

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/2984351>


# A Statistical Approach to Learning and Generalization in Layered Neural Networks

**Article** in *Proceedings of the IEEE* · November 1990  
DOI: 10.1109/5.58339 · Source: IEEE Xplore

CITATIONS  
218


READS  
451

3 authors:




**Esther Levin**  
Point 72 asset management  
**60** PUBLICATIONS **3,034** CITATIONS

SEE PROFILE



**Naftali Tishby**  
Hebrew University of Jerusalem  
**245** PUBLICATIONS **11,024** CITATIONS


SEE PROFILE




**Sara A Solla**  
Northwestern University  
**94** PUBLICATIONS **4,704** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Feature selection and extraction [View project](#)



Information constrained control and Reinforcement Learning [View project](#)

# A Statistical Approach to Learning and Generalization in Layered Neural Networks

ESTHER LEVIN, NAFTALI TISHBY, AND SARA A. SOLLA

*A general statistical description of the problem of learning from examples is presented. Our focus is on learning in layered networks, which is posed as a search in the network parameter space for a network that minimizes an additive error function of statistically independent examples. By imposing the equivalence of the minimum error and the maximum likelihood criteria for training the network, we arrive at the Gibbs distribution on the ensemble of networks with a fixed architecture. Using this ensemble, the probability of correct prediction of a novel example can be expressed, serving as a measure of the network's generalization ability. The entropy of the prediction distribution is shown to be a consistent measure of the network's performance. This quantity is directly derived from the ensemble statistical properties and is identical to the stochastic complexity of the training data. Our approach is a link between the information-theoretic model-order-estimation techniques, particularly minimum description length, and the statistical mechanics of neural networks. The proposed formalism is applied to the problems of selecting an optimal architecture and the prediction of learning curves.*

## 1. INTRODUCTION

Layered neural networks are nonlinear parametric models that can approximate any continuous input-output relation [1], [2]. The quality of the approximation depends on the architecture of the network used, as well as on the complexity of the target relation. The problem of finding a suitable set of parameters that approximate an unknown relation  $F$  is usually solved using supervised learning algorithms. Supervised learning requires a training set, that is, a set of input-output examples related through the relation  $F$ , as formalized by Valiant [3]. Learning the training set is often posed as an optimization problem by introducing an error measure. This error is a function of the training examples as well as of the network parameters, and it measures the quality of the network's approximation to the relation  $F$  on the restricted domain covered by the training set. The minimization of this error over the network's parameter space is called the training process. The task of learning, however, is to minimize that error for all possible examples related through  $F$ , namely, to generalize.

In this work we focus on a statistical description of the learning process by dealing with the ensemble of all net-

works with the same parameter space. The common method of parameter estimation is the maximum likelihood (ML) approach. By imposing the equivalence of the error minimization and the likelihood maximization we arrive at the Gibbs distribution on the parameter space for a canonical ensemble (see, for example, [4]) of networks with the given architecture. This distribution is interpreted as the post-training distribution, where the probability of arriving at a specific network decreases exponentially with the error of the network on the training examples. The imposed equivalence condition leaves only one free parameter for the training process: the ensemble temperature, which determines the level of acceptable training error as well as the level of stochasticity in the training algorithm. The normalization integral of the Gibbs distribution (that is, the partition function) measures the weighted volume of the configuration space of trained networks, and its functional form determines the average training error as well as the information gained (entropy) during training.

The training process selects network configurations that perform well on the restricted domain defined by the training examples. It is well known, however, that there can be little connection between the training error, restricted to the training set, and the network's ability to generalize outside of that set [5]–[7]. It is generally possible to get increasingly better performance on the training examples by increasing the complexity of the model, but such a procedure does not necessarily lead to a better generalization ability. Using the Gibbs posttraining distribution and the likelihood that results from the equivalence condition, we are able to express the probability of predicting a novel independent example. We show that this prediction probability induces a consistent measure of generalization, which can be viewed as an application to layered networks of the predictive minimum description length method (MDL) as proposed by Rissanen [8].

The problem of estimating the sufficient training set size for learning with layered networks has been previously discussed [9] in a distribution-free, worst case analysis. Our work is a step toward a typical case theory of generalization with layered networks. Though the general principles described can be applied to a wider class of parametric models, they are of special interest in the context of statistical mechanics of "neural networks."

Manuscript received September 25, 1989; revised March 21, 1990.  
E. Levin and N. Tishby are with AT&T Bell Laboratories, Murray Hill, NJ 07974, USA.  
S. A. Solla is with AT&T Bell Laboratories, Holmdel, NJ 07733, USA.  
IEEE Log Number 9039174.

