Asymptotic Statistical Theory of Overtraining and Cross-Validation

S. Amari[*] [†], N. Murata[†], K.-R. Müller[† §], M. Finke[‡], H. Yang[*]

METR 95-06 August 1995

[†] Department of Mathematical Engineering and Infomation. Physics, University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan.

[§] GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany. Klaus@first.gmd.de.

[‡] Institut für Logik, University of Karlsruhe, 76128 Karlsruhe, Germany.

[*] Lab. for Inf. Representation, RIKEN, Wakoshi, Saitama, 351-01, Japan.

Abstract

A statistical theory for overtraining is proposed. The analysis treats general realizable stochastic neural networks, trained with Kullback-Leibler loss 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 testing sets in order to obtain the optimum performance. However, cross-validated early stopping is useless in the asymptotic region, while it decreases the generalization error only in the non-asymptotic region. Our large scale simulations done on a CM5 are in nice agreement with our analytical findings.

1 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 stochastic gradient descent method minimizing a target function (Amari [1967], Rumelhart et al. [1986], and many others). In the case of on-line learning, there remain small fluctuations around it. The dynamical behavior of on-line learning was analyzed by Amari [1967] and Heskes and Kappen [1991] and recently deeper theories were presented by Barkai, Seung and Sompolinsky [1994] and Saad and Solla [1995].

When the number of examples is infinitely large and they are unbiased, the network parameters converge to one of the local minima of the risk function (expected loss) to be minimized. When the number of training examples is finite, the true risk function is different from the empirical risk function to be minimized. Thus, since the training examples are biased the network parameters converge to a biased solution. This is known as overfitting or overtraining, 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, regularization terms can be introduced (Poggio and Girosi[1990] and others), model selection methods (e.g., Murata, Yoshizawa and Amari [1994], Moody [1992], Rissanen [1986] and many others), and the usage of noise disturbance or finite temperature learning (Bös [1995]) can be applied.

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 (Hecht-Nielsen [1989], Hassoun [1995], Wang, Venkatesh and

Judd [1994], and others). To avoid overtraining, the following 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 testing such that the behavior of the trained network is evaluated by using the test examples and training is stopped at the point that minimizes the testing error. Wang, Venkatesh and Judd [1995] analyzed the average optimal stopping time without cross-validation in the case of linear Φ-machines. Barber, Saad and Sollich [1995a,b] considered the evaluation of the generalization error by cross-validation for linear perceptrons. It is widely believed that the cross-validation stopping is a simple and effective method. However, there are no general theories concerning this problem. For example the stopping time is generally unknown and often determined by trial-and-error.

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 Müller et al. [1995]). When t is smaller or nearly equal to m, the learning problem is ill posed because we do not have sufficient examples to determine the parameters. Overtraining is remarkable in this phase, in particular around $t \approx m$ and the generalization error even increases with t when $t < m$ (Bös et al. [1993], Bös [1995]). However, the splitting of an already small set of examples has serious disadvantages since it also decreases the scarce and valuable information in the small data set. In order to avoid overtraining in this case, we need to use other methods (regularization, model selection, finite-temperature learning, etc.) rather than cross-validation stopping.

When t is larger than m, simulations show that cross-validation stopping is effective in general. However, it is difficult to construct a general theory in this intermediate phase. 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 in the asymptotic phase of large t. We analyze the relation between the training error and testing error, and also the trajectory of learning. The effect of early stopping is studied on this basis. We show that the gain of the generalization error is small even if we could find the optimal stopping time, which is in general not possible. We then answer the question: In what ratio the examples should be divided into training and testing sets in order to obtain the optimum performance. 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 testing 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. However, we prove that the generalization error becomes worse in the asymptotic case of large t by the loss not using test examples for training but using them for obtaining the stopping time.

Our analysis treats 1) a realizable stochastic machine, 2) Kullback-Leibler loss (negative of the log likelihood loss), 3) asymptotic behavior where the number t of examples is sufficiently large (compared with the number m of parameters).

Our results were confirmed by large-scale computer simulations of three-layer feedforward networks where the number m of modifiable parameters are $m = 100$. When $t > 30m$, the theory fits well with simulations, showing cross-validation is not necessary, because the generalization error becomes worse by using test 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.

2 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 is the network specified by$\boldsymbol{w}$ 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

$$p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}) = q(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{w}). \tag{2.1}$$

We assume

A1. There exists a teacher network $N(\boldsymbol{w}_0)$ which generates training examples.

