UNIVERSIDADE DE SÃO PAULO

INSTITUTO DE FÍSICA
CAIXA POSTAL 66318
05389-970 SÃO PAULO - SP
BRASIL

PUBLICAÇÕES

IFUSP/P-1156

A LEARNING ALGORITHM WHICH GIVES THE BAYES GENERALIZATION LIMIT FOR PERCEPTRONS

Osame Kinouchi and **Nestor Caticha**
Instituto de Física, Universidade de São Paulo

Maio/1995

A Learning Algorithm which gives the Bayes Generalization Limit for Perceptrons

Osame Kinouchi* and **Nestor Caticha†**
Instituto de Física, Universidade de São Paulo
Caixa Postal 66318
CEP 05389-970 São Paulo, SP, Brazil

Abstract

A variational approach to learning a linearly separable rule by a single layer perceptron leads to a learning algorithm with the Bayes generalization ability calculated by Opper and Haussler. This is done by finding, through the Gardner-Derrida replica method, the student-teacher overlap R as a functional of the algorithm cost function and maximizing this functional. The resulting modulation weight function is closely related to the optimal function for on-line learning.

Key words: *neural networks, generalization, perceptron, learning algorithms, Bayes.*

PACS. #: 02.50, 05.90, 75.10 Hk, 87.10.e+10, 87.70.+c.

*E-mail: osame@if.usp.br
†E-mail: nestor@if.usp.br

We consider the problem of generalization by a single layer perceptron undergoing supervised learning from examples generated by a teacher network with the same architecture. There is a vast literature on this subject (*e.g.* [1] - [4]), which ranges from the numerical simulation and analytical calculation of the generalization ability of different algorithms, for either iterated schemes or single on-line presentation of the examples, to the determination of the best possible mean performance by Opper and Haussler [5]. Their determination of the best possible generalization ability is based on Bayes theorem and is thus sometimes referred to as the Bayes curve or the Bayes algorithm. This last name is however a bit misleading since they did not obtain a prescription of how the student's weights ought to be determined from the information contained in the input-output examples learning set.

The proof that there is at least a perceptron which actually gives the Bayes performance was given by Watkin [6]. The determination of the weights, as proposed by Watkin, amounts to calculating the center of mass of the version space, *i.e.* the set of students weight vectors consistent with all the examples present in the learning set. At least from a computational point of view, Watkin's algorithm is not, however, a realistic way of obtaining such vector, since it implies in randomly determining l independent members of the version space and then taking $l \to \infty$. The generality of such approach however makes it of special interest in dealing with more sophisticated networks where Bayes-like bounds might not be readily available.

In this letter we present a *gradient descent* algorithm which gives *exactly* the Opper-Haussler-Bayes (OHB) limit. The main idea in obtaining optimal generalization algorithms is to treat the learning problem as a variational one. This has been previously done in [7]-[13] for the on-line version of learning, where the examples are used only once and are thereafter discarded. We here extend the use of the variational approach to the off-line scenario of learning.

The generalization ability is calculated in general, for any algorithm with a non degenerate ground state, using the standard replica approach of Gardner's space of interactions. The optimization of such ability determines the algorithm, and the optimized ability is exactly the OHB curve. In performing such general calculation we rely heavily on the streamlined method of Bouten, Schietse and Van den Broeck (BSB) [14]. The optimal off-line learning algorithm is found to be the same as the optimal on-line algorithm.

Learning can in general be thought of as a gradient descent process. The derivative, with respect to the stability of an example, of the training energy function is related to the weight function that modulates the Hebbian correction of the synaptic couplings. A subtle point that arises from our result is that, in accordance to the interpretation of the learning process in the light of cavity methods [14, 15], the optimization procedure furnishes the modulation function in terms of the prior to learning quantities, whereas the energy function itself depends on the post learning stability. This is indeed necessary for a practical algorithm, since it would be of no use if the post-learning quantities were needed in order to proceed with the learning dynamics.

Using a numerical variational method, but allowing only for variations within a very restricted class of functions which push with varying degrees of strength away from the border of the version space, BSB have found a remarkably simple algorithm with an asymptotical behavior very close to the OHB curve. Being, as usual $\alpha = P/N$, the ratio of P, the number of patterns in the learning set to N, the number of inputs to the perceptron, the generalization error decays asymptotically, in the limit of infinite N, as C/α, where the algorithm dependent constant C is $C_{OHB} = 0.442\ldots$ and $C_{BSB} = 0.443\ldots$, so at least in the asymptotic limit a very efficient practical method has been obtained. For small α the BSB algorithm is not as good. The question of the existence of a gradient descent algorithm that leads to the optimal performance for every α remains and we now deal with it.

