Generalization Performance of Complex Adaptive Tasks

E. Eisenstein and I. Kanter
Department of Physics, Bar-Ilan University, Ramat-Gan 52900, Israel
(Received 11 February 1993)

Optimal strategies for predicting correctly the output of a few new random inputs, when various feedforward networks are trained by noise-free random training examples, are examined analytically and numerically. The existence of a universal strategy for various generalization tasks is discussed, and indicates that the Bayes algorithm is not always the optimal strategy.

PACS numbers: 87.10.+e, 02.50.Le

The new statistical mechanics approach suggested by Gardner [1,2] was commonly used for estimating the statistical nature of the fixed points of various networks, which for sophisticated researchers is measured by the quantity capacity [3]. This line of research has recently been extended for the calculation of other functions of networks. The aspect which attracts much attention is the process of learning from examples in the architecture of feedforward networks [4].

In this process, a set of random examples (random inputs) is chosen from a network which is called the teacher. Another network, which in the simplest case has the same architecture as the teacher, is trained by a learning algorithm to predict correctly the classification (outputs) of the teacher on this set of examples. The set of the examples is called the training examples and the trained network is called the student. The generalization performance of the student is defined as the average probability that the prediction of the student on a new random example is correct. This probability depends on the strategy (learning algorithm) which is adopted by the student.

The cornerstone of all the known strategies is the version space (VS), i.e., the set of all the weight vectors that are consistent with the training examples. One well discussed strategy is the Boltzmann algorithm, which assumes a constant density for any weight vector belonging to the VS and zero everywhere else [5,6]. The probability of predicting correctly is therefore the average probability over all vector weights in the VS. However, it is proven that the Boltzmann algorithm is not the best strategy [7]. The best strategy is known as the Bayes optimal classification algorithm, and was recently investigated and compared both analytically and numerically to the Boltzmann algorithm [7,8]. This optimal strategy is based on the fact that without any prior knowledge, the teacher can be any weight vector in the VS with equal probability. Hence, the optimal prediction for a new random example is fixed by the majority vote among the weight vectors which belong to the VS. These two strategies, especially the former, were examined exhaustively for the case of learning from examples of a new random pattern for various types of perceptrons and for some limited classes of multilayer networks (MLN) [4].

In this Letter, the best strategy for complicated tasks of learning from examples is examined both analytically and numerically. The question of predicting correctly the classification of *one* new random input is only a prototype aspect of possible generalization tasks. The optimal strategy for predicting correctly the classification of l new random inputs after the training of p examples is at the center of the following discussion. Practically, the process of learning from examples is an expensive process whose aim it is to predict the classification of many events (inputs). In such a case, an important task might be to maximize the probability of a correct prediction over some function of all the events. Basically, the number of such plausible different functions (tasks) increases with l, the number of events. Nevertheless, one can distinguish between the following three main categories: (a) to maximize the probability that the whole l events are predicted correctly, (b) to maximize the average number of events which are predicted correctly, and (c) to maximize the average number of correct predictions as in (b), under the constraint that at least k events (among l) are predicted correctly. Such questions are relevant to any set of events which are required to be predicted correctly, and in particular they are important in the case where an association among the sequence (or a subsequence) of events has a logical content such as words, codes, etc. It is interesting to examine whether the Bayes algorithm is universal in the sense that it is the optimal strategy for any generalization task, independent of the architecture of the network, or whether the optimal strategy depends on the details of the task and the network.

The following discussion concentrates mainly on categories (a) and (b) where the discussed architectures are the perceptron and the committee machine (CM) with three hidden units and with nonoverlapping receptive fields [9, 10]. The generalization of the results to other architectures is straightforward. More precisely, the perceptron is a single layer classifier,

$$f(\mathbf{w}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{x}), \tag{1}$$

where the input vector $\mathbf{x}$ is a real vector of N dimensions, $\mathbf{w}$ is a weight vector, and the classification of the perceptron is defined by $f(\mathbf{w})$. In the discussed CM

0031-9007/93/70(23)/3667(4)$06.00

$$f(\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3) = \mathrm{sgn}[\mathrm{sgn}_1 + \mathrm{sgn}_2 + \mathrm{sgn}_3], \quad (2)$$

where sgn_i stands for $\mathrm{sgn}(\mathbf{w}_i \cdot \mathbf{x}_i)$, $\mathbf{x}_i$ is a real vector of N dimensions, and f defines the classification of the network.

