Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

Noise Robustness in Multilayer Neural Networks

M. Copelli, R. Eichhorn, O. Kinouchi, M. Biehl
R. Simonetti, P. Riegler, and N. Caticha

Ref.: WUE-ITP-96-022
e-mail : biehl@physik.uni-wuerzburg.de
ftp : ftp.physik.uni-wuerzburg.de

Noise Robustness in Multilayer Neural Networks

M. Copelli[1], R. Eichhorn[2], O. Kinouchi[3], M. Biehl[2],
R. Simonetti[3], P. Riegler[2], and N. Caticha[3] *

[1] Limburgs Universitair Centrum
B-3590 Diepenbeek, Belgium
[2] Institut für Theoretische Physik
Universität Würzburg, Am Hubland
D-97074 Würzburg, Germany
[3] Instituto de Física, Universidade de São Paulo
CP 66318, 05389-970 São Paulo, SP, Brazil

Abstract

The training of multilayered neural networks in the presence of different types of noise is studied. We consider the learning of realizable rules in nonoverlapping architectures. Achieving optimal generalization depends on knowledge of the noise level, however its misestimation may lead to partial or complete loss of the generalization ability. We demonstrate this effect in the framework of online learning and present the results in terms of noise robustness phase diagrams. While for additive (weight) noise the robustness properties depend on the architecture and size of the networks, this is not so for multiplicative (output) noise. In this case we find a universal behaviour independent of the machine size for both the tree parity and committee machines.

The essential ingredients brought by the Statistical Mechanics [1, 2, 3] approach to the theory of learning are the consideration of large networks (the thermodynamic limit) and the possibility of performing averages over the disorder introduced by the random nature of the training data. Statistical Mechanics aims at describing typical behaviour and thus complements Computational Learning Theory [4] results.

Of the different learning scenarios that can be studied we will concentrate on online learning in networks with threshold units [5–10]. In this framework only the latest from a sequence of examples is used for updating the network parameters.

*The ordering of the authors has been determined randomly.

It reduces the storage needs and computational effort since examples are not presented repeatedly. This does not necessarily translate into poor performance since its simplicity has permitted to devise optimized learning algorithms which compete well with memory based offline schemes.

As these optimized algorithms depend on certain usually unknown parameters, they may be thought only as useful for the derivation of lower bounds to generalization errors. However, the constructive nature of the optimization procedure points out the relevant features necessary for efficient learning. Whether an arbitrarily constructed *ad hoc* algorithm performs satisfactorily, hinges on how well it reproduces the set of relevant characteristics.

The fact that the bounds are only saturated if the correct values of the parameters are used, suggests that the next step in algorithm design should concentrate on methods to estimate them efficiently. Any measure of this efficiency has to take into account how robust the algorithms are with respect to parameter precision.

Our aim in this letter is to investigate this question of robustness in the context of feedforward multilayer neural networks learning a rule in the presence of noise. We have studied two different architectures, the tree parity machine (TPM) and the tree committee machine (TCM) with K hidden units, and the types of noise that we consider are characterized by just a single parameter, the noise level. The influence of noisy data in supervised learning has been studied by several authors [2,9–16].

We present results in terms of "robustness diagrams", which show the regions in the space of true versus estimated noise level where different learning behaviour is obtained. These diagrams are rich in structure with transitions among perfect, imperfect, and impossible learning.

We consider the case where the teacher and student networks have the same architecture, in the absence of noise this corresponds to perfectly realizable rules. The machines with nonoverlapping receptive fields which we will study are composed of $K \sim \mathcal{O}(1)$ branches each with N/K input units and weight vectors $(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_K)$ for the teacher and $(\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_K)$ for the student. For each input vector $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_K)$ the teacher furnishes the correct classification label Σ_B which is given by

$$\Sigma_B(\boldsymbol{\xi}) = \begin{cases} \displaystyle\prod_{k=1}^{K} \sigma_B^k & \text{for the TPM} \\ \\ \operatorname{sign}\left(\displaystyle\sum_{k=1}^{K} \sigma_B^k\right) & \text{for the TCM} \end{cases} \tag{1}$$

where $\sigma_B^k = \operatorname{sign}(\mathbf{B}_k \cdot \boldsymbol{\xi}_k)$ denotes the internal representations in the teacher hidden units. The student input-output relations $\Sigma_J(\boldsymbol{\xi})$ are defined in the same fashion but with $\mathbf{J}_k$ substituting for $\mathbf{B}_k$.

