CRITERIA FOR SPECIFYING MACHINE COMPLEXITY IN LEARNING

Changfeng Wang and Santosh S. Venkatesh[*]
Department of Electrical Engineering,
University of Pennsylvania,
Philadelphia, PA 19104
Email: fwang@pender.ee.upenn.edu; venkatesh@ee.upenn.edu

Abstract

We consider the archetypal learning problem where a finite sample of examples generated by an underlying random process is made available to the learner who generates a hypothesis in a model class by gradient descent over the empirical loss function. In this context, we derive two criteria for machine size selection for a class of general nonlinear machines which includes feedforward neural networks as a subclass. The first criterion yields simultaneous estimates of optimal machine size and optimal stopping time for the gradient descent learning algorithm and may be viewed as a formal extension of Akaike's information criterion (AIC) to include general models and the learning process *per se*. This criterion results in optimal generalization—in the sense of minimizing the loss function—but may not lead to consistent estimation. The second criterion admits of a selection of machine size which leads to consistent estimation but is provably nonoptimal. The latter criterion has the same asymptotic form as Rissanen's minimum description length principle (MDL). A study of the properties of the two criteria sheds light on the effects of AIC and MDL on generalization performance, and provides guidelines in effecting a choice between the two types of model size selection criteria.

1 INTRODUCTION

A central problem in learning from examples is the selection of machine complexity (size) for a given learning problem. At the root of the problem is the bias/variance dilemma: a model that is too small will not be able to capture the salient features of the data (bias); on the other hand, a model that is too large may lead to poor performance on new data (because of large variance). Both situations lead to poor generalization. Machine complexity and generalization performance are hence intimately linked.

Machine training geared towards minimizing the empirical or training error on the data usually gives an unduly optimistic estimate of the generalization error, more so for machines of larger size. Therefore, essentially all machine size selection criteria modify the empirical error by penalizing larger machines in some way (for instance, by adding an additional term reflecting the machine complexity to the empirical error). Examples include Akaike's information criterion (AIC) [1], Barron's complexity regularization method [2], and the minimum description length (MDL) principle of Rissanen [3]. Bayesian methods modify the empirical error estimates by considering the *a priori* distribution of learning parameters.

While the above-mentioned criteria exhibit various optimality properties (for instance, the consistency of MDL), in practice, when the number of examples is finite, these criteria do not necessarily yield optimal (or even close to optimal) generalization performance. In fact, although MDL has been widely used in the learning community, its properties are not fully understood; the application of AIC on the other hand is justified only in some restricted settings, and it is not clear how to extend it to more general settings. The imposition of a systematic complexity penalty that yields optimal generalization performance is possible only when a sufficiently precise link between training and generalization errors is forged.

For a given model class (such as, for example, a nested

[*]The research reported here was supported by the Air Force Office of Scientific Research under grant F49620–93–1–0120.

COLT''95 Santa Cruz, CA USA© 1995 ACM 0-89723-5/95/0007..$3.50

class of neural networks) the number of parameters is one indication of the complexity of a machine class, the domain of the parameter space being another. The aforementioned machine selection methods in general impose complexity penalties by penalizing large numbers of parameters and/or restricting the domain of the parameter space. Typically the specifics of the learning algorithm and the training time are not taken into account. Indeed, in the application of these methods it is often assumed that for a machine of given size, the best fit to the data has been achieved. However, recent numerical experiments and theoretical studies on learning in feedforward neural networks reveal that the degree to which the data fits the machine plays an important rôle, as also the time dynamics of the training *process*, in determining the generalization performance. In particular, it is not always advantageous to maximally fit the data.

In this paper we present two criteria for choosing machine complexity. The first criterion simultaneously prescribes the optimal machine size as well as the optimal time at which learning should be stopped. This criterion can be viewed as an extension of AIC in two directions: (1) taking into account the learning process in time; and (2) including more general data generating models. We show that this criterion leads to near-optimal generalization when there are a finite number of examples though the estimates may not be consistent. The second criterion is similar to MDL in spirit and prescribes machine complexity which leads to consistent learning, but does not yield best generalization with a finite number of examples.

The main theorems are proved by the joint application of the uniform convergence of probability measures in empirical processes (based on ideas introduced by Vapnik and Červonenkis—the VC-method) and the notion of differentiable statistical functionals (due to Von Mises): the VC-method provides an initial bound while the latter technique yields a precise second-order approximation of the generalization error.

2 THE LEARNING PROBLEM

We consider the problem of learning from examples a relation between two vectors x and y determined by a *fixed but unknown* probability distribution $P(x, y)$. For definiteness, we shall assume that the relation is described by

$$y = g(x, \xi),$$

where g is some *unknown* function of x and ξ which are random vectors on the same probability space. The vector x can be viewed as the input to an unknown system, ξ a random noise term (possibly dependent on x), and y the system's output.

The hypothesis class $\mathcal{H}_d$ is a family of functions (vectors) indexed on a subset Θ_d of d-dimensional Euclidean space: $\mathcal{H}_d = \{ f(x, \theta) : \theta \in \Theta_d \subseteq \mathbb{R}^d \}$. For example, if $x \in \mathbb{R}^m$ and y is a scalar, $\mathcal{H}_d$ can be the class of functions computed by a feedforward neural network with one hidden layer comprised of h neurons and activation function ψ:

$$f(x, \theta) = \psi\left(\sum_{i=1}^{h} \theta_i \psi\left(\sum_{j=1}^{m} \theta_{ij} x_j + \theta_{i0} \right) + \theta_0 \right).$$

In the above, $d = (m + 2)h + 1$ denotes the number of adjustable parameters.

