A multifractal phase-space analysis of perceptrons with biased patterns

1998 J. Phys. A: Math. Gen. 31 2509

(http://iopscience.iop.org/0305-4470/31/11/004)

J. Phys. A: Math. Gen. **31** (1998) 2509–2517. Printed in the UK PII: S0305-4470(98)88639-3

A multifractal phase-space analysis of perceptrons with biased patterns

J Berg† and A Engel‡

Institut für Theoretische Physik, Otto-von-Guericke-Universität Magdeburg, PSF 4120, 39016 Magdeburg, Germany

Received 22 October 1997

Abstract. We calculate the multifractal spectrum of the partition of the coupling space of a perceptron induced by random input–output pairs with non-zero mean. From the results we infer the influence of the input and output bias respectively on both the storage and generalization properties of the network. It turns out that the value of the input bias is irrelevant as long as it is different from zero. The generalization problem with output bias is new and shows an interesting two-level scenario. To compare our analytical results with simulations we introduce a simple and efficient algorithm to implement Gibbs learning.

1. Introduction

The properties of simple networks of formal neurons can be quantitatively described by characterizing the partition of the coupling space induced by the required input–output mappings [1–4]. In some cases the geometrical properties of this partition can be concisely specified by the multifractal spectrum of the distribution of cells in coupling space corresponding to the different output sequences that can be generated by the system for given inputs [5, 6]. This approach has been used for a number of investigations of both single-layer as well as multilayer feed-forward neural nets [7–10] and revealed a number of interesting new aspects. The simplest case of the perceptron allows a rather detailed analysis also highlighting the problems and subtleties of the method [11]. All investigations carried out so far have, however, considered symmetric statistics of both inputs and outputs.

In this paper we analyse the multifractal properties of the phase space of a single-layer perceptron induced by input patterns with biased statistics. A possible bias of the outputs is taken care of by considering a special subset of cells only. The investigation is motivated by the fact that the storage properties of a perceptron are known to depend markedly on the statistics of the patterns. A similar influence can be expected for the generalization behaviour which has to our knowledge not been studied for biased outputs before.

The analysis is performed using the methods that have been employed for the case of unbiased patterns already. With the help of the replica trick the multifractal spectrum $f(\alpha)$ averaged over the distribution of inputs is calculated analytically for different pattern set sizes γ. The storage and generalization properties are determined by the points with slope 0 and 1 respectively of these curves [3, 6]. The results for the storage capacity are compared with previous findings [2], whereas those for the generalization behaviour are

† E-mail address: johannes.berg@physik.uni-magdeburg.de
‡ E-mail address: andreas.engel@physik.uni-magdeburg.de

0305-4470/98/112509+09$19.50

checked against numerical simulations. In order to efficiently explore the version space in these simulations, we introduce a randomized variant of the perceptron learning algorithm.

2. Calculation of the cell spectrum

We consider a spherical perceptron specified by a real coupling vector $\boldsymbol{J}$ obeying $\sum_i J_i^2 = N$ and a set of $p = \gamma N$ input patterns $\boldsymbol{\xi}^\mu$ with components ξ_i^μ drawn independently of each other from the distribution

$$P(\xi) = \frac{1-m}{2}\delta(\xi+1) + \frac{1+m}{2}\delta(\xi-1). \tag{1}$$

To every input $\boldsymbol{\xi}^\mu$ the perceptron determines an output σ^μ according to

$$\sigma^\mu = \operatorname{sgn}\left(\frac{1}{\sqrt{N}}\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu\right) = \operatorname{sgn}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} J_i\xi_i^\mu\right). \tag{2}$$

Every set of input patterns therefore divides the coupling space into 2^p cells

$$C(\boldsymbol{\sigma}) = \{\boldsymbol{J}|\sigma^\mu = \operatorname{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu)\forall\mu\} \tag{3}$$

labelled by the sequence of outputs σ^μ. Note that some of the cells may be empty. The size $P(\boldsymbol{\sigma}) = V(\boldsymbol{\sigma})/\sum_{\boldsymbol{\tau}} V(\boldsymbol{\tau})$ of the cell gives the fraction of the coupling space that will produce the outputs σ^μ given the inputs ξ_i^μ.