In the case of a perceptron and $l = 2$, each random input divides the VS into two regimes $A_i(\pm)$. These symbols, $A_i(\pm)$, stand for the regime's size of the VS which is signed by $\pm$ for the ith random input, and from normalization of the VS to 1, $A_i(+) = 1 - A_i(-)$. To simplify the discussion and without the loss of any generality, assume that $A_i(+) > 1/2$ for any i. In another description, the VS splits into four regimes denoted by $A(\pm\pm)$; each one of them stands for the possible outputs of the first and second random inputs, respectively. It is clear that the optimal strategy for predicting correctly the two inputs is to predict the answer following the largest regime among the four, similar to the sense of the Bayes algorithm. This algorithm is called in the following Bayes_2 algorithm, or Bayes_l algorithm in the general case of l random inputs where the VS splits into 2^l regimes. Let us prove now that in the thermodynamic limit and in the simple architecture of the perceptron, applying Bayes_1 twice is equivalent to applying Bayes_2 [11]. Since the inputs are *uncorrelated* then $A(\pm\pm) = A_1(\pm)A_2(\pm)$, respectively. Hence, the largest area among the four is $A(++)$ and the answers of the two algorithms are identical. It is easy to verify that the same identity occurs between Bayes_1 and Bayes_l in the general case of l random inputs. In a similar way one can also verify that Bayes_1 maximizes the average number of inputs which are predicted correctly [which is equal to the average area of $2A(++) + A(+-) + A(-+)$]. Hence, Bayes_1 algorithm is the *universal strategy* for the architecture of the perceptron with random inputs and in the thermodynamic limit, but for finite systems fluctuations are significant, $A(\pm\pm) \neq A_1(\pm)A_2(\pm)$, and the optimal algorithm is Bayes_l.

One may conclude from the simplest classifier that Bayes_1 algorithm is the universal strategy at least in the thermodynamic limit for any architecture. Nevertheless, in the following it is proven that this conclusion is wrong. Before the details are discussed let us introduce the notion of a *pocket*. In the case of p training examples and CM with three hidden units, the VS of the student is constructed from 4^p legal internal representation (LIR). Each LIR is a set of p internal representations (IR) for each of the examples, which gives the same output as the teacher. Each weight vector (of $3N$ dimensions) in the VS implements only one LIR. All the weight vectors which belong to one LIR are defined as a pocket, and it is clear that some of the 4^p pockets can be empty. Note that the notion pocket is *not conjugate* to a replica symmetry breaking phase [12].

In the following we concentrate on cases where only a finite number of pockets are available to the student. Let us mention now some of the cases where this assumption is realistic. (a) In simulations, only finite numbers of pockets are examined. (b) In cases where there is some prior knowledge on the state of the hidden units of the teacher, besides the desired output. In the case where all the IR of the teacher are known, the VS is constructed only from one pocket. However, when this knowledge is noisy, a few pockets are available for the student. (c) It is possible that there are many pockets which have the same extensive entropy, but the corrections to the extensive entropy lead to the fact that only a few pockets are dominant. (d) Asymptotically, when the size of the training examples diverges, the size of the VS shrinks to zero. It is still not known whether as the VS shrinks to zero the number of LIR also decreases to zero, or whether the number of LIR is exponential with the size of the training examples, but the size of each pocket shrinks to zero. (e) Even if we assume that asymptotically the number of pockets is large, they are strongly correlated (see discussion below). Hence, it is possible that effectively the number of pockets is finite.