A2. The Fisher information matrix $G = (G_{ij})$ defined by

$$G_{ij}(\boldsymbol{w}) = E\left[\frac{\partial}{\partial w_i}\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})\frac{\partial}{\partial w_j}\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w})\right] \tag{2.2}$$

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

A3. The training set

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

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

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

$$l(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}) = -\log p(\boldsymbol{x},\boldsymbol{y};\boldsymbol{w}). \tag{2.3}$$

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

$$R(\boldsymbol{w}) = -E_0[\log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})], \tag{2.4}$$

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

$$R(\boldsymbol{w}) = H_0 + D(\boldsymbol{w}_0 \| \boldsymbol{w}), \tag{2.5}$$

where

$$H_0 = -E_0[\log p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)] \tag{2.6}$$

is the entropy of the teacher network and

$$D(\boldsymbol{w}_0 \| \boldsymbol{w}) = E_0 \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)}{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})} \right] \tag{2.7}$$

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 $\boldsymbol{w} = \boldsymbol{w}_0$.

We give two typical examples of multilayer stochastic networks.

Example 1. Multilayer analog perceptron.

When input is $\boldsymbol{x}$, an l-layer perceptron calculates

$$z_i^{(k)} = f(\sum w_{ij}^{(k)} z_j^{(k-1)}), \quad k = 1, \cdots, l \tag{2.8}$$

recursively from $k = 1$ to $k = l$, where $z_j^{(0)} = x_j$ and $w_{ij}^{(k)}$ is the connection weights from neuron j of the $(k-1)$th layer to neuron i of the kth layer. As usual, $w_{i0}^{(k)}$ are the thresholds and $z_0^{(k)} = 1$. The output $\boldsymbol{y} = (y_i)$ of the network is a noise contaminated version of $\boldsymbol{z}^{(l)} = (z_i^{(l)})$,

$$y_i = z_i^{(l)} + n_i,$$

where n_i are the noise terms. A typical example is that n_i are independently distributed subject to the normal distribution $N(0, \sigma^2)$ with mean 0 and variance σ^2. The probability $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w})$ is given in this case by

$$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^k \exp\left\{-\frac{1}{2\sigma^2}|\boldsymbol{y} - \boldsymbol{z}^{(l)}|^2\right\}, \tag{2.9}$$

so that the risk is

$$R(\boldsymbol{w}) = k \log\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2\sigma^2} E_0[|\boldsymbol{y} - \boldsymbol{z}^{(l)}|^2],$$

where k is the number of the output units and and $\boldsymbol{z}^{(l)}$ is calculated by the network $N(\boldsymbol{w})$ by (2.8) when $\boldsymbol{x}$ is input. Hence, the loss is the ordinary squared error.

Example 2. Stochastic binary multilayer perceptron.

In this case, outputs y_i take on the binary values 0 and 1. The probability of $y_i = 1$ is given by

$$\text{Prob}\{y_i = 1\} = \varphi(z_i^{(l)}), \tag{2.10}$$

where φ is the sigmoid function and $\boldsymbol{z}^{(l)}$ is given by (2.8).

3 Asymptotic analysis of learning

The training set of examples $D_t = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_t, \boldsymbol{y}_t)\}$ is an identically and independently distributed sample from the probability distribution $p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{w}_0)$. The maximum likelihood estimator (m.l.e.) $\hat{\boldsymbol{w}}$ is the one that maximizes the likelihood of producing D_t, $\prod_{i=1}^{t} p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w})$, or equivalently minimizing the empirical risk function

$$R_{\rm train}(\boldsymbol{w}) = -\frac{1}{t}\sum_{i=1}^{t} \log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w}). \tag{3.1}$$

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 $R_{\rm gen}(\boldsymbol{w})$ 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,

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

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 $R_{\rm train}(\boldsymbol{w})$ in the neighborhood of $\boldsymbol{w}_0$.

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

$$R_{\rm gen}(\boldsymbol{w}) = H_0 + \frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_0)^T G(\boldsymbol{w}_0)(\boldsymbol{w}-\boldsymbol{w}_0) + O\left(\frac{1}{t^2}\right) \tag{3.2}$$

$$\begin{aligned} R_{\rm train}(\boldsymbol{w}) &= H_0 - \frac{1}{2}(\hat{\boldsymbol{w}}-\boldsymbol{w}_0)^T G(\boldsymbol{w}_0)(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) \\ &\quad + \frac{1}{2}(\boldsymbol{w}-\hat{\boldsymbol{w}})^T G(\boldsymbol{w}_0)(\boldsymbol{w}-\hat{\boldsymbol{w}}) + O\left(\frac{1}{t^2}\right), \end{aligned} \tag{3.3}$$

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

The relation (3.2) is the Taylor expansion of (2.4), 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 (3.3) is given in Appendix A.

By putting $\boldsymbol{w} = \hat{\boldsymbol{w}}$ in (3.2) and (3.3), we have the asymptotic evaluations of the generalization and training errors of $N(\hat{\boldsymbol{w}})$. They depend on the examples D_t from which the m.l.e. is calculated. We denote by $\langle \quad \rangle$ the average over the examples D_t which determine $\hat{\boldsymbol{w}}_t$. We then obtain the following universal relation concerning the generalization error and training error. This was first proved by Amari and Murata [1990]. A similar but different universal property is proved by Amari [1990] for deterministic dichotomy machines.

