LARGE SCALE SIMULATIONS FOR LEARNING CURVES

K.-R. Müller[†+*], M. Finke[‡], N. Murata[†], K. Schulten[+], S. Amari[†]

[†] *Dept. of Math. Engineering and Inf. Physics, University of Tokyo*
Hongo 7-3-1, Bunkyo-ku, Tokyo 113, Japan, E-mail: Klaus@first.gmd.de

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

[+] *Beckman Institute, University of Illinois, 405 North Mathews Av., Urbana IL., USA*

ABSTRACT

The universal asymptotic scaling laws proposed by Amari et al. [2,11] are studied in large scale simulations using a CM5. Small stochastic feed-forward networks trained with back-propagation and conjugate gradient descent are investigated. In the range of a large number of training patterns t, the predicted asymptotic $1/t$ scaling is observed. For a medium range t a faster scaling in the number of training patterns t than $1/t$ is observed. This effect is explained by using higher order corrections of the likelihood expansion. For small t it is shown, that the scaling law changes drastically, when the network undergoes a transition from permutation symmetric to permutation symmetry broken phase. This effect is related to previous theoretical work [15,3,17,16,8].

1. Introduction

Recently a growing interest in learning curves, i.e. scaling laws for the asymptotic behaviour of the learning and generalization ability of neural networks has emerged [2,11,4,14,12,13]. 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 in general. The large-scale simulations presented in this paper are addressing the question of scaling laws for training and generalization errors in small feed-forward networks with so far up to 256 parameters, trained on a finite number of training samples of up to 32768 patterns.

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 [12,13,9,18,7]. Several authors have observed a phase transition, when training with small to medium sized sample sets. For example, the generalization of the committee machine first scales as N/t in a so-called permutation symmetric phase whereas for more patterns a phase transition takes place and the system scales as NH/t in the permutation symmetry broken phase [15,3,17,16,8]. Here, parameter N denotes the number of inputs, while H denotes the number of hidden units. If we assume a student network learning from a teacher network by examples, then this student can – for small t – in the so called per-

*Permanent address: GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany.

mutation symmetric phase find a large set of possible solutions. This large number of solutions is due to the equality of all permutations of the teachers hidden units which basically lead to the same result in terms of the objective function to be minimized. But, from a certain number of patterns on a phase transition occurs, where the set of possible solutions is getting smaller and every student hidden unit has to decide for a teacher hidden unit. Thus, the permutation is fixed and the permutation symmetry is broken. The transition which takes place in a non-asymptotic range of t is not accessible to methods of statistical inference, which have been used to understand the asymptotic learning behaviour of general stochastic machines. This statistical approach is based on an asymptotic expansion of the likelihood of the estimating machines, always assuming a maximum likelihood estimator [2,11].

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 [4,14,6].

We will now state the major results of this paper. The first purpose of our investigation was to study, whether the well-known universal asymptotic scaling laws found by Amari et al. can be observed in a simulation with 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 [2,11]. 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. The second goal was to study whether the breaking of permutation symmetry also occurs in continuous networks with continuous patterns. Our simulations are using standard continuous feed-forward networks, backpropagation and a conjugate gradient descent in the Kullback-Leibler divergence (see section 2). In the simulations (see sections 3 and 5) we distinguish between three ranges of t:

1. small t: in this range of t, we observe a phase transition from $1/t$ scaling (cf. eq.(1)) towards a faster scaling for medium sized example sets (i.e. transition from permutation symmetric to permutation symmetry broken phase).

2. medium t: 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 here (section 4).

3. large t (asymptotic range): the asymptotics underlying eq.(1) are observed in the range of a large number of patterns. Nevertheless this regime shrinks for large networks, since only a maximum number of patterns ($t = 32768$) can be simulated for technical reasons.

We would like to stress the fact, that in almost all practical situations a faster scaling law than $1/t$ will be observed, 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. As the asymptotic range is reached slowly, the higher order terms loose their importance and the law stated in eq.(1) is approached.