Let $\mathbf{B}$ and $\mathbf{J}$ be respectively the unknown coupling vector of the teacher perceptron and that of the student. Let $\mathcal{L} = \{\mathbf{S}^\mu, \sigma_B^\mu\}_{\mu=1,\ldots,P}$ be the training set generated by the action of the teacher $\mathbf{B}$ on the input vectors $\mathbf{S}^\mu$:

$$\sigma_B^\mu = \mathrm{sign}(\mathbf{B}\cdot\mathbf{S}^\mu)\,. \tag{1}$$

Let $R = \mathbf{J}.\mathbf{B}/\|\mathbf{B}\|\|\mathbf{J}\|$. The generalization error is known to be a monotonic function of R:

$$e_G = \frac{1}{\pi}\arccos R\,. \tag{2}$$

Opper and Haussler have shown that the largest typical value of R that can be obtained from αN examples is $R_B(\alpha)$, which satisfies

$$\frac{R_B^2(\alpha)}{\sqrt{1-R_B^2(\alpha)}} = \frac{\alpha}{\pi}\int_{-\infty}^{\infty} Dt\,\frac{\exp(-R_B^2(\alpha)\,t^2/2)}{H(R_B(\alpha)t)}\,. \tag{3}$$

The asymptotic behavior ($\alpha\to\infty$) of the Bayes limit generalization error is given by:

$$e_g^{Bayes}(\alpha) = \frac{1}{\alpha}\left(\int_{-\infty}^{\infty} Dt\,\frac{\exp(-t^2/2)}{H(t)}\right)^{-1} + O(\alpha^{-2}) \approx \frac{0.442}{\alpha}\,. \tag{4}$$

The process of learning is that of determining the $\mathbf{J}$ vector such that the student is able to implement the map defined by eq.(1) to some measure. This can be achieved by the minimization of a cost function, which leads in a natural way to the introduction of the ideas of statistical mechanics of the space of interactions, as pioneered by E. Gardner. We write the cost function as a sum over the training set

$$E(\mathbf{J}) = \sum_{\mu=1}^{P} V(\lambda_\mu)\,, \qquad \lambda_\mu = \frac{1}{\sqrt{N}}\,\mathbf{J}\cdot\mathbf{S}^\mu\sigma_B^\mu\,. \tag{5}$$

If the spherical constraint, $\mathbf{J}\cdot\mathbf{J} = N$, is imposed, the partition function is given by

$$Z_{\mathcal{L}} = \int d\mathbf{J}\,\delta(\sum_j^N J_j^2 - N)\,e^{-\beta E(\mathbf{J})} \tag{6}$$

As usual, the quenched average over the training set leads to the evaluation of

$$-\beta f = \lim\frac{1}{N}\langle\ln Z_{\mathcal{L}}\rangle_{\mathcal{L}}\,. \tag{7}$$

under the assumption of replica symmetry, which can be checked to be stable

$$\begin{aligned} f &= -\mathrm{Extr}_{q,R}\left\{\frac{q-R^2}{2\beta(1-q)} + \frac{\ln(1-q)}{2\beta}\right. \\ &+ \left.\frac{\alpha}{\beta}\int Dt_1\,Dt_2\ln\int\frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp\left[-\beta\left(V(\lambda\,\mathrm{sgn}\,t_2) - \frac{(\lambda - Rt_2 - \sqrt{q-R^2}\,t_1)^2}{2\beta(1-q)}\right)\right]\right\}, \end{aligned} \tag{8}$$

where Dt_i is the Gaussian measure $(2\pi)^{-1/2}\exp(-t_i^2/2)\,dt$ and integrations limits are $(-\infty,\infty)$ unless otherwise stated. The order parameter q is the typical overlap between different students, whereas R is the typical overlap between a student and the teacher.

The fact that the best possible student is unique permits us to use the streamlined formalism of BSB, which was developed to treat the case of nondegenerate ground states. That is, as $\beta\to\infty$ then $q\to 1$ in such a manner that $x = \beta(1-q)$ is finite. The free energy can be written as

$$f = -\mathrm{Extr}_{x,R}\left\{\frac{1-R^2}{2x} - 2\alpha\int Dt_1\int_0^\infty Dt_2\min_\lambda\left[V(\lambda) + \frac{(\lambda-t)^2}{2x}\right]\right\}, \tag{9}$$

where $t \equiv Rt_2 + \sqrt{1-R^2}\,t_1$.

