Multilayer perceptrons may learn simple rules quickly

R. Urbanczik
Institut für theoretische Physik, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany
Received 13 August 1997

Zero-temperature Gibbs learning is considered for a connected committee machine with K hidden units. For large K, the scale of the learning curve strongly depends on the target rule. When learning a perceptron, the sample size P needed for optimal generalization scales so that $N \ll P \ll KN$, where N is the dimension of the input. This holds even for a noisy perceptron rule if a new input is classified by the majority vote of all students in the version space. When learning a committee machine with M hidden units, $1 \ll M \ll K$, optimal generalization requires $\sqrt{MK}N \ll P$. [S1063-651X(98)03908-7]

PACS number(s): 87.10.+e, 05.90.+m, 64.60.Cn

Supervised learning in neural networks has been studied from a wide range of theoretical perspectives. In statistics one may obtain bounds on the learning behavior that indicate that the sample size should be on the order of the VC dimension of the network to enable good generalization [1]. Under some regularity assumptions, one may use information geometric ideas to determine the asymptotics of the learning curve in the limit where the number of training examples is large [2]. This yields that the sample size should be on the order of the number of free parameters in the architecture, and for feedforward networks with threshold units this is the same as the VC dimension (up to a factor that is at most logarithmic) [3]. In particular, both approaches suggest that the sample size must be increased with the capabilities of the learner and that this is quite independent of the rule that is to be learned.

In statistical mechanics one has to make rather detailed assumptions about the learning problem, but can in turn calculate the learning behavior exactly in the thermodynamic limit. While this has given rise to important qualifications to the above theories, e.g., the discontinuous transition to perfect generalization in Ising networks [4,5], the above scaling of the learning curve has to date been observed in statistical mechanics as well. Indeed, generic arguments that the scale of the learning curve must be set by the number of free parameters in the thermodynamic limit have been brought forth in [6].

However, in some practical applications, the generalization properties of feedforward networks have been found to be startlingly good in view of these theoretical expectations [7,8]. The purpose of this paper is to point out that for a specific multilayer network, the fully connected committee machine, the scale of the learning curve depends strongly on the target rule.

This machine is characterized by K weight vectors $J_i \in \mathbb{R}^N$, $|J_i|=1$, and given an N-dimensional input ξ it computes $\sigma_J(\xi)=\mathrm{sgn}[\Sigma_{i=1}^K \mathrm{sgn}(J_i^T \xi)]$. We shall consider a situation where the target rule or teacher is a simpler committee machine $\sigma_B(\xi)$ with M weight vectors B_l and $M<K$. The high-temperature limit of a related scenario ($M=1, K=3$, binary synapses) has been discussed in [9]. Here we focus on the case $M \ll K \ll N$ since this not only is technically simpler than finite K but separates the scales of having a sample size of, e.g., $O(N)$ or $O(KN)$. In real world applications the architecture of the student will make it impossible to implement the teacher perfectly and such a situation shall be modeled by considering a noisy teacher.

We consider Gibbs learning at zero temperature since this can be shown to converge, for any teacher, to the optimal student in the limit of large sample size [10]. A well known strategy in machine learning is to combine the predictions of different classifiers. Instead of just picking a student from the Gibbs ensemble, we thus also consider classifying a new input by the output of the majority of the students in the Gibbs ensemble. Under suitable assumptions on priors (which do not hold in the present case), this is the Bayes algorithm [11].

More formally, let Ω be the set of inputs and A be a probability distribution on $\Omega \times \{-1,1\}$ representing the (stochastic) teacher. For a binary function $f \in \{-1,1\}^\Omega$ we may then define the generalization error $\epsilon_g(f) = \langle \theta(-\sigma f(\xi)) \rangle_{(\xi,\sigma)}$ as the probability with respect to A that $\sigma \neq f(\xi)$ for an input/output pair (ξ,σ). Let $\mathcal{F}$, a set of binary functions, be the class of students and μ a probability distribution on $\mathcal{F}$, representing our confidence in the generalization ability of a student. Denote by $r(\xi,\sigma)=\langle \theta(\sigma f(\xi)) \rangle_f$ the probability with respect to μ that $\sigma = f(\xi)$. We then obtain a classifier that averages over all students by setting $h_\mu(\xi) = \mathrm{sgn}[r(\xi,1)-r(\xi,-1)]$. The generalization error of this classifier ϵ_{ens} and the average generalization error ϵ_{smp} committed by simply sampling from μ are then

$$\epsilon_{\mathrm{smp}} = \langle \epsilon_g(f) \rangle_f = \langle r(\xi,-\sigma) \rangle_{(\xi,\sigma)},$$

$$\epsilon_{\mathrm{ens}} = \epsilon_g(h_\mu) = \langle \theta(2r(\xi,-\sigma)-1) \rangle_{(\xi,\sigma)}. \qquad (1)$$