We suppose that we are given a nested family of hypothesis classes $\bigcup_d \mathcal{H}_d$ where $\mathcal{H}_d \subset \mathcal{H}_{d'}$ for $d < d'$. Thus, for instance, we could consider the nested family of feedforward neural networks with one hidden layer and a variable number $h = 1, 2, \ldots$ of neurons in that layer (whence d ranges through values $(m + 2)h + 1$ as h ranges through $1, 2, \ldots$).

The goal of learning within the nested family of hypothesis classes $\bigcup_d \mathcal{H}_d$ is to find the best approximation of the relation between x and y in $\bigcup_d \mathcal{H}_d$ from a finite set of n examples $\mathcal{D}_n = \{ (x_1, y_1), \ldots, (x_n, y_n) \}$ drawn by independent sampling from the distribution $P(x, y)$. In particular, based on the finite n-sample $\mathcal{D}_n$, we wish to determine the optimal number of parameters d^* as well as the best choice of function $f(x, \theta^*) \in \mathcal{H}_{d^*}$.

3 GENERALIZATION ERROR

In practical learning situations such as learning in neural networks, one first selects a network of fixed structure (a fixed hypothesis class $\mathcal{H}_d$), and then determines the "best" weight vector θ^* (or equivalently, the best function $f(x, \theta^*)$ in this class) using some training algorithm. The proximity of an approximation $f(x, \theta)$ to the target function $y = g(x, \xi)$ at each point x is measured by a loss function $q(y, f(x, \xi))$. For a given hypothesis class, the function $f(\cdot, \theta)$ is completely determined by θ. With g fixed, the loss function may be written, with a slight abuse of notation, as a map $q(x, y, \theta)$. Examples of the forms of loss functions include

$$q(x, y, \theta) = \left(y - f(x, \theta)\right)^2 \qquad \text{(Square-law loss)},$$

$$q(x, y, \theta) = \ln \frac{p(y \mid f(x, \theta))}{p(y \mid x)} \qquad \text{(Kullbak-Leibler)},$$

$$q(x, y, \theta) = \delta(|y - f(x, \theta)|) \qquad \text{(0-1 loss)},$$

where the familiar square-law loss function is commonly used in regression and learning in network settings, the Kullback-Leibler distance (with $p(y \mid z)$ denoting the conditional density of y given z) appears frequently in density estimation, and the 0-1 loss function (with $\delta(\cdot)$

denoting the Kronecker function) is usual in pattern recognition.

Let us fix d for the nonce and write $\Theta = \Theta_d$ for simplicity. For given d, the "goodness of fit" of $f(\cdot, \theta)$ to $g(\cdot)$ is measured by the expected loss or error

$$\mathcal{E}(\theta, d) \triangleq \int q(x, y, \theta)\, P(dx, dy).$$

We write simply $\mathcal{E}(\theta)$ if the dimension d of the parameter space is clear from context. The optimal approximation $f(\cdot, \theta^*)$ is such that

$$\mathcal{E}(\theta^*, d) = \min_{\theta \in \Theta} \mathcal{E}(\theta, d).$$

When θ^* exists, it is in general not unique.

We define the corresponding empirical error on the n-sample $\mathcal{D}_n$ by

$$\mathcal{E}_n(\theta, d) \triangleq \frac{1}{n} \sum_{i=1}^{n} q(x_i, y_i, \theta) = \int q(x, y, \theta)\, P_n(dx, dy),$$

where P_n is the empirical distribution that assigns mass $1/n$ to each pair (x_i, y_i). Again, we write simply $\mathcal{E}_n(\theta)$ when the dimension d is readily inferred from the context. The global minimum of the empirical error over Θ is denoted by $\hat{\theta}$, namely

$$\hat{\theta} = \hat{\theta}(\mathcal{D}_n) \triangleq \arg\min_{\theta \in \Theta} \mathcal{E}_n(\theta, d).$$

The quantity

$$\mathcal{E}\big(\hat{\theta}, d \mid \mathcal{D}_n\big) = \int q\big(x, y, \hat{\theta}\big)\, P(dx, dy)$$

is referred to as the generalization error of the model $f\big(\cdot, \hat{\theta}\big)$ given the data $\mathcal{D}_n$. We will abuse notation slightly and write $\hat{\theta}$ to denote both the random vector $\hat{\theta}(\mathcal{D}_n)$ which minimizes the empirical loss function and the function $\hat{\theta}$ itself viewed as a map from the sample into $\mathbb{R}^d$. Thus,

$$\mathcal{E}\big(\hat{\theta}, d\big) = \mathbb{E}\, \mathcal{E}\big(\hat{\theta}, d \mid \mathcal{D}_n\big) = \int \mathcal{E}\big(\hat{\theta}(\mathcal{D}_n), d \mid \mathcal{D}_n\big)\, dP,$$

where the expectation is with respect to the product distribution of the sample, connotes the expected generalization error obtained at the end of training a machine of size d given a sample of size n.

Typically, a supervised learning algorithm is invoked to minimize the empirical error $\mathcal{E}_n(\theta, d)$ or a modification of it, the archetypal algorithm being gradient descent which prescribes an iterative update of the hypothesis in $\mathcal{H}_d$ according to

$$\theta_{t+1} = \theta_t - \epsilon \frac{\partial \mathcal{E}_n}{\partial \theta}(\theta_t, d) \qquad (t \geq 0)$$

with $\epsilon > 0$ denoting the rate of learning. In the sequel we focus on the sequence of hypotheses $\{\theta_t, t \geq 0\}$ generated by the gradient descent algorithm.