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_i of the ith output unit is calculated via the softmax squashing function $o_i = \exp(h_i)/\sum_k \exp(h_k)$, where $h_i = \sum_j w_{ij}^O s_j - \vartheta_i^O$ is the local field potential. The network parameters consist of biases ϑ and weights $\vec{w}$. $\vec{s}$ denotes the state of hidden neurons which is computed using the Fermi function, i.e. $s_i = [1+\exp(-\sum_j w_{ij}^H x_j - \vartheta_i^H)]^{-1}$, where $\vec{x}$ is the input to the network. The input layer is connected to the hidden layer, the hidden layer is connected to the output layer, but no short-cut connections are present. Although the network is completely deterministic, it is constructed to approximate class conditional probabilities [5]. 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 $q(C_i|\vec{x})$ over the classes C_i $(i = 1 \dots M)$ given a random input $\vec{x}^p$. We use the same network topology for teacher and student. Thus, we assume that the model is faithful, i.e. the teacher distribution can be exactly represented by a student $q(C_i|\vec{x}) = p(C_i|\vec{x}, \vec{w}_T)$.

A training and test set of the form $\mathcal{S} = \{(\vec{x}^p, \vec{y}^p)|p = 1 \dots t\}$ is generated randomly, by drawing samples of $\vec{x}$ from a normal resp. uniform distribution and forward propagating $\vec{x}^p$ through the teacher network. Then, according to the teachers' outputs $q(C_i^p|\vec{x}^p)$ one output unit is set to one stochastically and all others are set to zero leading to the target vector $\vec{y}^p = (0, \dots, 1, \dots, 0)$. For training the student network $\vec{w}$ we use a backpropagation algorithm with conjugate gradient descent to minimize our objective function: the Kullback-Leibler difference

$$D(q, p(\vec{w})) = \int d\vec{x} \sum_{i=1}^{M} q(\vec{x}) q(C_i|\vec{x}) \ln \frac{q(C_i|\vec{x})}{p(C_i|\vec{x}, \vec{w})}.$$

Here $q(C_i|\vec{x})$ denotes the class conditionals, respectively outputs of the teacher and $p(C_i|\vec{x}, \vec{w})$ are the class posteriors as approximated by the student network. The Kullback-Leibler difference 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 difference 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 $q(C_i|\vec{x})$. In practical applications there is typically no such

knowledge. So only the empirical Kullback-Leibler difference

$$D(q^*, p(\vec{w})) = -\frac{1}{t}\sum_p \ln p(c^p|\vec{x}^p, \vec{w}) \tag{2}$$

will be available, where q^* denotes the empirical distribution and c^p refers to the correct class associated to $\vec{x}^p$. The results found for this case have practical importance, since as mentioned above in general practical problems **only** the empirical distribution is known. A better approximation to the KL difference is computationally more intensive, but all necessary ingredients are known

$$D(q^*, p(\vec{w})) = -\frac{1}{t}\sum_p \sum_{i=1}^{M} q(C_i|\vec{x}^p) \ln p(C_i|\vec{x}^p, \vec{w}). \tag{3}$$

So given a random uniformly distributed input, we can use the a-posteriori probabilities $q(C_i|\vec{x}^p)$, which are exactly the output values given by the teacher networks on the presentation of an input vector $\vec{x}^p$. In our simulation both measures (3) and (2) are studied.

3. 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 set). 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 and the output distribution $q(C_i|\vec{x})$ is generated by the previously chosen teacher $\vec{w}_T$ and the class target vectors $\vec{y}$ are generated stochastically.

3. The generalization ability is measured on the test set in two ways: On the one hand we use the **empirical** Kullback-Leibler difference (2), so no information from the teacher distribution is actually used, and on the other hand we measure (3) since we assumed uniformly distributed inputs and the a-posteriori probabilities are simply the output activities of the teacher network.

4. Conjugate gradient learning on the empirical Kullback-Leibler distance (2) is applied starting from the teacher configuration $\vec{w}_T$ or from some random initial vector. Given we have reached a local minimum of that training error we assess the generalization ability on the test set. This solution is assumed to be very close to the maximum likelihood solution used in Amari's Universal Scaling Law.

