Statistical mechanics of learning in the presence of outliers

1998 J. Phys. A: Math. Gen. 31 9131

(http://iopscience.iop.org/0305-4470/31/46/005)

J. Phys. A: Math. Gen. **31** (1998) 9131–9147. Printed in the UK PII: S0305-4470(98)92336-8

Statistical mechanics of learning in the presence of outliers

Rainer Dietrich† and Manfred Opper‡

† Physikalisches Institut, Julius-Maximilians-Universität, Am Hubland, D-97074 Würzburg, Germany

‡ Department of Computer Science and Applied Mathematics, Aston University, Birmingham B4 7ET, UK

Received 9 March 1998, in final form 22 June 1998

Abstract. Using methods of statistical mechanics, we analyse the effect of outliers on the supervised learning of a classification problem. The learning strategy aims at selecting informative examples and discarding outliers. We compare two algorithms which perform the selection either in a soft or a hard way. When the fraction of outliers grows large, the estimation errors undergo a first-order phase transition.

1. Introduction

The analysis of algorithms which allow one to learn a rule from random examples is an active and fascinating topic in the area of statistical mechanics. For an overview see e.g. [1–3]. Many models, where examples are *correctly* classified by ideal experts (often called teachers), seem to be well understood. Now there is a great deal of interest in nonideal, but more realistic models, which incorporate the influence of different types of noise in learning.

In this paper, we study a model where not all examples carry information about the unknown rule, but where a nonzero fraction of them are just outliers. Naively learning *all* examples may considerably deteriorate the ability to infer the rule in such a case. As with learning with noisy data, some knowledge about the stochastic data generating mechanism can be helpful. Based on such a stochastic model, a good algorithm could try to select the informative examples and discard the remaining ones. Since, however, only partial information is available, such a selection can only be performed approximately and it is natural to try a *soft*, probabilistic selection.

Our model leads naturally to such a selection method. It consists of a classification problem, where data which come from two distributions (classes) centred at different points are mixed at random with outliers. A Bayesian approach, which aims at calculating the most probable values for the class centres by minimizing a specific *training energy* is combined with the so-called expectation maximization (EM) algorithm of Dempster *et al* [4], which nicely deals with the problem of hidden parameters (the knowledge which of the data are informative) in data mixtures. This procedure leads to an algorithm which iteratively computes the probability that an example is informative and weights each example in predicting the unknown class centres of the data generating distributions. Our model may also be considered as a simple version of the *mixtures of experts* models [5] which are frequently studied in the neural network literature. In these models, a complicated task is learnt by a division of labour among several simple learning machines (experts), where

0305-4470/98/469131+17$19.50

each expert learns from different subsets of examples. Our model would correspond to two experts where only one is able to extract information from the examples.

The paper is organized as follows. After an introduction of the learning problem, two learning strategies are defined in section 2. Section 3 gives the statistical mechanics formulation of the problem which, based on a replica calculation, leads to a computation of the learning performance in the thermodynamic limit. In section 4 the algorithmic implementation of the learning methods using the EM algorithm is explained. Section 5 presents the results of the statistical mechanics calculations and of numerical simulations and concludes with a discussion. Details of the replica calculations are given in the appendices.

2. The learning problem

We assume that the examples $\{\boldsymbol{\xi}^\mu, S^\mu\}$ ($\boldsymbol{\xi}^\mu \in \mathbb{R}^N, S^\mu \in \{\pm 1\}$), $\mu = 1, \ldots, \alpha N$, are generated alternatively by two different processes. For the first process, the input $\boldsymbol{\xi}^\mu$ is selected at random from one of two Gaussian clusters (labelled by the outputs $S^\mu = \pm 1$) which are chosen with equal probability. The clusters are centred at $\pm \boldsymbol{B}$ and have equal variance $1/\gamma$. $\boldsymbol{B}$ is an N-dimensional vector with $\boldsymbol{B}^2/N = 1$. The joint probability for inputs and outputs corresponding to this process can be written as

$$\mathcal{P}(\boldsymbol{\xi}^\mu, S^\mu | \boldsymbol{B}) \propto \exp\left[-\frac{\gamma}{2} \sum_j \left(\xi_j^\mu - \frac{1}{\sqrt{N}} S^\mu B_j \right)^2 \right].$$

The data from this process represent classified examples in a noisy (because the Gaussian clusters overlap) two-class problem.

In the second process, the inputs come from a single Gaussian centred at zero with the same variance and the output (chosen ± 1 with equal probability) is completely independent from the input. For this case, we make the ansatz

$$\mathcal{P}(\boldsymbol{\xi}^\mu, S^\mu | \boldsymbol{B}) \propto \exp\left[-\frac{\gamma}{2} \sum_j (\xi_j^\mu)^2 \right].$$

The data from the second process may be understood as representing outliers which do not contain any information about the two spatially structured classes of inputs and come from a 'garbage' class and are classified purely by random guessing. In order to distinguish the two processes, we introduce decision variables $V^\mu \in \{0, 1\}$, where $V^\mu = 1$ stands for the first process and $V^\mu = 0$ for the outliers. The joint set of decision variables is denoted by $\{V^\mu\}_\mu$. Putting conditions on these variables, we can write the probability distribution for the joint set of αN data $\mathbb{D} := \{\boldsymbol{\xi}^\mu, S^\mu\}_\mu$, $\mu = 1, \ldots, p = \alpha N$ within the single equation

$$\mathcal{P}(\mathbb{D} | \{V^\mu\}_\mu, \boldsymbol{B}) = \frac{1}{2^{\alpha N}} \left(\frac{\gamma}{2\pi} \right)^{\alpha N^2/2} \prod_{\mu, j} \exp\left[-\frac{\gamma}{2} (\xi_j^\mu)^2 + \frac{\gamma}{\sqrt{N}} V^\mu \xi_j^\mu S^\mu B_j - \frac{\gamma}{2N} V^\mu {B_j}^2 \right]. \tag{1}$$

In order to model the fact that outliers occur at random with a fixed rate, we will assume that both processes (structure, outliers) are chosen independently at random. The probability of having the value V^μ is written as

$$\mathcal{P}(V^\mu) = \frac{\exp[-\eta V^\mu]}{1 + \exp[-\eta]}. \tag{2}$$

Using the 'chemical potential' η, we can adjust the average fraction of structured data

$$\overline{V^\mu} = \frac{1}{\exp[\eta] + 1}.$$

For $\eta = -\infty$ all examples have $V^\mu = 1$, but with increasing η, fewer examples carry information. For $\eta = 0$, only half of the examples come from the structure and for $\eta = \infty$ all examples are outliers.

A learner tries to infer the vector $\boldsymbol{B}$ from the αN examples and makes an estimate $\boldsymbol{J}$ for $\boldsymbol{B}$. We will assume that the fraction of outliers is known to the learner. Although in our final results we mostly deal with the case that also the parameter γ is known precisely, we shall be more general in the basic definitions and assume that the learner uses $\tilde{\gamma}$ instead, with $\gamma \neq \tilde{\gamma}$. Hence, if the $\{V^\mu\}_\mu$ were known, the likelihood of the data based on the estimate $\boldsymbol{J}$ would be given by

$$\mathcal{P}(\mathbb{D}|\{V^\mu\}_\mu, \boldsymbol{J}) = \frac{1}{2^{\alpha N}} \left(\frac{\tilde{\gamma}}{2\pi}\right)^{\alpha N^2/2} \prod_{\mu,j} \exp\left[-\frac{\tilde{\gamma}}{2}(\xi_j^\mu)^2 + \frac{\tilde{\gamma}}{\sqrt{N}} V^\mu \xi_j^\mu S^\mu J_j - \frac{\tilde{\gamma}}{2N} V^\mu {J_j}^2\right].$$

