Learning from Examples in Large Neural Networks

H. Sompolinsky\(^{(a)}\) and N. Tishby

\textit{AT&T Bell Laboratories, Murray Hill, New Jersey 07974}

H. S. Seung

\textit{Department of Physics, Harvard University, Cambridge, Massachusetts 02138}

(Received 29 May 1990)

A statistical mechanical theory of learning from examples in layered networks at finite temperature is studied. When the training error is a smooth function of continuously varying weights the generalization error falls off asymptotically as the inverse number of examples. By analytical and numerical studies of single-layer perceptrons we show that when the weights are discrete the generalization error can exhibit a discontinuous transition to perfect generalization. For intermediate sizes of the example set, the state of perfect generalization coexists with a metastable spin-glass state.

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

Understanding how systems can be efficiently trained to perform tasks is of fundamental importance. A central issue in learning theory is the rate of improvement in the processing of novel data as a function of the number of examples presented during training, i.e., the generalization curve.\(^{1-7}\) Numerical results on training in layered neural networks indicate that the generalization error improves gradually in some cases, and sharply in others.\(^{3,5,8,9}\) In this work we use statistical mechanics to study generalization curves in large layered networks. We will first discuss the general theory and then present results for learning in a single-layer perceptron.

The computational function of layered neural networks is described in terms of the input-output relations that they generate. We consider here a multilayer network with \(M\) input nodes, whose states are denoted by \(S_i, i=1,\ldots,M\), and a single output node denoted by \(\sigma=\sigma(W,S)\) where the \(W_i, i=1,\ldots,N\), denote the synaptic weights of the network. The network is trained by adjusting its weights to approximate or reproduce, if possible, a target function \(\sigma_0(S)\) on the input space. This is achieved by providing a set of examples consisting of \(P\) input-output pairs \((S_i',\sigma_0(S_i'))\), \(i=1,\ldots,P\). We assume that the inputs \(S_i\) are chosen at random from the entire input space.

The training process is often described as the minimization of a training energy

\[
E(W) = \sum_{i=1}^{P} \epsilon(W;S_i').
\]

where \(\epsilon(W;S)\) is some measure of the deviation of the network output \(\sigma(W;S)\) from the target output \(\sigma_0(S)\), e.g., \(\epsilon(W;S)=[\sigma(W;S) - \sigma_0(S)]^2\). We consider a stochastic training process that leads at long times to a Gibbs distribution of networks

\[
P(W) = Z^{-1} \exp[-\beta E(W)],
\]

where

\[
Z = \int dW \exp[-\beta E(W)], \quad dW = \prod_{i=1}^{N} dW_i P_0(W),
\]

and \(P_0(W)\) is the \textit{a priori} measure on the weight space. The temperature \(T=\beta^{-1}\) denotes the level of stochastic noise in the training. In the limit \(T \to 0\) the training corresponds to finding the global minimum of \(E\).

The training energy \(E\) depends on the fixed random choice of examples \(S_i'\). Focusing on average quantities, we will first average over \(P(W)\) with fixed \(S_i'\). This \textit{thermal} average will be denoted by \(\langle \cdots \rangle_T\). We will then perform a \textit{quenched average}, i.e., average over the distribution of example sets, \(\langle \cdots \rangle = \int dS' \Pi \Pi dS'.\) Here \(dS\) represents the normalized measure over the space of inputs. The average training error is \(E(T,P) = P^{-1} \times \langle \langle (E(W))_T \rangle \rangle\). The performance of a network on the whole input space is measured by the \textit{generalization error}, defined as \(\epsilon(W) = \int dS \sigma(W,S)\). The average generalization error, after training with \(P\) examples, is \(\epsilon_T(T,P) = \langle \langle \epsilon(W) \rangle_T \rangle \rangle\). The deviations of the typical values of these quantities from their thermal and quenched averages are expected to vanish as \(N \to \infty\).

We apply the replica method\(^{10}\) to evaluate quenched averages. The average free energy \(F\) is given by

\[
-\beta F = \lim_{n \to 0} \ln \mathbb{E} \left[ (\langle Z^n \rangle - 1) \right].
\]

Using Eq. (1) we find

\[
\langle Z^n \rangle = \int d\sigma \prod_{\alpha=1}^{n} dW_{\sigma} \exp( - P G(W_{\sigma}), \rangle \rangle,
\]

(2)

where

\[
G = -\ln \int dS \exp \left[ - \beta \sum_{\sigma=1}^{n} \epsilon(W_{\sigma}, S) \right].
\]

The variables \(W_{\sigma}\) represent the weights of \(n\) copies (replicas) of the system. The importance of the form (2) lies in the fact that the effective (nonrandom) Hamiltonian \(G\) is intensive, and does not depend on the number of examples \(P\). A direct consequence is that the correct thermodynamic limit \((N \to \infty)\) is achieved when the number of examples scales as the total number of (independently determined) weights, i.e., \(P = aN\). This scaling guarantees that both the entropy and the energy...
are proportional to $N$.

