Asymptotic Statistical Theory of Overtraining and Cross-Validation

Shun-ichi Amari, *Fellow, IEEE,* Noboru Murata, Klaus-Robert Müller,
Michael Finke, and Howard Hua Yang, *Member, IEEE*

***Abstract*— A statistical theory for overtraining is proposed. The analysis treats general realizable stochastic neural networks, trained with Kullback–Leibler divergence in the *asymptotic* case of a large number of training examples. It is shown that the asymptotic gain in the generalization error is small if we perform early stopping, even if we have access to the optimal stopping time. Considering cross-validation stopping we answer the question: In what ratio the examples should be divided into training and cross-validation sets in order to obtain the optimum performance. Although cross-validated early stopping is useless in the asymptotic region, it surely decreases the generalization error in the nonasymptotic region. Our large scale simulations done on a CM5 are in nice agreement with our analytical findings.**

***Index Terms*—Asymptotic analysis, cross-validation, early stopping, generalization, overtraining, stochastic neural networks.**

I. Introduction

MULTILAYER NEURAL networks improve their behavior by learning a set of examples showing the desired input–output relation. This training procedure is usually carried out by a gradient descent method minimizing a target function ([1], [27], and many others).

When the number of examples is infinitely large and they are unbiased, the network parameters converge to one of the local minima of the empirical risk function (expected loss) to be minimized. When the number of training examples is finite, the true risk function (generalization error) is different from the empirical risk function. Thus, since the training examples are biased, the network parameters converge to a biased solution. This is known as overfitting or overtraining,[1] because the parameter values fit too well the speciality of the biased training examples and are not optimal in the sense of minimizing the generalization error given by the risk function.

There are a number of methods of avoiding overfitting. For example, model selection methods (e.g., [23], [20], [26] and many others), regularization ([25] and others), and early stopping ([16], [15], [30], [11], [4] and others) or structural risk minimization (SRM, cf. [32]) can be applied.

Here we will consider early stopping in detail. There is a folklore that the generalization error decreases in an early period of training, reaches a minimum and then increases as training goes on, while the training error monotonically decreases. Therefore, it is considered better to stop training at an adequate time, a technique often referred to as early stopping . To avoid overtraining, the following simple stopping rule has been proposed based on cross-validation: Divide all the available examples into two disjoint sets. One set is used for training. The other set is used for validation such that the behavior of the trained network is evaluated by using the cross-validation examples and the training is stopped at the point that minimizes the error on the cross-validation set. Note that dividing the available examples into two fixed sets is a strongly simplified implementation of k-fold cross-validation (cf. [12]).[2] In our study we will consider only the above described two set cross-validation and we will refer to it as cross-validation in the following.

Wang *et al.* [30] analyzed the average optimal stopping time without cross-validation in the case of linear Φ-machines. For the regression case Sjöberg and Ljung [29] calculated asymptotically that the number of efficient parameters is linked 1) to the regularization parameter if a specific regularization is applied and 2) to the number of iterations of the learning algorithm if early stopping is used. They denote early stopping as implicit regularization. Bishop [9] showed that regularization and early stopping lead to similar solutions and stressed the analogy between the number of iterations and the regularization parameter. Barber *et al.* [6], [7] considered the evaluation of the generalization error by cross-validation for linear perceptrons.

Recently Guyon [14] and Kearns [18] derived a VC bound for the optimal split between training and validation set, which shows the same scaling as our result. The VC result scales inversely with the square root of the VC dimension (cf. [31])

Manuscript received September 11, 1995; revised October 21, 1996 and May 10, 1997. K.-R. Müller was supported in part by the EC S & T fellowship (FTJ 3-004). This work was supported by the National Institutes of Health (P41RRO 5969) and CNCPST Paris (96JR063).

S. Amari is with the Lab. for Inf. Representation, RIKEN, Wakoshi, Saitama, 351-01, Japan. He is also with the Department of Mathematical Engineering and Information Physics, University of Tokyo, Tokyo 113, Japan.

N. Murata is with the Department of Mathematical Engineering and Information Physics, University of Tokyo, Tokyo 113, Japan.

K.-R. Müller is with the Department of Mathematical Engineering and Information Physics, University of Tokyo, Tokyo 113, Japan. He is also with GMD FIRST, 12489 Berlin, Germany.

M. Finke is with the Institut für Logik, University of Karlsruhe, 76128 Karlsruhe, Germany.

H. H. Yang is with the Lab. for Inf. Representation, RIKEN, Wakoshi, Saitama, 351-01, Japan.

Publisher Item Identifier S 1045-9227(97)06051-7.

[1] The concept of overfitting refers to fitting specific examples too much, thus loosing generality. Overtraining addresses the issue of using too many iterations in the learning procedure, which leads to overfitting (see, e.g., [29]).

[2] For example in leave one out cross-validation, a pattern is drawn randomly, its error is validated, then another pattern is drawn and so on. Finally the validation error is determined by averaging over all chosen patterns.

of the network, which in the case of realizable rules coincides with the number of parameters m. Their result was achieved by bounding the probability of error of the recognizer selected by cross-validation using both classical and VC bounds; the resulting bound is then optimized with respect to the split between training and validation set.

There are various phases in the overtraining phenomena depending on the ratio of the number t of examples to the number m of the modifiable parameters (see [24]). When t is smaller or nearly equal to m, the examples can in principle be memorized and overfitting is remarkable in this phase, in particular around $t \approx m$ ([19], [10]). However, the application of simple two set cross-validation stopping has serious disadvantages in this case. The simple splitting of an already small set of examples decreases the scarce and valuable information in the small data set. In order to avoid overtraining in this case, we need to use global methods like the above mentioned regularization, SRM or k-fold cross-validation rather than simple two set cross-validation stopping.

In an intermediate phase (cf. [24]) of t larger than m, simulations show that cross-validation stopping is effective in general. However, it is difficult to construct a general theory in this phase (see also Section VI for discussion).

In the asymptotic phase where t is sufficiently large, the asymptotic theory of statistics is applicable and the estimated parameters are approximately normally distributed around the true values.

As the first step toward elucidation of overtraining and cross-validation, the present paper gives a rigorous mathematical analysis of overtraining phenomena for 1) a realizable stochastic machine (Section II); 2) Kullback-Leibler divergence (negative of the log likelihood loss); and 3) a sufficiently large number t of examples (compared with the number m of parameters).

We analyze the relation between the training error and cross-validation error, and also the trajectory of learning using a quadratic approximation of the risk function around the optimal value in the asymptotic region (Section III). The effect of early stopping is studied on this basis. It is shown that asymptotically we gain little by early stopping even if we had access to the optimal stopping time (Section IV). Since we never have access to the optimal stopping time, the generalization error becomes asymptotically worse, which means that the gain achieved through early stopping is asymptotically smaller than the loss of not using the cross-validation examples for training.

We then answer the question: In what ratio the examples should be divided into training and cross-validation sets in order to obtain the optimum performance (Section V). We give a definite analytic answer to this problem. When the number m of network parameters is large, the best strategy is to use almost all t examples in the training set and to use only $t/\sqrt{2m}$ examples in the cross-validation set, e.g., when $m = 100$, this means that only 7% of the training patterns are to be used in the set determining the point for early stopping.

Our results are confirmed by large-scale computer simulations of three-layer feedforward networks where the number m of modifiable parameters is $m \sim 100$. When $t > 30m$, the theory fits well with simulations, showing cross-validation is not necessary, because the generalization error becomes worse by using cross-validation examples to obtain an adequate stopping time. For an intermediate range, where $t < 30m$ overtraining occurs surely and the cross-validation stopping improves the generalization ability strongly (Section VII). Finally, concluding remarks are given in Section VIII.

II. Stochastic Feedforward Networks

Let us consider a stochastic network which receives input vector $\boldsymbol{x}$ and emits output vector $\boldsymbol{y}$. The network includes a modifiable vector parameter $\boldsymbol{w} = (w_1, \cdots, w_m)$ and the network specified by $\boldsymbol{w}$ is denoted by $N(\boldsymbol{w})$. The input–output relation of the network $N(\boldsymbol{w})$ is specified by the conditional probability $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{w})$. It is assumed that input $\boldsymbol{x}$ is randomly chosen from an unknown probability $q(\boldsymbol{x})$. The joint probability of $(\boldsymbol{x},\boldsymbol{y})$ of $N(\boldsymbol{w})$ is given by

$$p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}) = q(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{w}). \tag{1}$$

We assume the following.

1) There exists a teacher network $N(\boldsymbol{w}_0)$ which generates training examples.
2) The Fisher information matrix $G = (G_{ij})$ defined by

$$G_{ij}(\boldsymbol{w}) = E\left[\frac{\partial}{\partial w_i}\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})\frac{\partial}{\partial w_j}\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})\right] \tag{2}$$

has a full rank and is differentiable with respect to $\boldsymbol{w}$, where E denotes the expectation with respect to $p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})$.
3) The training set

$$D_t = \{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_t,\boldsymbol{y}_t)\}$$

consists of t i.i.d. examples generated by the distribution $p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}_0)$ of $N(\boldsymbol{w}_0)$.