Since $\epsilon_{\mathrm{ens}} \leqslant \langle 2r(\xi,-\sigma)\theta(2r(\xi,-\sigma)-1) \rangle_{(\xi,\sigma)}$, one has $\epsilon_{\mathrm{ens}} \leqslant 2\epsilon_{\mathrm{smp}}$ and it is straightforward to construct unusual situations (for any $\epsilon_{\mathrm{smp}} \leqslant \frac{1}{2}$) where the inequalities are tight. However, below we shall encounter cases where ϵ_{ens} is much smaller than ϵ_{smp} and even smaller than the generalization error of the best student in the support of μ.

Let $\mathcal{T}$ be a training set of P pairs (ξ^ν, σ^ν) picked independently from A and assume that μ is such that any student f picked from μ lies in the version space, i.e., f has minimal training error $\Sigma_\nu \theta(-\sigma^\nu f(\xi^\nu))$. Then ϵ_{smp} will converge to ϵ_{min}, the minimal generalization error attainable in $\mathcal{F}$, as $P \to \infty$. However, only in realizable cases $\epsilon_{\mathrm{min}}=0$ does this

imply $\epsilon_{\rm ens}\to\epsilon_{\rm min}$. If the optimal student is unique, however, under weak assumptions on $\mathcal{F}$ and the input distribution, the version space will shrink to a point for large P and then trivially $\epsilon_{\rm ens}\to\epsilon_{\rm smp}\to\epsilon_{\rm min}$.

In our present case the teacher A is given by the noisy committee machine

$$\eta^*\sigma_B(\gamma\xi+\sqrt{1-\gamma^2}\,\eta). \tag{2}$$

The components of the input vector ξ and of the (input) noise vector η shall be picked independently from the normal distribution. A second source of (output) noise is due to $\eta^*\in\{-1,1\}$, which equals 1 with probability γ^*. Thus, for $\gamma=\gamma^*=1$ the teacher is deterministic and the learning problem realizable and for $\gamma=0$ or $\gamma^*=\frac{1}{2}$ the teacher is random.

For zero-temperature Gibbs learning μ is given by the uniform distribution on the parameters (J) of the functions in version space. Thus the key quantity to consider is the version space volume $V(\mathcal{T})$. Note that we shall consider only a sample size P for which zero training error is achievable. Hence the calculation of the replicated version space volume for large N leads to the following extremal problem:

$$\frac{1}{N}\ln\langle V^n(\mathcal{T})\rangle_{\mathcal{T}}=\operatorname*{extr}_{\mathbf{q},\mathbf{R}}\frac{P}{N}\ln G_r^{(n)}(\mathbf{q},\mathbf{R})+\ln G_s^{(n)}(\mathbf{q},\mathbf{R}). \tag{3}$$

Here the matrix $\mathbf{q}=(q_{ij}^{ab})$ is given by the overlaps of the weight vectors of the students and $\mathbf{R}=(R_{li}^a)$ by the overlaps between these and the weight vectors of the teacher. We assume that the teacher has orthonormal weight vectors and then the entropy term $G_s^{(n)}$ is

$$G_s^{(n)}=\det\begin{pmatrix}\mathbf{1} & \mathbf{R}^T\\ \mathbf{R} & \mathbf{q}\end{pmatrix}^{1/2}. \tag{4}$$

The solution of Eq. (3) will require symmetry assumptions about the extremal values of $\mathbf{q}$ and $\mathbf{R}$. Subject to such an assumption the determinant in $G_s^{(n)}$ may be evaluated by recursively applying the following relations for block matrices: $\det\left(\begin{smallmatrix}\mathbf{a} & \mathbf{b}\\ \mathbf{c} & \mathbf{d}\end{smallmatrix}\right)=\det\mathbf{a}\,\det(\mathbf{d}-\mathbf{c}\mathbf{a}^{-1}\mathbf{b})$ and $\det M_k(\mathbf{u},\mathbf{v})=\det(\mathbf{u}-\mathbf{v})^{k-1}\det[\mathbf{u}+(k-1)\mathbf{v}]$. Here $M_k(\mathbf{u},\mathbf{v})$ denotes a square matrix with k diagonal entries $\mathbf{u}$ and off-diagonal entries $\mathbf{v}$. For the more complicated parametrizations of $\mathbf{q}$ and $\mathbf{R}$ the calculations become tedious and are best left to a computer program capable of symbolic algebra.