As mentioned above we refer to the Kullback-Leibler divergence to measure the distance between $q(C|\vec{x},\vec{w}_T)$ (teacher) and $p(C|\vec{x},\vec{w})$ (student). Since these are basically the same parameterized distributions the Kullback-Leibler divergence can be equivalently considered as measuring the distance between the parameter vectors $\vec{w}_T$ and $\vec{w}$ with respect to p and q. In order to examine the relation between teacher $\vec{w}_T$ and student $\vec{w}$ in more detail, we introduce another measure of overlap that is supposed to be independent of the probability distribution p. So, contrary to the overlap $R_{ij} = 1/N\vec{w}_{Ti} \cdot \vec{w}_j$ as defined for a committee machine, we have to consider all permutations σ of the hidden units in the multi layer perceptron case and make the overlap independent of the actual permutation. Let $\vec{w}^H_{Ti\bullet}$ and $\vec{w}^H_{\sigma(i)\bullet}$ be the vectors of all weights from the input layer into hidden unit i for teacher and student respectively, and let $\vec{w}^H_{Ti\bullet} \cdot \vec{w}^H_{\sigma(i)\bullet} = \sum_j w^H_{Tij} w^H_{\sigma(i)j}$ denote the inner product of the two vectors. Based on this notation we define two measures for the correlation of the hidden units as

$$\mathrm{rH} = \max_\sigma \frac{1}{H} \sum_i^H \frac{\vec{w}^H_{Ti\bullet} \cdot \vec{w}^H_{\sigma(i)\bullet}}{||\vec{w}^H_{Ti\bullet}||||\vec{w}^H_{\sigma(i)\bullet}||} \qquad \text{and} \qquad \mathrm{rHmax} = \max_\sigma \max_i \frac{\vec{w}^H_{Ti\bullet} \cdot \vec{w}^H_{\sigma(i)\bullet}}{||\vec{w}^H_{Ti\bullet}||||\vec{w}^H_{\sigma(i)\bullet}||},$$

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 hidden units are maximally correlated. In order to detect the transition of the first pair of hidden units (teacher and student unit) into a correlated phase, we have defined rHmax where the sum over all overlaps is replaced by a maximum over all hidden units. Equivalently, we define the output units overlap independent of the permutation as

$$\mathrm{rO} \;=\; \max_\sigma \frac{1}{H} \sum_i^H \frac{\vec{w}^O_{T\bullet i} \cdot \vec{w}^O_{\bullet\sigma(i)}}{||\vec{w}^O_{T\bullet i}||||\vec{w}^O_{\bullet\sigma(i)}||},$$

where $\vec{w}^O_{T\bullet i}$ denotes the vector of weights from hidden unit i to all output units. This overlap of the output units does not consider the correlation of the hidden units involved. Therefore, the following overlap rHO combines the two perspectives of rH and rO by calculation the maximum correlation of the products of hidden and output overlaps, i.e.

$$\mathrm{rHO} \;=\; \max_\sigma \frac{1}{H} \sum_i^H \frac{\vec{w}^H_{Ti\bullet} \cdot \vec{w}^H_{\sigma(i)\bullet}}{||\vec{w}^H_{Ti\bullet}||||\vec{w}^H_{\sigma(i)\bullet}||} \frac{\vec{w}^O_{T\bullet i} \cdot \vec{w}^O_{\bullet\sigma(i)}}{||\vec{w}^O_{T\bullet i}||||\vec{w}^O_{\bullet\sigma(i)}||}.$$

In order to interpret the quantitative behaviour of these overlap measures which are all based on scalar products or equivalently on angles between teacher and student vectors we have to consider the following. Since the surface area of n dimensional hypersphere can be calculated by

$$\int_0^{2\pi} d\theta_1 \int_0^\pi \sin\theta_2 d\theta_2 \cdots \int_0^\pi \sin^{n-2}\theta_{n-1} d\theta_{n-1} \int_0^\alpha \sin^{n-1}\theta_n d\theta_n.$$

we can get full measure of the complete hypersphere by setting $\alpha = \pi$. The ratio of correlated area can be calculated by

$$\frac{\int_0^\alpha \sin^{n-1}\theta d\theta}{\int_0^\pi \sin^{n-1}\theta d\theta}.$$