On notation: Vectors are assumed to be column vectors and primes $(\cdot)'$ are used to denote matrix and vector transpose. If $\theta = (\theta_1, \ldots, \theta_d)' \in \mathbb{R}^d$, we write

$$\frac{\partial \phi}{\partial \theta} \triangleq \left(\frac{\partial \phi}{\partial \theta_1}, \ldots, \frac{\partial \phi}{\partial \theta_d}\right)'$$

for the gradient of the function $\phi(\theta)$ and, likewise,

$$\frac{\partial^2 \phi}{\partial \theta^2} \triangleq \left[\frac{\partial^2 \phi}{\partial \theta_i \partial \theta_j}\right]$$

for the Hessian of $\phi(\theta)$. We write $\langle \theta, \theta' \rangle \triangleq \sum_i \theta_i \theta_i'$ for the usual Euclidean inner-product and $|\theta - \theta'| \triangleq \langle \theta, \theta' \rangle^{1/2}$ for the induced metric; for a positive definite matrix A, we also write $|\theta - \theta'|_A \triangleq \langle \theta, A\theta' \rangle^{1/2}$ for the weighted Euclidean distance; finally, we write $\|\theta\|_\infty \triangleq \max_i |\theta_i|$ for the L^∞-vector norm and $\|A\|$ for the strong matrix norm of a matrix A.

4 OPTIMAL GENERALIZATION

To simplify the problem, let us first consider the traditional approach towards choice of machine complexity. In this case, the post-training generalization performance is of concern but not the training process *per se*. Since the structure of machine is given (i.e., the form of the function $f(x, \theta)$ is known), the only factor that affects the complexity of the machine is the number of parameters. The problem of complexity trade-off is equivalent to selecting the correct number of parameters.

We have hitherto considered learning within the confines of a particular hypothesis class $\mathcal{H}_d$ (for some fixed, but arbitrary, d). Let us now return to a consideration of the entire sequence of nested hypothesis classes $\bigcup_d \mathcal{H}_d$. For a given nested hypothesis class, there in general exists a *minimal realization* of the target function, i.e., a machine of minimal machine size d^* which achieves the best approximation of the process generating the examples.[1] Define

$$\theta^*(d) \triangleq \arg\min_{\theta \in \Theta_d} \mathcal{E}(\theta, d).$$

In line with the accords of Occam's razor, we can then define the best machine size as the size of the minimal realization:

$$d^* \triangleq \min\{\, d : \mathcal{E}(\theta^*(d)) \leq \mathcal{E}(\theta^*(d')) \text{ for all } d' \,\}.$$

[1] While the minimal realization is well-defined, the notion of "true size" is not; note that for a nested class of models a larger machine (containing the minimal realization) will do equally well.

However, even if d^* is known, one should not in general expect machines of this size to yield best generalization performance given a finite number of examples as, in the limit of training, we only have access to the machines in the subset $\{\hat{\theta}(d), d \geq 1\}$ rather than the entire hypothesis class. In analogy with its ensemble counterpart $\theta^*(d)$, denote by $\hat{\theta}(d)$ the parameter vector obtained in the limit of training as $t \to \infty$. The (minimal) *optimal machine size* $d_{opt}(\infty)$ at the limit of training satisfies

$$d_{opt}(\infty) \triangleq \min\{\, d : \mathcal{E}(\hat{\theta}(d)) \leq \mathcal{E}(\hat{\theta}(d')) \text{ for all } d' \,\}. \quad (\star)$$

Note that $d_{opt}(\infty) \neq d^*$, in general; an artifact of the finite size n of the training sample.

Matters can be improved somewhat by a consideration of the entire training process in time which results in a larger set of available hypotheses $\bigcup_d \{\theta_t(d) : t \geq 0\}$, where, for each d, $\{\theta_t(d), t \geq 0\}$ is the sequence of hypotheses in $\mathcal{H}_d$ generated by the gradient descent algorithm operating on the corresponding machine of size d. For any fixed d, is there any advantage to considering the whole sequence of hypotheses $\{\theta_t(d), t \geq 0\}$? Consider the corresponding sequence of generalization errors $\{\mathcal{E}(\theta_t(d)), t \geq 0\}$. In analogy with our notation in the limit as $t \to \infty$, take the liberty of identifying $\theta_t(d)$ both with the random vector $(\theta_t(d))(\mathcal{D}_n)$ at epoch t as well as the function $\theta_t(d)$ mapping samples $\mathcal{D}_n$ into vectors in $\mathbb{R}^d$. We may now define the *optimal stopping time* $t_{opt}(d)$ by

$$t_{opt}(d) \triangleq \min\{\, t : \mathcal{E}(\theta_t(d)) \leq \mathcal{E}(\theta_{t'}(d)) \text{ for all } t' \geq 0 \,\}, \quad (\star\star)$$

where, in analogy with our convention in the limit as $t \to \infty$,

$$\begin{aligned}\mathcal{E}(\theta_t(d)) &\triangleq \mathbb{E}\,\mathcal{E}(\theta_t(d) \mid \mathcal{D}_n) \\ &= \int \mathcal{E}((\theta_t(d))(\mathcal{D}_n) \mid \mathcal{D}_n)\, dP,\end{aligned}$$

the integration again with respect to the sampling distribution P. Surprisingly, it transpires that the optimal stopping time is, in general, finite, i.e., best generalization for any fixed machine occurs at a finite time during the training process and not in the limit of training [5, 6, 7]. The optimal machine size d_{opt} is obtained by further optimizing the generalization performance over the nesting parameter d:

$$\begin{aligned} d_{opt} &\triangleq \min\{\, d : \mathcal{E}(\theta_{t_{opt}(d)}(d)) \leq \mathcal{E}(\theta_{t_{opt}(d')}(d')) \text{ for all } d' \,\}, \\ t_{opt} &\triangleq t_{opt}(d_{opt}). \end{aligned} \quad (\star\star\star)$$

Thus, for a training sample of size n, optimal generalization attains if training is optimally stopped at time t_{opt} on a machine of optimal size d_{opt}:

$$\begin{aligned}\mathcal{E}(\theta_{t_{opt}}(d_{opt})) &= \min_{d,t} \mathcal{E}(\theta_t(d)) \\ &= \min_{d,t} \mathbb{E}\,\mathcal{E}(\theta_t(d) \mid \mathcal{D}_n).\end{aligned}$$

In the sequel we show how the optimal stopping time t_{opt} and optimal machine size d_{opt} may be determined in practice.

