IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

A fully connected committee machine learning unrealizable rules

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1995 J. Phys. A: Math. Gen. 28 7097

(http://iopscience.iop.org/0305-4470/28/24/010)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.68
The article was downloaded on 02/06/2010 at 00:58

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. 28 (1995) 7097–7104. Printed in the UK

A fully connected committee machine learning unrealizable rules

R Urbanczik†

Nordita, Blegdamsvej 17, DK-2100 Copenhagen Ø, Denmark

Received 17 May 1995

Abstract. We study generalization in a large fully connected committee machine with continuous weights trained on patterns with outputs generated by a teacher of the same structure but corrupted by noise. The corruption is due to additive Gaussian noise applied in the input layer or the hidden layer of the teacher. Contrary to related cases, in the presence of input noise the generalization error ϵ_g is not minimized by the teacher's weights. For small values of the load parameter α the student is in a permutation-symmetric phase. As α increases three additional phases emerge. The large-α theory of the stable phase is similar to the tree committee machine. In particular, at zero temperature in the presence of noise ϵ_g does not approach its minimal value $\epsilon_{\min}$ and the student's weights do not converge to those of the teacher. For a positive temperature $\epsilon_g - \epsilon_{\min}$ decays as a power of α, the exponent being the same as in the corresponding case of the tree. However, for all values of α an at least metastable phase exists which is permutation symmetric with respect to the teacher.

1. Introduction

The calculation of the generalization ability of feedforward neural networks has been a subject of considerable interest. Here we extend this work to the case of the fully connected committee machine learning specific instances of an unrealizable rule. The corresponding realizable case has been discussed within the annealed approximation in [4, 7] and using the replica formalism in [6]. The analogous questions for the unrealizable case have been considered in [8] for the tree committee machine and in [3] for the perceptron.

The connected machine has N real inputs (ξ_j) and K hidden units, each characterized by a weight vector $J_i \in \mathbb{R}^N$. Its output is given by

$$\tau_J(\xi) = \mathrm{sign}\left(\sum_{i=1}^{K} \mathrm{sign}(J_i^T \xi)\right). \tag{1}$$

We may assume that $|J_i| = 1$. The weight vectors are to be chosen such that τ_J approximates well a target concept (the teacher) which in this paper will be assumed to be given by a machine of a similar structure. The teacher is also a committee machine with K hidden units and orthonormal weight vectors J_i^0 but its output

$$\tau_{J^0}(\xi, \eta) = \mathrm{sign}\left(\eta_{K+1} + \sum_{i=1}^{K} \mathrm{sign}(\eta_i + J_i^{0^T} \xi)\right) \tag{2}$$

† E-mail address: urbanczik@urz.unibas.ch

0305-4470/95/247097+08$19.50 © 1995 IOP Publishing Ltd

is corrupted by noise. The noise terms η_k are assumed to be independent zero-mean Gaussian random variables with variance:

$$\langle\eta_k^2\rangle = \begin{cases} \dfrac{1-\gamma_1^2}{\gamma_1^2} & k \leqslant K \\ \dfrac{1-\gamma_2^2}{\gamma_2^2}K & k = K+1 \end{cases} \tag{3}$$

for values of the γ_i between 0 and 1. If $\gamma_1\gamma_2 = 1$ the output of the teacher is deterministic and we recover the realizable case. For $\gamma_1\gamma_2 = 0$ it is independent of the input ξ and we have the random map problem, some aspects of which have been discussed in [1, 2]. The noise in the input layer can be thought of as stemming from Gaussian noise added to each of the inputs ξ_j and in this case the assumed independence of the η_i is a consequence of the orthogonality of the teacher vectors.

A training set of P examples (ξ^μ, τ^μ) is constructed by independently picking inputs ξ_j^μ (from the normal distribution) as well as noise terms η_k^μ and assigning outputs t^μ by (2). The training energy $E(J) = \sum_\mu \theta(-\tau^\mu\tau_J(\xi^\mu))$, where θ is the Heaviside step function, then measures the performance of a student with weight vectors J_i on the training set. This is used to define on the space of students a probability density $p(J)$ with Boltzmann weight $e^{-\beta E(J)}$, where $\beta = 1/T$ is the inverse temperature. One hopes that for sufficiently large P a student picked from $p(J)$ will perform well on new input/output pairs constructed in the same manner as those in the training set. So the student should have minimal generalization error $\epsilon_g(J)$, where ϵ_g is the average of $\theta(-\tau_J(\xi)\tau_{J^0}(\xi,\eta))$ over noise terms η_k and normally distributed inputs ξ_j.

