united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
**abdus salam**
international centre for theoretical physics

SMR/1157 - 3

*SCHOOL ON*
*NEURAL INFORMATION PROCESSING*

*3 - 28 May 1999*

# STATISTICAL MECHANICS OF LEARNING

Andreas ENGEL
Institute of Theoretical Physics
Otto-von-Guericke University
Universitatsplatz 2
Postfach 4120
D-39016 Magdeburg
GERMANY

# Statistical Mechanics of Learning

Andreas Engel
Institute for Theoretical Physics
Otto-von-Guericke-University Magdeburg
(andreas.engel@physik.uni-magdeburg.de)

- Getting started
- The Gardner analysis
- Learning by minimizing cost functions
- Noisy teachers
- Variations of perceptron learning

# 1 First lecture: Getting started

## 1.1 Introduction

The aim of these lectures is to give a rather detailed account on how methods of statistical mechanics can be used to quantitatively investigate learning from examples in artificial neural networks. Some types of networks are shown in figure 1 below.

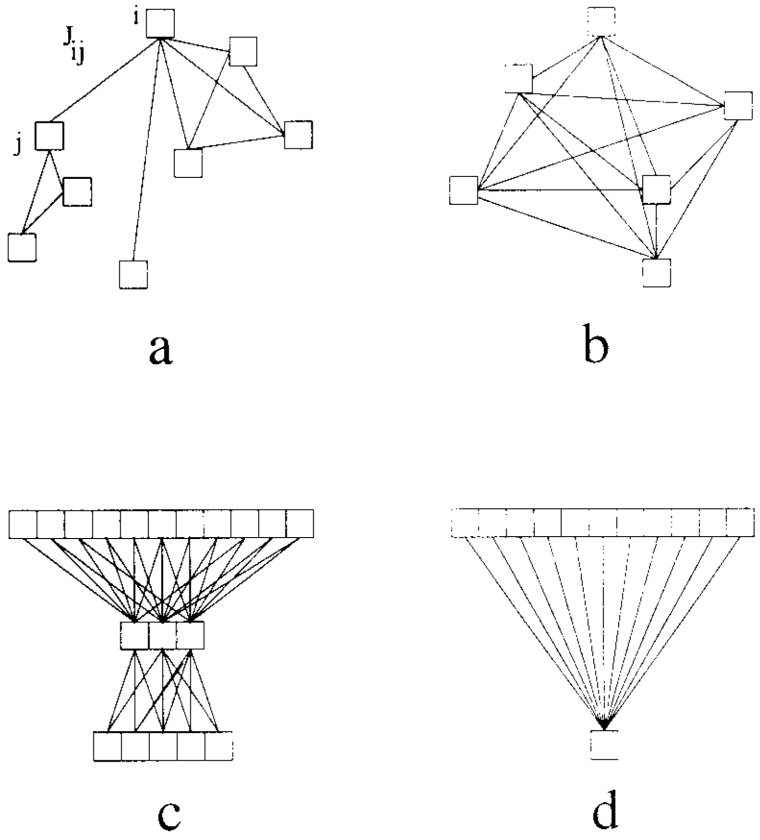

Figure 1: Different types of networks of formal neurons. a) general architecture, b) fully connected attractor neural network, c) feed-forward network with one hidden layer, d) single layer perceptron.

A mathematical analysis is, however, possible for some extreme architectures only. Two types of connectivities will be of special interest. In the first one every neuron is connected with every other neuron, see fig.1b. The dynamics is then highly recurrent and will in general result in a chaotic sequence of different activity patterns of the neurons. The other extreme type of architecture, called *feed-forward neural network*, is shown in fig.1c. In such a network, the neurons can be arranged in layers $l = 1, \ldots, L$ such that every neuron in layer $l$ only receives inputs from neurons of layer $(l-1)$ and in turn only feeds neurons in layer $(l+1)$. The first layer, $l = 1$, is called the ***input*** layer, the last one, $l = L$ the *output* layer, whereas all layers with $1 < l < L$ are referred to as *hidden* layers.

Due to the absence of feedback loops the dynamics is very simple: the input is mapped to the output via successive time steps. The network therefore performs a *classification* of the input strings into classes labeled by the different configurations of the output layer. This architecture is well suited for learning from examples. In particular the simplest feed-forward neural net, the perceptron, having no hidden layers at all, as shown in fig.1d can be analyzed in great detail.