The energy term $G_r^{(n)}$ in the extremal problem (3) is given by

$$G_r^{(n)}=\left\langle 2\,\theta(\eta^*\mathcal{Y})\prod_{a=1}^{n}\theta(\mathcal{Z}^a)\right\rangle_{Y_l,Z_i^a,\eta^*},$$

$$\mathcal{Z}^a=K^{-1/2}\sum_{i=1}^{K}\operatorname{sgn}(Z_i^a),$$

$$\mathcal{Y}=M^{-1/2}\sum_{l=1}^{M}\operatorname{sgn}(Y_l)\quad\text{if}\quad M>1,$$

$$\mathcal{Y}=Y_1\quad\text{if}\quad M=1. \tag{5}$$

The Z_i^a and Y_l are zero mean Gaussian, with covariances $\langle Z_i^aZ_j^b\rangle=q_{ij}^{ab}$, $\langle Y_lZ_i^a\rangle=\gamma R_{li}^a$, and $\langle Y_lY_k\rangle=\delta_{lk}$. For all the parametrizations of $\mathbf{q}$ and $\mathbf{R}$ we shall consider, one may show as in [12] that the joint distribution of $\mathcal{Z}^a$ and $\mathcal{Y}$ is Gaussian in the limit $K\to\infty$ when M is equal to 1 or when M is large as well. Since the covariances of the $\mathcal{Z}^a$ and $\mathcal{Y}$ are readily calculated from those of Z_i^a and Y_l this greatly simplifies $G_r^{(n)}$.

Combining Eqs. (1) and (3) allows us to calculate the generalization behavior: Almost by definition the typical, with respect to training sets, value of the nth moment of the random variable $r(\xi,\sigma)$ is given by $G_r^{(n)}$ evaluated at the typical values of $\mathbf{q},\mathbf{R}$. These may be obtained from the extremal condition in Eq. (3) for small n. Consequently, $\epsilon_{\rm smp}$ is given by

$$1-\gamma^*+\frac{2\gamma^*-1}{\pi}\arccos\rho \tag{6}$$

for $\rho=R_e/\sqrt{v_e}$. Here $R_e=\langle\mathcal{Y}\mathcal{Z}^a\rangle$ and $v_e=\langle\mathcal{Z}^a\mathcal{Z}^a\rangle$. In contrast to $\epsilon_{\rm smp}$, the generalization error of the ensemble depends on the geometry of the version space. Within a replica symmetric ansatz one finds

$$G_r^{(n)}=\left\langle 2H\left(\frac{R_e\eta^*x}{\sqrt{q_e-R_e^2}}\right)H\left(\sqrt{\frac{q_e}{v_e-q_e}}\,x\right)^n\right\rangle_{x,\eta^*}. \tag{7}$$

The distribution of x is normal, $q_e=\langle\mathcal{Z}^a\mathcal{Z}^b\rangle$, and for the typical values of the order parameters (7) equals $\langle r(\xi,\sigma)^n\rangle_{(\xi,\sigma)}$. As pointed out in [13], it is thus easy to evaluate $\langle F(r(\xi,\sigma))\rangle_{(\xi,\sigma)}$ if F is, or can arbitrarily well be approximated by, a polynomial. Since this holds for the θ function in Eq. (1), a simple calculation yields that $\epsilon_{\rm ens}$ is given by Eq. (6) for $\rho=R_e/\sqrt{q_e}$. However, already one step of replica symmetry breaking would yield a much more complicated right-hand side of Eq. (7). It thus seems very difficult to perform a similar calculation of $\epsilon_{\rm ens}$ when replica symmetry is broken.

We first consider $M=1$. In this case a site symmetric parametrization of $\mathbf{q}$ and $\mathbf{R}$ should be sufficient and we set $R_{li}^a=r^a/\sqrt{K}$ and $q_{ij}^{ab}=p^{ab}/K+\delta_{ij}q^{ab}$. The scaling of the order parameters with K is such that the contribution of r^a and p^{ab} to the covariance matrix of $\mathcal{Z}^a$ and $\mathcal{Y}$ stays finite in the large K limit. The best achievable generalization error is given by Eq. (6) for $\rho=\gamma$.