In the case of a CM where only one pocket is available for the student, then for the discussed generalization tasks the optimal strategy, as for the perceptron, is Bayes_1 which is identical to Bayes_l. Assume that $A_{ik}(\pm)$ is the size of the VS of the kth hidden unit where the ith pattern gives the sign $\pm$, respectively. One can easily verify now that the part of the pocket which is signed by $+$ is given by

$$B_i(+) = A_{i1}(+)A_{i2}(+) + A_{i1}(+)A_{i3}(+) + A_{i2}(+)A_{i3}(+) - 2A_{i1}(+)A_{i2}(+)A_{i3}(+), \quad (3)$$

and the output is fixed by the maximal regime among $B_i(\pm)$. Furthermore, in the case of $l = 2$, for instance, one can show that

$$B(\pm\pm) = B_1(\pm)B_2(\pm), \quad (4)$$

and therefore the results for the perceptron case are valid also to the general case of one pocket in a CM with k hidden units.

The case where many pockets are available for the student is straightforward but, for simplicity, detailed results are presented only for the case of two pockets. In the discussed CM with only two available pockets, the prediction of Bayes_1 algorithm is according to the largest joint regime of the two pockets. Assume that the VS of each pocket is normalized to 1 and assume that $B_i^\rho(\pm)$ stands for the signed $\pm$ regime of the ith input in the ρ's pocket. The output of Bayes_1 for each one of the two inputs is fixed by the maximal signed regime in the two pockets

$$\max\{B_i^1(+) + B_i^2(+),\ B_i^1(-) + B_i^2(-)\}, \quad (5)$$

where $i = 1, 2$. In the case of Bayes_2 algorithm, the VS of the joint pockets splits into four regimes, $B(\pm\pm)$, and

the prediction is fixed following the maximal one

$$\begin{aligned} B(++) &= B_1^1(+)B_2^1(+) + B_1^2(+)B_2^2(+), \\ B(+-) &= B_1^1(+)B_2^1(-) + B_1^2(+)B_2^2(-), \\ B(-+) &= B_1^1(-)B_2^1(+) + B_1^2(-)B_2^2(+), \\ B(--) &= B_1^1(-)B_2^1(-) + B_1^2(-)B_2^2(-). \end{aligned} \tag{6}$$

Surprisingly, Eq. (6) is not always consistent with Eq. (5), and therefore the predictions of Bayes_1 and Bayes_2 are not identical. It is now clear that the optimal strategy for predicting correctly two random inputs is Bayes_2, and hence Bayes_1 is not the universal optimal strategy. Running numerically on the allowed values for $B_i^\rho(\pm)$, one finds that the fraction of cases where the prediction of Bayes_1 differs from Bayes_2 is 0.14. Note that although 14% is not a negligible fraction, it is not the right quantity to measure the differences between the quality of these algorithms for the following reasons. First, each partition of the VS into two regimes of the sizes $B_i^\rho(\pm)$ does not have the same probability. Hence, the counting of these cases should be multiplied by their probability. Second, the quality of the two algorithms is not only a function of the probability that their predictions are different, but is also a function of how much they differ. More precisely, assume that Bayes_1 does not choose the largest regime among the four, as does Bayes_2. Since the probability that the two inputs are predicted correctly is proportional to the size of the chosen regime, the ratio between the regimes chosen by the two algorithms is also an important quantity in comparing the algorithms. The following discussion is devoted to estimating these differences quantitatively, both analytically and by simulations.

The cornerstone of the analytic work is the quantity $P(A)$, which stands for the probability that the VS of a perceptron splits by a new random input into two regimes of the sizes A and $1-A$, independent of their sign. A similar quantity was recently calculated [8], and from this one finds

$$P(A) = \gamma(\alpha)^{-1}\exp\left(-\frac{t^2}{2}[1-\gamma(\alpha)]^2\right), \tag{7}$$

where t is fixed by the solution of $A = H(t\gamma)$, $H(z) = \int_z^\infty Dt$, $Dt = (2\pi)^{-1/2}\exp(-t^2/2)$, and $p = \alpha N$ is the size of the training examples. The explicit dependence of γ on α for some special cases, such as the spherical perceptron, is known analytically [4,5]. It is clear that $P(A)$ is symmetric around $1/2$, since the sign of the output is unbiased. It is also clear that there are average nonvanishing correlations among the IR of the hidden units in each pocket. Nevertheless, this effect does not create correlations among the probabilities, $P(A)$, in different hidden units, since the architecture consists of nonoverlapping receptive fields. Hence, the probability that the VS of each pocket splits by a new random example into two regimes, B and $1-B$, is given for the CM with $K=3$ by

$$Q(B) = \int_0^1 \prod_{i=1}^{3} dA_i\, P(A_i)\delta(-2A_1A_2A_3 + A_1A_2 + A_1A_3 + A_2A_3 - B). \tag{8}$$