4.1 Optimal Stopping and Optimal Size Selection

We make the following assumptions:

A1 The loss function $q(x, y, \theta)$ is twice continuously differentiable with respect to θ, and each element of the Hessian $\frac{\partial^2 q}{\partial \theta^2}$ is continuous in (x, y).

A2 There exists a unique optimal point θ^* in the set $\Theta = \{\theta : \|\theta_i\|_\infty \leq A\}$ for some constant A (which, without lost of generality, is assumed to be 1).

A3 There exists a constant $\rho > 0$ such that the matrix $\Phi(\theta^*) \triangleq \frac{\partial^2 \mathcal{E}}{\partial \theta^2}(\theta^*)$ is nonsingular for all θ in a ρ-neighborhood of θ^*.

A4 The joint distribution of $Z = (X, Y)$ has compact support.

While these assumptions are fairly mild, a few comments may be in order. Assumption A1 is necessary if a gradient based algorithm is to be applicable. For non-differentiable loss functions, a differentiable approximation can be used: for example, the 0-1 loss function can be approximated arbitrarily well by the function $q(f, g) = 1 - e^{-\alpha(f-g)^2}$ for a proper choice of constant $\alpha > 0$. Assumption A4 is readily justifiable in typical learning situations as observable quantities are typically bounded. In fact, this assumption is much stronger than needed for the proof of the theorems to follow. Both Assumption A2 and A3 deal with the topology of the error surface. Assumption A2 tacitly focuses on the local behavior of the learning process when the initial hypothesis is close to θ^*. Such an assumption is often necessary in order for the learning algorithm to converge. Note that Assumption A2 does not exclude the possibility that the error surface may contain local minima (in fact, θ^* is allowed to be a local minimum). It does, however, exclude the situation where the error surface around θ^* flats out, a situation warranting separate study. Assumption A3 is closely related to Assumption A2 and essentially requires that there be no nuisance parameters (irrelevant attributes) in the model, namely, that all the parameters have an effect on the output of the model.

Let us first set up some more preliminary notation. Denote by $\lambda_{ni}(t)$ the ith largest eigenvalue of the matrix

$$\Phi_n(\theta_t) \triangleq \frac{\partial^2 \mathcal{E}_n}{\partial \theta^2}(\theta_t).$$

Let $T_n \triangleq T_n(\theta_t)$ be the corresponding diagonalizing matrix for $\Phi_n(\theta_t)$, i.e.,

$$T_n \Phi_n(\theta_t) T_n' = \mathrm{diag}(\lambda_{n1}(t), \ldots, \lambda_{nd}(t)),$$

and let $\nu_{ni}(t)$ the ith diagonal entry of the matrix

$$\Omega_n(\theta_t) \triangleq \tfrac{1}{n} \sum_{i=1}^{n} T_n \tfrac{\partial q}{\partial \theta}(x_i, \theta_t) \tfrac{\partial q}{\partial \theta}(x_i, \theta_t)' T_n'.$$

Theorem 1 *Let $0 < \delta \leq \frac{1}{2}$ be fixed and consider gradient descent training on a machine of size d. Then under Assumptions A1–A4, the generalization error of the machine at epoch t is governed by*

$$\begin{aligned} \mathcal{E}(\theta_t, d) &= \mathbb{E}\{\mathcal{E}_n(\theta_t, d)\} \\ &\quad + \mathbb{E}\left\{ \frac{1}{n} \sum_{i=1}^{d} \frac{\nu_{ni}(t)}{\lambda_{ni}(t)} \left[1 - (1 - \epsilon\lambda_{ni}(t))^t\right] \right\} \\ &\quad + \mathcal{O}(n^{-3\delta}). \end{aligned} \tag{3}$$

uniformly for all initial hypotheses θ_0 in a $\mathcal{O}(n^{-\delta})$-neighborhood centered at θ^* and for all epochs $t \geq 0$.

PROOF: We only sketch the main ideas here. We seek a sufficiently accurate connection between generalization error, training error, and machine complexity at each epoch of training t and each size d. As a first step we start with the Taylor expansion of $\mathcal{E}(\theta_t)$ around $\mathcal{E}(\theta^*)$; a consideration of the dynamics of the training process allows us to then estimate the latter quantity in terms of the empirical error at epoch t. This process leads to an estimate of the form

$$\begin{aligned} \mathcal{E}(\theta_t) &= \mathbb{E}\,\mathcal{E}_n(\theta_t) + \left|\hat{\theta} - \theta^*\right|^2_{(I - \Delta_n(t))\Phi'} \\ &\quad + \langle \theta_0 - \theta^*, \hat{\theta} - \theta^* \rangle_{\Delta_n(t)\Phi'} + \mathcal{O}(n^{-1-\delta}), \end{aligned}$$

with

$$\Delta_n(t) \triangleq \prod_{i=1}^{t} (I - \Phi_n(\theta_i')) \quad \text{and} \quad \Phi' \triangleq \tfrac{\partial^2 \mathcal{E}}{\partial \theta^2}(\theta_0'),$$

where θ_0' is a random vector on the line joining θ_t to $\hat{\theta}$, and, for $1 \leq i \leq t$, θ_i' is a random vector on the line joining the successive hypotheses θ_{i-1} and θ_i returned by the gradient descent algorithm at epochs $i-1$ and i, respectively.