It is instructive to consider the high-temperature limit of the theory. Expanding Eq. (3) in powers of $\beta$ we find the form

$$ P_G = N \left\{ a_\beta \sum_{x=1}^n \epsilon(W, \sigma) + a_\beta^2 \sum_{x=0}^n g(W, W, \sigma) + O(\beta^3) \right\}, $$

(4)

which is similar to the replica spin-glass Hamiltonian.\textsuperscript{10} The first term represents the nonrandom part of the energy. The second term represents coupling between replicas induced by the randomness of the examples. Keeping only the leading order in $\beta$ yields

$$ P(W) = Z^{-1} \exp[-N \beta e(W)], \quad \beta \text{ small}. $$

(5)

It should be emphasized that even in the high-$T$ limit $e_G$ can decrease to $e_{min}$ since by increasing $a$, the effective temperature of the system, $T/a$, can be made small.

Equation (5) implies that learning at high $T$ is simple in that the effects of random sampling of the examples are negligible, and the training error is identical to the generalization error. Furthermore, the training process should be relatively fast, since the effective energy function $e_G(W)$ is a smooth function of $W$. The price is the necessity of a large training set as $a$ must be at least of order $T$.\textsuperscript{11}

As the temperature is lowered, the fluctuating contributions to the training energy, represented by the second- and higher-order terms in Eq. (4), become increasingly important. As a result the average generalization error becomes increasingly higher than the average training error. In general, these fluctuations can also affect the asymptotic behavior of $e_G$. In addition, the dynamics at low temperatures will be slowed down by the emergence of spurious minima separated by barriers of considerable height. Depending on the details of the problem, a genuine spin-glass phase may appear at low temperatures. In such a phase, the barrier heights separating the many stable states diverge with $N$, breaking the ergodicity of the system.

The general form (2) can be used to study the behavior at large values of $a$. As $a$ increases to infinity the dominant contribution to the integrals over the weight space in Eq. (2) comes from the minimum of $G$. Denoting by $W^0$ the values of the $W_\sigma$ at this minimum and taking the $n \rightarrow 0$ limit one has

$$ n^{-1} G_{min} = -n^{-1} \ln \int dS \exp[-\beta e(W^0, S)] = \beta e(W^0). $$

This implies that for any temperature both the training error and the generalization error approach the optimal value $e_{min} = e(W^0)$ as $a \rightarrow \infty$.

A particularly simple case is when the weights vary continuously and $e(W, S)$ is an analytic function of $W$ near $W^0$. Then the dominant deviations of $G$ from $G_{min}$ (in the $a \rightarrow \infty$ limit) are quadratic in $W_\sigma - W^0$. Expanding in powers of $W_\sigma - W^0$ we obtain\textsuperscript{12}

$$ e_G(T, a) = e_{min} + (T + A)/2a. \quad a \rightarrow \infty, $$

(6)

where $A = N^{-1} \text{Tr}(U W^{-1})$. Here $V$ and $U$ are the $N \times N$ matrices $\sigma_i, \sigma_j e(W^0)$ and $\int dS \sigma_i e(W^0, S) \times \sigma_i e(W^0, S)$, respectively, where $\sigma_i = \delta_{i, j} W_i$. We will show below that when the weights are discrete valued the falloff of $e_G(a)$ can be much sharper than $a^{-1}$. Furthermore, even when the asymptotic behavior of $e_G(a)$ is given by Eq. (6), its dominant feature may actually be a sharp drop at $a$ of order 1.

We consider a single-layer perceptron which implements a binary function $f = \text{sgn}[S \cdot W]$, where $X \cdot Y \equiv N^{-1} \sum_{i=1}^N X_i Y_i$. The task is to learn the input-output relation generated by a teacher perceptron with weights $W^t$.\textsuperscript{7,13} The training examples are input patterns $[S']$ paired with their corresponding labels $\sigma_0 (S') = \text{sgn}[S' \cdot W^t]$. The input states $S'$ are chosen according to a Gaussian distribution with variance unity. We assume here that the teacher weights are binary valued ($W^t = \pm 1$) and the $a$ priori distribution of the weights of the trained system is

$$ P_0(W) = \prod_{i=1}^N \left\{ \exp \left[ -\frac{(W_i - 1)^2}{2a^2} \right] + \exp \left[ -\frac{(W_i + 1)^2}{2a^2} \right] \right\} \delta(W \cdot W - 1). $$

(7)

The natural measure of error for a threshold function is $e(W, S^t) = \theta(-S^t \cdot W^t)$, where $\theta(x)$ is the Heaviside function, $\theta(x) = 1$ for $x > 0$, and $\theta(x) = 0$ for $x \leq 0$. The training energy, Eq. (1), in this case is the number of misclassified examples in the training set. Averaging the error over the distribution of $S^t$ yields the following expression for the generalization error for the perceptron $W$:

$$ e(W) = e(m) = \pi^{-1} \arccos m, $$

(8)

where $m \equiv W \cdot W^t$ is the overlap between the network weight vector and the teacher.

