Learning algorithm that 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, São Paulo, Brazil
(Received 1 May 1995; revised manuscript 25 April 1996)

A variational approach to the study of learning a linearly separable rule by a single-layer perceptron leads to a gradient descent learning algorithm with exactly the same generalization ability as the Bayes limit calculated by Opper and Haussler [Phys. Rev. Lett. **66**, 2677 (1991)]. 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 cost function is closely related to the optimal cost function derived for on-line learning. [S1063-651X(96)51507-2]

PACS number(s): 02.70.−c, 87.10.+e, 02.50.−r, 05.90.+m.

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 vast literature on this subject (for reviews, see [1]), which ranges from the numerical simulation and analytical calculation of the generalization error of different algorithms to the determination of the best possible (''Bayes'') performance by Opper and Haussler.

The proof that there is a perceptron which actually gives the Bayes performance was given by Watkin (in [1]). This optimal perceptron corresponds to the center of mass in student space calculated from the posterior distribution produced by the Gibbs algorithm. Watkin suggests determining the optimal perceptron by sampling this distribution by independently training $l\to\infty$ students.

In this paper we show that there exists a training energy with a nondegenerate minimum that gives *exactly* the optimal perceptron. The main idea in obtaining optimal generalization algorithms is to treat the learning problem as a variational one. This has been previously done in [2]–[6] for on-line learning, where the examples are used only once and are thereafter discarded. Here we extend the use of the variational approach to the off-line learning scenario. The generalization ability is calculated, in general, for any algorithm with a nondegenerate ground state, using the standard Gardner replica analysis of the space of interactions. The optimization of such ability determines the algorithm, and the optimized ability is exactly the Bayes curve. In performing, such a general calculation we rely heavily on the streamlined method of Bouten, Schietse, and Van den Broeck (BSB) [7].

By numerical optimization inside a limited class of functions, BSB have found a remarkably simple algorithm with a learning behavior very close to the Bayes curve [7]. Optimization of the so called relaxation algorithm also approximates the Bayes limit [8]. However, all these are somewhat *ad hoc* approaches and only work in the absence of noise. The question of the existence of a gradient descent algorithm that leads *exactly* to the optimal performance for every α remains and we now deal with it.

Let $\mathbf{B}$ and $\mathbf{J}\in\mathbb{R}^N$ be, respectively, the 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 composed by input vectors $\mathbf{S}^\mu$ and output data $\sigma_B^\mu=\mathrm{sgn}(\mathbf{B}\cdot\mathbf{S}^\mu)$. We take the input data to be independent random vectors uniformly distributed on the N-dimensional sphere. This particular situation is considered only for the purpose of illustrating the variational approach, being easily generalized to other distributions. In terms of $R=\mathbf{J}\cdot\mathbf{B}/\|\mathbf{B}\|\|\mathbf{J}\|$, the teacher-student overlap, which is a self-averaging quantity in the thermodynamic limit, the average generalization error is a monotonic function $e_g=(1/\pi)\arccos R$ [1].

The process of learning is that of iterative determination of the coupling vector $\mathbf{J}$ such that the student is able to approximate the map defined by the teacher. This can be achieved by a stochastic minimization of a cost function or *training energy*, which leads in a natural way to the introduction of the ideas of statistical mechanics in the space of interactions [1].

We write the training energy as a sum over the training set $E(\mathbf{J})=\Sigma_{\mu=1}^{P}V(\lambda_\mu)$, where $\lambda_\mu\equiv N^{-1/2}\mathbf{J}\cdot\mathbf{S}^\mu\sigma_B^\mu$ is the example stability. The quenched average over the training data is done by the replica method. As usual, the free energy will depend, under the assumption of replica symmetry, on the order parameters q, the typical overlap between different students, and R, 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 [7], which was developed to treat the case of nondegenerate ground states. That is, as $\beta\to\infty$ then $q\to1$ in such a manner that $x=\beta(1-q)$ is finite. The free energy can be written as

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

where $\mathfrak{E}$ is the extremum function and $t\equiv Rt_2+\sqrt{1-R^2}t_1$.

