Universität Würzburg

Institut für Theoretische Physik

Am Hubland, D–97074 Würzburg, Germany

Perceptron learning: the largest version space

Michael Biehl and Manfred Opper

Ref.: WUE-ITP-95-001
e-mail : biehl@physik.uni-wuerzburg.de
ftp : ftp.physik.uni-wuerzburg.de

PERCEPTRON LEARNING: THE LARGEST VERSION SPACE

MICHAEL BIEHL and MANFRED OPPER
Institut für Theoretische Physik
Julius–Maximilians–Universität
D–97074 Würzburg, Germany
E-mail: biehl@physik.uni-wuerzburg.de
opper@physik.uni-wuerzburg.de

ABSTRACT

We revisit the learning of a linearly separable rule with a single layer perceptron. The rule is taken to be correlated with a set of random training inputs, such that the concept is located in the largest of all version spaces. We formulate the corresponding statistical mechanics problem and study the model using the replica method. Replica symmetry is found to be broken, but the zero entropy approximation is interpreted as an estimate for the groundstate properties of the system. We investigate the typical overlap and generalization error in the largest version space and compare with the results for a typical random rule. The learning curves differ significantly, but preliminary studies indicate that the asymptotic decay of the generalization error with the number of examples could be the same apart from possible logarithmic corrections.

1. Introduction

Feedforward neural networks[1,2] can serve as a tool for *classification*: an output value is assigned to any possible input of the network. The adaptation of the network parameters, aiming at the realization of a specific classification scheme or *rule*, is termed *learning*. Typically, the correct output according to the unknown rule is known explicitly for only a limited set of example inputs. The extraction of the underlying concept from this *training set* is called *generalization* and the aim of training is to choose network parameters which provide the correct answer to novel input data with high probability.

Statistical mechanics methods have been applied successfully to simple models of such learning from examples. For recent reviews of the field see e.g. references 3,4,5, and this volume for an up–to–date overview. Usually, the generalization behavior of very large networks is studied on average over "quenched disorder" provided by a *randomness* in the training set, the parametrized rule, additional noise, etc.

In this paper we revisit the statistical mechanics of the single layer *perceptron*[6,7]. This model has been particularly attractive because of its simplicity and clearity. We will study the learning of a *linearly separable* concept but refrain from defining it through random parameters which are completely independent from the training set. The introduction of specific correlations between the rule and the actual example inputs enables us to study concepts of varying difficulty.

In the next section we specify the model and introduce the size of the *version space* as a possible "measure of difficulty" for linearly separable rules. Section 3 explains the basic formalism and analytic treatment of the problem. Well known results from the statistical mechanics of the perceptron can be rederived as special cases of our model. We present our results for generalization in the largest version space in section 4. The replica symmetric treatment is proven to fail and an alternative zero entropy approach is used to obtain better estimates for the groundstate properties. Finally we give a short summary and discuss possible extensions of our considerations.

2. The model

2.1. The perceptron

The by far most thoroughly studied feedforward neural network is the *single layer perceptron*[6,7]. For any N–dimensional input $\boldsymbol{\xi} \in I\!R^N$ the network output is given by

$$\sigma(\boldsymbol{\xi}) = \mathrm{sign}(\mathbf{J} \cdot \boldsymbol{\xi}) \qquad \in \{-1, +1\}, \tag{1}$$

where $\mathbf{J} \in I\!R^N$ is the vector of adjustable network weights. Obviously the length of the weight vector can be fixed without changing the classification, we will set $\mathbf{J}^2 = N$ in the following.

The geometric interpretation of equation (1) is that the two classes of inputs are separated by a hyperplane perpendicular to $\mathbf{J}$ and through the origin. $\sigma(\boldsymbol{\xi})$ in (1) is called a *linearly separable* function of the inputs.

On the other hand, a dual picture describes the situation in weight space: A certain input $\boldsymbol{\xi}$ defines a hyperplane above (below) which all vectors $\mathbf{J}$ would produce an output $\sigma(\boldsymbol{\xi}) = +1$ (-1 respectively).

Now consider a set of input/output–pairs $\{\boldsymbol{\xi}^\mu, \sigma^\mu\}_{\mu=1,\ldots p}$. The hyperplanes given by the $\boldsymbol{\xi}^{\boldsymbol{\mu}}$ define, together with their orientations σ^μ, the set

$$\begin{aligned} \mathcal{V}(\vec{\sigma}) &= \left\{ \mathbf{J} \mid \mathrm{sign}(\mathbf{J} \cdot \boldsymbol{\xi}^{\boldsymbol{\mu}}) = \sigma^\mu \ \text{ for } \mu = 1, \ldots p, \quad \mathbf{J}^2 = N \right\} \\ &= \left\{ \mathbf{J} \mid \mathbf{J} \cdot \boldsymbol{\xi}^{\boldsymbol{\mu}} \sigma^\mu > 0 \ \text{ for } \mu = 1, \ldots p, \quad \mathbf{J}^2 = N \right\} \end{aligned} \tag{2}$$

$\mathcal{V}(\vec{\sigma})$ contains all normalized weight vectors consistent with the particular labelling $\vec{\sigma} = (\sigma^1, \ldots, \sigma^p)^T$. We can now define the volume of $\mathcal{V}(\vec{\sigma})$ as

$$V(\vec{\sigma}) = \int_{\mathcal{V}(\vec{\sigma})} d\mathbf{J} = \int d\mathbf{J}\ \delta(\mathbf{J}^2 - N) \ \prod_{\mu=1}^{p} \Theta(\mathbf{J} \cdot \boldsymbol{\xi}^{\boldsymbol{\mu}} \sigma^\mu). \tag{3}$$

This quantity was first introduced in the pioneering work of E. Gardner[8] where the typical value of $\ln V(\vec{\sigma})$ for a random labelling $\vec{\sigma}$ is calculated. We will shortly return to this problem later in the paper.

