some a_i and b_i in S_n with $a_i < b_i$. Without loss of generality, we can assume that a_i and b_i are in $\{0, 1, \ldots, n\}$ for $i = 1, 2, \ldots, d$. (Otherwise we could perturb t slightly at each of the points x_i and adjust the offending a_i or b_i appropriately, since $\mathcal{H}_0$ is finite.) Set $\Delta_i = b_i - a_i$. Since Ψ is an interval distinguisher, we can find a function α_i in Ψ_n such that, for some

$m_i \in \{0, 1, \ldots, n - \Delta\}$, α_i maps one of m_i and $m_i + \Delta_i$ to 0 and the other to 1. Defining t' as $t'(x_i) = t(x_i) + 2\eta(a_i - m_i)$, we have

$$\left\{\left(\alpha_1(h''_{[\eta,t']}(x_1)), \ldots, \alpha_d(h''_{[\eta,t']}(x_d))\right) : h \in \mathcal{H}_0\right\} = \{0,1\}^d,$$

which implies $\Psi\text{-dim}_{\mathcal{H}}(\eta) \geq d$.

To prove the second inequality, suppose $\Psi\text{-dim}_{\mathcal{H}}(\eta) \geq d$. As above, there is a function $t : X \to \Re$, and sequences $\mathbf{x} = (x_1, \ldots, x_d) \in X^d$ and $\psi = (\psi_1, \ldots, \psi_d) \in \Psi_n^d$ such that

$$\left\{\left(\psi_1(h''_{[\eta,t]}(x_1)), \ldots, \psi_d(h''_{[\eta,t]}(x_d))\right) : h \in \mathcal{H}\right\} \supseteq \{0,1\}^d.$$

By the definition of Ψ_B, we can find functions β_i in $\Psi_{B,n}$ which are equal to ψ_i on S_n. It follows that $\Psi_B\text{-dim}_{\mathcal{H}}(\eta) \geq d$.

Now, suppose $\Psi_B\text{-dim}_{\mathcal{H}}(\eta) \geq d$. Then there is a function $t : X \to \Re$, sequences $\mathbf{x} = (x_1, \ldots, x_d) \in X^d$ and $\psi = (\psi_1, \ldots, \psi_d) \in \Psi_B^d$, and a subset $\mathcal{H}_0$ satisfying $|\mathcal{H}_0| = 2^d$ such that

$$\{0,1\}^d \subseteq \left\{\left(\psi_1(h''_{[\eta,t]}(x_1)), \ldots, \psi_d(h''_{[\eta,t]}(x_d))\right) : h \in \mathcal{H}_0\right\}.$$

Clearly, $\mathcal{H}''_{0[\eta,t]}$ is a set of S_n-valued functions, and so we can consider the set of restrictions to S_n of functions in $\Psi_{B,n}$. Applying Theorem 6 gives

$$4.67 \log_2(2n+2)\Psi_{\mathrm{Nat},n}\text{-dim}(\mathcal{H}''_{[\eta,t]}) \geq \Psi_{B,n}\text{-dim}(\mathcal{H}''_{[\eta,t]}) \geq d,$$

and the third inequality follows. □

We can represent $\mathrm{Pdim}(\mathcal{H})$ and $\mathrm{d}^*_{\mathcal{H},\Re^X}(\eta)$ as dimensions of this form.

Let $\Psi_P = \{\Psi_{P,n} : n \in \mathbb{N}\}$ be the sequence of function classes $\Psi_{P,n} = \{\alpha_{n,m} : m \in S_n\}$ with $\alpha_{n,m} : S \to \{0, 1, *\}$ defined by

$$\alpha_{n,m}(y) = \begin{cases} 0 & \text{if } -1/2 \leq y \leq m \\ 1 & \text{if } m < y \leq n \\ * & \text{otherwise.} \end{cases}$$

Let Ψ^* be the sequence of function classes $\Psi^*_n = \{\beta_{n,m} : m \in \{0, 1, \ldots, n\}\}$ with $\beta_{n,m} : S \to \{0, 1, *\}$ defined by

$$\beta_{n,m}(y) = \begin{cases} 0 & \text{if } -1/2 \leq y \leq n \text{ and } y \neq m \\ 1 & \text{if } y = m \\ * & \text{otherwise.} \end{cases}$$

Clearly, $\Psi_P\text{-dim}_{\mathcal{H}}(\eta) = \mathrm{Pdim}(\mathcal{H})$ and $\Psi^*\text{-dim}_{\mathcal{H}}(\eta) = \mathrm{d}^*_{\mathcal{H},\Re^X}(\eta)$. Furthermore, Ψ_P and Ψ^* are interval distinguishers, so we can apply Theorem 12.

Theorem 13 *Suppose $\mathcal{H}$ is a class of $[0,1]$-valued functions defined on a set X. Then for all $\eta > 0$*

$$\mathrm{Pdim}(\mathcal{H}) < 4.67 \log_2 \left(2 \left\lceil \frac{1}{2\eta} \right\rceil + 2 \right) \mathrm{d}^*_{\mathcal{H},\Re^X}(\eta).$$

Furthermore, for any $\eta > 0$ and any sufficiently large set X, there is a class $\mathcal{H}$ of $[0,1]$-valued functions defined on X such that $\mathrm{Bdim}_{\mathcal{H}}(\eta) \leq 1$ but $\mathrm{Pdim}(\mathcal{H}) \geq \left\lfloor \log_2 \left(\frac{1}{4\eta} \right) \right\rfloor$.