It is convenient to characterize the cell sizes by a *crowding index* $\alpha(\boldsymbol{\sigma})$ defined by

$$P(\boldsymbol{\sigma}) = 2^{-N\alpha(\boldsymbol{\sigma})}. \tag{4}$$

The entropy of the distribution of cell sizes in the thermodynamic limit averaged over the input patterns is then given by

$$f(\alpha) = \lim_{N\to\infty}\frac{1}{N}\langle\langle\log_2\sum_{\boldsymbol{\sigma}}\delta(\alpha-\alpha(\boldsymbol{\sigma}))\rangle\rangle \tag{5}$$

where $\langle\langle\cdots\rangle\rangle$ denotes the average over the input pattern distribution (1). Involving a trace over *all* output sequences $\boldsymbol{\sigma}$ this quantity cannot be used to elucidate the dependencies on the output bias. In fact an explicit calculation shows that it is also independent of the input pattern distribution giving results for $f(\alpha)$ identical to those for $m = 0$. The intuitive reason for this fact is that a cell chosen at random from the above ensemble will, with probability 1, lie in the $(N-1)$-dimensional subspace of the coupling space which is orthogonal to the direction of the bias. However, the projection of the cell structure onto this subspace—whose properties dominate the cell spectrum in the thermodynamic limit—carries no bias.

Hence in order to study the influence of the input and output statistics on the geometry of the phase space we have therefore restricted the $\boldsymbol{\sigma}$-trace to outputs with magnetization m' according to

$$\begin{aligned} f(\alpha) &= \lim_{N\to\infty}\frac{1}{N}\left\langle\!\!\left\langle\log_2{\sum_{\boldsymbol{\sigma}}}'\delta(\alpha-\alpha(\boldsymbol{\sigma}))\right\rangle\!\!\right\rangle \\ &= \lim_{N\to\infty}\frac{1}{N}\left\langle\!\!\left\langle\log_2\sum_{\boldsymbol{\sigma}}\delta\left(\frac{1}{\gamma N}\sum_{\mu=1}^{\gamma N}\sigma^\mu - m'\right)\delta(\alpha-\alpha(\boldsymbol{\sigma}))\right\rangle\!\!\right\rangle. \end{aligned} \tag{6}$$

In the literature of multifractals $f(\alpha)$ is called the *multifractal spectrum*. It can be calculated by using its analogy with the microcanonical entropy of a spin system $\boldsymbol{\sigma}$ with Hamiltonian

$\alpha(\boldsymbol{\sigma})$ and the free energy

$$\tau(q) = -\lim_{N\to\infty}\frac{1}{N}\langle\langle \log_2 Z\rangle\rangle = -\lim_{N\to\infty}\frac{1}{N}\left\langle\!\!\left\langle \log_2 {\sum_{\boldsymbol{\sigma}}}' P^q(\boldsymbol{\sigma})\right\rangle\!\!\right\rangle$$
$$= -\lim_{N\to\infty}\frac{1}{N}\left\langle\!\!\left\langle \log_2 {\sum_{\boldsymbol{\sigma}}}' 2^{-qN\alpha(\boldsymbol{\sigma})}\right\rangle\!\!\right\rangle. \tag{7}$$

The quenched average over input patterns with magnetization m is performed using the pattern statistics (1). The entropy $f(\alpha)$ can now be obtained by a Legendre transformation with respect to the inverse temperature q

$$f(\alpha) = \min_q[\alpha q - \tau(q)]. \tag{8}$$

For the perceptron we have

$$P(\boldsymbol{\sigma}) = \int \mathrm{d}\mu(\boldsymbol{J}) \prod_{\mu=1}^{p} \theta\left(\frac{1}{\sqrt{N}}\sigma^\mu \boldsymbol{J}\cdot\boldsymbol{\xi}^\mu\right) \tag{9}$$

with the integral measure

$$\mathrm{d}\mu(\boldsymbol{J}) = \prod_i \frac{\mathrm{d}J_i}{\sqrt{2\pi e}}\delta(N - \boldsymbol{J}^2) \tag{10}$$

ensuring the spherical constraint for the coupling vectors and the total normalization over all cells $\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) = 1$. $\theta(x)$ is the Heaviside step function. In order to average $\log\sum P^q$ over the input patterns we introduce two sets of replicas; one set labelled $a = 1,\ldots,n$ for the standard replica trick to replace the log and one set labelled $\alpha = 1,\ldots,q$ representing the qth power of P in (7). Thus we arrive at the replicated partition function