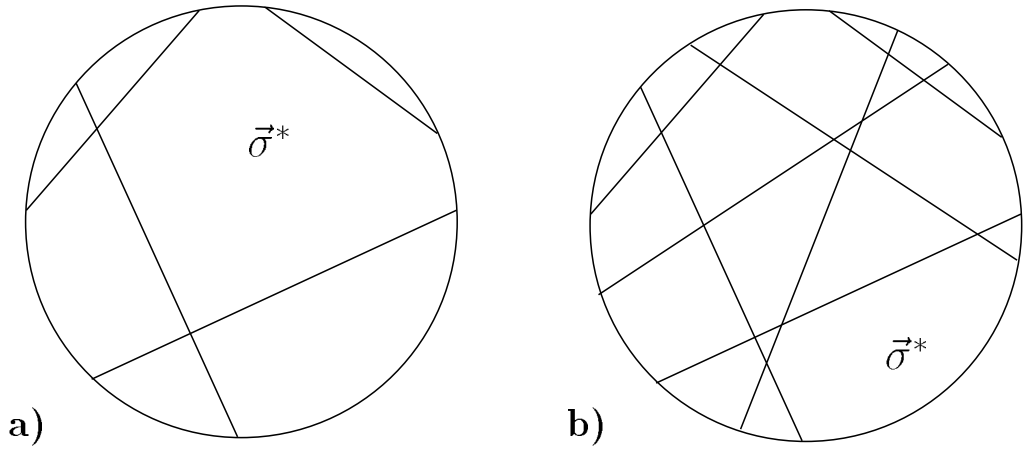

Figure 1: Illustration of the geometric interpretation of learning in the perceptron weight space: A set of example inputs defines volumes of weight vectors with the same output $\vec{\sigma}$. Figure **a)** sketches the situation for a low number of examples, the largest version space is labelled $\vec{\sigma}^*$. With more examples known (Fig.**b)**) the largest volume is to be found in a completely different region of $\mathbf{J}$–space.

A set of p input vectors partitions the space of normalized weights into at most 2^p such volumes. Note that many of the labellings might in fact not be linearly separable, corresponding to $\mathcal{V}(\vec{\sigma}) = \emptyset$ and $V(\vec{\sigma}) = 0$. Figure 1 shows a sketch in the sense of the above geometric interpretation.

2.2. Learning a linearly separable rule

Now consider a perceptron learning a linearly separable rule, defined through an unknown teacher vector $\mathbf{J}^*$, $(\mathbf{J}^*)^2 = N$. The only available information is contained in the training set $\{\boldsymbol{\xi}^{\boldsymbol{\mu}}, \sigma^{*\mu} = \mathrm{sign}(\mathbf{J}^* \cdot \boldsymbol{\xi}^{\boldsymbol{\mu}})\}_{\mu=1,\dots p}$ and $\mathcal{V}(\vec{\sigma}^*)$ is the set of all students with zero *training error*

$$\epsilon_t = \frac{1}{p} \sum_{\mu=1}^{p} \Theta(-\mathbf{J} \cdot \boldsymbol{\xi}^{\boldsymbol{\mu}} \sigma^{*\mu}). \tag{4}$$

Each of its elements can be considered a possible hypothesis for $\mathbf{J}^* \in \mathcal{V}(\vec{\sigma}^*)$, hence the term *version space* for this set.

A measure of a particular student's performance is the so–called *generalization error*, the probability for disagreement between student and rule output for an arbitrary input:

$$\epsilon_g = \langle \Theta(-(\mathbf{J} \cdot \boldsymbol{\xi})(\mathbf{J}^* \cdot \boldsymbol{\xi}) \rangle_{\xi}, \tag{5}$$

where $\langle \dots \rangle_{\xi}$ denotes an average over the distribution of possible input vectors. In the following we consider only independent, identically distributed random components

ξ_j^μ with zero mean and unit variance. In this simple case one obtains

$$\epsilon_g = \frac{1}{\pi} \arccos\left(\frac{1}{N}\mathbf{J}\cdot\mathbf{J}^*\right), \tag{6}$$

which is easily understood in terms of the above geometric interpretation[4]. If an input is drawn from a uniform distribution, the probability for misclassification is proportional to the angle between student and teacher vector.

Most statistical mechanics results[3,4,5] have been derived for a "typical teacher", i.e. a vector $\mathbf{J}^*$ randomly drawn according to some *a priori* measure $d\mu(\mathbf{J})$ which is independent of the actual training inputs. The simplest such measure is constant everywhere on the sphere $\mathbf{J}^2 = N$, but it would be straightforward to incorporate more complicated $d\mu(\mathbf{J})$ and redefine for example $V(\vec{\sigma})$ accordingly.

