Extending Occam's Razor

Kevin S. Van Horn & Tony R. Martinez
3361 TMCB
Computer Science Department
Brigham Young University
Provo, UT 84602

email: kevin@bert.cs.byu.edu, martinez@cs.byu.edu

This paper will appear in the *Proceedings of the Third Golden West International Conference on Intelligent Systems*, Las Vegas, Nevada, 6 June 1994.

An extended version of this paper that also examines "loose Occam Algorithms" appears as Chapter 3 of

K. S. Van Horn
Learning as Optimization
Ph.D. dissertation, Computer Science Dept.
Brigham Young University, Provo, UT
July 1994

EXTENDING OCCAM'S RAZOR

KEVIN S. VAN HORN & TONY R. MARTINEZ
Computer Science Department
Brigham Young University
Provo, UT 84602

Abstract: Occam's Razor states that, all other things being equal, the simpler of two possible hypotheses is to be preferred. A quantified version of Occam's Razor has been proven for the PAC model of learning, giving sample-complexity bounds for learning using what Blumer et al. call an Occam algorithm [1]. We prove an analog of this result for Haussler's more general learning model, which encompasses learning in stochastic situations, learning real-valued functions, etc.

1. Introduction

Work in machine learning often makes use of the Occam's Razor principle: given two explanations of the data, all other things being equal, the simpler of the two is preferable. Occam's Razor can be applied once we have a measure of the complexity of a hypothesis. Such a measure is obtained by choosing some means of representing hypotheses, and defining the complexity of a hypothesis in terms of the size of its smallest representation.

Blumer et al. [1] have formalized this intuition for a restricted model of learning. They show that the problem of learning under the PAC model, with a hypothesis space of large or infinite VC dimension, can be solved with an *Occam algorithm*: an approximate minimization algorithm that finds a near-simplest hypothesis correctly classifying all the training examples. The PAC model, however, has a number of limitations. It assumes that the problem is to learn the correct classification of instances, and that some member of the given hypothesis space correctly classifies all instances. This rules out problems such as learning real-valued functions, probability distributions, class probability distributions as a function of the instance to be classified, and the Bayes-optimal classifier in a stochastic setting.

Haussler [3] has generalized the PAC model to deal with these situations and others. In this paper we prove, for Haussler's model, an analog of the Occam algorithm result. The approach we analyze is essentially the "hold-out" method often used in applied statistics. That is, one tries to minimize the empirical risk (a measure of error on the training sample) with different bounds on the complexity

of the hypotheses to be considered, then uses a separate hold-out set to choose the best complexity bound. As in [1], attention is paid to avoiding exact minimization and its attendant intractability — a limited increase in hypothesis complexity is allowed over what is strictly needed to attain a given level of empirical risk. We obtain sample complexity bounds similar to those in [1], but for Haussler's more general learning model.

2. Background

2.1. The PAC model and Occam's Razor

The PAC (probably approximately correct) model of learning [1, 4–6] is a simplified learning model first introduced by Valiant [6]. The elements of the model are an *instance space* X and a "stratified" *hypothesis space* $H = (H_i)_{i \geq 1}$, where $H_i \subseteq H_{i+1}$ and we write H for $\bigcup_i H_i$. The elements of H are functions from X to $\{0, 1\}$. H_n can be thought of as the set of all hypotheses of size at most n, where size is defined in terms of some means of representing the elements of H.

In the PAC model it is assumed that there is some unknown *target* $f \in H$ which determines the classification of instances, and some unknown distribution D governing the frequency of occurrence of instances. The *error* of a hypothesis $h \in H$ is the probability that $h(x) \neq f(x)$ (h misclassifies x) when x is drawn at random from the distribution D. The goal of learning is to find a hypothesis with small error.

An algorithm for learning H takes as input a sequence of *training examples* $(x_i, f(x_i))$, where the x_i are randomly and independently chosen according to D, and outputs a hypothesis $h \in H$. Its *sample complexity* $m(\epsilon, \delta, n)$ is the worst-case number of examples needed to have confidence $1-\delta$ of returning a hypothesis with error at most ϵ, when the target f may be any element of H_n and D may be any distribution over X.

Blumer et al. [1] discuss the Occam's Razor approach to learning: find a near-smallest *consistent* hypothesis (a consistent hypothesis correctly classifies all m training examples). Their result is framed in terms of the *VC dimension* of H_i, a combinatorial parameter that is generally related to the number of bits or parameters required to specify an arbitrary hypothesis in H_i (see [1]). Suppose we have an algorithm A, functions $p(s)$ and $b(s, m)$, and constant a satisfying the following:

1. $0 \leq a < 1$;
2. $p(s) \geq 1$ and $p(s)$ is bounded by a polynomial in s;
3. the VC dimension of $H_{b(s,m)}$ is at most $p(s)m^a$;
4. given m training examples as input, A outputs a consistent hypothesis $h \in H$ of size at most $b(s, m)$, where s is the size of the smallest consistent hypothesis.