A different measure of the student's performance is the distance between its weight vectors and those of the teacher (after a suitable reordering). This will be closely related to ϵ_g if $\epsilon_g(J^0) = \min_J \epsilon_g(J)$. This is the case for the perceptron [3] and the tree committee. It will turn out not to be true for the fully connected machine in the presence of input noise since the student can adapt to the fact that $\tau_{J^0}(\xi, 0)$ and $\tau_{J^0}(\xi,\eta)$ may be anticorrelated for some inputs ξ.

2. Order parameters and generalization error

To study typical properties of the student space we calculate the quenched average of the n-times replicated Gardner volume. This leads to a symmetric $(K+nK, K+nK)$-matrix

$$q_{ij}^{ab} = J_i^{aT} J_j^b \qquad a, b = 0, \ldots, n \tag{4}$$

where $a = 1, \ldots, n$ indexes the replicas and $a = 0$ the teacher. The order parameters are the non-constant entries of this matrix. As in the realizable case the symmetries of the problem suggest a site-symmetric parametrization of the order parameter matrix:

$$q_{ij}^{ab} = \tilde{p}^{ab} + \delta_{ij} q^{ab}\,. \tag{5}$$

We shall call the $\tilde{p}^{ab}$ the permutation symmetric and the q^{ab} the specialized overlaps. In the appendix it is shown that using this parametrization the expression for the replicated Gardner volume becomes quite similar to the one for the perceptron in the limit of large K and with the scaling assumption $\tilde{p}^{ab} = \mathcal{O}(1/K)$ which arises naturally in the course of the derivation.

A subsequent parametrization with one step of replica-symmetry breaking leads to the specialized order parameters q_0, q_1, R and to rescaled permutations symmetric ones $p_0, p_1, \bar{p}, \bar{R}$. Here R and $\bar{R}$ denote the student/teacher overlaps, q_i and p_i the student/student overlaps in different replicas, and $\bar{p}$ is the overlap between hidden units in the same replica. Note, that p_0, p_1, $\bar{R}$ can have any real value because of the rescaling (A7) and that $\bar{p} \geqslant -1$. It is convenient to introduce the abbreviations:

$$
\begin{aligned}
q_{i\mathrm{e}} &= \frac{2}{\pi}(p_i + \arcsin(q_i)) & q_i^{\mathrm{e}} &= p_i + q_i \\
R_{\mathrm{e}} &= \frac{2}{\pi}\gamma_2(\gamma_1 \bar{R} + \arcsin(\gamma_1 R)) & R^{\mathrm{e}} &= \bar{R} + R\,.
\end{aligned} \tag{6}
$$

The generalization error depends only on the overlaps between the student and the teacher and is for large K:

$$
\epsilon_g(J) = \frac{1}{\pi}\arccos\left(\frac{R_{\mathrm{e}}}{\sqrt{1+\frac{2}{\pi}\bar{p}}}\right). \tag{7}
$$

Applying the Cauchy–Schwarz inequality to $\sum_i J_i^T \sum_i J_i^0$ shows that $R^{\mathrm{e}2} \leqslant 1 + \bar{p}$. From this it is easy to see that the minimum $\epsilon_{\min}$ of ϵ_g is attained at $R = 1$, $\bar{p} = \bar{p}_s$ and is given by

$$
\begin{aligned}
\epsilon_{\min} &= \frac{1}{\pi}\arccos\left(\gamma_1\gamma_2\sqrt{\frac{2}{\pi} + \left(1 - \frac{2}{\pi}\right)\frac{1}{1+\bar{p}_s}}\right) \\
\bar{p}_s &= \left(\frac{\pi}{2} - 1\right)^2 (\gamma_1^{-1}\arcsin\gamma_1 - 1)^{-2} - 1\,.
\end{aligned} \tag{8}
$$

The optimal value $\bar{p}_s$ is zero if $\gamma_1 = 1$ and diverges for $\gamma_1 \to 0$. So the teacher's weights give an optimal generalization error only if there is no input noise. Moreover, the dependence of ϵ_g on R vanishes as $\bar{p} \to \infty$. So in the limit of high input noise we may think of the optimal student as being the perceptron obtained by averaging the teacher's weight vectors.