### A. Layered Networks

We would like to "learn," or model, an input-output relation  $F$  by a feedforward layered network, consisting of  $L$  layers of processing. The architecture is fixed and determined by the number  $\{N_l, 0 \leq l \leq L\}$  of processing elements per layer and by their connectivity. The elements of the  $(l+1)$ th layer are connected to the previous layer and their state is determined through the recursion relation

$$\begin{cases} u_i^{(l+1)} = \sum_{j=1}^{N_l} w_{ij}^{(l+1)} v_j^{(l)} + w_i^{(l+1)} \\ v_j^{(l+1)} = h(u_j^{(l+1)}), \quad 1 \leq i \leq N_{l+1}. \end{cases} \quad (1)$$

The input to the  $i$ th element of the  $l$ th layer  $u_i^{(l)}$  determines the state  $v_i^{(l)}$  of the element through a sigmoid nonlinearity, such as  $h(x) = 1/(1 + \exp(-x))$ . The parameters of the network are the connections  $\{w_{ij}^{(l)}, 1 \leq j \leq N_{l-1}, 1 \leq i \leq N_l, 1 \leq l \leq L\}$  and the biases  $\{w_i^{(l)}, 1 \leq i \leq N_l, 1 \leq l \leq L\}$ , corresponding to a point  $\omega$  in the  $D = \sum_{l=1}^L N_l(N_{l-1} + 1)$  dimensional Euclidean space  $\mathbf{R}^D$ . For every point  $\omega$  in the network configuration space  $\mathbf{W} \subset \mathbf{R}^D$ , the network (1) is a realization of a deterministic mapping from an input  $x \in X \subset \mathbf{R}^p$  to an output  $y \in Y \subset \mathbf{R}^q$ , provided that  $N_0 = p$  and  $N_L = q$ . We denote this mapping by  $y = F_\omega(x)$ ,  $F_\omega: \mathbf{R}^p \times \mathbf{W} \rightarrow \mathbf{R}^q$ . In what follows we discuss the ensemble of all networks in the configuration space  $\mathbf{W}$ .

We focus on the problem of learning an unknown input-output relation  $F$  from examples: a training set of  $m$  input-output pairs, related through the unknown relation  $F$ ,  $\xi^{(m)} \equiv \{\xi_i, 1 \leq i \leq m\}$ , where  $\xi \equiv (x, y)$ ,  $x \in X \subset \mathbf{R}^p$ , and  $y \in Y \subset \mathbf{R}^q$ . The relation  $F$  can be generally described by the probability density function defined over the space of input-output pairs  $X \otimes Y \subset \mathbf{R}^{p+q}$ :  $P_F(\xi) = P_F(x)P_F(y|x)$ , where  $P_F(x)$  defines the region of interest in the input space and  $P_F(y|x)$  describes the functional or the statistical relation between the inputs and the outputs. The training set consists of examples drawn independently according to this probability density function. Learning the training set by a layered network is posed as an optimization problem by introducing a measure of quality of the approximation of the desired relation  $F$  by the mapping  $F_\omega$  realized by the network. The additive error function

$$E^{(m)}(\omega) \equiv E(\xi^{(m)}|\omega) = \sum_{i=1}^m e(y_i|x_i, \omega) \quad (2)$$

measures the dissimilarity between  $F$  and  $F_\omega$  on the restricted domain covered by the training set. The error function  $e(y|x, \omega)$  is a distance measure on  $\mathbf{R}^q$  between the target output  $y$  and the output of the network on the given input  $x$ , that is,  $e(y|x, \omega) = d(y, F_\omega(x))$ .

## II. PROBABILITY INFERENCE IN THE NETWORK SPACE

### A. Error Minimization as a ML Approach

Our first goal is to introduce a statistical description of the training process. The statistical modeling problem is usually posed as one of finding a set of parameters  $\omega$  that maximizes the likelihood of the training set of  $m$  independent examples

$$\text{Max}_{\omega \in \mathbf{W}} P(\xi^{(m)}|\omega) = \prod_{i=1}^m P_F(x_i) \cdot \text{Max}_{\omega \in \mathbf{W}} \prod_{i=1}^m p(y_i|x_i, \omega), \quad (3)$$

where the conditional probability  $p(y|x, \omega)$  should be considered as a measure of the "reasonable expectation" of the compatibility of the pair  $(x, y)$  to the network  $\omega$ , rather than relative frequency in some sample space [10].

Our primary requirement is that the maximization of the likelihood (3) be equivalent to the minimization of the additive error (2), for every set of independent training examples  $\xi^{(m)}$ . These two optimization criteria can be equivalent only if they are directly related through an arbitrary monotonic and smooth function  $\phi$ , namely,

$$\prod_{i=1}^m p(y_i|x_i, \omega) = \phi \left( \sum_{i=1}^m e(y_i|x_i, \omega) \right), \quad (4)$$

assuming that the derivatives w.r.t.  $\omega$  vanish at the extreme points of both the likelihood and the error. The only solution to the functional equation (4) is given by [11], [12]

$$\begin{aligned} p(y|x, \omega) &= \frac{1}{z(\beta)} \exp[-\beta e(y|x, \omega)] \\ z(\beta) &= \int_Y \exp[-\beta e(y|x, \omega)] dy \end{aligned} \quad (5)$$

where  $\beta$  is a positive integration constant which determines the sensitivity of the probability  $p(y|x, \omega)$  to the error value. The mean error  $\bar{e} = \int e(y|x, \omega) p(y|x, \omega) dy$ , is a measure of the acceptable error level, and is related to  $\beta$  through

$$\bar{e} = -\frac{\partial \log z}{\partial \beta}; \quad \frac{\partial \bar{e}}{\partial \beta} < 0. \quad (6)$$

We assume that the normalization constant  $z(\beta)$  or, equivalently, the mean error  $\bar{e}$ , is not an explicit function of the specific network  $\omega$  or the input  $x$ . This assumption is justified considering that the integration in (5) is performed over all possible output values, and is rigorously correct if the error is invariant under translations in the range  $Y$ .