Then A is called an *Occam algorithm*. Blumer et al. show that any Occam algorithm serves as a learning algorithm with sample complexity

$$O\left(\epsilon^{-1}\log\delta^{-1} + (p(s)\epsilon^{-1}\log\epsilon^{-1})^{\frac{1}{1-\alpha}}\right).$$

Note that exact minimization of the hypothesis size is not needed — even a very weak approximation algorithm will do. This is important, because exact minimization is often NP-hard.

2.2. Haussler's Extension of the PAC Model

Haussler [3] has extended the PAC model to handle more general learning situations. The elements of his model are an *instance space* X, an *outcome space* Y, a *decision space* Y', a *hypothesis space* H, an *assumption space* $\mathcal{A}$, and a *loss function* $l : Y' \times Y \to [0, M]$ (for some $M > 0$). The elements of H are functions $h : X \to Y'$, training examples are from $X \times Y$, and the elements of $\mathcal{A}$ are probability distributions over $X \times Y$. $l(y', y)$ is the loss incurred when a hypothesis outputs y' for an instance x whose outcome is y. The PAC model can be considered a special case of this model, with $Y = Y' = \{0, 1\}$, $l(y', y) = 1$ if $y' \neq y$ and 0 otherwise, and $\mathcal{A}$ being the set of all distributions over $X \times Y$ satisfying $\mathbf{Pr}[y = f(x)] = 1$ for some $f \in H$.

Similar to the PAC model, it is assumed that there is some unknown *target* $D \in \mathcal{A}$ which determines the frequency of occurrence of instances $x \in X$ and their outcomes $y \in Y$. The *risk* $\mathbf{r}(h)$ of a hypothesis h is the expected value of $l(h(x), y)$ when (x, y) is drawn at random from the distribution D. A learning algorithm takes training examples drawn from this same distribution. The goal of learning is to find a hypothesis whose error is close to the minimum possible for hypotheses from H.

Haussler defines sample complexity in terms of the following family of metrics on the real numbers: for any real $\nu > 0$, d_ν is a measure of relative distance defined by

$$d_\nu(r, s) = \frac{|r - s|}{\nu + r + s}$$

for any $r, s \geq 0$. The condition $d_\nu(r, s) \leq \alpha$ is equivalent to

$$\frac{1-\alpha}{1+\alpha}r - \frac{\alpha\nu}{1+\alpha} \leq s \leq \frac{1+\alpha}{1-\alpha}r + \frac{\alpha\nu}{1-\alpha};$$

for small α this says that s differs from r by at most a multiplicative factor of about $1 + 2\alpha$ and an additive term of $\alpha\nu$. The sample complexity $m(\alpha, \nu, \delta)$ of a learning algorithm is then the worst-case number of examples needed to have confidence $1 - \delta$ of returning a hypothesis h satisfying

$$d_\nu(\mathbf{r}(h), \inf\{\mathbf{r}(h') : h' \in H\}) \leq \alpha,$$

when the target distribution may be any $D \in \mathcal{A}$.

To our knowledge the literature contains no analog, for Haussler's model, of Blumer et al.'s Occam algorithm result. (Nor does it contain such an analog for the simpler PAC extension of just allowing stochastic targets.) However, Haussler gives useful sample complexity results for the case that H has a finite *pseudodimension*. The pseudodimension of H, denoted $\mathrm{pdim}(H)$, depends on the loss function l and may be considered a generalization of the VC dimension; in fact, $\mathrm{pdim}(H) = \mathrm{VCdim}(H)$ if the hypotheses of H are Boolean-valued and $l(y', y) = (y' \neq y)$. (Note: in this paper we abuse terminology by writing $\mathrm{pdim}(H)$ when we really mean $\mathrm{pdim}\{f_h : h \in H\}$, where $f_h(x, y) = l(h(x), y)$.) Details on the pseudodimension may be found in [3].

Given a sequence of examples $\bar{z} = (x_1, y_1), \ldots, (x_m, y_m)$, the *empirical risk* $\hat{\mathbf{r}}(h; \bar{z})$ of h is the average loss of h on $\bar{z}$, i.e. $\hat{\mathbf{r}}(h; \bar{z}) = \frac{1}{m}\sum_{i=1}^{m} l(h(x_i), y_i)$. We write $\hat{\mathbf{r}}(h)$ when $\bar{z}$ is understood. A learning algorithm is said to use *empirical risk minimization* [3, 7, 8] if it returns a hypothesis whose empirical risk is the minimum possible for hypotheses from H.