This probability is calculated numerically and serves as a preprocessing for problems with more than one pocket. Note that this preprocessing reduces the complexity of the calculations and makes the numerical calculations feasible for two and three pockets. The analytical results for predicting correctly $l=2$ random inputs for one pocket and for the spherical case are presented in Fig. 1.

For the general case of ρ pockets (even correlated) and l random inputs one can show analytically that

$$T_{1+1\cdots+1} = T_1^l, \tag{9}$$

where the symbol $T_{i_1+i_2+\cdots+i_j}$ stands for the probability of predicting correctly i_1 random inputs by Bayes_{i_1} and additional i_2 random inputs by Bayes_{i_2}, etc. The starting point of the proof is that for $\rho = 2$, $T_1 = \int dB_1\, dB_2\, Q(B_1)Q(B_2)\Theta(B_1+B_2-1)(B_1+B_2)/2$, where $Q(B_1)$ and $Q(B_2)$ can be correlated. Assume that $Q(A_i)$ has the same meaning as $Q(B_i)$, but for another input. It is clear that $Q(A_i)$ is uncorrelated with $Q(B_j)$ since we deal with random inputs, and from that the derivation of Eq. (9) is straightforward. The average number of inputs which are predicted correctly is $\sum \binom{l}{k} T_1^l (1-T_1)^{l-k} = lT_1$. Hence, Bayes_1 is the *universal optimal strategy* to maximize the average number of correct predictions.

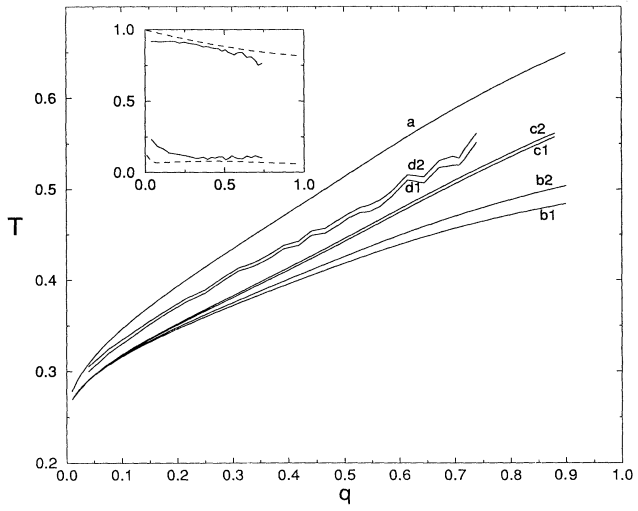

FIG. 1. Results for T_{1+1} and T_2 as a function of q for the discussed CM with a spherical constraint on the weights [$\alpha(q)$ is given explicitly in Ref. [5]]. (a) Analytical results for T_2, for one pocket, derived from Eq. (8). Two uncorrelated pockets, T_{1+1} (b_1) and T_2 (b_2). Two correlated pockets, Eq. (10), T_{1+1} (c_1) and T_2 (c_2). Simulations of T_{1+1} (d_1) and T_2 (d_2). Inset: The probability $P(T_{1+1} \neq T_2)$ (lower two lines) and the average T_{1+1}/T_2 (upper two lines) vs q. Analytical results for correlated pockets, Eq. (10) (dashed lines), and simulations (full lines).