In general, however, the learner does not know which of the examples contain information and which are outliers. Hence, to the learner the $\{V^\mu\}_\mu$ are *hidden variables* which are not observed but need to be averaged over. Hence, the actual ansatz for the distribution of data will be given by the *mixture distribution*

$$\mathcal{P}(\mathbb{D}|\boldsymbol{J}) = \sum_{\{V^\mu\}_\mu} \mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu|\boldsymbol{J}) \tag{3}$$

where

$$\begin{aligned}\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu|\boldsymbol{J}) &= \mathcal{P}(\mathbb{D}|\{V^\mu\}_\mu, \boldsymbol{J})\mathcal{P}(\{V^\mu\}_\mu) \\ &= \frac{1}{2^{\alpha N}} \left(\frac{\tilde{\gamma}}{2\pi}\right)^{\alpha N^2/2} \frac{1}{(1+\exp[-\eta])^{\alpha N}} \exp\left[-\frac{\tilde{\gamma}}{2}\sum_{\mu,j}(\xi_j^\mu)^2 - \sum_\mu V^\mu f_\mu(\boldsymbol{J})\right]\end{aligned} \tag{4}$$

and where we have defined

$$f_\mu(\boldsymbol{J}) := -\frac{\tilde{\gamma}}{\sqrt{N}} \sum_j \xi_j^\mu S^\mu J_j + \frac{\tilde{\gamma}}{2N} \sum_j {J_j}^2 + \eta.$$

One possible way of getting an estimate for the unknown vector $\boldsymbol{B}$, would be the *maximum likelihood* method, i.e. one would use the vector $\boldsymbol{J}$ which maximizes the likelihood (3). A second possibility is given by a Bayesian approach, where the learner supplies some *prior knowledge* about reasonable estimates $\boldsymbol{J}$ within a *prior distribution*. We will use a distribution which on average gives the correct length of the unknown vector but does not favour any spatial direction

$$\mathcal{P}(\boldsymbol{J}) = \left(\frac{1}{2\pi}\right)^{N/2} \exp\left[-\frac{1}{2}\sum_j {J_j}^2\right]. \tag{5}$$

Based on the prior and the likelihood of the data, the learner can construct the posterior distribution, using Bayes rule

$$\mathcal{P}(\boldsymbol{J}|\mathbb{D}) = \frac{\mathcal{P}(\mathbb{D}|\boldsymbol{J})\mathcal{P}(\boldsymbol{J})}{\mathcal{P}(\mathbb{D})}. \tag{6}$$

There are several ways of using the information contained in the posterior (6). For example, simply taking the *posterior mean* as the estimate for $\boldsymbol{B}$ will minimize the expected average (with respect to the posterior) squared error. Unfortunately, for a high-dimensional space, such expectations will not be easy to calculate exactly, and one has to resort to Monte Carlo sampling. A simpler estimate, which should not perform too poorly, is given by the vector

$\boldsymbol{J}$, which has maximal *a posteriori* probability (MAP), i.e. the one which maximizes (6). Actually, if there are enough data available, one can expect that the posterior will be close to a Gaussian, and both estimates will come close.

In order to maximize the posterior $\mathcal{P}(\boldsymbol{J}|\mathbb{D})$ with respect to $\boldsymbol{J}$, we can equivalently minimize the 'training' energy function

$$\mathcal{H}(\boldsymbol{J}) = -\ln \mathcal{P}(\mathbb{D}, \boldsymbol{J}) = -\ln \sum_{\{V^\mu\}_\mu} \mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu|\boldsymbol{J})\mathcal{P}(\boldsymbol{J}). \tag{7}$$

As we shall see in section 4, there is a simple algorithm to calculate the MAP, based on a recursive estimation of the (posterior) expected decision variables $\{V^\mu\}_\mu$. Since examples are weighted by their probability of being informative rather than being kept or discarded from the training set, we call this method a *soft selection* of examples.

As an alternative to the MAP approach for $\boldsymbol{J}$, we shall also discuss an algorithm which calculates the MAP for the hidden variables $\{V^\mu\}_\mu$. Since these variables take the values 0 and 1 only, the result will be a *hard* selection of informative examples, rather than a soft weighting. We look for the values of $\{V^\mu\}_\mu$ which maximize

$$\mathcal{P}(\{V^\mu\}_\mu|\mathbb{D}) = \frac{\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu)}{\mathcal{P}(\mathbb{D})}. \tag{8}$$

Equivalently, we can maximize the numerator of this expression, which can be written as a mixture probability

$$\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu) = \int \mathrm{d}\boldsymbol{J}\, \mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu, \boldsymbol{J}) \tag{9}$$

resulting in a training energy

$$\mathcal{H}_h(\{V^\mu\}_\mu) = -\ln \int \mathrm{d}\boldsymbol{J}\, \mathcal{P}(\mathbb{D}, \boldsymbol{J}, \{V^\mu\}_\mu). \tag{10}$$

Finally, after minimization, we can use the expectations

$$\langle J_j \rangle_{\boldsymbol{J}} = \frac{\int \mathrm{d}\boldsymbol{J}\, J_j \mathcal{P}(\mathbb{D}, \boldsymbol{J}, \{V^\mu\}_\mu)}{\int \mathrm{d}\boldsymbol{J}\, \mathcal{P}(\mathbb{D}, \boldsymbol{J}, \{V^\mu\}_\mu)} \tag{11}$$

as an estimate for the unknown B_j.

3. Analysis by statistical mechanics

In this section, we study the performance of both MAP estimates analytically in the thermodynamic limit $N \to \infty$ using a statistical mechanics framework. We begin first with the soft selection. There are different ways of measuring how well the learner, equipped with the MAP estimate, has learnt the structured distribution. An obvious idea is to measure the quadratic deviation between the true vector $\boldsymbol{B}$ and the MAP:

$$\Delta = \frac{1}{N}\langle (\boldsymbol{J} - \boldsymbol{B})^2 \rangle = Q - 2R + 1 \tag{12}$$

where we have defined the order parameters

$$\begin{aligned} R &= \frac{1}{N}\langle \boldsymbol{J} \cdot \boldsymbol{B} \rangle \\ Q &= \frac{1}{N}\langle \boldsymbol{J} \rangle^2. \end{aligned} \tag{13}$$

It is also useful to calculate the angle $\Phi = \angle(\boldsymbol{J}, \boldsymbol{B})$ between estimate and $\boldsymbol{B}$. This angle Φ, normalized by $1/\pi$, is given in terms of the order parameters by

$$\Phi = \frac{1}{\pi} \arccos \frac{\boldsymbol{J} \cdot \boldsymbol{B}}{\|\boldsymbol{J}\| \|\boldsymbol{B}\|} \tag{14}$$

$$= \frac{1}{\pi} \arccos \frac{R}{\sqrt{Q}}. \tag{15}$$

The order parameters for the soft selection MAP algorithm can be derived from a partition function Z where the corresponding Hamiltonian is given by $\mathcal{H}(\boldsymbol{J})$ from (7). Assuming that the inverse temperature β is an integer, we define

$$\begin{aligned} Z &= \int \mathrm{d}\boldsymbol{J} \ \exp[-\beta \mathcal{H}(\boldsymbol{J})] \\ &= \int \mathrm{d}\boldsymbol{J} \ \exp[\beta \ln \mathcal{P}(\mathbb{D}, \boldsymbol{J})] \\ &= \int \mathrm{d}\boldsymbol{J} \ (\mathcal{P}(\mathbb{D}, \boldsymbol{J}))^{\beta} \\ &= \int \mathrm{d}\boldsymbol{J} \left\{ \sum_{\{V^{\mu}\}_{\mu}} \mathcal{P}(\mathbb{D}, \{V^{\mu}\}_{\mu}, \boldsymbol{J}) \right\}^{\beta} \\ &= \int \mathrm{d}\boldsymbol{J} \sum_{\{V_b^{\mu}\}_{\mu}} \prod_{b=1}^{\beta} \mathcal{P}(\mathbb{D}, \{V_b^{\mu}\}_{\mu}, \boldsymbol{J}). \end{aligned} \tag{16}$$