When learning, the student uses at time μ the information contained in the input-output pair $(\boldsymbol{\xi}^\mu, \widetilde{\Sigma}_B^\mu)$, where, due to noise, the example label $\widetilde{\Sigma}_B^\mu$ may differ from the correct classification $\Sigma_B^\mu = \Sigma_B(\boldsymbol{\xi}^\mu)$. We consider one of two different stochastic mechanisms responsible for this corruption. We first study the effect of multiplicative (output) noise and then that of additive (weight) noise.

In this paper we consider the inputs to be vectors with i.i.d. random components with zero mean and unit variance. The success of learning is measured by the probability of disagreement between the student and teacher on an independently drawn such vector. The generalization error is defined as $e_g = \langle\Theta(-\Sigma_J\Sigma_B)\rangle_{\boldsymbol{\xi}}$, which compares the student output to the true rule, while the prediction error $e_p = \left\langle\Theta(-\Sigma_J\widetilde{\Sigma}_B)\right\rangle_{\boldsymbol{\xi},\widetilde{\Sigma}_B}$ is an average over the input distribution and the stochastic noise process.

Online learning proceeds by updating the student parameters at each time step μ according to

$$\mathbf{J}_k^{\mu+1} = \mathbf{J}_k^\mu + \frac{1}{N} F_k \;\; \boldsymbol{\xi}_k^\mu. \tag{2}$$

This is modified Hebbian learning [1] where the modulation functions F_k define the particular learning algorithm.

In the thermodynamic limit the generalization and prediction error become simple monotonic functions of the overlaps [17, 18]

$$\rho_k = \frac{\mathbf{J}_k \cdot \mathbf{B}_k}{J_k B_k} \tag{3}$$

where J_k and B_k are the lengths of the respective branch weight vectors. Without loss of generality we take $B_k = 1$ in the following. This suggests looking at the evolution of the overlaps ρ_k and the J_k in the course of learning, which can be obtained from eq. (2). As $N \to \infty$ these are self-averaging quantities with respect to the randomness in the training data. In terms of the continuous time $\alpha = \mu/N$ the dynamics is then described by a set of $2K$ first order ordinary differential equations [8]:

$$\frac{d\rho_k}{d\alpha} = \rho_k \left\langle \frac{F_k}{J_k}\left(\frac{y_k}{\rho_k} - x_k - \frac{F_k}{2KJ_k}\right)\right\rangle_{\boldsymbol{\xi},\widetilde{\Sigma}_B} , \tag{4}$$

$$\frac{dJ_k}{d\alpha} = J_k \left\langle\left(\frac{(F_k)^2}{2KJ_k^2} + \frac{F_k x_k}{J_k}\right)\right\rangle_{\boldsymbol{\xi},\widetilde{\Sigma}_B} \tag{5}$$

where the normalized internal fields are $x_k = \mathbf{J}_k \cdot \boldsymbol{\xi}_k / J_k$ and $y_k = \mathbf{B}_k \cdot \boldsymbol{\xi}_k$. In the following we assume symmetric initial conditions $\rho_k = \rho$ and $J_k = J$ for all k, this symmetry clearly will be preserved by the dynamics. Due to the above mentioned monotonicity, the optimal generalization ability is achieved for the modulation function $F_k^\star$ which maximizes the rate $d\rho_k/d\alpha$, which is given by [8]

$$F_k^\star = KJ_k \left\langle\left(\frac{y_k}{\rho_k} - x_k\right)\right\rangle_{y_k|\widetilde{\Sigma}_B,\{x_i\}} . \tag{6}$$

Here, the average is to be performed over the conditional probability of the unknown teacher internal field y_k in branch k, given the set of all $\{x_i\}_{i=1,\dots K}$, the noisy training label $\widetilde{\Sigma}_B$, and the noise level. Note that the resulting learning prescription is nonlocal, in the sense that information about all other branches of the student is used when updating one of the $\mathbf{J}_k$ according to eq. (2). Whereas this formal prescription is machine independent, it actually incorporates the specific details of the architecture in the way that it uses the available information.

The use of the optimal weight function requires explicit knowledge of both the actual noise level and the teacher-student overlap ρ, which in general are not immediately accessible. To address these issues, adaptive algorithms for online estimation of these quantities have been proposed in [9, 19].

