TECHNICAL REPORTS

A Numerical Study on Learning Curves in Stochastic Multi-Layer Feed-Forward Networks

K.-R. Müller[†#+], M. Finke[‡], N. Murata[†],

K. Schulten[+], S. Amari[†*]

METR 95-03 May 1995

MATHEMATICAL ENGINEERING SECTION

DEPARTMENT OF MATHEMATICAL ENGINEERING AND INFORMATION PHYSICS

FACULTY OF ENGINEERING, THE UNIVERSITY OF TOKYO

BUNKYO-KU, TOKYO, JAPAN

Abstract

The universal asymptotic scaling laws proposed by Amari et al. are studied in large scale simulations using a CM5. Small stochastic multi-layer feed-forward networks trained with back-propagation are investigated. In the range of a large number of training patterns t, the asymptotic generalization error scales as $1/t$ as predicted. For a medium range t a faster $1/t^2$ scaling is observed. This effect is explained by using higher order corrections of the likelihood expansion. It is shown for small t that the scaling law changes drastically, when the network undergoes a transition from ineffective to effective learning.

1 Introduction

Recently a growing interest in learning curves, i.e. scaling laws for the asymptotic behavior of the learning and generalization ability of neural networks has emerged (Amari et al. 1993, Barkai et al. 1992, Baum et al. 1989, Haussler et al. 1994, Murata et al. 1993, Opper et al. 1990 and 1995, Saad et al. 1995, Seung et al. 1992, Sompolinski et al. 1990). Clearly, as soon as learning is applied, we observe the characteristics and the performance of the learning algorithms in terms of generalization and training error. Therefore, it is important to study the bounds on how fast we can learn as a function of the number of parameters in general. The large-scale simulations presented in this paper are addressing the question of scaling laws for training and generalization errors in small multilayer feed-forward networks with so far up to 256 parameters, trained on a finite number of training samples of up to 32768 patterns.

We address the teacher-student situation, i.e. given a teacher network, a student network of the same architecture learns from the examples generated by the teacher.

So far a number of groups have used statistical mechanics and the replica trick in order to find the scaling properties of the generalization ability, first for simple perceptron systems, and recently for tree-like networks with hidden units (for reviews see Heskes et al. 1991, Opper et al. 1995, Saad et al. 1995, Seung et al. 1992 and Watkin et al. 1993).

A further approach for estimating asymptotic learning curves is the computational one, where the VC dimension is used to measure the complexity of a given problem (Baum et al. 1989, Haussler et al. 1994, Opper et al. 1991).

We would like to adopt the viewpoint of information geometry which provides an alternative method for estimating the asymptotic behavior of learning based on an asymptotic expansion of the likelihood of the estimating machines, always assuming a maximum likelihood estimator (Amari et al. 1993, Murata et al. 1993).

In this paper we studied, whether the well-known universal asymptotic scaling laws found by Amari et al. can be observed in a simulation of a finite continuous network and a finite number of continuous training patterns. According to this theory the scaling law

$$\epsilon_g = H_0 + \frac{m}{2t}, \tag{1}$$

holds for general stochastic machines (Amari et al. 1993, Murata et al. 1993).

The quantity ϵ_g denotes the averaged likelihood (generalization ability), m is the number of parameters of the model (bias + weights) and t is the number of training examples presented to the network. Emphasis is set to the issue of evaluating, whether these asymptotic results have an impact on the practical user of neural networks. Also the question, where asymptotics starts, is addressed.
A further point of interest is to get insights about the dynamics of the hidden units during the learning process.
In our simulations we are using standard multi-layer continuous feed-forward networks, trained with backpropagation and a conjugate gradient descent in the Kullback-Leibler divergence.
The next section describes the model investigated. The technical details of our simulations are given in section 3 and higher order corrections to Eq.(1) are presented in section 4. Section 5 discusses the numerical scaling results and finally a conclusion is given.

2 The Model

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

$$p(\vec{y} = C_l|\vec{x};\vec{w}) = O_l = \frac{\exp(h_l^O)}{1 + \sum_k \exp(h_k^O)}, \quad l = 1,\ldots,M, \tag{2}$$

where

$$O_0 = \frac{1}{1 + \sum_k \exp(h_k^O)},$$

and 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 an input pattern being in class C_l, O_0 denotes a zero class for normalization purposes. The m network parameters consist of biases $\vartheta = (\vartheta^H, \vartheta^O)$ and weights $\vec{w} = (\vec{w}^H, \vec{w}^O)$. When $\vec{x} = (x_1,\ldots,x_N)$ is input, the activity $\vec{s} = (s_1,\ldots,s_H)$ is computed as

$$s_j = [1 + \exp(-\sum_{k=1}^{N} w_{jk}^H x_k - \vartheta_j^H)]^{-1}, \quad j = 1,\ldots,H. \tag{3}$$

The input layer is connected to the hidden layer via $\vec{w}^H$, the hidden layer is connected to the output layer via $\vec{w}^O$, but no short-cut connections are present.

Although the network is completely deterministic, it is constructed to approximate class conditional probabilities (Finke et al. 1993). In this sense it is considered a stochastic machine randomly generating class labels for M different classes given the input. Therefore, each randomly generated teacher $\vec{w}_T$ represents by construction a multinomial probability distribution $q(C_l|\vec{x},\vec{w}_T) = \text{Prob}\{\vec{x} \in C_l\}$ over the classes C_l $(l = 1 \ldots M)$ given a random input $\vec{x}$. 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 $q(C_l|\vec{x}) = p(C_l|\vec{x},\vec{w}_T)$.
A training and test set of the form $\mathcal{S} = \{(\vec{x}^p, c^p)|p = 1 \ldots t\}$ is generated randomly, by drawing samples of $\vec{x}$ from a uniform distribution and forward propagating $\vec{x}^p$ through the teacher network. Then, according to the teachers' outputs $q(C_l^p|\vec{x}^p)$ one output unit is set to one stochastically and all others are set to zero leading to the target vector $\vec{y}^p = (0,\ldots,1,\ldots,0)$. A student network $\vec{w}$ is then trying to approximate the teacher given the example set $\mathcal{S}$. For training the student network $\vec{w}$ we use a backpropagation algorithm with conjugate gradient descent to minimize our objective function: the Kullback-Leibler divergence

$$D(q,p(\vec{w})) = \int d\vec{x} \sum_{l=0}^{M} q(\vec{x}) q(C_l|\vec{x}) \ln \frac{q(C_l|\vec{x})}{p(C_l|\vec{x},\vec{w})}. \tag{4}$$

Here $q(C_l|\vec{x})$ denotes the class conditionals, respectively outputs of the teacher and $p(C_l|\vec{x},\vec{w})$ are the class posteriors as approximated by the student network. The Kullback-Leibler divergence is the natural objective function to measure the degree of coincidence of the teacher and student distributions q and p. To measure the Kullback-Leibler divergence one has to know the stochastic source underlying the data-set which can be decomposed into the input generating part $q(\vec{x})$ and the output probability distribution $q(C_l|\vec{x})$. In practical applications there is typically no such knowledge. So in our training procedure only the log likelihood

$$\epsilon_T = -\frac{1}{t} \sum_{p} \ln p(c^p|\vec{x}^p,\vec{w}) \tag{5}$$

will be available, using the empirical joint distribution

$$q^*(\vec{x},C_m) = \frac{1}{t} \sum_{p=1}^{t} \begin{cases} 1 & : \quad \vec{x} = \vec{x}^p \text{ and } C_m = c^p \\ 0 & : \quad \text{otherwise} \end{cases}$$

to evaluate (4); c^p refers to the correct class label associated to $\vec{x}^p$.[1]
Our results based on training with (5) have practical importance, since as mentioned above, in general practical problems only the empirical distribution is known. On the test set we use a better approximation to the KL divergence by sampling (4) for which all necessary ingredients are known

$$\epsilon_g = -\frac{1}{\#\text{test set}} \sum_p \sum_{l=0}^{M} \left[q(C_l|\vec{x}^p) \ln p(C_l|\vec{x}^p, \vec{w}) - q(C_l|\vec{x}^p) \ln q(C_l|\vec{x}^p, \vec{w}) \right]. \quad (6)$$

So given a random uniformly distributed input, we can use the a-posteriori probabilities $q(C_l|\vec{x}^p)$, which are exactly the output values given by the teacher networks on the presentation of an input vector $\vec{x}^p$.