The procedure for obtaining the overlap R is very simple [7]. We must look for the function $\lambda_0(t,x)$ that minimizes $\mathcal{E}(\lambda)\equiv V(\lambda)+[(\lambda-t)^2/2x]$. The extremum conditions for R and x lead to

*Electronic address: osame@curie.if.usp.br
†Electronic address: nestor@if.usp.br

$$\frac{R}{\alpha}=2\int Dt_1\int_0^\infty Dt_2[\lambda_0(t,x)-t]\frac{\partial t}{\partial R}, \tag{2}$$

$$\frac{1-R^2}{\alpha}=2\int Dt_1\int_0^\infty Dt_2[\lambda_0(t,x)-t]^2. \tag{3}$$

At the minimum of $\mathcal{E}(\lambda)$ we have

$$\lambda_0-t=F,\quad F\equiv -x\frac{\partial V(\lambda)}{\partial\lambda}\bigg|_{\lambda_0}. \tag{4}$$

Substituting in (2) leads to

$$\frac{R}{\alpha}=\sqrt{\frac{2}{\pi}}\int Dt\widetilde{F}, \tag{5}$$

$$\frac{\sqrt{1-R^2}}{\alpha}=2\int Dt\frac{\widetilde{F}^2}{g}, \tag{6}$$

where $\widetilde{F}$ is F in the transformed variable $t'=\sqrt{1-R^2}t$ and $g\equiv e^{-R^2t^2/2}/H(-Rt)$. Equations (5) and (6) lead to

$$\frac{R^2}{\sqrt{1-R^2}}=\frac{\alpha}{\pi}\frac{\langle Gg\rangle_t^2}{\langle G^2g\rangle_t}, \tag{7}$$

where $G=\widetilde{F}/g$ and $\langle(\ )\rangle_t\equiv\int Dt(\ )$.

The solution of Eq. (7) determines the performance of any gradient descent algorithm, defined at this point by G, with a nondegenerate ground state. We have managed to reduce the problem to a point where the application of the variational ideas is now trivial. From a Schwartz-like inequality, the right-hand side of Eq. (7) is found to be maximized when G does not depend on t. The performance of the resulting algorithm is then given by the solution of the transcendental equation:

$$\frac{R^2}{\sqrt{1-R^2}}=\frac{\alpha}{\pi}\int Dt\frac{e^{-(1/2)R^2t^2}}{H(-Rt)}. \tag{8}$$

Equation (8) is exactly the Opper and Haussler [1] Bayes result. Thus the variational method constitutes not only an alternative and more direct procedure for obtaining the Bayes curve but will also give the form of the optimal training energy (see below). This enables us to study the properties of the optimal perceptron (stability distribution, classification error, etc.) by the standard Gardner method [9].

By referring to Eqs. (5) and (6) again, it can be found that $G=\sqrt{\Gamma/2\pi}$, with $\Gamma=(1-R^2)/R^2$. Then, the optimal F function is

$$F_{opt}=\sqrt{\frac{\Gamma}{2\pi}}\frac{e^{-(1/2)t^2/\Gamma}}{H(-t/\sqrt{\Gamma})}. \tag{9}$$

The function F_{opt} [Fig. 1(a)] is the difference between the prelearning and postlearning stabilities [7]. It has precisely the form of the modulation function of the optimal algorithm for on-line learning and it depends on the prelearning stability t of the example [2]. The modulation depends on the order parameter Γ. As it has been shown in [3] this can be

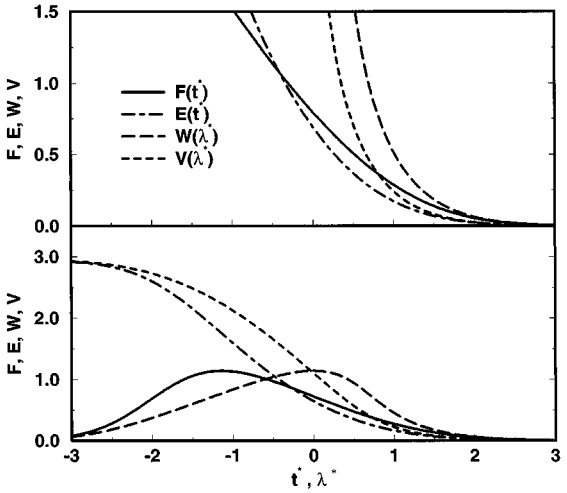

FIG. 1. Rescaled on-line modulation function $F_{opt}(t^\star)/\Gamma^{1/2}$ (solid), on-line energy $E_{opt}(t^\star)/\Gamma$ (dot-dashed), off-line modulation function $W_{opt}(\lambda^\star)/\Gamma^{1/2}$ (long-dashed), and off-line potential $V_{opt}(\lambda^\star)/\Gamma$ (dashed) for learning (a) without noise, (b) with noise level $\chi=0.05$.

viewed as a factor dependent of the length of the training set $[\Gamma\equiv\Gamma(\alpha)]$ for stationary rules or as performance dependent $[\Gamma\equiv\tan(\pi e_g)]$ for nonstationary environments. Note that as it is, the off-line variational calculation does not determine the value of x because it plays the role of a simple multiplicative constant to the optimal energy, being irrelevant for the equilibrium properties.

