# Bounding Sample Size with the Vapnik-Chervonenkis Dimension

John Shawe-Taylor
Department of Computer Science

and

Martin Anthony
Department of Mathematics
Royal Holloway and Bedford New College
Egham Surrey TW20 0EX UK

and

N.L. Biggs
Department of Mathematical Sciences
London School of Economics
Houghton St
London WC2A 2AE

9th November, 1989

For reprint requests contact John Shawe-Taylor at the above address. Tel: (0784) 439021.

Running title: Vapnik-Chervonenkis Dimension

# Bounding Sample Size with the Vapnik-Chervonenkis Dimension

## Abstract

A proof that a concept is learnable provided the Vapnik-Chervonenkis dimension is finite is given. The proof is more explicit than previous proofs and introduces two new parameters which allow bounds on the sample size obtained to be improved by a factor of approximately $4\log_2(e)$.

Keywords: learning, Vapnik-Chervonenkis dimension, Probably Approximately Correct (PAC), sample.

## 1 Introduction

The problem of deciding what sample size is needed to guarantee accurate learning in Valiant's Probably Approximately Correct [Vali84] sense has received much attention in the literature. This has been brought into focus by the current interest in connectionist models of learning. Many experimental results have been obtained with these models but there is little theory to justify their generality. The first inroads into this area were made by Pitt and Valiant [PiVa] and Blumer, Ehrenfeucht, Haussler and Warmuth [BEHW87], who pointed out that if the number $r$ of hypotheses is finite then the probability that any hypothesis with error larger than $\epsilon$ is consistent with the target concept on a sample of size $m$, is less than $(1-\epsilon)^m r$. We include their proof in Section 2 for completeness and as motivation for the proof of our main theorem. Their result led them to introduce *Occam-algorithms*. The definition of an Occam-algorithm includes the requirement that a learning algorithm exist for the concept which runs in polynomial time. This requirement is not necessary for the error bound which holds independently of how the hypothesis was arrived at or indeed which hypothesis it is, providing only that it is consistent with the given sample.

In the case when the number of hypotheses is actually or potentially infinite, which includes most realistic examples of learning and certainly feedforward Neural Networks with real weights, the extent of the hypothesis space has to be bounded in some way before anything can be said independently of the learning algorithm. This is because if all functions are allowed then however large a (finite) sample we have we can choose the function to be as inaccurate as we like on the rest of the (infinite) input space. The breakthrough in the case of infinite hypothesis spaces came with the application of the Vapnik-Chervonenkis dimension [VaCe71]. Using this measure of the size of the hypothesis space it was shown [BEHW89], [HaWe87], that the sample size needed can be bounded in terms of the Vapnik-Chervonenkis dimension, again independently of any learning algorithm.

This paper gives a proof of this result, which introduces two additional parameters and so improves the bound obtained for the sample size by a factor of approximately $4\log_2(e)$.

## 2 Definitions and Preliminary Results

We first introduce the framework within which we study learning. Let $(X, \varepsilon, \mu)$ be a probability space. This is the space from which the samples and test examples will be chosen and can be thought of as the space of "naturally occurring" phenomena. A *concept* defined over $X$ is a mapping $c$:

$$c \in \mathcal{C}(X) = \big\{ f | f : X \to \big\{+1, -1\big\} \text{ and } f \text{ is measurable} \big\},$$

identifying the elements of $X$ as positive or negative examples. Fix a concept $c$, calling it the *target* concept. The task of a learning algorithm is to find a good approximation to $c$ from some given subset $H \subseteq \mathcal{C}(X)$, called the set of *hypotheses*. The approximation must be found by using information about the target concept's behaviour on a sample of inputs.

Given any $f \in \mathcal{C}(X)$ we define the *actual error* of $f$ (with respect to $c$) as

$$\mathrm{er}_\mu(f) = \mu\big\{x \in X | f(x) \neq c(x)\big\}.$$

Further for a set $H$ of hypotheses we define the *haziness* of $H$ (with respect to $c$) as

$$\mathrm{haz}_\mu(H) = \sup\big\{\mathrm{er}_\mu(h) | h \in H\big\}.$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$ and an $h \in \mathcal{C}(X)$ the *observed error* of $h$ on $\mathbf{x}$ (with respect to $c$) is

$$\mathrm{er}_{\mathbf{x}}(h) = \frac{1}{m} \left| \big\{ i | h(x_i) \neq c(x_i) \big\} \right|.$$

The subset of a set $H$ of hypotheses *consistent* with $\mathbf{x}$ (with respect to $c$) is

$$H[\mathbf{x}] = \big\{ h \in H | \mathrm{er}_{\mathbf{x}}(h) = 0 \big\}.$$

We say that $H$ can *approximate* $c$ if for all $m$ and all $x \in X^m$, $H[\mathbf{x}] \neq \emptyset$: this is certainly true if $c \in H$. We can now introduce the concept of *learnable* in the Probably Approximately Correct sense.