A consideration of the rate of convergence of $\hat{\theta}$ to θ^* now shows that

$$\begin{aligned} &\mathbb{E}\left|\hat{\theta} - \theta^*\right|^2_{(I - \Delta_n(t))\Phi'} \\ &\quad = \tfrac{1}{n} \mathbb{E}\,\mathrm{tr}\left\{\Phi^{-1}\Omega(I - \Delta(t))\right\} + \mathcal{O}(n^{-1-\delta}), \end{aligned}$$

where Ω is the almost sure limit of $\Omega_n(\theta_t)$ as $n \to \infty$ and $t \to \infty$. Write $K_n = \Delta_n(t)\Phi'$ and denote by K the almost sure limit of K_n as $n \to \infty$. We then obtain

$$\begin{aligned} &\left|\mathbb{E}\langle \theta_0 - \theta^*, \hat{\theta} - \theta^* \rangle_{K_n}\right| \\ &\qquad \leq \left|\langle \theta_0 - \theta^*, \mathbb{E}(\hat{\theta} - \theta^*) \rangle_K\right| \\ &\qquad\quad + \mathbb{E}\left\{\left|\theta_0 - \theta^*\right| \left|\hat{\theta} - \theta^*\right| \|K_n - K\|\right\} \\ &\qquad = \mathcal{O}(n^{-\frac{3}{2}}) + \mathcal{O}(n^{-\frac{1}{2} - 2\delta}). \end{aligned}$$

Now, the substitution of Φ and Ω_n by $\Phi_n(\theta_t)$ and $\Omega_n(t)$, respectively, in (4) occasions an error of order $\mathcal{O}(n^{-\frac{1}{2} - 2\delta})$. Consequently, we obtain

$$\begin{aligned} \mathcal{E}(\theta_t) &= \mathbb{E}\{\mathcal{E}_n(\theta_t)\} + \mathbb{E}\left\{\tfrac{1}{n}\,\mathrm{tr}\left[I - (I - \Delta_n(t))^t\right]\right. \\ &\qquad \left. \times\, \Phi_n(t)^{-1}\Omega_n(t)\right\} + \mathcal{O}(n^{-3\delta}). \end{aligned}$$

An orthogonal transformation of the second term yields the desired result. ▌

Now writing

$$C(n, d, t) \triangleq \tfrac{1}{n} \sum_{i=1}^{d} \frac{\nu_{ni}(t)}{\lambda_{ni}(t)} \left[1 - (1 - \epsilon\lambda_{ni}(t))^t\right],$$

the integrand in (3) can be written as

$$\underbrace{\mathcal{E}_n(\theta_t, d)}_{\substack{\text{approximation} \\ \text{error}}} + \underbrace{C(n, d, t)}_{\substack{\text{complexity} \\ \text{error}}},$$

which gives the decomposition of the generalization error in terms of the empirical approximation error and the contribution due to the *effective* machine complexity *at epoch* t. For fixed d let $t_{n,\mathrm{opt}}(d)$ denote the smallest value of t for which $\mathcal{E}_n(\theta_t, d) + C(n, d, t)$ is minimized (compare with (⋆⋆)), and assume that $t_{n,\mathrm{opt}}(d)$ is measurable. Then for all $t \geq 0$,

$$\begin{aligned} \mathcal{E}(\theta_{t_{n,\mathrm{opt}}(d)}(d)) &= \mathbb{E}\{\mathcal{E}_n(\theta_{t_{n,\mathrm{opt}}(d)}(d))\} \\ &\quad + \mathbb{E}\{C(n, d, t_{n,\mathrm{opt}}(d))\} + \mathcal{O}(n^{-3\delta}) \\ &\leq \mathbb{E}\{\mathcal{E}_n(\theta_t(d))\} \\ &\quad + \mathbb{E}\{C(n, d, t)\} + \mathcal{O}(n^{-3\delta}) \\ &= \mathcal{E}(\theta_t(d)) + \mathcal{O}(n^{-3\delta}). \end{aligned}$$

In particular, we have

$$\mathcal{E}(\theta_{t_{n,\mathrm{opt}}(d)}(d)) \leq \mathcal{E}(\theta_{t_{\mathrm{opt}}}(d)) + \mathcal{O}(n^{-3\delta}),$$

which implies that $t_{n,\mathrm{opt}}(d)$ is approximately the optimal stopping time given the examples. Since $t_{n,\mathrm{opt}}(d)$ may be determined solely from the examples, this provides a practical method for estimating the optimal stopping time $t_{\mathrm{opt}}(d)$ for a given machine size d.

Theorem 1 also forms the basis for determining the optimal machine size. Suppose that d ranges over a bounded set in the nested class of hypotheses $\bigcup_d \mathcal{H}_d$. An argument similar to the one above then gives the optimal machine size at each epoch t. In particular, as $t \to \infty$, we obtain $\mathcal{E}_n(\theta_t, d) \to \mathcal{E}(\hat{\theta}, d)$ and $C(n, d, t) \to \frac{1}{n}\sum_{i=1}^{d} \frac{\nu_{ni}(\infty)}{\lambda_{ni}(\infty)}$, where $\nu_{ni}(\infty)$ is the ith diagonal element of $\frac{\partial^2 \mathcal{E}_n}{\partial \theta^2}(\hat{\theta})$, and $\lambda_{ni}(\infty)$ is the ith eigenvalue of $\Phi_n(\hat{\theta})$. The optimal machine size at the end of training is given by $d_{n,\mathrm{opt}}(\infty)$ which satisfies

$$d_{n,\mathrm{opt}}(\infty) = \arg\min_{d}\left\{ \mathcal{E}_n(\hat{\theta}, d) + \frac{1}{n}\sum_{i=1}^{d} \frac{\nu_{ni}(\infty)}{\lambda_{ni}(\infty)} \right\}.$$