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

$$\langle R_{\rm gen}(\hat{\boldsymbol{w}})\rangle = H_0 + \frac{m}{2t} + O\left(\frac{1}{t^2}\right), \tag{3.4}$$

$$\langle R_{\rm train}(\hat{\boldsymbol{w}})\rangle = H_0 - \frac{m}{2t} + O\left(\frac{1}{t^2}\right), \tag{3.5}$$

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, Yoshizawa and Amari [1993], [1994] obtained more general versions and proposed the NIC (network information criterion) for model selection by generalizing the AIC (Akaike [1974]). Let us consider the gradient descent learning rule, where the parameter $\hat{\boldsymbol{w}}(n)$ at the n-th step is modified by

$$\hat{\boldsymbol{w}}(n+1) = \hat{\boldsymbol{w}}(n) - \varepsilon\frac{\partial R_{\rm train}(\hat{\boldsymbol{w}}(n))}{\partial \boldsymbol{w}}, \tag{3.6}$$

where ε is a small positive constant. This is batch learning where all the training examples are used for each iteration of modifying $\hat{\boldsymbol{w}}(n)$. We can alternatively use on-line learning[1],

$$\hat{\boldsymbol{w}}(n+1) = \hat{\boldsymbol{w}}(n) + \varepsilon \frac{\partial \log p(\boldsymbol{x}_n, \boldsymbol{y}_n; \boldsymbol{w}_n)}{\partial \boldsymbol{w}}, \tag{3.7}$$

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 $\hat{\boldsymbol{w}}$, 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 (3.3) as

$$\frac{\partial R_{\text{train}}(\hat{\boldsymbol{w}}(n))}{\partial \boldsymbol{w}} = G(\boldsymbol{w}_0)\{\hat{\boldsymbol{w}}(n) - \hat{\boldsymbol{w}}\} + O\left(\frac{1}{t^2}\right).$$

Hence, by neglecting the term of order $1/t^2$, (3.6) 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,

$$G(\boldsymbol{w}_0) = I \tag{3.8}$$

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

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

[1] Its dynamical behavior was studied by Amari [1967], Heskes and Kappen [1993], and recently by Barkai, Seung, and Sompolinsky [1995] and Solla and Saard [1995].

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

An interesting question is from which direction $\hat{\boldsymbol{w}}(n)$ approaches $\hat{\boldsymbol{w}}$. Even if the initial $\boldsymbol{w}(0)$ is uniformly distributed, we cannot say that $\hat{\boldsymbol{w}}(n)$ approaches $\hat{\boldsymbol{w}}$ isotropically, since dynamics (3.6) 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. We evaluate the averages $\langle R_{\rm gen}(\hat{\boldsymbol{w}}(n))\rangle$ and $\langle R_{\rm train}(\hat{\boldsymbol{w}}(n))\rangle$, where the ensemble average $\langle\ \ \rangle$ is taken over the training examples D_t. The ensemble average $\langle\ \ \rangle$ over D_t is, in the present evaluation, to take average over the m.l.e. $\hat{\boldsymbol{w}}$, which is isotropically distributed in the $(1/\sqrt{t})$-neighborhood of $\boldsymbol{w}_0$, since its covariance matrix is $G(\boldsymbol{w}_0) = I$. This implies that $\langle R\rangle$ is invariant under the orthogonal transformation of $\boldsymbol{w}$ around $\boldsymbol{w}_0$ or $\hat{\boldsymbol{w}}$. Let us consider a trajectory starting at $\boldsymbol{w}(0)$. For a fixed $\hat{\boldsymbol{w}}$, the trajectory approaches the $\hat{\boldsymbol{w}}$ in a definite direction. For estimation obtained from D_t', $\hat{\boldsymbol{w}}'$, other than $\hat{\boldsymbol{w}}$, we can transform $\hat{\boldsymbol{w}}'$ into $\hat{\boldsymbol{w}}$ by an orthogonal transformation. This transformation changes the trajectory. The learning trajectory approaches $\hat{\boldsymbol{w}}$ after the transformation from a different direction. Since $\hat{\boldsymbol{w}}'$ are isotropically distributed around $\boldsymbol{w}_0$, the transformed trajectories are isotropically distributed around the $\hat{\boldsymbol{w}}$. This true when the initial value $\boldsymbol{w}(0)$ is located far from $\boldsymbol{w}_0$ and $\hat{\boldsymbol{w}}$, since the isotropy is around $\boldsymbol{w}_0$ but we assume the isotropy around $\hat{\boldsymbol{w}}$. Thus, we have an important lemma.