Combining Lemma 1 and Theorem 7 of Haussler [3] gives this result: any learning algorithm that uses empirical risk minimization has a sample complexity that is

$$O((\alpha^2\nu)^{-1}(\mathrm{pdim}(H)\log(\alpha\nu)^{-1} + \log\delta^{-1})). \tag{1}$$

3. Summary of Results

We begin with some definitions.

Definition 1 *We say that hypothesis class H is* stratified by dimension *if $H = (H_i)_{i\geq 1}$, where $H_i \subseteq H_{i+1}$ and $\mathrm{pdim}(H_i) \leq i$ for all i. We define the* size *of $h \in H$, denoted* $\mathrm{siz}(h)$, *to be* $\min\{i : h \in H_i\}$.

Definition 2 *Let $p(s) \geq 1$ be monotonic and polynomially bounded in s, and let $0 \leq a < 1$. We say that A is a* loose ERM (empirical risk minimization) algorithm for $H = (H_i)_{i\geq 1}$ with bound $p(s)m^a$ *if*

- *A takes as input a sequence of examples $\bar{z}$ and an integer* size bound *i satisfying $1 \leq i \leq m \stackrel{\mathrm{def}}{=} |\bar{z}|$;*
- *A outputs a hypothesis $h \in H_i$ satisfying $\hat{\mathbf{r}}(h; \bar{z}) \leq \hat{\mathbf{r}}(h'; \bar{z})$ for all $h' \in H$ s.t. $p(\mathrm{siz}(h'))m^a < i + 1$.*

Thus a loose ERM algorithm loosens the requirements of empirical risk minimization in hopes of making the problem tractable. Associated with each loose ERM algorithm for H is a learning algorithm for H:

Definition 3 *Let A be a loose ERM algorithm for H, $0 < c < 1$, and $0 < \rho \leq 1$; then $HO[A, c, \rho]$ is the following algorithm:*

1. *Input a sequence of examples* $\bar{z}$.

2. *Split* $\bar{z}$ *into two sequences:* $\bar{z}_1$ *(the* training sample*), containing the first* $\lfloor |\bar{z}|/(1+\rho) \rfloor$ *elements of* $\bar{z}$*, and* $\bar{z}_2$ *(the* test sample*), containing the remaining elements of* $\bar{z}$*. Let* $m \stackrel{\text{def}}{=} |\bar{z}_1|$.

3. *For all* $1 \leq i \leq \lfloor \log m / \log(1/c) \rfloor$*, compute* $h_i \stackrel{\text{def}}{=} A(\bar{z}_1, \lfloor mc^i \rfloor)$.

4. *Output that* h_i *minimizing* $\hat{\mathbf{r}}(h_i; \bar{z}_2)$.

The above is essentially the "hold-out" method often used in applied statistics (see [2], for example). Our contribution is to analyze the sample complexity of $\mathrm{HO}[A, c, \rho]$ given the weakened requirement that A be a loose ERM algorithm (as opposed to doing strict empirical risk minimization), and show that we get sample complexity bounds analogous to those for Blumer et al.'s Occam algorithms.

We modify Haussler's definition of sample complexity to take into account hypothesis complexity by adding a size parameter n. We also need to take into account the possibility that the hypothesis returned has size greater than n, and hence might have a risk less than the minimum achievable by hypotheses from H_n.

Definition 4 $d'_\nu(r,s) \stackrel{\text{def}}{=} d_\nu(r, \max\{r,s\})$.

Definition 5 *The sample complexity* $m(\alpha, \nu, \delta, n)$ *of a learning algorithm for* $H = (H_i)_{i \geq 1}$ *is the worst-case number of examples needed to have confidence* $1-\delta$ *of returning a hypothesis* h *satisfying*

$$d'_\nu(\inf\{\mathbf{r}(h') : h' \in H_n\}, \mathbf{r}(h)) \leq \alpha$$

when the target distribution may be any $D \in \mathcal{D}$.

Our result is that if H is stratified by dimension and A is a loose ERM algorithm for H with bound $p(s)m^a$, then $\mathrm{HO}[A, c, \rho]$ is a learning algorithm for H with sample complexity

$$O((\alpha^2\nu)^{-1} \ln \delta^{-1} + ((\alpha^2\nu)^{-1} p(n) \ln(\alpha\nu)^{-1})^{\frac{1}{1-a}}).$$

4. Proof of Result

Our proof will require some theorems from Haussler and two lemmas. We will assume throughout that M is a bound on the loss function for whatever hypothesis class H is being discussed. The theorems from Haussler give sample complexity bounds for hypothesis classes of finite cardinality or finite pseudodimension.