The MAP, which is the minimum of the energy $\mathcal{H}(\boldsymbol{J})$, is derived from the limit $\beta \to \infty$. The case $\beta = 1$ would correspond to Gibbs learning, where a vector $\boldsymbol{J}$ is drawn at random from the posterior. As usual, order parameters are found from an average of the free energy $f = -\frac{1}{\beta N} \ln Z$ over the distribution of the examples. To perform the average, we utilize the replica trick

$$\begin{aligned} \langle f \rangle &= -\frac{1}{\beta N} \langle \ln Z \rangle \\ &= -\frac{1}{\beta N} \lim_{n \to 0} \frac{\partial}{\partial n} \ln \langle Z^n \rangle \end{aligned} \tag{17}$$

where $\langle \ldots \rangle$ denotes the average over the distribution (see (1) and (2))

$$\mathcal{P}(\xi_j^{\mu}, S^{\mu} | \boldsymbol{B}) = \frac{1}{2} \left(\frac{\gamma}{2\pi} \right)^{1/2} \frac{1}{1 + \mathrm{e}^{-\eta}} \sum_{V^{\mu}} \exp \left[-\frac{\gamma}{2} \left(\xi_j^{\mu} - \frac{1}{\sqrt{N}} V^{\mu} S^{\mu} B_j \right)^2 - \eta V^{\mu} \right].$$

The replicated partition function is now written as

$$Z^n = \sum_{\{V_{ab}^{\mu}\}_{\mu}} \int \prod_a \mathrm{d}\boldsymbol{J}^a \prod_{a,b} \mathcal{P}(\mathbb{D}, \{V_{ab}^{\mu}\}_{\mu}, \boldsymbol{J}^a) \tag{18}$$

where the decision variables contain *two replica* indices. Here, the index a runs from 1 to n, whereas b runs from 1 to β. For the subsequent calculations we have assumed the correct parameters $\gamma = \tilde{\gamma}$ and have made a *replica symmetric ansatz* with respect to the indices a. We think that this should be at least a good approximation, because our model is an example of a *teacher–student* learning scenario, where student and teacher match in the sense that the student uses the right statistical model for the data. For the Gibbs learning scenario ($\beta = 1$), where the symmetry of student and teacher becomes perfect in the replica calculation (this can be seen by introducing a further average over $\boldsymbol{B}$, using the prior (5)),

replica symmetry is usually considered to be exact (although no general proof has been given so far). Hence, assuming that the effects of replica symmetry-breaking are small, we have refrained from performing a replica stability analysis.

The treatment of the replica indices b is much simpler, because the order parameters (see appendix A) do not depend on them. Hence, as long as β is an integer, no further symmetry assumptions are required for the b's. Although we do not have a proof that the continuation to noninteger β is unique, we expect that the limit $\beta \to \infty$ exists and can be safely calculated using a sequence of integers.

The *hard selection problem* of decision variables is treated similarly using the (zero temperature) free energy which is defined from the partition function

$$Z_h = \sum_{\{V^\mu\}_\mu} \mathrm{e}^{-\beta \mathcal{H}_h(\{V^\mu\}_\mu)} \tag{19}$$

with the energy (10). The averages which are necessary for the calculation of error measures, e.g.

$$\Phi = \frac{1}{\pi} \arccos \frac{\sum_j \langle J_j \rangle_{\boldsymbol{J}} B_j}{\sqrt{\sum_j \langle J_j{}^2 \rangle_{\boldsymbol{J}}} \sqrt{N}} \tag{20}$$

can be found in a standard way from derivatives of the free energy with respect to appropriate external fields, e.g.

$$\sum_j \langle J_j \rangle_{\boldsymbol{J}} B_j = -\lim_{\lambda \to 0} \frac{\partial}{\partial \lambda} \lim_{\beta \to \infty} \frac{1}{\beta} \ln \sum_{\{V^\mu\}_\mu} \mathrm{e}^{-\beta \mathcal{H}_h(\{V^\mu\}_\mu, \lambda)} \tag{21}$$

where

$$\mathcal{H}_h(\{V^\mu\}_\mu, \lambda) = -\ln \int \mathrm{d}\boldsymbol{J}\, \mathcal{P}(\mathbb{D}, \boldsymbol{J}, \{V^\mu\}_\mu) \exp\left[-\lambda \sum_j J_j B_j \right].$$

Explicit calculations of the free energies and order parameters for both cases are given in the appendices.

4. The EM algorithm

Unfortunately, the maximization of the posterior distributions cannot be carried out in closed form and must be done numerically. Usually, nonlinear optimization problems are solved by gradient descent algorithms which require a tuning of the step sizes. However, for the type of (generalized) maximum likelihood problem for mixture distributions such as (3) and (9), there is a simpler and well known algorithm which has been developed by Dempster *et al* [4]. The EM algorithm guarantees that the (generalized) likelihood is nondecreasing for every iteration step and converges to a local maximum. To explain the idea for the soft selection problem, let us assume for the moment that the hidden variables $\{V^\mu\}_\mu$ are actually known. Then the corresponding log-likelihood $\ln[\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu | \boldsymbol{J}) \mathcal{P}(\boldsymbol{J})]$ can be maximized in closed form. In the EM algorithm, the true values of the hidden variables are replaced iteratively by suitable averages. At iteration i, in the *expectation step*, the function

$$A(\boldsymbol{J}, \boldsymbol{J}^{(i)}) := \langle \ln[\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu | \boldsymbol{J}) \mathcal{P}(\boldsymbol{J})] \rangle_{\mathcal{P}(\{V^\mu\}_\mu | \mathbb{D}, \boldsymbol{J}^{(i)})} \tag{22}$$

is calculated, which is the log-likelihood of observed and hidden data averaged over the posterior distribution of the hidden data, given the old estimate $\boldsymbol{J}^{(i)}$. In the *maximization step*, (22) is maximized with respect to $\boldsymbol{J}$ in order to obtain the new iteration $\boldsymbol{J}^{(i+1)}$.

We will not give the proof of convergence here, as it is relatively simple and can be found in many textbooks (see e.g. [6]). However, we can easily see that a fixed point of the algorithm is also a local extremum of (7). At the maximum of (22), we have

$$0 = \frac{\partial}{\partial J_k} A(\boldsymbol{J}, \boldsymbol{J}^{(i)}) = \frac{\partial}{\partial J_k} \langle \ln[\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu | \boldsymbol{J}) \mathcal{P}(\boldsymbol{J})] \rangle_{\mathcal{P}(\{V^\mu\}_\mu | \mathbb{D}, \boldsymbol{J}^{(i)})}$$
$$= \sum_{\{V^\mu\}_\mu} \frac{\frac{\partial}{\partial J_k} \mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu, \boldsymbol{J}) \mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu, \boldsymbol{J}^{(i)})}{\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu, \boldsymbol{J}) \mathcal{P}(\mathbb{D}, \boldsymbol{J}^{(i)})}.$$

Hence, at the fixed point, where $\boldsymbol{J}^{(i)} = \boldsymbol{J}$, we also have $\frac{\partial}{\partial J_k} \ln \mathcal{P}(\mathbb{D}, \boldsymbol{J}) = 0$. For the explicit calculation, we need the conditional distribution of the hidden variables, given the data and $\boldsymbol{J}$

$$\mathcal{P}(\{V^\mu\}_\mu | \mathbb{D}, \boldsymbol{J}) = \frac{\mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu, \boldsymbol{J})}{\mathcal{P}(\mathbb{D}, \boldsymbol{J})}$$
$$= \prod_\mu \frac{\exp[-V^\mu f_\mu(\boldsymbol{J})]}{1 + \exp[-f_\mu(\boldsymbol{J})]}. \tag{23}$$

Using the distribution (2), we get

$$\frac{\partial}{\partial J_k} A(\boldsymbol{J}, \boldsymbol{J}^{(i)}) = -\tilde{\gamma} \sum_\mu \langle V^\mu \rangle \left(-\frac{1}{\sqrt{N}} \xi_k^\mu S^\mu + \frac{1}{N} J_k \right) - J_k$$
$$\stackrel{!}{=} 0$$

which gives

$$\boldsymbol{J} = \frac{\sqrt{N} \sum_\mu \langle V^\mu \rangle \boldsymbol{\xi}^\mu S^\mu}{\sum_\mu \langle V^\mu \rangle + N/\tilde{\gamma}} \tag{24}$$

where

$$\langle V^\mu \rangle = \sum_{V^\mu = 0,1} V^\mu \mathcal{P}(V^\mu | \mathbb{D}, \boldsymbol{J}^{(i)})$$
$$= \frac{1}{\exp[f_\mu(\boldsymbol{J}^{(i)})] + 1}. \tag{25}$$

Hence, the estimate $\boldsymbol{J}$ for $\boldsymbol{B}$ is of the form of a *weighted Hebbian* sum, where each example has a weight which is proportional to the estimated probability $\langle V^\mu \rangle$, that the example is not an outlier. It is interesting to look at the limiting case $\eta \to -\infty$, i.e. where all examples are from the double cluster and where no outliers are present. In this case the EM iteration stops after one step and we get

$$\langle V^\mu \rangle = 1 \qquad \text{for all } \mu$$
$$\boldsymbol{J} = \frac{1}{\sqrt{N}} \frac{\sum_\mu \boldsymbol{\xi}^\mu S^\mu}{\alpha + 1/\tilde{\gamma}} \tag{26}$$

which is the usual Hebbian vector.