3. One-step RSB theory

Within the one-step ansatz the free energy F per weight can be written as

$$
\begin{aligned}
-\beta F &= \mathrm{extr}\frac{P}{KN}G_r(R_{\mathrm{e}}, q_{0\mathrm{e}}, q_{1\mathrm{e}}, \bar{p}, m) + G_s(R, \bar{R}, (q_i), (p_i), \bar{p}, m) \\
G_r &= \frac{2}{m}\int \mathrm{D}x\, H\left(\frac{R_{\mathrm{e}}}{\sqrt{q_{0\mathrm{e}} - R_{\mathrm{e}}^2}}x\right)\ln\int \mathrm{D}y\left[\mathrm{e}^{-\beta} + (1-\mathrm{e}^{-\beta})H(z)\right]^m \\
z &\equiv \frac{\sqrt{q_{0\mathrm{e}}}x - \sqrt{q_{1\mathrm{e}} - q_{0\mathrm{e}}}y}{\sqrt{1+\frac{2}{\pi}\bar{p} - q_{1\mathrm{e}}}} \\
G_s &= \frac{K-1}{2K}S(R, q_0, q_1, \bar{p}/(1-K)) + \frac{1}{2K}S(R^{\mathrm{e}}, q_0^{\mathrm{e}}, q_1^{\mathrm{e}}, \bar{p}) \\
S(R, q_0, q_1, \bar{p}) &= \frac{q_0 - R^2}{1 + \bar{p} - q_1 + m(q_1 - q_0)} + \frac{m-1}{m}\ln(1 + \bar{p} - q_1) \\
&\quad + \frac{1}{m}\ln(1 + \bar{p} - q_1 + m(q_1 - q_0))\,.
\end{aligned} \tag{9}
$$

If the permutation-symmetric parameters are zero, these expression become identical to the ones found for the tree committee machine [8]. Similarly as for the RS equations [6] the stationarity conditions for p_0, p_1 and $\bar{p}$ imply

$$1 + \bar{p} - q_1^{\mathrm{e}} + m(q_1^{\mathrm{e}} - q_0^{\mathrm{e}}) = \mathcal{O}(1/K) \tag{10}$$

$$q_0^{\mathrm{e}} - R^{\mathrm{e}2} = \mathcal{O}(1/K)\,. \tag{11}$$

The remaining equation admits the solution

$$q_1^{\mathrm{e}} - q_0^{\mathrm{e}} = \mathcal{O}(1/K)\,. \tag{12}$$

This is the physical solution since $1 + \bar{p} - q_i^{\mathrm{e}}$ is non-negative (it is the length of the vector $K^{-1/2}\sum_k(J_k^a - J_k^b)$ if $J_k^{aT}J_k^b = q_i$) and this, together with (10), implies (12). For large K and a finite value of the load parameter $\alpha = \frac{P}{KN}$ these relations allow us to eliminate three of the four permutation symmetric parameters. Further, stationarity with respect to $\bar{R}$ yields

$$2R^{\mathrm{e}}\left(\frac{\partial}{\partial p_0} + \frac{\partial}{\partial p_1} + \frac{\partial}{\partial \bar{p}}\right)G_r + \frac{\partial}{\partial \bar{R}}G_r = 0 \tag{13}$$

which is independent of α.

As a consequence of (10)–(12) for $q_1 \to 1$ the same asymptotic relationship holds between the one-step and the replica-symmetric (RS) theory as in the tree committee [8]. In particular, for $T = 0$ and in the presence of noise the RS theory gives for large α an asymptotically correct ϵ_g but an incorrect value of the free energy. For finite β the one-step theory becomes equivalent to the RS theory at inverse temperature $m\beta$ as $\alpha \to \infty$, where m must be chosen such that the RS entropy decreases only logarithmically with α.