Various strategies for picking a specific hypothesis from the version space have been considered. For example, a so–called Gibbs procedure[9] would accept any $\mathbf{J} \in \mathcal{V}(\vec{\sigma}^*)$ with equal probability, whereas more sophisticated algorithms[5,10,11] prefer students close to the "center of mass" of $\mathcal{V}(\vec{\sigma}^*)$. Recently, also the "worst possible student", i.e. the $\mathbf{J} \in \mathcal{V}(\vec{\sigma}^*)$ with the highest expectation value for ϵ_g, was studied[12].

In the thermodynamic limit $N \to \infty$, $p = \alpha N$, all these prescriptions yield a generalization error which decays like

$$\epsilon_g(\alpha) \propto \alpha^{-1} \quad \text{as} \quad \alpha \to \infty. \tag{7}$$

Note that we are dealing with the specific case of a *learnable* rule: the student network can achieve perfect generalization by properly choosing its weights. Therefore, the restriction of $\mathbf{J}$ to the version space is a useful training strategy here, whereas in more general cases the minimization of ϵ_t can lead to *overtraining* effects[3,4,5,9]. Note furthermore, that zero training error seems to be sufficient for an asymptotic behavior of the form (7), but is not a necessary condition[13,14,15].

2.3. The largest version space

Consider a specific realization of a set of example inputs. Intuitively, the volume $V(\vec{\sigma}^*)$ can be interpreted as a measure for how difficult a linearly separable rule $\mathbf{J}^* \in \mathcal{V}(\vec{\sigma}^*)$ would be to infer from the training set.

If the teacher was to be found in a rather small volume we would expect a large overlap $\mathbf{J}\cdot\mathbf{J}^*/N$ for a hypothesis $\mathbf{J}$ which is for example randomly drawn from the same volume. In a larger version space the same Gibbs procedure should clearly yield a higher generalization error.

This rather hand–waving argument disregards the fact that also the actual shape of the version space plays an important role. It holds, more strictly speaking, only for perfectly symmetric volumes, e.g. spheres.

In the following, we are particularly interested in the largest version space. It corresponds to a rule about which the given training inputs reveal relatively little information. In order to study this problem, we have to consider a teacher which is highly correlated to the given examples. It cannot be drawn independently according to some fixed measure. Figure 1 illustrates this point: as the number of examples increases, the largest version space might be found in a completely different region of weight space.

In the next section we will formulate the corresponding statistical mechanics problem. An energy is assigned to all possible labellings $\vec{\sigma}$ such that its minimum corresponds to the largest of all volumes $V(\vec{\sigma})$. The formalism enables us also to derive more general results by controlling the (average) version space volume through a temperature–like parameter.

3. Statistical mechanics

3.1. The general formalism

The method was recently introduced by Monasson and O'Kane[16] in a different context. The authors study domains of different hidden representations in a specific multilayer neural network, which correspond to the same classification of the examples. Here, we will consider the perceptron outputs themselves as the thermodynamic variables.

For fixed input examples $\{\boldsymbol{\xi}^{\mu}\}_{\mu=1,\dots p}$ we assign a cost function to every labelling $\vec{\sigma}$ of the inputs:

$$\mathcal{H}(\vec{\sigma}) = -\ln V(\vec{\sigma}) \tag{8}$$

with V defined in equation (3). If we interpret the output labels σ^{μ} as the p interacting degrees of freedom in a system with energy $\mathcal{H}(\vec{\sigma})$ the corresponding partition function reads

$$\mathcal{Z} = \sum_{\vec{\sigma}\in\{\pm 1\}^p} \exp\left[-\beta\mathcal{H}(\vec{\sigma})\right] = \sum_{\vec{\sigma}\in\{\pm 1\}^p} V^{\beta}(\vec{\sigma}), \tag{9}$$

at the inverse "temperature" β. Formally, we will consider only integer values of β, which allows for rewriting

$$\mathcal{Z} = \sum_{\vec{\sigma}} \left(\int \prod_b d\mathbf{J}_b \right) \prod_b \delta(\mathbf{J}_b^2 - N) \prod_{\mu,b} \Theta(\mathbf{J}_b \cdot \boldsymbol{\xi}^{\mu} \sigma^{\mu}), \tag{10}$$

where all products over b are from 1 to β, products over μ from 1 to p.

This is in principle identical with the replication in reference 8, which was used to calculate $\ln V$ in the limit $\beta \to 0$. Here, however, the intention is to consider $\beta \to \infty$, and the restriction to integer β should not limit the validity of the results for the groundstate of the system.

At a given temperature, thermal averages can be calculated as proper derivatives of the free energy $-\ln \mathcal{Z}/(N\beta)$. In order to obtain typical results, this quantity has to be averaged over the quenched disorder provided by the randomness in the inputs $\{\boldsymbol{\xi}^\mu\}$. This yields the quenched free energy

$$f = -\frac{1}{\beta N}\langle \ln \mathcal{Z}\rangle_\xi \tag{11}$$

on average over the uniform input distribution (section 2.2).