We first consider what has been termed output noise [12]. In this case, each example is independently subject to a random inversion of the training label according to the conditional probability

$$P\left(\widetilde{\Sigma}_B^\mu \mid \Sigma_B^\mu\right) = \lambda_o \delta(\widetilde{\Sigma}_B^\mu, -\Sigma_B^\mu) + (1-\lambda_o)\delta(\widetilde{\Sigma}_B^\mu, \Sigma_B^\mu) \tag{7}$$

The maximal inversion rate $\lambda_o = 1/2$ would correspond to completely uncorrelated labels which contain no information about the rule.

In this model, the dependence on the random inputs in eq. (2) is only through the student and teacher internal fields, which by the central limit theorem are Gaussian correlated random variables with zero means, $\langle x_k^2 \rangle = \langle y_k^2 \rangle = 1$ and $\langle x_k y_k \rangle = \rho_k$. In addition, the average over the data corruption has to be performed explicitely according to eq. (7).

Although the optimized modulation functions for the TPM and TCM are different (for a comparison see [16, 20]) they share certain common behaviours. A universal characteristic we want to stress is that the resulting optimal algorithms for the TCM and TPM, for any K, present the same asymptotical behaviour, independently of the architecture and hidden layer size. For details see [8, 9, 10, 16]. For sufficiently large α the generalization errors decrease like

$$e_g(\alpha) = \frac{2}{I(\lambda_o)} \frac{1}{\alpha} \tag{8}$$

where $I(\lambda_o) = (1-2\lambda_o)^2 \int Dt e^{-t^2}/\hat{H}(t)$ and $Dt = (2\pi)^{-1/2}\exp(-t^2/2)dt$, $H(t) = \int_t^\infty Dx$ and $\hat{H}(t) = \lambda_o + (1-2\lambda_o)H(t)$.

In order to approximate the optimal behaviour obtained with $F_k^\star(K, \lambda_o)$ given by eq. (6), a constant estimate Λ_o could be used in its stead, but the performance of the training algorithm which uses the modulation function $F_k = F_k^\star(K, \Lambda_o)$ depends critically on the quality of this estimation.

Three regimes, as shown in figure 1 a), can be identified by analysis of the fixed point properties of $d\rho/d\alpha$ as a function of ρ. An overestimation of the unknown noise level ($\Lambda_o > \lambda_o$) still enables the system to achieve perfect generalization in the limit

$\alpha \to \infty$. The generalization error decays as $C(\lambda_o, \Lambda_o)/\alpha$ where the coefficient C for given λ_o attains its maximal value $2/I(\lambda_o)$ at $\Lambda_o = \lambda_o$. This phase of asymptotically perfect generalization extends into the region of underestimated noise levels ($\Lambda_o < \lambda_o$) and is bounded by the solid line in figure 1 a) which is numerically obtained from the condition $1/C(\lambda_o, \Lambda_o) = 0$. We call this phase the *robust learning regime*; outside this region, i.e., for worse underestimation of the noise level, the ability to generalize perfectly is lost. The boundary shows the continuous transition where the fixed point $\rho = 1$ of eq. (4) becomes repulsive and signals the appearance of a new attractive fixed point at an intermediate value of ρ and $e_g > 0$. This phase of *imperfect learning* extends all the way to the origin ($\lambda_o = \Lambda_o = 0$), thus for any nonzero noise level total confidence ($\Lambda_o = 0$) of the student on the teacher information will impair its ability to extract the rule completely.

It can be proved that imperfect learning will occur unless Λ_o is less than $2\lambda_o - 1/2$ where $d\rho/d\alpha|_{\rho=0}$ becomes negative. The corresponding dashed line in figure 1 a) marks the onset of a third phase characterized by total lack of generalization corresponding to $\rho = 0$ being an attractive fixed point.

It is rather remarkable that the phase boundaries are independent of the machine architecture. Thus, the robustness properties of the TCM and the TPM are the same and independent of K in the presence of multiplicative noise, details of the proof will be presented in a forthcoming publication [21]. Of course, for $K = 1$ both architectures reduce to the simple single layer perceptron for which the same diagram holds.