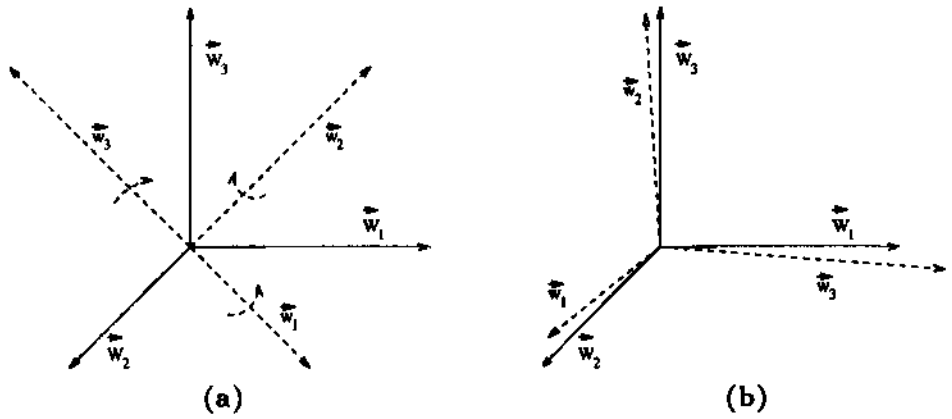


Figure 1: *Schematic picture of the weight vectors of the student and the teacher (a) before and (b) after the transition from uncorrelated to correlated learning.*

3 Order Parameters

For the committee machine several authors have observed a phase transition, where the generalization error first scales as N/t in a so-called symmetric phase whereas for more patterns a transition takes place and the system scales as NH/t in the symmetry broken phase (Barkai et al. 1992, Schwarze et al. 1993, Seung

[1] We use Eq.(5) instead of Eq.(4) because minimizing the KL divergence and minimizing $-\int d\vec{x} \sum_{l=0}^{M} q(\vec{x}) q(C_l|\vec{x}) \ln p(C_l|\vec{x}, \vec{w})$ differs only by a constant and is therefore equivalent. In the learning situation, only the set of training examples is available, so we have to use Eq.(5).

et al. 1992, Kang et al. 1993, Saad et al. 1995). Below the transition all hidden units learn uncorrelated to each other and to all the teacher hidden units (see fig.1a). Above the transition every student hidden unit decides for one teacher hidden unit and is maximally uncorrelated with the other teacher hidden units (see fig.1b).
We would like to study whether this transition also occurs in continuous multi-layer feed-forward networks being trained with continuous patterns. We therefore define a set of order parameters which allow a more careful inspection of the correlations between student and teacher than the Kullback-Leibler divergence.

3.1 Angle Based Order Parameters

In the committee machine the overlap

$$R_{ik} = \frac{1}{N}\sum_{k=1}^{N} w_{Tik}^{H} w_{jk}^{H} = \frac{1}{N} w_{Ti\bullet}^{H} \cdot w_{j\bullet}^{H}$$

and the self-overlap describe the dynamics of the hidden units during learning, where we used the abreviation $w_{Ti\bullet}^{H} = (w_{Ti1}^{H}, \ldots, w_{TiN}^{H})$. To have only one parameter we consider all permutations σ of the hidden units in the multi-layer perceptron case to make the overlap independent of the actual permutation. In our case the weights have to be normalized, since our system is not binary. Let $w_{Ti\bullet}^{H}$ and $w_{\sigma(i)\bullet}^{H}$ be the vectors of all weights from the input layer into hidden unit i for teacher and student respectively, and let $w_{T\bullet i}^{O}$ and $w_{\bullet\sigma(i)}^{O}$ denote the weight vectors from hidden unit i to all output units. Based on this notation we define two measures for the correlation of the weight vectors

$$\mathrm{rH} = \max_{\sigma} \frac{1}{H}\sum_{j=1}^{H} \frac{w_{Tj\bullet}^{H} \cdot w_{\sigma(j)\bullet}^{H}}{||w_{Tj\bullet}^{H}||||w_{\sigma(j)\bullet}^{H}||} \quad \text{and} \quad \mathrm{rO} = \max_{\sigma} \frac{1}{H}\sum_{j=1}^{H} \frac{w_{T\bullet j}^{O} \cdot w_{\bullet\sigma(j)}^{O}}{||w_{T\bullet j}^{O}||||w_{\bullet\sigma(j)}^{O}||}, \quad (7)$$

where $\max_{\sigma}$ is the maximum over all possible permutations σ of the hidden units. In other words, we consider the overlap of the hidden units given a permutation such that the weights of the hidden units of the teacher and the student are maximally correlated. A transition from uncorrelated to correlated learning, as mentioned above, would be detected as a change of the angles between teacher and student vectors.

3.2 Length Based Order Parameters

The order parameters introduced in the last section essentially measure the angle between teacher and student machine. Now we have to take into account that we do not deal with binary weights, which are nicely normalized, but with students who can change the lengths of their parameter vectors quite drastically in the dynamics of the learning process. We therefore introduce a new set of order parameters based on the ratio between the teacher and student weights

$$\text{ratio H} = \max_{\sigma} \frac{1}{H} \sum_{j=1}^{H} \frac{||w^H_{\sigma(j)\bullet}||}{||w^H_{Tj\bullet}||} \quad \text{and} \quad \text{ratio O} = \max_{\sigma} \frac{1}{H} \sum_{j=1}^{H} \frac{||w^O_{\bullet\sigma(j)}||}{||w^O_{T\bullet j}||}. \tag{8}$$

3.3 Correlation Based Order Parameters

Since the hidden units implement functions, we measure additionally the functional $\mathcal{L}^2$ norm, which corresponds to the correlations between the hidden units activities

$$\text{Act H}_{ij} = \frac{1}{\#\text{test set}} \sum_{p} (s_{Ti}(\vec{x}^p) - s_j(\vec{x}^p))^2 . \tag{9}$$

The sum is taken over the test set and s_{Ti} denotes the activity of the ith hidden unit of the teacher while s_j is the student's activity at hidden unit j (cf. Eq. 3). A value of Act $\text{H}_{ij} \sim 0$ corresponds to a maximal correlation between student and teacher. This parameter gives a very clear picture of the dynamics of the functional distance between teacher and student hidden units during learning.