The potential $V_{opt}(\lambda)$ can be obtained, up to irrelevant additive and multiplicative constants, by integrating Eq. (4). Note that $F(t)+t\geq 0$, thus $\lambda\geq 0$, which, by the way, shows this to be a *consistent* algorithm. This means that $V_{opt}(\lambda)$ is infinite for negative arguments. For positive values of λ use Eq. (4) to obtain

$$V_{opt}(\lambda)=V_{opt}(\lambda_1)-x^{-1}\int_{t(\lambda_1)}^{t(\lambda)}F_{opt}\left(1+\frac{dF_{opt}}{dt'}\right)dt'.$$

But F_{opt} is the derivative, with respect to t, of the on-line optimal energy

$$E_{opt}(t)=-\Gamma\ln H(-t/\sqrt{\Gamma}), \tag{10}$$

so the integrand is just a total derivative. Then, the optimal energy function $V_{opt}(\lambda)$ is

$$V_{opt}(\lambda)=x^{-1}\{E_{opt}[t(\lambda)]-\tfrac{1}{2}F^2[t(\lambda)]\}, \tag{11}$$

where the value of $t(\lambda)$ is obtained from Eq. (4).

Thus, the optimal off-line potential $V_{opt}(\lambda)$ is not identical to the optimal on-line energy $E_{opt}(t)$, but is closely related to it. From the cavity method perspective [7], the term $F^2/2x=(\lambda-t)^2/2x$ appears as an additional energy contribution, due to the other examples, to the potential $V_{opt}(\lambda)$. It is this combined cost energy $\mathcal{E}=V_{opt}+(\lambda-t)^2/2x$ which is nicely related to the optimal on-line energy $E_{opt}(t)$. We call $W_{opt}\equiv-\partial V_{opt}/\partial\lambda$ the modulation function for the optimal off-line case. The relevant variables in the modulation functions are the ratios $t^\star\equiv t/\sqrt{\Gamma}$ and $\lambda^\star\equiv\lambda/\sqrt{\Gamma}$. We also can

rescale these function by the factor $\sqrt{\Gamma}$ in order to make them independent of R. The rescaled modulation functions and learning potentials are shown in Fig. 1(a).

Concerning the practical implementation of an optimal algorithm, we note that the dependence on the order parameter R, far from being disturbing, turns out to be of high theoretical importance, as we discuss below and in [9].

We stress that this work only illustrates, within a simple but paradigmatic situation, a general method for obtaining optimal learning curves and optimal cost functions. For each machine and learning environment (example distribution, noise distribution, etc.) there exists a corresponding optimal algorithm, which can be found by the method introduced here. It is currently being used for determining optimal algorithms for noisy environments [9] and unsupervised learning situations [10]. We only give the results (see [9]) for learning in the presence of output noise. Let χ be the flip probability; then the modulation function is changed to

$$F_{opt}=(1-2\chi)\sqrt{\frac{\Gamma}{2\pi}}\frac{e^{-1/2\,t^2/\Gamma}}{\hat{H}(-t/\sqrt{\Gamma})}, \tag{12}$$

where $\hat{H}(-t/\sqrt{\Gamma})=\chi+(1-2\chi)H(-t/\sqrt{\Gamma})$. The corresponding Bayes curve is given by

$$\frac{R^2}{\sqrt{1-R^2}}=\frac{\alpha}{\pi}(1-2\chi)^2\int Dt\frac{e^{-1/2\,R^2t^2}}{\hat{H}(-Rt)}, \tag{13}$$

which agrees with the results of Opper and Haussler [1]. W_{opt} and V_{opt} are shown in Fig. 1(b).

The difference between on-line and off-line can be summarized as follows. Off-line learning is described by a Langevin equation, i.e., an energy gradient descent process plus noise, where the energy $E=\Sigma_\mu V(\lambda_\mu)$ is defined over the whole set of examples. On-line learning, on the other hand, while also being a gradient process, has a cost function $E(t_\mu)$ that depends only on the latest example.

For optimal learning, we have shown that the two types of energies, although related, are not the same. The main qualitative difference lies in that they depend on different variables λ and t, the postlearning and prelearning stabilities, this difference being clarified by the cavity interpretation of the learning process [7,9]. The optimal on-line energy has been deduced for several different architectures (boolean perceptron [2,3], linear perceptron [5,9], tree committee machine [6], parity machine [11], and unsupervised learning [10]).