Lemma 2 *When we calculate the ensemble averages $\langle R_{\rm gen}(\hat{\boldsymbol{w}}(n))\rangle$ and $\langle R_{\rm train}(\hat{\boldsymbol{w}}(n))\rangle$ for large n, $\hat{\boldsymbol{w}}(n)$ is supposed to approach $\hat{\boldsymbol{w}}$ isotropically.*

We call the trajectory $\hat{\boldsymbol{w}}(n)$ a ray linearly approaching $\hat{\boldsymbol{w}}$ when n is large. The rays are isotropically distributed around $\hat{\boldsymbol{w}}$.

4 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

$$R(n) = R_{\rm gen}(\hat{\boldsymbol{w}}(n)), \quad n = 1, 2, \cdots. \tag{4.1}$$

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 $n_{\rm opt}$ is a random variable depending on $\hat{\boldsymbol{w}}$ and the initial $\boldsymbol{w}(0)$. We evaluate the ensemble average of $\langle R(n_{\rm opt})\rangle$.

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

$$d^2 = |\hat{\boldsymbol{w}} - \boldsymbol{w}_0|^2 \tag{4.2}$$

and

$$E_0[d^2] = E_0[(\hat{\boldsymbol{w}} - \boldsymbol{w}_0)^T G^{-1} (\hat{\boldsymbol{w}} - \boldsymbol{w}_0)] = \frac{1}{t}{\rm tr}(G^{-1}G) = \frac{m}{t}. \tag{4.3}$$

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

$$R(n) = H_0 + \frac{1}{2}|\hat{\boldsymbol{w}}(n) - \boldsymbol{w}_0|^2 \tag{4.4}$$

is given by the following lemma.

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

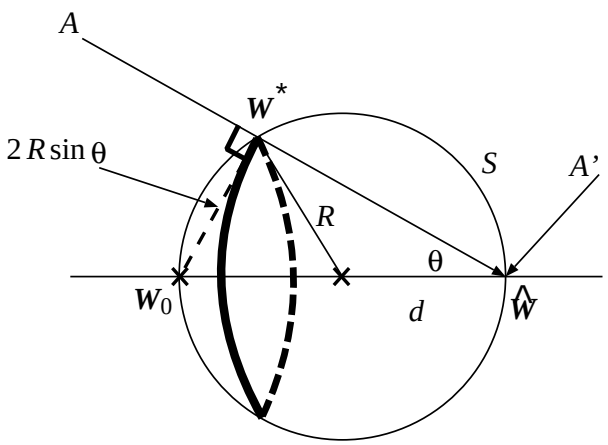


Figure 1: Geometrical picture to determine the optimal stopping point $\boldsymbol{w}^*$.

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.

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

$$r(\theta) = \frac{1}{I_{m-2}} \sin^{m-2} \theta, \tag{4.5}$$

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 R approaches $\hat{\boldsymbol{w}}$ isotropically (Lemma 2), its intersection to S' is uniformly distributed on S'. 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

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

∎

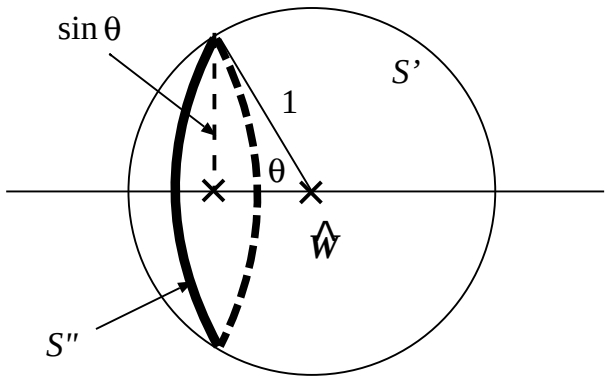


Figure 2: Distribution of the angle θ.

Now we have the following theorem.

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

$$\langle R(n_{\rm opt})\rangle = H_0 + \frac{1}{2t}(m - \frac{1}{2}). \tag{4.6}$$

Proof When ray A is at angle θ ($0 \leq \theta < \pi/2$), 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

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

Since we have,

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

When θ is $\pi/2 \leq \theta \leq \pi$, 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 (4.4), 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 $1/2t$, 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, Venkatesh and Judd [1995] calculated $\langle R(n)\rangle$ in the case of linear Φ-machines and defined the optimal average stopping time $\bar{n}_{\rm opt}$ that minimizes $\langle R(n)\rangle$. This is different from the present

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

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

We can prove

$$\langle R(\bar{n}_{\rm opt})\rangle = \langle R(\hat{\boldsymbol{w}})\rangle - O\left(\frac{1}{t^2}\right)$$

in agreement with Wang, Venkatesh and Judd [1995]. This shows that the gain becomes much smaller by using the average stopping time $\bar{n}_{\rm opt}$. 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.