**Definition 2.1 :** *Given a set $H$ of hypotheses, the target concept $c$ is learnable in context $H$, if $H$ can approximate $c$ and, given $\epsilon, \delta \in (0,1)$, there is a positive integer $m_0 = m_0(\epsilon, \delta)$ such that*

$$\mu^m \big\{ \mathbf{x} \in X^m | \mathrm{haz}_\mu(H[\mathbf{x}]) \leq \epsilon \big\} > 1 - \delta, \quad \textit{for all } m \geq m_0.$$

Note that this definition says nothing about how hard it is to find an element in $H[\mathbf{x}]$, or about the rate of growth of the function $m_0(\epsilon, \delta)$. Many definitions of learnability include stipulation of a polynomially bounded algorithm to find an element in $H[\mathbf{x}]$ and also the requirement that $m_0(\epsilon, \delta)$ should be polynomial in $1/\epsilon$ and $1/\delta$. The bound on $m_0$ for learnability given a finite Vapnik-Chervonenkis dimension certainly satisfies this second condition, but we do not address the problem of the algorithms needed to find hypotheses fitting a given sample.

We now give the result for finite $H$. Note that, we use ln to denote natural logarithm and log to denote logarithm to the base 2.

**Theorem 2.2 :** *If $H$ is a finite set of hypotheses, then $c \in H$ is learnable in context $H$ and a suitable value for $m_0$ is*

$$\frac{1}{\epsilon}\ln(\frac{|H|}{\delta}).$$

**Proof:** Let $B_\epsilon = \{h \in H | \mathrm{er}_\mu(h) > \epsilon\}$ and fix $h \in B_\epsilon$. First note that by the definition of actual error,

$$\mu\{x \in X | h(x) = c(x)\} < 1 - \epsilon.$$

But then for a sequence $\mathbf{x} = (x_1, \ldots, x_m)$ of independently selected samples the probability that $\mathrm{er}_{\mathbf{x}}(h) = 0$ can be bounded;

$$\mu^m\{\mathbf{x} \in X^m | \mathrm{er}_{\mathbf{x}}(h) = 0\} \leq (1 - \epsilon)^m.$$

Hence by the subadditivity of the probability measure the probability that at least one $h \in B_\epsilon$ has as low an error as this is given by

$$\mu^m\{\mathbf{x} \in X^m | \exists h \in B_\epsilon \text{ such that } \mathrm{er}_\mu(h) = 0\} < |B_\epsilon|(1 - \epsilon)^m.$$

Hence

$$\begin{aligned}\mu^m\{\mathbf{x} \in X^m | \mathrm{haz}[\mathbf{x}] \leq \epsilon\} &> 1 - |B_\epsilon|(1 - \epsilon)^m \\ &> 1 - |H|\exp(-\epsilon m).\end{aligned}$$

The result follows from setting this quantity greater than or equal to $1 - \delta$. ∎

## 3 Vapnik-Chervonenkis Dimension

We now consider generalising the result from finite sets of hypotheses to sets of hypotheses with finite Vapnik-Chervonenkis dimension.

To introduce the Vapnik-Chervonenkis dimension, it is useful to consider first the following number for a given sequence $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$:

$$\Pi_H(\mathbf{x}) = \left|\left\{(h(x_1), \ldots, h(x_m)) | h \in H\right\}\right| .$$

This is the number of different output sequences that can be obtained by applying all of the hypotheses to a fixed input $\mathbf{x}$. The maximum that can be obtained for all $m$-tuple inputs is a function $\Pi_H(m)$ of $m$:

$$\Pi_H(m) = \max_{\mathbf{x} \in X^m} \Pi_H(\mathbf{x}).$$

This function is called the *growth function* of $H$. Clearly $\Pi_H(m) \leq 2^m$, since there are only $2^m$ possible sequences. For a set $H$ of hypotheses the *Vapnik-Chervonenkis dimension*, denoted VCdim($H$) is defined as

$$\mathrm{VCdim}(H) = \begin{cases} \infty; & \text{if } \Pi_H(m) = 2^m \text{ for all } m; \\ \max\left\{m | \Pi_H(m) = 2^m\right\}; & \text{otherwise.} \end{cases}$$

For $m \leq \mathrm{VCdim}(H)$, we have by definition $\Pi_H(m) = 2^m$. The next Lemma gives a bound on the size of $\Pi_H(m)$ for $m > \mathrm{VCdim}(H)$. It can be found in [Haus88].