## 1.2 A simple example

It is certainly appropriate to introduce learning from examples by discussing a simple example. Consider the perceptron shown in fig.2. It has $N = 20$ input units $S_i$ each connected directly to the single output $\sigma$ by real valued couplings $J_i$. For any input vector $\mathbf{S}$ the output is determined by the rule

$$\sigma = \mathrm{sgn}(\sum_i J_i S_i) \quad . \tag{1}$$

We would like to use the network to rank 10-digits dual numbers. To this end we require the output to be $+1$ $(-1)$ if the dual number represented by the left ten input bits is larger (smaller) than the one given by the right ten inputs [1]. For simplicity we ignore for the moment the possibility of the two numbers being equal.

With some thought, it is easy to construct a set of couplings that do the job perfectly. Consider the coupling values

$$\begin{aligned} J_i^{perf} &= 2^{10-i} && \text{if} \quad i = 1, \ldots, 10 \\ J_i^{perf} &= -J_{i-10}^{perf} && \text{if} \quad i = 11, \ldots, 20. \end{aligned} \tag{2}$$

displayed also in fig.3. This choice gives, as it should, a larger weight in

[1] A 3-digit dual number with dual code (-1,1,-1) is equal to $0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0$.

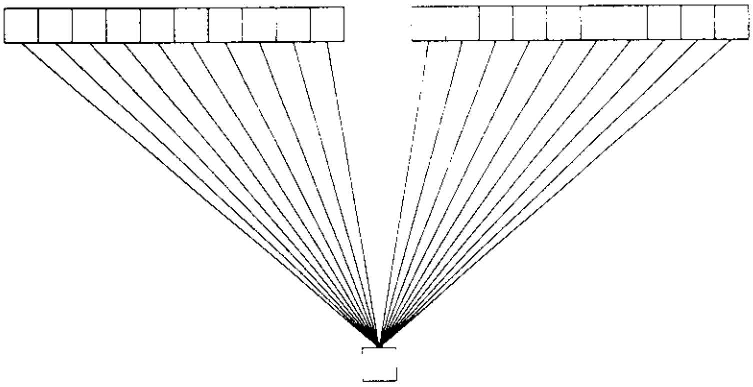

Figure 2: Simple perceptron used to rank dual numbers

the superposition (1) to the leftmost bits in the two subfields of the input. On the other hand it ensures that less significant bits are able to tip the balance if the first bits of the two numbers coincide. The above problem is simple enough to guess the appropriate values of the couplings. Doing so is an example of explicit programming as used in almost all present-day computers.

However, we can apply another, more interesting procedure to solve the problem, namely by learning from examples. Let us first initialize the couplings $J_i$ at random. We then select, out of the total of $2^{20} \simeq 10^6$ different input strings, a given number $p$ of input vectors $\boldsymbol{\xi}^\mu, \mu = 1, \ldots, p$ at random and for each case provide the *correct output* which we denote by $\sigma_T^\mu$. Next, we *train* the network with this set $\{\boldsymbol{\xi}^\mu, \sigma_T^\mu\}$. To this end we sequentially present each of the input vectors to the network and verify whether the resulting network output $\sigma^\mu$ given through (1) is correct, i.e. coincides with $\sigma_T^\mu$. If so, which will initially happen for roughly half of the cases, we simply proceed to the next example. If however $\sigma^\mu \neq \sigma_T^\mu$ we modify the couplings in such a way that the example under consideration is less likely to be misclassified upon the next presentation. Various rules that achieve this goal are available, in the simulations shown below we used the randomized perceptron learning rule. We iterate this procedure until all examples of the training set are reproduced correctly. The fact that the procedure converges is a priori not obvious, but it does so for the problem under consideration: ranking numbers is a *learnable* problem for the perceptron.