Similarly, to apply the EM algorithm to the hard selection problem with the mixture distribution (9), we take $\boldsymbol{J}$ as the hidden quantity. In each iteration step, we have to maximize

$$\hat{A}(\{V^\mu\}_\mu, \{V^\mu\}_\mu^{(i)}) := \langle \ln \mathcal{P}(\mathbb{D}, \{V^\mu\}_\mu, \boldsymbol{J}) \rangle_{\mathcal{P}(\boldsymbol{J} | \mathbb{D}, \{V^\mu\}_\mu^{(i)})}$$
$$= -\frac{\tilde{\gamma}}{2} \sum_{\mu,j} (\xi_j^\mu)^2 + \frac{\tilde{\gamma}}{\sqrt{N}} \sum_{\mu,j} V^\mu \xi_j^\mu S^\mu \langle J_j \rangle - \frac{\tilde{\gamma}}{2N} \sum_{\mu,j} V^\mu \langle J_j{}^2 \rangle$$

$$-\eta \sum_{\mu} V^{\mu} - \tfrac{1}{2} \sum_{j} \langle {J_j}^2 \rangle \tag{27}$$

with respect to $\{V^{\mu}\}_{\mu}$. Defining

$$\begin{aligned} a &:= \frac{\tilde{\gamma}}{N} \sum_{\mu} V^{\mu(i)} + 1 \\ b_j &:= \frac{\tilde{\gamma}}{\sqrt{N}} \sum_{\mu} V^{\mu(i)} \xi_j^{\mu} S^{\mu} \end{aligned} \tag{28}$$

we obtain for the expectations at step i

$$\begin{aligned} \langle J_j \rangle &= \frac{b_j}{a} \\ &= \frac{\sqrt{N} \sum_{\mu} V^{\mu(i)} \xi_j^{\mu} S^{\mu}}{\sum_{\mu} V^{\mu(i)} + N/\tilde{\gamma}} \\ \langle {J_j}^2 \rangle &= \frac{{b_j}^2}{a^2} + \frac{1}{a}. \end{aligned} \tag{29}$$

Finally, after convergence, we use $\langle J_j \rangle$ as an estimate for B_j.

5. Results and discussion

5.1. Soft selection

Solving for the order parameters and assuming that $\tilde{\gamma} = \gamma$ we find that for fixed η, as expected, both error measures Φ and Δ decrease towards 0 with an increasing number αN of examples, showing that the algorithm is able to find the true structure vector $\boldsymbol{B}$. Since for the EM algorithm both error measures show qualitatively the same behaviour; we shall concentrate mainly on the angle Φ.

Figure 1 shows $\Delta(\alpha)$ for $\eta = 0$. The second curve gives the performance of the Hebbian rule (26). It demonstrates the importance of selecting informative examples. If all examples are weighted equally (and $\eta \neq \infty$), then the true vector $\boldsymbol{B}$ cannot be recovered for $\alpha \to \infty$. In figure 2, $\Phi(\alpha)$ (EM algorithm) is shown for $\eta = 0$ and $\eta = 4$. Since it was harder to perform simulations for $\eta = 4$, where only about 1.8% of the examples are informative, we have shown simulations only for $\eta = 0$. Asymptotically one finds a decrease of the error along the lines of

$$\Phi \overset{\alpha \to \infty}{\simeq} \frac{1}{\pi R_{\infty}} \sqrt{\frac{c}{\alpha}} \tag{30}$$

where R_{∞} is the asymptotic value of the order parameter R and both R_{∞} and c depend on η.

As expected, for fixed α, the error increases with η, i.e. with a growing number of outliers. More interesting is the nonsmooth behaviour of the second curve, which gives a sudden drop of the error as η is varied. This phase transition can be observed in more detail in the relief plot of the order parameters R and Q in figures 3(a) and (b). In regions of large η or large α, the saddle-point equations have three solutions. Taking the solution with the smallest free energy leads to a jump of the order parameters. It is easier to investigate the transition by simulations as a function of η, for fixed α. This is shown in figure 4, together with the predictions of the theory.

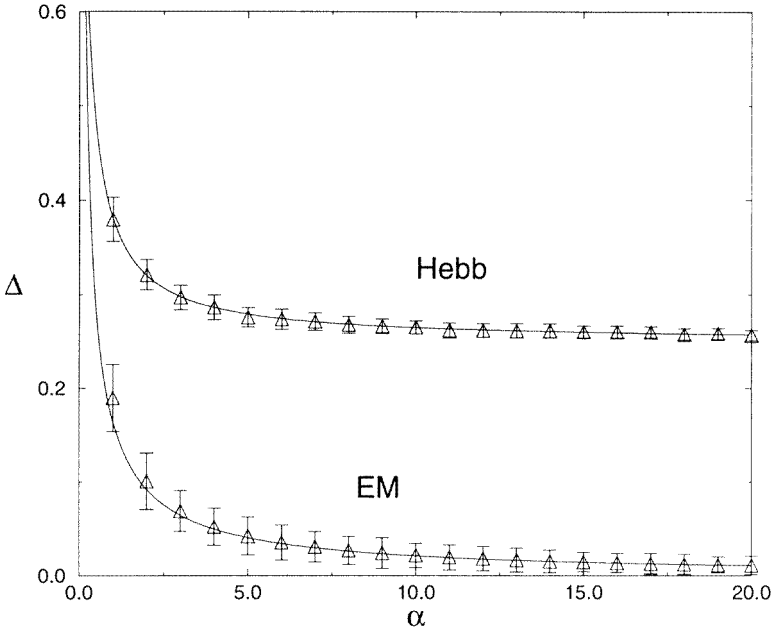

Figure 1. Comparison between the EM algorithm and naive Hebbian rule. Parameters are $\eta = 0$, $\gamma = \tilde{\gamma} = 10$. The full curves show the theoretical results. Simulations were done with $N = 500$. Here and in following figures, bars mark standard deviations over 100 runs.

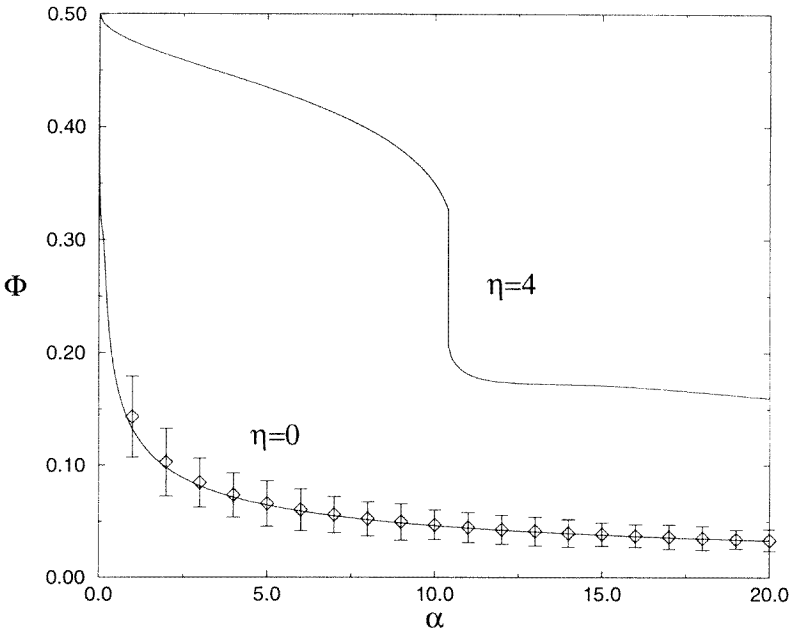

Figure 2. $\Phi(\alpha)$ for $\eta = 0$ and $\eta = 4$, respectively (MAP estimate). The simulations at $\eta = 0$ were performed with $N = 500$; results were averaged over 100 runs.