Proof: Since Ψ_P and Ψ^* are interval distinguishers, Theorem 12 implies that $\Psi_{\mathrm{Nat}}\text{-}\dim_{\mathcal{H}}(\eta) \leq \mathrm{d}^*_{\mathcal{H},\Re^X}(\eta)$ and

$$\mathrm{Pdim}(\mathcal{H}) \leq \Psi_B\text{-}\dim_{\mathcal{H}}(\eta) < 4.67 \log_2 \left(2 \left\lceil \frac{1}{2\eta} \right\rceil + 2 \right) \Psi_{\mathrm{Nat}}\text{-}\dim_{\mathcal{H}}(\eta)$$

for all $\eta > 0$

To show that this bound cannot be improved asymptotically by more than a constant factor, let $N = \left\lfloor \log_2 \left(\frac{1}{4\eta} \right) \right\rfloor$. If $\eta > 1/8$, the second part of the theorem is trivially true, so assume $\eta \leq 1/8$ and hence $N \geq 1$. Define $\mathcal{H} = \{h_b : b \in 0, \ldots, 2^N - 1\}$ where $h_b : \{1, 2, \ldots, N\} \to [0,1]$ is defined by

$$h_b(n) = \begin{cases} 2^{-N-1} b & \text{if } \mathrm{bit}_n(b) = 0 \\ 1/2 + 2^{-N-1} b & \text{if } \mathrm{bit}_n(b) = 1, \end{cases}$$

where $\mathrm{bit}_n(b)$ is the nth bit from the right in the binary representation of b. Of course, for any sufficiently large X, $\mathcal{H}$ is isomorphic to some function class defined on X. For any distinct $b_1, b_2 \in \{0, 1, \ldots, 2^N - 1\}$ we have

$$|h_{b_1}(n) - h_{b_2}(n)| \geq 2^{-N-1} |b_1 - b_2| \geq 2\eta.$$

Clearly, $\mathrm{Bdim}_{\mathcal{H}}(\eta) = \mathrm{d}^*_{\mathcal{H},\Re^X}(\eta) = 1$ but $\mathrm{Pdim}(\mathcal{H}) = N$. □

We now state the following result, which follows immediately, and which completes the proof of Theorem 2.

Theorem 14 *Suppose that $\mathcal{H}$ is a set of functions from a set X to $[0,1]$. Then $\mathcal{H}$ validly generalises from approximate interpolation if and only if $\mathcal{H}$ has finite pseudo-dimension. Furthermore, there are constants $c_1, c_2 > 0$ such that if $\mathcal{H}$ has*

finite pseudo-dimension $\mathrm{Pdim}(\mathcal{H})$ *then a sufficient sample length function for generalisation from approximate interpolation is*

$$\frac{c_1}{\epsilon}\left(\mathrm{Pdim}(\mathcal{H})\ln\left(\frac{1}{\epsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right),$$

and any suitable sample length function must satisfy

$$m_0(\eta,\epsilon,\delta)\geq c_2\frac{1}{\epsilon}\left(\frac{\mathrm{Pdim}(\mathcal{H})}{\log(1/\eta)}+\log\left(\frac{1}{\delta}\right)\right)$$

for all $\eta>0$ *and* $\epsilon,\delta\in(0,1)$.

We say that a sequence Ψ of functions from S to $\{0,1,*\}$ *characterises valid generalisation from approximate interpolation* if, for all classes $\mathcal{H}$ of $[0,1]$-valued functions defined on X, $\mathcal{H}$ validly generalises from approximate interpolation if and only if $\Psi\text{-dim}_{\mathcal{H}}(\eta)$ is finite for all $\eta>0$. As in the proof of Theorem 13, we can use Theorem 12 to show that any interval distinguisher characterises valid generalisation from approximate interpolation. The following theorem also shows that the interval distinguishers are the *only* such admissible function sequences, giving a characterisation of those admissible function sequences that characterise valid generalisation from approximate interpolation. The proof is in the Appendix.

Theorem 15 *For any admissible sequence* Ψ *of* $\{0,1,*\}$*-valued functions,* Ψ *characterises valid generalisation from approximate interpolation if and only if* Ψ *is an interval distinguisher.*

It is reasonable to consider only dimensions of $\mathcal{H}$ that can be expressed as $\Psi\text{-dim}_{\mathcal{H}}(\eta)$ for some admissible Ψ, since only discrete properties of $\mathcal{H}$ in relation to intervals of width 2η are relevant to the definition of valid generalisation from η-approximate interpolation. The Ψ-dimensions capture all properties of $\mathcal{H}$ when it is quantised in all possible ways with quantisation width 2η. As Proposition 10 shows, requiring that Ψ be admissible is only a notational convenience.