$$Z_n = \langle\langle Z^n\rangle\rangle = \left\langle\!\!\left\langle \sum_{\{\sigma_\mu^a\}}\prod_a \delta\left(\sum_\mu \sigma_\mu^a - N\gamma m'\right)\int\prod_{a,\alpha}\mathrm{d}\mu(\boldsymbol{J}_a^\alpha)\prod_{\mu,a,\alpha}\theta\left(\frac{\sigma_\mu^a}{\sqrt{N}}\boldsymbol{J}_a^\alpha\cdot\boldsymbol{\xi}^\mu\right)\right\rangle\!\!\right\rangle. \tag{11}$$

Averaging over the quenched disorder results in the order parameters

$$Q_{a,b}^{\alpha,\beta} = \frac{1}{N}\boldsymbol{J}_a^\alpha\cdot\boldsymbol{J}_b^\beta$$
$$M_a^\alpha = \frac{1}{\sqrt{N}}\sum_i J_i^{a,\alpha}$$

as well as their conjugates $\hat{Q}_{a,b}^{\alpha,\beta}$ and $\hat{M}_a^\alpha$. In the following M_a^α will be referred to as the *weak* magnetization since the mean value of the couplings, the *strong* magnetization $1/N\sum_i J_i^{a,\alpha}$, vanishes in the thermodynamic limit. The appearance of an additional order parameter describing a ferromagnetic bias of order $\sqrt{N}$ was to be expected. To produce biased outputs, the local fields $\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}$ must have an average of order 1. Given $\langle\langle\xi_i^\mu\rangle\rangle = m$ this requires an average of the J_i of order $\sqrt{N}$. The integral representation of the delta function restricting the set of outputs introduces a further set of order parameters R_a.

In this paper we will only discuss the results obtained within the replica symmetric (RS) ansatz [5, 6]

$$Q_{a,b}^{\alpha,\beta} = \begin{cases} 1 & (a,\alpha) = (b,\beta) \\ Q_1 & a = b,\ \alpha\neq\beta \\ Q_0 & a\neq b \end{cases} \tag{12}$$
$$M_a^\alpha = M$$
$$R_a = R.$$

This ansatz represents the two-replica structure of the problem: Q_1 is assigned to the overlap between coupling vectors belonging to *identical cells* labelled by σ_μ^a, Q_0 is assigned to the overlap between *different* cells. For the cell spectrum without bias, it gives the correct result for $0 \leqslant q \leqslant 1$, which is the interval of interest here. Nevertheless it remains plagued by divergences for $q < 0$ and continuous replica symmetry breaking for $q > 1$, for details see [11].

Eliminating the conjugate order parameters one finds $Q_0 = 0$ is always an extremum of $\tau(q)$ since for $Q_0 = 0$ the saddle-point equation in Q_0 coincides with that in M. The interpretation of this result is that two randomly chosen coupling vectors with the same weak magnetization do not have an overlap of order 1.

We thus arrive at the free energy

$$\begin{aligned}\tau(q) = -\frac{1}{\ln 2}\mathrm{extr}_{Q_1,M,R}\Bigg[&\frac{q-1}{2}\ln(1-Q_1) + \frac{1}{2}\ln(1-(1-q)Q_1) - \gamma m' R \\ &+\gamma \ln\left(\mathrm{e}^R \int \mathrm{D}s\, H_+^q + \mathrm{e}^{-R}\int \mathrm{D}s\, H_-^q\right)\Bigg] \\ H_+ = H\left(\frac{\sqrt{Q_1}s - mM}{\sqrt{1-Q_1}}\right)& \\ H_- = H\left(\frac{-\sqrt{Q_1}s + mM}{\sqrt{1-Q_1}}\right)&\end{aligned} \tag{13}$$

where $\mathrm{D}s = \frac{\mathrm{d}s}{\sqrt{2\pi}}\exp(-s^2/2)$ is the Gaussian integral measure and $H(x) = \int_x^\infty \mathrm{D}s$.