The calculation of $\langle \ln \mathcal{Z}\rangle_\xi$ is performed by applying the replica trick once more:

$$\langle \ln \mathcal{Z}\rangle_\xi = \lim_{n\to 0}\frac{\partial}{\partial n}\ln\langle \mathcal{Z}^n\rangle_\xi = \lim_{n\to 0}\frac{\partial}{\partial n}\ln\left\langle\left(\sum_{\vec{\sigma}} V^\beta(\vec{\sigma})\right)^n\right\rangle_\xi . \tag{12}$$

$\mathcal{Z}^n$ is the partition function of an n–fold replicated system with np degrees of freedom $\sigma_a^\mu, a = 1,\dots n$. For integer n it can be written as

$$\mathcal{Z}^n = \prod_a \sum_{\vec{\sigma}_a} \int \prod_{a,b}\left(d\mathbf{J}_{a,b}\delta(\mathbf{J}_{a,b}^2 - N)\right)\prod_{\mu,a,b}\Theta(\mathbf{J}_{a,b}\cdot\boldsymbol{\xi}^\mu\sigma_a^\mu) \tag{13}$$

Note that the limit $n \to 0$ requires the extrapolation to real n, which might cause subtle difficulties[17].

Performing the average over the input examples[8] leads to the introduction of order parameters

$$q_{a\tilde{a}b\tilde{b}} = \frac{1}{N}\mathbf{J}_{a,b}\cdot\mathbf{J}_{\tilde{a},\tilde{b}}, \quad \text{with} \quad q_{aabb} = 1 \tag{14}$$

which represent the typical overlap between two weight vectors. These quantities have to be determined by means of a saddle point integration.

The (thermal and quenched) average energy of the system is then given by

$$\frac{1}{N}\Big\langle\langle\mathcal{H}\rangle_\beta\Big\rangle_\xi = -\frac{1}{N}\Big\langle\langle\ln V\rangle_\beta\Big\rangle_\xi = \frac{\partial}{\partial\beta}(-\beta f), \quad \text{and} \quad s = \beta^2\frac{\partial f}{\partial\beta} \tag{15}$$

is the corresponding entropy at temperature $1/\beta$.

3.2. Replica symmetry

In a replica symmetric ansatz[16,17] we distinguish only two order parameters:

$q = q_{aab\tilde{b}}$	the typical overlap of two weight vectors in the same version space $\mathcal{V}(\vec{\sigma}_a)$ $(b \neq \tilde{b})$, and
$r = q_{a\tilde{a}b\tilde{b}}$	the overlap of two weight vectors from different volumes $a \neq \tilde{a}$, at arbitrary "temperatures" $b, \tilde{b}$.

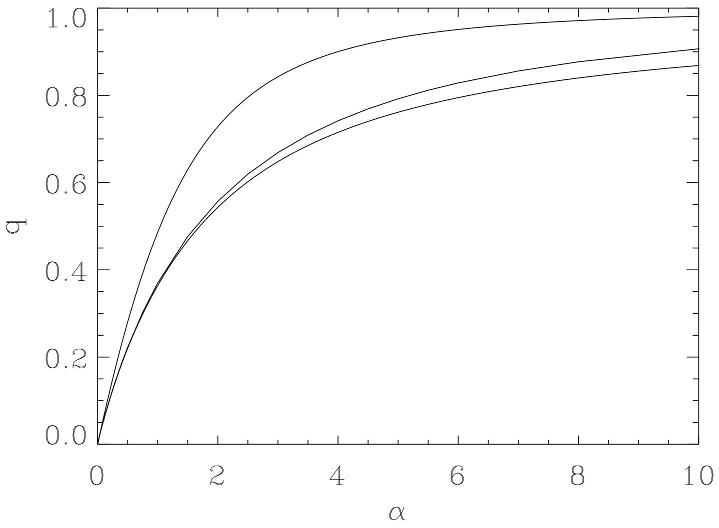

Figure 2: The overlap q in a (one) version space as a function of α. The upper curve corresponds to the learning of a typical rule by a Gibbs student[9]. The second curve is the overlap in the largest version space as obtained in the zero entropy formalism, and the lowest is the result of the naive $\beta \to \infty$ limit.

Within this simplified scheme the free energy is the extremum of

$$\begin{aligned} f &= -\frac{1}{2}(1+\ln(2\pi)) - \frac{1}{2}\frac{r}{1-q+\beta(q-r)} - \frac{1}{2}\frac{\beta-1}{\beta}\ln(1-q) \\ &\quad -\frac{1}{2\beta}\ln(1-q+\beta(q-r)) - \frac{\alpha}{\beta}\int Dz \ln\left(\sum_{\sigma=\pm1}\int Dt\Phi^{\beta}\left[-\frac{\sqrt{r}\sigma z+\sqrt{q-r}t}{\sqrt{1-q}}\right]\right) \end{aligned} \tag{16}$$

with the abbreviation $Dt = dt\exp[-t^2/2]/\sqrt{2\pi}$ and the function $\Phi[y] = \int_{-\infty}^{y} Dt$. The order parameters q and r are determined by the conditions $\partial f/\partial q = 0$ and $\partial f/\partial r = 0$, respectively.

The inverse temperature β fixes a value of $\ln V$ (on average), and the order parameter q measures the overlap of two random vectors found in a (one) version space of this particular size. For a Gibbs student this is also the typical overlap with a teacher vector that could be anywhere in the same volume with constant probability. Thus, q translates into the generalization error

$$\epsilon_g = \frac{1}{\pi}\arccos(q) \tag{17}$$

which can now be calculated for varying version space size.

