Scaling Laws in Learning of Classification Tasks

N. Barkai, H. S. Seung, and H. Sompolinsky
Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel
and AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974
(Received 14 January 1993)

The effect of the structure of the input distribution on the complexity of learning a pattern classification task is investigated. Using statistical mechanics, we study the performance of a winner-take-all machine at learning to classify points generated by a mixture of $K$ Gaussian distributions ("clusters") in $R^N$ with intercluster distance $u$ (relative to the cluster width). In the separation limit $u \gg 1$, the number of examples required for learning scales as $N K u^{-p}$, where the exponent $p$ is 2 for zero-temperature Gibbs learning and 4 for the Hebb rule.

PACS numbers: 87.10.+e, 02.50.-r, 05.20.-y

A fundamental problem in learning theory is the understanding of the determinants of generalization, the ability of a learner to respond correctly to inputs not previously seen in training. Naively, one expects that low generalization error can only be achieved if the learner is trained on more examples than there are degrees of freedom adjusted in the learning process. In other words, the scale of the learning curve (the generalization error as a function of the number of examples) should be the number of free parameters. This expectation is borne out in solutions of several neural network models using statistical mechanics [1,2]. Likewise, results from computational learning theory suggest that the scale of the learning curve is the Vapnik-Chervonenkis dimension, which for feedforward neural networks is closely related to the number of parameters [3]. However, empirical studies of learning in layered neural networks often show low generalization error with fewer examples than parameters [4,5]. It has been suggested [6] that such empirical results might be explained by a learning theory for more "realistic" input distributions. The present work investigates the effect of the shape of the input distribution on the learning of pattern classification tasks. We focus on the interesting case of high-dimensional inputs, and define an appropriate thermodynamic limit for the problem. We study the case of an input distribution that is a mixture of Gaussian clusters, and calculate the dependence of the learning curve on the separation of the cluster centers.

In the Gaussian mixture model, points $x \in R^N$ are generated by a two step procedure. First a class $l = 1, \ldots, K$ is selected with class probability $P(l)$. Then $x$ is chosen with conditional probability

$$P(x|l) = \frac{1}{(2\pi)^{N/2}} \exp\left(-|x - U_l|^2/2\right),$$  

i.e., according to a Gaussian distribution with mean $U_l$. For simplicity we assume $P(l) = 1/K$ for all $l$, and that the cluster centers $U_l$ are orthogonal to each other and have equal magnitudes, i.e., $U_l \cdot U_m = \delta_{lm}$. Since only the ratio of mean to fluctuation matters, we have assumed that each component of $x$ has unit variance.

Given a point $x$, the problem of pattern classification is to decide which class it came from. Since the clusters (1) are overlapping, this cannot be done with perfect accuracy. If a classifier is denoted by $C(x)$, a function taking on the values $1, \ldots, K$, its probability of error can be written as

$$\epsilon(C) = \sum_{k=1}^{K} P(k) \epsilon_k(C),$$

where $\epsilon_k(C)$ is the class-conditional error, i.e.,

$$\epsilon_k(C) = 1 - \langle \delta_{k,C(x)} \rangle_k,$$

and $\langle \cdots \rangle_k$ means an integration over $x$ with the measure $P(x|k)$. In the context of learning theory, $\epsilon(C)$ is the generalization error of the classifier $C$.

We first discuss the properties of the optimal classifier, i.e., the one that minimizes $\epsilon(C)$. For the case of equal prior probabilities, $P(l) = 1/K$, it is straightforward to prove that the optimal classifier is the one which maximizes $P(x|l)$ with respect to $l$ [7]. For our model (1), this is equivalent to assigning each $x$ to the nearest cluster center. Hence the optimal classifier induces a Voronoi tessellation of the input space into $K$ cells, each containing a cluster center. We can write it as

$$C_{\text{opt}}(x) = \text{argmax}_k \{ P_k(x) \} = \text{argmax}_k \{ U_k \cdot x \},$$

where we have used the fact that all $U_k$ have the same norm.