In case of 8-8-* networks, we have $n = 9$. Here, the size of the area in which the angle is less than $\pi/3$ ($\cos\theta > 0.5$) is about 0.06, if the angle is less than $\pi/4$ (0.71) the size is 0.007 and in case the angle is less than $2\pi/5$ (0.31) we have 0.17. Therefore, even small numbers of overlap mean a reasonable high correlation between teacher and student since the ratio of the correlated area tends to zero as the dimension n goes to infinity.

4. 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 difference) around the teacher $\vec{w}_T$ following [1,2,11,19]. 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.

$$\epsilon_g = H_0 + \frac{m}{2t} + \frac{A}{t^2} + \text{higher order terms.} \tag{4}$$

The $1/t^2$ corrections have a prefactor $A = \mathcal{O}(m^2)$, which is very complicated and strongly model dependent. The first $m/2t$ term is model independent. The variance of the first order term in ϵ_g has the form $\sigma = (m/2t^2)^{-1/2}$. A complete discussion of the variance and the correction term A goes beyond the scope of this contribution (see [10] for details).

5. Results

5.1. Permutation Symmetry Breaking

Our numerical results show a picture of a transition from permutation symmetry to broken permutation symmetry (see fig. 1a). Plotted is the Kullback-Leibler difference found in the simulation for a 108 parameter network (8-8-4). Clearly, the slope 24 of the interpolation for small t in the figure shows a change compared the slope 54, which would be expected from $m/2t$. Since the interpolation is smooth and linear in $1/t$, the number of effective parameters m^* in this range of t seems to involve only **part** of the network. From the qualitative breakdown in figure 1(a) we would estimate the transistion to happen close to $t = 400$, where the $1/t$ behaviour changes to a faster

scaling law. Measuring the overlap parameters we encounter a smoother picture. The plot of the different overlap measures in figure 3 shows a change in the hidden-output overlap rHO from 0.35 at $t = 100$ to 0.9 at $t = 8000$, where the student permutation of the hidden units is found to be fixed with respect to the teacher. But considering the change in our correlation measures and regarding the high dimensionality of the space, this turns out to be a significant transition. For larger networks both Kullback Leibler difference and correlation measures steepen up.

5.2. Medium range – many examples –

After a critical range close to $t \sim 400$, we observe a change towards a faster scaling than $1/t$, while we enter the range of medium 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.(4) can explain this effect. To have a better impression of the quality of the t^{-1} and t^{-2} scaling, we subtracted $108/2t$ from the data points and clearly see $\epsilon_g = H_0$ for $t > 3000$ while for $t > 500$ a t^{-2} fit can can be nicely applied. Note that practical applications have usually access to a data size $> 5m^*$, where m^* is the number of effective parameters in the network. So both: a knee in the learning curve and a faster scaling than $1/t$ should be observed in most practical situations.

5.3. Asymptotic Behaviour

All networks studied exhibit a $m/2t$ scaling in their asymptotic range[a]. In the figures 2(a) and (b) we show the 8-8-4 results 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 considered in our simulation. In this case higher order corrections of the scaling law (4) always have to be taken into account.

5.4. 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 neigbourhood of the maximum likelihood estimator in the asymptotic case, the teacher would be a good starting condition for training. Figure 4(a) shows the complete learning curve of a 8-8-4 network comparing this initialization of the student to a random one. In the range of few examples both initializations yield the

[a]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.

same results. From this we conclude that no matter where we start in phase space, the dynamics of learning is always attracted to a symmetric solution. This means, the symmetric solution is a stable attractive region from which it is difficult to escape until a certain number of patterns is reached.
The detailed picture of the asymptotic range is given in figure 4(b). 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.4b). Note however, that the learning speed is increased by 20% using the teacher as initial starting point of learning.