Before we investigate generalization in the largest version space, let us resort to two special cases which correspond to well known results from the statistical mechanics

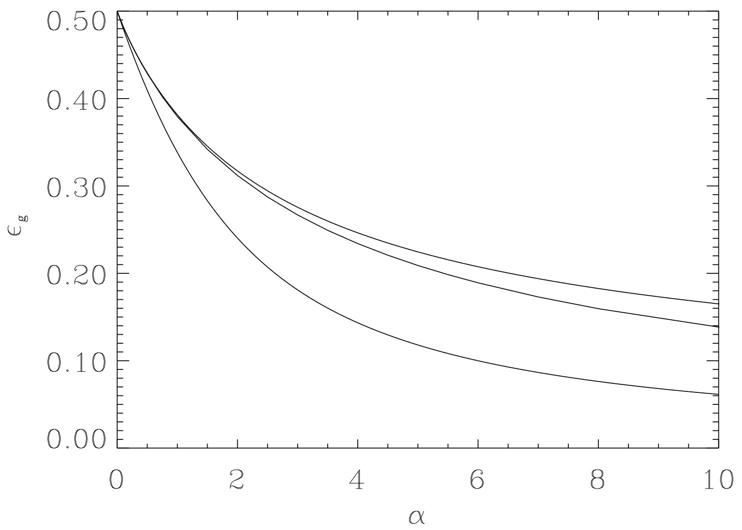

Figure 3: The generalization error $\epsilon_g = \arccos(q)/\pi$ corresponding to the overlaps shown in figure 2. From bottom to top: typical rule, zero entropy estimate for the largest version space, replica symmetric $\beta \to \infty$ solution.

of the perceptron.

3.3. The capacity problem, $\beta \to 0$

At infinite temperature, the energy of a labelling becomes irrelevant and the thermal average reduces to summing over all possible $\vec{\sigma}$ with an equal, volume independent weight 2^{-p}. For instance, the average (negative) energy reads

$$\begin{aligned} -\frac{1}{N}\Big\langle \langle \mathcal{H} \rangle_{\beta=0} \Big\rangle_\xi &= \frac{1}{N}\Big\langle \langle \ln V \rangle_{\beta=0} \Big\rangle_\xi = \lim_{\beta \to 0} \frac{\partial}{\partial \beta} \langle \ln \mathcal{Z}(\beta) \rangle_\xi = \\ &= \lim_{\beta \to 0} \left\langle \frac{1}{\mathcal{Z}(0)} \sum_{\vec{\sigma}} \frac{\partial}{\partial \beta} V^\beta(\vec{\sigma}) \right\rangle_\xi \propto \langle \langle \ln V(\vec{\sigma}) \rangle_\sigma \rangle_\xi, \end{aligned} \tag{18}$$

with $\langle \ldots \rangle_\sigma$ denoting an average over independent and unbiased random outputs. The partition function $\mathcal{Z}(0) = \sum_{\vec{\sigma}} V^0(\vec{\sigma})$ simply "counts" all non–empty version spaces, i.e. linearly separable $\vec{\sigma}$. Naively inserting $\beta = 0$ in equation (16) yields

$$\frac{1}{N} \ln \mathcal{Z}(\beta = 0) = - \beta f(\beta)|_{\beta=0} = \alpha \ln 2 = s(\beta = 0). \tag{19}$$

This seems to indicate [a] that all $2^p = \mathcal{Z}(0)$ labellings are linearly separable, regardless of the actual value of $p = \alpha N$.

On the other hand, an expansion for $\beta \approx 0$ yields

$$f(\beta \approx 0) = f(0) + \beta \left. \frac{\partial f}{\partial \beta} \right|_{\beta=0} = \left. \frac{\partial}{\partial \beta} (\beta f(\beta)) \right|_{\beta=0} = \frac{1}{N} \Big\langle \langle \ln V \rangle_{\beta=0} \Big\rangle_\xi . \tag{20}$$

[a] It also implies that the "$\propto$" in equation (18) can be replaced by an equals sign.

Thus, the free energy for small β is given by the "Gardner entropy" (at zero stability)[8] which we recover for $r = 0$:

$$\begin{aligned}\frac{1}{N}\Big\langle \langle \ln V \rangle_{\beta=0} \Big\rangle_{\xi} &= \frac{1}{2}(1+\ln(2\pi)) + \frac{1}{2}\frac{q}{1-q} + \frac{1}{2}\ln(1-q) \\ &+\alpha \int Dx \ln \Phi\left[-\sqrt{\frac{q}{1-q}}x\right]. \qquad (21)\end{aligned}$$

The well known result of an extremization with respect to q is that the overlap tends to 1 as $\alpha \to \alpha_c = 2$. Above this so–called *storage capacity* of the perceptron the typical version space volume shrinks to zero ($\langle\langle \ln V \rangle\rangle \to -\infty$).

3.4. The typical teacher, $\beta \to 1$

For $\beta = 1$ the partition function adds up all non–zero volumes in weight space $\mathcal{Z} = \sum_{\vec{\sigma}} V(\vec{\sigma})$ and should simply equal the surface of the sphere $\mathbf{J}^2 = N$. In fact, the saddle point equations are satisfied for $r = 0$ and arbitrary q, yielding

$$f(\beta = 1) = -\frac{1}{N}\langle \ln \mathcal{Z} \rangle_{\xi} = -\frac{1}{2}(1+\ln(2\pi)) \qquad (22)$$

as expected.