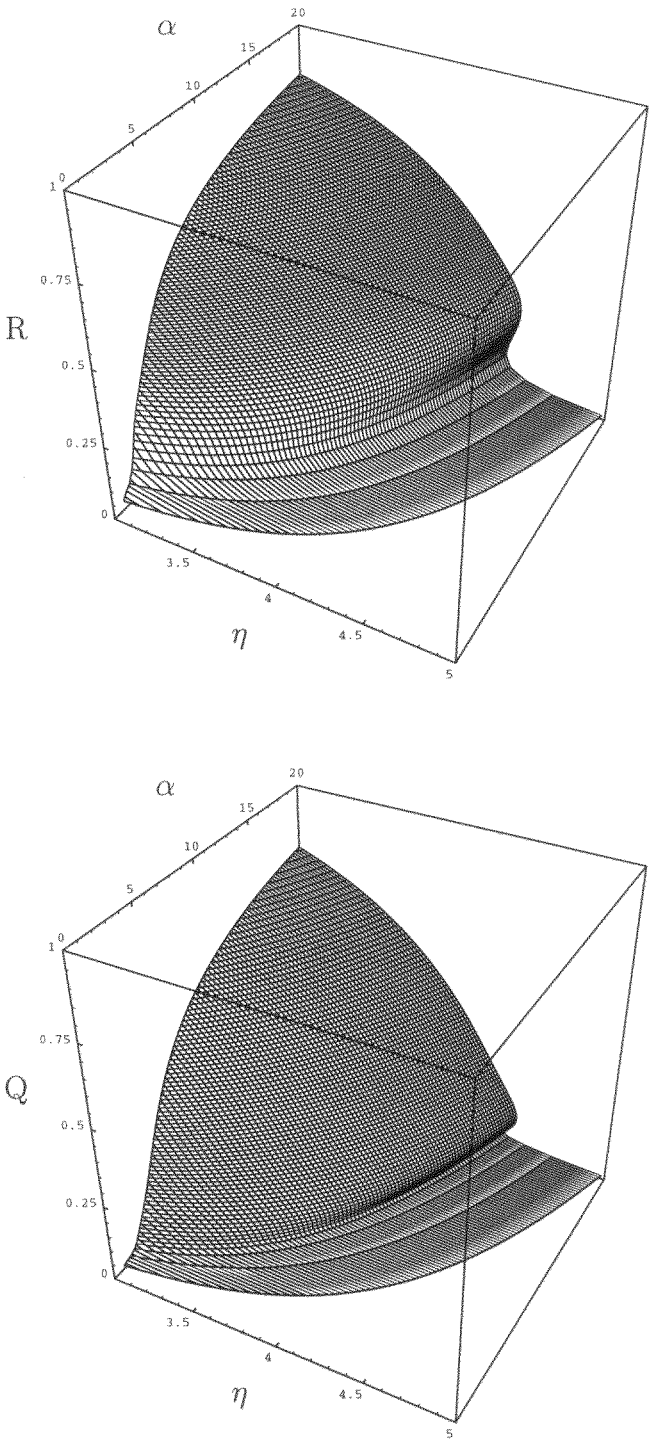

Figure 3. Order parameters $R(\alpha, \eta)$ (top) and $Q(\alpha, \eta)$ (bottom) for MAP. As in figure 2 and subsequently, we set $\gamma = \tilde{\gamma} = 10$.

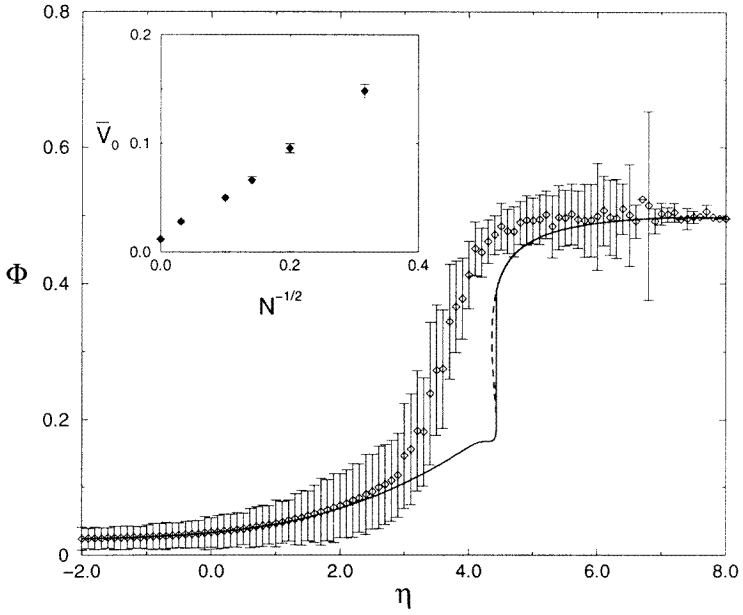

Figure 4. Error Φ for soft selection versus amount of outliers, represented by η. The relative number of data is fixed at $\alpha = 20$. The broken length of the theoretical curve denotes the region where three solutions of the saddle-point equations exist. The full curve follows the solution with minimal free energy. Simulations were results from 100 runs with $N = 100$. Note that, for finite N, the transitions of the two order parameters do not coincide. The error measure Φ roughly follows the overlap R between solution vector and structure axis, whereas the drop in Q gives rise to the increased standard deviation at $\eta = 6.8$. The inset shows a finite-size scaling of the phase transition as described in the text. The corresponding dimensions of the data are $N = 10, 25, 50, 100, 1000$ respectively.

We have simulated the EM algorithm starting from random initial conditions and averaged the order parameters over many samples of random inputs. Fixing α, the simulations show a good agreement with the theory for small and large values of η, but discrepancies show up close to the predicted transition. Since the average fraction $\bar{V}$ of informative data points decreases exponentially with η, finite-size effects play a crucial role in the simulations. For example, for $\eta = 4$, less than two examples out of $N = 100$ are informative on average whereas the replica theory is based on infinitely many examples from the structured clusters. Hence, we have performed a finite-size scaling to determine the critical value $\bar{V}_0$, where the transition sets in. Since for small η (large $\bar{V}$), the simulations show rather small statistical fluctuations around a value of R close to 1, we have (for each N) defined $\bar{V}_0$ as the point, where the distribution of the observed values for R significantly broadens, indicating the onset of transitions to different values of R. A simple linear extrapolation to $N = \infty$ as shown in the inset of figure 4 gives a value for $\bar{V}_0$ which is in good agreement with the predicted value for the phase transition. The large error bar at $\eta = 6.8$ is explained from the fact that the values for Φ (equation (14)) have been obtained by using the sample averages of R and Q which (for finite N) show a transition at slightly different values of η.