The success on the training set, however, does not tell us whether the

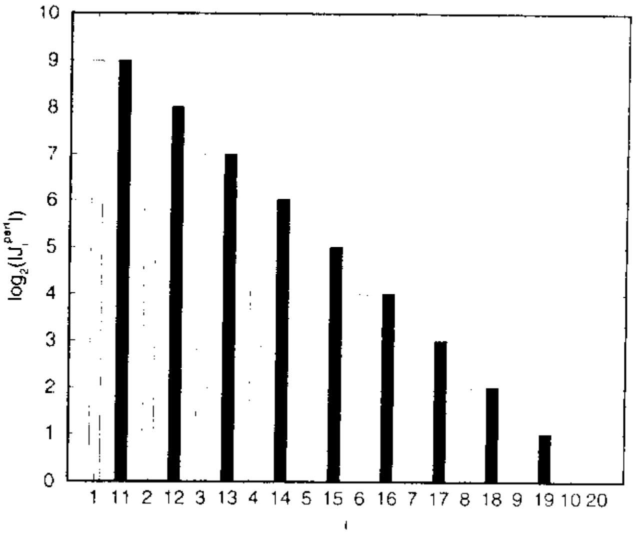

Figure 3: Graphical representation of the perfect couplings for a perceptron to rank dual numbers as given by (2). $J_1 \dots J_{10}$ (white) are positive, $J_{11} \dots J_{20}$ (black) are negative. cf. (2).

network has really learned the *rule* behind the the examples. To answer this question the performance on *so far unseen* inputs has to be investigated. A quantitative measure of the degree of generalization from the examples to the rule can be obtained by determining the fraction of wrong outputs when running through the *complete* set of $2^{20}$ different inputs. This fraction is called the generalization error $\varepsilon$ and is one of the central quantities in the analysis of learning problems.

Fig.4 shows $\varepsilon$ as a function of the size $p$ of the training set as resulting from simulations as described above. Note that $\varepsilon$ is a random variable which depends on the particular choice of the training set. In fig.4, we have reproduced the average over 1000 random realizations of the training set.

The general behaviour is as expected. For $p = 0$ the network has no information at all about the target rule. By chance half of the examples are classified correctly, $\varepsilon = .5$, which is the known success rate for pure guessing. With increasing $p$ the generalization error decreases monotonically and for $p \to \infty$ it must, of course, vanish. However, the surprising fact is that the

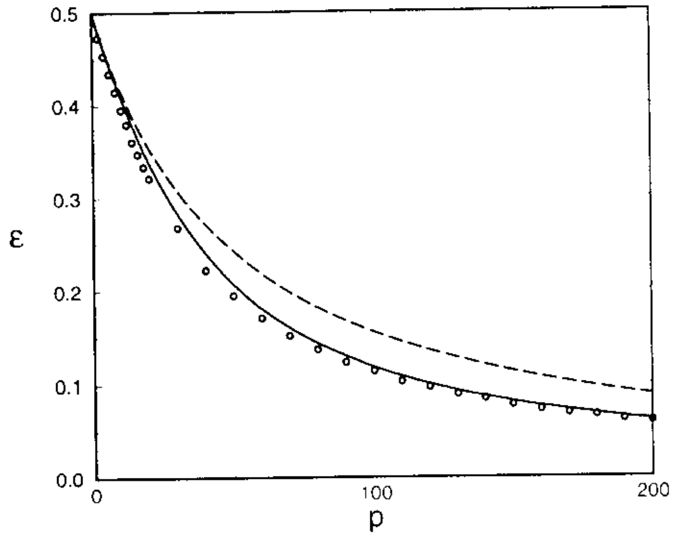

Figure 4: Simulation results (circles) for the generalization error of a perceptron learning from examples to rank dual numbers. The results are averaged over 1000 realizations of the training set. the statistical error is smaller than the symbol size. The full line gives the analytic result of the quenched calculation, the dashed line that of the annealed approximation.

generalization error becomes rather small already for $p$ of the order of a few hundred, which is *much less* than the total number of different input vectors! In other words, the network is able to generalize rather well in the sense that it can approximate the desired rule on the basis of a very limited set of examples.

In a similar way, one can show that a somewhat more complicated network made of Boolean gates is able to learn the addition of numbers from examples [1]. Another striking demonstration of learning from examples in artificial neural networks is the ability of a multi-layer neural net to read English text aloud [2] and many more examples have been documented [3].