Theorem 1 *(Theorem 1 of [3].) Let H be a finite set of hypotheses; let a sequence $\bar{z}$ of m examples be drawn randomly and independently from the distribution D; and let $\nu > 0$ and $0 < \alpha < 1$. Then*

$$\mathbf{Pr}[\exists h \in H.\, d_\nu(\hat{\mathbf{r}}(h;\bar{z}), \mathbf{r}(h)) > \alpha] \leq 2|H| \exp(-\alpha^2 \nu m/M).$$

For $\delta > 0$ and $m \geq M(\alpha^2\nu)^{-1}(\ln|H| + \ln(2/\delta))$ this probability is at most δ.

Theorem 2 *(Theorem 7 of [3].) Let H be a set of hypotheses with $\mathrm{pdim}(H) = n$ for some $1 \leq n < \infty$; let a sequence $\bar{z}$ of $m \geq 1$ examples be drawn randomly and independently from the distribution D; and let $0 < \nu \leq 8M$ and $0 < \alpha < 1$. Then*

$$\mathbf{Pr}[\exists h \in H.\, d_\nu(\hat{\mathbf{r}}(h;\bar{z}), \mathbf{r}(h)) > \alpha] \leq 8 \left(\frac{16eM}{\alpha\nu} \ln \frac{16eM}{\alpha\nu}\right)^n \exp\left(-\frac{\alpha^2\nu m}{8M}\right).$$

The following lemma is used to bound the risk of the hypothesis output by a loose ERM algorithm. In what follows we write $\mathrm{lln}(x)$ for $\ln(x \ln x)$. Note that $\mathrm{lln}(x) = O(\ln x)$.

Lemma 3 *Let m be a positive integer and $0 \leq a < 1$; let H be a set of hypotheses with $\mathrm{pdim}(H) \leq nm^a$; let a sequence $\bar{z}$ of m examples be drawn randomly and independently from the distribution D; and let $0 < \alpha < 1$, $\delta > 0$ and $0 < \nu \leq 8M$. Then*

$$\mathbf{Pr}[\exists h \in H.\, d_\nu(\hat{\mathbf{r}}(h;\bar{z}), \mathbf{r}(h)) > \alpha] \leq \delta$$

whenever

$$m \geq \max\left\{\left(\frac{16M}{\alpha^2\nu} n \,\mathrm{lln}\, \frac{16eM}{\alpha\nu}\right)^{\frac{1}{1-a}},\ \frac{16M}{\alpha^2\nu} \ln\frac{8}{\delta}\right\}. \tag{2}$$

Proof. By Theorem 2 it suffices to show that

$$8\left(\frac{16eM}{\alpha\nu} \ln \frac{16eM}{\alpha\nu}\right)^{nm^a} \exp\left(-\frac{\alpha^2\nu m}{8M}\right) \leq \delta$$

when the bound (2) on m holds. Taking logarithms of both sides of the above inequality and rearranging yields

$$m \geq \frac{8M}{\alpha^2\nu}\left(nm^a \,\mathrm{lln}\, \frac{16eM}{\alpha\nu} + \ln\frac{8}{\delta}\right). \tag{3}$$

From the bound (2) on m we have that

$$\frac{m}{2} \geq \frac{8M}{\alpha^2\nu} \ln\frac{8}{\delta};$$

thus (3) will be satisfied if

$$\frac{m}{2} \geq Bm^a \quad \text{where} \quad B \stackrel{\text{def}}{=} \frac{8M}{\alpha^2\nu} n \,\mathrm{lln}\, \frac{16eM}{\alpha\nu},$$

which can be rewritten as $m^{1-a} \geq 2B$ and then as $m \geq (2B)^{\frac{1}{1-a}}$. This latter inequality follows directly from (2). □

The next lemma is used to bound the error incurred in step 4 of HO$[A, c, \rho]$.