Extremizing $\tau(q)$ with respect to Q_1,M, and R yields the saddle-point equations

$$\begin{aligned}\gamma\frac{\int \mathrm{D}s\,(\mathrm{e}^R H_+^{q-2} + \mathrm{e}^{-R}H_-^{q-2})G^2}{\int \mathrm{D}s\,(\mathrm{e}^R H_+^q + \mathrm{e}^{-R}H_-^q)} &= \frac{Q_1}{1-(1-q)Q_1} \\ \frac{\int \mathrm{D}s\,(\mathrm{e}^R H_+^{q-1} - \mathrm{e}^{-R}H_-^{q-1})G}{\int \mathrm{D}s\,(\mathrm{e}^R H_+^q + \mathrm{e}^{-R}H_-^q)} &= 0 \\ \frac{\int \mathrm{D}s\,(\mathrm{e}^R H_+^q - \mathrm{e}^{-R}H_-^q)}{\int \mathrm{D}s\,(\mathrm{e}^R H_+^q + \mathrm{e}^{-R}H_-^q)} &= m'\end{aligned} \tag{14}$$

where $G = 1/\sqrt{2\pi}\exp(-\frac{(\sqrt{Q_1}s - mM)^2}{2(1-Q_1)})$.

The cell spectrum $f(\alpha)$ can now be evaluated using (8).

3. Discussion

For $m' = 0$ the cell spectrum of unbiased patterns is recovered for any m. For $m = 0$ no coupling vector with a weak magnetization which produces a sequence of outputs with $m' \neq 0$ exists. We thus turn to the case $m' \neq 0, m \neq 0$. A transformation $mM \to M$ in the free energy (13) would remove the magnetization of the inputs from the saddle-point equations (14). Thus the properties of the cell spectrum for non-zero m and m' do not depend on m but only on γ and m' whereas the weak magnetization of the couplings M scales with m^{-1} for fixed m' and γ. Hence we may restrict the discussion to the case $m = m' \neq 0$ without loss of generality.

Figure 1 shows the cell spectrum at several loading capacities γ and magnetizations m'. For any given $q \equiv \mathrm{d}f/\mathrm{d}\alpha$ the number of cells decreases exponentially with increasing m'. This is in accordance with the fact that the maximal possible number of cells with output bias m' scales as $2^{Nf_{\mathrm{tot}}}$ with $f_{\mathrm{tot}} = \gamma((1-m')/2\log_2(1-m')/2 + (1+m')/2\log_2(1+m')/2)$.

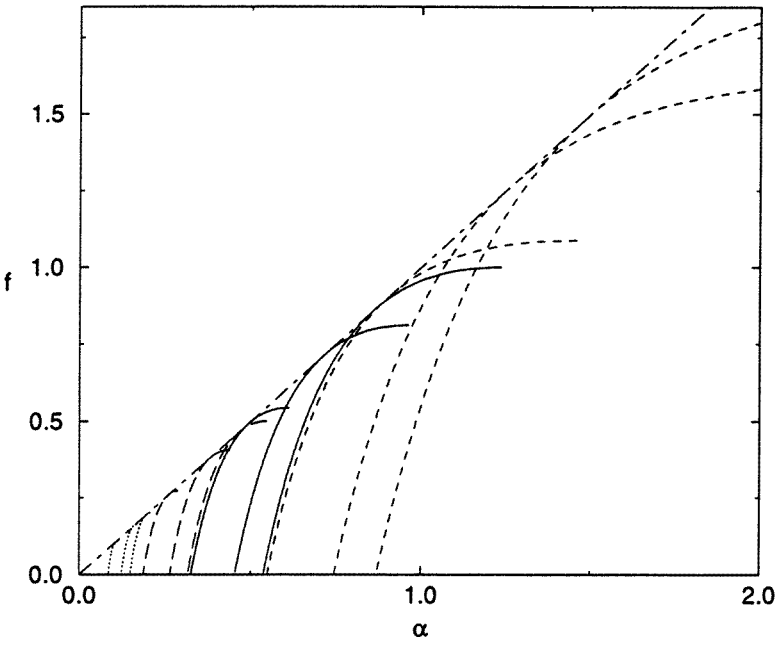

Figure 1. The multifractal spectrum $f(\alpha)$ for various values of the loading parameter γ =0.2 (dotted), 0.5 (long broken), 1 (full), and 2 (broken) and values of the magnetization $m' = 0, 0.5, 0.75$ from top to bottom respectively. The parts with negative slope have been omitted since their interpretation is presently not clear (cf [11]).