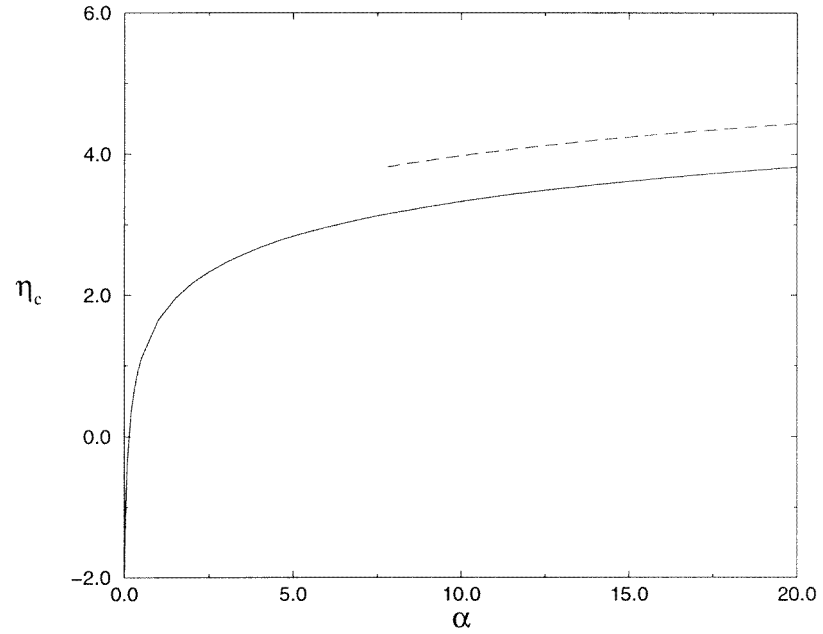

Figure 5. Critical fraction of outliers for hard selection as a function of α (full curve). The broken curve represents the phase transition for soft selection.

5.2. Hard selection

Solving the order-parameter equations for the free energy (B1) at zero temperature, we find similar first-order transitions as for the method of soft selection. For η small enough, there is only one solution which has a nonzero overlap to the teacher vector $\boldsymbol{B}$. Increasing η (and thereby the number of outliers) beyond a value η_0, another solution with $\hat{R} = \hat{Q} = \hat{z} = 0$ (see equation (B1)) appears, i.e. where all $V^\mu = 0$ and all data are considered to be outliers. Here

$$\eta_0 = -\frac{\tilde{\gamma}}{2} + \frac{\tilde{\gamma}^2}{4\gamma}. \tag{31}$$

Between η_0 and a second parameter value η_c, however, this trivial solution has a higher free energy $f_h = 0$ than the nontrivial one. Finally, for $\eta > \eta_c$, the trivial solution with zero order parameters, giving rise to $\Phi = \frac{1}{2}$, is the one with lowest free energy. Figure 5 shows this critical η as a function of α.

So, unlike in the soft selection case, we have, for a large range of η, two solutions of the order-parameter equations. This is reflected in the simulations, the single runs clearly tending to either of these two optima. Effects of metastability (which would be a sign of a rugged energy landscape and indicate strong effects of replica symmetry-breaking) could not be observed. However, a finite-size scaling for the transition point did not lead to a satisfactory agreement with the theory. We think that the observed discrepancy is a dynamical effect, where the EM algorithm, starting from a random initial condition, is unable to reach the global minimum and converges only to the local one, thus shifting the phase transition to smaller values of η. We have balanced this effect to some extent by keeping only those simulations (as long as they occur) where the EM algorithm converges to the solution with nonzero overlap to the vector $\boldsymbol{B}$.

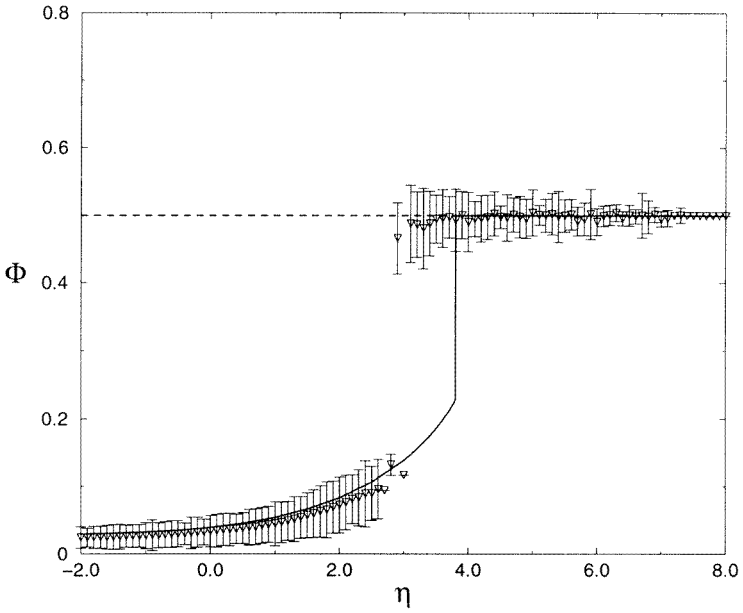

Figure 6. Error Φ for hard selection versus amount of outliers, represented by η. As in figure 4, $\alpha = 20$ and simulations were performed with $N = 100$ and 100 runs. The full curve indicates the theoretical result for the global optimum, the broken line for the local one.

Figure 6 shows the performance of the hard selection for $\alpha = 20$. Comparision to figure 4 suggests that the soft selection should be preferred. The difference between the performance of the two algorithms becomes more drastic for $\alpha \to \infty$: the soft selection algorithm is able to tolerate an *arbitrary fraction* of outliers as long as enough data are available. Eventually, it will always find the true teacher vector $\boldsymbol{B}$. On the other hand, for hard selection, the explicit solution of the order-parameter equations for $\alpha \to \infty$ shows that there is always a critical fraction of outliers (corresponding to a parameter η_c (B3)) where learning is no longer possible even if inifinitely many examples are available. It is also interesting to investigate the influence of the overlap of the two Gaussian clouds in the structured input distribution on the transition parameter η_c. Figure 7 shows η_c for $\alpha = \infty$ as a function of γ, which gives the inverse squared width of each Gaussian and so measures the distinguishability of the clouds. Somewhat surprisingly, if γ is below 0.278, the critical η jumps discontinuously to zero, i.e. if the overlap of the two clouds is above a certain value, only 50% outliers can be tolerated.

Phase transitions in the performance of learning algorithms have been observed frequently in the statistical mechanics of neural networks. Since such effects do not occur in asymptotic (in the sense of large α) expansions or in the exact bounds known in statistics they seem to be one of the major contributions of statistical mechanics to the field of computational learning theory. Phase transitions occur in multilayer networks, where they are can be related to the breaking of symmetries which are related to the network architecture [12, 11]. Other examples include models with a so-called student–teacher mismatch [13], models with discrete adjustable parameters [7, 8] and models of unsupervised learning [9, 10]. For the present supervised learning model, where the basic adjustable parameters are continuous variables and where the learner matches with the distribution of the data, the phase transition was unexpected. It will be interesting to apply recently developed

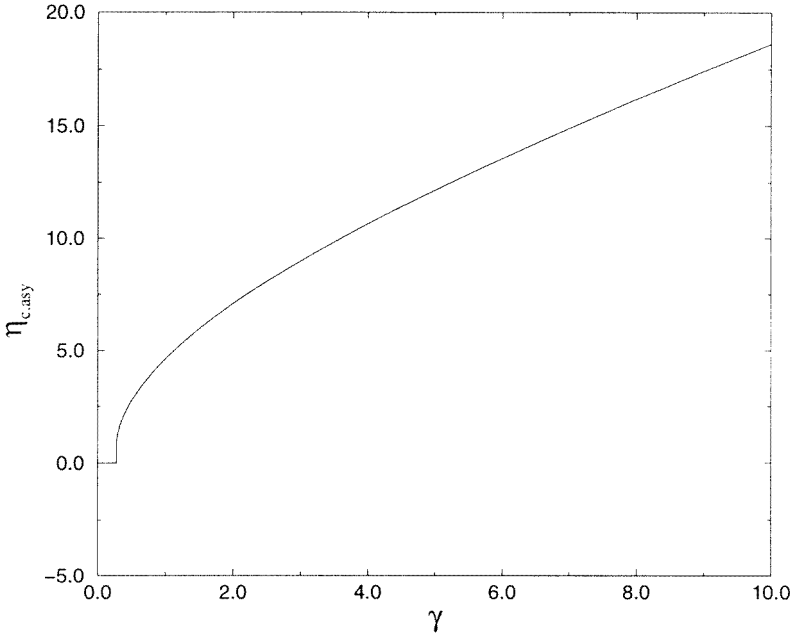

Figure 7. Asymptotic critical fraction of outliers for hard selection, plotted against inverse squared width of the Gaussian clusters.

combinations of statistical mechanics techniques and methods of information theory [14] to establish the existence of phase transitions in mixture models in more general circumstances.