Let us define the risk and loss functions. When an input–output pair $(\boldsymbol{x},\boldsymbol{y})$ is an example generated from network $N(\boldsymbol{w})$, its loss or error is given by the negative of the likelihood,

$$l(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}) = -\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}). \tag{3}$$

The risk function $R(\boldsymbol{w})$ of network $N(\boldsymbol{w})$ is the expectation of loss with respect to the true distribution

$$R(\boldsymbol{w}) = -E_0[\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})] \tag{4}$$

where E_0 denotes the expectation with respect to $p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}_0)$. The risk function $R(\boldsymbol{w})$ is called the generalization error, since it is evaluated by the expectation of $l(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})$ where the test pair $(\boldsymbol{x},\boldsymbol{y})$ is newly generated by $N(\boldsymbol{w}_0)$. It is easy to show

$$R(\boldsymbol{w}) = H_0 + D(\boldsymbol{w}_0||\boldsymbol{w}) \tag{5}$$

where

$$H_0 = -E_0[\log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)] \tag{6}$$

is the entropy of the teacher network and

$$D(\boldsymbol{w}_0||\boldsymbol{w}) = E_0\left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)}{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})}\right] \tag{7}$$

is the Kullback–Leibler divergence from probability distribution $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)$ to $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$ or the divergence of $N(\boldsymbol{w})$ from $N(\boldsymbol{w}_0)$. Obviously, $D(\boldsymbol{w}_0||\boldsymbol{w}) \geq 0$ and the equality holds when, and only when, $\boldsymbol{w} = \boldsymbol{w}_0$. Hence, the risk $R(\boldsymbol{w})$ measures the divergence between the true distribution $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)$ and the distribution $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$ of $N(\boldsymbol{w})$ except for a constant term H_0 which denotes the stochastic uncertainty of the teacher machine itself. Minimizing $R(\boldsymbol{w})$ is equivalent to minimizing $D(\boldsymbol{w}_0||\boldsymbol{w})$, and the minimum is attained at $\boldsymbol{w} = \boldsymbol{w}_0$.

In the case of multilayer perceptrons with additive Gaussian noise, the output $\boldsymbol{y}$ is written as

$$\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{w}) + \boldsymbol{n} \tag{8}$$

where $f(\boldsymbol{x}; \boldsymbol{w})$ is the analog function calculated by the multilayer perceptron $N(\boldsymbol{w})$ with a set of parameters $\boldsymbol{w}$ and $\boldsymbol{n}$ is Gaussian noise. When its components n_i are independent subject to $N(0, \sigma^2)$, we have

$$p(\boldsymbol{y}|\boldsymbol{x}; w) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^k \exp\left\{-\frac{1}{2\sigma^2}|\boldsymbol{y} - f(\boldsymbol{x}; \boldsymbol{w})|^2\right\} \tag{9}$$

Hence the loss is the ordinary squared error

$$l(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \frac{1}{2\sigma^2}|\boldsymbol{y} - f(\boldsymbol{x}; \boldsymbol{w})|^2 + c(\boldsymbol{x}) \tag{10}$$

where $c(\boldsymbol{x})$ does not depend on $\boldsymbol{w}$.

III. Asymptotic Analysis of Learning

The maximum likelihood estimator (m.l.e.) $\hat{\boldsymbol{w}}_t$ maximizes the likelihood $\Pi_{i=1}^t p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w})$ of producing the training set D_t, or equivalently minimizing the empirical risk function

$$R_{\text{train}}(\boldsymbol{w}) = -\frac{1}{t}\sum_{i=1}^{t} \log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}). \tag{11}$$

This empirical risk is called the training error since it is evaluated by the training examples themselves. In order to avoid confusion, $R(\boldsymbol{w})$ is denoted by $R_{\text{gen}}(\boldsymbol{w})$ when necessary.

The asymptotic theory of statistics proves that the m.l.e. is asymptotically subject to the normal distribution with mean $\boldsymbol{w}_0$ and variance G^{-1}/t

$$\hat{\boldsymbol{w}}_t - \boldsymbol{w}_0 \sim N\left(0, \frac{1}{t}G^{-1}\right)$$

under certain regularity conditions, where G^{-1} is the inverse of the Fisher information matrix G.

By expanding the risk functions, we have the following asymptotic evaluations of $R(\boldsymbol{w})$ and $R_{\text{train}}(\boldsymbol{w})$ in the neighborhood of $\boldsymbol{w}_0$.

Lemma 1: When $\boldsymbol{w}$ belongs to the $(1/\sqrt{t})$-neighborhood of $\boldsymbol{w}_0$

$$R_{\text{gen}}(\boldsymbol{w}) = H_0 + \tfrac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_0)^T G(\boldsymbol{w}_0)(\boldsymbol{w} - \boldsymbol{w}_0) + \mathcal{O}(t^{-3/2}) \tag{12}$$

$$R_{\text{train}}(\boldsymbol{w}) = \hat{H}_0 - \tfrac{1}{2}(\hat{\boldsymbol{w}}_t - \boldsymbol{w}_0)^T G(\boldsymbol{w}_0)(\hat{\boldsymbol{w}}_t - \boldsymbol{w}_0) + \tfrac{1}{2}(\boldsymbol{w} - \hat{\boldsymbol{w}}_t)^T G(\boldsymbol{w}_0)(\boldsymbol{w} - \hat{\boldsymbol{w}}_t) + \mathcal{O}_p(t^{-3/2}) \tag{13}$$

where $\boldsymbol{w}^T$ denotes the transpose of the column vector $\boldsymbol{w}$ and

$$\hat{H}_0 = -\frac{1}{t}\sum_{i=1}^{t} \log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}_0) = R_{\text{train}}(\boldsymbol{w}_0)$$

$$\langle \hat{H}_0 \rangle = H_0, \quad \hat{H}_0 = H_0 + \mathcal{O}_p(1/\sqrt{t})$$

and $\langle \quad \rangle$ represents the average with respect to the distribution of the sample D_t [3].

The relation (12) is the Taylor expansion of (4), where the identity

$$E_0\left[\frac{\partial^2}{\partial w_i \partial w_j} \log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)\right] = -G(\boldsymbol{w}_0)$$

is used. The proof of (13) is given in Appendix A.

By putting $\boldsymbol{w} = \hat{\boldsymbol{w}}_t$ in (12) and (13), we have the asymptotic evaluations of the generalization and training errors of $N(\hat{\boldsymbol{w}}_t)$. They depend on the examples D_t from which the m.l.e. is calculated. We denote by $\langle \quad \rangle$ the average with respect to the distribution of the sample D_t which determines $\hat{\boldsymbol{w}}_t$. We then obtain the following universal relation concerning the generalization error and training error. This was first proved by [3]. A similar but different universal property is proved by Amari [2] for deterministic dichotomy machines.

Corollary 1: For the m.l.e. $\hat{\boldsymbol{w}}_t$, the average training error and generalization error are asymptotically evaluated by the AIC[3] type criterion [3]

$$\langle R_{\text{gen}}(\hat{\boldsymbol{w}}_t) \rangle = H_0 + \frac{m}{2t} + \mathcal{O}(t^{-3/2}) \tag{14}$$

$$\langle R_{\text{train}}(\hat{\boldsymbol{w}}_t) \rangle = H_0 - \frac{m}{2t} + \mathcal{O}(t^{-3/2}) \tag{15}$$

independently of the architecture of networks, where m is the number of modifiable parameters (dimension number of $\boldsymbol{w}$) and t is the number of training patterns.

Murata *et al.* [22], [23] obtained more general versions and proposed the NIC (network information criterion) for model selection by generalizing the AIC [5].

Let us consider the gradient descent learning rule, where the parameter $\hat{\boldsymbol{w}}(n)$ at the nth step is modified by

$$\hat{\boldsymbol{w}}(n+1) = \hat{\boldsymbol{w}}(n) - \varepsilon \frac{\partial R_{\text{train}}(\hat{\boldsymbol{w}}(n))}{\partial \boldsymbol{w}} \tag{16}$$

where ε is a small positive constant. More precisely, $\hat{\boldsymbol{w}}(n)$ should be denoted by $\hat{\boldsymbol{w}}_t(n)$ since it depends on D_t, but we omit the subscript t for the sake of simplicity. This is batch learning where all the training examples are used for each

[3] Akaike's information criterion.

iteration of modifying $\hat{\boldsymbol{w}}(n)$. We can alternatively use on-line learning[4]

$$\hat{\boldsymbol{w}}(n+1) = \hat{\boldsymbol{w}}(n) + \varepsilon \frac{\partial \log p(\boldsymbol{x}_n, \boldsymbol{y}_n; \boldsymbol{w}_n)}{\partial \boldsymbol{w}} \tag{17}$$

where $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ is randomly chosen at each step from the training data set D_t. The batch process is deterministic and $\hat{\boldsymbol{w}}(n)$ converges to $\hat{\boldsymbol{w}}$,[5] provided the initial $\boldsymbol{w}(0)$ is included in its basin of attraction. For large $n, \hat{\boldsymbol{w}}(n)$ is in the $(1/\sqrt{t})$-neighborhood of $\hat{\boldsymbol{w}}$, and the gradient of R_{train} is approximated from (13) as

$$\frac{\partial R_{\text{train}}(\hat{\boldsymbol{w}}(n))}{\partial \boldsymbol{w}} = G(\boldsymbol{w}_0)\{\hat{\boldsymbol{w}}(n) - \hat{\boldsymbol{w}}\} + \mathcal{O}_p(t^{-3/2}).$$

Hence, by neglecting the term of order $1/t^{3/2}$, (16) is approximated by

$$\hat{\boldsymbol{w}}(n+1) = \hat{\boldsymbol{w}}(n) - \varepsilon G\{\hat{\boldsymbol{w}}(n) - \hat{\boldsymbol{w}}\}.$$

This gives the asymptotic evaluation

$$\hat{\boldsymbol{w}}(n) = (I - \varepsilon G)^{n-n'}\{\hat{\boldsymbol{w}}(n') - \hat{\boldsymbol{w}}\} + \hat{\boldsymbol{w}}$$

where I is the identity matrix and $n'(<n)$ is assumed to be large.

In order to make the analysis easier, we take the coordinate system such that the Fisher information matrix G is equal to the identity matrix I

$$G(\boldsymbol{w}_0) = I \tag{18}$$

at $\boldsymbol{w}_0$. This is possible without loss of generality, and the results of the following analysis are the same whichever coordinate system we use. Under this coordinate system, we have

$$\hat{\boldsymbol{w}}(n) = (1 - \varepsilon)^{n-n'}\{\hat{\boldsymbol{w}}(n') - \hat{\boldsymbol{w}}\} + \hat{\boldsymbol{w}}$$

showing that the trajectory $\hat{\boldsymbol{w}}(n)$ linearly approaches $\hat{\boldsymbol{w}}$ in the neighborhood of $\hat{\boldsymbol{w}}$.