An important example is the quadratic error function  $e(y|x, \omega) = (y - F_\omega(x))^2$ . The resulting  $p(y|x, \omega)$  in this case is the Gaussian distribution

$$p(y|x, \omega) = (2\pi\sigma^2)^{-1/2} \exp[-(y - F_\omega(x))^2/(2\sigma^2)] \quad (7)$$

with  $\beta = 1/(2\sigma^2)$ . Then  $z(\beta) = \sqrt{(\pi/\beta)}$ , and  $\bar{e} = \sigma^2$ , both independent of the network  $\omega$  and the input  $x$ .

### B. The Gibbs Distribution

The conditional likelihood (3) can now be inverted using the Bayes formula to induce a distribution on the network configuration space  $\mathbf{W}$  given the set of input-output pairs  $\xi^{(m)}$

$$\rho^{(m)}(\omega) \equiv P(\omega|\xi^{(m)}) = \frac{\rho^{(0)}(\omega) \prod_{i=1}^m p(y_i|x_i, \omega)}{\int_{\mathbf{W}} \rho^{(0)}(\omega) \prod_{i=1}^m p(y_i|x_i, \omega) d\omega}, \quad (8)$$

where  $\rho^{(0)}$  is a nonsingular prior distribution on the configuration space.

Writing (8) directly in terms of the training error  $E(\xi^{(m)}|\omega)$ , we arrive at the "Gibbs canonical distribution" on the ensemble of networks

$$\rho^{(m)}(\omega) = \frac{1}{Z^{(m)}} \rho^{(0)}(\omega) \exp[-\beta E^{(m)}(\omega)], \quad (9a)$$

where the normalization integral

$$\mathbf{Z}^{(m)}(\beta) = \int_{\mathbf{W}} \rho^{(0)}(\omega) \exp [-\beta E^{(m)}(\omega)] d\omega \quad (9b)$$

is the error moment generating function, known in statistical mechanics as the partition function, and it measures the weighted accessible volume in configuration space. Equation (9) has a clear intuitive meaning as the *posttraining* distribution in  $\mathbf{W}$ : The probability of each point  $\omega$  is reduced exponentially with the error of the network on the training set  $\xi^{(m)}$ . Though this distribution may appear unlikely for some training methods, it arises naturally for stochastic algorithms, such as simulated annealing, [13] which essentially implement the Gibbs distribution in configuration space. It is the only distribution that corresponds directly to the error minimization, it is the most probable distribution for large networks, and it is well utilized in statistical mechanics (see, for example, [4]).

Learning the training examples results in a modification of the probability distribution over the network's parameters space. In the Gibbs formulation, training on an additional independent example  $\xi_{m+1}$  is equivalent to multiplying the distribution  $\rho^{(m)}$  by the factor  $\exp(-\beta e(y_{m+1} | x_{m+1}, \omega))$  and renormalizing, that is,

$$\begin{aligned} \mathbf{Z}^{(m+1)} &= \int_{\mathbf{W}} \rho^{(0)}(\omega) \exp [-\beta E^{(m)}(\omega) - \beta e(y_{m+1} | x_{m+1}, \omega)] d\omega \\ &\leq \int_{\mathbf{W}} \rho^{(0)}(\omega) \exp [-\beta E^{(m)}(\omega)] d\omega = \mathbf{Z}^{(m)}. \end{aligned} \quad (10)$$

Training thus results in a reduction of the weighted volume in configuration space or, equivalently, in a monotonic increase of the *ensemble free energy*  $\beta f \equiv -\log \mathbf{Z}^{(m)}$  with the size  $m$  of the training set. It is this free energy, as a function of the parameter  $\beta$ , which determines the average training error

$$\langle E^{(m)} \rangle = \int_{\mathbf{W}} \rho^{(m)}(\omega) E^{(m)}(\omega) d\omega = -\frac{\partial \log \mathbf{Z}^{(m)}}{\partial \beta} \geq 0, \quad (11)$$

as well as the ensemble fluctuations around this error

$$\frac{\partial \langle E \rangle}{\partial \beta} = -\frac{\partial^2 \log \mathbf{Z}}{\partial \beta^2} = -\langle (E - \langle E \rangle)^2 \rangle < 0. \quad (12)$$

The average training error is thus a decreasing function of the sensitivity parameter  $\beta$ , as expected.

### C. Information Gain and Entropy

An important characterization of the learning process is the amount of information gained during training. A common way of quantifying this information gain is by the statistical distance between the pre- and posttraining distributions, given by the relative entropy of the ensemble with respect to the prior distribution (Kullback-Leibler distance [14]), that is,

$$S^{(m)} \equiv D[\rho^{(m)} | \rho^{(0)}] = \int_{\mathbf{W}} \rho^{(m)}(\omega) \log \frac{\rho^{(m)}(\omega)}{\rho^{(0)}(\omega)} d\omega \geq 0. \quad (13)$$

The familiar thermodynamic relation

$$S^{(m)} = -\log \mathbf{Z}^{(m)} - \beta \langle E^{(m)} \rangle \quad (14)$$

identifies this quantity as the thermodynamic entropy, which decreases by reducing the weighted configuration

volume  $\mathbf{Z}^{(m)}$ , as well as by decreasing the training error  $\langle E^{(m)} \rangle$ . Equation (14) provides another meaning for the parameter  $\beta$ , namely the Lagrange multiplier for the constrained average training error  $\langle E^{(m)} \rangle$  during minimization of the relative entropy (13), and the inverse of  $\beta$  plays the role of the ensemble temperature, as in statistical mechanics.<sup>1</sup>