5 The Restricted Problem: $\mathcal{C}=\mathcal{H}$

In this section, we concentrate on the case in which $\mathcal{C}=\mathcal{H}$. From the previous results, a necessary and sufficient condition for $\mathcal{H}$ to validly generalise $\mathcal{H}$ from approximate interpolation is that $\mathrm{d}^*_{\mathcal{H},\mathcal{H}}(\eta)<\infty$ for all $\eta>0$. We shall henceforth

denote $\mathrm{d}^*_{\mathcal{H},\mathcal{H}}(\eta)$ simply by $\mathrm{Ddim}_{\mathcal{H}}(\eta)$. The main purpose of this section is to show that this measure of dimension is different from other measures of dimension which have occurred in the learning theory of real functions.

We have already discussed the pseudo-dimension and have seen that finiteness of the pseudo-dimension is a necessary and sufficient condition for the (unrestricted) problem of valid generalisation from approximate interpolation. We now show, however, that finiteness of the pseudo-dimension is *not* a necessary condition for the restricted problem.

Proposition 16 *There is a set $\mathcal{H}$ of functions from the set $\mathbb{N}$ of positive integers to $[0,1]$ such that $\mathcal{H}$ validly generalises $\mathcal{H}$ from approximate interpolation, but $\mathcal{H}$ has infinite pseudo-dimension.*

Proof: For each positive integer i, let h_i be the function from $\mathbb{N}$ to $[0,1]$ given by

$$h_i(n) = \begin{cases} 1/(n+1) & \text{if } \mathrm{bit}_n(i) = 1; \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathrm{bit}_n(i)$ is the nth bit from the right in the binary representation of i. Let $\mathcal{H} = \{h_i : i \in \mathbb{N}\}$. For any $k \geq 1$, the set $\{1, 2, \ldots, k\}$ is shattered: take $\mathbf{r} = (1/2k, 1/2k, \ldots, 1/2k)$ in the definition of shattering and, for $\mathbf{b} \in \{0,1\}^k$, let $h_{\mathbf{b}}$ be h_i where i is the integer whose binary expansion is $b_n b_{n-1} \ldots b_1$. It follows that $\mathcal{H}$ has infinite pseudo-dimension. To show that $\mathcal{H}$ generalises $\mathcal{H}$ from approximate interpolation, we show that $\mathrm{Ddim}_{\mathcal{H}}(\eta)$ is finite for all $\eta > 0$. Fix η. If $d \geq 1/\eta$, then for all $h_j \in \mathcal{H}$, $h_j(d)$ is either 0 or $1/(d+1)$. In either case, $0 \leq h_j(d) < \eta$. Thus, for all $t = h_i \in \mathcal{H}$ and for all j, $|t(d) - h_j(d)| < \eta$. It follows that d cannot belong to any subset of $\mathbb{N}$ which is shattered by $\mathcal{H}_{[\eta,t]}$. Since this is true for any $d \geq 1/\eta$, we have $\mathrm{VCdim}(\mathcal{H}_{[\eta,t]}) < 1/\eta$ for $\eta > 0$. But this is true for all $t \in \mathcal{H}$ and hence

$$\mathrm{Ddim}_{\mathcal{H}}(\eta) = \max_{t \in \mathcal{H}} \mathrm{VCdim}(\mathcal{H}_{[\eta,t]}) < 1/\eta,$$

and so, since $\mathrm{Ddim}_{\mathcal{H}}(\eta)$ is finite for all $\eta > 0$, $\mathcal{H}$ generalises $\mathcal{H}$ from interpolation. □

This result shows that the restricted problem is easier than the unrestricted problem. Moreover, it shows that, while $\mathrm{Pdim}(\mathcal{H}) < \infty$ implies $\mathrm{Ddim}_{\mathcal{H}}(\eta) < \infty$ for all $\eta > 0$, the converse is false.

Another measure of dimension which has been important in the development of the theory of learning real functions is a 'scale-sensitive' version of the pseudo-dimension.

This dimension was introduced by Kearns and Schapire [16] in their work on the learnability of p-concepts. Here, we use the notation and terminology of [9]. Suppose that $\mathcal{H}$ is a set of functions from X to $[0,1]$ and that $\gamma > 0$. We say that the finite subset $S = \{x_1, x_2, \ldots, x_d\}$ of X is *γ-shattered* if there is $\mathbf{r} = (r_1, r_2, \ldots, r_d) \in \Re^d$ such that for every $\mathbf{b} = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $h_{\mathbf{b}} \in \mathcal{H}$ with

$$h_{\mathbf{b}}(x_i) \begin{cases} \geq r_i + \gamma & \text{if } b_i = 1 \\ \leq r_i - \gamma & \text{if } b_i = 0. \end{cases}$$