Appendix A. Free energy and order parameters for soft selection

Upon averaging, we obtain

$$\langle Z^n\rangle = \sum_{\{V^\mu_{ab}\}_\mu} \int \prod_{a,j} \mathrm{d}J^a_j \, \exp\left[-\sum_{a,b}\left(\frac{\tilde\gamma}{2N}\sum_\mu V^\mu_{ab} + \frac{1}{2}\right)\sum_j (J^a_j)^2 - \eta \sum_{a,b}\sum_\mu V^\mu_{ab}\right] \times\left\langle \exp\left[-\frac{\tilde\gamma n\beta}{2}\sum_{\mu,j}(\xi^\mu_j)^2 + \frac{\tilde\gamma}{\sqrt{N}}\sum_{a,b}\sum_{\mu,j} V^\mu_{ab}\xi^\mu_j S^\mu J^a_j\right]\right\rangle.$$

Within replica symmetry, the introduction of the order parameters

$$R = \frac{1}{N}\langle \boldsymbol{J}\cdot\boldsymbol{B}\rangle = \frac{1}{N}\sum_j J^a_j B_j$$

$$q = \frac{1}{N}\langle \boldsymbol{J}^2\rangle = \frac{1}{N}\sum_j J^a_j J^{\tilde a}_j$$

$$Q = \frac{1}{N}\langle \boldsymbol{J}\rangle^2 = \frac{1}{N}\sum_j (J^a_j)^2$$

together with their conjugates yields

$$\langle Z^n\rangle \propto \int \prod_{a,j} \mathrm{d}J^a_j \, \exp\left[\mathrm{i}N\Phi\left(\frac{1}{N}\sum_{j,a} J^a_j B_j - nR\right)\right]$$

$$\times \prod \exp\left[\mathrm{i}N\omega\left(\frac{1}{N}\sum_{j,a,\tilde{a}\neq a} J_j^a J_j^{\tilde{a}} - n(n-1)q\right)\right]$$
$$\times \exp\left[\mathrm{i}N\Omega\left(\frac{1}{N}\sum_{j,a}(J_j^a)^2 - nQ\right)\right]$$
$$\times \sum_{\{V_{ab}^\mu\}_\mu}\left(\prod_{a,b}\exp\left[-\left(\frac{\tilde{\gamma}}{2}\sum_\mu V_{ab}^\mu + \frac{1}{2}N\right)Q - \eta\sum_\mu V_{ab}^\mu\right]\right)$$
$$\times\left(\prod_\mu \exp\left[\frac{1}{1+n\beta\tilde{\gamma}/\gamma}\left(-\frac{1}{2}\tilde{\gamma}n\beta(V^\mu)^2 + \tilde{\gamma}\sum_{a,b} V_{ab}^\mu V^\mu R\right.\right.\right.$$
$$\left.\left.\left.+\frac{\tilde{\gamma}^2}{2\gamma}\sum_{a,\tilde{a}\neq a}\sum_{b,\tilde{b}} V_{ab}^\mu V_{\tilde{a}\tilde{b}}^\mu q + \frac{\tilde{\gamma}^2}{2\gamma}\sum_a\sum_{b,\tilde{b}} V_{ab}^\mu V_{a\tilde{b}}^\mu Q\right) - \eta V^\mu\right]\right).$$

In this expression (and in the following one) the order parameters have to be taken at their saddle-point values. After a lengthy calculation, we arrive at an expression for the free energy

$$f = \frac{1}{\beta}\frac{R^2-Q}{2(Q-q)} - \frac{1}{2\beta}\ln(Q-q) + \frac{1}{2}Q - \frac{\alpha}{\beta}M(R,q,Q) + \text{constant} \tag{A1}$$

with

$$M(R,q,Q) = \frac{1}{1+\mathrm{e}^{-\eta}}\int \mathrm{D}x\left\{\ln\left(\int \mathrm{D}y\left(1+\exp\left[-\frac{\tilde{\gamma}}{2}Q - \eta + \tilde{\gamma}\sqrt{\frac{q}{\gamma}}x\right.\right.\right.\right.$$
$$\left.\left.\left.+\tilde{\gamma}\sqrt{\frac{Q-q}{\gamma}}y\right]\right)^\beta\right) - \frac{1}{2}\mathrm{e}^{-\eta}\tilde{\gamma}\rho^2\beta$$
$$+\mathrm{e}^{-\eta}\ln\left(\int \mathrm{D}y\left(1+\exp\left[-\frac{\tilde{\gamma}}{2}Q - \eta + \tilde{\gamma}R\right.\right.\right.$$
$$\left.\left.\left.+\tilde{\gamma}\sqrt{\frac{q}{\gamma}}x + \tilde{\gamma}\sqrt{\frac{Q-q}{\gamma}}y\right]\right)^\beta\right)\right\}.$$

For $\beta\to\infty$ we have to take the limit $q\to Q$. With the ansatz $(Q-q)\beta =: z = \mathcal{O}(1)$, we get in the limit

$$f = \frac{R^2-Q}{2z} + \frac{1}{2}Q - \frac{\alpha}{1+\mathrm{e}^{-\eta}}\left(\hat{I}_5 + \frac{b}{2}\hat{I}_1\right) - \frac{\alpha\mathrm{e}^{-\eta}}{1+\mathrm{e}^{-\eta}}\left(I_5 + \frac{b}{2}I_1\right) + \text{constant.} \tag{A2}$$

This yields the saddle-point equations

$$0 \stackrel{!}{=} \frac{\partial f}{\partial R} = \frac{R}{z} - \frac{\alpha\mathrm{e}^{-\eta}}{1+\mathrm{e}^{-\eta}}\left(I_6 + \frac{b}{2}I_2\right)\tilde{\gamma}$$
$$0 \stackrel{!}{=} \frac{\partial f}{\partial z} = \frac{Q-R^2}{2z^2} - \frac{\alpha}{1+\mathrm{e}^{-\eta}}\hat{I}_4\frac{\tilde{\gamma}^2}{2\gamma} - \frac{\alpha\mathrm{e}^{-\eta}}{1+\mathrm{e}^{-\eta}}I_4\frac{\tilde{\gamma}^2}{2\gamma}$$
$$0 \stackrel{!}{=} \frac{\partial f}{\partial Q} = -\frac{1}{2z} + \frac{1}{2} - \frac{\alpha}{1+\mathrm{e}^{-\eta}}\left(\frac{\tilde{\gamma}}{2}\left(\hat{I}_6 + \frac{b}{2}\hat{I}_2\right) + \frac{\tilde{\gamma}}{2\sqrt{\gamma Q}}\left(\hat{I}_7 + \frac{b}{2}\hat{I}_3\right)\right)$$
$$-\frac{\alpha\mathrm{e}^{-\eta}}{1+\mathrm{e}^{-\eta}}\left(\frac{\tilde{\gamma}}{2}\left(I_6 + \frac{b}{2}I_2\right) + \frac{\tilde{\gamma}}{2\sqrt{\gamma Q}}\left(I_7 + \frac{b}{2}I_3\right)\right)$$

where

$$I_1 := \int \mathrm{D}x\,\frac{1}{\mathrm{e}^{-2a} + 1 + (2-b)\mathrm{e}^{-a}}$$

$$I_2 := \int \mathrm{D}x\ \frac{2\mathrm{e}^{-2a} + (2-b)\mathrm{e}^{-a}}{(\mathrm{e}^{-2a} + 1 + (2-b)\mathrm{e}^{-a})^2}$$

$$I_3 := \int \mathrm{D}x\ \frac{2\mathrm{e}^{-2a} + (2-b)\mathrm{e}^{-a}}{(\mathrm{e}^{-2a} + 1 + (2-b)\mathrm{e}^{-a})^2} x$$