### D. Training Without Errors

The interesting case of error-free learning can now be recovered by taking the limit  $\beta \rightarrow \infty$ , that is, the zero temperature limit. The Gibbs measure in this limit is simply the prior distribution restricted to the zero error region in  $\mathbf{W}$ , namely,

$$\rho^{(m)}(\omega) = \frac{\rho^{(0)}(\omega) \prod_{i=1}^m \theta(\omega, \xi_i)}{\mathbf{Z}^{(m)}} \quad (15)$$

where  $\theta(\omega, \xi)$  is a "masking function" that is equal to 1 when the error  $e(y | x, \omega)$  is zero, and which vanishes elsewhere. Equation (15) becomes meaningless when the number of examples increases beyond the capacity [5], [9], [18], [19] of the network, that is, when the training set cannot be learned without errors due to noisy examples or a nonlearnable input-output relation  $F$ . In such cases, reaching the capacity is not an indication of the best generalization, which generally can be improved by training with finite error on additional examples. In this sense our formalism generalizes the work of Denker *et al.* [15] to the more common case of finite average training error [16].

The conditional probability (5) in this case is given by

$$p(y | x, \omega) = \mathbf{Z}^{-1} \theta(\omega, \xi) \quad (16)$$

suggesting another interpretation regarding the probability that the pair  $\xi$  was included in a training set leading to the network  $\omega$ .

## III. GENERALIZATION AND THE STOCHASTIC COMPLEXITY

### A. The Prediction Probability

The learning process selects network configurations that have small error on the restricted domain defined by the training examples. Whether the learning process leads to successful rule extraction—in that the resulting network configurations  $\omega$  implement the desired relation  $F$ —can only be tested through performance on *novel* patterns not belonging to the training set. It is generally possible to get increasingly better performance of the models on the training set, but such a procedure does not necessarily lead to better generalization.

The generalization ability can be measured by the probability that networks trained on  $m$  examples, given an input  $x$ , will correctly predict the output  $y$ , where the pair  $(x, y)$  is an independent sample from  $P_F$ . The quantity of interest is the conditional probability

$$\begin{aligned} P(y | x, \xi^{(m)}) &\equiv p^{(m)}(y | x) = \int_{\mathbf{W}} \rho^{(m)}(\omega) p(y | x, \omega) d\omega \\ &= \frac{\mathbf{Z}^{(m+1)}(\xi)}{\mathbf{Z}^{(m)} \mathbf{Z}} \end{aligned} \quad (17)$$

<sup>1</sup>The usual entropy  $S = -\int_{\mathbf{W}} \rho(\omega) \log \rho(\omega) d\omega$  is maximized by the Gibbs distribution, subject to the  $\langle E^{(m)} \rangle$  constraint.

where  $p^{(m)}(\omega) = P(\omega | \xi^{(m)})$  of Eq. (9) is the posttraining probability of network  $\omega$ , and  $p(y | x, \omega)$  of Eq. (5) is the probability that the new pair  $\xi$  is compatible with the network  $\omega$ . The prediction probability has a clear intuitive meaning, since the ratio  $(Z^{(m+1)}/Z^{(m)})$  describes the relative volume of networks that are compatible with all  $m + 1$  examples among those that are compatible with the  $m$  training examples.

Since  $Z^{(m)}$  is a monotonic function of  $m$ , the statistical prediction error [8]  $-\log p^{(m)}(\xi)$  follows from Eq. (17) and can be expressed as a free energy derivative

$$-\log p^{(m)} \approx -\frac{\partial \log Z^{(m)}}{\partial m} + \log z. \quad (18)$$

A reliable estimate of the generalization ability requires the prediction of a large number of independent points  $\xi^{(i)}$ , distributed according to the underlying probability function  $P_f(\xi)$ . The average statistical prediction error can be shown to be a consistent measure of the generalization ability using the Gibbs inequality, that is,

$$\begin{aligned} & -\frac{1}{T} \log \prod_{i=1}^T p^{(m)}(y_i | x_i) \\ &= -\frac{1}{T} \sum_{i=1}^T \log p^{(m)}(y_i | x_i) \xrightarrow{T \rightarrow \infty} \langle -\log p^{(m)} \rangle \\ &= -\int P_f(\xi) \log p^{(m)}(y | x) d\xi \\ &\geq -\int P_f(\xi) \log P_f(y | x) d\xi. \end{aligned} \quad (19)$$

This measure is called the *stochastic complexity* [8] and it was introduced by Rissanen for the purpose of estimating the generalization in statistical models. The maximal statistical generalization ability, or minimal stochastic complexity, is obtained if and only if the prediction probability  $p^{(m)}(y | x)$  of the trained networks equals  $P_f(y | x)$ . In that case the trained networks implement the underlying relation  $F$ .

Another important property of the statistical prediction error  $-\log p^{(m)}$  is that it is bounded by the generalization and training errors on the example  $\xi$ . Using the positivity of the relative entropies

$$\begin{aligned} \int_{\omega} p^{(m+1)}(\omega) \log \frac{p^{(m+1)}(\omega)}{p^{(m)}(\omega)} d\omega &\geq 0; \\ \int_{\omega} p^{(m)}(\omega) \log \frac{p^{(m)}(\omega)}{p^{(m+1)}(\omega)} d\omega &\geq 0 \end{aligned}$$

we obtain

$$\beta \langle e(y | x) \rangle_{p^{(m+1)}} \leq -\log p^{(m)}(y | x) - \log z \leq \beta \langle e(y | x) \rangle_{p^{(m)}} \quad (20)$$

with equalities if, and only if, the two errors are equal. Thus  $-1/\beta \log p^{(m)}$  is bounded by the more natural measure of generalization: the average error on examples not in the training set, that is, the *generalization error*. Note that minimizing the statistical prediction error amounts to minimizing the distance between the ensemble posttraining prediction distribution and the *true* distribution of the data. From (20) it is clear that this is a weaker form of generalization than minimizing the pretraining or the generalization error directly.