(Compare with $(\star)$.) Since the influence of the initial point is eliminated at the end of training, we can take $\delta = \frac{1}{2}$ in Theorem 1, so that optimality here is at the order of $\mathcal{O}(n^{-\frac{3}{2}})$; namely, for all d,

$$\mathcal{E}(\hat{\theta}(d_{n,\mathrm{opt}}(\infty))) \leq \mathcal{E}(\hat{\theta}(d)) + \mathcal{O}(n^{-\frac{3}{2}}),$$

and, in particular,

$$\mathcal{E}(\hat{\theta}(d_{n,\mathrm{opt}}(\infty))) \leq \mathcal{E}\{\hat{\theta}(d_{\mathrm{opt}}(\infty))\} + \mathcal{O}(n^{-\frac{3}{2}}).$$

Optimizing both machine size and stopping time by a completely analogous argument enables us to obtain simultaneous estimates $t_{n,\mathrm{opt}}$ and $d_{n,\mathrm{opt}}$ of the optimal stopping time t_{opt} and optimal machine size d_{opt}, respectively.

Criterion O *Optimal generalization is achieved for a choice of* $t = t_{n,\mathrm{opt}}$ *and* $d = d_{n,\mathrm{opt}}$ *according to the criterion*

$$\mathcal{E}_n(\theta_{t_{n,\mathrm{opt}}}(d_{n,\mathrm{opt}})) = \min_{t,d}\{\mathcal{E}_n(\theta_t(d)) + C(n, d, t)\};$$

the optimality is at the order of $\mathcal{O}(n^{-3\delta})$.

(Compare with $(\star\star\star)$)

4.2 Relation with AIC

Consider as choice of loss function the Kullback-Leibler distance

$$q(x, y, \theta) = \ln \frac{p(y \mid f(x, \theta))}{p(y \mid x)}$$

and an additive data generating model where the noise variable ξ has variance σ^2 and is independent of x. If $g(x, \xi) = f(x, \theta^*) + \xi$, it can be shown that (cf. [7])

$$\mathcal{E}(\hat{\theta}, d) = \mathcal{E}(\theta^*, d) + \frac{\sigma^2 d}{2n} + o(n^{-1}).$$

The complexity penalty of $\sigma^2 \frac{d}{2n}$ prescribed by Akaike's information criterion [1] coincides with that of Criterion O in the limit of large time. Criterion O can hence be viewed as a formal extension of AIC to more general settings while taking into account the training process in time.

The optimality of model size selection via Criterion O holds when the model size d varies over a bounded range in the nested family of hypothesis classes. Criterion O may not be optimal, however, if d varies unboundedly or if the choice of d is allowed to grow with the number of examples, and indeed, just as for AIC, the criterion may even lead to inconsistent learning in such cases.[2]

5 CONSISTENT LEARNING

In this section we derive another machine size selection criterion which results in consistent learning based on error bounds on the generalization error involving the complexity of the hypothesis class. The basic idea as in structural risk minimization [4] is to minimize the upper bound of the generalization error in the form of a sum of an empirical error and a term which is proportional to the complexity of the machine. To get sharper bounds, however, we explore the local properties of the error function. As we will see, this results in a complexity term of order $\mathcal{O}(d \ln n / n)$.[3]

Write $z = (x, y)$ for brevity, and define the finite, nonzero quantities

$$c_1 \triangleq \sup_{z,\theta} \left| \frac{\partial q}{\partial \theta}(z, \theta) \right|,$$

$$c_2 \triangleq \sup_{z,\theta} \left\| \frac{\partial^2 q}{\partial \theta^2}(z, \theta) \right\|^2,$$

$$C(\epsilon) \triangleq \inf_{|\theta - \theta^*| > \epsilon} \{\mathcal{E}(\theta) - \mathcal{E}(\theta^*)\}.$$

Lemma 1 *Under Assumptions A1–A4, the inequality*

$$\mathbb{P}\{|\hat{\theta} - \theta^*| > \epsilon\} \leq \left[8\left(\tfrac{2c_1}{\epsilon}\right)^d + 1\right] e^{-nC^2(\epsilon)/K} + \mathcal{O}(e^{-nB(\rho)}),$$

where $K = 512 \sup_{z,\theta} q(z, \theta)$ *and* $B(\rho)$ *is a constant depending only on* ρ *(Assumption A3), holds for all* $n \geq \epsilon^2/8$.

PROOF: Again, we only sketch the principal ideas in the proof. Define the event $E \triangleq \{|\hat{\theta} - \theta^*| \leq \rho\}$. An elementary conditioning argument now shows that

$$\mathbb{P}\{|\hat{\theta} - \theta^*| > \epsilon\} \leq \mathbb{P}\{|\hat{\theta} - \theta^*| > \epsilon \mid E\} + \mathbb{P}\{|\hat{\theta} - \theta^*| > \rho\}.$$

[2] We may allow d to grow at a modest rate while still preserving the optimality of Criterion O; for instance, $d_n = \mathcal{O}(n^{1/3})$ works [5].

[3] This does not fly in contradiction of the Devroye-Lagosi uniform lower bound $\Omega(1/\sqrt{n})$ on probability of errors. Our bound is not uniform over the whole hypothesis class but rather is specialized to the point of special interest $\hat{\theta}$ and depends critically on our assumptions of the error surface.