3.4 Output Measure

As a last order parameter we consider the extremality of the output activities

$$\text{Ext} = \frac{1}{\#\text{set}} \sum_{p} \frac{1}{O} \sum_{l=1}^{O} \min\{(1 - O_l(\vec{x}^p))^2, (O_l(\vec{x}^p) - 0)^2\}. \tag{10}$$

The sum over p is taken either over the training or the test set and we normalize over the cardinality of the respective set (denoted by #set). The quantity Ext measures how strongly the network fits the extreme values of the targets, so if the network outputs are close to either 0 or 1 we obtain Ext ~ 0. In this sense Ext is a measure of overfitting, assuming a smooth posterior $q(C_l|\vec{x}, \vec{w}_T)$ of the teacher. As Ext takes non-zero values the student network starts to provide a smooth estimate of the a-posteriori distribution of the teacher.

4 The Simulation

The simulations were performed on a parallel computer (CM5). Every curve in the figures takes about 3-5h of computing time on a 128 respectively 256 partition of the CM5. This setting enabled us to do the statistics for a single teacher over 128-512 samples (different training sets). The exact conditions under which our simulations were performed are

1. A teacher network $\vec{w}_T$ is chosen at random, where weights and biases are normally distributed with zero mean and variance 1.

2. Then a random training set of size t and test set with fixed size 100000 is drawn by choosing $\vec{x}^p$ from a uniform distribution of appropriate width. The output distribution $q(C_l|\vec{x}, \vec{w}_T)$ is generated by the previously chosen teacher $\vec{w}_T$ and the 1 out of M class target vectors $\vec{y}^p$ are generated stochastically.

3. A student $\vec{w}$ is initialized randomly or as the teacher configuration $\vec{w}_T$. Conjugate gradient learning with linesearch on the log likelihood (5) is applied. Given the student has reached a local minimum of the training error (5) we assess the different order parameters of Eq. (7)-(10).

4. Furthermore the generalization ability of the student is measured on the test set via Eq. (6).

5 Higher Order Corrections

To obtain the asymptotic theory for the learning curve of the student networks $\vec{w}$ we have to expand the likelihood function (KL divergence) around the teacher $\vec{w}_T$ following (Amari 1985, Amari et al. 1993, Murata et al. 1993, Akahira et al. 1981). We now give the results for the higher order corrections to the asymptotic expansion yielding a refined scaling law, not only consisting of eq.(1), but of higher order terms, responsible for the deviations seen in the simulation.

$$\epsilon_g = H_0 + \frac{m}{2t} + \frac{A}{t^2} + \text{higher order terms.} \tag{11}$$

The $1/t^2$ corrections have a prefactor A, which is very complicated and unfortunately strongly model dependent. The first $m/2t$ term is model independent.

The variance of the first order term in ϵ_g has the form $\sigma = (m/2t^2)^{-1/2}$. The complete correction term A is discussed in the appendix.

6 Results

In our simulations we can distinguish between three ranges of t, which will be described subsequently. First we summarize the general picture and then we relate this picture to the numerics.

1. small t: in this range we observe strong overfitting, which induces diverging weights and generalization error, whereas the simulations typically show a finite generalization error due to finite numerical precision and the flatness of the error surface.

2. medium t: a $1/t^2$ scaling is observed. So far, neither the statistical physics predictions nor statistical considerations have addressed the scaling of learning curves in a medium range of t. We propose necessary higher order corrections that have to be taken into account to explain the phenomena.

3. large t (asymptotic range): the asymptotics underlying eq.(1) are observed in the range of a large number of patterns.

6.1 Few Examples – overfitting –

In the following we will first give a theoretical explanation of the small t range and then report on our experimental findings.

6.1.1 Why overfitting? – theoretical considerations –

For small t we are below storage capacity. A network is considered to operate below storage capacity if the student can reproduce the correct labeling on the training set with probability 1 and can therefore classify all given training patterns without error. The best and global solution of the learning problem in this case is: one output set to 1, all others equal to zero and diverging weights. If the weights diverge, also the generalization error is bound to go to infinity.
For a fixed architecture the limit of storage capacity depends on the specific sample. Above storage capacity – as the student cannot classify all training patterns correctly for a given sample – a minimum with finite generalization

error and finite weights becomes favorable. In fig. 3 we plot the probability r for finding a finite minimum, computed by averaging over a large number of samples. As we see for $t > 2m$ all students end up in a finite minimum with probability $r = 1$. At $t \sim m$ about half of the students are giving perfect classifications ($r = 1/2$), and therefore diverging generalization error. So r is not only a good parameter to detect the limit of the storage capacity of the classifier, but $r < 1$ can be used as an indication for a diverging generalization error.
So for $t < 2m$, the *averaged* generalization error should be always be infinity, according to our theoretical considerations. On single samples we can of course obtain a finite generalization error, if a student cannot classify all training patterns correctly. In the range of small t an analogy to the transition found in the binary commitee machine (Barkai et al. 1992, Schwarze et al. 1993, Seung et al. 1992, Kang et al. 1993, Saad et al. 1995), could be the transition from infinite to finite weights, respectively KL divergence.

6.1.2 Experimental Results

Plotted in figure 2a is the Kullback-Leibler divergence found in the simulation for a 108 parameter network (8-8-4)[2]. Obviously the generalization error is not diverging. This result is typical for a practical simulation which is limited due to finite precision and the flatness of the error surface.
For $t < m$ the student overfits strongly with outputs tending to take the extreme values 0 or 1 in order to imitate the empirical distribution $q^*(\vec{x}, C_m)$. As one student output tends to 1, the others tend to zero. The value for the extremality parameter – also observed in fig. 3 of the simulation – in this situation is Ext ~ 0 before the bend of the Kullback-Leibler divergence (near $t \sim m$) and Ext > 0 after the bend. Taking extreme output values is only possible if the student weights increase drastically. Although we cannot see the expected diverging weight values we observe in figure 4, that the size of the student weights is very large, until after the transition point, it approaches a similar magnitude as the teacher's weights. The measure ratio O shows a nice agreement with the shape of the generalization error, while ratio H approaches its maximum value after the transition near $t \sim m$. As more examples are learned and the point $t = m$ is passed, we observe a knee in the learning curve and a decrease of the absolute values of the student

[2]For the 8-8-4 network we compute the number of free parameters as $m = (N+1)H + (H+1)M = 108$.

weights. For larger networks we find even stronger bends in the learning curve. In the region of the bend in the KL divergence at $m < t < 4m$ we find a change in the scaling behavior towards a faster scaling law. In this range the outputs start to take non-extreme values and the parameter Ext shows a sharp bend, since more examples are provided to give a smoother estimate of the a-posteriori distribution of the teacher. Also the probability r of finding a finite solution tends to 1 and for $t > 2m$ numerical effects do not have to be considered anymore. We measure that the activities and angles of the teacher and student hidden units are still uncorrelated, i.e. the student hidden units do not correlate to specific teacher hidden units.
We conclude that overfitting effects dominate the small t region to a large extent. They can be measured through the order parameters ratio O and Ext. The region where the average generalization error actually diverges theoretically can be estimated by r. We would like to emphasize that below storage capacity numerical effects that act as regularizers depending on implementation details[3] will typically be observed and are hard to be circumvented.