The stochastic complexity  $\langle -\log p^{(m)} \rangle$  has an additional meaning as the average number of bits required to encode

a novel example, given a system trained on the  $m$  examples. The network with minimal prediction error is thus the one that provides maximal average compression of the data. The training of networks that minimize the statistical prediction error is an application to layered networks of the principle of minimum stochastic complexity, also known as the minimum description length principle (MDL) [8].

## B. Sample and Ensemble Averages

The partition function and all the quantities derived from it are functions of the random choice of a specific training set, and as such are still random variables. The typical performance of the network must be estimated by averaging these random variables over all possible training sets of the given size  $m$ . This averaging, denoted by  $\langle \langle \rangle \rangle$ , should be done at the end with respect to the *external* measure  $P_f(\xi)$  and is different from the ensemble average over the Gibbs measure. As is evident from Eqs. (11), (12), and (18), the important function from which the interesting quantities are derived is the free energy  $-\log Z^{(m)}(\beta)$ . Due to the external nature of  $P_f(\xi)$ —it is independent of  $\beta$  or  $m$ —we can interchange partial derivatives with sample averaging, and the basic problem is reduced to the calculation of the *quenched* free energy  $\langle \langle \log Z(m, \beta) \rangle \rangle$ . This generally very difficult problem can be solved in special cases using the “replica method” [17]–[19], which has become an almost standard tool in the study of random systems in statistical physics. The basic equations of our theoretical framework are thus summarized by

$$\langle \langle \log Z(m, \beta) \rangle \rangle = \int P_f(\xi^{(m)}) \log Z(\xi^{(m)}, \beta) d\xi^{(m)} \quad (21a)$$

$$\langle \langle E^{(m)} \rangle \rangle = -\frac{\partial \langle \langle \log Z(m, \beta) \rangle \rangle}{\partial \beta} \quad (21b)$$

$$\langle \langle -\log p^{(m)} \rangle \rangle \approx -\frac{\partial \langle \langle \log Z(m, \beta) \rangle \rangle}{\partial m} + \log z(\beta) \quad (21c)$$

with

$$P_f(\xi^{(m)}) d\xi^{(m)} = \prod_{i=1}^m P_f(\xi_i) d\xi_i.$$

## C. The Role of $\beta$

The parameter  $\beta$  plays an important role in our framework. It is the ensemble inverse temperature and it controls the amount of stochasticity in the training algorithm. In addition, the presence of finite  $\beta$  regularizes the error-free case and allows the formal derivation of the training and prediction errors from the ensemble-free energy. The most significant benefit of keeping the parameter  $\beta$  finite is that it can be optimized to obtain the best generalization ability. Indeed, as has been recently shown [18]–[20], a variable  $\beta$  strategy gives superior generalization for various learning problems. A simple linear learning problem illustrates this effect in the next section.

## IV. APPLICATIONS

### A. Learning in the Noisy Linear Map

The ideas are now illustrated by analyzing a simple example: the linear learning problem. Consider a linear network: the output of the network is given by  $F_{\omega}(x) = \omega^T x$ , where  $x$  and  $\omega$  are  $D$ -dimensional real column vectors and  $\omega^T$  denotes

the transpose of  $\omega$ . The prior on the network configuration space is taken to be a  $D$ -dimensional symmetric Gaussian distribution  $p^{(0)}(\omega) = N(0, R_\omega)$ , with  $R_\omega = \sigma_\omega^2 \cdot I_D$ ,  $\sigma_\omega \gg 1$ , and  $I_D$  is the  $D$ -dimensional unit matrix.

The examples are generated by a similar linear map corrupted by additive white Gaussian noise:  $y = \omega_0^T x + \eta$ , where the noise  $\eta \sim N(0, \sigma_\eta^2)$ . The distribution over the domain  $X$  is taken to be also a  $D$ -dimensional Gaussian  $x \sim N(0, R_x)$ ;  $R_x = \sigma_x^2 \cdot I_D$ . These two distributions determine the underlying probability distribution of the examples  $P_F(\xi)$ . Learning proceeds using a quadratic error function:  $e(y|x, \omega) = (y - \omega^T x)^2 = ((\omega - \omega_0)^T x - \eta)^2$ .