Thus, S is γ-shattered if it is shattered, with a 'width of shattering' of at least γ. We define $\mathrm{fat}_{\mathcal{H}}(\gamma)$ as the largest cardinality of a γ-shattered set, or infinity if there is no bound on the cardinalities of such sets. The *fat-shattering function* is the function $\mathrm{fat}_{\mathcal{H}}(\gamma)$ of γ, from $\Re^+$ to $\mathbb{N}_0 \cup \{\infty\}$. It is easy to see that $\mathrm{Pdim}(\mathcal{H}) = \lim_{\gamma \to 0} \mathrm{fat}_{\mathcal{H}}(\gamma)$. It should be noted, however, that it is possible for the pseudo-dimension to be infinite, even when $\mathrm{fat}_{\mathcal{H}}(\gamma)$ is finite for all $\gamma > 0$. We shall say that $\mathcal{H}$ has *finite fat-shattering function* whenever it is the case that for all $\gamma > 0$, $\mathrm{fat}_{\mathcal{H}}(\gamma)$ is finite. Kearns and Schapire [16] proved that if a class of p-concepts is learnable, then the class has finite fat-shattering dimension. Alon *et al.* [1] proved, conversely, that if a class of p-concepts has finite fat-shattering function, then it is learnable. This follows from a more general result they obtained, classifying classes that satisfy a certain uniform convergence property (the Glivenko-Cantelli classes) as those with finite fat-shattering function. Bartlett, Long and Williamson [9] proved that finiteness of the fat-shattering function is a necessary and sufficient condition for a standard model of function learning in the presence of (certain forms of) random noise. We have the following result, which shows that finiteness of the fat-shattering function is *not* a sufficient condition for restricted valid generalisation from approximate interpolation.

Proposition 17 *There is a set $\mathcal{H}$ of functions from $[0,1]$ to $[0,1]$ such that $\mathcal{H}$ has finite fat-shattering function but such that $\mathcal{H}$ does* not *validly generalise $\mathcal{H}$ from approximate interpolation.*

Proof: Let $\mathcal{H}$ be the set of all functions $h : [0,1] \to [0,1]$ which are 1-Lipschitz-continuous. Thus, $\mathcal{H}$ is the set of all functions h such that

$$|h(x) - h(y)| \leq |x - y| \quad \text{for all } x, y \in [0,1].$$

Then, it is easily seen that $\mathcal{H}$ has finite fat-shattering function. However, $\mathcal{H}$ does not validly generalise $\mathcal{H}$ from approximate interpolation. To see this, we can show that $\mathrm{Ddim}_{\mathcal{H}}(\eta)$ is infinite for some η. (Fix $\eta < 1/2$ and $t : x \mapsto 1/2$ and consider the subset of $\mathcal{H}$ containing functions that take values close to $\eta + 1/2$.)

We provide an alternative proof that illustrates why $\mathcal{H}$ does not validly generalise from interpolation. Take t to be the identically-0 function and $\mathcal{P}$ to be the uniform distribution on $[0,1]$. Let m be any positive integer and suppose that a sample $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in [0,1]^m$ is given and (without loss) suppose that $x_1 < x_2 < \cdots < x_m$. For convenience, let $x_0 = 0$ and $x_{m+1} = 1$. We now define a function h piecewise, on each of the intervals $[x_i, x_{i+1}]$ for $0 \leq i \leq m$. On the interval $[x_i, x_{i+1}]$, let $h(x) = \min(1, g(x))$, where

$$g(x) = \begin{cases} \eta - \alpha + (x - x_i) & \text{if } x_i \leq x \leq (x_i + x_{i+1})/2 \\ \eta - \alpha + (x_{i+1} - x) & \text{if } (x_i + x_{i+1})/2 \leq x \leq x_{i+1}, \end{cases}$$

with

$$0 < \alpha \leq \min_{0 \leq i \leq m} (x_{i+1} - x_i)/4$$

and $\alpha \leq \eta$. Clearly, for $1 \leq i \leq m$,

$$|h(x_i) - t(x_i)| = |h(x_i)| \leq \eta - \alpha < \eta.$$

It is easily checked that $h \in \mathcal{H}$ and that

$$\mathcal{P}\left(\{x \in [0,1] : |h(x) - t(x)| \geq \eta\}\right) = \mathcal{P}\left(\{x : h(x) \geq \eta\}\right) \geq 1/2.$$

Since m was arbitrary, this shows that $\mathcal{H}$ does not generalise $\mathcal{H}$ from interpolation. □

This result shows that finiteness of the fat-shattering function does not imply finiteness of $\mathrm{Ddim}_{\mathcal{H}}(\gamma)$ for all γ. The results of this section therefore show that the dimension function Ddim_H is quite distinct from two important dimensions which have proven to be useful in other forms of function learning. In particular, since finite fat-shattering function is a sufficient condition for function learning [9], we see that (restricted) valid generalisation from approximate interpolation is a *strictly* stronger condition than learnability, a fact briefly mentioned earlier in the paper. Finiteness of the pseudo-dimension implies finiteness of $\mathrm{Ddim}_{\mathcal{H}}(\gamma)$ for all γ, while it is not true that finiteness of the fat-shattering function implies finiteness of $\mathrm{Ddim}_{\mathcal{H}}(\gamma)$ for all γ. It is natural to ask whether, in some sense, $\mathrm{Ddim}_{\mathcal{H}}(\gamma)$ lies 'between' the pseudo-dimension and the fat-shattering function. In fact, this is so; in [4], a relationship is derived which shows that if $\mathrm{Ddim}_{\mathcal{H}}(\gamma)$ is finite for all $\gamma > 0$ then $\mathcal{H}$ has finite fat-shattering function. In other words, we have

$$\mathrm{Pdim}(\mathcal{H}) < \infty \Longrightarrow \forall \gamma > 0,\ \mathrm{Ddim}_{\mathcal{H}}(\gamma) < \infty \Longrightarrow \forall \gamma > 0,\ \mathrm{fat}_{\mathcal{H}}(\gamma) < \infty,$$

with neither implication reversible. The proof of the second implication is given in [4].