6.2 Medium range – many examples –

For $4m < t < 30m$ we find a scaling law of $1/t^2$ which is faster than $1/t$. Yet, the exponent is slowly decreasing towards t^{-1} as t is growing towards the large t regime. The higher order corrections of eq.(11) can explain this effect: the farther we are away from the $1/t$ asymptotics the more prominent are the correction terms of Eq.(11).
Note that the above mentioned value $4m$ for the onset of the $1/t^2$ asymptotic region is specific to the example (8-8-4) used frequently in this paper, since the parameter A from Eq.(11) - determining the onset - is unfortunately strongly model dependent (see appendix). In figure 6 we can see the asymptotic region for a number of different networks as a function of m/t. Clearly the range of the $1/t$ regime is completely different for different network configurations.
To have a better impression of the quality of the t^{-1} and t^{-2} scaling, we subtracted $108/2t$ from the data points in figure 5 and clearly see $\epsilon_g = 0$ for $t > 3000$ while for $t > 400$ a t^{-2} fit can be nicely applied.
In the following we will use the term correlation synonymously to the functional distance or the angle. In the t^{-2} range, quantitatively the correlations (angle rH)

[3] In our case the linesearch and the bracketing subroutines have tolerance bounds for the gradients respectively the log likelihood (5). These act implicitely as regularizers.

between teacher and student weights show a transition from a state where the hidden units of the student and the teacher are initially correlated to a certain extent (rH = 0.63) towards asymptotic alignment (rH =1; cf. Fig. 7). Furthermore, if we consider the functional distances Act H_{ij} in fig. 7, we observe an initial overall similar functional distance between student and teacher hidden units ranging from 0.15 to 0.4. For larger t this distance is decreased to zero for one hidden unit, while the others maintain a similar magnitude ranging from 0.15 to 0.35 as before. This effect would also be a candidate for the transition in the binary committee machine (Barkai et al. 1992, Schwarze et al. 1993, Seung et al. 1992, Kang et al. 1993, Saad et al. 1995), although it is by no means similarily abrupt and has to be observed in several order parameters (angle, functional distance and ratio) as proposed above (see also section 6.1.1).
Note that practical applications have usually access to a data size $> 5m^*$, where m^* is the number of effective parameters in the network. So under the conditions pointed out in section 3 we will observe in most practical situations a knee in the learning curve and a faster scaling than $1/t$, i.e. the exponent of t is smaller than -1 and higher order correction terms have to be taken into account to explain this effect.

6.3 Asymptotic Behavior – extensively many examples –

As the asymptotic range is reached slowly, the higher order terms loose their importance and the law stated in eq.(1) is approached. All networks studied exhibit a $m/2t$ scaling in their asymptotic range[4]. In the figures 8a and b we show in particular the 8-8-4 result with an interpolated slope of 57 and the 16-10-4 net (212 parameters) with a slope 104 respectively. Clearly the interpolated region of $m/2t$ is reached at higher t ($t > 5000$) in the larger system. In even larger networks (e.g. 16-12-4) the asymptotic region will shrink and will eventually not be reached for the maximum number of patterns 32768 considered in our simulation. In this case one always has to rely on higher order corrections of the scaling law Eq. (11). In figure 6 we plotted the KL divergence as a function of m/t. For large t all curves coincide with a slope of 1/2.

[4]E.g. 16-4-4 slope: 47, 16-8-4, slope: 98, 16-10-4 slope: 104, 8-8-4 start from teacher slope: 57, 8-8-4 start from random initialization slope: 56.

6.4 Initialization

Most of the figures report on the simulation scenario, where we trained the student network starting from the teacher configuration $\vec{w}_T$. The idea was, that since we consider a local neighbourhood of the maximum likelihood estimator in the asymptotic case, the teacher would be a good starting condition for training. Figure 2a shows the complete learning curve of a 8-8-4 network comparing this initialization of the student to a random one. Except for the asymptotic range both initializations always yield very similar results. From this we conclude that no matter where we start in phase space, the dynamics of learning is always attracted to a local minimum of similar quality as in the case of a start from $\vec{w}_T$. The detailed picture of the asymptotic range is given in figure 2b. Clearly, starting from a random initial state makes the learning converge to a higher local minimum in the generalization error only in the asymptotic range. Nevertheless, since the asymptotic theory is valid in any local minima close to the teacher, we observe the same asymptotic $m/2t$ scaling for the random initialization as for a start from the teacher (cf. fig.2b). Note however, that the learning speed is increased by 20% using the teacher as initial starting point of learning.

7 Discussion and Outlook

In our numerical study we observed a rich structure in the learning curves of continuous feed-forward networks. For a small number of patterns we find a phase of strong overfitting, where the outputs take extreme values in their estimate of $q(C_l|\vec{x}, \vec{w}_T)$ (fig. 3) and the student can classify all training patterns correctly. We are below storage capacity of the classifier, so the weights and the generalization ability should theoretically diverge. This fact is not observed in a typical simulation due to numerical effects of finite precision (inducing an implicit regularization) and the flatness of the error surface. As the number of patterns increases beyond storage capacity, the Kullback-Leibler divergence reaches also theoretically the finite value found in the simulation and the outputs start estimating smoother probabilities. The size of the student weights becomes comparable to the teacher weights. The bend of the learning curve is followed by a region of $1/t^2$ scaling when t is increased. Asymptotically we confirm the $m/2t$ behavior.

From our results it seems important to reach the $1/t^2$ phase as fast as possible

to learn efficiently without overfitting and to obtain a smooth estimate of the a-posteriori distribution. Furthermore, as a smooth estimate is obtained, the network is finally free to learn in a collective manner, i.e. the activity of one student hidden unit becomes highly correlated to one specific teacher hidden unit (figs. 7).
Practical applications have usually access to datasets large enough to enter the $1/t^2$ range. If maximum likelihood training and no early stopping method is used – according to our results – typically both, a knee and a faster scaling in the learning curve should be observed. Yet, the range of the asymptotic $1/t$ scaling seems to be too far from realistic sizes of data sets available to most practical users of neural nets.
We would like to emphasize that we *always* find a faster scaling than $1/t$ between the small t overfitting phase and the asymptotic phase. For this reason model selection criteria which are usually based on asymptotic or a certain overall assumptions on the smoothness of learning curves are likely to perform weakly, since they do neither capture the transition encountered nor the faster scaling observed (see also Kearns et al. 1995).
Further investigation is focussed on the measurement of scaling laws in a real practical application and on algorithms that use early stopping to avoid overlearning or overfitting effects (Amari et al. 1995).

8 Acknowledgments

We would like to thank the participants of the NNSMP and the Snowbird workshop for fruitful and stimulating discussions. K. -R. M. thanks for valuable discussions with S. Bös, T. Heskes, A. Herz and for warm hospitality during his stay at the Beckman Instititute in Urbana, Illinois. We further gratefully acknowledge computing time on the CM5 in Urbana (NCSA) and in Bonn. This work was supported by the National Institutes of Health (P41RRO 5969) and K. -R. M. is supported by the European Communities S & T fellowship under contract FTJ3-004.

A Appendix

We now describe the details of the asymptotic theory for the higher order corrections. The conditions for an asymptotic evaluation of ε_g are $t \to$ large and a

realizable teacher machine which has parameter $\mathbf{w}_T$. The present framework can be readily extend to unrealizable cases.