At first sight it may seem somewhat enigmatic that a system as simple as the perceptron should be "intelligent enough" to decipher a rule behind examples. Nevertheless the explanation is rather simple: the perceptron can only implement a very limited set of mappings between input and output, and the ranking of numbers happens to be one of them. Given this limitation it is therefore comparatively easy to select the proper mapping on the basis of examples. These rather vague statements will be made more precise in the following lectures.

To get a more concrete idea of how the perceptron proceeds in the above problem, it is instructive to look at the evolution of the couplings $J_i$ as a function of the size $p$ of the training set. In fig.5 the couplings are shown for $p = 50$ and $p = 200$. In both cases we have normalized them such that $J_1 = 2^9$ in order to facilitate comparison with the target values given in (2) and fig.3. As one easily realizes, the relation between the most important couplings

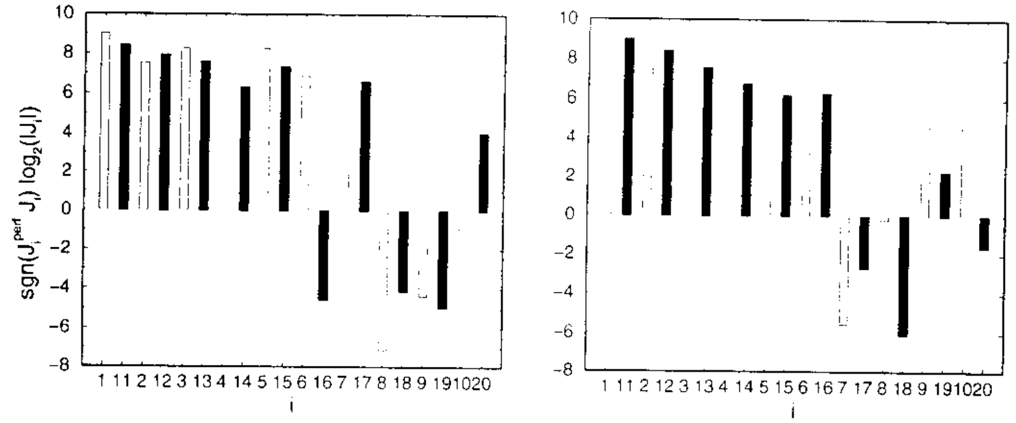

Figure 5: Graphical representation of the perceptron couplings after learning 50 (left) and 200 (right) examples respectively. The signs of the columns indicate whether the couplings have the same sign as the perfect couplings of fig.3 or not.

$J_1, J_2, J_3, J_{11}, J_{12}, J_{13}$ is fixed first. This is because they decide about the output for the large majority of input patterns, both in the training and in the complete set. Considering that correct values for $J_1, J_2, J_3, J_{11}, J_{12}$ and $J_{13}$ yield a correct output for 15/16 of all patterns already, one understands how the initial efficiency of the learning process is achieved. By the same token, one expects that inputs which give information about the couplings $J_9, J_{10}, J_{19}$ and $J_{20}$ are rare, with a rather slow asymptotic decay of the generalization error to zero as a result.

## 1.3 Annealed analysis of Gibbs learning

To begin with let us introduce a simple geometric interpretation of the classification sgn(**JS**) of an input **S** by a perceptron **J**. It is +1 or −1 depending on whether the angle between **S** and **J** is smaller or larger than $\pi/2$. Hence, the collection of points at which the classification switches from +1 to −1, also referred to as the *decision boundary*, is just the hyperplane orthogonal to **J** through the origin. For this reason, the classifications implementable by a perceptron are called *linearily separable*.

It is clear that the *length* of **J** and **S** have no impact on the classification. To avoid the possibility of different coupling vectors yielding identical classifications it has become customary to normalize both couplings and inputs according to

$$\mathbf{J}^2 = \sum_{i=1}^{N} J_i^2 = N \qquad \text{and} \qquad \mathbf{S}^2 = \sum_{i=1}^{N} S_i^2 = N \tag{3}$$