5 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 testing set for evaluating the generalization error. The training error monotonically decreases with iterations, but according to the folklore the generalization error evaluated by the testing 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 testing 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 testing set, where

$$r + r' = 1. \tag{5.1}$$

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 $r't$ testing examples, that is $\hat{\boldsymbol{w}}$ and $\tilde{\boldsymbol{w}}$ minimize the training error function

$$R_{\text{train}}(\boldsymbol{w}) = \frac{1}{rt}\sum_i \{-\log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w})\} \tag{5.2}$$

and testing error function

$$R_{\text{test}}(\boldsymbol{w}) = \frac{1}{r't}\sum_i \{-\log p(\boldsymbol{x}_i, \boldsymbol{y}_i; \boldsymbol{w})\} \tag{5.3}$$

respectively, where summations are taken over rt trainig examples and $r't$ testing examples. Since the training examples and testing examples are independent, both $\hat{\boldsymbol{w}}$ and $\tilde{\boldsymbol{w}}$ are 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 testing error is the point on A that is closest to $\tilde{\boldsymbol{w}}$, since the testing error defined by (5.3) can be expanded from (3.3) as

$$R_{\text{test}}(\boldsymbol{w}) = H_0 - \frac{1}{2}|\tilde{\boldsymbol{w}} - \boldsymbol{w}_0|^2 + \frac{1}{2}|\boldsymbol{w} - \tilde{\boldsymbol{w}}|^2. \tag{5.4}$$

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

$$d = |\hat{\boldsymbol{w}} - \tilde{\boldsymbol{w}}|. \tag{5.5}$$

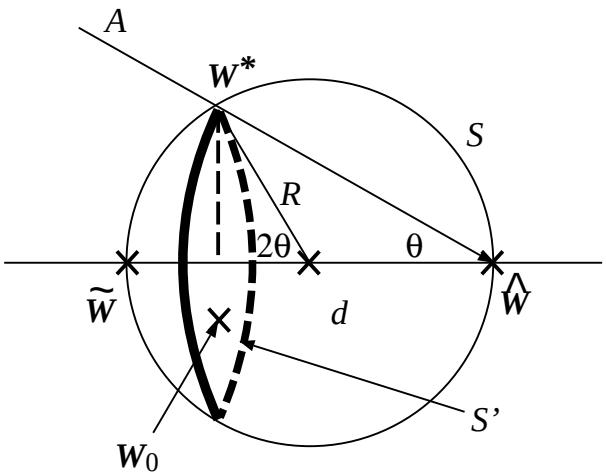


Figure 3: Optimal stopping point $\boldsymbol{w}^*$ by cross-validation.

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 (3.2), 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*

$$\langle R(\boldsymbol{w}^*, r)\rangle = H_0 + \frac{2m-1}{4rt} + \frac{1}{4r't}. \tag{5.6}$$

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

Theorem 2 *The average generalization error is minimized at*

$$r_{\rm opt} = 1 - \frac{\sqrt{2m-1}-1}{2(m-1)}. \tag{5.7}$$

The theorem shows the optimal division of examples into training and testing sets. When m is large,

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

showing that only $(1/\sqrt{2m}) \times 100\%$ of examples are to be used for 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 testing. From this we also have

Theorem 3 *The optimal generalization error is*

$$\langle R(\boldsymbol{w}^*, r_{\rm opt})\rangle = H_0 + \frac{1}{4t}(\sqrt{2m-1}+1)^2. \tag{5.8}$$

When m is large, we have

$$\langle R(\boldsymbol{w}^*, r_{\rm opt})\rangle = H_0 + \frac{m}{2t}\left(1+\sqrt{\frac{2}{m}}\right). \tag{5.9}$$

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,

$$\langle D(\boldsymbol{w}_0\|\boldsymbol{w}^*)\rangle > \langle D(\boldsymbol{w}_0\|\hat{\boldsymbol{w}})\rangle, \tag{5.10}$$

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

6 Simulations

We use standard feed-forward classifier networks with N inputs, H sigmoid hidden units and M softmax outputs (classes). The output activity O_l of the lth output unit is calculated via the softmax squashing function

$$\begin{aligned} p(\boldsymbol{y} = C_l|\boldsymbol{x};\boldsymbol{w}) = O_l &= \frac{\exp(h_l)}{1+\sum_k \exp(h_k)}, \quad l = 1,\cdots,M, \\ O_0 &= \frac{1}{1+\sum_k \exp(h_k)}, \end{aligned}$$

where $h_l^O = \sum_j w_{lj}^O s_j - \vartheta_l^O$ is the local field potential. Each output O_l codes the a-posteriori probability of being in class C_l, O_0 denotes a zero class for normalization purposes. The m network parameters consist of biases ϑ and weights $\boldsymbol{w}$. When $\boldsymbol{x}$ is input, the activity of the j-th hidden unit is

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

The input layer is connected to the hidden layer via $\boldsymbol{w}^H$, the hidden layer is connected to the output layer via $\boldsymbol{w}^O$, but no short-cut connections are present. Although the network is completely deterministic, it is constructed to approximate class conditional probabilities (Finke and Müller [1994]).