A.1 Asymptotic distribution of the m.l.e. $\hat{\mathbf{w}}$

Let us normalize the maximum likelihood estimator (m.l.e.) $\hat{\mathbf{w}}$ as

$$\tilde{\mathbf{w}} = \sqrt{t}(\hat{\mathbf{w}} - \mathbf{w}_T).$$

Then, the error $\tilde{\mathbf{w}}$ is asymptotically normally distributed

$$p(\tilde{\mathbf{w}}; \mathbf{w}_T) = \varphi(\tilde{\mathbf{w}}; G) + \mathcal{O}\left(\frac{1}{\sqrt{t}}\right), \text{where } \varphi(\tilde{\mathbf{w}}; G) = \frac{1}{\sqrt{(2\pi)^m |G|}} \exp\{-\frac{1}{2}\tilde{\mathbf{w}}^T G^{-1} \tilde{\mathbf{w}}\}$$

with mean 0 and variance matrix (g^{ij}), where (g^{ij}) is the inverse of the Fisher information matrix $G = (g_{ij})$

$$g_{ij} = E\left[\frac{\partial^2 \log p(C_i, \mathbf{x}; \mathbf{w}_T)}{\partial w^j \partial w^i}\right].$$

The higher-order Edgeworth expansion gives

$$\begin{aligned} p(\tilde{\mathbf{w}}; \mathbf{w}_T) &= \varphi(\tilde{\mathbf{w}}; G)\{1 + A_t(\tilde{\mathbf{w}})\} \\ A_t(\tilde{\mathbf{w}}) &= \frac{1}{6\sqrt{t}}(K_i h^i + K_{ijk} h^{ijk}) \\ &\quad \frac{1}{4t}\left\{C_{ij} h^{ij} + \frac{1}{6} C_{ijkl} h^{ijkl} + \frac{1}{18} K_{ijk} K_{lmn} h^{ijklmn}\right\} \\ &\quad + \mathcal{O}\left(\frac{1}{t\sqrt{t}}\right). \end{aligned}$$

Here, h^i, h^{ij}, h^{ijk}, h^{ijkl}, $\cdots$ are the tensorial Hermite polynomials with the metric $G = (g_{ij})$. For example

$$\begin{aligned} h^i &= \tilde{w}^i \\ h^{ij} &= \tilde{w}^i \tilde{w}^j - g^{ij}, \\ h^{ijk} &= \tilde{w}^i \tilde{w}^j \tilde{w}^k - (g^{ij} \tilde{w}^k + g^{ik} \tilde{w}^j + g^{kj} \tilde{w}^i), \\ h^{ijkl} &= \tilde{w}^i \tilde{w}^j \tilde{w}^k \tilde{w}^l - 6 g^{(ij} \tilde{w}^k \tilde{w}^{l)} + 3 g^{(ij} g^{kl)}, \end{aligned}$$

etc., where ( ) attached to indices denotes symmetrization with respect to the indices inside the bracketts ( ).

The Edgeworth expansion of asymptotic distributions of m.l.e. was given by many researchers in the eighties, e.g., Akahira and Takeuchi (1981), Amari (1984). Amari gave its geometrical interpretation in the framework of curved exponential families.
From this, we have the moments of the error in parameter space $\tilde{\mathbf{w}} = \sqrt{t}(\hat{\mathbf{w}} - \mathbf{w}_T)$.

$$\begin{aligned}
E[\tilde{\mathbf{w}}] &: E[\tilde{w}^i] = \frac{1}{\sqrt{t}} K^i \\
E[\tilde{\mathbf{w}}\tilde{\mathbf{w}}^T] &: E[\tilde{w}^i \tilde{w}^j] = g^{ij} + \frac{1}{t} A^{ij}, \\
E[\tilde{\mathbf{w}}\tilde{\mathbf{w}}\tilde{\mathbf{w}}] &: E[\tilde{w}_i \tilde{w}^j \tilde{w}^k] = \frac{1}{\sqrt{t}} A^{ijk}, \\
E[\tilde{\mathbf{w}}\tilde{\mathbf{w}}\tilde{\mathbf{w}}\tilde{\mathbf{w}}] &: E[\tilde{w}^i \tilde{w}^j \tilde{w}^k \tilde{w}^l] = 3g^{(ij} g^{kl)} + \frac{1}{t} A^{ijkl},
\end{aligned}$$

where A's are given explicitly in Akahira and Takeuchi (1981) and Amari (1985).

A.2 Expansion of the Kullback-Leibler Divergence

Let $\mathbf{w} = \mathbf{w}_T + \Delta\mathbf{w}$. Then, by Taylor expansion, we have

$$\begin{aligned}
D(\mathbf{w}_T, \mathbf{w}) &= \int p(\mathbf{x}, \mathbf{w}_T) \log \frac{p(\mathbf{x}, \mathbf{w}_T)}{p(\mathbf{x}, \mathbf{w})} d\mathbf{x} \\
&= E_{\mathbf{w}_T}[l(\mathbf{w}_T) - l(\mathbf{w})],
\end{aligned}$$

where $\mathbf{x}$ implies hereafter the pair $(\mathbf{x}, C_i)$ and $l(\mathbf{w}) = \log p(\mathbf{x}, \mathbf{w})$. By expansion, we have

$$\begin{aligned}
l(\mathbf{w}) &= l(\mathbf{w}_T + \Delta\mathbf{w}) \\
&= l(\mathbf{w}_T) + \frac{\partial l}{\partial w_i}\Delta w_i + \frac{1}{2}\sum \frac{\partial^2 l}{\partial w_i \partial w_j}\Delta w_i \Delta w_j \\
&\quad + \frac{1}{6}\sum \frac{\partial^3 l}{\partial w_i \partial w_j \partial w_k}\Delta w_i \Delta w_j \Delta w_k \\
&\quad + \frac{1}{24}\sum \frac{\partial^4 l}{\partial w_i \partial w_j \partial w_k \partial w_l}\Delta w_i \Delta w_j \Delta w_k \Delta w_l + \mathcal{O}(|\Delta\mathbf{w}|^5).
\end{aligned}$$

Hence we arrive at

$$\begin{aligned}
D(\mathbf{w}_T, \mathbf{w}) &= -\frac{1}{2} L_{ij} \Delta w^i \Delta w^j - \frac{1}{6}\sum L_{ijk} \Delta w_i \Delta w_j \Delta w_k \\
&\quad - \frac{1}{24}\sum L_{ijkl} \Delta w_i \Delta w_j \Delta w_k \Delta w_l + \mathcal{O}(|\Delta w|^5),
\end{aligned}$$

where for example L_{ij} is given by

$$L_{ij} = E_{\mathbf{w}_T}\left[\frac{\partial^2 \log p(C_i, \mathbf{x}; \mathbf{w}_T)}{\partial w_i \partial w_j}\right].$$

Therefore the expansion of ε_g is given as

$$\begin{aligned}
\varepsilon_g &= E[D(\mathbf{w}_T, \hat{\mathbf{w}})] = E[D(\mathbf{w}_T, \mathbf{w}_T + \frac{1}{\sqrt{t}}\tilde{\mathbf{w}})] \\
&= -\frac{1}{2}\sum L_{ij} E[\frac{1}{t}\tilde{w}^i \tilde{w}^j] \\
&\quad -\frac{1}{6}\sum L_{ijk} E[\frac{1}{t\sqrt{t}}\tilde{w}^i \tilde{w}^j \tilde{w}^k] \\
&\quad -\frac{1}{24}\sum L_{ijkl} E[\frac{1}{t^2}\tilde{w}^i \tilde{w}^j \tilde{w}^k \tilde{w}^l] + \mathcal{O}\left(\frac{1}{t^2\sqrt{t}}\right) \\
&= \frac{1}{2t}\sum g_{ij}(g^{ij} + \frac{1}{t}A^{ij}) \\
&\quad -\frac{1}{6}\sum L_{ijk} \cdot \frac{1}{t^2} A^{ijk} - \frac{1}{24}\sum L_{ijkl}\frac{1}{t^2} \cdot 3 g^{ij} g^{kl} \\
&\quad +\mathcal{O}\left(\frac{1}{t^2\sqrt{t}}\right) \\
&= \frac{m}{2t} + \frac{A}{t^2} + \mathcal{O}\left(\frac{1}{t^2\sqrt{t}}\right),
\end{aligned}$$

where

$$A = \sum g_{ij} A^{ij} - \frac{1}{6} L_{ijk} A^{ijk} - \frac{1}{8}\sum L_{ijkl} g^{ij} g^{kl}.$$