The stationarity conditions for the specialized parameters admit the permutation symmetric solution $q_i = R = 0$. This solution is locally stable against fluctuations in the specialized parameters since $\frac{\pi}{2}q_{i\mathrm{e}} - q_i^{\mathrm{e}}$ is proportional to q_i^3 for small q_i and similarly for R^{e} and R_{e}. In view of (12) it must be replica-symmetric. Using the analogy to the perceptron described in the appendix and the results in [3] this may easily be confirmed by evaluating the AT condition. The generalization error of the permutation symmetric solution is independent of α and equals

$$\epsilon_g = \begin{cases} \dfrac{1}{\pi}\arccos\left(\gamma_1^2\gamma_2^2\dfrac{2}{\pi}\right) & T = 0 \\[2ex] \dfrac{1}{\pi}\arccos\left(\gamma_1\gamma_2\sqrt{\dfrac{2}{\pi}}\right) & T \to \infty\,. \end{cases} \tag{14}$$

As already observed in [6] the overlap $\bar{p}$ between different hidden units is zero for $\gamma_1\gamma_2 = 1$ and $T = 0$. It increases with $\gamma_1\gamma_2$ and at $\gamma_1\gamma_2 = T = 0$ one finds $\bar{p} = -1$. A similar anticorrelation of the hidden units has been found in the random map problem of the $K = 3$ committee machine [1].

The argument for the local stability of the permutation symmetric phase allows any combination of the specialized parameters to be zero as long as this does not violate the stability of the entropy term. There are four possibilities which can all yield locally stable solutions:

$$\begin{array}{lll} \text{A:} & q_1 = q_0 = R = 0 & \text{permutation symmetric,} \\ \text{B:} & q_1 > q_0 = R = 0 & \\ \text{C:} & q_1 \geqslant q_0 > R = 0 & \\ \text{D:} & q_1 \geqslant q_0 \geqslant R > 0 & \text{specialized}\,. \end{array} \tag{15}$$

Note, that solutions of type B are, by definition, not replica-symmetric while C and D can be. We shall not attempt a full description of this rich phase structure here but highlight some of the main points, focusing on $T = 0$ and the random map problem as well as the realizable case.

In the random map problem we find a transition from A to B at $\alpha \approx 4.91$ with $m = 1$ and q_1 close to 1 at the critical point. A similar continuous transition from a locally stable replica-symmetric phase to one with broken replica symmetry has been found previously in the random map problem of the binary perceptron [5] at positive temperatures. At $\alpha \approx 15.4$ a transition from B to a phase C with broken replica symmetry occurs, accompanied by a discontinuous increase in q_0. Even in this last phase $\bar{p}$ is equal to its minimal possible value -1. So the anticorrelation of the hidden units maximizes the storage capacity.

In the realizable case a discontinuous transition to a replica-symmetric phase D was found at $\alpha \approx 7.65$ for $T = 0$ in [6]. For higher α this solution describes the stable state. However, the permutation symmetric solution does not describe the metastable state correctly for large α. Assuming replica symmetry $q_1 = q_0 = q$ we find that the maximum with respect to q of the free energy at $q = 0$ is only a local one above $\alpha \approx 17.0$. So a transition to phase C occurs, accompanied by a small but discontinuous increase in ϵ_g. Indeed, ϵ_g continues to rise and for $\alpha \to \infty$ we find $\bar{R} \approx 0.681$ as compared to the permutation-symmetric prediction $\bar{R} = 1$. Further, $\bar{p}$ is negative in phase C, so the student is finding a compromise between having a high overlap with the average of the teacher's vectors and maximizing its storage capacity. The replica-symmetric learning curve is shown in figure 1.

Considering the full one-step equations, still for $\gamma_1\gamma_2 = 1$, we find that permutation symmetry is broken in the metastable state already above $\alpha \approx 7.68$. Here a transition to phase B occurs, with $m = 1$ at the critical point as in the random map problem. The asymptotic relationship between the one-step and the RS theory shows that for some higher α there will be a transition from B to a phase C with broken replica symmetry. The generalization error will approach the value of the RS prediction for this phase as $\alpha \to \infty$. So even in the one-step description, a student staying in the metastable state will display a non-monotonic ϵ_g.