**Lemma 3.1 :** *If* $\mathrm{VCdim}(H) = d$ *and* $m \geq d \geq 1$, *then* $\Pi_H(m) \leq (em/d)^d$, *where* $e$ *is the base of the natural logarithm.*

We give a second Lemma which will prove useful later.

**Lemma 3.2 :** *For any* $\alpha > 0$, $\ln(x) \leq c + \alpha x$, *for* $x \geq 0$, *where* $c = \ln(1/\alpha) - 1$.

**Proof :** Consider the function $f(x) = c + \alpha x - \ln(x)$. Consider $f'(x) = \alpha - 1/x$. This will be positive for $x > 1/\alpha$ and negative for $x < 1/\alpha$. Setting $f(x) = 0$ for $x = 1/\alpha$ gives $c = \ln(1/\alpha) - 1$ as required. ∎

We are now ready to present our main theorem. The sample size bound will be given in a Corollary to this theorem.

**Theorem 3.3 :** *For a given hypothesis $c$ and set of hypotheses $H \subseteq \mathcal{C}(X)$ with finite Vapnik-Chervonenkis dimension $d > 1$, the probability that some function $h$ in $H$ which agrees with $c$ on $m$ independent random examples (chosen according to any fixed probability distribution $\mu$) has error $\mathrm{er}_\mu(h)$ greater than $\epsilon$ is less than*

$$2\left(\frac{em}{d}\right)^{2d}\epsilon^d e^{-\epsilon m} e^{2\sqrt{2d}},$$

*provided that $m \geq 4d/\epsilon$.*

**Proof :** Let $B_\epsilon = \{h \in H | \mathrm{er}_\mu(h) > \epsilon\}$. Following [HaWe] we define two subsets of vectors of points from $X$. The sets in our case are,

$$Q_\epsilon^m = \{\mathbf{x} \in X^m | \exists h \in B_\epsilon, \text{ such that } \mathrm{er}_\mathbf{x}(h) = 0\}$$

and

$$J_\epsilon^{m+k} = \{\mathbf{xy} \in X^{m+k} | \exists h \in B_\epsilon \text{ such that } \mathrm{er}_\mathbf{x}(h) = 0, \text{ and } \mathrm{er}_\mathbf{y}(h) > r\epsilon\}.$$

The parameters $k \geq 1$ and $r < 1$ have yet to be chosen. In [HaWe] these numbers are chosen to be $m$ and 0.5 respectively. Here we take them to satisfy

$$r = 1 - \sqrt{\frac{2}{\epsilon k}}$$
$$k = m\left(\frac{\epsilon r m}{d} - 1\right).$$

These two equations have a solution for some $k \geq m$, provided that $\epsilon m \geq 4d$. We will assume that $k$ is an integer and ignore the effects when this is not the case. As in [HaWe] the proof is divided into two stages proving the two inequalities,

$$\mu^m(Q_\epsilon^m) < 2\mu^{m+k}(J_\epsilon^{m+k})$$

and

$$\mu^{m+k}(J_\epsilon^{m+k}) \leq \left(\frac{em}{d}\right)^{2d}\epsilon^d e^{-\epsilon m} e^{2\sqrt{d}}.$$

The result will clearly follow.