6. Conclusion

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 symmetric phase which is stable and learning converges - almost independent of initialization - into a solution where the hidden units do not learn collectively. As the number of patterns is increased we can escape from the symmetric phase, the symmetry is broken and both the qualitative behaviour (fig.1a) seen in the Kullback-Leibler difference and the quantitative behaviour seen in the correlation parameter (fig.3) indicate a phase transition. From our results it seems important to reach the broken phase as fast as possible to have the network use all free parameters in the learning process. The clear bend of the learning curve is followed by a region of $1/t^2$ scaling when t is increased. Assymptotically we confirm the $m/2t$ behavior. More details on our study can be found in [10].
We would like to emphasize that we always find a faster scaling than $1/t$ between the symmetric phase (small t) and the asymptotic phase. For this reason model selection criteria which are usually based on a certain overall assumption on the smoothness of learning curves are likely to perform weakly, since they do neither capture the transition encountered nor the faster scaling observed.
Further investigation will be focussed on the measurement of scaling laws in a real practical application.

7. Acknowledgements

We would like to thank the participants of the NNSMP workshop for fruitful and stimulating discussions. K. -R. M. thanks for valuable discussions with 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.

8. References

1. Amari, S., Differential geometrical methods in statistics, Lecture Notes in Statistics No.28, Springer New York (1985)
2. Amari, S., Murata, N., Neural. Comp. 5, 140 (1993)
3. Barkai, E., Hansel, D., Sompolinsky, H., Phys.Rev.A, 45, 4146 (1992)
4. Baum, E.B., Haussler, D., Neural. Comp. 1, 151 (1989)
5. 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)
6. Haussler, D., Kearns, M., Seung, S., Tishby, N., Rigorous Learning Curve Bounds from Statistical Mechanics, preprint (1994)
7. Heskes, T.M., Kappen, B., Phys.Rev.A, 440, 2718 (1991)
8. Kang, K. Oh, J.-H., Kwon, C., Park, Y., Phys.Rev.E 48, 4805 (1993)
9. Kuhlmann, P., Müller, K.-R., J.Phys. A:Math.Gen., 27, 3759-3774 (1994)
10. Müller, K.-R., Murata, N., Finke, M., Amari, S., in preparation (1995)
11. Murata, N., Yoshizawa,S., Amari, S., NIPS 5, Morgan Kaufmann, San Mateo, 607 (1993)
12. Opper, M., Kinzel, W., Kleinz, J., Nehl, R., J.Phys. A:Math.Gen. 23, L581 (1990)
13. Opper, M., Kinzel, W., preprint
14. 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)
15. Schwarze, H., Hertz, J., Europhys. Lett. 21, 785 (1993)
16. Sompolinsky, H., Tishby, N., Seung, S., Phys.Rev.Lett. 65, 1683 (1990)
17. Seung, S., Sompolinsky, H., Tishby, N., Phys.Rev.A 45, 6056 (1992)
18. Watkin, T.L.H, Rau, A., Biehl, M., The Statistical Mechanics of Learning a Rule preprint (1992)
19. Akahira, M., Takeuchi, K., Asymptotic efficiency of statistical estimators: Concepts and higher order asymptotic efficiency, Springer New York (1981)