Quantitative results even for the case of two pockets are more involved. In the case where the correlations among the pockets are negligible, one can show that T_2 converges asymptotically ($\alpha \to \infty$) to 9/16, instead of 1. Nevertheless, the results for T_2 and T_{1+1} for this case of uncorrelated pockets are presented in Fig. 1. A more realistic model takes into account the correlations among the pockets, and the simplest type of correlations is such that

$$P(B_2|B_1) = CP(B_2)(1 - |B_2 - B_1|q), \tag{10}$$

where C is a normalization constant, $P(B_2|B_1)$ is a conditional probability for B_2 given B_1, and q is the average overlap between two vectors of N dimensions in the VS which belong to one hidden unit [4]. It is clear that $P(B_2|B_1)$ obeys the trivial limits. At $q = 0$, B_2 is independent of B_1 and as $q \to 1$, $P(B_1)$ is constructed from two delta functions with probability 1/2 at 0 and 1, and hence B_2 is fully correlated with B_1. In the intermediate regime, $0 < q < 1$, the probability of B_2 being in a distance $|x|$ from B_1 decreases as xq. The analyical results of this simple model for T_2 and T_{1+1} are presented also in Fig. 1. The results for both cases of correlated and uncorrelated pockets indicate that $T_2 \neq T_{1+1}$. Furthermore, finer quantities, defined above, can be used to compare both algorithms. Using Eqs. (7)–(10), the probability that both algorithms differ in their predictions, $P_{ne}(T_2 \neq T_{1+1})$ and the average T_{1+1}/T_2 for these cases were also calculated analytically and presented in the inset of Fig. 1. The results roughly indicate that for 10% of the samples the prediction of both algorithms differs by 10%.

In order to examine the assumption, Eq. (10), and the effects of finite size systems, heavy numerical simulations were carried out. The size of the input of each hidden unit was $16 \leq N \leq 48$ and the second pocket was found from random initial configuration by the least action algorithm [9, 10]. The VS of each pocket was sampled by 5000–20000 random movements and each point was averaged over 3000–5000 samples. The results for the spherical case are presented in Fig. 1 from which it is clear that the average T_{1+1}/T_2 and P_{ne} are close to the analytical predictions even for small systems. Significant deviations for small q are due to strong fluctuations at finite p, and decreases when the size of the system increases [13]. Note, however, that in the limits $q \to 0$ or 1 the differences between both algorithms vanish. The effect of correlations among the pockets is not negligible and even a simple type of correlation among the pockets, Eq. (10), gives a deviation of a few percent from the exact results, which depend on the whole structure of correlations. Furthermore, similar deviations were found between the exact analytical results and simulations (same size system) for the perceptron. Hence, it is expected that the results of Eq. (10) should be very close to the exact results in the thermodynamic limit.

For the third category of generalization tasks neither Bayes_1 nor Bayes_l is the optimal strategy. The optimal strategy for these cases is a table which indicates for each type of partition of the VS to make a prediction following a particular part of the VS, among 2^l. The major reason for such a behavior is that there is no definite order among the sizes of the 2^l parts of the VS, independent of the partition. The minimal size of the table is still in question. Note that the above-mentioned universality strategies also split into many tables where the inputs are correlated.

This research is supported in part by the Israeli Academy of Science and Humanities.

[1] E. Gardner, J. Phys. A **21**, 257 (1988).
[2] E. Gardner and D. Derrida, J. Phys. A **21**, 271 (1988).
[3] See, for example, J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Reading, MA, 1991).
[4] See, for instance, T. L.H. Watkin, A. Rau, and M. Biehl, "Statistical Mechanics of Learning a Rule," Rev. Mod. Phys. (to be published).
[5] G. Gyorgi and N. Tishby, in *Neural Networks and Spin Glasses*, edited by K. Thuemann and R. Koeberle (World Scientific, Singapore, 1990).
[6] H. Sompolinsky, N. Tishby, and H. S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).
[7] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).
[8] M. Opper and D. Haussler, in *Proceedings of the Fourth Workshop on Computational Learning Theory*, edited by M. K. Warmuth and L. G. Valliant (Morgan Kaufmann, San Mateo, 1991).
[9] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).
[10] A. Engel, H. M. Kohler, F. Tschepke, H. Vollmayr, and A. Zippelius, Phys. Rev. A **45**, 7590 (1992).
[11] See also, T. L. H. Watkin, Europhys. Lett. **21**, 871 (1993).
[12] M. Mezard, M. A. Virasoro, and G. Parisi, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[13] E. Eisenstein and I. Kanter (unpublished).