We call the trajectory $\hat{\boldsymbol{w}}(n)$ a ray which approaches $\hat{\boldsymbol{w}}$ linearly when n is large. An interesting question is from which direction the ray $\hat{\boldsymbol{w}}(n)$ approaches $\hat{\boldsymbol{w}}$. Even if the initial $\hat{\boldsymbol{w}}(0)$ is uniformly distributed, we cannot say that $\hat{\boldsymbol{w}}(n)$ approaches $\hat{\boldsymbol{w}}$ isotropically, since dynamics (16) is highly nonlinear in an early stage of learning. In other words, the distribution of $\hat{\boldsymbol{w}}(n')$ is not isotropic but may have biased directions.

Although the rays are not isotropically distributed around $\boldsymbol{w}_0$, the quantity $\hat{\boldsymbol{w}}$ is isotropically distributed around $\boldsymbol{w}_0$ because G is put equal to I at $\boldsymbol{w}_0$. This implies that the relative direction of a ray with respect to the isotropically distributed $\hat{\boldsymbol{w}}$ is isotropically distributed. This gives us the following lemma, which helps to calculate $\langle R_{\text{gen}}(\hat{\boldsymbol{w}}(n))\rangle$ and $\langle R_{\text{train}}(\hat{\boldsymbol{w}}(n))\rangle$.

Lemma 2: Although $\hat{\boldsymbol{w}}(n)$ does not necessarily approach $\hat{\boldsymbol{w}}$ isotropically, the ensemble averages $\langle R_{\text{gen}}(\hat{\boldsymbol{w}}(n))\rangle$ and $\langle R_{\text{train}}(\hat{\boldsymbol{w}}(n))\rangle$ are the same as those calculated by assuming that $\hat{\boldsymbol{w}}(n)$ approaches $\hat{\boldsymbol{w}}$ isotropically.

[4]Its dynamical behavior was studied by Amari [1], Heskes and Kappen [17], and recently by Barkai *et al.* [8] and Solla and Saad [28].

[5]Or to $\hat{\boldsymbol{w}}_t$, but the subscript is omitted hereafter.

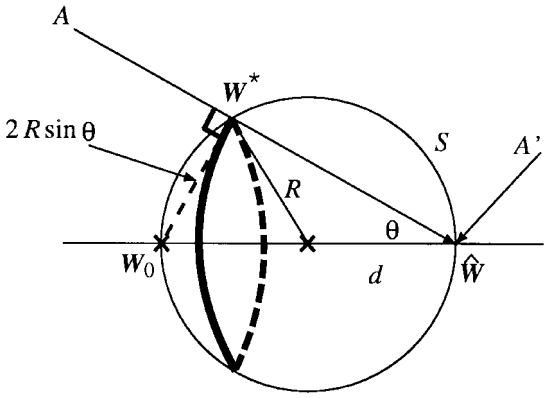


Fig. 1. Geometrical picture to determine the optimal stopping point $\boldsymbol{w}^*$.

Proof: The distribution of the ray $\hat{\boldsymbol{w}}(n)$ is not necessarily isotropic but $\hat{\boldsymbol{w}}$ is distributed isotropically around $\boldsymbol{w}_0$. The average $\langle R_{\text{gen}}(\hat{\boldsymbol{w}}(n))\rangle$ is the expectation with respect to the unknown initial $\boldsymbol{w}(0)$, which determines the ray $\hat{\boldsymbol{w}}(n)$ and with respect to D_t, which determines $\hat{\boldsymbol{w}}$. Let us fix a ray and take average with respect to D_t, that is with respect to $\hat{\boldsymbol{w}}$. Since the relative direction between the fixed ray and all possible $\hat{\boldsymbol{w}}$ is isotropically distributed, it follows that taking the average with respect to $\hat{\boldsymbol{w}}$ for a fixed ray is equivalent to taking the average with respect to isotropically distributed rays and a fixed $\hat{\boldsymbol{w}}$. Therefore we may calculate the averages by using the isotropically distributed rays instead of the isotropically distributed $\hat{\boldsymbol{w}}$. ■

IV. Virtual Optimal Stopping Rule

When the parameter $\hat{\boldsymbol{w}}(n)$ approaches $\hat{\boldsymbol{w}}$ as learning goes on, $n = 1, 2, \cdots$, the generalization behavior of network $N(\hat{\boldsymbol{w}}(n))$ is evaluated by the sequence

$$R(n) = R_{\text{gen}}(\hat{\boldsymbol{w}}(n)), \qquad n = 1, 2, \cdots. \tag{19}$$

It is believed that $R(n)$ decreases in an early period of learning but it increases later. Therefore, there exists an optimal stopping time n at which $R(n)$ is minimized. The stopping time n_{opt} is a random variable depending on $\hat{\boldsymbol{w}}$ and the initial $\boldsymbol{w}(0)$. We evaluate the ensemble average of $\langle R(n_{\text{opt}})\rangle$.

The true $\boldsymbol{w}_0$ and the m.l.e. $\hat{\boldsymbol{w}}$ are in general different, and their distance is of order $1/\sqrt{t}$. Let us compose a sphere S of which the center is at $(1/2)(\boldsymbol{w}_0 + \hat{\boldsymbol{w}})$ and which passes through both $\boldsymbol{w}_0$ and $\hat{\boldsymbol{w}}$, as is shown in Fig. 1. Its diameter is denoted by d where

$$d^2 = |\hat{\boldsymbol{w}} - \boldsymbol{w}_0|^2 \tag{20}$$

and

$$\begin{aligned} E_0[d^2] &= E_0[(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)^T(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)] \\ &= E_0[\text{tr}(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)^T] \approx \frac{1}{t}\text{tr}I = \frac{m}{t}. \end{aligned} \tag{21}$$

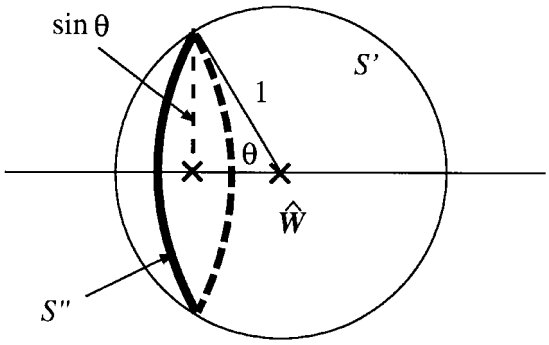


Fig. 2. Distribution of the angle θ.

Let A be the ray, that is the trajectory $\hat{\boldsymbol{w}}(n)$ starting at $\hat{\boldsymbol{w}}(0)$, which is far from the neighborhood of $\boldsymbol{w}_0$. The optimal stopping point $\boldsymbol{w}^*$ that minimizes

$$R(n) = H_0 + \tfrac{1}{2}|\hat{\boldsymbol{w}}(n) - \boldsymbol{w}_0|^2 + \mathcal{O}_p(t^{-3/2}) \tag{22}$$

is given by the following lemma.

Lemma 3: The optimal stopping point $\boldsymbol{w}^*$ is asymptotically the first intersection of the ray A and the sphere S.

Proof: Since $\boldsymbol{w}^*$ is the point on A such that $\boldsymbol{w}_0 - \boldsymbol{w}^*$ is orthogonal to A, it lies on the sphere S (Fig. 1). When ray A' is approaching $\hat{\boldsymbol{w}}$ from the opposite side of $\boldsymbol{w}_0$ (the right-hand side in the figure), the first intersection point is $\hat{\boldsymbol{w}}$ itself. In this case, the optimal stopping never occurs until it converges to $\hat{\boldsymbol{w}}$. ■

Let θ be the angle between the ray A and the diameter $\boldsymbol{w}_0 - \hat{\boldsymbol{w}}$ of the sphere S. We now show the distribution of θ when the rays are isotropically distributed relative to $\hat{\boldsymbol{w}}$.

Lemma 4: When ray A is approaching $\hat{\boldsymbol{w}}$ from the side in which $\boldsymbol{w}_0$ is included, the probability density of $\theta(0 \le \theta \le \pi/2)$, is given by

$$r(\theta) = \frac{1}{I_{m-2}} \sin^{m-2}\theta \tag{23}$$

where

$$I_m = \int_0^{\pi/2} \sin^m \theta \, d\theta.$$

Proof: Let us compose a unit $(m-1)$-dimensional sphere S' centered at $\hat{\boldsymbol{w}}$ (Fig. 2). Since ray A is considered to approach $\hat{\boldsymbol{w}}$ isotropically (Lemma 2), its intersection to S' is uniformly distributed on S', when A is approaching from the side of $\boldsymbol{w}_0$. Let us consider the area on S' such that the angles are between θ and $\theta + d\theta$. Then, the area is an $(m-2)$-dimensional sphere S'' on S' whose radius is $\sin\theta$ (Fig. 2). Hence, its volume is $c_{m-2}\sin^{m-2}\theta$, where c_{m-2} is the volume of a unit $(m-2)$-sphere. By normalization, the density of θ is

$$r(\theta) = \frac{1}{I_{m-2}} \sin^{m-2}\theta, \quad 0 \le \theta \le \frac{\pi}{2}. \qquad ■$$

Now we have the following theorem.