This gives the higher-order correction to the learning curve,

$$\begin{aligned}
\varepsilon_g &= \langle -\log p(C_i, \mathbf{x}|\hat{\mathbf{w}})\rangle = E_{\hat{w}} E_{(x,C_i)}\left[-\log p(C_i, \mathbf{x}|\hat{\mathbf{w}})\right] \\
&= H_0 + \frac{m}{2t} + \frac{A}{t^2} + \text{higher-order terms.}
\end{aligned}$$

The result is also confirmed by Komaki (1994), where he obtained the Kullback-Leibler divergence with the modification of the predictive distribution by the normal mixture direction. When the normal correction is put equal to 0, his result gives

$$D(\mathbf{w}_T, \hat{\mathbf{w}}) = \frac{m}{2t} + \frac{A}{4t^2} + \mathcal{O}\left(\frac{1}{t^2}\right),$$

where A is explicitly obtained. It includes the curvature terms, bias gradient terms, geometrical and the fourth cumulant terms etc., in agreement with that given by Amari (1985).

References

Akahira, M., Takeuchi, K., Asymptotic efficiency of statistical estimators: Concepts and higher order asymptotic efficiency, Springer New York (1981)

Amari, S., Differential geometrical methods in statistics, Lecture Notes in Statistics No.28, Springer New York (1985)

Amari, S., Murata, N., Neural. Comp. 5, 140 (1993)

Amari, S., Murata, N., Müller, K.-R., Finke, M., Yang, H., Statistical Theory of Overtraining - Is Cross-Validation Effective?, University of Tokyo Technical Report and submitted to Nips 95 (1995)

Barkai, E., Hansel, D., Sompolinsky, H., Phys.Rev.A, 45, 4146 (1992)

Baum, E.B., Haussler, D., Neural. Comp. 1, 151 (1989)

Finke, M., Müller, K.-R., 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 (1994)

Haussler, D., Kearns, M., Seung, S., Tishby, N., Rigorous Learning Curve Bounds from Statistical Mechanics, preprint (1994)

Heskes, T.M., Kappen, B., Phys.Rev.A, 440, 2718 (1991)

Kang, K. Oh, J.-H., Kwon, C., Park, Y., Phys.Rev.E 48, 4805 (1993)

Kearns, M., Mansour, Y., Ng, A.Y., Ron, D., An experimental and theoretical comparison of model selection methods, preprint to appear in COLT 95 (1995)

Komaki, F., On asymptotic properties of predictive distributions, METR94-21, University of Tokyo (1994)

Müller, K.-R., Finke, M., Murata, N., Schulten, K., Amari, S., On Large Scale Simulations for Learning Curves, in Proc. of the Workshop on the Theory of Neural Networks: The Statistical Mechanics Perspective, POSTECH, Pohang, Korea, to appear in World scientific Pub. (1995)

Murata, N., Yoshizawa, S., Amari, S., NIPS 5, Morgan Kaufmann, San Mateo, 607 (1993)

Opper, M., Kinzel, W., Kleinz, J., Nehl, R., J.Phys. A:Math.Gen. 23, L581 (1990)

Opper, M., Kinzel, W., Statistical Mechanics of Generalization, in Physics of neural networks III, (eds.) E. Domany, J.L. van Hemmen and K. Schulten, Springer Heidelberg (1995)

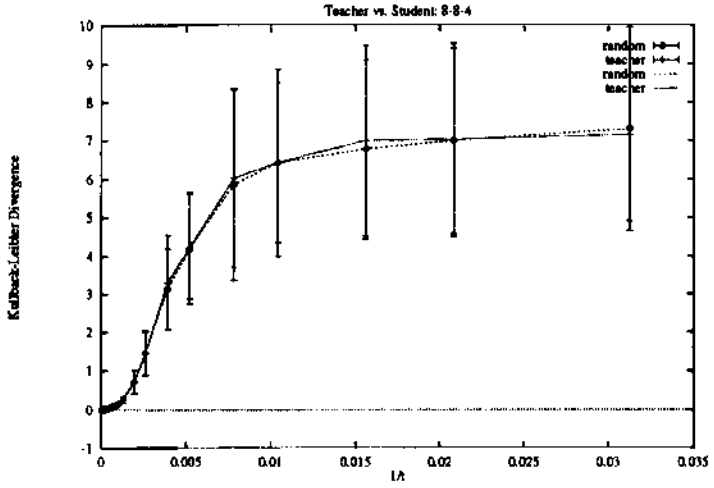
Opper, M., Haussler, D., Calculation of the Learning Curve of Bayes Optimal Classification Algorithm for Learning a Perceptron with Noise, in Proc. of COLT (1991)

Saad, D., Solla, S., preprint, submitted to Phys. Rev. E, also to appear on Phys. Rev Lett. (1995)

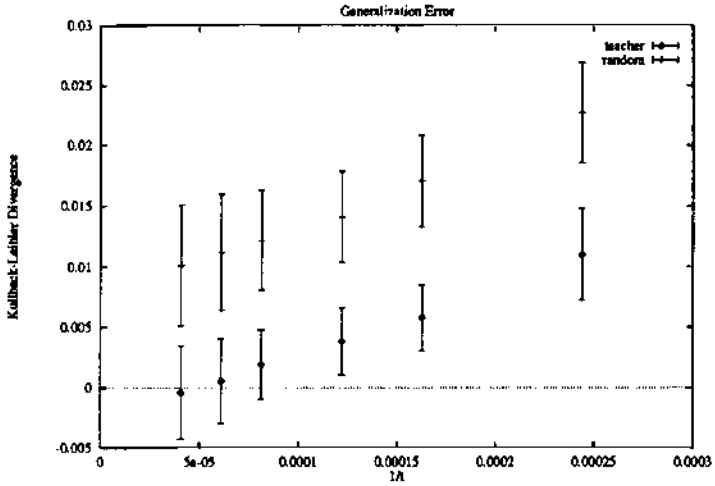
Schwarze, H., Hertz, J., Europhys. Lett. 21, 785 (1993)

Seung, S., Sompolinsky, H., Tishby, N., Phys.Rev.A 45, 6056 (1992)

Sompolinsky, H., Tishby, N., Seung, S., Phys.Rev.Lett. 65, 1683 (1990)
Watkin, T.L.H, Rau, A., Biehl, M., The Statistical Mechanics of Learning a Rule, Rev. Mod. Phys. 65, 499 (1993)



(a)



(b)

Figure 2: *Plotted are the simulated generalization values over $1/t$ for an 8-8-4 network. We compare the start from the teacher w_T and a random initialization (a) for the whole learning curve and (b) for the asymptotic area. Note that in the asymptotic range we find for the random started simulation higher values for the KL divergence, i.e. the simulation gets stuck earlier in local minima.*