The procedure to obtain the overlap R is very simple [14]. We must look for the function $\lambda_0(t,x)$ that minimizes

$$V(\lambda) + \frac{(\lambda-t)^2}{2x}\,. \tag{10}$$

The saddle point equations which furnish the values of R and x are

$$\frac{R}{\alpha} = \sqrt{\frac{2}{\pi}}\int Dt\,\lambda_0(\sqrt{1-R^2}\,t,x) \tag{11}$$

$$\frac{1-R^2}{\alpha} = 2\int DtH\left(\frac{-Rt}{\sqrt{1-R^2}}\right)[\lambda_0(t,x)-t]^2\,. \tag{12}$$

At the minimum of eq. (10) we have

$$\lambda_0 - h = xF\,, \qquad F \equiv -\left.\frac{\partial V(\lambda)}{\partial\lambda}\right|_{\lambda_0}\,. \tag{13}$$

Substituting in (11) leads to

$$x\sqrt{\frac{2}{\pi}}\int Dt\tilde{F} = \frac{R}{\alpha} \tag{14}$$

where $\tilde{F}$ is F in the transformed variable $t' = \sqrt{1-R^2}\,t$.

Doing the same substitution in equation (12), and when necessary $t \to t'$ and $t \to -t$

$$2x^2\int Dt\frac{1}{g}\tilde{F}^2 = \frac{\sqrt{1-R^2}}{\alpha}, \tag{15}$$

where

$$g = \frac{e^{-R^2t^2/2}}{H(-Rt)}. \tag{16}$$

Eliminating x leads to

$$\frac{R^2}{\sqrt{1-R^2}} = \frac{\alpha}{\pi}\frac{\langle gG\rangle_t^2}{\langle gG^2\rangle_t}, \tag{17}$$

where $G = \tilde{F}/g$ and

$$\langle(\ldots)\rangle_t \equiv \int Dt\,(\ldots). \tag{18}$$

It is easy to see that the right hand side of eq. (17) is maximized when G does not depend on t leading to

$$\tilde{F}_{opt} = G\,g \quad\Rightarrow\quad F_{opt} = G\frac{e^{-\frac{1}{2}t^2/\Gamma}}{H\left(-t/\sqrt{\Gamma}\right)}, \tag{19}$$

$$\frac{R^2}{\sqrt{1-R^2}} = \frac{\alpha}{\pi}\int Dt\frac{\exp-\frac{1}{2}R^2t^2}{H(-Rt)}, \tag{20}$$

$$x = \frac{\sqrt{\Gamma/2\pi}}{G}, \tag{21}$$

where $\Gamma = (1-R^2)/R^2$.

Since equation (20) is identical to equation (3) the performance of this algorithm is the best possible in the sense of generalization. Note that the off-line variational calculation does not determine G, since it refers to the properties of equilibrium and not to the dynamic question of optimizing relaxation times, an issue related to the learning rate G.

The functional dependence of F on t is the same as in the modulation function obtained by the on-line optimization [8].This is remarkable since it means that the optimization of the first step in the learning of αN examples leads to the optimal off-line algorithm. But the on-line calculation gives even more information since it determines the optimized learning rate $G_{OL} = \sqrt{\Gamma/2\pi}$ (incidentally, this gives $x = 1$). We conjecture that a lower bound for the relaxation times will be obtained by using a time-dependent learning step $G(\tau) = \sqrt{1-R^2(\tau)}/R(\tau)$ in the case of off-line learning (τ is the training time). However, the value of $R(\tau)$ is not an accessible quantity. This issue has to be addressed to determine a practical algorithm which approximates the bound, both in the sense of optimal performance, as well as in that of having optimal relaxation time.

In the *on-line unconstrained* learning case this represents no problem since it can be proved [10] that $R(\alpha) = \|\mathbf{J}(\alpha)\|$ so that we can use a $\Gamma(\|\mathbf{J}\|)$, which only depends on measurable quantities. For the off-line case [8] we have suggested substituting for $\Gamma(\tau)$ its equilibrium value, $\Gamma(\alpha) = \sqrt{1-R_B^2(\alpha)}/R_B(\alpha)$. Simulations led to a very close agreement with the OHB curve on a logarithmic scale [8]. The issue of optimal relaxation times, however, remains open.

This type of learning can be interpreted as a *modulated* Hebb algorithm. The modulation function depends on a balance of confidence and surprise, that is, the ratio $r = t/\sqrt{\Gamma}$ (see fig. 1): confidence of how well the student expects to perform in the new example, as measured by the factor $\Gamma = \tan(\pi e_g)$; surprise as indicated by the actual student performance on that example, given by the value of the stability t.