The maximum f_{max} of $f(\alpha)$ exponentially dominates the *number of cells* $\mathcal{N} = \int \mathrm{d}\alpha\, 2^{Nf(\alpha)}$, hence a randomly chosen output sequence will result in a cell of size $\alpha(q = 0)$, which is termed a *storage cell*. For values of the loading parameter below the critical storage capacity γ_c typically all possible cells may be realized, so $f_{\max} = f_{\mathrm{tot}}$. However, above the critical storage capacity only an exponentially small fraction of all possible cells may be realized hence $f_{\max} < f_{\mathrm{tot}}$. Note that although $f_{\max}$ decreases with increasing m' as shown in figure 1 it does so slower than f_{tot} so that the storage capacity *increases*. This can also be seen by comparing the curves of $f(\alpha)$ at $\gamma = 2$ for $m' = 0$ and $m' = 0.75$. The maximum of $f(\alpha)$ is attained for $m' = 0$ as $\alpha \to \infty$ indicating the cell volume shrinks to zero which signals the critical storage capacity. However, for $m' = 0.75$ the maximum of $f(\alpha)$ is reached at finite values of α indicating a finite size of a storage cell at $\gamma = 2$. In fact the limit $q \to 0$ of (13) can be taken explicitly yielding the saddle-point equations for the storage problem for magnetized patterns [2].

On the other hand the cells dominating the *volume* $\mathcal{V} = \int \mathrm{d}\alpha\, 2^{N(f(\alpha)-\alpha)}$ of the coupling space are characterized by $q \equiv \frac{\mathrm{d}f}{\mathrm{d}\alpha} = 1$. In general such a cell is taken to describe the generalization behaviour of the perceptron, since in the thermodynamic limit a randomly chosen teacher perceptron will lie within a cell of this size with probability one. The saddle-point equations (14) at this point are

$$Q_1 = \gamma \int \mathrm{D}s\, (H_+^{-1} + H_-^{-1})G^2 \tag{15}$$

$$R = 0 \tag{16}$$

$$m' = 1 - 2H(mM). \tag{17}$$

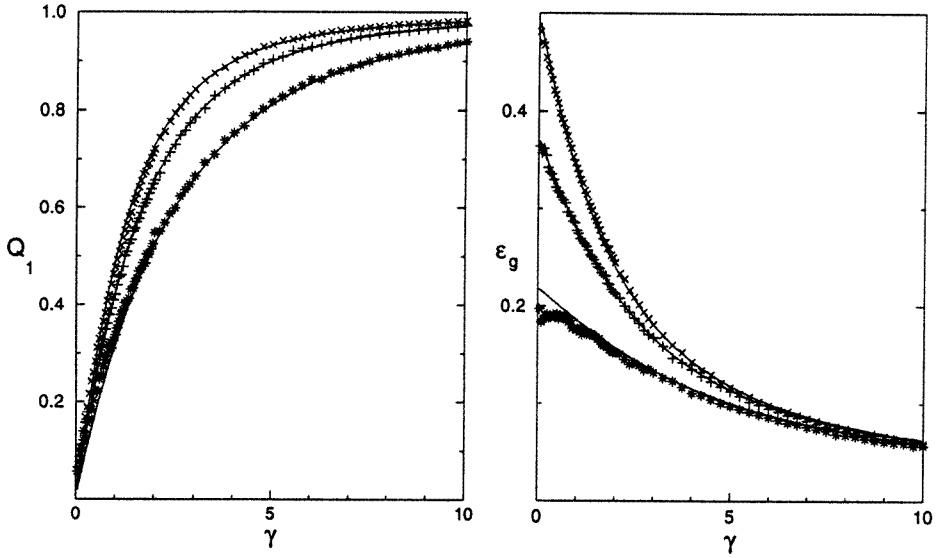

Figure 2. The teacher–student overlap Q_1 and the resulting generalization error ϵ_g as a function of γ for $m' = 0, 0.5, 0.75$ (from top to bottom). The uppermost curve corresponds to the results of Györgyi and Tishby [12]. The full lines are analytical curves whereas the symbols are the results of numerical simulations with $N = 200$ averaged over 200 patterns. The symbol size corresponds to 5 times the statistical error.