First, we discuss the high-temperature theory. From Eq. (4) it follows that in the high-$T$ limit, the training energy for a network with overlap $m$ is $Na e(m)$. The free energy per weight for a fixed $m$ is

$$ f = (a/\pi) \arccos m - s(m), $$

(9)

where the entropy function $s(m)$ depends on the specific constraints on the parameters $W_i$ and $W^t_0$. The parameter $\bar{a}$ is the inverse of the effective temperature $\bar{a} \equiv a/T$. The equilibrium state in the limit of large $N$ is determined by minimizing $f$ with respect to $m$.

We first consider the limit $a \rightarrow 0$, which corresponds to a binary perceptron $W_i = \pm 1$. This limit has been studied by Gardner and Derrida.\textsuperscript{13} For that case the en-
entropy is simply
\[ s(m) = -\frac{1}{2} [(1 - m) \ln(1 - m) + (1 + m) \ln(1 + m)] + \ln 2. \]

The resultant free energy has a local minimum at \( m = 1 \), for all values of \( \alpha \). For \( \alpha < 2.08 \) there is another local minimum, at \( m_0 < 1 \), given by the equation
\[ m_0 = \tanh \left[ \alpha/\pi (1 - m_0^2)^{1/2} \right]. \quad (10) \]

For \( \alpha < 1.70 \) the state \( m_0 \) is the global minimum of \( f \). For \( 1.70 < \alpha < 2.08 \) it is only a local minimum. However, starting from random initial weights \( (m \approx 0) \) the system will converge fast to \( m = m_0 \), and will reach the state \( m = 1 \) only at exponentially long times. Above \( \alpha = 2.08 \) there is no local minimum at \( m < 1 \) and the system will converge fast to the state \( m = 1 \).

The average generalization error as a function of \( \alpha \) is shown in Fig. 1. At \( \alpha = 2.08 \) there is a discontinuous drop of \( e_g \) from 0.2 to zero, i.e., a transition from poor to perfect generalization. We also present the training and generalization errors from Monte Carlo simulations of the binary perceptron with \( N = 75 \), fixed temperature \( T = 5 \), and \( \alpha \) up to \( \alpha = 4 \). They agree well with the predictions of the high-\( T \) theory.

The present discontinuous (first-order) transition is unusual in that the system collapses to the energy ground state \( W = W^0 \) even at finite temperature. The origin of this behavior is the singularity of \( e(m) \sim \sqrt{1 - m} \) near \( m = 1 \), implying that the energy of perceptrons having only one weight different in sign from the teacher is \( e(1 - 2/N) \approx 1/\sqrt{N} \). At high \( \alpha \), this large gap overwhelms any entropic terms, and the system remains in the ground state. This differs from the usual behavior of discrete systems, where the energy gap per degree of freedom is typically of order \( 1/N \). Qualitatively similar results hold in the more general case, e.g., a perceptron with a nonzero threshold.

In the case of continuous-valued weights, i.e., \( a \) finite, the asymptotic form of \( e_g \) is \( 1/a \). However, for \( a < 0.26 \), \( e_g \) still exhibits a discontinuous drop at critical value of \( a \), as shown in Fig. 1. This demonstrates that even networks with continuous weights can exhibit a sharp decrease of \( e_g \) at \( a \) of order 1.

Although the above results were derived using the high-\( T \) limit, the qualitative picture remains valid even at lower temperatures. We have solved the replica theory of the binary perceptron with an energy function of the form of Eq. (1) for all \( a \) and \( T \). The complete phase diagram is shown in the figure inset. The discontinuous transition to perfect generalization at finite \( a \) exists at all temperatures. At \( T = 0 \), for \( a < 1.24 \) there are states with \( m < 1 \) that have zero training error, i.e., they are degenerate with the \( m = 1 \) state. The entropy of these states, which is proportional to \( N \), decreases to zero as \( a \to 1.24 \). Above \( a = 1.24 \) the only ground state, i.e., state with zero training error, is the \( m = 1 \) state. However, for \( 1.24 < a < 1.63 \) metastable states with \( m_0 < 1 \) and positive training error exist. Above \( a = 1.63 \) the only stable state at \( T > 0 \) is that with \( m = 1 \), although strictly at \( T = 0 \) states that are stable to flips of single weights are expected to be present even at higher \( a \).