In [27] (Chapter 7), Vapnik showed that finiteness of a related dimension of $\mathcal{H}$ (that he called the capacity of $\mathcal{H}$) was sufficient for uniform convergence over $\mathcal{H}$ of

$$\frac{1}{m}\sum_{i=1}^{m}(h(x_i)-t(x_i))^2$$

to $E(h(x)-t(x))^2$.

Notice that $\mathrm{Bdim}_{\mathcal{H}}(\gamma) = \mathrm{VCdim}(\mathcal{H}_1(\gamma))$, where

$$\mathcal{H}_1(\gamma) = \left\{(x,y) \mapsto \begin{cases} 0 & \text{if } |h(x)-y| \geq \gamma \\ 1 & \text{otherwise} \end{cases} : h \in \mathcal{H}\right\}.$$

Vapnik's capacity can be expressed as the VC-dimension of $\bigcup_{\gamma>0}\mathcal{H}_1(\gamma)$. Obviously, finiteness of Vapnik's capacity implies finiteness of $\mathrm{Bdim}_{\mathcal{H}}(\gamma)$ for all γ. By Theorem 2, this implies finiteness of the pseudo-dimension of $\mathcal{H}$. Theorem 8 in [1] shows that finiteness of the fat-shattering function of $\mathcal{H}$ (a strictly weaker condition on $\mathcal{H}$ than finiteness of the pseudo-dimension) is sufficient for the uniform convergence property studied by Vapnik.

6 The Unbounded Case

In this section, we briefly discuss the case of classes of functions which are not uniformly bounded. Until now, we have dealt solely with classes of functions mapping into some fixed bounded interval. The definitions of generalisation from approximate interpolation still make sense when $\mathcal{H}$ does not map into a bounded set. Analysis of the proofs shows that the general results of Section 3 concerning generalisation of $\mathcal{C}$ from approximate interpolation remain true for such classes $\mathcal{H}$. In particular, $\mathcal{H}$ validly generalises $\Re^X$ from approximate interpolation if and only if $\mathrm{Bdim}_{\mathcal{H}}(\eta)$ is finite for all $\eta > 0$. The proof of Theorem 7 also remains valid if functions in $\mathcal{H}$ map to $\Re$, so finite pseudo-dimension is sufficient for valid generalisation from approximate interpolation in this case also. If $\mathcal{H}$ is a linear space, we can find tight bounds on the necessary sample size.

Proposition 18 *If $\mathcal{H}$ is a linear space of real-valued functions defined on X, then*

$$\dim(\mathcal{H}) \leq d^*_{\mathcal{H},\Re^X}(\eta) < 7.5\ \dim(\mathcal{H}),$$

where $\dim(\mathcal{H})$ *is the (linear) dimension of* $\mathcal{H}$.

Proof: Suppose $\dim(\mathcal{H}) \geq d$. Then it is possible to choose an $\mathbf{x} = (x_1, \ldots, x_d) \in X^d$ such that $\{(h(x_1), \ldots, h(x_d)) : h \in \mathcal{H}\} = \Re^d$, which implies the first inequality. The second inequality follows from Theorem 7 and the fact that $\mathrm{Pdim}(\mathcal{H}) = \dim(\mathcal{H})$ (see for example [14]). □

While finite pseudo-dimension is a sufficient condition for valid generalisation from approximate interpolation, it is *not* necessary in such cases. Indeed, consider the following example. For each positive integer i, let $f_i : \mathbb{N} \to \Re$ be defined by

$$f_i(n) = \begin{cases} i & \text{if } \mathrm{bit}_n(i) = 1; \\ -i & \text{otherwise,} \end{cases}$$

where $\mathrm{bit}_n(i)$ the n^{th} digit from the right in the binary encoding of i. Let $\mathcal{H} = \{f_i : i \in \mathbb{N}\}$. Then it is clear that $\mathcal{H}$ has infinite pseudo-dimension but, for all $\eta > 0$, $\mathcal{H}$ has finite η-band-dimension and hence generalises from approximate interpolation.

7 Conclusions

Figure 2 summarises the necessary and sufficient conditions for valid generalisation from approximate interpolation under various assumptions on the hypothesis and target classes. In all cases we have presented sample complexity bounds that cannot be improved by more than a $\log 1/\epsilon$ factor.

One obvious variant of the problem studied here is that in which there is an extra parameter $\gamma > 0$ and one demands that, with high probability, every η-interpolant be $(\eta + \gamma)$-close to the target on a set of measure at least $1 - \epsilon$ (rather than η-close there). This is a weakening of the generalisation from approximate interpolation condition. In [4], it is shown that finiteness of the fat-shattering function is necessary and sufficient for this weaker condition to hold.

Acknowledgements

This research was supported in part by the Australian Telecommunications and Electronics Research Board and by the European Union through the "Neurocolt" ESPRIT Working Group. Thanks to an anonymous reviewer for helpful suggestions.