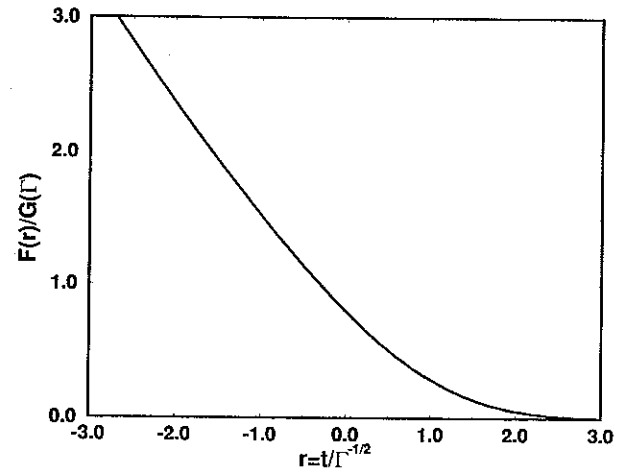

Figure 1: The rescaled weight function $F(r)/G(\Gamma)$ of the optimal algorithm.

The *simple* Hebb rule (or Hebb-Stent rule [17]) based only in temporal correlation between two neuron activities is not optimal, behaving asymptotically as $e_g \propto \alpha^{-1/2}$ [16]. For students in the early stage of learning (α small, $\Gamma \to \infty$) the optimal learning function $F(r)$ is almost simple Hebb, with a value around $F(0)$ in the interval of typical stabilities t. But the best performance ($e_g \propto \alpha^{-1}$) for experienced students (α large) is obtained when the student pays 'attention' (measured by the $F(r)$ function) only to the relevant and surprising examples (examples with high $r = t/\tan(\pi e_g)$ ratio. [1].

[1]Other standard algorithms (Perceptron, Adatron etc.) which give the $e_g \propto \alpha^{-1}$ behavior can be viewed as more or less reliable approximations of the optimal one.

Surprisingly, the *on-line* F_{opt} algorithm produces 'biological' side effects such as the well know *blocking phenomena* and *recency effects*, while the simple Hebb does not produce any of these phenomena. The modulation of Hebb mechanism at the synaptic level due to behavioral factors (attention, task relevance etc.) has been experimentally demonstrated recently [17]. The neurochemical basis of this modulation mechanism is an open and active topic in neurobiological studies. Discussing these topics is beyond the purpose of this paper.

We finish with some comments about the problem of overfitting. The optimal *on-line* algorithm is inconsistent in the sense that it has a nonzero training error. We have showed that the optimal *off-line* algorithm produces a vector within the *Bayes cone* defined by the angle $\arccos(R_B)$, but not necessarily belonging to the version space. Several iterative algorithms work by addressing the issue of consistency, they are iterated until the training error is zero. But this does not necessarily lead to better generalization [11, 13].

The class of algorithms that we have introduced do not work by dealing with the problem of consistency (the performance on past examples), but rather by minimizing an energy $E = -\Gamma \sum_\mu \ln P(\sigma_B^\mu \mid h_\mu)$ related to the loglikelihood of the data σ_B^μ given the present post synaptic field in the student $h_\mu = \mathbf{J} \cdot \mathbf{S}^\mu$ [8, 11, 12]. The teacher perceptron is at the border of the version space [14] and in trying to approximate it, the question of whether one is inside the version space or not should not be the *only* guiding force.

Acknowledgements: This research was partially supported by CNPq.

References

[1] Gardner E and Derrida B (1988) *J. Phys. A: Math. Gen.* **21** 271.

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

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

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

[5] Opper M and Haussler (1991) *Phys. Rev. Lett.* **66** 2677.

[6] Watkin T L H, (1993) *Europhys. Lett.* **21** 871.

[7] Kinouchi O and Caticha N (1992) *Physica* **185A** 411.

[8] Kinouchi O and Caticha N (1992) *J. Phys. A: Math. Gen.* **25** 6243.

[9] Kinouchi O and Caticha N (1993) *J. Phys. A: Math. Gen.* **26** 6161.

[10] Copelli M and Caticha N (1995) *J. Phys. A: Math. Gen.* **28** 1615

[11] Kinouchi O and Caticha N (1995) "On-line versus off-line learning in the linear perceptron: a comparative study" (to appear in *Phys. Rev. E.*)

[12] Copelli M, Kinouchi O and Caticha N (1995) "Equivalence between on-line learning in perceptrons with noisy examples and committee machines" (submitted to *J. Phys. A: Math. Gen.*)

[13] Kinouchi O and Caticha N (1995) "Preventing overfitting by on-line learning" (in preparation).

[14] Bouten M, Schietse J and Van den Broeck C (1994) "Gradient descent learning in perceptrons: a review of its possibilities", Limburgs Universitair Centrum *preprint.*

[15] Griniasty M (1993) *Phys. Rev. E* **47** 4496.

[16] Vallet F and Cailton J-G (1990) *Phys. Rev. A* **41** 3059.

[17] Ahissar E *et al.* (1992) *Science* **257** 1412.