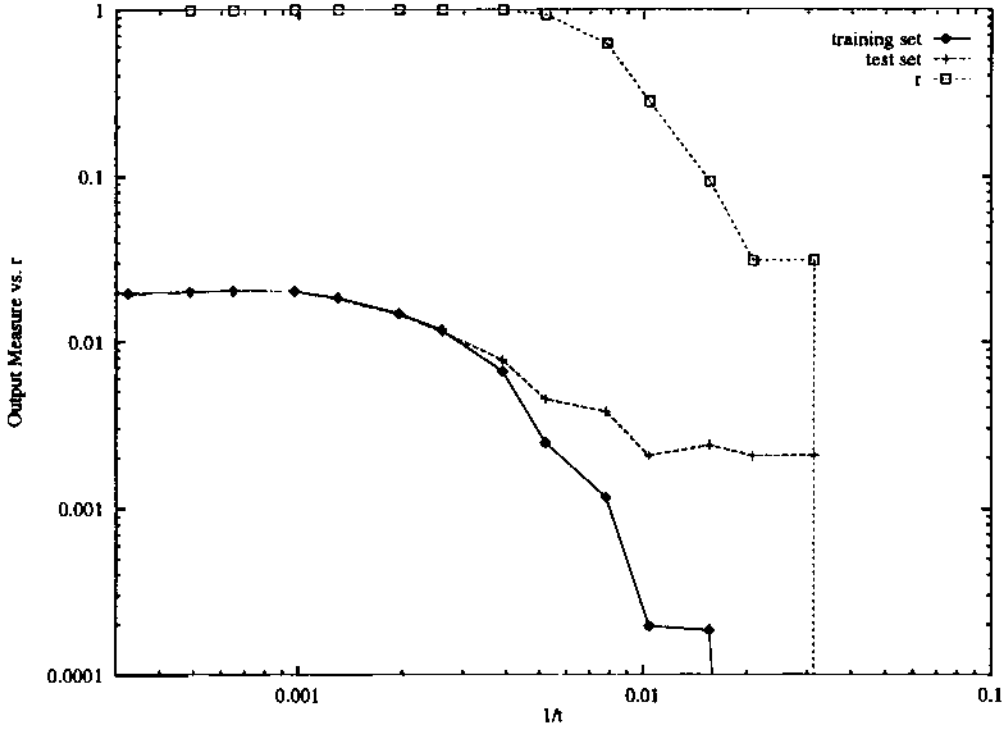


Figure 3: Ext *measured on training and test set indicates, whether the output activities take extreme values as a function of* $1/t$ *(8-8-4 net). A value of zero indicates extreme output values, i.e. 0 or 1. Compared to* Ext *is the probability* r *of wrong classification on the training set, for* $r = 0$ *only a diverging KL divergence is a valid solution, for* $r = 1$ *a finite minimum is more favourable.* r *is a good parameter to detect the limit of the storage capacity of the classifier.*

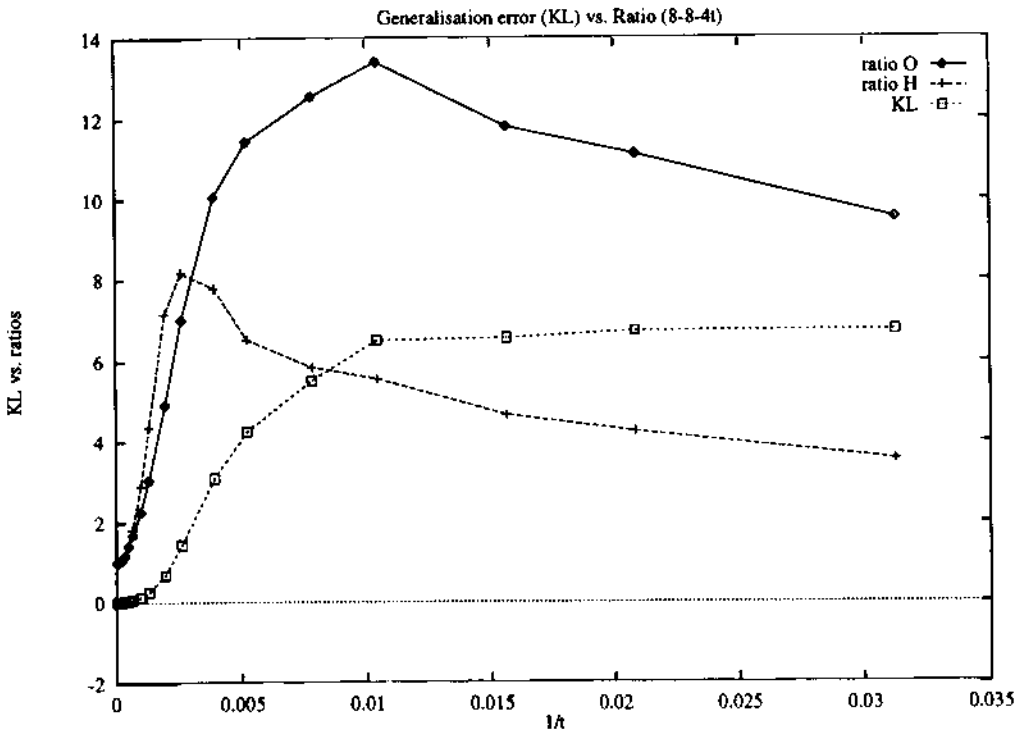


Figure 4: *Ratio of the student and teacher weights of hidden to output units (ratio O) and input to hidden units (ratio H) versus Kullback-Leibler divergence as a function of* $1/t$ *(8-8-4 net). Note the strong increase of ratio O at the bend of KL near* $t \sim m$.

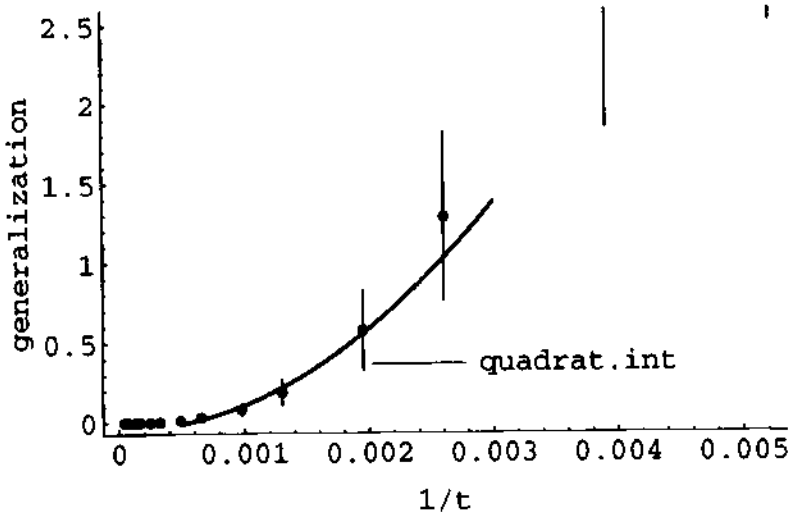


Figure 5: *Plotted are the simulated generalization values over* $1/t$ *for an 8-8-4 network. For large* t *an exponent of the scaling law smaller than* -1 *is observed. Shown are the simulated values minus* $m/2t$*. Above* $t = 3000$ *we find the scaling predicted in eq.(1), e.g. the points are on the line* $\epsilon_g = 0$*. Below* $t = 3000$ *an quadratic interpolation is applied, yielding the necessary higher order corrections of eq.(1).*