Lemma 4 *Let $\rho, C, m > 0$; let H be a set of $\lfloor C \ln m \rfloor$ hypotheses; let a sequence $\bar{z}$ of at least ρm examples be drawn randomly and independently from the distribution D; and let $\alpha^2 \nu \leq M/(3\rho)$; then*

$$\mathbf{Pr}[\exists h \in H.\ d_\nu(\hat{\mathbf{r}}(h; \bar{z}), \mathbf{r}(h)) > \alpha] \leq \delta$$

whenever

$$m \geq \frac{2M}{\alpha^2 \nu \rho} \max \left\{ 0.44 + \ln \ln \frac{2M}{\alpha^2 \nu \rho},\ \ln \frac{2C}{\delta} \right\} \tag{4}$$

Proof. By Theorem 1, the above-mentioned probability will be at most δ if

$$\rho m \geq \frac{M}{\alpha^2 \nu} (\ln(C \ln m) + \ln(2/\delta)),$$

which can be rewritten as

$$m \geq \frac{M}{\alpha^2 \nu \rho} (\ln \ln m + \ln(2C/\delta)).$$

This inequality in turn will hold if

$$m \geq \frac{2M}{\alpha^2 \nu \rho} \ln \frac{2C}{\delta} \tag{5}$$

and

$$m \geq B \ln \ln m \ \text{ where } \ B \stackrel{\text{def}}{=} \frac{2M}{\alpha^2 \nu \rho}. \tag{6}$$

From the bound (4) on m we see that (5) holds, so it remains only to show that (6) holds. We shall use the fact that $B \geq 6$, since $\alpha^2 \nu \leq M/(3\rho)$ from the statement of the lemma.

Let $f(b) \stackrel{\text{def}}{=} b \ln(1.55 \ln b)$. Since $0.44 > \ln 1.55$, we have by (4) that $m > f(B)$. In fact, $m = f(b')$ for some $b' > B$, since $f(b)$ is continuous and increasing for $b > 1$, and $f(b) \to \infty$ as $b \to \infty$. Thus (6) holds if $f(b) \geq B \ln \ln f(b)$ for all $b \geq B$; this in turn holds if

$$f(b) \geq b \ln \ln f(b) \ \text{ for all } \ b \geq 6 \tag{7}$$

Defining $g(b) \stackrel{\text{def}}{=} b^{0.55} / \ln(1.55 \ln b)$, for all $b \geq 6$ we have

$$\begin{aligned} & f(b) \geq b \ln \ln f(b) \ \Leftrightarrow \ 1.55 \ln b \geq \ln(b \ln(1.55 \ln b)) \\ & \Leftrightarrow \ b^{1.55} \geq b \ln(1.55 \ln b) \ \Leftrightarrow \ g(b) \geq 1. \end{aligned}$$

Writing $\mathrm{sgn}(x)$ for the sign of x and noting that $1.55 \ln b > 1$ for $b \geq 6$, we have that

$$\begin{aligned} \mathrm{sgn}(dg/db) &= \mathrm{sgn}(\ln(1.55 \ln b) \cdot 0.55 b^{-0.45} - b^{0.55} \cdot (1.55 \ln b)^{-1} 1.55 b^{-1}) \\ &= \mathrm{sgn}(s(b)) \text{ where } s(b) \stackrel{\mathrm{def}}{=} 0.55 \ln(1.55 \ln b) - (\ln b)^{-1}. \end{aligned}$$

But $s(6) > 0$ and $s(b)$ is an increasing function of b, so $s(b) > 0$ for all $b \geq 6$. Hence $dg/db > 0$ for all $b \geq 6$. But $g(6) \simeq 2.6 > 1$, hence $g(b) > 1$ for all $b \geq 6$. Thus (7) holds. □

We now mention some properties of d_ν and d'_ν. Haussler proves the following properties of d_ν:

1. d_ν is a metric on the nonnegative reals, i.e. $d_\nu(r,s) \geq 0$, $d_\nu(r,s) = 0$ iff $r = s$, $d_\nu(r,s) = d_\nu(s,r)$, and $d_\nu(r,t) \leq d_\nu(r,s) + d_\nu(s,t)$ (triangle inequality.)
2. d_ν is compatible with the ordering on the reals, i.e. if $r \leq s \leq t$ then $d_\nu(r,s) \leq d_\nu(r,t)$ and $d_\nu(s,t) \leq d_\nu(r,t)$.

We state without proof the following easily-verified properties of d'_ν:

1. if $r_2 \leq r_1$ then $d'_\nu(r_1, r_2) = 0$;
2. $0 \leq d'_\nu(r_1, r_2) \leq d_\nu(r_1, r_2)$;
3. $d'_\nu(r_1, r_3) \leq d'_\nu(r_1, r_2) + d'_\nu(r_2, r_3)$ (triangle inequality.)

Finally, we are ready for the main theorem.