The class-conditional error $\epsilon_k$ of the optimal classifier is the probability that an input $x$ that is generated by $P(x|k)$ will be nearer to a cluster center $U_l$, $l \neq k$, than to $U_k$. Hence,

$$\epsilon_k(C_{\text{opt}}) = 1 - \left( \prod_{l=1, l \neq k}^{K} \Theta(x \cdot (U_k - U_l)) \right)_k,$$

where $\Theta(x)$ is the Heaviside step function. To carry out
the integration \( \langle \cdot \rangle_k \) one notices that the argument of the step functions contains \( K - 1 \) independent Gaussian variables \( \{ x \cdot U_l / \sqrt{N} \} , l \neq k \), which have zero mean (because of the orthogonality of the centers) and variance \( u^2 \). In addition, the term \( x \cdot U_k / \sqrt{N} \) is a common Gaussian variable with both mean and variance equaling \( u^2 \). Integrating over these Gaussian variables one obtains

\[
\epsilon_{\text{opt}} = 1 - \int Dz \left[ H(z - u) \right]^{K-1} ,
\]

where \( H(z) \equiv \int_{-\infty}^{\infty} dt \) and \( Dt \equiv dt \exp(-t^2/2) / \sqrt{2\pi} \). This residual classification error is a measure of the overlap between the Gaussian distributions.

The problem of learning a pattern classification arises when the classifying system does not have complete knowledge of the input distributions, and hence can only approximate the action of the optimal classifier (4). We assume that the learner has access to a training set of \( P \) classified examples \( \{x^\mu, l^\mu\} , \mu = 1, \ldots, P \), and using them must construct a classifier with low error. One method of learning from these examples is to estimate the conditional distributions and then construct a classifier using (4). Another method is to bypass the intermediate step of density estimation, and search directly for a classifier that classifies the training examples with low error, with the hope that this classifier will generalize well. This second method is what concerns us here.

Other than the work of [8] on problem-average error rates, there have been few exact calculations of learning curves for classification tasks. In this Letter, we use statistical mechanics to perform such exact calculations. For this purpose, we propose below two limits: the thermodynamic limit and the separation limit. The first limit is relevant to the classification of high-dimensional inputs that are generated by relatively few clusters. The second limit specializes to the case where the overlap between the clusters is relatively small.

**Thermodynamic limit.**—We define the thermodynamic limit of the classification problem as the limit where the dimension \( N \) of \( x \) goes to infinity, while both the number of clusters \( K \) and the separation parameter \( u \) remain finite. Thus, within a cluster the ratio between the mean and the typical fluctuation of each component of \( x \) is roughly \( u / \sqrt{N} \), which is vanishingly small in the thermodynamic limit just defined. The motivation for this definition of the thermodynamic limit is twofold. First, keeping \( u \) finite guarantees that the overlap between the clusters remains of order unity, and hence that the optimal error remains finite; see Eq. (6). Second, and more importantly, in this limit the number \( P \) of examples required to achieve a good classification performance scales linearly with the number of weights \( NK \), as is the case with other learning problems in high dimensions. Thus in the thermodynamic limit \( \alpha \equiv P/NK \) is held finite while both \( P, N \rightarrow \infty \).

**Separation limit.**—Even in the thermodynamic limit, theoretical study of the performance of classifiers can be complicated. However, the calculations are considerably simplified in the limit where the clusters are well separated, i.e., in the case where \( u \gg 1 \) (but still also \( u \ll \sqrt{N} \)). In the \( u \rightarrow \infty \) limit the minimal error (error of the optimal classifier) Eq. (6) is negligible. In fact, for any fixed number of examples, the generalization error tends to zero as \( u \rightarrow \infty \). However, if the number of examples is reduced while the separation increases, a meaningful limit can be defined. We define the separation limit to be \( \alpha \rightarrow 0, u \rightarrow \infty \) with \( \alpha u^p \) held constant, such that the generalization error remains finite. The power \( p \) will depend on the particular model and learning algorithm.