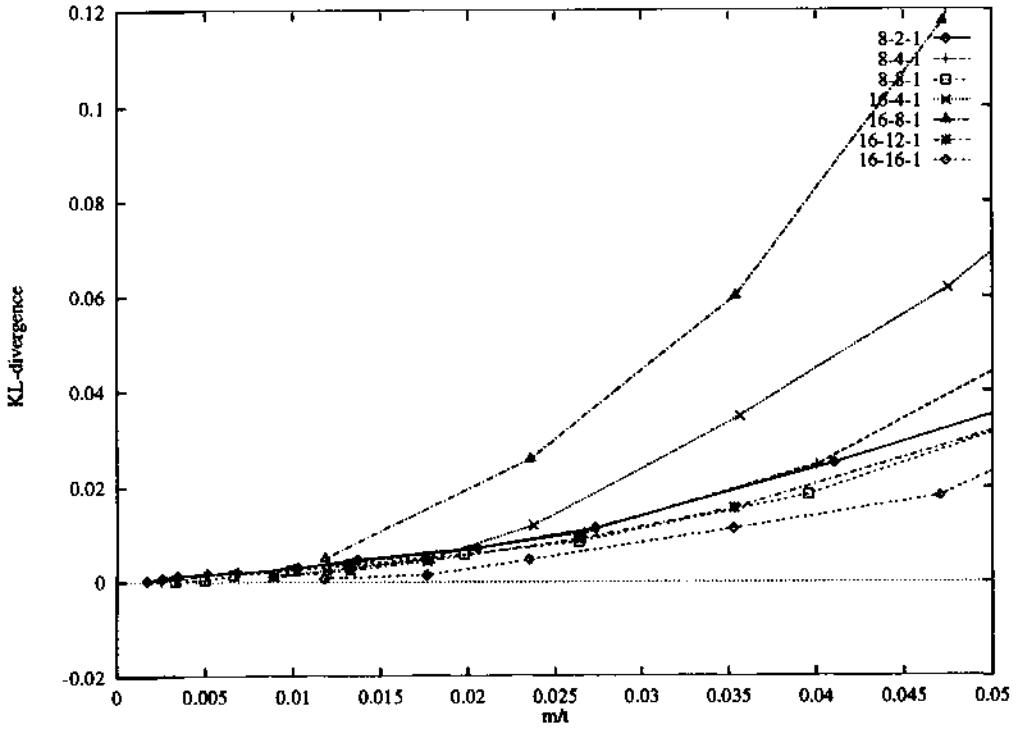


Figure 6: *Kullback-Leibler divergence as a function of m/t for different network sizes as indicated in the caption. Asymptotically all curves coincide. Furthermore, note the different onset of the $1/t^2$ region for the different network sizes.*

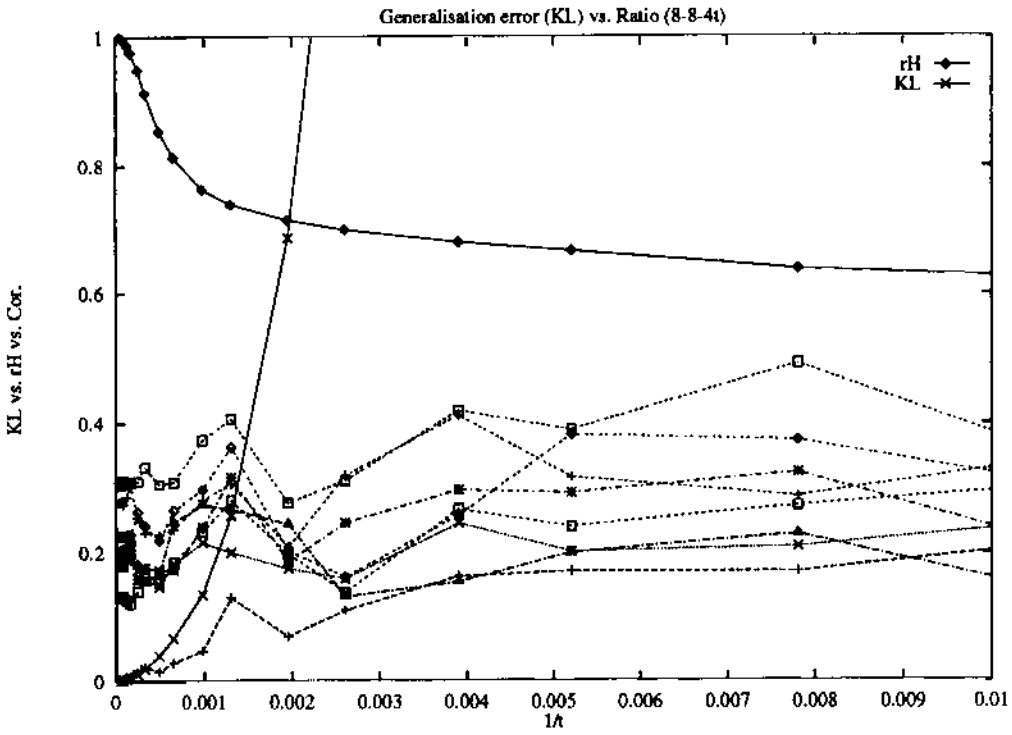
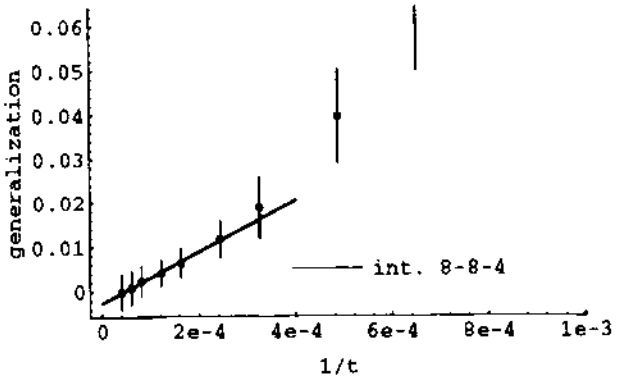
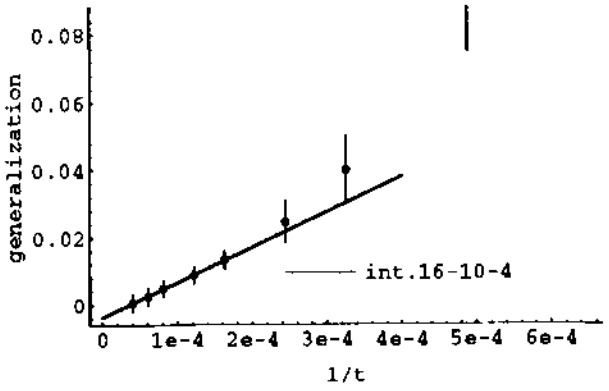


Figure 7: *Angle between student and teacher weights of input to hidden units (rH) versus $\mathcal{L}^2$ functional distance between the activity of student hidden unit 1 and all teacher hidden unit activities versus Kullback-Leibler divergence (KL) as a function of $1/t$ (8-8-4 net).*



(a)



(b)

Figure 8: *Plotted are the simulated generalization values in the asymptotic range for (a) the 8-8-4 network (108 parameters) and (b) for the 16-10-4 network (212 parameters). In both cases a clear scaling as* $1/t$ *is seen.*