The large-α theory of the stable state in the general case $0 < \gamma_1\gamma_2 < 1$ is similar to the one of the tree committee machine. In particular, at zero temperature the specialized overlap R does not approach 1 as $\alpha \to \infty$ and thus neither $\epsilon_g \to \epsilon_{\min}$ nor $J \to J^0$. In contrast to the tree, a positive value of R is only achieved at $T = 0$ for low levels of noise. For higher levels, even as $\alpha \to \infty$, the stable state has $R = 0$ and is of type C. Examples of this behaviour are shown in figure 2. For $T > 0$ we do find $R \to 1$ with increasing α, and for $q_i = 1$ condition (13) requires $\bar{p}$ to be chosen so as to minimize ϵ_g. So the same exponent in the power law for the decrease of $\epsilon_g - \epsilon_{\min}$ is found as for the tree. In

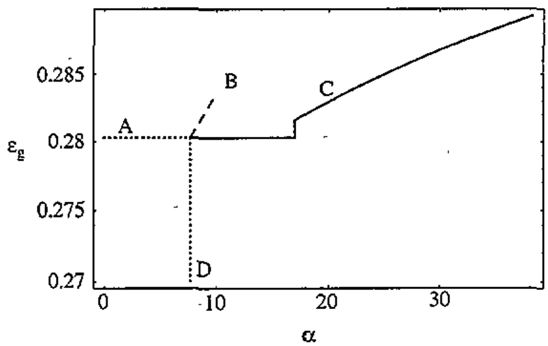

Figure 1. Replica-symmetric learning curves for $\gamma_1\gamma_2 = 1$ and $T = 0$. The dotted line corresponds to the stable, the full curve to the metastable state. The broken line hints at the one-step corrections for the metastable state. The broken line and full curve meet at $\alpha = \infty$ and $\epsilon_g \approx 0.321$.

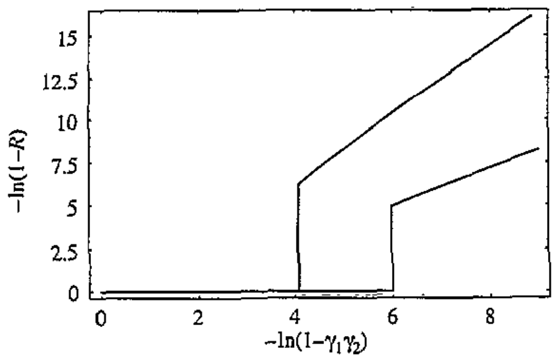

Figure 2. Asymptotic value of R for $\alpha \to \infty$ as a function of $\gamma_1\gamma_2$. The upper curve is for $\gamma_1 = 1$, the lower one for $\gamma_2 = 1$. The transition to $R = 0$ occurs at $\gamma_2 \approx 0.983$ in the upper and at $\gamma_1 \approx 0.9977$ in the lower curve. Note the logarithmic scales used in the plot.

particular, the one-step equations yield:

$$\epsilon_g - \epsilon_{\min} \propto \begin{cases} \alpha^{-2/5} & \gamma_1 < 1 \\ \alpha^{-2/3} & \gamma_1 = 1 \end{cases} \tag{16}$$

and thus, in contrast to the tree, the student's weights converge to those of the teacher only if $\gamma_1 = 1$.

This last point, along with the rich phase structure, is perhaps the most striking difference to the tree committee machine. It should be pointed out that, since the the teacher vectors are orthogonal, in the present case we may even assume the teacher to be a tree committee. So the target concept can be thought of as being the same in the two cases. The difference arises from the fact that in the present case the student space is not constrained to orthogonal vectors and this can allow the student to improve on the noiseless teacher in the unrealizable setting.

Acknowledgments

I thank J Hertz for stimulating discussions and his valuable advice. This work was supported by the E Batschelet-Mader Foundation and the M Geldner Foundation.

Appendix

By standard arguments the quenched average of the replicated Gardner volume $\langle V^n \rangle$ is for large N:

$$\ln\langle V^n \rangle \sim N \operatorname*{extr}_{q^{ab}_{ij}} \frac{P}{N} G_r^{(n)}(q) + G_s^{(n)}(q)\,. \tag{A1}$$

We first discuss $G_r^{(n)}$ which may be written as

$$\begin{aligned} G_r^{(n)} &= \ln\left\langle \mathcal{E}\left(K^{-1/2} \sum_{i=1}^{K} \operatorname{sign}(Z_i^a) \right) \right\rangle_{Z,\eta_{K+1}} \\ \mathcal{E}((x^a)) &\equiv \prod_{a=1}^{n} \exp\left(-\beta\theta\left(-x^a\left(K^{-1/2}\eta_{K+1} + x^0\right)\right)\right). \end{aligned} \tag{A2}$$