Since $\mathcal{E}_n(\hat{\theta}) \leq \mathcal{E}_n(\theta^*)$, the event $\mathcal{E}(\hat{\theta}) - \mathcal{E}(\theta^*) > \delta$ implies the event

$$\mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta}) + \mathcal{E}_n(\theta^*) - \mathcal{E}(\theta^*) > \delta,$$

whence,

$$\begin{aligned}\mathbb{P}\{\mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta}) + \mathcal{E}_n(\theta^*) - \mathcal{E}(\theta^*) > \delta \mid \mathrm{E}\} \\ \leq \mathbb{P}\{\mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta}) > \tfrac{1}{2}\delta \mid \mathrm{E}\} \\ + \mathbb{P}\{\mathcal{E}_n(\theta^*) - \mathcal{E}(\theta^*) > \tfrac{1}{2}\delta \mid \mathrm{E}\}.\end{aligned}$$

Denoting by $S(\theta^*, \epsilon) \triangleq \{\theta : |\theta - \theta^*| \leq \epsilon\}$, it is easy to see that

$$C(\epsilon) = \inf_{\Theta \backslash S(\theta^*, \epsilon)} \left\{\mathcal{E}(\theta) - \mathcal{E}(\theta^*)\right\} > 0,$$

and that

$$\left\{|\hat{\theta} - \theta^*| > \epsilon\right\} \subseteq \left\{\mathcal{E}(\hat{\theta}) - \mathcal{E}(\theta^*) > C(\epsilon)\right\}.$$

Setting $\delta = C(\epsilon)$, we have

$$\begin{aligned}\mathbb{P}\{|\hat{\theta} - \theta^*| > \epsilon \mid \mathrm{E}\} \\ \leq \mathbb{P}\{\mathcal{E}(\hat{\theta}) - \mathcal{E}_n(\hat{\theta}) > \tfrac{1}{2}C(\epsilon) \mid \mathrm{E}\} \\ + \mathbb{P}\{\mathcal{E}_n(\theta^*) - \mathcal{E}(\theta^*) > \tfrac{1}{2}C(\epsilon) \mid \mathrm{E}\}.\end{aligned}$$

Hoeffding bounds and moment inequalities may now be deployed to obtain exponential decay for the various terms. ▌

Lemma 1 gives the rate of convergence of $\hat{\theta}$ to θ^* in terms of the machine complexity and the parameter $C(\epsilon)$ which plays an important rôle in characterizing the complexity of the error surface near θ^*. We may lower bound $C(\epsilon)$ as follows: since θ^* is an interior point of Θ and $\frac{\partial \mathcal{E}}{\partial \theta}(\theta^*) = 0$, we obtain the Taylor expansion

$$\mathcal{E}(\theta) = \mathcal{E}(\theta^*) + (\theta - \theta^*)' \Phi(\tilde{\theta})(\theta - \theta^*)$$

valid for all $\theta \in \Theta \backslash S(\theta^*, \epsilon)$. In the above, $\tilde{\theta}$ is a point on the line joining $\hat{\theta}$ and θ^*. It follows that

$$\begin{aligned}C(\epsilon) &= \inf_{\Theta \backslash S(\theta^*, \epsilon)} \mathcal{E}(\theta) - \mathcal{E}(\theta^*) \\ &= \inf_{\Theta \backslash S(\theta^*, \epsilon)} (\theta - \theta^*)' \tfrac{\partial^2 \mathcal{E}}{\partial \theta^2}(\tilde{\theta})(\theta - \theta^*) \\ &\geq \lambda_{\min} \inf_{\Theta \backslash S(\theta^*, \epsilon)} |\theta - \theta^*|^2 \geq \lambda_{\min} \epsilon^2,\end{aligned}$$

where $\lambda_{\min}$ denotes the smallest eigenvalue of $\Phi(\theta)$ over $\Theta \backslash S(\theta^*, \epsilon)$. Since $\Phi(\theta)$ is continuous in θ, its eigenvalues are continuous in θ, and as Θ is compact, it follows that $\lambda_{\min} > 0$.

All is now set for the criterion. Consider first the case when there exists a constant $\lambda > 0$ such that $C(\epsilon) > \lambda \epsilon$. A Taylor expansion of $\mathcal{E}(\cdot, d)$ around θ^* yields

$$\mathcal{E}(\theta, d) \leq \mathcal{E}(\theta^*, d) + \tfrac{1}{2} c_2 |\theta - \theta^*|^2.$$

On the other hand,

$$\mathcal{E}_n(\theta^*, d) \leq \mathcal{E}_n(\hat{\theta}, d) + \tfrac{1}{2} c_2 \left|\hat{\theta} - \theta^*\right|^2.$$

Thus

$$\mathcal{E}(\hat{\theta}, d) \leq \mathbb{E}\{\mathcal{E}_n(\hat{\theta}) + c_2 \left|\hat{\theta} - \theta^*\right|^2\}.$$

In order to minimize $\mathcal{E}(\hat{\theta}, d)$ we seek to minimize the integrand on the right hand side of the last inequality.

The idea now is to obtain the smallest distribution-free upper bound for $\hat{\theta} - \theta^*$. This is achieved by a consideration of the convergence rate evidenced in Lemma 1. The tail bound of the lemma may be rewritten as

$$8 \exp\{\tfrac{1}{2}(\ln c_1^2 - \ln \epsilon^2 - \tfrac{2n}{Kd}\epsilon^2)\} + \exp\{-n\epsilon^2/K\}.$$

For the above bound to go to zero as $n \to \infty$, ϵ cannot be smaller than the order $\mathcal{O}\left(\sqrt{\frac{\ln^\delta n}{n}}\right)$ for any $\delta \leq 1$; and indeed, precisely at the order of $\mathcal{O}\left(\sqrt{\frac{\ln n}{n}}\right)$, we have $\left|\hat{\theta} - \theta^*\right|^2 \leq \frac{Kd \ln n}{2n}$ with probability approaching one as $n \to \infty$ for the minimizing choice $\epsilon = \sqrt{\frac{Kd \ln n}{2n}}$. It follows that we require $d = o(\ln n / n)$ if ϵ is to approach zero. Let $\mathcal{J}$ denote the range of values over which d is allowed to range in the nested family of hypothesis classes $\bigcup_d \mathcal{H}_d$, and denote $\mathcal{J}' \triangleq \{d \in \mathcal{J} : d = o(n / \ln n)\}$. In the absence of any further information on the distance between $\hat{\theta}$ and θ^*, we may elect to choose machine size in accordance with the following

Criterion C *Consistent learning is achieved with a selection of machine size*

$$d_{n,c} \triangleq \arg\min_d \left\{\mathcal{E}_n(\hat{\theta}, d) + \frac{\mu d \ln n}{4n} : d \in \mathcal{J}'\right\},$$

where $\mu = Kc_2$.