The interpretation of this saddle point may not be immediately obvious, since we have specified the magnetization of the outputs, but no properties of a teacher that will produce such a set of outputs on a given input pattern. However, equation (17) provides an explicit relation between m, m' and M. It gives the weak magnetization M of the couplings of a teacher or student required to produce a magnetization m' of the outputs given a set of inputs with magnetization m. In the context of a teacher with magnetization M acting on a set of inputs with magnetization m (17) simply follows from the central-limit theorem and Q_1 describes the overlap between teacher and student after the student has learned to classify γN examples correctly. The subsequent generalization error ϵ_g is given by

$$\epsilon_g = 1 - \frac{2}{\sqrt{1-Q_1^2}} \int_0^\infty \int_0^\infty \frac{\mathrm{d}\chi\, \mathrm{d}\sigma}{2\pi} \exp\left(-\frac{\sigma^2 - 2\sigma\chi Q_1 + \chi^2 + 2m^2M^2(1-Q_1)}{2(1-Q_1^2)}\right) \times \cosh\left(mM\frac{\chi+\sigma}{1+Q_1}\right). \tag{18}$$

The full lines in figure 2 show the overlap Q_1 between student and teacher and the corresponding generalization error ϵ_g as a function of γ after the student has learned γN examples for different magnetizations $m' = m$. The generalization error is found to decrease for increasing m' at fixed γ. The effect is most pronounced at low values of γ; in particular we find $\epsilon_g(\gamma = 0, m') = \frac{1-m'^2}{2}$. In fact the weak magnetization of the couplings M is independent of the loading parameter and already takes on its non-zero value (for $m \neq 0$) at $\gamma = 0$. However, two independent output strings σ_T^μ and σ_J^μ with the same average m' differ on average in $(1 - m'^2)/2$ bits only. Hence $\epsilon_g < 0.5$ even for zero teacher–student overlap Q_1. Qualitatively this means that the student learns the correct *bias* of the outputs after a *non-extensive number* of examples already. A plausibility argument underlining

this effect is a follows: by the central-limit theorem the sum $1/\sqrt{N}\sum_i J_i\xi_i$ is a Gaussian variable with mean mM and variance 1. Hence, equation (17) needs to be fulfilled if an output of the same sign as m' is to be produced with probability $(1+m')/2$. If the number of examples p becomes infinite in the thermodynamic limit the number of outputs of the same sign as m' tends to p times this probability. This implies that it is sufficient for the number of examples to scale with N^δ, $0 < \delta < 1$, so the number of examples is infinite in the thermodynamic limit, for M to take on its saddle-point value.

4. Simulation algorithm and numerical results

Gibbs learning at $T = 0$ is a convenient tool for the analytical study of generalization problems, since it characterizes the typical performance of a compatible student. It is, however, not completely straightforward to implement in numerical simulations. The necessary average over the compatible students cannot be performed directly because the version space is only an exponential fraction of the high-dimensional coupling space of the perceptron. Several methods to circumvent this problem have been suggested, including a random walk in the version space of the student [13], a billiard in version space [14], or a variant of the Adatron algorithm, where in each realization a few randomly chosen patterns are learnt in addition to the examples [15].

Here we propose the *randomized perceptron algorithm* as a new method to effectively simulate Gibbs learning and apply it to the specific problem of biased patterns. Starting with all couplings equal to zero, the randomized perceptron algorithm runs over all examples, leaving the couplings unchanged if the pattern is classified correctly by their present values. If a pattern is not classified correctly, the algorithm adds the standard Hebbian term as well as a random vector with components chosen independently from a Gaussian distribution. Since for large N the random vectors are all orthogonal to each other, the coupling vector will end up in version space even though the amplitudes of the Hebbian and the random terms are of comparable magnitude. This procedure slows down the convergence of the perceptron algorithm but leads to more reliable results for the generalization error. The standard deviation of the noise term was 4.8γ (compared with the magnitude of the Hebbian term of 2γ) but no strong dependence on the standard deviation was observed on the interval 2γ to 10γ.