This particular temperature corresponds to averaging over all version spaces with a weight proportional to their volume, or equivalently, averaging over all possible rules with a constant measure anywhere on the N–dimensional sphere[10]. More formally:

$$\lim_{\beta\to 1} \Big\langle \langle \ln V \rangle_{\beta=1} \Big\rangle_{\xi} = \left\langle \lim_{\beta\to 1} \frac{\partial}{\partial\beta} \ln \sum_{\vec{\sigma}} V^{\beta}(\vec{\sigma}) \right\rangle_{\xi} = \left\langle \frac{1}{\mathcal{Z}(\beta=1)} \sum_{\vec{\sigma}} V(\vec{\sigma}) \ln V(\vec{\sigma}) \right\rangle_{\xi}. \qquad (23)$$

In the vicinity of $\beta = 1$ one obtains from (16) with $r = 0$

$$\begin{aligned} f(\beta \approx 1) &= f(1) + s(1)\frac{\beta - 1}{\beta^2} \quad \text{where} \\ s(\beta = 1) &= -\frac{q}{2} - \ln(1-q) - 2\alpha \int Dx \Phi\left[-\sqrt{\frac{q}{1-q}}\right] \ln \Phi\left[-\sqrt{\frac{q}{1-q}}\right]. \qquad (24)\end{aligned}$$

This is identical to the entropy (free energy respectively) calculated in reference 9 for the (noise free) learning of a typical rule. Extremization of the expression with respect to q yields the well known results for the Gibbs student with the asymptotics

$$\epsilon_g(\alpha) = \frac{1}{\pi} \arccos\left(q(\alpha)\right) \approx \frac{0.62}{\alpha} \quad \text{for} \quad \alpha \to \infty. \qquad (25)$$

The overlap $q(\alpha)$ and the generalization error $\epsilon_g(\alpha)$ are plotted in figures 2 and 3 respectively.

4. Generalization in the largest version space

4.1. The limit $\beta \to \infty$

Assuming the existence of a unique groundstate the two order parameters q and r should coincide as $\beta \to \infty$. At zero temperature only vectors from a single (the largest) volume are admitted.

We make a corresponding ansatz introducing the non–negative order parameter

$$w = \beta(q - r). \tag{26}$$

Now the free energy is given as the extremum (with respect to q and w) of

$$\begin{aligned} f(\beta \to \infty) &= \frac{1}{2}(1 + \ln(2\pi)) + \frac{1}{2}\frac{q}{1-q+w} + \frac{1}{2}\ln(1-q) \\ &\quad + 2\alpha \int_0^\infty Dt \max_z \left\{ -\frac{1-q}{2w} z^2 + \ln \Phi\left[\sqrt{\frac{q}{1-q}}t - z\right]\right\}. \end{aligned} \tag{27}$$

The numerical solution of the saddle point equations yields q and w as a function of α. Figure 2 shows $q(\alpha)$ in comparison with the result for the typical teacher[9], in figure 3 the corresponding generalization errors are plotted. As expected, ϵ_g in the largest version space is much higher than for the typical rule.

It is important to keep in mind, that this "learning curve" does not describe any realistic training process with a fixed rule being learnt. For each value of α the teacher is assumed somewhere in the currently largest volume which – as explained above – may vary drastically with the number of examples.

The limit $\alpha \to \infty$ is of particular interest. The question is whether the $1/\alpha$ asymptotic decay of the generalization error is valid only for the typical case or holds as well for particularly "difficult" rules.

As α increases, the size of any non–empty (and also the largest) version space will decrease since more and more conditions are to be satisfied. In turn, the weight space is of course partitioned into an increasing number of such volumes. This is reflected in the fact that the typical overlap approaches its maximum value $q = 1$ as α tends to infinity.

The numerical solution suggests a behavior

$$w \to \text{const.} \quad \text{and} \quad (1-q)\ln(1-q) \propto -\alpha^{-2/3} \quad \text{as} \quad \alpha \to \infty. \tag{28}$$

Using this ansatz we can solve the saddle point equations selfconsistently and obtain

$$w \to \frac{3}{2} \quad \text{and} \quad (1-q) \approx \left(\frac{3^{5/3}\pi^{1/3}}{2^{7/3}}\right)\frac{\alpha^{-2/3}}{\ln \alpha} \tag{29}$$

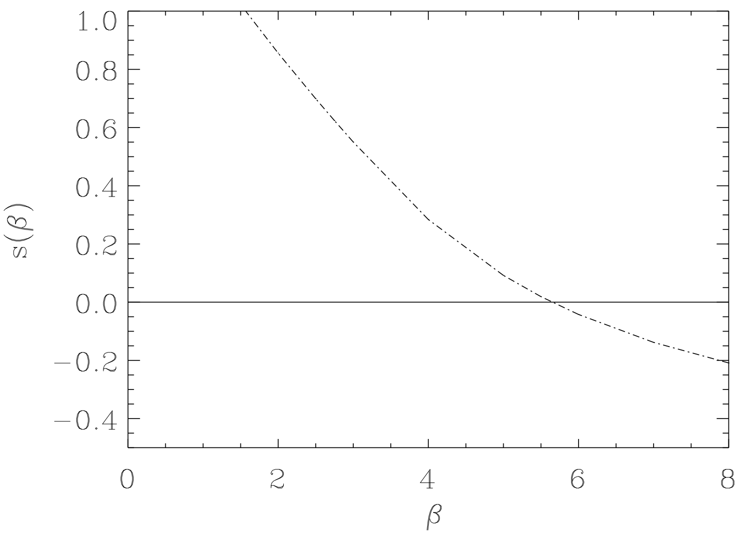

Figure 4: The replica symmetric entropy $s = \beta^2 \partial f / \partial \beta$ as a function of the inverse temperature for $\alpha = 2.5$; it becomes negative at $\beta_o \approx 5.65$ and diverges like $-\ln \beta$ for $\beta \to \infty$.