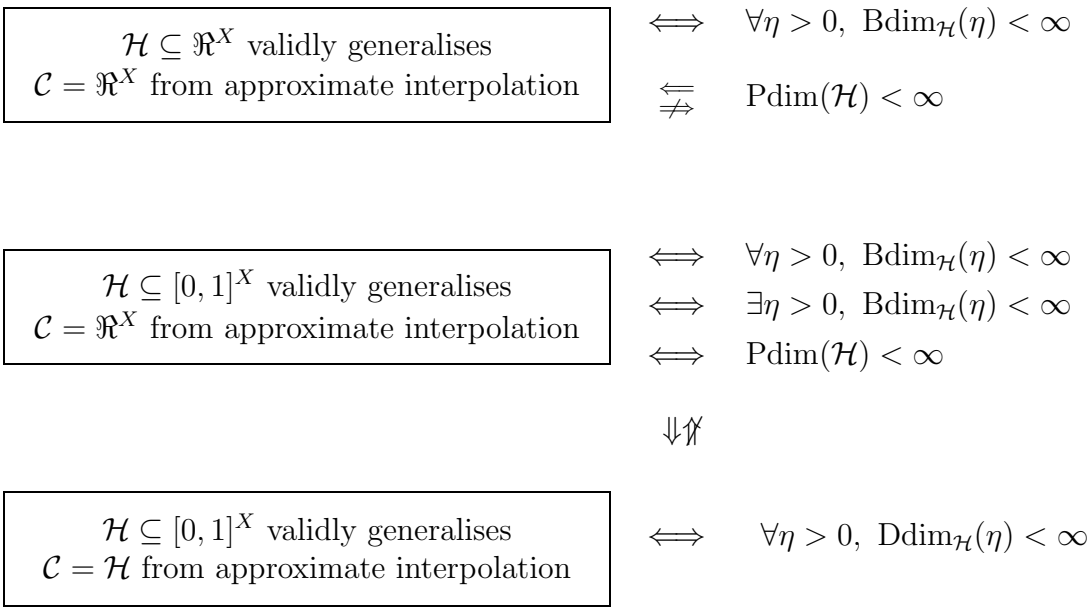

Figure 2: Necessary and sufficient conditions for valid generalisation from approximate interpolation.

References

[1] Alon, N. , Ben-David, S. , Cesa-Bianchi, N. and Haussler, D. (1993). Scale-sensitive dimensions, uniform convergence, and learnability. In *Proceedings of the 1993 IEEE Symposium on Foundations of Computer Science*, IEEE Press.

[2] Angluin, D. (1988) Queries and concept learning. *Machine Learning* 2(4): 319–342.

[3] Angluin, D. and Valiant, L. (1979). Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, 18: 155-193.

[4] Anthony, M. and Bartlett, P.L. (1994). Function learning from interpolation. In preparation.

[5] Anthony, M. and Biggs, N. (1992). *Computational Learning Theory: An Introduction*, Cambridge University Press.

[6] Anthony, M. , Biggs, N. and Shawe-Taylor, J. (1990) The learnability of formal concepts. In *COLT'90, Proceedings of the Third Annual Workshop on Computational Learning Theory*, Morgan Kaufmann, San Mateo, CA.

[7] Anthony, M. and Shawe-Taylor, J. (1993). A result of Vapnik with applications, *Discrete Applied Mathematics*, 47: 207–217.

[8] Anthony, M. and Shawe-Taylor, J. (1994). Valid generalisation of functions from close approximations on a sample. In *Computational Learning Theory: EUROCOLT'93* (ed. J. Shawe-Taylor and M. Anthony), Oxford University Press.

[9] Bartlett, P.L. , Long, P.M. and Williamson, R.C. (1994). Fat-shattering and the learnability of real-valued functions. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory*, ACM Press, New York.

[10] Ben-David, S. , Benedek, G. M. and Mansour, Y. (1989). A parameterization scheme for classifying models of learnability. In *COLT'89, Proceedings of the Second Workshop on Computational Learning Theory*, Morgan Kaufmann, San Mateo, CA.

[11] Ben-David, S. , Cesa-Bianchi, N. , Haussler, D. and Long, P. (1992). Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. Technical Report. (An earlier version appeared in *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, ACM Press, New York.) To appear, *Journal of Computer and System Sciences*.

[12] Blumer, A. , Ehrenfeucht, A. , Haussler, D. and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension, *Journal of the ACM* 36(4): 929–965.

[13] Ehrenfeucht, A. , Haussler, D. , Kearns, M. and Valiant, L. (1989). A general lower bound on the number of examples needed for learning, *Information and Computation* 82: 247-261.

[14] Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Information and Computation*, 100: 78–150.

[15] Haussler, D. and Long, P. (1990). A generalization of Sauer's lemma, Technical Report UCSC-CRL-90-15, University of California at Santa Cruz.

[16] Kearns, M.J. and Schapire, R. E. (1990). Efficient distribution-free learning of probabilistic concepts, in *Proceedings of the 1990 IEEE Symposium on Foundations of Computer Science*, IEEE Press.

[17] Natarajan, B.K. (1989). On learning sets and functions, *Machine Learning*, 4: 67–97.