The average free energy can be calculated [21] for  $m > D$  and  $\sigma_\omega \gg 1$ , and is

$$\begin{aligned} \langle \langle \log Z(m, \beta) \rangle \rangle &= -D \log \sigma_\omega - \frac{\omega_0^T \omega_0}{2\sigma_\omega^2} - \frac{1}{2} D \log (2\beta\sigma_x^2 m) \\ &\quad - \beta m \sigma_\eta^2 + \beta D \sigma_\eta^2 + O\left(\frac{1}{m}\right). \end{aligned} \quad (22)$$

The training and generalization errors follow from (21)

$$\langle \langle E^{(m)} \rangle \rangle = \frac{D}{2\beta} + (m - D)\sigma_\eta^2 \quad (23a)$$

$$\langle \langle -\log p^{(m)} \rangle \rangle \approx \frac{D}{2m} + \beta\sigma_\eta^2 + \log \sqrt{\frac{\pi}{\beta}} + O\left(\frac{1}{m^2}\right). \quad (23b)$$

It is interesting to note that the optimal  $\beta$ , that is, the one that gives minimal prediction error, can be easily determined from Eq. (23b) to be  $\beta_0 = 1/2\sigma_\eta^2$ , and is just the  $\beta$  corresponding to the level of noise in the examples. For this value of  $\beta$  the average prediction error reaches asymptotically its lowest possible value (Eq. (19)), which is the entropy of the noise in the examples. The optimal value  $\beta_0$  corresponds to a training error which is simply a constant per training example (Eq. (23a)), but not minimal! Reducing the training error by increasing the value of  $\beta$  beyond  $\beta_0$  increases the prediction error, due to an *overfitting* of the model to the noisy data. The asymptotic  $D/2m$  decrease of the prediction error is of the form obtained for other learning problems by various authors [18]–[20], [22].

### B. Architecture Selection in the Contiguity Problem

The sum over the generalization errors during the training process

$$\frac{1}{T} \sum_{t=1}^T \langle e(y_t | x_t) \rangle_{p^{(t-1)}} \geq \frac{1}{T} \sum_{t=1}^T \frac{-1}{\beta} \log p^{(t)}(y | x) \quad (24)$$

is an upper bound on the statistical prediction error and can be used as the practical generalization measure. By exchanging the averaging over the network ensemble with the summation over the training set we get an effective estimate  $\hat{G}^{(T)}$  of the generalization measure from the training samples  $\xi^{(T)}$  alone

$$\begin{aligned} \hat{G}^{(T)} &= \frac{1}{T} \sum_{m=0}^{T-1} \frac{1}{T-m} \sum_{t=m+1}^T e(y_t | x_t, \omega_m) \\ &\approx \frac{1}{T} \sum_t \langle e(y_t | x_t) \rangle_{p^{(t-1)}} \end{aligned} \quad (25)$$

where the ensemble averaging is approximated by randomly selecting the initial training points  $\omega_m$  for each  $0 \leq m \leq T-1$  [23].

To demonstrate the utility of this generalization measure (Eq. 25) for determining a sufficient size of the training set, as well as selecting the optimal architecture of the network, we focus on a simple Boolean mapping known as the “clumps” or contiguity problem [24]. For binary patterns of 10 bits, 792 of the 1024 patterns contain two or three continuous blocks of ones, or “clumps” separated by zeros. The Boolean function that separates this set into the subsets of two and three clumps cannot be implemented by a single-layer perceptron [25] and two layers of units are needed. There are, however, several two-layer networks that can implement the mapping, varying only in the connectivity between the first and second layer, the receptive width, as depicted in Fig. 1.

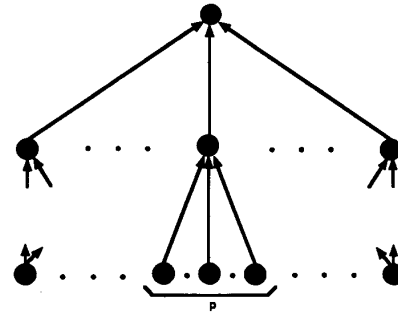


Fig. 1. The network architecture used for the contiguity problem. The receptive field width  $p$  is the size of the input field connected to each hidden unit. By varying the receptive field we observe a significant change in the generalization ability of the network.

Though the training error can be reduced on a small training set of 150 patterns for almost all the different architectures, these networks have a very different generalization ability at the end of the training process. To illustrate this point, the network was trained on increasing subsets of the training set, starting with 10 patterns, and increasing the number of patterns by 10 after convergence of the training algorithm on the previous subset. Using the generalization measure—as given by Eq. (25)—estimated within the training set, we were able to evaluate the generalization ability of the network and determine the optimal architecture. A more detailed description of these experiments can be found in Tishby *et al.* [23]. In Fig. 2 we plot the generalization

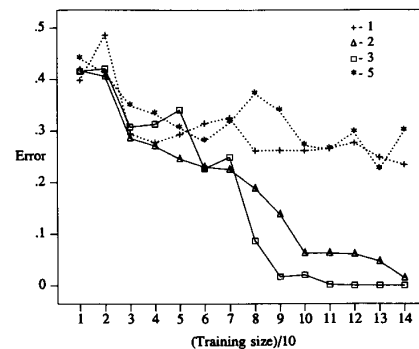


Fig. 2. Prediction errors at the end of training as a function of the training size, for various receptive fields. The difference in generalization ability is evident.

measure, estimated from the training set alone, and the average generalization error, calculated from all 792 patterns, as a function of the current training size, for the various receptive widths given in Fig. 1. We observe that, although for receptive widths of 1 and 5 the prediction error remains high throughout the training set, it drops sharply after about half of the patterns for the receptive widths of 2 and 3, indicating the superior generalization ability of these networks. The method described here is similar to the cross-validation techniques commonly used in pattern recognition [26].