The examples $\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_t,\boldsymbol{y}_t)\}$ are produced randomly, by drawing $\boldsymbol{x}_i$, $i = 1,\cdots,t$, from a uniform distribution independently and producing the labels $\boldsymbol{y}_i$ stochastically from the teacher classifier. Conjugate gradient learning with line-search on the training error function (5.2) is applied, starting from some random initial vector, and the testing error (5.3) is measured on the cross-validation set to stop learning. The generalization ability (2.4)is approximately estimated by (3.1) on a large test set (50000 patterns). We compare the generalization error for the settings: exhaustive training (no stopping), early stopping (controlled by the cross-validation

set) 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 8h 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. Amari et al. [1995]).

Fig. 5 shows the results of simulations, where $N = 8$, $H = 8$, $M = 4$, so that the number m of modifiable parameters is $m = (N + 1)H + (H + 1)M = 108$. In Fig. 5a we see the intermediate range of patterns $t < 30m$ (see Müller et al. [1995]), where early stopping improves the generalization ability to a large extent. From Fig. 5b 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. In the asymptotic area (Fig. 5b) we observe that the smaller the percentage of the training set, which is used to determine the point of early stopping, the better the performance of the generalization ability. When we use cross-validation, the optimal size of the test set is about 7% of all the examples, as the theory predicts.

In Fig. 6a 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. 6b 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 (3.1) on the training patterns. Therefore the gradient terms in the likelihood expansion 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.

7 Concluding Remarks

We proposed an asymptotic theory for overtraining. The analysis treats realizable stochastic neural networks, trained with Kullback-Leibler loss.

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 showed for large m that optimally only $r'_{opt} = 1/\sqrt{2m}$ 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 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-validated stopping the generalization error is always increased comparing to exhaustive training. Nevertheless note, that this range is due to the limited size of the data sets often unaccessible in practical applications.

In the non-asymptotic region simulations show that cross-validated early stopping always helps to enhance the performance since it decreases the generalization error. In this intermediate range our theory also gives a good estimate of the optimal size of the early stopping set.

It is interesting to see why cross-validation early stopping is effective in the intermediate range. Let D_0 be the distance between the initial $\boldsymbol{w}(0)$ and the true $\boldsymbol{w}_0$. Our analysis is based on the assumption that D_0 is sufficiently large so that the isotropy lemma (Lemma 2) holds. However, when D_0 is small compared to the distance of $\boldsymbol{w}_0$ and $\hat{\boldsymbol{w}}$, it does not hold. When both $\boldsymbol{w}(0)$ and $\boldsymbol{w}_0$ are randomly chosen, it is natural to assume that $\boldsymbol{w}(0)$ is isotropically distributed around $\boldsymbol{w}_0$. Assume that D_0 is small and is of the order of $1/\sqrt{t}$. Then, the ray A starting at the

distance D_0 from $\boldsymbol{w}_0$ intersects the sphere S with high probability and it is seldom that ray A approaches $\hat{\boldsymbol{w}}$ from the opposite side (right side in Fig.1). In the extreme case where D_0 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_0 is very large. However, a natural prior distribution is the Jeffrey non-informative distribution[2] 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_0 into account. However, the theory is valid where t is not asymptotically large so that the non-linear effect of learning trajectories cannot be neglected and we need higher-order corrections to the asymptotic properties of the estimator $\hat{\boldsymbol{w}}$. This will be in the focus of our future studies.

Acknowledgements

We 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, Illinois. We acknowledge computing time on the CM5 in Urbana (NCSA) and in Bonn, supported by the National Institutes of Health (P41RRO 5969). K. -R. M. was partly supported by the EC S & T fellowship and thanks for warm hospitality at RIKEN during the completion of this work.

[2]Note also that the Bayes estimator $\hat{\boldsymbol{w}}_{\mathrm{Bayes}}$ 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.

References

Amari, S. [1967], Theory of adaptive pattern classifiers, *IEEE Trans.*, **EC-16**, 299–307.

Amari, S., Murata, N. [1993], Statistical Theory of Learning Curves under Entropic Loss Criterion, *Neural Computation*, **5**, 140–153.

Amari, S., Murata, N., Müller, K.-R., Finke, M., Yang, H. [1995], Statistical Theory of Overtraining - Is Cross-Validation Effective?, submitted to NIPS'95.

Akaike [1974], A new look at statistical model identification, *IEEE Trans.*, **AC-19**, 716–723.

Barber, D., Saad, D., Sollich, P. [1995a], Test Error fluctuations in finite linear perceptrons, *Neural Computation* **7**, 809–821.

Barber, D., Saad, D., Sollich, P. [1995b], Finite-size effects and optimal test size in linear perceptrons, *Journal of Physics A*, **28**, 1325–1334.