yielding an asymptotic generalization error

$$\epsilon_g = \left(\frac{3^{5/6}}{2^{2/3} \pi^{5/6}} \right) \frac{\alpha^{-1/3}}{\sqrt{\ln \alpha}} \approx \frac{0.606}{\alpha^{1/3} \sqrt{\ln \alpha}} \quad \text{for} \quad \alpha \to \infty. \tag{30}$$

This would in fact imply a decay much slower than for the typical case and suggest that among the linearly separable rules a spectrum of very different "degrees of difficulty" could be found.

A more careful analysis, however, shows that the results obtained in the replica symmetric scheme cannot be valid for large β. In the next subsection we discuss a more sophisticated approximation of the true groundstate.

4.2. The zero entropy solution

In order to check for a possible violation of replica symmetry (RS) we calculate the entropy

$$s = \beta^2 \frac{\partial f}{\partial \beta}$$

for general β from the free energy (16). Since we are dealing with a system of discrete degrees of freedom $\sigma^\mu = \pm 1$, the entropy is required to be non–negative at any temperature.

For large enough β, however, we find this condition violated as shown in figure 4 for $\alpha = 2.5$. In the limit $\beta \to \infty$ the result is to leading order in β

$$s(\beta \to \infty) \approx -\frac{\alpha}{2} \ln \beta \to -\infty. \tag{31}$$

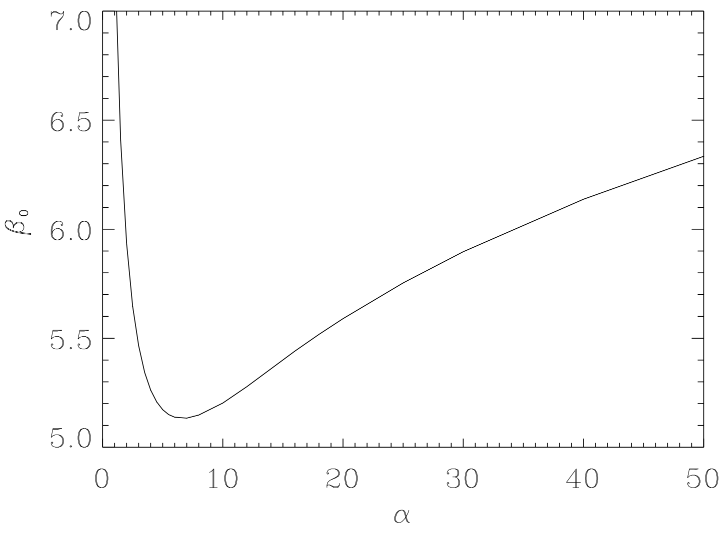

Figure 5: The critical temperature β_o as defined by $s(\beta_o) = 0$ vs. α. Numerical results suggest an asymptotic growth $\beta_o \propto \ln \alpha$ for $\alpha \to \infty$.

This drastic violation indicates that our RS ansatz is too simple for studying the low temperature behavior of the system.

A correct treatment would require replica symmetry breaking (RSB) according to Parisi's scheme[17]. At this point we resort to a heuristically motivated approximation of the groundstate which has proven useful in other models[2,16,18,19]. A possible scenario is the freezing of the system to its groundstate at a non–zero temperature $1/\beta_o$ given by $s(\beta_o) = 0$. In physical terms it corresponds to the point where the number of allowed states becomes finite (less than exponential in N). In figure 4 the critical temperature $\beta_o(\alpha)$ is plotted for our model, it appears to grow logarithmically with α.

We assume that the solution found at β_o persists for lower temperatures. The approximation should be exact, if the correct $s(\beta)$ was available for $\beta \leq \beta_o$, that is if RS was broken only for $\beta > \beta_o$. This was shown to be true for the calculation of the storage capacity of a so–called Ising perceptron[18]. In other cases, the method yields at least better estimates than the RS–($\beta \to \infty$)–solution[16,19,20].

The overlap $q(\alpha)$ as obtained by this approximation is plotted in figure 2. The comparison shows that its increase with α is still slower than for the learning of a typical rule but significantly faster than suggested by the naive $\beta \to \infty$ result. Figure 3 displays the corresponding generalization error $\epsilon_g(\alpha)$.

Preliminary studies of the asymptotics $\alpha \to \infty$ seem to indicate that the basic behavior is

$$(1-q) \propto \mathcal{O}(\alpha^{-2}) \quad \text{and} \quad \epsilon_g \propto \mathcal{O}(\alpha^{-1}), \tag{32}$$

but subject to logarithmic corrections. At this point however, the version space of possible decays is too large to draw a definite conclusion.