**Stage 1:** This stage relies on Chebyshev's inequality to prove that for $h \in B_\epsilon$

$$\mu^k\{\mathbf{y} \in X^k | \mathrm{er}_\mathbf{y}(h) > r\epsilon\} > 0.5,$$

by showing that

$$\mu^k\{\mathbf{y} \in X^k | \mathrm{er}_\mathbf{y}(h) \leq r\epsilon\} < 0.5.$$

The expected number of indices for which $h(\mathbf{y}_i) \neq c(\mathbf{y}_i)$ is $pk$ where

$$p = \mathrm{er}_\mu(h) > \epsilon,$$

while the variance is $p(1-p)k$. Hence by Chebyshev the probability that of $k$ independent choices fewer than $r\epsilon k$ fall in this set is less than or equal to

$$\frac{p(1-p)k}{((p-r\epsilon)k)^2} < \frac{1}{(1-r)^2\epsilon k}$$

since $p > \epsilon$. Hence since

$$r = 1 - \sqrt{\frac{2}{\epsilon k}}$$

we have

$$\mu^k\{\mathbf{y} \in X^m | \mathrm{er}_{\mathbf{y}}(h) \leq r\epsilon\} < \frac{1}{(1-r)^2\epsilon k} = \frac{1}{2},$$

as required.

**Stage 2:** Consider the transformations

$$\pi_1, \ldots, \pi_{(m+k)!},$$

of the vectors $\mathbf{xy} \in X^{m+k}$ obtained by permuting the $m+k$ indices. We can sum the measure of the space $J_\epsilon^{m+k}$ over all of $(m+k)!$ copies of $X^{m+k}$ obtained by permuting the indices:

$$(m+k)!\mu^{m+k}(J_\epsilon^{m+k}) = \sum_{i=1}^{(m+k)!} \int_{X^{m+k}} \chi_{J_\epsilon^{m+k}}(\pi_i(\mathbf{xy}))d\mu^{m+k}(\mathbf{xy}).$$

Interchanging the summation and integration gives

$$(m+k)!\mu^{m+k}(J_\epsilon^{m+k}) = \int_{X^{m+k}} \sum_{i=1}^{(m+k)!} \chi_{J_\epsilon^{m+k}}(\pi_i(\mathbf{xy}))d\mu^{m+k}(\mathbf{xy}).$$

The proof now involves finding a bound on the inner sum

$$\sum_{i=1}^{(m+k)!} \chi_{J_\epsilon^{m+k}}(\pi_i(\mathbf{xy}))$$

which is independent of the vector $\mathbf{xy}$. This is a matter of counting for a particular $\mathbf{xy}$ how many rearrangements of it lie in $J_\epsilon^{m+k}$. If we can bound this by a fraction $\theta$ of all $(m+k)!$ rearrangements, we can then bound

$$\mu^{m+k}(J_\epsilon^{m+k}) \leq \theta.$$

Consider $\mathbf{xy} \in X^{m+k}$ and the possible values $h(\mathbf{xy})$ for $h \in H$. By Lemma 3.1 there are at most

$$\Pi_H(m+k) \leq \left(\frac{e(m+k)}{d}\right)^d$$

such sequences since $H$ has finite VC dimension $d$. For each such sequence if the number of indices in which the sequence disagrees with $c$ is $\ell$ then the fraction of permutations which will place it in $J_\epsilon^{m+k}$ is given by

$$\begin{aligned}\frac{\binom{k}{\ell}}{\binom{m+k}{\ell}} &= \frac{k(k-1)\dots(k-\ell+1)}{(m+k)(m+k-1)\dots(m+k-\ell+1)} \leq \left(\frac{k}{m+k}\right)^{\ell} \\ &= \left(1-\frac{m}{m+k}\right)^{\ell} \\ &\leq e^{-\frac{m}{m+k}\ell}.\end{aligned}$$

Since if $\ell < r\epsilon k$ no rearrangement lies in $J_\epsilon^{m+k}$, we can bound the proportion of permutations by

$$e^{-\frac{m}{m+k}r\epsilon k}.$$

This is the fraction of permutations which were included for a particular sequence, hence the total number can be bounded by the sum over all possible sequences

$$\mu^{m+k}(J_\epsilon^{m+k}) \leq \left(\frac{e(m+k)}{d}\right)^{d} e^{-\frac{m}{m+k}r\epsilon k}.$$

Finally we substitute the value of $k$

$$k = m\left(\frac{\epsilon r m}{d} - 1\right)$$

giving

$$\mu^{m+k}(J_\epsilon^{m+k}) \leq \left(\frac{em}{d}\right)^{2d} \epsilon^d e^{-\epsilon m\left(1-\sqrt{\frac{2}{\epsilon k}}\right)}.$$