Barkai, N. and Seung, H. S. and Sompolinsky, H. [1995], On-line learning of dichotomies, in *Advances in Neural Information Processing Systems* NIPS 7, (eds. Tesauro G., Touretzky, D.S. and Leen T.K.), MIT Press.

Bös, S., Kinzel, W. and Opper, M. [1993], Generalization ability of perceptrons with continuous outputs, *Physical Review*, **E47**, 1384–1391.

Bös, S. [1995], Avoiding overfitting by finite temperature learning and cross-validation, in ICANN'95: Proceedings of the international conference on artificial neural networks, Paris.

Finke, M. and Müller, K.-R. [1994], in proc. of the 1993 Connectionist Models summer school, Mozer, M., Smolensky, P., Touretzky, D.S., Elman, J.L. and Weigend, A.S. (Eds.), Hillsdale, NJ: Erlenbaum Associates, 324.

Hassoun, M. H. [1995], Fundamentals of Artificial Neural Networks, MIT Press.

Hecht-Nielsen, R. [1989], Neurocomputing, Addison-Wesley.

Heskes, T. and Kappen, B. [1991], Learning process in neural networks, *Physical Review*, **A44**, 2718–2762.

Moody, J. E. [1992], The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems, in *Advances in Neural Information Processing Systems* NIPS 4, J. E. Moody, S. J. Hanson, and R. P. Lippmann, eds., Morgan Kaufmann, Sanmateo, CA.

Murata, N., Yoshizawa, S., Amari, S. [1993], Learning curves, model selection and complexity of neural networks, in *Advances in Neural Information Processing Systems* NIPS 5, S.J.Hanson et al. eds., Morgan Kaufmann, Sanmateo, CA.

Murata, N., Yoshizawa, S., Amari, S. [1994], Network information criterion — Determining the number of hidden units for an artificial neural network model, *IEEE Trans.*, **NN5**, 865–872.

Müller, K.-R., Finke, M., Murata, N., Schulten, K. and Amari, S. [1995], A numerical study on learning curves in stochastic multilayer feed-forward networks, Univ. of Tokyo Tech. Report METR 03-95 and submitted to *Neural computation.*

Poggio, T. and Girosi, F. [1990], Regularization algorithms for learning that are equivalent to multilayer networks, *Science*, **247**, 978–982.

Rissanen, J. [1986], Stochastic complexity and modeling, *Ann, Statist.*, **14**, 1080–1100.

Rumelhart, D., Hinton, G. E., Williams, R. J. [1986], in PDP, Vol.**1**, MIT Press, Cambridge, MA. Learning internal representations by error propagation, in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol.**1**, *Foundations*, MIT Press, Cambridge, MA.

Saad, D., Solla, S. A. [1995], On-line learning in soft committee machines, preprint.

Wang, Ch., Venkatesh, S. S., Judd, J. S. [1994], Optimal stopping and effective machine complexity in learning, to appear, (revised and extended version of NIPS'93).

A Proof of (3.3)

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

$$R_{\rm train}(\boldsymbol{w}) = R_{\rm train}(\hat{\boldsymbol{w}}) + \frac{1}{2}(\boldsymbol{w}-\hat{\boldsymbol{w}})^T G(\boldsymbol{w}-\hat{\boldsymbol{w}}) + O\left(\frac{1}{t^2}\right). \tag{A.2}$$

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

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

Hence,

$$\frac{\partial R_{\rm train}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}}(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) = -(\hat{\boldsymbol{w}}-\boldsymbol{w}_0)^T G(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) + O\left(\frac{1}{t^2}\right). \tag{A.4}$$

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

$$R_{\rm train}(\hat{\boldsymbol{w}}) = H_0 - \frac{1}{2}(\hat{\boldsymbol{w}}-\boldsymbol{w}_0)^T G(\hat{\boldsymbol{w}}-\boldsymbol{w}_0) + O\left(\frac{1}{t^2}\right).$$

Again by substituting this in (A.2), we have (3.3).

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 (4.5). 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'.

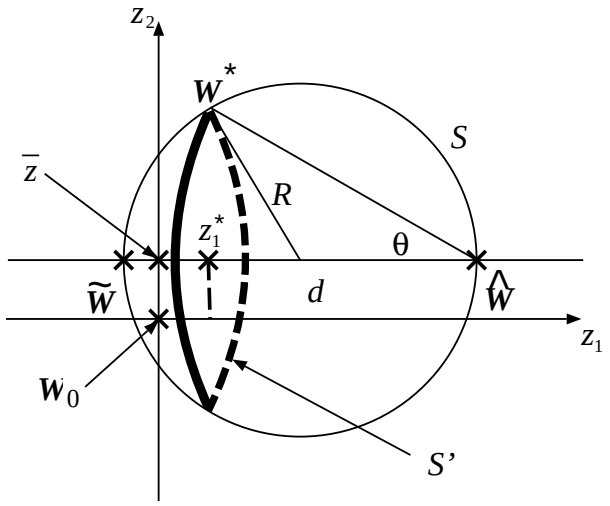


Figure 4: New coordinate system z.