Theorem 5 *Let $H = (H_i)_{i \geq 1}$ be a hypothesis space stratified by dimension; let n be a positive integer and $r_0 = \inf\{\mathbf{r}(h) : h \in H_n\}$; let A be a loose ERM algorithm for H with bound $p(s)m^a$; let $0.05 \leq c < 1$ and $0 < \rho \leq 1$; let $\bar{z}$ be a sequence of m examples drawn randomly and independently from the distribution D; let $h_{out} = HO[A, c, \rho](\bar{z})$; and let $0 < \delta \leq 1$, $0 < \alpha < 1$, and $0 < \nu \leq 8M$. Then*

$$\mathbf{Pr}[d'_\nu(r_0, \mathbf{r}(h_{out})) > \alpha] \leq \delta \tag{8}$$

whenever $m \geq (1+\rho)(B+1)$, where

$$B = \max\left\{ \frac{K}{\rho}\left(0.44 + \ln\ln\frac{K}{\rho}\right), \frac{K}{\rho}\ln\frac{2(C+6)}{\delta}, \left(\frac{K}{c}p(n)\,\mathrm{lln}\,\frac{81M}{\alpha\nu}\right)^{\frac{1}{1-a}} \right\},$$

$K = 55M/(\alpha^2\nu)$, and $C = 1/\ln(1/c)$.

Before giving the proof, we have a few comments. Recall that $\mathrm{HO}[A, c, \rho]$ runs A with various size bounds of the form $\lfloor mc^i \rfloor$, where i goes from 1 to a maximum value. The requirement $0.05 \leq c < 1$ merely says that the size bound decreases as i increases, but not by more than a factor of 20 at each step. Recall also that A

is run on the first $\lfloor |\bar{z}|/(1+\rho) \rfloor$ examples, with the remaining examples used to choose the best size bound. Thus the requirement $0 < \rho \leq 1$ says that at least (about) half — but not all — of the examples are used as input to A. Finally, note that a, c, ρ, and M are constants, $\mathrm{lln}(x) = O(\ln x)$, and

$$\ln \ln \frac{55M}{\alpha^2 \nu \rho} = O(\ln \frac{1}{\alpha\nu}),$$

so that $B = O((\alpha^2\nu)^{-1} \ln \delta^{-1} + ((\alpha^2\nu)^{-1} p(n) \ln(\alpha\nu)^{-1})^{\frac{1}{1-a}})$.

Proof. HO$[A, c, \rho]$ will be run with $m_* = \lfloor (1+\rho)^{-1} m \rfloor$ training examples and $m_{**} = m - m_*$ test examples. Since $m \geq (1+\rho)(B+1)$, we have that

$$m_* \geq \lfloor (1+\rho)^{-1}(1+\rho)(B+1) \rfloor = \lfloor B+1 \rfloor > B$$

and

$$m_{**} = m - m_* \geq m - \frac{m}{1+\rho} = \frac{\rho}{1+\rho} m \geq \rho \lfloor (1+\rho)^{-1} m \rfloor = \rho m_*.$$

Thus it will suffice to show that (8) holds whenever steps 3 and 4 of HO$[A, c, \rho]$ are run with $m_* > B$ training examples and $m_{**} \geq \rho m_*$ test examples. We write $\bar{z}_1$ for the training examples and $\bar{z}_2$ for the test examples.

Viewing B as a function of α, we have that $B(\alpha)$ is continuous and decreasing in α, and $B(\alpha) \to \infty$ as $\alpha \to 0$; hence $m_* > B(\alpha)$ implies that $m_* = B(\alpha_*)$ for some $\alpha_* < \alpha$. By the definition of r_0 as an infimum, there are hypotheses $h \in H_n$ with risk arbitrarily close to r_0; in particular, there exists some $h_* \in H_n$ with $d'_\nu(r_0, \mathbf{r}(h_*)) \leq \alpha - \alpha_*$. By the triangle inequality for d'_ν, it then suffices to show that

$$\mathbf{Pr}[d'_\nu(\mathbf{r}(h_*), \mathbf{r}(h_{out})) > \alpha_*] \leq \delta \tag{9}$$

whenever steps 3 and 4 of HO$[A, c, \rho]$ are run with $m_* = B(\alpha_*)$ training examples and $m_{**} \geq \rho m_*$ test examples.

It will be useful to have a simpler lower bound on m_*. By the requirements that $\nu \leq 8M$ and $\alpha, c < 1$ we have that $55M/(\alpha_*^2 \nu c) > 55/8$ and $\mathrm{lln}(81M/(\alpha\nu)) > \mathrm{lln}(81/8)$; hence using $m_* = B(\alpha_*)$ we have

$$m_* > \left(\frac{55}{8} p(n) \,\mathrm{lln} \frac{81}{8} \right)^{\frac{1}{1-a}} > (21.6\, p(n))^{\frac{1}{1-a}} \tag{10}$$

Let us define the following:

- $h_i = A(\bar{z}_1, \lfloor m_* c^i \rfloor)$ for $1 \leq i \leq \lfloor C \ln m_* \rfloor$, as in the definition of HO$[A, c, \rho]$.
- $j = \max\{i \in \mathcal{Z} : m_* c^i \geq p(n) m_*^a\}$.
- $n' = \lfloor m_* c^j \rfloor$.

