ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 181-186

Noise robustness in the perceptron

Mauro Copelli *

Limburgs Universitair Centrum
B-3590 Diepenbeek, Belgium

Abstract. Within the framework of online supervised learning, algorithms which lead to optimal generalization ability in the boolean perceptron are obtained. Restrictions on the available information during learning lead to different optimal algorithms. Knowledge of the noise level is required for optimal performance to be achieved, and misestimation of that quantity may lead to partial or complete loss of the generalization ability. Results are shown in terms of robustness phase diagrams.

1. Introduction

Online supervised learning in feedforward boolean neural networks [6, 4, 5, 1, 7] has been lately studied by physicists. The statistical mechanics approach to these problems has shown to be very fruitful, since analytical solutions can be found in some cases, and optimal algorithms can eventually be found on the grounds of a simple variational argument.

This paper will concentrate on the problem of online supervised learning in the perceptron (even though some of the results are found to be exactly the same for some multilayer machines – see [2] for details). The scenario is that of a student boolean perceptron learning from a teacher with the same architecture. The examples are randomly generated, and outputs are corrupted by output (multiplicative) noise.

A short introduction to online learning is given in section 2. In section 3. the variational optimization scheme is sketched. For each learning condition (to be defined below) there is a corresponding optimal algorithm. Four different learning conditions are studied here, and the performances of the corresponding algorithms are given in section 4.

The robustness of those algorithms with respect to a misestimation of the noise level is studied in section 5. A given algorithm can lead to different dynamical behaviours, thus giving rise to a *noise robustness phase diagram.*

Acknowledgements: I would like to thank the FWO, Flemish Government and the IUAP, Prime Minister's office for financial support.

*email address: copelli@luc.ac.be

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 181-186

2. Online learning

The aim of the learning process is to change the student vector $\mathbf{J}$ in such a way that it approaches the teacher vector $\mathbf{B}$. Both vectors have N real components, and we can set $\mathbf{B} \cdot \mathbf{B} = 1$ without loss of generality. Input vectors $\mathbf{S}$ are independently drawn from a distribution which satisfies $\langle S_i \rangle = 0$ and $\langle S_i S_j \rangle = \delta(i,j)$, where $\delta(i,j)$ is the Kronecker delta. Upon presentation of a particular input vector $\mathbf{S}^\mu$, the student perceptron gives an output $\sigma_J^\mu = \mathrm{sign}(\mathrm{h}^\mu)$, where $h^\mu = \mathbf{J} \cdot \mathbf{S}^\mu / J$, $J = \sqrt{\mathbf{J} \cdot \mathbf{J}}$. The teacher final output is denoted by ξ^μ, which is a noisy version of $\sigma_B^\mu = \mathrm{sign}(\mathrm{b}^\mu)$, $b^\mu = \mathbf{B} \cdot \mathbf{S}^\mu$. The *noise level* ϵ is defined as twice the probability of flipping σ_B: $P(\xi \mid \sigma_B) = (\epsilon/2)\delta(\xi, -\sigma_B) + (1 - \epsilon/2)\delta(\xi, \sigma_B)$.

The learning set is made up of αN noisy examples $\{(\mathbf{S}^\mu, \xi^\mu)\}$, $\mu = 1, \ldots, \alpha N$, which are presented sequentially. Each new example induces a *single* change in the student vector,

$$\mathbf{J}(\mu + 1) = \mathbf{J}(\mu) + \frac{1}{N} F \mathbf{S}^\mu \,, \tag{1}$$

being *discarded* after that (note that in this case the index μ can be used as a discrete time counter). F is called the *modulation function*, and it defines the algorithm. Its explicit form is to be determined by the optimization process to be described in the next section.

The error measure to be studied here is the *generalization error* $e_g = \pi^{-1} \arccos(\rho)$. e_g is the probability that $\sigma_J \neq \sigma_B$ upon presentation of a new random input vector. The overlap $\rho = \mathbf{J} \cdot \mathbf{B}/J$ is thus the relevant order parameter for this problem. Its evolution during the learning process can be described by a differential equation when the limit $N \to \infty$ is taken. From eq. 1, keeping terms of $\mathcal{O}(N^{-1})$, one obtains

$$\frac{d\rho}{d\alpha} = \rho \left\langle \frac{F}{J} \left(\frac{b}{\rho} - h - \frac{F}{2J} \right) \right\rangle \,, \tag{2}$$

where the brackets denote average over the randomness of the examples. Since the right hand side of the above equation depends only on the fields h and b, the Central Limit Theorem can be used, yielding the following simplification: $\langle (\cdots) \rangle = \sum_\xi \int P(\xi \mid b) P_0(h,b)\, dh\, db (\cdots)$, where P_0 is a Gaussian with $\langle h \rangle = \langle b \rangle = 0$, $\langle h^2 \rangle = \langle b^2 \rangle = 1$ and $\langle hb \rangle = \rho$.