[18] Natarajan, B.K. (1991). *Machine Learning: A Theoretical Approach*, Morgan Kaufmann, San Mateo, CA.

[19] Natarajan, B.K. (1993). Occam's razor for functions. In *Proceedings of the Sixth ACM Workshop on Computational Learning Theory, July 1993*, ACM Press.

[20] Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer-Verlag.

[21] Shawe-Taylor, J. and Anthony, M. (1991). Sample sizes for multiple-output threshold networks, *Network: Computation in Neural Systems*, 2: 107–117.

[22] Shawe-Taylor, J. , Anthony, M. and Biggs, N.L. (1993). Bounding sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 41: 65–73.

[23] Simon, H.U. (1993). General bounds on the number of examples needed for learning probabilistic concepts. In *Proceedings of the Sixth ACM Workshop on Computational Learning Theory, July 1993*, ACM Press. To appear, *Journal of Computer and System Sciences*.

[24] Simon, H. U. (1994). Bounds on the number of examples needed for learning functions. In *Computational Learning Theory: EUROCOLT'93* (ed. J. Shawe-Taylor and M. Anthony), Oxford University Press.

[25] Sontag, E.D. (1992). Feedforward nets for interpolation and classification, *Journal of Computer and System Sciences*, 45: 20–48.

[26] Valiant, L.G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11): 1134–1142.

[27] Vapnik, V.N. (1982). *Estimation of Dependences Based on Empirical Data*, Springer-Verlag.

[28] Vapnik, V.N. and Chervonenkis, A.Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2): 264-280.

Appendix: Proof of Theorem 15

Theorem 15 *For any admissible sequence Ψ of $\{0,1,*\}$-valued functions, Ψ characterises valid generalisation from approximate interpolation if and only if Ψ is an interval distinguisher.*

Proof: If Ψ is an interval distinguisher, Theorem 12 shows that $\Psi\text{-dim}_{\mathcal{H}}(\eta)$ is within a factor of $\log\frac{1}{\eta}$ of $\text{Bdim}_{\mathcal{H}}(\eta)$. So Theorem 5 implies that finiteness of $\Psi\text{-dim}_{\mathcal{H}}(\eta)$ for all $\eta > 0$ is necessary and sufficient for valid generalisation from approximate interpolation.

Conversely, suppose $\Psi = \langle \Psi_1, \Psi_2, \ldots \rangle$ is not an interval distinguisher. Then there is an $\eta > 0$ and a $\Delta \in \{1, 2, \ldots, n\}$ (with $n = \left\lceil \frac{1}{2\eta} \right\rceil$) such that, for all $\psi \in \Psi_n$ and all $m \in \{0, 1, \ldots, n - \Delta\}$, $\psi(m) = *$ or $\psi(m + \Delta) = *$ or $\psi(m) = \psi(m + \Delta)$.

Suppose that $\eta \geq 1/2$, which implies $n = 1$. Let $\mathcal{H} = \{h_m : m \in \mathbb{N}\}$ be the set of functions from $X = \mathbb{N}$ to $[0, 1]$ with

$$h_m(k) = \begin{cases} \frac{1}{m+4/\eta-1} & \text{if } \text{bit}_k(m) = 0 \\ 1 - \frac{1}{m+4/\eta-1} & \text{if } \text{bit}_k(m) = 1 \end{cases}$$

where $\text{bit}_k(m)$ is the kth bit from the right in the binary representation of m. Now, suppose that there are points $x_1, x_2 \in X$, a function $t : \{x_1, x_2\} \to \Re$, and functions $\psi_1, \psi_2 \in \Psi_1$ such that

$$\{0, 1\}^2 \subseteq \left\{ \left(\psi_1 \left(\phi \left(\frac{h_m(x_1) - t(x_1)}{2\eta} \right) \right), \psi_2 \left(\phi \left(\frac{h_m(x_2) - t(x_2)}{2\eta} \right) \right) \right) : m \in \mathbb{N} \right\}.$$

Without loss of generality, we may assume that $\psi_1(0) = \psi_1(1) = 1$. Then if

$$\psi_1 \left(\phi \left(\frac{h_m(x_1) - t(x_1)}{2\eta} \right) \right) = 0$$

we must have $h_m(x_1) = t(x_1)$ or $h_m(x_1) = t(x_1) - 2\eta$. But this can be true only for two values of m. It follows that $\Psi\text{-dim}_{\mathcal{H}}(\eta) \leq 2$. However, it is clear that $\text{Pdim}(\mathcal{H}) = \infty$, so finiteness of $\Psi\text{-dim}_{\mathcal{H}}(\eta)$ does not imply that $\mathcal{H}$ validly generalises from η-approximate interpolation.