We wish to ensure that $1 \leq j \leq \lfloor C \ln m_* \rfloor$, so that $h_j = A(\bar{z}_1, n')$ will be one of the hypotheses considered in step 4 of HO$[A, c, \rho]$. We will have $j \geq 1$ if

$m_* c \geq p(n) m_*^a$, i.e. $c \geq p(n) m_*^{a-1}$. From (10) and the fact that $a < 1$ we have that

$$p(n) m_*^{a-1} < p(n) (21.6\, p(n))^{\frac{a-1}{1-a}} = 21.6^{-1};$$

but $c \geq 0.05$ (from the statement of the theorem) $> 21.6^{-1} > p(n) m_*^{a-1}$, so $j \geq 1$. For the upper bound on j, note that since $m_* c^j \geq p(n) m_*^a$, we have

$$j \leq \ln(p(n) m_*^{a-1}) / \ln c = \ln(p(n)^{-1} m_*^{1-a}) / \ln(1/c) \leq \ln m_* / \ln(1/c)$$

(the last inequality uses $p(n) \geq 1$, $m_* > 1$, and $a \geq 0$); but since j is an integer we have

$$j \leq \lfloor \ln m_* / \ln(1/c) \rfloor = \lfloor C \ln m_* \rfloor.$$

By the triangle inequality for d'_ν, we will have $d'_\nu(\mathbf{r}(h_*), \mathbf{r}(h_{out})) \leq \alpha_*$ if the following hold for appropriate positive values a_i summing to 1:

1. $d'_\nu(\mathbf{r}(h_*), \hat{\mathbf{r}}(h_*; \bar{z}_1)) \leq a_1 \alpha_*$;
2. $d'_\nu(\hat{\mathbf{r}}(h_*; \bar{z}_1), \hat{\mathbf{r}}(h_j; \bar{z}_1)) = 0$;
3. $d'_\nu(\hat{\mathbf{r}}(h_j; \bar{z}_1), \mathbf{r}(h_j)) \leq a_2 \alpha_*$;
4. $d'_\nu(\mathbf{r}(h_j), \hat{\mathbf{r}}(h_j; \bar{z}_2)) \leq a_3 \alpha_*$;
5. $d'_\nu(\hat{\mathbf{r}}(h_j; \bar{z}_2), \hat{\mathbf{r}}(h_{out}; \bar{z}_2)) = 0$;
6. $d'_\nu(\hat{\mathbf{r}}(h_{out}; \bar{z}_2), \mathbf{r}(h_{out})) \leq a_4 \alpha_*$.

Thus we prove (9) by showing that conditions 2 and 5 always hold, and that the following hold for appropriate positive values f_i summing to 1:

$$\begin{array}{lcl} \mathbf{Pr}[\text{Condition 1 doesn't hold}] & \leq & f_1 \delta \\ \mathbf{Pr}[\text{Condition 3 doesn't hold}] & \leq & f_2 \delta \\ \mathbf{Pr}[\text{Condition 4 doesn't hold}] & \leq & f_3 \delta \\ \mathbf{Pr}[\text{Condition 6 doesn't hold}] & \leq & f_4 \delta. \end{array} \tag{11}$$

By definition, $\hat{\mathbf{r}}(h_{out}; \bar{z}_2) \leq \hat{\mathbf{r}}(h_j; \bar{z}_2)$, so condition 5 holds. Since $h_* \in H_n$ we have $\mathrm{siz}(h_*) \leq n$, hence using the monotonicity of p and the definition of j,

$$p(\mathrm{siz}(h_*)) m_*^a \leq p(n) m_*^a \leq m_* c^j < \lfloor m_* c^j \rfloor + 1 = n' + 1.$$

But $h_j = A(\bar{z}_1, n')$ and A is a loose ERM algorithm, so $\hat{\mathbf{r}}(h_j; \bar{z}_1) \leq \hat{\mathbf{r}}(h_*; \bar{z}_1)$, and condition 2 holds.

We now specify the values f_i summing to 1 and a_i summing to 1:

- $f_1 = f_3 = (C+6)^{-1}$, $f_2 = 4(C+6)^{-1}$, and $f_4 = C(C+6)^{-1}$.
- $a_1 = a_3 = (6+\sqrt{2})^{-1}$, $a_2 = 4(6+\sqrt{2})^{-1}$, and $a_4 = \sqrt{2}(6+\sqrt{2})^{-1}$.