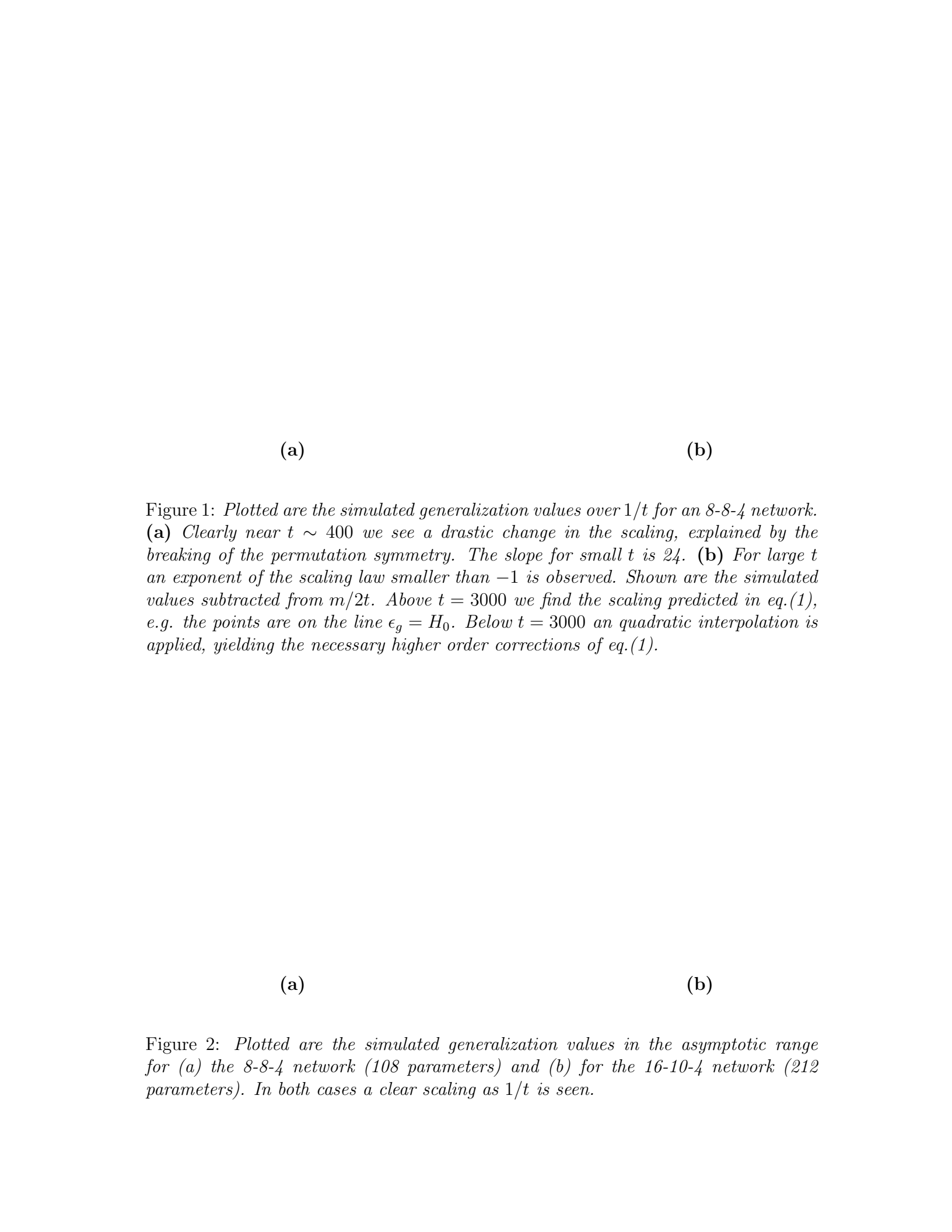

Figure 1: *Plotted are the simulated generalization values over* $1/t$ *for an 8-8-4 network.* **(a)** *Clearly near* $t \sim 400$ *we see a drastic change in the scaling, explained by the breaking of the permutation symmetry. The slope for small* t *is 24.* **(b)** *For large* t *an exponent of the scaling law smaller than* -1 *is observed. Shown are the simulated values subtracted from* $m/2t$*. Above* $t = 3000$ *we find the scaling predicted in eq.(1), e.g. the points are on the line* $\epsilon_g = H_0$*. Below* $t = 3000$ *an quadratic interpolation is applied, yielding the necessary higher order corrections of eq.(1).*

Figure 2: *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.*

(a)

(b)

Figure 3: *Plotted are the different overlap measures for an 8-8-4 network for (a) the start from the teacher w_T and (b) random initialization (see section 5).*