More precisely, we have shown the following

Theorem 2 *Suppose* μ *and* c_1 *are constants independent of* d, *and suppose Assumptions A1–A4 are satisfied, and furthermore,* $C(\epsilon) \geq \lambda \epsilon$ *for some* $\lambda > 0$. *Then Criterion C leads to consistent learning in the nested family of hypothesis classes* $\bigcup_{d=1}^{\infty}\{f(x, \theta) : \theta \in \mathbb{R}^d\}$.

Consistency here comes about as a consequence of restrictions on the range on d. The minimization with respect to d of the upper limit bound of $\mathcal{E}(\hat{\theta}, d)$ is in the hope of reducing the error when there are only a finite number of examples. This is in the same spirit as structural risk minimization. In general, however, the criterion does not yield the optimal choice of machine size in the sense that the generalization error is not minimized with a finite number of examples.

5.1 Relation with MDL

If $C(\epsilon) \approx \lambda\epsilon^{\kappa}$, with $\kappa \leq 2$, we obtain a complexity penalty of order $\mathcal{O}((\ln n)^{1/\kappa})$. In particular, when $\kappa = 2$, we recover the order given by structural risk minimization.

Note that the complexity penalty of Criterion C is similar to that of MDL (cf. [3]). Indeed, if the loss function is $q(x, y, \theta) = -\ln p\big(y \mid f(x, \theta)\big)$, and $\mu c_1 = 1$, the criterion has the same asymptotic form as MDL. Since the complexity penalty of order $\mathcal{O}(d \ln n/n)$ is about the smallest distribution-free complexity penalty that can be added to the empirical error, we conclude that MDL in effect results in the smallest complexity penalty required for consistent learning.

For learning to be consistent, typically we need to restrict the rate of growth of machine size at an order $o(n/\ln n)$. On the other hand, a necessary condition for MDL to lead to consistent learning is that the complexity penalty $d \ln n/n$ approaches zero as $n \to \infty$. This is equivalent to requiring d to be of order $o(n/\ln n)$!

6 CONCLUSIONS

An examination of the fine structure of the Criterion O and Criterion C shows that when d is relatively small compared to n, the effect of machine size on generalization error is at the order of $\Theta(1/n)$; thus in this case MDL would seem to penalize machine size too severely to allow of an optimal trade-off between training error and machine complexity.

There is a need for both kinds of complexity penalties. In cases where $\mathcal{E}(\theta^*)$ decreases very slowly with increase of machine size d, i.e., in cases where the machine structure does not fit the underlying process well, the principal approach toward reduction of generalization error is to obtain a smaller $\mathcal{E}(\theta^*)$ by choosing a larger machine. In this case, d must grow fairly fast as n increases, and we need to impose restrictions on the growth of machine size so that consistent learning is achieved. This is the MDL principle at work. On the other hand, if a very small d is required to obtain a small approximation error $\mathcal{E}(\theta^*)$, (say $d = \mathcal{O}(n^r)$ for $r \ll 1/3$), consistency will be automatic, and we find nearly optimal trade-off by penalizing the empirical error with a complexity term of order $\Theta(1/n)$. This is the AIC type of complexity penalty at work.

In summary, the kind of complexity penalty needed depends on the approximation rate $\mathcal{E}(\theta^*)$. To assure consistency for Criterion O, one only needs to restrict the choice of d to the subset of indices $\mathcal{J}' \subset \mathcal{J}$. This has the effect of automatically switching between the two kinds of criteria: AIC and MDL.

As a final note, our derivation of Criterion C shows that as far as consistency of learning is concerned, the complexity penalty could be rather arbitrary: any penalty larger than $\mathcal{O}(\frac{d \ln n}{n})$ will also lead to consistent learning. When optimality of generalization given a finite number of examples is sought, however, a penalty of order $\mathcal{O}(\frac{\ln n}{n})$ is magnitudes larger than needed!

References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Second International Symposium on Information Theory,* Ed. B.N. Krishnaiah, North Holland, Amsterdam, pp. 27–41, 1974.

[2] A. Barron, "Complexity regularization with applications to artificial neural networks," *Nonparametric Function Estimation*, pp. 561–576. Boston, MA: Kluwer Academic Publishers, 1990.

[3] J. Rissanen, "Stochastic complexity," *J. Royal Statistical Society*, Series B, vol. 49, no. 3, pp. 223–265, 1987.

[4] V. Vapnik, *Estimation of Dependencies by Empirical Data.* New York: Springer-Verlag, 1982.

[5] C. Wang, *A Theory of Generalization in Learning Machines.* Ph. D. Thesis, University of Pennsylvania, 1994.

[6] C. Wang and S. S. Venkatesh, "Temporal dynamics of Generalization in Learning in Neural Networks," in *Proceedings of NIPS'94*, in press.

[7] C. Wang, S. S. Venkatesh, and J. S. Judd, "Optimal stopping and effective machine size in learning," in *Proceedings of NIPS'93.*