The Z_i^a are zero-mean Gaussian random variables (independent of η_{K+1}) with a covariance matrix $\bar{q}$ given by

$$\langle Z_i^a Z_j^b \rangle = \begin{cases} \gamma_1 q_{ij}^{ab} & a > 0 \quad b = 0 \\ q_{ij}^{ab} & \text{otherwise}. \end{cases} \tag{A3}$$

The assumption of site symmetry (5) then implies that a coordinate transformation exists such that

$$Z_i = A x_i + B K^{-1/2} \sum_{k=1}^{K} x_k \tag{A4}$$

where Z_i denotes the vector $(Z_i^0, Z_i^1, \ldots, Z_i^n)$ and the x_i^a are independent and normally distributed random variables. The $(n+1, n+1)$-matrices A and B need to satisfy

$$\begin{aligned} \langle Z_1 Z_1{}^T \rangle - \langle Z_1 Z_2{}^T \rangle &= A A^T \\ \langle Z_1 Z_1{}^T \rangle + (K-1)\langle Z_1 Z_2{}^T \rangle &= (A + K^{1/2} B)(A + K^{1/2} B)^T . \end{aligned} \tag{A5}$$

These equations have real solutions since (5) implies that eigenvalues of the matrices on the LHS of these equations are also eigenvalues of the entire covariance matrix $\bar{q}$.

We may thus rewrite $G_r^{(n)}$ as

$$\begin{aligned} G_r^{(n)} &= \ln \prod_{a=0}^{n} \int \mathrm{d}u^a \left\langle \prod_a \delta\left(u^a - \left(B K^{-1/2} \sum_i x_i\right)^a\right) \mathcal{E}\left(K^{1/2} m^a(u^a) + \chi^a(u^a)\right)\right\rangle_{x,\eta_{K+1}} \\ \chi^a(u^a) &\equiv K^{-1/2} \sum_i \left(\mathrm{sign}\left(u^a + (A x_i)^a\right) - m^a(u^a)\right). \end{aligned} \tag{A6}$$

The $m^a(u^a)$ should be chosen such that the mean of $\chi^a(u^a)$ is zero. Since the x_i^a are independent, the joint distribution of χ and $\hat{\chi} = K^{-1/2} \sum_i x_i$ will approach for large K a Gaussian one with covariance matrix $C(u)$. But if the u^a are of order 1, the argument of $\mathcal{E}$ will be dominated by $K^{1/2} m^a(u^a)$ in this limit. We assume this not to be the case and take the u^a to be of order $K^{-1/2}$ which is equivalent to the reparametrization

$$p^{ab} = K \tilde{p}^{ab} \qquad p^{aa} = (K-1)\tilde{p}^{aa} \qquad (a > b). \tag{A7}$$

Further, this implies $C(u) \to C(0)$ as $K \to \infty$ and the integrals over the u^a in (A6) can be easily done. In the end we find for large K:

$$G_r^{(n)} \sim \ln \left\langle \prod_{a=1}^{n} \mathrm{e}^{-\beta\theta(-\mathcal{Z}^a \mathcal{Z}^0)} \right\rangle_{\mathcal{Z}} \tag{A8}$$

for zero-mean Gaussian random variables with covariance matrix

$$\langle \mathcal{Z}^a \mathcal{Z}^b \rangle = \begin{cases} 1 & a = b = 0 \\ 1 + \dfrac{2}{\pi} p^{aa} & a = b > 0 \\ \dfrac{2}{\pi}(p^{ab} + \arcsin q^{ab}) & a > b > 0 \\ \dfrac{2}{\pi}\gamma_2(\gamma_1 p^{ab} + \arcsin(\gamma_1 q^{ab})) & a > b = 0. \end{cases} \tag{A9}$$

But for the different dependence of the covariances on the order parameters, $G_r^{(n)}$ has the same form as in the case of the perceptron.