Consider the value of $r$. Since $k \geq m$ and $\epsilon m \geq 4d$, we have that

$$r \geq 1 - \sqrt{\frac{2}{4d}} = 1 - \frac{1}{\sqrt{2d}}.$$

But $d \geq 2$, and so $r \geq 0.5$. We now have

$$\begin{aligned}k = m\left(\frac{\epsilon r m}{d} - 1\right) &\geq \frac{\epsilon m^2}{2d} - m \\ &\geq \frac{\epsilon m^2}{4d}, \\ \text{since} \quad \frac{\epsilon m^2}{4d} - m &\geq m\,(1-1) \geq 0,\end{aligned}$$

again using $\epsilon m \geq 4d$. Hence

$$\mu^{m+k}(J_\epsilon^{m+k}) \leq \left(\frac{em}{d}\right)^{2d} \epsilon^d e^{-\epsilon m\left(1-2\sqrt{2d}/\epsilon m\right)}.$$

as required. ∎

**Corollary 3.4 :** *If $d = \mathrm{VCdim}(H) > 1$ is finite and $H$ can approximate $c$, then $c$ is learnable in context $H$ and a suitable $m_0$ is*

$$\frac{1}{\epsilon(1-\sqrt{\epsilon})}\left[2d\ln(6/\epsilon)+\ln(2/\delta)\right].$$

**Proof :** Since $H$ can approximate $c$, we must check only the existence of $m_0$. By the Theorem, for given $\epsilon$ and $\delta$, we must choose $m$ such that

$$2\left(\frac{em}{d}\right)^{2d}\epsilon^d e^{-\epsilon m}e^{2\sqrt{2d}} < \delta,$$

provided this is greater than or equal to $4d/\epsilon$. Taking logarithms and regrouping terms gives,

$$m\epsilon > d\left(2+2\ln(m)-2\ln(d)+\ln(\epsilon)+\frac{2\sqrt{2}}{\sqrt{d}}\right)+\ln(2/\delta).$$

By Lemma 3.2, if we choose

$$c = \ln(2)+\ln(d)+\ln(1/\epsilon)+\ln(1/\alpha)-1$$

for some $\alpha$ between 0 and 1, then $\ln(m) \le c + \alpha\epsilon m/2d$, and so it is sufficient to choose

$$m > \frac{1}{\epsilon(1-\alpha)}\left[d\left(\ln(1/\epsilon)+2\ln(2)+2\ln(1/\alpha)+\frac{2\sqrt{2}}{\sqrt{d}}\right)+\ln(2/\delta)\right].$$

Choosing $\alpha = \sqrt{\epsilon}$ gives the result since $\ln(2)+\sqrt{\frac{2}{d}} < \ln(6)$, for $d > 1$. This value for $m$ certainly satisfies the requirement that $m \ge 4d/\epsilon$. A more optimal choice for $\alpha$ is $(1+\ln(1/\epsilon))^{-1}$, but the expression generated is less readable. The optimal $\alpha$ for given $\epsilon$, ignoring the effect of the $2\sqrt{2}/\sqrt{d}$ term, is obtained by solving the recurrence

$$y = 1+\ln(y)+\ln(1/\epsilon)/2,$$

and setting $\alpha = 1/y$. ∎

The upper bound for $m$ given by Haussler in [Haus] is

$$\frac{4}{\epsilon}\left[2d\log_2(13/\epsilon)+\log_2(2/\delta)\right].$$

This upper bound for $m$ we have obtained is smaller than this bound by a factor larger than $4(1-\sqrt{\epsilon})\log_2(e)$.

## 4 References

[BEHW] : Anselm Blumer, Andrzej Ehrenfeucht, David Haussler and Manfred K. Warmuth, Occam's Razor, Information Processing Letters, 24 (1987) 377–380.

[Haus] : David Haussler, Quantifying Inductive Bias, AI Learning Algorithms and Valiant's Learning Framework, Artificial Intelligence, 36 (1988) 177–221.

[HaWe] : David Haussler and Emo Welzl, $\epsilon$-Nets and Simplex Range Queries, Discrete Comput. Geom., 2 (1987) 127–151.

[PiVa] : L. Pitt and L.G. Valiant, Computational Limits on Learning from Examples, Tech. Rept., Dept. of Computer Science, Harvard Univerisity, Cambridge, MA (1986); also: J. ACM, to appear.

[Vali] : L.G. Valiant, A Theory of the Learnable, Communications of the ACM, 27 (1984) 1134–1142.

[VaCe] : V.N. Vapnik and A.Ya. Chervonenkis, On the Uniform Convergence of Relative Frequencies of Events to their Probabilities, Theor. Probab. Appl., 16 (2) (1971) 264–280.