Theorem 1: The average generalization error at the optimal stopping point is asymptotically given by

$$\langle R(n_{\rm opt})\rangle \simeq H_0 + \frac{1}{2t}\left(m - \frac{1}{2}\right). \tag{24}$$

Proof: When ray A is at angle $\theta(0 \le \theta < \pi/2)$, the optimal stopping point $\boldsymbol{w}^*$ is on the sphere S. It is easily shown that

$$|\boldsymbol{w}^* - \boldsymbol{w}_0| = d\sin\theta.$$

This is the case where A approaches $\hat{\boldsymbol{w}}$ from the left-hand side in Fig. 1, which occurs with probability 0.5, and the average of $(d\sin\theta)^2$ is

$$E_0[(d\sin\theta)^2] = \frac{1}{I_{m-2}} E_0[d^2] \int_0^{\pi/2} \sin^2\theta \sin^{m-2}\theta \, d\theta$$
$$= \frac{m}{t}\frac{I_m}{I_{m-2}}.$$

Since we have

$$\frac{I_m}{I_{m-2}} = 1 - \frac{1}{m}$$
$$E_0[(d\sin\theta)^2] = \frac{m}{t}\left(1 - \frac{1}{m}\right).$$

When θ is $\pi/2 \le \theta \le \pi$, that is A approaches $\hat{\boldsymbol{w}}$ from the opposite side, it does not stop until it reaches $\hat{\boldsymbol{w}}$, so that

$$|\boldsymbol{w}^* - \boldsymbol{w}_0|^2 = |\hat{\boldsymbol{w}} - \boldsymbol{w}_0| = d^2.$$

This also occurs with probability 0.5. Hence, from (22), we proved the theorem. ■

The theorem shows that, when we could know the optimal stopping time $n_{\rm opt}$ for each trajectory, the generalization error decreases by $1/4t$, which has an effect of decreasing the effective dimensions by $1/2$. This effect is negligible when m is large. The optimal stopping time is of order $\log t$. However, it is impossible to know the optimal stopping time. If we stop learning at an estimated optimal time $\hat{n}_{\rm opt}$, we have some gain when ray A is from the same side as $\boldsymbol{w}_0$ but we have some loss when ray A is from the opposite direction.

Wang *et al.* [30] calculated $\langle R(n)\rangle$ in the case of linear Φ-machines and defined the optimal average stopping time $\overline{n}_{\rm opt}$ that minimizes $\langle R(n)\rangle$. This is different from the present $n_{\rm opt}$, since our $n_{\rm opt}$ is defined for each trajectory A. Hence it is a random variable depending on $\boldsymbol{w}(0)$ and $\hat{\boldsymbol{w}}$. Our average

$$\langle R(n_{\rm opt})\rangle = \langle R(\hat{\boldsymbol{w}}(n_{\rm opt}))\rangle$$

is different from $\langle R(\overline{n}_{\rm opt})\rangle$, since $\overline{n}_{\rm opt}$ is common to all the trajectories while $n_{\rm opt}$ are different. We can show

$$\langle R(n_{\rm opt})\rangle < \langle R(\overline{n}_{\rm opt})\rangle.$$

We can prove

$$\langle R(\overline{n}_{\rm opt})\rangle = \langle R(\hat{\boldsymbol{w}})\rangle - \mathcal{O}(t^{-2})$$

in agreement with Wang *et al.* [30]. This shows that the gain becomes much smaller by using the average stopping time $\overline{n}_{\rm opt}$. However, the point is that there is no direct means to estimate $n_{\rm opt}$ except for cross-validation. Hence, we need to analyze cross-validation early stopping.

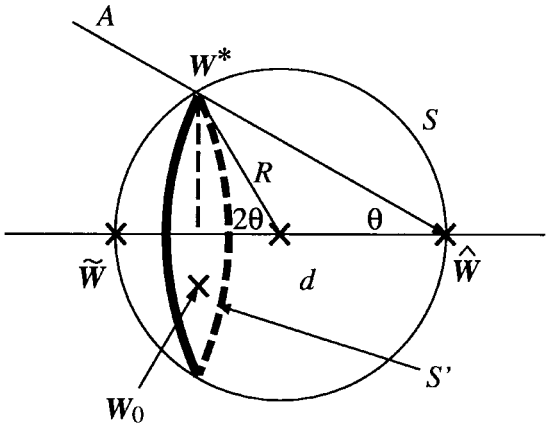


Fig. 3. Optimal stopping point $\boldsymbol{w}^*$ by cross-validation.

V. Optimal Stopping by Cross-Validation

In order to find the optimal stopping time for each trajectory, an idea is to divide the available examples into two disjoint sets; the training set for learning and the cross-validation set for evaluating the generalization error. The training error monotonically decreases with iterations, but according to the folklore the generalization error evaluated by the cross-validation set decreases in an early period but it increases after a critical period. This gives the optimal time to stop training. The present section studies two fundamental problems: 1) Given t examples, how many examples should be used in the training set and how many in the cross-validation set? 2) How much gain can one expect by the above cross-validated stopping?

Let us divide t examples into rt examples of the training set and $r't$ examples of the cross-validation set, where

$$r + r' = 1. \tag{25}$$

Let $\hat{\boldsymbol{w}}$ be the m.l.e. from rt training examples, and let $\tilde{\boldsymbol{w}}$ be the m.l.e. from the other $r't$ cross-validation examples, that is $\hat{\boldsymbol{w}}$ and $\tilde{\boldsymbol{w}}$ minimize the training error function

$$R_{\text{train}}(\boldsymbol{w}) = -\frac{1}{rt}\sum_i \log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}) \tag{26}$$

and cross-validation error function

$$R_{\text{cv}}(\boldsymbol{w}) = -\frac{1}{r't}\sum_i \log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}) \tag{27}$$

respectively, where summations are taken over rt training examples and $r't$ cross-validation examples. Since the training examples and cross-validation examples are independent, both $\hat{\boldsymbol{w}}$ and $\tilde{\boldsymbol{w}}$ are asymptotically subject to independent normal distributions with mean $\boldsymbol{w}_0$ and covariance matrices $G^{-1}/(rt)$ and $G^{-1}/(r't)$, respectively.

Let us compose the triangle with vertices $\boldsymbol{w}_0, \hat{\boldsymbol{w}}$ and $\tilde{\boldsymbol{w}}$ (Fig. 3). The trajectory A starting at $\boldsymbol{w}(0)$ enters $\hat{\boldsymbol{w}}$ linearly in the neighborhood of $\hat{\boldsymbol{w}}$. The point $\boldsymbol{w}^*$ on the trajectory A which minimizes the cross-validation error is the point on A that is closest to $\tilde{\boldsymbol{w}}$, since the cross-validation error defined by (27) can be expanded from (13) as

$$R_{\text{cv}}(\boldsymbol{w}) \simeq H_0 - \tfrac{1}{2}|\tilde{\boldsymbol{w}} - \boldsymbol{w}_0|^2 + \tfrac{1}{2}|\boldsymbol{w} - \tilde{\boldsymbol{w}}|^2. \tag{28}$$

Let S be the sphere whose center is at $(\hat{\boldsymbol{w}}+\tilde{\boldsymbol{w}})/2$ and which passes through both $\hat{\boldsymbol{w}}$ and $\tilde{\boldsymbol{w}}$. Its diameter is given by

$$d = |\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}}|. \tag{29}$$

Then, the optimal stopping point $\boldsymbol{w}^*$ is given by the intersection of the trajectory A and sphere S. When the trajectory comes from the opposite side of $\tilde{\boldsymbol{w}}$ (right-hand side in the figure), it does not intersect S until it converges to $\hat{\boldsymbol{w}}$, so that the optimal point is $\boldsymbol{w}^* = \hat{\boldsymbol{w}}$ in this case.

The generalization error of $\boldsymbol{w}^*$ is given by (12), so that we calculate the expectation of $|\boldsymbol{w}^* - \boldsymbol{w}_0|^2$.

Lemma 5:

$$E[|\boldsymbol{w}^* - \boldsymbol{w}_0|^2] = \frac{m}{tr} - \frac{1}{2t}\left(\frac{1}{r} - \frac{1}{r'}\right).$$

Proof is given in Appendix B. It is immediate to show Lemma 6.

Lemma 6: The average generalization error by the optimal cross-validated early stopping is asymptotically

$$\langle R(\boldsymbol{w}^*, r)\rangle \simeq H_0 + \frac{2m-1}{4rt} + \frac{1}{4r't}. \tag{30}$$

We can then calculate the optimal division rate r_{opt} of examples which minimizes the generalization error.

Theorem 2: The average generalization error is minimized asymptotically at

$$r_{\text{opt}} = 1 - \frac{\sqrt{2m-1}-1}{2(m-1)}. \tag{31}$$

The theorem shows the optimal division of examples into training and cross-validation sets. When m is large

$$r_{\text{opt}} \simeq 1 - \frac{1}{\sqrt{2m}} \tag{32}$$

showing that only $(1/\sqrt{2m}) \times 100\%$ of examples are to be used for cross-validation testing and remaining most examples are used for training. When $m = 100$, this shows that 93% of examples are to be used for training and only 7% are to be kept for cross-validation. From this we obtain

Theorem 3: The asymptotically optimal generalization error is

$$\langle R(\boldsymbol{w}^*, r_{\text{opt}})\rangle \simeq H_0 + \frac{1}{4t}(\sqrt{2m-1}+1)^2. \tag{33}$$

When m is large, we have

$$\langle R(\boldsymbol{w}^*, r_{\text{opt}})\rangle \simeq H_0 + \frac{m}{2t}\left(1 + \sqrt{\frac{2}{m}}\right). \tag{34}$$

This shows that the generalization error increases slightly by cross-validation early stopping compared with learning which uses all the examples for training. That is

$$\langle D(\boldsymbol{w}_0||\boldsymbol{w}^*)\rangle > \langle D(\boldsymbol{w}_0||\hat{\boldsymbol{w}})\rangle \tag{35}$$

for the optimal cross-validated $\boldsymbol{w}^*$ and the m.l.e. $\hat{\boldsymbol{w}}$ based on all the examples without cross-validation.