The replica symmetric theory will be sufficient for $P=\tilde{\alpha}N$, when the sample size is an infinitesimal fraction of the number of free parameters. The resulting power laws for the generalization error as $\tilde{\alpha}\to\infty$ are summarized in Table I. Only in the noiseless case does $\epsilon_{\rm smp}$ decay to $\epsilon_{\rm min}$. For identical values of $\epsilon_{\rm min}$, the asymptotic value of $\epsilon_{\rm smp}$ is higher in the case of output noise than for input noise. The generalization error of the ensemble becomes minimal in all cases. For input noise, the $1/\tilde{\alpha}$ decay of the ensemble quite remarkably equalizes the decay in the Bayesian algorithm that is optimized for this specific class of teachers [14].

The great difference between the ensemble and sampling may be explained quite simply. Let $\hat{B}=K^{-1/2}\Sigma_iJ_i$ be the

TABLE I. For large $\tilde{\alpha}$ the generalization error decays to $\epsilon_{\min}$ as $d\tilde{\alpha}^{-k}$ for the values of k and d given in this table. The value of c_1 is $c_1 = -\int_{-\infty}^{0} dx \ln H(x)$.

Model		k	d
no noise	ϵ_{smp}	1/3	$\frac{(\pi-2)^{1/3}}{\sqrt{2}\pi^{5/6}c_1^{1/3}}$
no noise	ϵ_{ens}	2/3	$\frac{\sqrt{\sqrt{2}-c_1\sqrt{\pi}}(\pi-2)^{1/6}}{2\pi^{5/12}c_1^{1/6}}$
input noise	ϵ_{smp}	0	$\frac{1}{\pi}\arccos\gamma^2-\epsilon_{\min}$
input noise	ϵ_{ens}	1	$\frac{1}{4\gamma}$
output noise	ϵ_{smp}	0	no explicit form
output noise	ϵ_{ens}	1/2	no explicit form

(rescaled) average weight vector of a typical student in version space. Then in the limit $|\hat{B}|\to\infty$ the output of the large committee σ_J is equal to the perceptron with weight vector $\hat{B}$ on almost all inputs. Further, $\hat{B}$ becomes parallel to the teacher for large $\tilde{\alpha}$. However, only in the noiseless case does the length of $\hat{B}$ diverge (as $\tilde{\alpha}^{1/3}$). This length influences the performance of a single student but is immaterial for the ensemble since the specialized overlaps q^{ab} are zero.

We next consider the more conventional scaling of the sample size $P=\alpha KN$. If there is no noise, the generalization error vanishes. In the noisy cases up to a critical point the generalization behavior is the same as for $\tilde{\alpha}\to\infty$. Since $\epsilon_{\text{ens}}=\epsilon_{\min}$ the ensemble agrees with the noiseless teacher on almost all inputs but, in contrast to the noiseless teacher, it has zero training error. With increasing α the version space shrinks rapidly and above a critical α specialized correlations between the students emerge, i.e., $q^{ab}=0$ no longer holds. While this can be seen in the replica symmetric theory, a correct description requires the breaking of replica symmetry. The critical value α_{RSB} as function of the noise is shown in Fig. 1. At the transition ϵ_{smp} increases and, due to the specialized correlations, the error of the ensemble will increase as well. For large α the version space shrinks to a point ($q^{ab}\to 1$) and thus $\epsilon_{\text{ens}}\to\epsilon_{\text{smp}}$. The asymptotic value of ϵ_{smp} on this scale is higher than the one found for large $\tilde{\alpha}$. The generalization error will decrease again when the training set size is on the order of the storage capacity of the student, that is, on the order of $\sqrt{\ln K}KN$ [15].