Using $m_* = B(\alpha_*)$ and $55 > (6+\sqrt{2})^2$ we obtain

$$m_* \geq \frac{M}{a_1^2 \alpha_*^2 \nu \rho} \ln \frac{2}{f_1 \delta}.$$

Now apply Theorem 1 with a singleton hypothesis set, $a_1\alpha_*$ for α and $f_1\delta$ for δ; then using the facts that $\rho \leq 1$ and $d'_\nu(r,s) \leq d_\nu(r,s)$ we obtain

$$\mathbf{Pr}[\text{Condition 1 doesn't hold}] \leq f_1\delta.$$

Using $a_1 = a_3$ and $f_1 = f_3$ we also obtain

$$m_* \geq \frac{M}{a_3^2 \alpha_*^2 \nu \rho} \ln \frac{2}{f_3 \delta}.$$

Since the choice of h_j is independent of $\bar{z}_2$, and $|\bar{z}_2| = m_{**} \geq \rho m_*$, we can again apply Theorem 1 to obtain

$$\mathbf{Pr}[\text{Condition 4 doesn't hold}] \leq f_3\delta.$$

Using $m_* = B(\alpha_*)$, $55 > (6+\sqrt{2})^2$, $81 > 4e(6+\sqrt{2})$, and $\rho \leq 1$, we obtain

$$m_* \geq \max\left\{ \left(\frac{16M}{a_2^2 \alpha_*^2 \nu} c^{-1} p(n) \operatorname{lln} \frac{16eM}{a_2 \alpha_* \nu} \right)^{\frac{1}{1-a}},\ \frac{16M}{a_2^2 \alpha_*^2 \nu} \ln \frac{8}{f_2 \delta} \right\}$$

Referring back to the definitions of j and n', we note that $m_* c^{j+1} < p(n) m_*^a$ (recall that $0 < c < 1$), hence

$$n' = \lfloor m_* c^j \rfloor \leq m_* c^j < c^{-1} p(n) m_*^a.$$

Then applying Lemma 3 with this upper bound on n' and the preceding lower bound on m_* we obtain

$$\mathbf{Pr}[\exists h \in H_{n'} .\, d_\nu(\hat{\mathbf{r}}(h; \bar{z}_1), \mathbf{r}(h)) > a_2\alpha_*] \leq f_2\delta.$$

Since $h_j = A(\bar{z}_1, n')$ and hence $h \in H_{n'}$, we then obtain

$$\mathbf{Pr}[\text{Condition 3 doesn't hold}] \leq f_2\delta.$$

Finally, we look at condition 6. In step 4 of the algorithm we look at $k \stackrel{\text{def}}{=} \lfloor C \ln m_* \rfloor$ hypotheses, which are tested on $m_{**} \geq \rho m_*$ examples. Using the facts that $\nu \leq 8M$ and $\alpha_* < 1$, we have

$$(a_4\alpha_*)^2 \nu < \frac{2}{(6+\sqrt{2})^2} 8M < M/3 \leq M/(3\rho).$$

Furthermore, using $m_* = B(\alpha_*)$ and $55 > (6+\sqrt{2})^2$ we obtain

$$m_* \geq \frac{2M}{a_4^2 \alpha_*^2 \nu \rho} \max\left\{ 0.44 + \ln\ln \frac{2M}{a_4^2 \alpha_*^2 \nu \rho},\ \ln \frac{2C}{f_4 \delta} \right\}.$$

Thus the conditions of Lemma 4 hold; applying this lemma, we obtain

$$\mathbf{Pr}[\exists 1 \leq i \leq k.\ d_\nu(\hat{\mathbf{r}}(h_i; \bar{z}_2), \mathbf{r}(h_i)) > a_4\alpha_*] \leq f_4\delta.$$

and hence $\mathbf{Pr}$[Condition 6 doesn't hold] $\leq f_4\delta$.

Thus we have proven that the inequalities (11) hold, which completes the proof of (9), and hence of the theorem itself. □

References

[1] Blumer, A., et al. (1989.) Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM* 36, 929–965.

[2] Devroye, L. (1988.) Automatic pattern recognition: a study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10, 530–543.

[3] Haussler, D. (1990.) Decision theoretic generalizations of the PAC model for neural net and other learning applications. Technical report UCSC-CRL-91-02, Baskin Center for Computer Engineering and Information Sciences, UC Santa Cruz.

[4] Kearns, M. (1990.) *The Computational Complexity of Machine Learning.* Cambridge, MA: MIT Press.

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

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

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

[8] Vapnik, V. N. (1989.) Inductive principles of the search for empirical dependences. In *Proceedings of the 2nd Annual Workshop on Computational Learning Theory.* San Mateo, CA: Morgan Kaufmann.