However, this universality does not generally carry over to other types of noise in the training set. To show this, we consider in the following the corruption of examples by *weight noise* in the parity machine. The same behaviour would arise in the presence of input noise or of additive noise in the internal fields of the teacher.

The effect of weight noise is that for each training example μ the true teacher vectors $\{\mathbf{B}_k\}$ are replaced by normalized random vectors $\{\widetilde{\mathbf{B}}_k^\mu\}$ uniformly distributed on the cones defined by $\mathbf{B}_k \cdot \widetilde{\mathbf{B}}_k^\mu = \omega$. The probability for a given example's label being inverted by this process is strongly dependent on its actual teacher fields $\{y_k\}$. For given ω the inversion rate λ_w, the expected fraction of flipped training labels averaged over the input distribution, is given by

$$\lambda_w = \frac{1}{2}\left[1 - \left(1 - \frac{2}{\pi}\arccos\omega\right)^K\right], \tag{9}$$

as can be seen from the equivalence of this expression to the generalization error of a student with branch overlaps $\rho_k = \omega$ [16]. Note that the same weight noise level ω results in different inversion rates for different network sizes K.

Even if the inversion rate is perfectly known, the asymptotic decay of the generalization error is significantly slower than in the previous case: $e_g \propto 1/\sqrt{\alpha}$. This has been found for the simple perceptron previously [9]. Furthermore it is not universal, because the coefficient is K-dependent for fixed λ_w or fixed ω.

In figure 1 b) we present the robustness diagram of the TPM in the presence of weight noise in terms of the true inversion λ_w and a constant estimate Λ_w. The latter corresponds to the inversion rate that would result if the weight noise level was Ω in analogy to eq. (9).

Qualitatively, the diagram resembles the one for output noise in that the same three phases can be found. However, the phase boundaries are K-dependent as shown in figure 1 b) for $K = 1, 2, 3$. The structure of the problem in the presence of output noise permits to factor out a common kernel, independent of K and architecture, which is responsible for the fixed point properties in the vicinity of $\rho = 1$. The problem is not factorizable in the same manner in the case of weight noise. Note also, that the boundaries between the regions of perfect and imperfect generalization here approach the origin $\lambda_w = \Lambda_w = 0$ with a nonzero slope, in contrast to the case of output noise. On the other hand, in the large noise limit the robustness diagrams coincide for both types of noise.

The transition line between the region of imperfect learning and the phase with total lack of generalization can be calculated analytically: $\Omega = 2^{1/K}\omega$. This result is K-dependent also in terms of the inversion rates as displayed in figure 1 b) for $K = 1, 2, 3$.

Loss of generalization due to noise has been found in other models [13, 22]. Efficient *ad hoc* algorithms can be interpreted as more or less crude approximations of the optimal scheme for a certain noise level. Our analysis explains why there will exist an actual noise level for which a non–adaptive *ad hoc* algorithm will fail. This suggests the use of adaptive algorithms which include the determination of unknown parameters in the learning process in order to avoid the deterioration due to their misestimation.

In a forthcoming publication [21] we will present, besides the details of this work, analogous studies for nonoptimal *ad hoc* algorithms and different network architectures as well as further work into the online determination of unavailable parameters. This step is necessary in order to devise practical learning schemes which are guaranteed to work under more general circumstances.

Note that the observed boundaries distinguish different behaviours of a dynamical system. Nevertheless, the same program can be carried out in the framework of offline learning where corresponding boundaries will truly separate equilibrium phases.

Acknowledgment: This work has been supported by the Deutsche Forschungsgemeinschaft (DFG), the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), the Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the Nationaal Fonds voor Wetenschappelijk Onderzoek (NFWO).

References

[1] J.A. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison–Wesley (Redwood City, CA), 1991.

[2] H.S. Seung, H. Sompolinsky, and N. Tishby, Phys.Rev.A **45**, 6056 (1992).

[3] T.L.H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys.**65**, 499 (1993).

[4] M. Kearns and U. Vazirani, *An Introduction to Computational Learning Theory*, MIT Press (Cambridge, MA), 1994

[5] W. Kinzel and P. Ruján, Europhys.Lett. **13**, 473 (1990).

[6] O. Kinouchi and N. Caticha, J.Phys. **A25**, 6243 (1992).

[7] M. Biehl and H. Schwarze, J.Phys. **A26**, 2651 (1993).