5. Summary and outlook

We have outlined a formalism in which it is possible to study linearly separable concepts of definite difficulty. The focus has been on learning a rule within the largest of all version spaces for a given set of training inputs. The replica symmetric treatment turns out to be incorrect, but can be refined in terms of a zero entropy consideration. Its validity should be checked by means of an RSB analysis.

Our preliminary results indicate that the variation of the learning curve asymptotics is rather limited within the space of linearly separable rules. Even for a rule in the largest version space the basic dependence appears to be

$$\epsilon_g \propto \alpha^{-1} \quad \text{for} \quad \alpha \to \infty$$

with possible logarithmic corrections. A more thorough investigation remains to be done.

The formalism allows for investigations beyond the scope of this paper. We have already discussed the special cases $\beta \to 0$ and $\beta \to 1$ respectively, but admitting arbitrary temperatures (including $\beta < 0$)[16] enables us to work out the entropy of the system as a function of the average $\ln V$. Details will be published elsewhere.

An application of the method to the discretized weight space of the Ising perceptron seems also promising and should help to gain new insights into this model.

The statistical mechanics developed here is strongly reminiscent of the recent analysis of unsupervised learning[21,22,23]. In fact, the analogy is more than formal, in particular with the concept of unsupervised maximal stability[21]. Given only a set of example inputs, the search for the labelling with the largest volume in weight space could be a useful strategy for the extraction of information about the input distribution. We will investigate this point by introducing a structure in the input data[21,22,23].

As already discussed at length, the largest version space corresponds to a rule which is rather difficult to be inferred from the training set. An investigation of the true *worst case* (given random examples) should consider the pair $(\mathbf{J}, \mathbf{J}^*)$ with minimal mutual overlap but perfect agreement on all examples. A model in which both vectors $\mathbf{J}$ and $\mathbf{J}^*$ represent the annealed variables is currently studied.

Acknowledgements
We would like to thank M. Bouten, C. van den Broeck, A. Engel, W. Kinzel, B. Lautrup, A. Scharnagl, S.A. Solla, O. Winther, and the participants of this workshop for many useful and stimulating discussions. We thank R. Monasson and D. O'Kane for communicating the results of ref. 16 prior to publication.

Both authors acknowledge financial support from the Deutsche Forschungsgemeinschaft.

References

1. J.A. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation* (Addison–Wesley, Redwood City CA, 1991).
2. B. Müller and J. Reinhardt, *Neural Networks* (Springer, Berlin, 1990)
3. H.S. Seung, H. Sompolinsky, and N. Tishby, *Phys. Rev.* **A45** (1992) 6056.
4. T.L.H. Watkin, A. Rau, and M. Biehl, *Rev. Mod. Phys.* **65** (1993) 499.
5. M. Opper and W. Kinzel, in *Physics of Neural Networks III*, series editors E. Domany, J.L. van Hemmen, and K. Schulten (Springer, Berlin, in press).
6. F. Rosenblatt, *Principles of Neurodynamics* (Spartan, Ney York N.Y., 1962).
7. M.L. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge MA, 1969 and 1988).
8. E. Gardner, *J. Phys. A: Math. Gen.* **21** (1988) 257.
   E. Gardner and B. Derrida, *J. Phys. A: Math. Gen.* **21** (1988) 271.
9. G. Györgyi and N. Tishby, in *Neural Networks and Spin Glasses* , eds. W.K. Theumann and R. Koeberle (World Scientific, Singapore, 1990).
   G. Györgyi, *Phys. Rev. Lett.* **64** 2957.
10. M. Opper and D. Haussler, *Phys. Rev.* **E 66** (1991) 2677.
11. T.L.H. Watkin, *Europhys. Lett.* **21** (1993) 871.
12. A. Engel and C. van den Broeck, *Phys. Rev. Lett.* **71** (1993) 1772.
    A. Engel and W. Fink, *J. Phys. A: Math. Gen.* **26** (1993) 6893.
13. O. Kinouchi and N. Caticha, *J. Phys. A: Math. Gen.* **25** (1992) 6243.
14. M. Biehl and P. Riegler, *Europhys. Lett.* **28** (1994) 525.
15. N. Barkai, H.S. Seung, and H. Sompolinsky, On–line Learning of Dichotomies (1994) preprint, see also this volume.
16. R. Monasson and D. O'Kane, *Europhys. Lett.* **27** (1994) 85.
17. M. Mezard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and beyond*, (World Scientific, Singapore, 1987).
18. W. Krauth and M. Mezard, *J. Phys. (Paris)* **50** (1989) 3057.
19. R. Meir and J.F. Fontanari, *J. Phys. A: Math. Gen.* **26** (1993) 1077.
20. H. Schwarze, *J. Phys. A: Math. Gen.* **26** (1993) 5781.
21. M. Biehl and A. Mietzner, *J. Phys. A: Math. Gen.* **27** (1994) 1885.
    A. Mietzner, M. Opper, and W. Kinzel, *J. Phys. A: Math. Gen.* in press.
22. T.L.H. Watkin and J.-P. Nadal, *J. Phys. A: Math. Gen.* **27** (1994) 1899.
23. C. Marangi, M. Biehl, and S.A. Solla, *Supervised learning from structured input examples*, (1994) preprint.