Assume now that $\eta \in (0, 1/2)$. Consider the function class $\mathcal{H} = \{h_m : m \in \mathbb{N}\}$ where $h_m : \mathbb{N} \to [0, 1]$ is defined by

$$h_m(k) = \begin{cases} \frac{1}{m+c/\eta-1} & \text{if } \text{bit}_k(m) = 0 \\ 2\Delta\eta + \frac{1}{m+c/\eta-1} & \text{if } \text{bit}_k(m) = 1, \end{cases}$$

where $c > 1$ and

$$c > \eta + \frac{1}{2\left(1 - \left\lceil \frac{1}{2\eta} \right\rceil + \frac{1}{2\eta}\right)}.$$

It is easy to show that these conditions imply that h_m maps to $[0,1]$ and that

$$\frac{1}{m + c/\eta - 1} < \eta.$$

Now, suppose that there are points $x_1, x_2 \in X$, a function $t : \{x_1, x_2\} \to \Re$, and functions $\psi_1, \psi_2 \in \Psi_n$ such that

$$\{0,1\}^2 \subseteq \left\{ \left(\psi_1 \left(\phi \left(\frac{h_m(x_1) - t(x_1)}{2\eta} \right) \right), \psi_2 \left(\phi \left(\frac{h_m(x_2) - t(x_2)}{2\eta} \right) \right) \right) : m \in \mathbb{N} \right\}. \tag{1}$$

Consider, for any fixed η, the set

$$A = \left\{ \frac{h_m(x_1) - t(x_1)}{2\eta} : m \in \mathbb{N} \text{ and } \mathrm{bit}_{x_1}(m) = 0 \right\}.$$

For some $a_1 \in \Re$, A is a subset of the interval $(a_1, a_1 + 1/2)$. Similarly, the set

$$B = \left\{ \frac{h_m(x_1) - t(x_1)}{2\eta} : m \in \mathbb{N} \text{ and } \mathrm{bit}_{x_1}(m) = 1 \right\}$$

is a subset of the interval $(\Delta + a_1, \Delta + a_1 + 1/2)$. For $\{x_1\}$ to be shattered by $\mathcal{H}''_{[\eta,t]}$, there must be numbers $m_1, m_2 \in \mathbb{N}$ for which

$$\phi \left(\frac{h_{m_1}(x_1) - t(x_1)}{2\eta} \right) - \phi \left(\frac{h_{m_2}(x_1) - t(x_1)}{2\eta} \right) \tag{2}$$

is not in $\{0, \Delta\}$ (whatever the values of $\mathrm{bit}_{x_1}(m_1)$ and $\mathrm{bit}_{x_1}(m_2)$).

If no integer falls in the interval $(a_1, a_1 + 1/2)$, then for all m_1 and m_2 in $\mathbb{N}$, (2) is either 0 or Δ, and $\{x_1, x_2\}$ is not Ψ_n-shattered by $\mathcal{H}''_{[\eta,t]}$. So assume that there is a $k_1 \in \mathbf{Z}$ satisfying $k_1 \in (a_1, a_1 + 1/2)$. Without loss of generality, we may assume that ψ_1 satisfies

$$\psi_1(\phi(\alpha)) = \begin{cases} 0 & \text{if } \alpha \in (a_1, k_1) \text{ or } \alpha \in (\Delta + a_1, \Delta + k_1) \\ 1 & \text{if } \alpha \in (k_1, a_1 + 1/2) \text{ or } \alpha \in (\Delta + k_1, \Delta + a_1 + 1/2). \end{cases}$$

Then since (1) is true, there must be an m_1 and m_2 in $\mathbb{N}$ satisfying

$$\frac{h_{m_1}(x_1) - t(x_1)}{2\eta} \in (a_1, k_1] \cup (\Delta + a_1, \Delta + k_1]$$

and

$$\frac{h_{m_2}(x_1) - t(x_1)}{2\eta} \in [k_1, a_1 + 1/2) \cup [\Delta + k_1, \Delta + a_1 + 1/2).$$

These conditions imply that $m_1 \leq D_1$ and $m_2 \geq D_1$, where

$$D_1 = 1 - \frac{c}{\eta} - \frac{1}{t(x_1) + 2\eta k_1}.$$

Defining a_2 and k_2 in the same way for x_2, we can assume without loss that ψ_2 satisfies

$$\psi_2(\phi(\alpha)) = \begin{cases} 0 & \alpha \in (a_2, k_2) \text{ or } \alpha \in (\Delta + a_2, \Delta + k_2) \\ 1 & \alpha \in (k_2, a_2 + 1/2) \text{ or } \alpha \in (\Delta + k_2, \Delta + a_2 + 1/2). \end{cases}$$

In that case, for (1) to be true there must be four distinct numbers $m_1, m_2, m_3, m_4 \in \mathbb{N}$ satisfying

$$\begin{array}{ll} m_1 \leq D_1 & m_1 \leq D_2 \\ m_2 \geq D_1 & m_2 \leq D_2 \\ m_3 \leq D_1 & m_3 \geq D_2 \\ m_4 \geq D_1 & m_4 \geq D_2 \end{array}$$

where D_2 depends on c, η, $t(x_2)$ and k_2, and is defined in the same way as D_1. These inequalities imply $D_1 \leq m_2 \leq D_2$ and $D_2 \leq m_3 \leq D_1$, so $m_2 = m_3 = D_1 = D_2$. But this contradicts the assumption that the four numbers are distinct. It follows that $\Psi\text{-dim}_{\mathcal{H}}(\eta) \leq 1$. However, it is obvious that $\text{Pdim}(\mathcal{H}) = \infty$, so $\mathcal{H}$ does not validly generalise from approximate interpolation. □