### C. Learning Curves in the Annealed Approximation

The calculation of the sample average free energy for the general layered network is a very hard problem. A useful approximation, introduced by Schwartz *et al.* [27], becomes valid when the partition function itself is a "self-averaging" quantity, namely, when the random  $Z^{(m)}(\beta)$  converges for large  $m$  to a deterministic function and an "annealed" average can be used. In this case the partition function is asymptotically independent of the specific training set  $\xi^{(m)}$  and can be evaluated directly by the average

$$\begin{aligned} \langle\langle Z^{(m)}(\beta) \rangle\rangle &= \int_W d\omega \rho^{(0)}(\omega) \prod_{i=1}^m \int_{X \otimes Y} P_F(\xi) \\ &\quad \cdot \exp[-\beta e(y|x, \omega)] d\xi \\ &\equiv z^m \cdot \int_W \rho^{(0)}(\omega) g^m(\omega, \beta) d\omega \end{aligned} \quad (26)$$

where

$$g(\omega, \beta) \equiv \int_{X \otimes Y} P_F(\xi) \frac{\exp[-\beta e(y|x, \omega)]}{z} d\xi = \langle\langle p(y|x, \omega) \rangle\rangle$$

is the sample average of the likelihood (5), and is a measure of the compatibility of the network  $\omega$  to the modeled relation  $F$ .

Within the annealed approximation, the average prediction probability after training on  $m$  examples can be written as a ratio of two successive moments of a well-defined prior distribution.

$$\langle\langle p^{(m)} \rangle\rangle \approx \frac{\langle\langle Z^{(m+1)}(\beta) \rangle\rangle}{\langle\langle Z^{(m)}(\beta) \rangle\rangle z} = \frac{\langle g^{m+1} \rangle_{\rho^{(0)}(g)}}{\langle g^m \rangle_{\rho^{(0)}(g)}}. \quad (27)$$

The density  $\rho^{(0)}(g) \equiv \int \rho^{(0)}(\omega) \delta(g(\omega) - g) d\omega$ , where  $\delta(x)$  is the Dirac delta function, is the prior  $\rho^{(0)}(\omega)$  expressed as a function of the single variable  $g(\omega, \beta)$ , and contains all the information about the configuration space  $W$  and about the task (the desired relation  $F$ ) through the definition of  $g(\omega, \beta)$ . Learning curves,  $\langle -\log p^{(m)} \rangle$  versus  $m$ , can now be obtained simply by calculating the moments of such prior densities [23], [27].

The asymptotic behavior of the moments ratio (27) is determined solely by the functional form of  $\rho^{(0)}(g)$  near  $g = 1$ . If, for example,  $\rho^{(0)}(g) \sim (1 - g)^d$  as  $g \rightarrow 1$ , for some exponent  $d \geq 0$

$$p^{(m)} \approx 1 - \frac{d+1}{m}; \quad -\log p^{(m)} \approx \frac{d+1}{m} \quad (28)$$

for large  $m$ . The asymptotic form (28) is in agreement with the  $1/m$  decrease of the generalization error found earlier [18]–[20], [22]. This annealed approximation suggests an

interesting possible relation between the value of the exponent  $d$  and the VC-dimension [9], [28] of the learning system.

### V. SUMMARY

By using a few simple and plausible assumptions we show that the Gibbs formulation of statistical mechanics is well suited for the typical case analysis of the problem of learning from examples in layered networks. We propose a statistical measure of generalization that can be derived directly from the free energy of the network's ensemble, and which is shown to be equivalent to the stochastic complexity of the training data. By doing this we link the statistical mechanics of neural networks with the modern methods of statistical estimation theory. The formalism can be applied to analytic and numerical evaluation of learning curves, as well as be a practical method for training networks that generalize well.

### ACKNOWLEDGMENT

The authors give special thanks to Neri Merhav for many illuminating discussions, for his help in the calculations of the linear example, and for critically reading the manuscript. Useful discussions with Géza Gyögyi, Dan Schwartz, Andrew Ogielski, and John Denker are greatly appreciated.

### REFERENCES

- [1] R. P. Lippmann, "An introduction to computing with neural nets," *ASSP Magazine*, vol. 4, no. 2, pp. 4–22, 1987.
- [2] G. Cybenko, "Continuous valued neural networks with two hidden layers are sufficient," Tufts University preprint, 1988.
- [3] L. G. Valiant, "A theory of the learnable," *Comm. ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.
- [4] L. D. Landau and E. M. Lifshitz, *Course of theoretical physics*, vol. 5, 3rd ed. Pergamon, 1980.
- [5] T. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications to patterns recognition," *IEEE Trans. Electron. Comput.*, vol. 14, pp. 326–334, 1965.
- [6] E. B. Baum, "On the capabilities of multilayer perceptrons," *J. Complexity*, 1989.
- [7] H. White, "Learning in artificial neural networks: A statistical perspective," *Neural Computation*, vol. 1, pp. 425–464, 1989.
- [8] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, pp. 1080–1100, 1986.
- [9] E. B. Baum and D. Haussler, "What size net gives valid generalization," *Neural Computation*, vol. 1, pp. 151–160, 1989.
- [10] R. T. Cox, "Probability, frequency and reasonable expectation," *Amer. J. Phys.*, vol. 14, pp. 1–26, 1946.
- [11] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterizations*. Academic Press, 1975, p. 16.
- [12] Y. Tikhonchinsky, N. Tishby, and R. D. Levine, "Alternative approach to maximum entropy inference," *Phys. Rev. A*, vol. 30, pp. 2638–2644, 1984.
- [13] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, pp. 671–680, 1983.
- [14] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.
- [15] J. Denker *et al.*, "Large automatic learning, rule extraction, and generalization," *Complex Systems*, vol. 1, pp. 877–922, 1987.
- [16] P. Carnevali and S. Patranello, "Learning networks of neurons with Boolean logic," *Europhys. Lett.*, vol. 4, pp. 503–508, 1999–1204, 1987.
- [17] E. Gardner, "The space of interactions of neural networks models," *J. Phys. A*, vol. 21, pp. 257–270, 1988; E. Gardner and B. Derrida, *J. Phys. A*, vol. 21, pp. 271–284, 1988.
- [18] D. Hansel and H. Sompolinsky, "Learning from examples in a single-layer neural network," *Europhys. Lett.*, 1990.
- [19] G. Gyögyi and N. Tishby, "Statistical theory of learning a rule," in *Neural Networks and Spin Glasses*. W. Theumann and R. Kobrele Eds. New Jersey: World Scientific, 1990.