The calculation of the entropy term $G_s^{(n)}$ involves a symmetric (nK, nK)-matrix $\hat{q}$ of order parameters conjugate to the overlaps between students q_{ij}^{ab} $(a, b > 0)$ as well as a (nK, K)-matrix $\hat{R}$ of conjugates to the q_{ij}^{a0} $(a > 0)$. One finds

$$G_s^{(n)} = \operatorname*{extr}_{\hat{q},\hat{R}} -\tfrac{1}{2}nK + \sum_{a,i,j} \hat{R}_{ij}^a q_{ij}^{a0} + \tfrac{1}{2} \sum_{a,b,i,j} \hat{q}_{ij}^{ab} q_{ij}^{ab} - \tfrac{1}{2} \ln\det\hat{q} + \tfrac{1}{2}\operatorname{Tr}(J^0 \hat{R}^T \hat{q}^{-1} \hat{R} J^{0^T}) \tag{A10}$$

where J^0 is the (N, K)-matrix of teacher vectors. Since they are orthonormal the trace may immediately be simplified to $\operatorname{Tr}(\hat{R}^T \hat{q}^{-1} \hat{R})$. We assume the conjugate parameters to be site-symmetric as well:

$$\hat{q}_{ij}^{ab} = \hat{q}_P^{ab} + \delta_{ij}\hat{q}_S^{ab} \qquad \hat{R}_{ij}^a = \hat{R}_P^a + \delta_{ij}\hat{R}_S^a . \tag{A11}$$

Thinking of $\mathbb{R}^{nK}$ as $(\mathbb{R}^n)^K$, this implies that the subspace $\{(x, x, \dots, x) | x \in \mathbb{R}^n\}$ is stable under $\hat{q}$. This also holds for the subspaces $(x, -x, 0, \dots, 0)$, $(x, 0, -x, 0, \dots, 0)$, $\dots, (x, 0, \dots, 0, -x)$. Similarly $\hat{R}^T \hat{q}^{-1} \hat{R}$ has eigenvectors $(1, 1, \dots, 1)$, $(1, -1, 0, \dots, 0)$ etc and hence

$$\begin{aligned} \operatorname{Tr}(\hat{R}^T \hat{q}^{-1} \hat{R}) &= (\hat{R}_S + K\hat{R}_P)^T (\hat{q}_S + K\hat{q}_P)^{-1} (\hat{R}_S + K\hat{R}_P) + +(K-1)\hat{R}_S^T \hat{q}_S^{-1} \hat{R}_S \\ \det\hat{q} &= \det(\hat{q}_S + K\hat{q}_P) \det\hat{q}_S^{K-1} . \end{aligned} \tag{A12}$$

A linear transformation in the conjugate parameters then leads to

$$\begin{aligned} G_s^{(n)} &= (K-1) G_{s,\mathrm{P}}((q^{ab})) + G_{s,\mathrm{P}}((q^{ab}) + K(\bar{p}^{ab})) \\ G_{s,\mathrm{P}}^{(n)}((q^{ab})) &= \operatorname*{extr}_{\hat{q}_S, \hat{R}_S} -\tfrac{1}{2}n + \sum_a \hat{R}_S^a q^{a0} + \tfrac{1}{2}\sum_{a,b} \hat{q}_S^{ab} q^{ab} - \tfrac{1}{2}\ln\det\hat{q}_S + \tfrac{1}{2}\hat{R}_S^T \hat{q}_S^{-1} \hat{R}_S . \end{aligned} \tag{A13}$$

where $G_{s,\mathrm{P}}^{(n)}$ is essentially the entropy term of the perceptron.

References

[1] Barkai E, Hansel D and Sompolinsky H 1992 Broken symmetries in multilayered perceptrons *Phys. Rev.* A **45** 4146–61

[2] Engel A, Köhler H M, Tschepke F, Vollmayr H and Zippelius A 1992 Storage capacity and learning algorithms for two layer neural networks *Phys. Rev.* A **45** 7590

[3] Györgyi G and Tishby N 1990 *Statistical Theory of Learning a Rule Neural Networks and Spin Glasses* ed K Thuemann and R Köberle (Singapore: World Scientific) pp 3–36

[4] Kang K, Oh J-H, Kwon C and Park Y 1993 Generalization in a two layer network *Phys. Rev.* E **48** 4805–9

[5] Krauth W and Mézard M 1989 Storage capacity of memory networks with binary couplings *J. Physique* **50** 3057–66

[6] Schwarze H 1993 Learning a rule in a multilayer neural network *J. Phys. A: Math. Gen.* **26** 5781–94

[7] Schwarze H and Hertz J 1993 Generalization in fully connected committee machines *Europhys. Lett.* **21** 785–90

[8] Urbanczik R 1995 A large committee machine learning unrealizable rules *Neural Comput.* submitted