$$I_4 := \int \mathrm{D}x\ \frac{\mathrm{e}^{-2a} + 1 + 2\mathrm{e}^{-a}}{(\mathrm{e}^{-2a} + 1 + (2-b)\mathrm{e}^{-a})^2}$$

$$I_5 := \int \mathrm{D}x\ \ln(1 + \mathrm{e}^{a})$$

$$I_6 := \int \mathrm{D}x\ \frac{1}{\mathrm{e}^{-a} + 1}$$

$$I_7 := \int \mathrm{D}x\ \frac{x}{\mathrm{e}^{-a} + 1}.$$

For the $\hat{I}_j$, a has to be replaced by $\hat{a}$, where

$$a := -\frac{\tilde{\gamma}}{2} Q - \eta + \tilde{\gamma} R + \tilde{\gamma}\sqrt{\frac{Q}{\gamma}}\, x$$

$$\hat{a} := -\frac{\tilde{\gamma}}{2} Q - \eta + \tilde{\gamma}\sqrt{\frac{Q}{\gamma}}\, x$$

$$b := \frac{\tilde{\gamma}^2}{\gamma} z.$$

Appendix B. Free energy and order parameters for hard selection

The Hamiltonian (10) is explicitly given by

$$\begin{aligned}\mathcal{H}_h(\{V^\mu\}_\mu) &:= -\ln \int \mathrm{d}\boldsymbol{J}\, \mathcal{P}(\mathbb{D}, \boldsymbol{J}, \{V^\mu\}_\mu) \\ &= -\left[\frac{\tilde{\gamma}^2}{2N(\tilde{\gamma}\hat{Q} + 1)} \sum_{\mu,\nu} V^\mu V^\nu \sum_j \xi_j^\mu \xi_j^\nu S^\mu S^\nu - \frac{\tilde{\gamma}}{2} \sum_{\mu,j} (\xi_j^\mu)^2 - \eta \sum_\mu V^\mu\right] \\ &\quad + (N/2)\ln(\tilde{\gamma}\hat{Q} + 1) - \ln C\end{aligned}$$

where

$$C := \frac{1}{2^{\alpha N}} \left(\frac{\tilde{\gamma}}{2\pi}\right)^{\alpha N^2/2} \frac{1}{(1 + \exp[-\eta])^{\alpha N}} \left(\frac{1}{2\pi}\right)^{N/2}$$

with the order parameters

$$\hat{R} := \frac{1}{N} \sum_\mu V_a^\mu V^\mu$$

$$\hat{q} := \frac{1}{N} \sum_\mu V_a^\mu V_{\tilde{a}}^\mu$$

$$\hat{Q} := \frac{1}{N} \sum_\mu (V_a^\mu)^2 = \frac{1}{N} \sum_\mu V_a^\mu.$$

Averaging the partition function (19) yields

$$\langle Z_h^n\rangle = \left(\frac{1}{1+\mathrm{e}^{-\eta}}\right)^{\alpha N}\left(\frac{1}{1+n\beta\tilde\gamma/\gamma}\right)^{\alpha N^2/2}\sum_{\{V_a^\mu,V^\mu\}_\mu}\int\prod_{a,j}\mathrm{D}y_j^a$$
$$\times\exp\Bigg[-\eta\sum_\mu V^\mu+\frac{1}{2(1+n\beta\tilde\gamma/\gamma)}\Bigg(-n\beta\tilde\gamma\sum_\mu V^\mu$$
$$+\frac{2\tilde\gamma\sqrt{\beta}}{\sqrt{\tilde\gamma\hat Q+1}}\sum_j\sum_a y_j^a B_j\hat R+\frac{\tilde\gamma^2\beta}{\gamma(\tilde\gamma\hat Q+1)}\sum_j\sum_{a,\tilde a\neq a}y_j^a y_j^{\tilde a}\hat q$$
$$+\frac{\tilde\gamma^2\beta}{\gamma(\tilde\gamma\hat Q+1)}\sum_j\sum_a(y_j^a)^2Q\Bigg)-nN\beta\eta\hat Q\Bigg](\tilde\gamma Q+1)^{-nN\beta/2}C^{n\beta}.$$

The free energy f_h simplifies in the limit $\beta\to\infty$, where the scaling $\beta(\hat q-\hat Q)=:\hat z=\mathcal{O}(1)$ is used. We finally obtain f_h as a function of the actual order parameters at the saddle point:

$$f_h=\eta\hat Q+\frac{1}{2}\ln(\hat Q\tilde\gamma+1)-\frac{(\hat Q+2\hat R^2\gamma\rho^2)(\hat Q\tilde\gamma+1)\gamma\tilde\gamma^2}{4(\hat Q\gamma\tilde\gamma+\hat z\tilde\gamma^2+\gamma)^2}. \tag{B1}$$

A similar calculation using (21) yields the averages

$$\sum_j\langle J_j\rangle_{\boldsymbol{J}}B_j=N\frac{\hat R\tilde\gamma}{\hat Q\tilde\gamma+1}\left(1-\frac{\hat z\tilde\gamma^2}{\hat Q\gamma\tilde\gamma+\hat z\tilde\gamma^2+\gamma}\right)$$
$$\sum_j\langle {J_j}^2\rangle_{\boldsymbol{J}}=N\frac{\gamma\tilde\gamma^2(\hat Q+2\hat R^2\gamma)}{2(\hat Q\gamma\tilde\gamma+\hat z\tilde\gamma^2+\gamma)^2}+N\frac{1}{\hat Q\tilde\gamma+1}. \tag{B2}$$

In the limit $\alpha\to\infty$, the resulting order-parameter equations can be further simplified by making the scaling ansatze $\hat R=\alpha\hat R_0$, $\hat Q=\alpha\hat Q_0$, $\hat z=-\alpha\hat z_0$, where $\hat R_0,\hat Q_0,\hat z_0$ are independent of α as $\alpha\to\infty$. For $\gamma=\tilde\gamma$, the equation for the critical ratio of outliers η_c, where the trivial solution with zero order parameters has the global minimum of the free energy, is determined from

$$0=\eta-2\gamma\pi\eta\exp[\gamma+2\eta]\Phi^2\left[\sqrt{\gamma}-\sqrt{2\eta}\right]\Big/\left\{\exp\left[\sqrt{2\gamma\eta}\right]+\exp[\gamma/2+\eta]\right.$$
$$\left.+\sqrt{\pi\eta}\exp[\gamma/2+\eta]\left(-2\Phi\left[\sqrt{\gamma}-\sqrt{2\eta}\right]-2\mathrm{e}^{\eta}+2\mathrm{e}^{\eta}\Phi\left[\sqrt{2\eta}\right]\right)\right\}^2. \tag{B3}$$

References

[1] Seung H, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056
[2] Watkin T L H, Rau A and Biehl M 1993 *Rev. Mod. Phys.* **65** 499
[3] Opper M and Kinzel W Statistical mechanics of generalization *Physics of Neural Networks* ed J L van Hemmen, E Domany and K Schulten (Berlin: Springer) to appear
[4] Dempster A P, Laird N M and Rubin D B 1977 *J. R. Stat. Soc.* B **39** 1
[5] Jacobs R A, Jordan M I and Hinton G E 1991 *Neural Comput.* **3** 79
[6] Honerkamp J 1994 *Stochastic Dynamical Systems* (New York: VCH)
[7] Gardner E and Derrida B 1989 *J. Phys. A: Math. Gen.* **22** 1983
[8] Gyorgyi G 1990 *Phys. Rev.* A **41** 7097
[9] Biehl M and Mietzner A 1993 *Europhys. Lett.* **24** 421
[10] Barkai and Sompolinsky 1994 *Phys. Rev.* E **50** 1766
[11] Opper M 1994 *Phys. Rev. Lett.* **72** 2113
[12] Schwarze H 1993 *J. Phys. A: Math. Gen.* **26** 5781
[13] Gyorgyi G 1990 *Phys. Rev. Lett.* **64** 2957
[14] Opper M and Haussler D 1995 *Phys. Rev. Lett.* **75** 3772