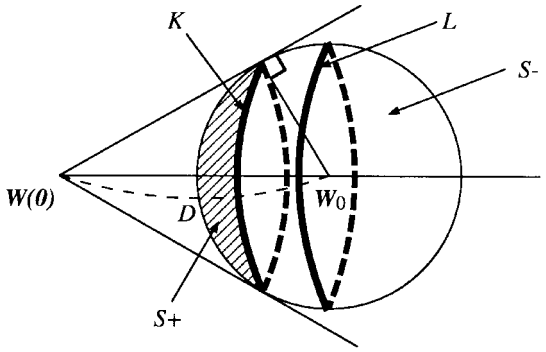


Fig. 4. Geometrical picture for the intermediate range.

VI. Intermediate Range

So far we have seen that cross-validation stopping is asymptotically not effective. Now we would like to discuss from several viewpoints why cross-validation early stopping is effective in the intermediate range. Note however that our explanations are not mathematically rigorous, but rather sketch three possible lines along which our theory can be generalized for the intermediate range: 1) a geometrical picture; 2) the distribution of the initial $\boldsymbol{w}(0)$; and 3) the nonlinearities of the trajectories.

In order to have intuitive understanding, we draw another picture (Fig. 4). Here, $\hat{\boldsymbol{w}}$ is distributed uniformly on the sphere S whose center is the true value. Let $\boldsymbol{w}(0)$ be the initial weight and let D be the distance between $\boldsymbol{w}(0)$ and $\boldsymbol{w}_0$. We draw tangent rays from $\boldsymbol{w}(0)$ to the sphere S. Then, the tangent points on S form an $(m-1)$-dimensional sphere K that divides S into two parts S_+ (shaded, left side) and S_- (right side). When $\hat{\boldsymbol{w}}$ lies on S_+, early stopping is not necessary, but when $\hat{\boldsymbol{w}}$ lies on S_- then early stopping improves the solution.

In the asymptotic range where t is very large, whatever D is, it is far larger than the radius $1/\sqrt{t}$ of S. This implies that $\boldsymbol{w}(0)$ is located almost infinitely far, so that the $(m-1)$-sphere K dividing S into S_+ and S_- is equal to the $(m-1)$-sphere L which is the vertical cut of S (the cut at $\boldsymbol{w}_0$ orthogonal to the line connecting $\boldsymbol{w}(0)$ and $\boldsymbol{w}_0$). In this case, S is divided into two parts S_+ and S_- with an equal volume. Moreover, when m is large, the most volume of S is concentrated in a neighborhood of L so that the effect of early stopping is not remarkable.

In the intermediate range where D is not so large, the sphere K is different from L and is located on the left side of L. Since most volume is concentrated in a neighborhood of L, the measure of S_+ is negligible in this case. This implies that early stopping improves $\hat{\boldsymbol{w}}$ with a probability close to one. In the extreme case where D is very small and $\boldsymbol{w}(0)$ is inside S, immediate stopping without any training is the best strategy.

This shows that, when t is not asymptotically large, we cannot neglect the distribution of the initial $\boldsymbol{w}(0)$ which is not so far from $\boldsymbol{w}_0$. Let $W = \{\boldsymbol{w}\}$ be the parameter space. What is the natural distribution of $\boldsymbol{w}_0$ and $\boldsymbol{w}(0)$? If we assume a uniform distribution over a very large convex region, D is very large. However, a natural prior distribution is the Jeffrey noninformative distribution[6] which is given by $\sqrt{g(\boldsymbol{w})}/V$ where $g = \det|G|$ is the Riemannian volume element of W. In most neural network architectures, the volume $V = \int \sqrt{g(\boldsymbol{w})}\, d\boldsymbol{w}$ is finite and this implies that the effect of the distribution of initial $\boldsymbol{w}(0)$ cannot be neglected when t is not asymptotically large.

It is possible to construct a theory by taking D into account. However, for the theory to be valid where t is not asymptotically large, the nonlinear learning trajectories cannot be neglected and we need higher-order corrections to the asymptotic properties of the estimator $\hat{\boldsymbol{w}}$ (cf. Amari[7]).

VII. Simulations

We use standard feedforward classifier networks with N inputs, H sigmoid hidden units and M softmax outputs (classes). The m network parameters $\boldsymbol{w}$ consist of biases $\vartheta = \{\vartheta_j^H, \vartheta_j^O\}$ and weights $\boldsymbol{w} = \{w_{ij}^H, w_{jk}^O\}$. The input layer is connected to the hidden layer via $\boldsymbol{w}^H$, the hidden layer is connected to the output layer via $\boldsymbol{w}^O$, and no shortcut connections are present. The output activity O_l of the lth output unit is calculated via the softmax squashing function

$$p(\boldsymbol{y} = C_l|\boldsymbol{x};\boldsymbol{w}) = O_l = \begin{cases} \dfrac{\exp(h_l)}{1+\sum_k \exp(h_k)}, & l = 1,\cdots,M-1, \\ \dfrac{1}{1+\sum_k \exp(h_k)}, & l = M \end{cases}$$

where $h_l = \Sigma_{j=1}^H w_{lj}^O s_j - \vartheta_l^O (l = 1,\cdots,M-1)$ is the local field potential and

$$s_j = \left[1+\exp\left(-\sum_{k=1}^N w_{jk}^H x_k + \vartheta_j^H\right)\right]^{-1}, \qquad j = 1,\cdots,H.$$

is the activity of the j-th hidden unit, given input $\boldsymbol{x}$.

Each output O_l codes the *a posteriori* probability of being in class C_l. Although the network is completely deterministic, it is constructed to approximate class conditional probabilities ([13]).

Therefore, each randomly generated teacher $\boldsymbol{w}_0$ represents by construction a multinomial probability distribution $q(C_l|\boldsymbol{x},\boldsymbol{w}_0) = \text{Prob}\{\boldsymbol{x} \in C_l\}$ over the classes C_l $(l = 1\cdots M)$ given a random input $\boldsymbol{x}$. We use the same network architecture for teacher and student. Thus, we assume that the model is faithful, i.e., the teacher distribution can be exactly represented by a student $q(C_l|\boldsymbol{x}) = p(C_l|\boldsymbol{x},\boldsymbol{w}_0)$.

A training and cross-validation set of the form $D_t = \{(\boldsymbol{x}^p, c^p)|p = 1,\cdots,t\}$ is generated randomly, by drawing samples of $\boldsymbol{x}$ from a uniform distribution and forward propagating $\boldsymbol{x}^p$ through the teacher network. Then, according to the teachers' outputs $q(C_l^p|\boldsymbol{x}^p)$ one output unit is set to one

[6]Note also that the Bayes estimator $\hat{\boldsymbol{w}}_{\text{Bayes}}$ with the Jeffrey prior $\sqrt{g}$ is better than the m.l.e. from the point of view of minimizing Kullback–Leibler divergence, although they are equivalent for large t.

[7]*Differential Genetical Methods in Statistics*. New York: Springer-Verlag, Lecture Notes in Statistics no. 28, 1985.

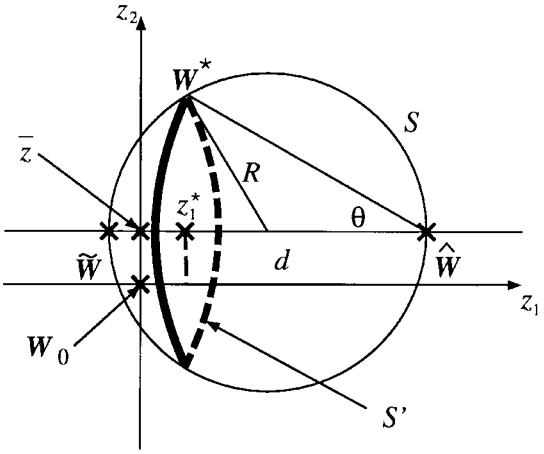


Fig. 5. New coordinate system z.

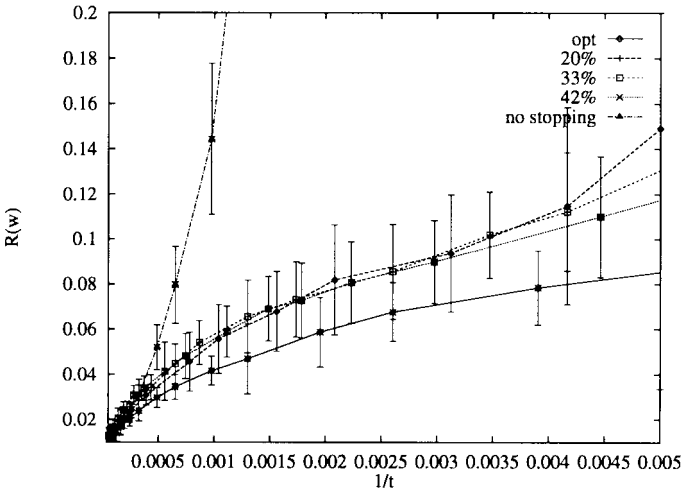
stochastically and all others are set to zero leading to the target vector $\boldsymbol{y}^p = (0, \cdots, 1, \cdots, 0)$. A student network $\boldsymbol{w}$ is then trying to approximate the teacher given the example set D_t. For training the student network $\boldsymbol{w}$ we use—within the backpropagation framework—conjugate gradient descent with a line-search to optimize the training error function (26), starting from some random initial vector. The cross-validation error (27) is measured on the cross-validation examples to stop learning. The average generalization ability (4) is approximately estimated by

$$R_{\text{test}}(t) = -\frac{1}{K}\sum_{i=1}^{K} p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}_0) \log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}) \qquad (36)$$

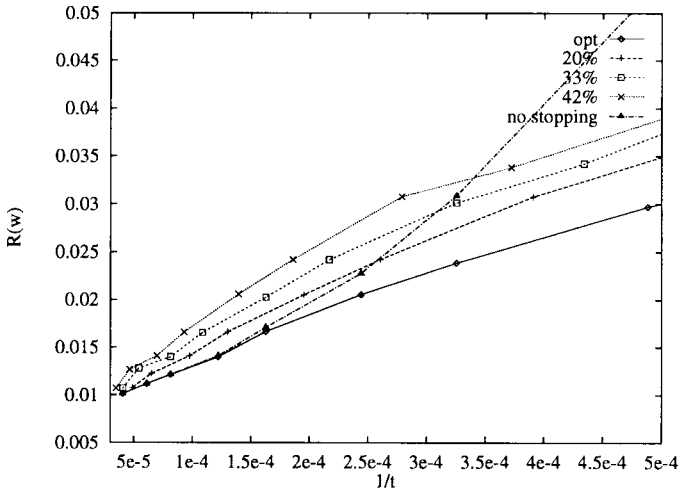
on a large test set ($K = 50\,000$ patterns) and averaged over 128–256 trials.[1] We compare the generalization error for the settings: exhaustive training (no stopping), early stopping (controlled by the cross-validation examples) and optimal stopping (controlled by the large test set). The simulations were performed on a parallel computer (CM5). Every curve in the figures takes about 8 h of computing time on a 128, respectively, 256 partition of the CM5, i.e., we perform 128–256 parallel trials. This setting enabled us to do extensive statistics (cf. [4], [24]).