We have used the above limits to analyze the performance of the winner-take-all (WTA) machine at learning the task of classifying inputs from the Gaussian mixture defined above. The WTA classifier is written as

\[
C(x) = \arg\max_k \{ W_k \cdot x \} , \quad (7)
\]

and is parametrized by the set of \( N \)-dimensional weight vectors \( W_k \), \( k = 1, \ldots, K \). We impose a spherical constraint \( W_k \cdot W_k = 1 \) on each of the \( K \) weight vectors. As with the optimal classifier (4), the \( K \) decision regions are the Voronoi cells corresponding to the weight vectors. This classifier can be thought of as a neural network in which classification is the result of a winner-take-all operation performed on the outputs of a layer of \( K \) linear hidden units. In this context the vector \( W_k \) specifies the connections between the input neurons and the \( k \)th hidden neuron. The WTA classifier has also been called a "linear machine" [7] and a "multiclass perceptron" [9].

Obviously, for \( W_k = U_k \), the WTA machine implements the optimal classifier (4), and will therefore have a generalization error given by Eq. (6). For arbitrary \( W_k \), the expression for the generalization error is cumbersome. For the statistical mechanical calculation below, it suffices to consider only weight vectors that satisfy the symmetric conditions

\[
W_k \cdot U_l = \begin{cases}
u R, & l = l', \\
u r, & l \neq l',
\end{cases} \quad (8)
\]

\[
W_k \cdot U_l = \begin{cases}1, & l = l', \\
Q, & l \neq l'.
\end{cases} \quad (9)
\]

Then the generalization error can be derived by a calculation similar to that leading to Eq. (6). The result is

\[
\epsilon(W_k) = 1 - \int Dz \left[ H \left( z - u(R - r) \right) \right]^{K-1} .
\]

(10)

The values of \( R, r, \) and \( Q \), and hence that of \( \epsilon(W_k) \), depend on the training algorithm used to determine the \( W_k \) from the examples, as will be discussed below.
The first training algorithm that we study is a stochastic training dynamics that results in an equilibrium Gibbs distribution in the space of \{W_k\}. The energy function of this distribution is the number of classification errors made on the examples in the training set. The temperature is an independent parameter characterizing the training process. Here we focus on the limit of zero temperature, also called zero-T learning. A zero-T learner chooses at random a set of \{W_k\} that minimizes the classification error on the training examples. In the separation limit, the relevant regime is that of a small number of examples, so that the minimal training error is zero. In this case, zero-T learning corresponds to choosing uniformly at random from those sets of \{W_k\} that are consistent with all given examples.

For any finite \(N\), the values of \(R\), \(r\), and \(Q\) (and hence also the value of \(\epsilon\)) will depend on the particular choice for random examples. However, in the thermodynamic limit, they are self-averaging, and therefore their typical values can be derived by quenched average of the \(P\) examples \((\mathbf{x}^\mu, l^\mu)\), \(\mu = 1, \ldots, P\), over the probability distribution \(P(\mathbf{x}^\mu, l^\mu) = P(\mathbf{x}^\mu | l^\mu)P(l^\mu)\). This average is denoted by \(\langle \cdots \rangle\). Furthermore, in the thermodynamic limit the average values of these order parameters can be evaluated by maximizing the average entropy per weight \(S(Q, R, r) = \langle \ln Z \rangle / NK\). Here \(Z\) is the volume of the space of all sets of \(K\) unit length vectors \(\mathbf{W}\) with zero training error, and having the values \(Q, R, r\) for the corresponding overlaps.