3. Variational optimization

Eq. 2 is valid for any modulation function F. Examples of algorithms usually found in the literature include the Hebbian rule ($F = \xi$), the Perceptron learning rule ($F = \xi\, \Theta(-\sigma_J \xi)$) or the Adatrom algorithm ($F = -h\xi\, \Theta(-h\xi)$), where Θ is the Heaviside function. The aim of the optimization process is to *obtain* an algorithm which leads to the smallest generalization error for a given number of examples, instead of proposing an *ad hoc* prescription.

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 181-186

The *learning condition* determines which stochastic quantities are available to the student perceptron during learning, and it should be specified *a priori*. Two sets of random variables can be defined: the set $\mathcal{X}$, which contains the unknown variables, and the set $\mathcal{Y}$, which contains the known ones. These sets are defined in such a way that they contain all the information concerning the randomness of the learning process. For instance, the learning conditions for the three mentioned algorithms can be described by: $\mathcal{X} = \{b, h\}, \mathcal{Y} = \{\xi\}$ (Hebb), $\mathcal{X} = \{b, |h|\}, \mathcal{Y} = \{\xi, \sigma_J\}$ (Perceptron) or $\mathcal{X} = \{b\}, \mathcal{Y} = \{\xi, h\}$ (Adatrom). With the learning condition defined prior to the optimization procedure, it is clear that *the resulting algorithm should depend only on available variables*. That is, $F = F(\mathcal{Y})$. Rewriting the average on the examples as $\langle(\cdots)\rangle = \int P(\mathcal{X} \mid \mathcal{Y})\, P(\mathcal{Y})\, d\mathcal{X}\, d\mathcal{Y}\, (\cdots)$, integration on $\mathcal{X}$ can be immediately performed on eq. 2.

Kinouchi and Caticha's optimization scheme [5] relies on two main points. Note that a) e_g is a monotonically decreasing function of ρ and b) the dependence of $d\rho/d\alpha$ on F is functional. The optimal algorithm F^* is then given by the condition

$$\frac{\delta}{\delta F}\left(\frac{d\rho}{d\alpha}[F]\right)\Big|_{F=F^*} = 0\,, \tag{3}$$

where $\delta/\delta F$ is a functional derivative. This simple calculation yields

$$F^*(\mathcal{Y}) = J\left(\frac{1}{\rho}\langle b\rangle_{\mathcal{X}|\mathcal{Y}} - \langle h\rangle_{\mathcal{X}|\mathcal{Y}}\right)\,. \tag{4}$$

4. Performance

Eq. 4 holds for any $\mathcal{X}$ and $\mathcal{Y}$. I now proceed to show the explicit results for four different learning conditions. The expressions of the corresponding optimal algorithms and their asymptotic performances (behaviour of e_g for $\alpha \to \infty$) are given below:

- $\mathcal{X} = \{b\}$, $\mathcal{Y} = \{\xi, h\}$: This is the learning condition where all the available information is used. The corresponding F_{opt} is thus the *best possible* algorithm [1, 3] for online learning in this scenario:

$$F_{opt}(\xi, h; \epsilon, \rho) = J\lambda\xi\frac{(1-\epsilon)}{\sqrt{2\pi}}\frac{e^{-h^2/2\lambda^2}}{\epsilon/2 + (1-\epsilon)H(-\xi h/\lambda)}\,, \tag{5}$$

where $\lambda = \rho^{-1}\sqrt{1-\rho^2}$, $H(x) = \int_x^\infty Dt$ and $Dt = dt(2\pi)^{-1/2}e^{-t^2/2}$. For this algorithm, $e_g \simeq \frac{2}{I(\epsilon)}\alpha^{-1}$, where $I(\epsilon) = (1-\epsilon)^2\int \frac{Dx\, e^{-x^2/2}}{\epsilon/2+(1-\epsilon)H(x)}$.

- $\mathcal{X} = \{b, \sigma_J\}$, $\mathcal{Y} = \{\xi, |h|\}$: In this learning condition, only the absolute value of the student field h is known, but not its sign. The optimal algorithm in this case corresponds to a more general version of the Symmetric Weight algorithm [4],

$$F_{sw}(\xi, |h|; \epsilon, \rho) = (1-\epsilon)\sqrt{\frac{2}{\pi}}J\lambda\xi e^{-|h|^2/2\lambda^2}\,, \tag{6}$$

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 181-186

and its asymptotic performance is given by $e_g \simeq \sqrt{2}(1-\epsilon)^{-2}\alpha^{-1}$.