The simulations whose results are shown in figure 2 were performed with a system size of $N = 200$. Sets of Gaussian distributed inputs with magnetization m were generated and the components of the couplings of potential teacher perceptrons were chosen from a Gaussian distribution with zero mean. In this way each teacher perceptron was given a weak magnetization. Teachers were generated until one of them produced an output magnetization m' on the given inputs. This teacher was used to generate the outputs used in the subsequent step. The resulting patterns were taught to the student using the randomized perceptron algorithm and its overlap with the teacher and the generalization error were evaluated.

Except at large values of the magnetization m', where finite-size effects are more noticeable, the numerical results are in very good agreement with the analytical curves.

5. Summary

In this paper we have investigated the influence of a bias in the distribution of inputs and outputs on the cell structure in the phase space of a perceptron. To this end the multifractal

spectrum $f(\alpha)$ was calculated analytically for different values of the storage ratio γ with the help of the methods used already for the case of unbiased patterns. Both the storage and the generalization properties can be read off from the behaviour of $f(\alpha)$.

For the storage problem we showed that the input bias has no influence on the storage capacity provided the output bias m' is equal to zero. If both input and output bias are non-zero the storage capacity γ_c increases with increasing output bias m' irrespectively of the value of the input bias. Biased patterns have been considered before in the context of the phase-space volume of attractor neural networks [2]. In this case it is natural to assume $m = m'$. Our results show that for the perceptron this case is in fact generic as for non-zero m the properties of the entire cell spectrum only depend on m'. The case $m = 0$ but $m' \neq 0$ cannot be realized by a perceptron with weak magnetization of the couplings without thresholds [16] and thus was not treated here. The behaviour of the maximum $f_{\max}$ of $f(\alpha)$ as a function of m' generalize the results about the number of dichomoties [1] to $m' \neq 0$.

For the generalization problem we found an interesting two-level scenario of learning. The student first determines the weak magnetization of the teacher couplings necessary to produce outputs of the required bias. This is accomplished after a non-extensive number of training examples already. The curves for the generalization error therefore start off at $\gamma = 0$ with values smaller than 0.5. In the second step, the student then reduces the generalization error further in the usual way. The asymptotic behaviour is not modified by the bias of the patterns. The analytical results are in excellent agreement with numerical simulations. We have found a randomized variant of the perceptron algorithm, a simple device for reliable simulations of Gibbs learning.

Biased patterns can be viewed as the simplest example of a hierarchy of inputs. It would be interesting to see whether the generalization strategy observed can also be found in the general case of hierarchically correlated patterns in the sense that the student first learns the classes of patterns and only then the individual representatives.

Acknowledgments

Many thanks to Martin Weigt for stimulating discussions and critical reading of the manuscript. JB gratefully acknowledges financial support by the *Studienstiftung des Deutschen Volkes*.

References

[1] Cover T M 1965 *IEEE Trans. Electron. Comput.* **EC-14** 326
[2] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[3] Derrida B, Griffith R B and Prügel-Bennett A 1991 *J. Phys. A: Math. Gen.* **24** 4907
[4] Opper M and Kinzel W 1996 *Models of Neural Networks* vol III, ed E Domany, J L van Hemmen and K Schulten (New York: Springer)
[5] Monasson R and O'Kane D 1994 *Europhys. Lett.* **27** 85
[6] Engel A and Weigt M 1996 *Phys. Rev.* E **53** R2064
[7] Monasson R and Zecchina R 1995 *Phys. Rev. Lett.* **75** 2432
[8] Biehl M and Opper M 1995 *Neural Networks: The Statistical Mechanics Perspective* ed Jong-Houn Oh, Chulan Kwon, and Sungzoon Cho (Singapore: World Scientific)
[9] Cocco S, Monasson R and Zecchina R 1996 *Phys. Rev.* E **54** 717
[10] Bex G J and van den Broeck C 1997 *Phys. Rev.* E **56** 870
[11] Weigt M and Engel A 1997 *Phys. Rev.* E **55** 4552
[12] Györgyi G and Tishby N 1990 *Workshop on Neural Networks and Spin Glasses* ed K Theumann and W K Koeberle (Singapore: World Scientific)

[13] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[14] Ruján P 1997 *Neural Comput.* **9** 99
[15] Watkin T 1993 *Europhys. Lett.* **21** 871
[16] Biehl M and Opper M 1991 *Phys. Rev.* A **44** 6888