Even in the thermodynamic limit, the evaluation of the quenched average \(\langle \cdots \rangle\) of \(\ln Z\) is complicated, since it usually involves not only the mean values of the \(W_m\), but also their fluctuations due to the randomness of the training set. However, we have found that in the separation limit, the quenched average simplifies greatly. This is because in the limit \(u \to \infty\) most of the learning occurs at small \(\alpha\) where the fluctuations of the weights due to different samplings of inputs from each cluster are negligible. Specifically we obtain that in the large \(u\) limit

\[
S = K^{-1} \sum_{i=1}^{K} S_i, \quad S_i = N^{-1} \ln \langle Z \rangle_i. \tag{11}
\]

Here \(\langle \cdots \rangle_i\) denotes averaging over all sets of \(P\) inputs with \(l^\mu = l\). Note that the \(S_i\) are computed by the annealed average where \(Z\) instead of \(\ln Z\) is averaged. The annealed averages are much simpler to compute than the quenched ones, but in general are only an uncontrolled approximation to the correct quenched theory. Here we have found that in the separation limit, the annealed approximation within each cluster becomes exact [10].

Focusing on the separation limit, we have evaluated Eq. (11) in the small \(\alpha\) limit. Although the overlaps \(R\) and \(r\) are of order \(1/u\), they cannot be neglected, since they appear in Eq. (10) through the product \(u(R - r)\), which is of order unity. The overlap \(Q\) between different vectors can be neglected, since \(Q \sim R^2\). Maximizing the entropy

\[
s = -\frac{R^2}{2K} - \frac{(K-1)^2}{2K} + \alpha \ln \int Dz \{H[z - u(R - r)]\}^{K-1} \tag{12}
\]

we find that \(r = -R/(K - 1)\), and that \(uR\) is a function of the scaled variable \(\tilde{\alpha}\), where

\[
\tilde{\alpha} = \alpha u^2. \tag{13}
\]

Substituting into Eq. (10) yields the generalization error

\[
\epsilon(\tilde{\alpha}) = 1 - \int Dz \{H[z - \tilde{R}(\tilde{\alpha})]\}^{K-1}, \tag{14}
\]

where \(\tilde{R} \equiv uR/(K - 1)\) is of order 1 when \(\tilde{\alpha}\) is of order 1. The function \(\epsilon(\tilde{\alpha})\) is a monotonically decreasing function of \(\tilde{\alpha}\), and approaches zero for large \(\tilde{\alpha}\) as

\[
\epsilon(\tilde{\alpha}) \approx \left(\frac{K-1}{K^2}\right)^{-\frac{1}{\tilde{\alpha}}} \tilde{\alpha} \to \infty. \tag{15}
\]

Thus in the case of zero-T learning the separation limit is defined by the limit \(u \to \infty\), \(\alpha \to 0\), while \(\tilde{\alpha} \equiv \alpha u^2\) is held fixed. In this limit the residual error (6) of the classification due to the overlap between the Gaussians is negligible. The error is dominated by the smallness of \(\alpha\), which allows only a poor estimate of the clusters’ boundaries.

Another possible learning algorithm estimates the location of the cluster centers by the sample mean of the examples generated by each of them,

\[
W_k = \frac{K}{P} \sum_{\mu=1}^{P} \mathbf{x}^\mu \delta_{kl^\mu}. \tag{16}
\]

In the context of the WTA network, this rule is equivalent to a Hebb learning rule where the weights of the \(k\)th neuron are a linear sum of the inputs arriving from the \(k\)th cluster. The Hebb rule is interesting because of its simplicity and biological plausibility. This learning rule is also equivalent to the classical algorithm of maximum likelihood estimation [7] of the cluster centers.

To evaluate the performance of the Hebb rule, we note that \(W_k\), as defined in Eq. (16), is a Gaussian variable with mean equal to \(U_k\) in the thermodynamic limit. Its components are uncorrelated and each has variance \(P^{-1}\).