For large but finite K the generalization performance will

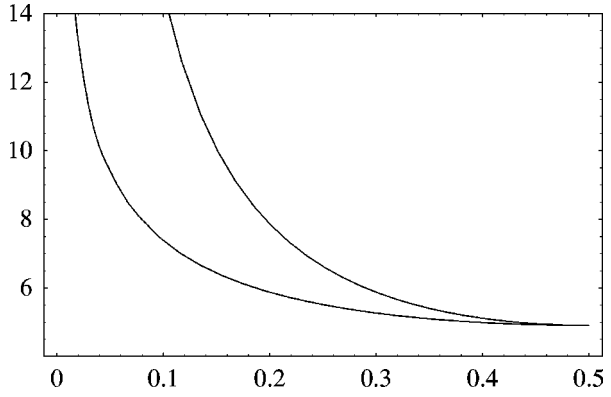

FIG. 1. Value of α at which replica symmetry breaks. The upper curve is for input noise, the lower one for output noise. One step of RSB was considered, at the transition the breakpoint parameter m decreases from 1, and one of the two values of q^{ab} is still 0 and the other is close to 1.

be well described by the above $\tilde{\alpha}$ theory as long as $P<\alpha_{\text{RSB}}KN$. So for finite K only a close to minimal generalization error is achievable in this phase. Nevertheless, the question arises whether the remarkable generalization performance for small sample sizes occurs only if $M=1$. Indeed, the analysis of, e.g., $M=3$ cannot be obtained by a simple extension from the perceptron case. If one-third of the hidden units in the student have small overlaps with the first unit in the teacher, this will reproduce only the hidden field $B_1^T\xi$, but not its sign.

It is not necessary, however, that all of the hidden units in the student specialize on some unit in the teacher. Let us assume that for each teacher unit there are λ^* units in the student that have specialized on it. The scale of λ^* is set by the requirement that the field produced by the specialized units should have the same order of magnitude as the entire field ($\mathcal{Z}^a$), that is, $\lambda^*=\lambda\sqrt{K/M}$. This in turn suggests that the learning curve should exhibit an interesting behavior when the size of the training set scales as $P=\hat{\alpha}\sqrt{KM}N$. Due to the broken site symmetry the calculations are rather involved and we shall consider only the noiseless case in the limit $M\to\infty$ but $M\ll K$.

Setting $h(i)=\lceil i/\lambda^*\rceil$, so that for $h(i)\leqslant M$ the ith hidden unit of the student has specialized of the $h(i)$th teacher unit, we arrive at the parametrization of the overlap matrices

$$R_{li}^a=\begin{cases} r_s^a/M+\delta_{lh(i)}R_s^a & \text{if } h(i)\leqslant M \\ r_u^a/K & \text{if } h(i)>M, \end{cases}$$

$$q_{ij}^{ab}=\begin{cases} p_{s2}^{ab}/M+\delta_{h(i)h(j)}p_{s1}^{ab}+\delta_{ij}q_s^{ab} & \text{if } h(i),h(j)\leqslant M \\ p_{su}^{ab}/\sqrt{MK} & \text{if } h(i)\leqslant M<h(j) \\ p_u^{ab}/K+\delta_{ij}q_u^{ab} & \text{if } M<h(i),h(j). \end{cases} \tag{8}$$

The replica symmetric theory will be sufficient for the above scaling of the training set size. Then the extremal problem (3) has a solution with $q_u^{ab}=0$ and where the following quantities are of order $1/\sqrt{KM}$:

$$1+p_u^{aa}-p_u^{ab},\ \ p_u^{ab}-(r_u^a)^2,\ \ p_{su}^{aa}-p_{su}^{ab},$$

$$p_{su}^{ab}-r_u^a(r_s^a+R_s^a),\ \ p_{s1}^{ab}+p_{s2}^{ab}-(r_s^a+R_s^a)^2,\ \ q_s^{ab}. \qquad (9)$$

These relations imply that the average weight vector of the teacher, the average of the specialized units, and the average of the unspecialized units in the student are parallel. Further one finds

$$O(1/M)=1-p_{s1}^{aa}+\lambda*(p_{s1}^{aa}-p_{s1}^{ab}+p_{s2}^{aa}-p_{s2}^{ab}),$$

$$O(\sqrt{M/K})=p_{s1}^{aa}-(R_s^a)^2,\quad O(\sqrt{M/K})=p_{s1}^{aa}-p_{s1}^{ab}. \qquad (10)$$