Fig. 6 shows the results of simulations, where $N = 8, H = 8, (M-1) = 4$, so that the number m of modifiable parameters is $m = (N+1)H + (H+1)(M-1) = 108$. In Fig. 6(a) we see the intermediate range of patterns $t < 30m$ (see [24]), where early stopping improves the generalization ability to a large extent, clearly confirming the folklore mentioned above. From Fig. 6 we note that the learning curves and variances are similar in the intermediate range no matter how the split is choosen. Only as we get to small numbers of patterns ($t < 3m$) we find a growing of the variances for the small splits, which is to be expected.

[1] Several sample sets have been used with changing initial vectors. In each trial a sample of size t is generated, the net is trained starting from a random initialization $\boldsymbol{w}(0)$. As the number of patterns is in subsequent experiments increased to t', the newly generated patterns are added to the old set of t patterns.



(a)



(b)

Fig. 6. Shown is $R(\boldsymbol{w})$ plotted for different sizes r' of the early stopping set for a 8-8-4 classifier network $(N = 8, H = 8, (M-1) = 4)$ (a) in the intermediate and (b) in the asymptotic regime as a function of $1/t$. An early stopping set of 20% means: 80% of the t patterns in the training set are used for training, while 20% of the t patterns are used to control the early stopping. opt. denotes the use of a very large test set (50 000) and no stopping addresses the case where 100% of the training set is used for exhaustive learning.

From Fig. 6(b) we observe clearly, that saturated learning without early stopping is the best in the asymptotic range of $t > 30m$, a range which is due to the limited size of the data sets often unaccessible in practical applications. Cross-validated early stopping does not improve the generalization error here, so that no overtraining is observed on the average in this range. This result confirms similar findings by Sjöberg and Ljung [29]. In the asymptotic area [Fig. 6(b)] we observe that the smaller the percentage of the cross-validation set, which is used to determine the point of early stopping, the better the performance of the generalization ability. Fig. 7 shows that the learning curves for different sizes of the cross-validation set are in good agreement with the theoretical prediction of (30).

Three systematic contributions to the randomness arise 1) random examples; 2) initialization of the student weights; and 3) local minima. The part of the variance given by

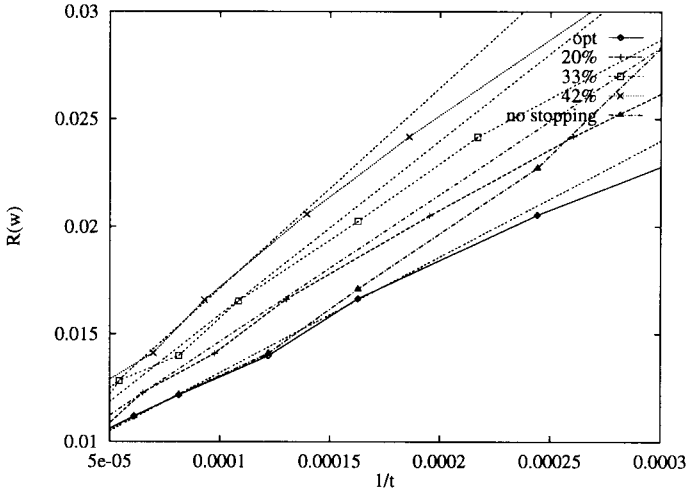


Fig. 7. Shown is $R(\boldsymbol{w})$ from the simulation plotted for different sizes r' of the cross-validation stopping set for a 8-8-4 classifier network in the asymptotic regime as a function of $1/t$. The straight line interpolations are obtained from (30). Note the nice agreement between theory and experiment.

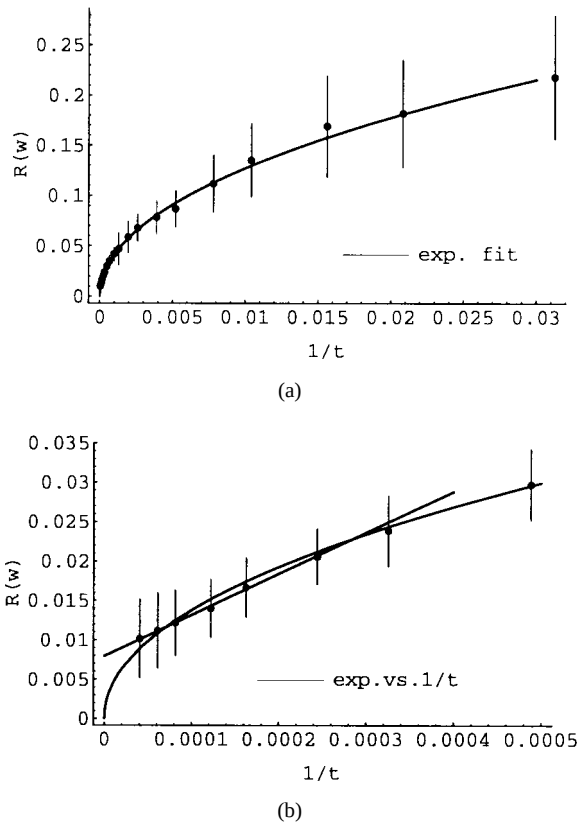


Fig. 8. $R(\boldsymbol{w})$ plotted as a function of $1/t$ for optimal stopping. (a) Exponential fit $t^{-0.49}$ in the whole range of t and (b) comparison of the exponential and the $m/2t$ fit in the asymptotic regime. Shown is data for a 8-8-4 classifier network.

the examples is bounded by the Cramer–Rao bound $(1/t)$. Initialization gives only a small contribution since the results do not change qualitatively if we start with initial weights far enough outside the circle S of Fig. 1. Finally there is a contribution to the variance from local minima distributed around the m.l.e. solution. Note that local minima do not change the validity of our theory. Our simulations essentially measure a coarse distribution of local minima solutions around the m.l.e. and contains a variance due to this fact (for a further discussion on local minima in learning curve simulations see [24]).

In Fig. 8(a) we show an exponential interpolation of the learning curve over the whole range of examples in the situation of optimal stopping (controlled by the large test set). The fitted exponent of $t^{-0.49}$ indicates a $1/\sqrt{t}$ scaling. In the asymptotic range as seen from Fig. 8(a) and (b) the $1/\sqrt{t}$ fit fails to hold and a $m/2t$ scaling gives much better interpolation results. An explanation of this effect can be obtained by information geometrical means: early stopping gives per definition a solution, which is not a global or local minimum of the empirical risk function (11) on the training patterns. Therefore the gradient terms in the likelihood expansion (see Appendix A) are contributing and have to be considered carefully. For an intermediate range the gradient term in the expansion, which scales as $1/\sqrt{t}$ gives the dominant contribution. Asymptotically the gradient term fails to give large contributions because the solution taken is very close to a local minimum and thus a $m/2t$ scaling dominates.

VIII. Concluding Remarks

We proposed an asymptotic theory for overtraining. The analysis treats realizable stochastic neural networks, trained with Kullback–Leibler divergence. However a generalization to unrealizable cases and other loss functions can be obtained along the lines of our reasoning.

It is demonstrated both theoretically and in simulations that *asymptotically* the gain in the generalization error is small if we perform early stopping, even if we have access to the optimal stopping time. For cross-validation stopping we computed the optimal split between training and validation examples and showed for large m that optimally only $r'_{\text{opt}} = 1/\sqrt{2m}$ examples should be used to determine the point of early stopping in order to obtain the best performance. For example, if $m = 100$ this corresponds to using roughly 93% of the t training patterns for training and only 7% for testing where to stop. Yet, even if we use r_{opt} for cross-validation stopping the generalization error is always increased comparing to exhaustive training. Nevertheless note, that this asymptotic range is often unaccessible in practical applications due to the limited size of the data sets.

In the nonasymptotic region our simulations confirm the folklore that cross-validated early stopping always helps to enhance the performance since it decreases the generalization error. We gave an intuitive explanation why this is observed (see Section VI, Fig. 4).

Furthermore for this intermediate range our asymptotic theory provides a guideline which can be used in practice as a heuristic estimate for the choice of the optimal size of the early stopping set in the same sense as NIC ([21]–[23]) is used as a guideline for model selection.

In future studies we would like to extend our theory—along the lines of Section VI—to incorporate the prior distributions of the initial weights and the nonlinear learning trajectories necessary to understand the intermediate range.