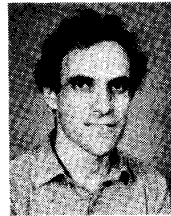
- [20] H. Sompolinsky, S. Seung, and N. Tishby, "Learning from examples in large neural networks," to be published.
- [21] E. Levin, N. Tishby, and S. A. Solla, in *Proc. 2nd Ann. Workshop on Computational Learning Theory (COLT'89)*, R. Rivest, D. Haussler, and M. K. Warmuth, Eds. San Mateo, CA: Morgan Kaufmann, 1989, pp. 245-260.
- [22] D. Haussler, N. Littlestone, and M. K. Warmuth, "Predicting  $\{0, 1\}$  Functions on Randomly Drawn Points," in *Proc. COLT'88 San Mateo, CA: Morgan Kaufmann, 1988*, pp. 280-295.
- [23] N. Tishby, E. Levin, and S. A. Solla, "Consistent inference of probabilities in layered networks: Predictions and generalization," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*, vol. 2, pp. 403-409, IEEE Press, Washington DC, June 1989. New York: IEEE Press, 1989.
- [24] T. Maxwell, C. L. Giles, and Y. C. Lee, "Generalization in neural networks, the contiguity problem," in *Proc. IEEE 1st Int. Conf. on Neural Networks*, San Diego, 1987; T. Grossman, R. Meir, and E. Domany, "Learning by choice of internal representations," *Complex Systems*, vol. 2, pp. 555-563, 1988.
- [25] M. Minsky, and S. Papert, *Perceptrons*. Cambridge, MA: M.I.T. Press, 1969.
- [26] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *J. R. Stat. Soc. Ser. B*, vol. 36, pp. 111-133, 1974.
- [27] D. B. Schwartz, V. K. Samalam, J. S. Denker, and S. A. Solla, "Exhaustive learning," *Neural. Comp.*, 1990.
- [28] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," UCSC-CRL 87-20 and *J. ACM*, 1988.



**Esther Levin** was born in Kiev, USSR, in 1960. She received the B.Sc., M.Sc., and D.Sc. degrees from the Technion-Israel Institute of Technology in 1982, 1985, and 1988, all in electrical engineering.

In 1988 she joined the Speech Research Department of AT&T Bell Laboratories, Murray Hill, NJ. Her current research interests include computational neural science, machine learning, and applications of neural networks to speech recognition. In

the latter area she studies combinations of neural networks with hidden Markov models for nonlinear prediction of the speech signal.



**Naftali Z. Tishby** was born in Jerusalem, Israel, in December 1952. He received the B.Sc. degree (cum laude) in physics and mathematics from the Hebrew University of Jerusalem in 1974, M.Sc. degree (cum laude) in physics from Tel-Aviv University in 1980, and the Ph.D. in theoretical physics from Hebrew University in 1985.

From 1974 to 1981 he was with the Israel Defense Forces (IDF), where he established and headed a research group in signal and speech processing. During 1984-1985 he served as a Vice President of Research in Sesame Systems Ltd. developing speech and speaker recognition systems. In 1985-1986 he was a postdoctoral fellow at the Massachusetts Institute of Technology, working on chaotic Hamiltonian dynamics. Since 1987 he has been a Member of Technical Staff (information principles laboratory) at AT&T Bell Laboratories, Murray Hill, NJ. His current research subjects include nonlinear dynamics and its applications to speech processing, stochastic processes, learning theory, and statistical mechanics of neural networks.

Dr. Tishby received the Eliyahu Golomb Israel Security Award in 1980 and the Chaim Weizmann fellowship in physics in 1985.



**Sara A. Solla** was born in Buenos Aires, Argentina. She received the B.Sc. degree in physics from the National University of Technology, Buenos Aires, in 1974, and the Ph.D. degree in theoretical physics from the University of Washington, Seattle, in 1982.

She has served as a lecturer in the Physics Department of the National University of Technology, Buenos Aires. She has done research as a postdoctoral associate at Cornell University, where she worked in the

areas of critical phenomena, metal-insulator transitions, and fracture of geological systems. Subsequently, her work as a visiting scientist at the IBM T. J. Watson Research Center, on the application of nonequilibrium statistical physics to the investigation of the dynamical properties of the simulated annealing algorithm, led to her current interest in neural networks. Since 1986, she has been a Member of the Technical Staff at AT&T Bell Laboratories, Holmdel, NJ. Her current research in the area of neural networks focuses on the theoretical description of learning and generalization, and includes applications to pattern recognition and control theory.