[8] M. Copelli and N. Caticha, J.Phys. **A28**, 1615 (1995).

[9] M. Biehl, P. Riegler, and M. Stechert, Phys.Rev.E **52**, R4624 (1995).

[10] M. Copelli, O. Kinouchi, and N. Caticha, Phys.Rev.E **53**, 6341 (1996).

[11] G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses*, edited by W.K. Theumann and R. Köberle, (World Scientific, Singapore, 1990).

[12] M. Opper, and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991), and in *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, edited by L.G. Valiant and M.K. Warmuth, (Morgan Kaufmann, San Mateo, CA, 1991).

[13] Y. Kabashima, J.Phys. **A27**, 1917 (1994).

[14] M. Opper and W. Kinzel, in *Models of Neural Networks III*, series editors E. Domany, J.L. van Hemmen, and K. Schulten (Springer, Berlin, 1996).

[15] J.W. Kim and H. Sompolinsky, Phys.Rev.Lett. **76**, 3021 (1996).

[16] R. Simonetti and N. Caticha, J.Phys. **A29**, 4859 (1996).

[17] G. Mato and N. Parga, J.Phys. **A25**, 5047 (1992).

[18] M. Opper, Phys.Rev.Lett. **72**, 2113 (1994).

[19] O. Kinouchi and N. Caticha, J.Phys. **A26**, 6161 (1993).

[20] M. Copelli and N. Caticha, in preparation

[21] in preparation

[22] P. Sollich, PhD Thesis, University of Edinburgh, 1995

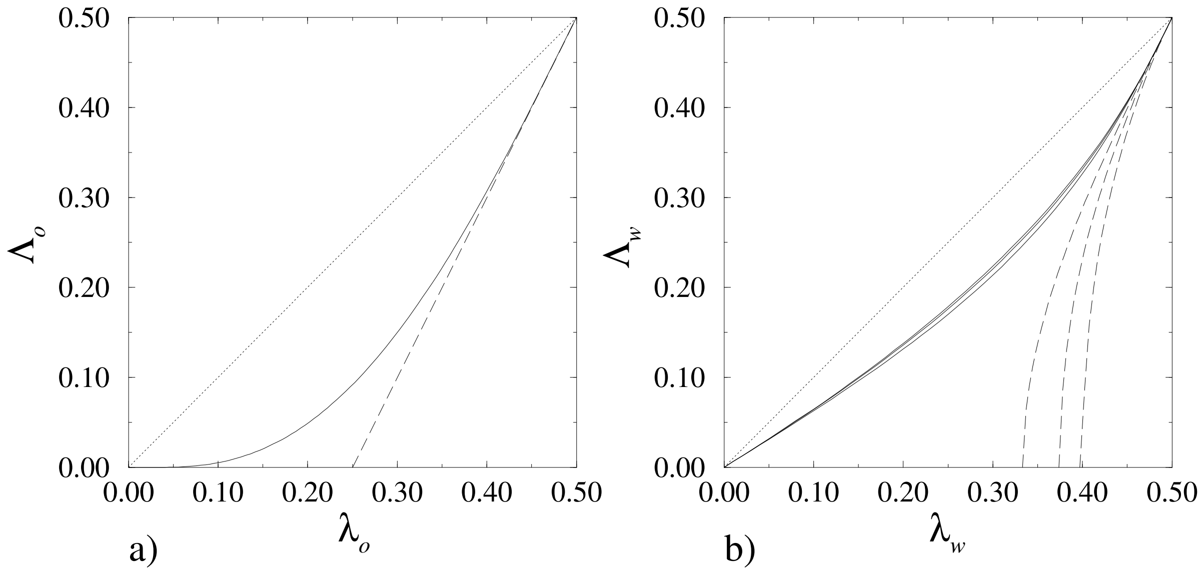

Figure 1: Robustness diagrams in terms of estimated (Λ) vs. true (λ) inversion rates for the TPM and the TCM in the presence of output noise (a) and for weight noise in the TPM(b). Performance is optimal on the diagonal. Solid lines separate the robust learning regime from the region of imperfect learning. Dashed lines correspond to the boundaries between the latter and the regions of total lack of generalization. In figure a) the diagram is the same for all K (TCM and TPM), in figure b) curves are shown for a TPM with $K = 1, 2, 3$ (solid lines: from bottom to top, dashed lines: from left to right).