APPENDIX A
PROOF OF (13)

Since $R_{\rm train}(\boldsymbol{w})$ is minimized at $\hat{\boldsymbol{w}}$, we have

$$\frac{\partial R_{\rm train}(\hat{\boldsymbol{w}})}{\partial \boldsymbol{w}} = 0. \tag{A.1}$$

Its second derivative

$$\frac{\partial}{\partial \boldsymbol{w}}\frac{\partial}{\partial \boldsymbol{w}} R_{\rm train}(\boldsymbol{w}_0) = -\frac{1}{t}\sum_{i=1}^{t}\frac{\partial}{\partial \boldsymbol{w}}\frac{\partial}{\partial \boldsymbol{w}}\log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}_0)$$

converges to its expectation $G(\boldsymbol{w}_0) = G$ by the law of large numbers. Hence, by expanding $R_{\rm train}(\boldsymbol{w})$ at $\hat{\boldsymbol{w}}$, we have

$$\begin{aligned} R_{\rm train}(\boldsymbol{w}) = & R_{\rm train}(\hat{\boldsymbol{w}}) + \tfrac{1}{2}(\boldsymbol{w}-\hat{\boldsymbol{w}})^T G(\boldsymbol{w}-\hat{\boldsymbol{w}}) \\ & + \mathcal{O}_p(t^{-3/2}). \end{aligned} \tag{A.2}$$

In order to evaluate $R_{\rm train}(\hat{\boldsymbol{w}})$. we expand it as

$$\begin{aligned} R_{\rm train}(\hat{\boldsymbol{w}}) = & R_{\rm train}(\boldsymbol{w}_0) + \frac{\partial R_{\rm train}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}}(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) \\ & + \frac{1}{2}(\hat{\boldsymbol{w}}-\boldsymbol{w}_0)^T G(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) + \mathcal{O}_p(t^{-3/2}). \end{aligned} \tag{A.3}$$

Expanding (A.1) around $\boldsymbol{w}_0$, we have

$$\begin{aligned} & \frac{\partial R_{\rm train}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}}(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) \\ & \quad = -(\hat{\boldsymbol{w}}-\boldsymbol{w}_0)^T G(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) + \mathcal{O}_p(t^{-3/2}). \end{aligned} \tag{A.4}$$

By substituting this in (A.3), we have

$$R_{\rm train}(\hat{\boldsymbol{w}}) = \hat{H}_0 - \tfrac{1}{2}(\hat{\boldsymbol{w}}-\boldsymbol{w}_0)^T G(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) + \mathcal{O}_p(t^{-3/2}).$$

Again by substituting this in (A.2), we have (13).

APPENDIX B
PROOF OF LEMMA 5

We have the triangle $\boldsymbol{w}_0\tilde{\boldsymbol{w}}\hat{\boldsymbol{w}}$, and let S be the $(m-1)$-sphere whose diameter is spanned by $\tilde{\boldsymbol{w}}$ and $\hat{\boldsymbol{w}}$. Let A be a ray approaching $\hat{\boldsymbol{w}}$ from the left-hand side, which intersects S at $\boldsymbol{w}^*$. This is the optimal stopping point. Let θ be the angle between the ray A and the diameter $\tilde{\boldsymbol{w}}\hat{\boldsymbol{w}}$. The probability density $r(\theta)$ of θ is given by (23). Let us consider the set S' of points on S whose angles are between θ and $\theta + d\theta$ when $\boldsymbol{w}^*$ is on S'. It is a $(m-2)$-sphere on S. We then calculate the average of the square of distance $|\boldsymbol{w}^* - \boldsymbol{w}_0|^2$ when $\boldsymbol{w}^*$ is on S' (that is, when the angle is θ)

$$s(\theta) = \int |\boldsymbol{w}^* - \boldsymbol{w}_0|^2 \frac{dS'}{|S'|}$$

where the integration is taken over S'.

This is the case where A is from the same side as $\boldsymbol{w}_0$. For calculation, we introduce an orthogonal coordinate system $\boldsymbol{z} = (z_i)$ in the space of $\boldsymbol{w}$ (Fig. 5) such that 1) its origin is put at $\boldsymbol{w}_0$ so that $\boldsymbol{w}_0 = 0$; 2) its z_1 and z_2 axes are on the plane of the triangle $\boldsymbol{w}_0\hat{\boldsymbol{w}}\tilde{\boldsymbol{w}}$, so that

$$\begin{aligned} \hat{\boldsymbol{w}} &= (\hat{z}_1, \hat{z}_2, 0, \cdots, 0) \\ \tilde{\boldsymbol{w}} &= (\tilde{z}_1, \tilde{z}_2, 0, \cdots, 0); \end{aligned}$$

3) the z_1 and z_2 axes are chosen such that

$$\hat{z}_2 = \tilde{z}_2 = \overline{z};$$

and (4) all the other axes are orthogonal to the triangle. Moreover, we assume $\hat{z}_1 > \tilde{z}_1$. The opposite case is analyzed in the same way, giving the same final result for symmetry reasons.

The sphere S' is written in this coordinate system as

$$S' = \{\boldsymbol{z} | z_1 = z_1^*, (z_2 - \overline{z})^2 + z_3^2 + \cdots + z_m^2 = R^2 \sin^2 2\theta\}$$

where $R\sin 2\theta$ is the radius of the sphere

$$R = \frac{d}{2} = \frac{1}{2}|\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}}|$$

and

$$z_1^* = \frac{\hat{z}_1 + \tilde{z}_1}{2} - R\cos 2\theta.$$

Hence

$$\begin{aligned} s(\theta) &= \int_{S'} \sum_{i=1}^{m} (z_i)^2 \frac{dS'}{|S|} \\ &= \int \{(z_1^*)^2 + (z_2 - \overline{z})^2 + z_3^2 + \cdots + z_m^2 \\ &\quad + 2z_2\overline{z} - \overline{z}^2\} \frac{dS'}{|S'|} \\ &= \int \{(z_1^*)^2 + R^2\sin^2 2\theta + 2(z_2 - \overline{z})\overline{z} + \overline{z}^2\} \frac{dS'}{|S'|} \\ &= \left(\frac{\hat{z}_1 + \tilde{z}_1}{2} - R\cos 2\theta\right)^2 + R^2 \sin^2 2\theta + \overline{z}^2 \\ &= \left(\frac{\hat{z}_1 + \tilde{z}_1}{2}\right)^2 + R^2 - (\hat{z}_1 + \tilde{z}_1)R\cos 2\theta + \overline{z}^2 \\ &= \|\frac{1}{2}(\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}})\|^2 + \|\frac{1}{2}(\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}})\|^2 \\ &\quad - \frac{1}{2}(\hat{z}_1^2 - \tilde{z}_1^2)\cos 2\theta \\ &= \frac{1}{4}\|\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}}\|^2 + \frac{1}{4}\|\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}}\|^2 \\ &\quad - \frac{1}{2}\{\|\hat{\boldsymbol{w}}\|^2 - \|\tilde{\boldsymbol{w}}\|^2\}\cos 2\theta \\ &= \frac{1}{2}\|\hat{\boldsymbol{w}}\|^2 + \frac{1}{2}\|\tilde{\boldsymbol{w}}\|^2 - \frac{1}{2}\{\|\hat{\boldsymbol{w}}\|^2 - \|\tilde{\boldsymbol{w}}\|^2\}\cos 2\theta. \end{aligned}$$

Here, we used the following properties:

$$\begin{aligned} \left(\frac{\hat{z}_1 + \tilde{z}_1}{2}\right) + \overline{z}^2 &= \|(\hat{\boldsymbol{w}} + \tilde{\boldsymbol{w}})\|^2 \\ \hat{z}_1^2 - \tilde{z}_1^2 &= (\hat{z}_1^2 + \overline{z}^2) - (\tilde{z}_1^2 + \overline{z}^2) = \|\hat{\boldsymbol{w}}\|^2 - \|\tilde{\boldsymbol{w}}\|^2 \\ R &= \frac{d}{2} = \frac{1}{2}(\hat{z}_1 - \tilde{z}_1). \end{aligned}$$

From

$$E[\|\hat{\boldsymbol{w}}\|^2] = \frac{m}{rt}, \quad E[\|\tilde{\boldsymbol{w}}\|^2] = \frac{m}{r't}$$

we obtain

$$\begin{aligned} E[s(\theta)] &= \frac{m}{2t}\left\{\frac{1}{r} + \frac{1}{r'} - \left(\frac{1}{r} - \frac{1}{r'}\right)\cos 2\theta\right\} \\ &= \frac{m}{2tr'}\{1 - (1-2r)\cos 2\theta\}. \end{aligned}$$

Therefore, from

$$\frac{1}{I_{m-2}}\int_0^{\pi/2}\cos 2\theta \sin^{m-2}\,d\theta = \frac{1}{I_m}\int_0^{\pi/2}(1-2\sin^2\theta)\sin^{m-2}\theta\,d\theta = \frac{1}{I_{m-2}}(I_{m-2}-2I_m) = 1-2\left(1-\frac{1}{m}\right) = -1+\frac{2}{m}$$

when ray A arrives from the left side, we get

$$E[||\boldsymbol{w}^*-\boldsymbol{w}_0||^2] = \int_0^{\pi/2}E[s(\theta)]r(\theta)\,d\theta = \frac{m}{trr'}\left\{r'-\frac{1-2r}{m}\right\}.$$

When ray A arrives from the right side, $\boldsymbol{w}^*=\hat{\boldsymbol{w}}$, so that

$$E[||\boldsymbol{w}^*-\boldsymbol{w}_0||^2] = E[||\hat{\boldsymbol{w}}-\boldsymbol{w}_0||^2] = \frac{m}{rt}$$

holds. Hence, by averaging the above two, we have

$$\langle R(\boldsymbol{w}^*,r)\rangle = \frac{m}{2trr'}\left\{2r'-\frac{1-2r}{m}\right\} = \frac{m}{tr}-\frac{1-2r}{2trr'} = \frac{m}{tr}-\frac{1}{2t}\left(\frac{1}{r}-\frac{1}{r'}\right).$$

In order to obtain $r_{\rm opt}$, we evaluate

$$\frac{d}{dr}\langle R(\boldsymbol{w}^*,r)\rangle = -\frac{m}{tr^2}+\frac{1}{2tr^2}+\frac{1}{2t}\frac{1}{(1-r)^2}=0$$

giving

$$r_{\rm opt} = 1-\frac{\sqrt{2m-1}-1}{2(m-1)}$$

which can be expanded for large m as

$$r_{\rm opt}\doteq 1-\frac{1}{\sqrt{2m}}.$$

ACKNOWLEDGMENT

The authors would like to thank Y. LeCun, S. Bös, and K. Schulten for valuable discussions. K.-R. M. thanks K. Schulten for warm hospitality during his stay at the Beckman Inst. in Urbana, IL. and for warm hospitality at RIKEN during the completion of this work. The authors acknowledge computing time on the CM5 in Urbana (NCSA) and in Bonn, Germany. They also acknowledge L. Ljung for communicating his results on regularization and early stopping prior to publication.