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.4) 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

$$\hat{z}_2 = \tilde{z}_2 = \bar{z},$$

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 because of the symmetry.

The sphere S' is written in this coordinate system as

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

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,

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

Here, we used the following properties;

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

From

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

we have

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

Therefore, from

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

when ray A arrives from the left side,

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

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

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

In order to obtain r_{opt}, we evaluate

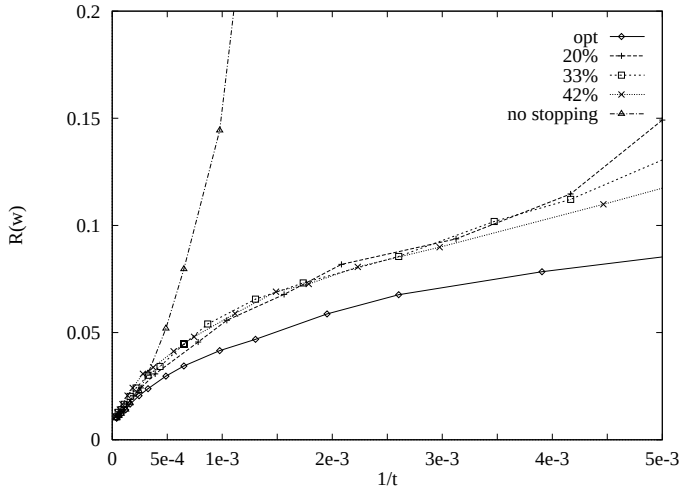
$$\frac{d}{dr}\langle R(\boldsymbol{w}^*, r)\rangle = -\frac{m}{tr^2} + \frac{1}{2tr^2} + \frac{1}{2t}\frac{1}{(1-r)^2} = 0,$$

giving

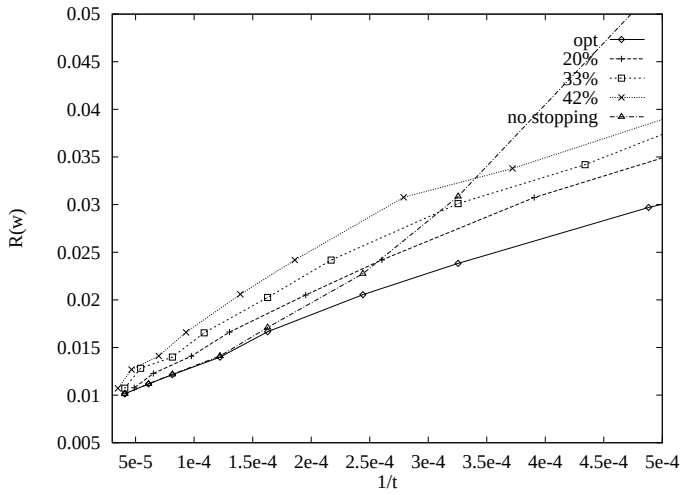
$$r_{\mathrm{opt}} = 1 - \frac{\sqrt{2m-1} - 1}{2(m-1)}.$$

and for large m we get

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

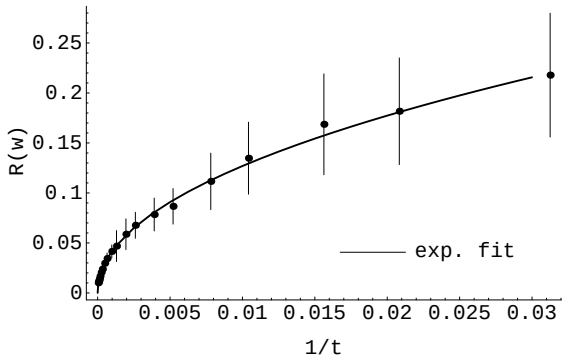


(a)

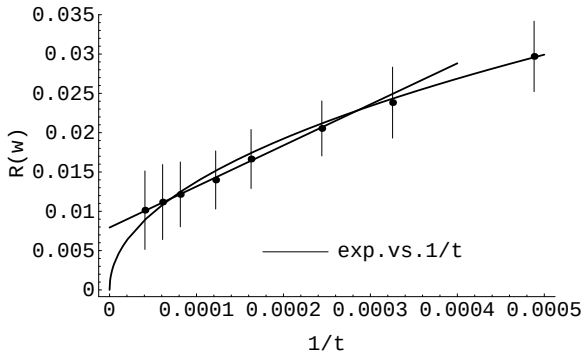


(b)

Figure 5: Shown is $R(\boldsymbol{w})$ plotted for different sizes r' of the early stopping set for a 8-8-4 classifier network (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 cross-validation set (50000) and no stopping adresses the case where 100% of the training set is used for exhaustive learning.



(a)



(b)

Figure 6: $R(\boldsymbol{w})$ 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.