With these relations the typical value of the version space volume can be obtained from

$$\frac{1}{\sqrt{KMN}}\langle \ln V(\mathcal{T})\rangle_{\mathcal{T}}=\operatorname*{extr}_{R_s,\lambda}\hat{\alpha}G_r+\frac{\lambda}{2}\ln(1-R_s^2), \qquad (11)$$

where $G_r=(\partial/\partial n)G_r^{(n)}{}_{|n=0}$ and $G_r^{(n)}$ is given by Eq. (7) for

$$R_e=2(r+\lambda \arcsin R_s)/\pi,$$

$$q_e=2[r(r+2\lambda R_s)+\lambda^2 \arcsin R_s^2]/\pi,$$

$$v_e=q_e+1-2/\pi. \qquad (12)$$

The value of $r=r_u^a+\lambda r_s^a$ is given by the constraint

$$2(r+\lambda R_s)\frac{\partial G_r}{\partial q_e}+\frac{\partial G_r}{\partial R_e}=0. \qquad (13)$$

For small values of $\hat{\alpha}$ the above extremal problem has the unspecialized solution $r=1,\lambda=0$. The generalization error is the same as for $M=K$ in this unspecialized phase [16]. However, above a critical sample size $\hat{\alpha}\approx 5.17$, the value of λ increases from zero and it diverges with growing $\hat{\alpha}$. The value of R_s is close to one already at the transition. Asymptotically, one finds

$$\epsilon_{\rm smp}=\frac{2^{1/4}(\pi-2)^{1/4}}{\pi\sqrt{c_2}}\sqrt{\frac{\ln\hat{\alpha}}{\hat{\alpha}}}, \qquad (14)$$

where $c_2=-\int dx H(x)\ln H(x)$. The ensemble improves only marginally on this performance, the decay in its generalization error being a factor $1/\sqrt{2}$ faster. This is related to the fact that $r>0$ for finite $\hat{\alpha}$. While this improves the performance of a single student, it creates a deviation from the teacher that is common to (almost) all students in the ensemble.

In summary, we have seen that the initial scale of the learning curve is not determined by the number of free parameters. It is determined by the number of constraints on the parameters that must be approximately satisfied to approximate an optimal student well [17]. For the large fully connected committee machine the two quantities can differ by orders of magnitude and this endows the machine with a built-in capability of model selection.

I thank M. Biehl and M. Opper for helpful discussions on the final version of the manuscript. This work was supported by the Deutsche Forschungsgemeinschaft.

[1] V. Vapnik, *Estimation of Dependences Based on Empirical Data* (Springer, Berlin, 1982). The maximal size of a set of inputs for which a given class of functions implements all possible input/output relations on the set is called the VC dimension of the class.
[2] S. Amari and S. Shinomoto, Neural Comput. **4**, 605 (1992).
[3] G. Mitchison and R. Durbin, Biol. Cybern. **60**, 345 (1989).
[4] G. Györgyi, Phys. Rev. A **41**, 7097 (1990).
[5] H. Sompolinsky, N. Tishby, and H. Seung, Phys. Rev. Lett. **65**, 1683 (1990).
[6] H. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. A **45**, 6056 (1992).
[7] G. Martin and J. Pittman, Neural Comput. **3**, 258 (1991).
[8] Y. LeCun *et al.*, in *Neural Networks: The Statistical Mechanics Perspective*, edited by J. Oh, C. Kwon, and S. Cho (World Scientific, Singapore, 1995), pp. 261–276.
[9] H. Schwarze, M. Opper, and W. Kinzel, Phys. Rev. A **46**, R6185 (1992).
[10] D. Haussler, Inf. Comput. **100**, 78 (1992).
[11] M. Opper and D. Haussler, Phys. Rev. Lett. **66**, 2677 (1991).
[12] R. Urbanczik, J. Phys. A **28**, 7097 (1995).
[13] M. Opper and D. Haussler, in *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, edited by L. Valiant and M. Warmuth (Kaufmann, San Mateo, 1991), pp. 75–87.
[14] M. Opper and W. Kinzel, in *Models of Neural Networks III*, edited by E. Domany, J. van Hemmen, and K. Shulten (Springer, New York, 1995), pp. 151–207. Note that the authors measure the generalization performance against the noiseless teacher. If one compares with the noisy teacher, their result is as quoted above.
[15] R. Urbanczik, J. Phys. A **30**, L387 (1997).
[16] H. Schwarze, J. Phys. A **26**, 5781 (1993).
[17] One may also say that the uniform prior on the parameters assumed in Gibbs learning yields a highly nonuniform prior on the functions implemented by these parameters. Note that given a distribution on inputs, $\mathcal{F}$ can be seen in a natural way as a metric space, defining the distance between two functions as the probability that they differ on some input [14].