REFERENCES

[1] S. Amari, "Theory of adaptive pattern classifiers," *IEEE Trans. Electron. Comput.*, vol. EC-16, pp. 299–307, 1967.
[2] ——, "A universal theorem on learning curves," *Neural Networks*, vol. 6, pp. 161–166, 1993.
[3] S. Amari and N. Murata, "Statistical theory of learning curves under entropic loss criterion," *Neural Computa.*, vol. 5, pp. 140–153, 1993.
[4] S. Amari, N. Murata, K.-R. Müller, M. Finke, and H. Yang, "Statistical theory of overtraining—Is cross-validation effective?," in *NIPS'95: Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996.
[5] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716–723, 1974.
[6] D. Barber, D. Saad, and P. Sollich, "Test error fluctuations in finite linear perceptrons," *Neural Computa.* vol. 7, pp. 809–821, 1995.
[7] ——, "Finite-size effects and optimal test size in linear perceptrons," *J. Phys. A*, vol. 28, pp. 1325–1334, 1995.
[8] N. Barkai, H. S. Seung, and H. Sompolinsky, "On-line learning of dichotomies," in *Advances in Neural Information Processing Systems* NIPS 7, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. Cambridge, MA: MIT Press, 1995.
[9] C. M. Bishop, "Regularization and complexity control in feedforward networks," Aston Univ., Tech. Rep. NCRG/95/022, 1995.
[10] S. Bös, W. Kinzel, and M. Opper, "Generalization ability of perceptrons with continuous outputs," *Phys. Rev.*, vol. E47, pp. 1384–1391, 1993.
[11] ——, "Avoiding overfitting by finite temperature learning and cross-validation," in *Proc. Int. Conf. Artificial Neural Networks ICANN'95*, Paris, 1995, pp. 111–116.
[12] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. London: Chapman and Hall, 1993.
[13] M. Finke and K.-R. Müller, "Estimating *a posteriori* probabilities using stochastic network models," in *Proc. 1993 Connectionist Models Summer School*, M. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1994, p. 324.
[14] I. Guyon 1996, "A scaling law for the validation-set training-set ratio," preprint.
[15] M. H. Hassoun, *Fundamentals of Artificial Neural Networks*. Cambridge, MA: MIT Press, 1995.
[16] R. Hecht-Nielsen, *Neurocomputing*. Reading, MA: Addison-Wesley, 1989.
[17] T. Heskes and B. Kappen, Learning process in neural networks, *Phys. Rev.*, vol. A44, pp. 2718–2762, 1991.
[18] M. Kearns, "A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split," in *NIPS'95: Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer and M. E. Hasselmo, Eds. Cambridge, MA: MIT Press, 1996.
[19] A. Krogh and J. Hertz, "Generalization in a linear perceptron in the presence of noise," *J. Phys. A*, vol. 25, pp. 1135–1147, 1992.
[20] J. E. Moody, "The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems," in *NIPS 4: Advances in Neural Information Processing Systems*, J. E. Moody, S. J. Hanson, and R. P. Lippmann, Eds. San Mateo, CA: Morgan Kaufmann, 1992.
[21] N. Murata, S. Yoshizawa, and S. Amari, "A criterion for determining the number of parameters in an artificial neural-network model," in *Artificial Neural Networks*, T. Kohonen, *et al.*, Eds.. Amsterdam, The Netherlands: Elsevier, 1991, pp. 9–14.
[22] ——, "Learning curves, model selection and complexity of neural networks," in *NIPS 5: Advances in Neural Information Processing Systems*, S. J. Hanson *et al.*, Eds. San Mateo, CA: Morgan Kaufmann, 1993.
[23] ——, "Network information criterion—Determining the number of hidden units for an artificial neural-network model," *IEEE Trans. Neural Networks*, vol. 5, pp. 865–872, 1994.
[24] K.-R. Müller, M Finke, N. Murata, K Schulten, and S. Amari, "A numerical study on learning curves in stochastic multilayer feedforward networks," Univ. Tokyo, Tech. Rep. METR 03-95, 1995, also *Neural Computa.*, vol. 8, pp. 1085–1106, 1996.
[25] T. Poggio and F. Girosi, "Regularization algorithms for learning that are equivalent to multilayer networks," *Science*, vol. 247, pp. 978–982, 1990.
[26] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, 1986.
[27] D. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1—Foundations*. Cambridge, MA: MIT Press, 1986.
[28] D. Saad, S. A. Solla, "On-line learning in soft committee machines," *Phys. Rev. E*, vol. 52, pp. 4225–4243, and "Exact solution for on-line learning in multilayer neural networks," *Phys. Rev. Lett.*, vol. 74, pp. 4337–4340, 1995.
[29] J. Sjöberg and L. Ljung, "Overtraining, regularization and searching for minimum with application to neural networks," Linköping Univ., Sweden, Tech. Rep. LiTH-ISY-R-1567, 1994.
[30] C. Wang, S. S. Venkatesh, J. S. Judd, "Optimal stopping and effective machine complexity in learning," to appear, 1994 (revised and extended version of NIPS vol. 6, pp. 303–310, 1995).

[31] V. Vapnik, *Estimation of Dependences Based on Emprirical Data*. New York: Springer-Verlag, 1992.
[32] ——, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.

Shun-ichi Amari (M'71–SM'92–F'94) received the bachelor's degree from the University of Tokyo in 1958 majoring in mathematical engineering, and received the Dr. Eng. degree from the University of Tokyo in 1963.

He was an Associate Professor at Kyushu University, Japan, and a Professor at the University of Tokyo. He is now the Director of the Brain Information Processing Group in the Frontier Research Program at the Institute of Physical and Chemical Research (RIKEN) in Japan. He has been engaged in research in many areas of mathematical engineering and applied mathematics. In particular, he has devoted himself to mathematical foundations of neural network theory. Another main subject of his research is the information geometry that he himself proposed.

Dr. Amari is the former President of the International Neural Network Society and a member of editorial or advisory editorial boards of more than 15 international journals. He was given the IEEE Neural Networks Pioneer Award in 1992, Japan Academy Award in 1995, and IEEE Emanuel R. Piore Award in 1997, among others.

Noboru Murata received the B. Eng., M. Eng., and Dr. Eng in mathematical engineering and information physics from the University of Tokyo in 1987, 1989, and 1992, respectively.

He was a Research Associate at the University of Tokyo, and he was a Visiting Researcher with the Research Institute for Computer Architecture and Software Technology of the German National Research Center for Information Technology (GMD FIRST) from 1995 to 1996 supported by Alexander von Humboldt Foundation. He is currently a Frontier Researcher with the Brain Information Processing Group in the Frontier Research Program at the Institute of Physical and Chemical Research (RIKEN) in Japan. His research interests include the theoretical aspects of neural networks, focusing on the dynamics and statistical properties of learning.

Klaus-Robert Müller received the Diplom degree in mathematical physics 1989 and the Ph.D. degree in theoretical computer science in 1992, both from University of Karlsruhe, Germany. From 1992 to 1994 he was a Postdoctoral fellow at the Research Institute for Computer Architecture and Software Technology of the German National Research Center for Information Technology (GMD FIRST) in Berlin.

From 1994 to 1995 he was a European Community STP Research Fellow at University of Tokyo. Since 1995 he has been with the GMD FIRST in Berlin. He is also lecturing at Humboldt University and the Technical University of Berlin. He has worked on statistical physics and statistical learning theory of neural networks and time-series analysis. His present interests are expanded to support vector learning machines and nonstationary blind separation techniques.

Michael Finke received the Diplom degree in computer science in 1993 from the University of Karlsruhe, Germany. Since then he has been working toward the PhD. degree at the Interactive Systems Laboratories in Karlsruhe.

He was a Research Associate at the Heidelberg Scientific Research Center of IBM from 1989 to 1993. Since 1995 he has been a Visiting Researcher at Carnegie Mellon University (CMU), Pittsburgh, PA. His primary research interests are the theory of neural networks, probabilistic models and information geometry of neural networks and graphical statistical models. His research at CMU is also focused on developing a speech recognition engine for large vocabulary conversational speech recognition tasks.

Howard Hua Yang (M'95) received the B.Sc. degree in applied mathematics from Harbin Shipbuilding Engineering Institute, in 1982 and the M.Sc. and Ph.D. degrees in probability and statistics in 1984 and 1989, respectively, from Zhongshan University, P.R. China.

From January 1988 to January 1990, he was a Lecturer in the Department of Mathematics at Zhongshan University, P.R. China. From April 1990 to March 1992, he was a Research Fellow in the Department of Computer Science and Computer Engineering at La Trobe University, Australia, where he did research and teaching in neural networks and computer science. From April 1992 to December 1994, he was a Research Fellow working in communications and digital signal processing group in the Department of Electrical and Electronic Engineering at the University of Melbourne, Australia. Since January 1995, he has been Frontier Researcher in the Laboratory for Information Representation in the Frontier Research Program of The Institute of Physical and Chemical Research (RIKEN) in Japan. His research interests include neural networks, statistical signal processing, blind deconvolution/equalization, and estimation theory.