• $\mathcal{X} = \{b, |h|\}$, $\mathcal{Y} = \{\xi, \sigma_J\}$: In this case only the sign of the student field is known, but not its absolute value. The corresponding Step algorithm resembles the Perceptron algorithm in its step-like shape. Yet it has a non-zero value for positive $\sigma_J\xi$, whose relative importance increases with the noise level ϵ and decreases with increasing ρ:

$$F_{step}(\xi, \sigma_J; \epsilon, \rho) = J\lambda^2\rho\xi\frac{(1-\epsilon)}{\sqrt{2\pi}}\left[\frac{\epsilon}{2} + \frac{(1-\epsilon)}{\pi}\arccos(-\rho\xi\sigma_J)\right]^{-1} . \quad (7)$$

There is an interesting transition in the performance of this algorithm. For $\epsilon = 0$ (i.e. the noiseless case), the asymptotic result is $e_g \simeq (4/\pi)\alpha^{-1}$. However, for $\epsilon \neq 0$ the decay is qualitatively different: $e_g \simeq |1-\epsilon|^{-1}\sqrt{\epsilon(2-\epsilon)/(2\pi)}\alpha^{-1/2}$.

• $\mathcal{X} = \{b, h\}$, $\mathcal{Y} = \{\xi\}$: This is the learning condition with least available information. The corresponding algorithm,

$$F_H(\xi; \epsilon, \rho) = J\xi\sqrt{\frac{2}{\pi}}\frac{(1-\rho^2)}{\rho}(1-\epsilon) , \quad (8)$$

was shown in [3] to correspond exactly to the Hebbian algorithm, despite the apparently time-dependent prefactor. Its performance [1, 3] is given by $e_g \simeq (2\pi)^{-1/2}|1-\epsilon|^{-1}\alpha^{-1/2}$.

5. Phase diagrams

The above algorithms are optimal for their specific learning conditions. They share some common features among themselves, like the explicit dependence on the noise level ϵ and the student performance (as parametrized by the overlap ρ). They lead to optimal generalization ability *only* if those quantities can be used during learning. The question to be now addressed is the following: how do the algorithms perform when the noise level ϵ is unknown?

Let η be a fixed estimate of the noise level in the system. Assuming that ρ is known, the student learns with a modulation function $F^*(\mathcal{Y}; \eta, \rho)$, which is optimal only when $\eta = \epsilon$. However, for $\eta \neq \epsilon$, sub-optimal performance is attained. For each algorithm, different dynamical behaviours can be seen, depending on the values of η and ϵ. Whenever the noise level is superestimated ($\eta > \epsilon$), the student is able to asymptotically learn the rule ($\rho \to 1$ when $\alpha \to \infty$). This is the *robust learning regime*, which can be reached even for some $\eta \leq \epsilon$. For given η, there is however a critical value $\epsilon = \epsilon_c$ above which perfect learning is no longer reached. The system converges to a new fixed point $\rho_0 < 1$. As can be seen in Figure 1, the critical line $\epsilon_c(\eta)$ is different for the four algorithms under study. A more detailed analysis of each phase diagram is given below.

Figure 1(a) depicts the noise robustness phase diagram for F_{opt}. In the robust learning regime (grey), the asymptotic behaviour is governed by the

ESANN'1997 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 16-17-18 April 1997, D-Facto public., ISBN 2-9600049-7-3, pp. 181-186