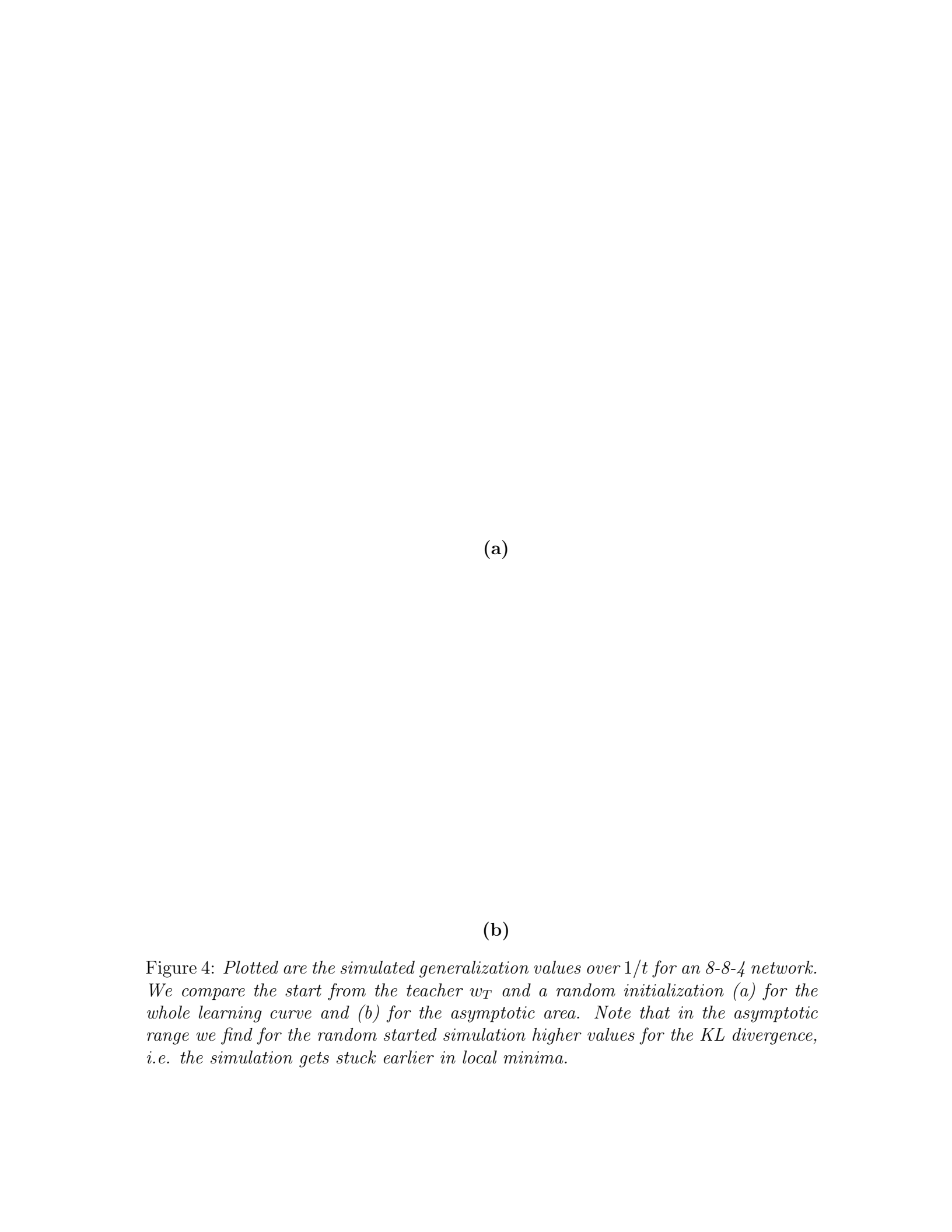

Figure 4: *Plotted are the simulated generalization values over* $1/t$ *for an 8-8-4 network. We compare the start from the teacher* w_T *and a random initialization (a) for the whole learning curve and (b) for the asymptotic area. Note that in the asymptotic range we find for the random started simulation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