The optimal algorithms have a rich structure and present some properties that turn out to be very interesting when examined from a biological perspective. Remember that the perceptron is not only a formal neuron model. It is a general model for associative learning and causal inference based on a weighted sum of signals, appearing in different contexts from information processing by protein networks [12] to classification tasks in animals including humans (Rescola-Wagner models [13] are single-layer perceptrons) and weighted voting in committees.

As our results indicate, for each environment there is an optimal way of doing this associative learning that corresponds to a very specific *modulated* Hebb mechanism that evolves along the learning process. The modulation function depends on a balance of confidence and surprise, that is, the ratio $t^\star=t/\sqrt{\Gamma}$: confidence of how well the student expects to perform in the new example, given its average performance 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 prelearning stability t. It also depends on the type and level of the noise in the examples; see Fig. 1(b) [6,14].

The optimal algorithm for each learning situation also suggests how practical algorithms can be constructed—they must mimic the properties of the optimal one—and gives a benchmark curve for their evaluation. For optimality to be achieved, the cost function must be time dependent [2,15], or performance dependent if the rule changes with time [3].

The fact that the optimal algorithms depend on various internal and external quantities may appear to be distressing, since those may not be readily available variables. We do not see this as a problem but as an inevitable theoretical result that provides important insights for a learning theory. It indicates that, in order to optimize learning algorithms, there exists a ''selection pressure'' for the development of ''modules'' which are needed to estimate the unknown but important quantities (type and level of noise, performance level, surprise level, etc.). These on-line estimators have already been developed for the perceptron [3,14,16].

To conclude, the variational approach for determining optimal learning curves has been extended to off-line learning. Minimization of $V_{opt}(\lambda)$ produces the optimal perceptron. This training energy is found to be closely related to the optimal on-line energy $E_{opt}(t)$. The optimal algorithms present complex and rich mechanisms for modulating the Hebbian term. It is important to stress that our variational replica calculation for the optimization of cost functions is general and can be extended to other distributions of examples, machines and perhaps even other optimization problems. The variational approach leads to a scenario where learning algorithms are not hard wired, but are the object of a (second-order) learning process; a scenario where students learn to learn in an optimal way.

Discussions with Chris Van den Broeck and Mauro Copelli are gratefully acknowledged. This research was partially supported by CNPq and FAPESP.

[1] T. L. H. Watkin, A. Rau, and M. Biehl, Rev. Mod. Phys. **65**, 499 (1993); M. Opper and W. Kinzel, in *Physics of Neural Networks III*, edited by E. Domany, J. L. Van. Hemmen, and K. Schulten (Springer, Berlin, 1994).

[2] O. Kinouchi and N. Caticha, J. Phys. A **25**, 6243 (1992).

[3] M. Biehl and H. Schwarze, J. Phys. A **26**, 2651 (1993); O. Kinouchi and N. Caticha, J. Phys. A **26**, 6161 (1993).

[4] M. Copelli and N. Caticha, J. Phys. A **28**, 1615 (1995).

[5] O. Kinouchi and N. Caticha, Phys. Rev. E **52**, 2878 (1995).
[6] M. Copelli, O. Kinouchi, and N. Caticha, Phys. Rev. E **53**, 6341 (1996).
[7] M. Bouten, J. Schietse, and C. Van den Broeck, Phys. Rev. E **52**, 1958 (1995).
[8] R. Meir and J. F. Fontanari, Phys. Rev. E **45**, 8874 (1993).
[9] O. Kinouchi and N. Caticha (unpublished).
[10] C. Van den Broeck and P. Reimann, Phys. Rev. Lett. **76**, 2188 (1996).
[11] R. Simonetti and N. Caticha, J. Phys. A (to be published).
[12] D. Bray, *Nature* **376**, 307 (1995).
[13] M. A. Gluck and G. H. Bower, J. Exp. Psychology General **117**, 227 (1988).
[14] M. Biehl, P. Riegler, and M. Stechert, Phys. Rev. E **52**, R4624 (1995).
[15] D. S. Chen and R. C. Jain, IEEE Trans. Neural Netw. **5**, 467 (1994).
[16] Heskes T, in *Proceedings of the ZiF Conference on Adaptive Behavior and Learning*, edited by J. Dean, H. Cruse, and H. Ritter (University of Bielefeld, Bielefeld, Germany, 1994).