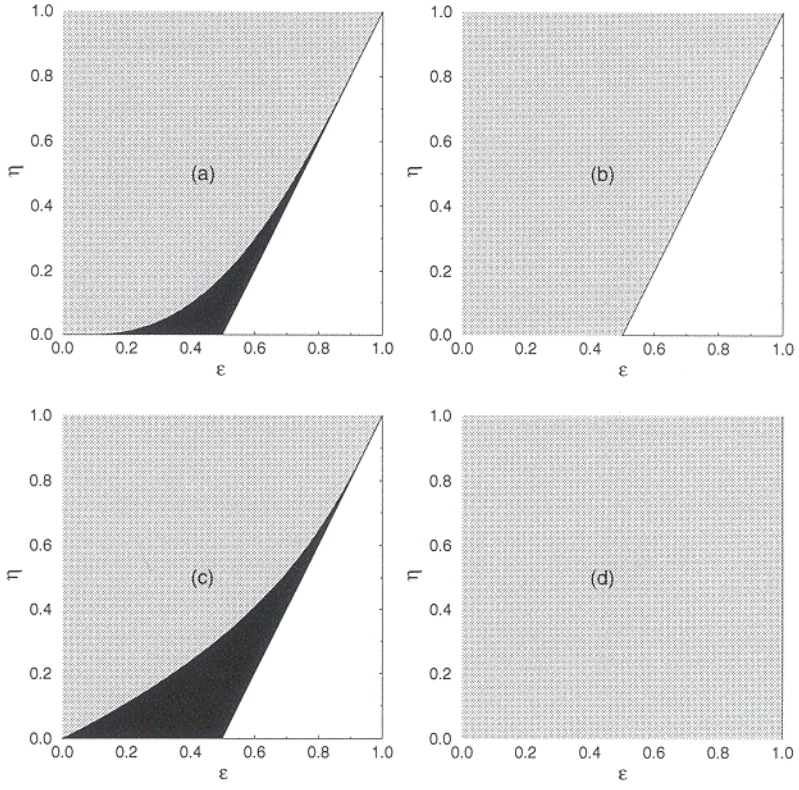

Figure 1: Noise robustness phase diagrams for (a) F_{opt}; (b) F_{sw}; (c) F_{step} and (d) F'_H (see text for details). When $\alpha \to \infty$, the overlap ρ tends either to 1 (grey), 0 (white) or ρ_0, with $0 < \rho_0 < 1$ (black).

same exponent as for the optimal performance described in section 4., $e_g \simeq C(\epsilon, \eta)\alpha^{-1}$. The critical line $\epsilon_c(\eta)$ is numerically obtained by imposing the condition $1/C(\epsilon_c, \eta) = 0$. For $\epsilon > \epsilon_c$ (black) the system converges to $e_g \neq 0$ (*imperfect learning*), still with the same exponent. However, a worse underestimation of the noise level may lead to *total loss of generalization*: $\rho_0 = 0$ ($e_g = 1/2$) becomes an attractive fixed point of the dynamics for $\epsilon > (1+\eta)/2$ (white).

The diagram for F_{sw} (Figure 1(b)) does not present the imperfect learning phase. The region where perfect learning occurs extends until the line $\epsilon = (1+\eta)/2$, which again is the border of the region with total loss of generalization. The asymptotic value of ρ drops discontinuously from 1 to 0. Note that even though F_{sw} has a worse performance than F_{opt} under ideal conditions ($\eta = \epsilon$), its robust phase is bigger in the $\epsilon \times \eta$ plane. Like in the previous case, the asymptotic convergence in this region is governed by the same exponent, $e_g \sim$

α^{-1}.

The apparent trade-off between optimal performance and robustness, which was seen in the previous two cases, does not hold for F_{step}. Apart from the noiseless case, F_{step} performs qualitatively worse than F_{sw} for $\epsilon = \eta$. Yet its perfect learning phase is smaller than that of F_{sw}. As can be seen in Figure 1(c), even F_{opt} is more robust than F_{step}. The critical line in this case is simply given by $2\epsilon_c = 4\eta - 3\eta^2 + \eta^3$, while the region with $\rho_0 = 0$ remains bounded by $\epsilon = (1+\eta)/2$. Both perfect and imperfect learning are governed by the same exponent, $\rho - \rho_0 \sim \alpha^{-1}$.

As shown in [3], the prefactor in eq. 8 is just a constant when the real noise level ϵ is used. This suggests the study of a slightly modified version of F_H, namely $F'_H = (1-\epsilon)\xi$. This is the simple way of writing Hebb's rule, which leads to Figure 1(d). Written in this way, the algorithm leads to perfect learning in the whole plane. However, when the ρ dependence of F_H is mantained (as originally written in eq. 8), the corresponding phase diagram is given by Figure 1(b). In this case, F_H and F_{sw} differ only in the exponent governing the asymptotic convergence. For F_H, $e_g \sim \alpha^{-1/2}$.

In conclusion, there is a price paid for the optimality of the algorithms obtained via variational optimization. They depend critically on parameters that are not always readily available. The results presented here strongly motivate the development of methods for estimating relevant parameters during learning, as done in [1].

References

[1] M. Biehl, P. Riegler, and M. Stechert. Learning from noisy data: an exactly solvable model. *Phys. Rev. E*, 52:R4624, 1995.

[2] M. Copelli, R. Eichorn, O. Kinouchi, M. Biehl, R. Simonetti, P. Riegler, and N. Caticha. Noise robustness in multilayer neural networks. (to appear in Europhys. Lett.).

[3] M. Copelli, O. Kinouchi, and N. Caticha. Equivalence between on-line learning in noisy perceptrons and tree committee machines. *Phys. Rev. E*, 53:6341, 1996.

[4] O. Kinouchi and N. Caticha. Biased learning in boolean perceptrons. *Physica A*, 185:411, 1992.

[5] O. Kinouchi and N. Caticha. Optimal generalization in perceptrons. *J. Phys. A*, 25:6243, 1992.

[6] W. Kinzel and P. Ruján. Improving a network generalization ability by selecting examples. *Europhys. Lett.*, 13:473, 1990.

[7] C. Van den Broeck and P. Reimann. Unsupervised learning by examples: on-line versus off-line. *Phys. Rev. Lett.*, 76:2188, 1996.