If we normalize each weight vector (16) to the spherical constraint and then calculate the overlaps in (8), we find that \(R = u/\sqrt{u^2 + \alpha^{-1}}\), and \(Q = r = 0\). This yields

\[
\epsilon(\alpha) = 1 - \int Dz \left[ H\left(z - \frac{u^2}{\sqrt{u^2 + \alpha^{-1}}}ight) \right]^{K-1}. \tag{17}
\]

Note that as \(\alpha \to \infty\), this approaches the optimal error (6).

The preceding formula was valid for all \(\alpha\) and \(u\). Defin-
ing the separation limit to be \( \alpha \to 0 \), \( u \to \infty \), with

\[
\hat{\alpha} = au^4
\]  

(18)

held constant, we obtain the form

\[
\epsilon(\hat{\alpha}) = 1 - \int Dz \left[ H(z - \sqrt{\hat{\alpha}}) \right]^{K-1}.
\]  

(19)

At large \( \hat{\alpha} \), this becomes

\[
\epsilon(\hat{\alpha}) = (K-1) \frac{e^{-\hat{\alpha}/2}}{\sqrt{2\pi \hat{\alpha}}}. \]

(20)

In summary, we have derived a scaling form for the generalization error,

\[
\epsilon(\alpha, u) \approx \epsilon(\alpha u^p),
\]  

(21)

where \( p = 2 \) for zero-temperature learning, and \( p = 4 \) for the Hebb rule. The formula is exact in the separation limit \( \alpha \to 0 \), \( u \to \infty \), \( \alpha u^p = \text{const} \), and should also be a good approximation for large but finite separation. The number of examples required to obtain a fixed level of error (of order 1) scales like

\[
P \sim \frac{NK}{u^p}. \]

(22)

These scaling laws show how the sample complexity of learning depends on the form of the input distribution. Although our results are for the particular problem of classification of Gaussian mixtures using two specific learning algorithms, we believe that similar scaling forms hold for more general input distributions and learning algorithms. For example, it can be shown that the maximal stability algorithm [11] leads to a scaling form (21) with \( p = 4 \), much like the Hebb rule [10]. We also believe that such results might lead to a better understanding of the relation between learning theory and applications.

An important consequence of our results is that for well-separated inputs, the simple Hebb rule is far superior to zero-\( T \) learning, since it achieves the same generalization performance with a training set that is a factor of \( 1/u^2 \) smaller. A similar effect has been observed for learning a perceptron rule with a uniform input distribution [12], but in that model the superiority of the Hebb rule is not as marked.

In conclusion, it should be noted that the behavior of the system in the separation limit is completely different from the predictions based on the asymptotic behavior of the generalization error in the limit of large sample size, i.e., a large number of examples (with other parameters held fixed). To demonstrate this point, we have solved the case \( K = 2 \) for all values of \( \alpha \) and \( u \) by taking the \( T \to 0 \) limit of the finite temperature statistical mechanics [2]. Taking the large \( \alpha \) limit (holding \( u \) fixed) we find

\[
\epsilon(\alpha) = H(u/\sqrt{2}) + \frac{\sqrt{2}}{3\pi^{1/4}} \frac{e^{-u^2/8}}{\sqrt{\alpha u}}, \quad \alpha u \gg e^{u^2/4}. \]

(23)

This result is probably only approximate, since it comes from the replica symmetric solution, which may be unstable. The result for the Hebb learner can be obtained from Eq. (17),

\[
\epsilon(\alpha) = H(u/\sqrt{2}) + \frac{1}{2\sqrt{\pi}} \frac{e^{-u^2/4}}{\alpha u}, \quad \alpha u \gg 1. \]

(24)

Comparing these results with the behavior in the separation limit, e.g., Eqs. (15) and (20), it is evident that in cases where \( u \) is large the usual limit of large sample size is irrelevant, since it holds only in a region where the error is extremely small. Our results suggest that the limit of large sample size, the traditional focus of statistical learning theory, may sometimes be irrelevant for understanding the learning of realistic tasks.

We thank D. Haussler, M. Kearns, and N. Tishby for very helpful discussions.