In contrast to the high-\( T \) limit, in the darker region of the phase diagram the metastable state represents a spin-glass phase. The presence of this phase implies that there is an enormous number of metastable states separated by energy barriers which diverge with \( N \), rendering the convergence to \( m = 1 \) extremely slow. In fact, our numerical simulations of small sizes (\( N = 25 \)) show a substantial slowing down already at \( T \approx 1 \).

The properties of the spin-glass phase were derived within the replica symmetry-breaking spin-glass theory, and will be reported elsewhere. A unique feature of this phase is that thermal fluctuations are essentially frozen within this phase, and the entropy is zero. The value of the energy, i.e., the training error, as well as \( m \), is independent of \( T \) for fixed \( a \). A similar frozen spin-glass phase has been found for the storing of random patterns in a binary perceptron.

In conclusion, we have shown that when the training energy is a smooth function of real-valued weights, the average generalization error \( e_g(a) \) falls off asymptotically as \( 1/a \). When the spectrum of the training energy is discrete, the asymptotic behavior depends on the size of
its gap for large $a$. In the case of the binary perceptron where the gap per example is of order $1/\sqrt{N}$ we have found a discontinuous transition at a critical $a$ from poor to perfect generalization. This transition reflects the fact that there are no networks other than that with $e_g = 0$ that have small generalization error and are consistent with the examples. A discontinuous transition to a perfect generalization is probably a general feature of learning in networks with a gap per example that is larger than order $1/N$. When the gap vanishes like $1/N$ or faster the asymptotic behavior of $e_g$ may depend on the details of the model as well as on the temperature during training. Previous studies indicate that to obtain an optimal $e_g(a)$ requires tuning $T$ with $a$. The discreteness of the weights of the system is not sufficient to guarantee the sharp transition in the learning process. Particularly, we find that when the minimal generalization error that can be achieved with a given architecture is positive the improvement in the generalization may be gradual even if the network is discrete.

Helpful discussions with P. C. Hohenberg and D. A. Huse are gratefully acknowledged. We also thank Thinking Machines, Inc. and Carl Diegert at Sandia National Laboratories for allowing the use of their Connection Machines. One of us (H.S.S.) acknowledges support from NSF Grant No. DMR88-17291 and the Harvard Materials Research Laboratory, and the hospitality of AT&T Bell Laboratories.

Note added.—Recently, Gy"orgyi has independently studied the first-order transition in binary-perceptron learning at $T = 0$.  

\(\text{VOLUME 65, NUMBER 13} \quad \text{PHYSICAL REVIEW LETTERS} \quad 24 \text{ SEPTEMBER 1990}\)

\(\text{[Proc.}\ 22\text{, 1983 (1989).} \)

\(\text{[Proc.}\ 50\text{, 3057 (1989).} \)

\(\text{[Proc.}\ 41\text{, 7097 (1990).} \)

\(\text{[Proc.}\ 27\text{, 1134 (1984); Philos.} \)

\(\text{[Proc.}\ 312\text{, 441 (1984).} \)

\(\text{[Proc.}\ 82\text{, 1983 (1989).} \)

\(\text{[Proc.}\ 8\text{, 1983 (1989).} \)

\(\text{[Proc.}\ 4, 1199 (1987).} \)

\(\text{[Proc.}\ 4, 503 (1987).} \)

\(\text{[Proc.}\ 4, 441 (1984).} \)

\(\text{[Proc.}\ 2, 403 (1989)]; \}

\(\text{[Proc.}\ 1, 1983 (1989).} \)

\(\text{[Proc.}\ 1, 1983 (1989).} \)

\(\text{[Proc.}\ 1, 1983 (1989).} \)

\(\text{[Proc.}\ 1, 1983 (1989).} \)
FIG. 1. Generalization curves $e_p(\bar{a})$ for the perceptron at different values of the width parameter $a$ defined in Eq. (7), calculated by the high-$T$ theory. Shown also are data points for the training and generalization errors from Monte Carlo simulations for the binary perceptron with $N=75$ and $T=5$. Inset: Phase diagram of learning in the Ising perceptron in the $(T,a)$ plane. At high $T$ and small $a$ the equilibrium value of the overlap with the teacher, $m$, is small, implying a high generalization error. In the shaded area the state with lowest free energy is at $m=1$, but other metastable states exist. Beyond the solid line the only stable state (at $T>0$) is that with $m=1$. The darker region marks a spin-glass metastable phase that appears at low temperatures.