respectively. Hence both types of vectors lie on the surface of a $N$-dimensional sphere with radius $\sqrt{N}$ which in the following we will call the $N$-sphere. In order to compare the classifications performed by a teacher perceptron **T** and a student perceptron **J**, we project the input examples onto the plane spanned by the coupling vectors of teacher and student (see fig.6). One easily realizes that the projections lying in the shaded region originate from inputs that are classified differently by teacher and student. If the inputs are chosen at random, the probability of disagreement, which is precisely the generalization error $\varepsilon$, is just the probability for falling into this region. Since the decision lines are orthogonal to the vectors **T** and **J**, we conclude that $\varepsilon = \theta/\pi$, where $\theta$ is the angle between **T** and **J**. The generalization error is therefore proportional to the geodesic distance between the two points on the $N$-sphere that correspond to the teacher and the student perceptron. It is convenient to introduce the so-called *teacher-student overlap*:

$$R = \frac{\mathbf{JT}}{N} \quad . \tag{4}$$

Since we fixed the lengths of the vectors equal to $\sqrt{N}$, $R$ is nothing but the cosine of the angle $\theta$ between **J** and **T**, and the generalization error can be written as:

$$\varepsilon = \frac{1}{\pi} \arccos R \quad . \tag{5}$$

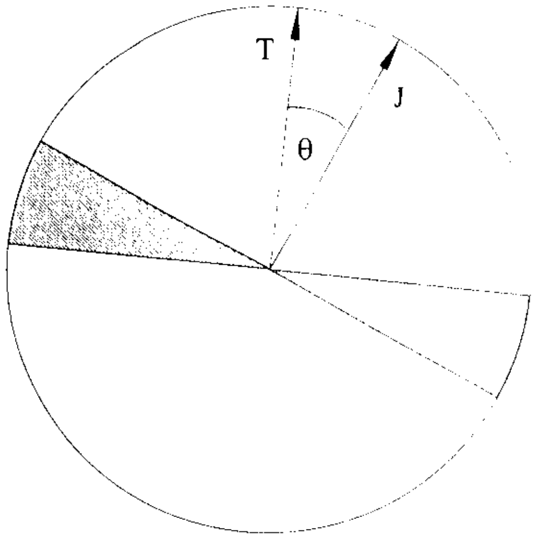

Figure 6: Projection of the input space to the plane spanned by the coupling vectors of teacher and student. Patterns with projection in the shaded region are classified wrongly by the student.

We now turn to a simple strategy for fixing the student couplings during learning. Consider the coupling vectors **J** that score on the examples exactly like the teacher. The set of these *compatible* students is called the *version space.* We ask for the generalization error of a vector **J** drawn *at random* from this version space. This simple prescription is called Gibbs learning. It is interesting since the results to be found for this learning rule characterize the *typical performance of a compatible student.*

Within the framework of Gibbs learning the generalization error decreases with increasing training set size because more and more couplings **J** are rejected as incompatible with the examples causing the version space to shrink in size. If we were able to quantify the "survival chance" of a coupling when presenting a new example, we could infer the average behaviour of the generalization error as the training proceeds. This is indeed possible if we group the couplings into classes with respect to their overlap with the teacher. For all couplings **J** with overlap $R$ defined in (4), the chance of producing the same output on a randomly chosen input as the teacher is $1 - (\arccos R)/\pi = 1 - \varepsilon$ by the very definition of the generalization error $\varepsilon$. Let $\Omega_0(\varepsilon)$ be the volume of coupling vectors **J** with overlap $R = \cos(\pi\varepsilon)$ before training has taken place. Since the examples are assumed to be independent, and each example will reduce this number on average by a factor $1-\varepsilon$, we conclude that the *average* volume $\Omega_p(\varepsilon)$ of compatible students with generalization error $\varepsilon$ after presentation of $p$ training examples is given by:

$$\Omega_p(\varepsilon) = \Omega_0(\varepsilon)(1-\varepsilon)^p \quad . \tag{6}$$

In the limit $N \to \infty$ a simple calculation gives to leading order in $N$:

$$\Omega_0(\varepsilon) = \int d\mathbf{J}\delta(\mathbf{J}^2 - N)\delta(\frac{\mathbf{JT}}{N} - \cos(\pi\varepsilon)) \sim \exp\{\frac{N}{2}[1 + \ln 2\pi + \ln\sin^2(\pi\varepsilon)]\} \tag{7}$$

and, from (6) and (7) we find using the scaling $p = \alpha N$ of the training set size

$$\Omega_p(\varepsilon) \sim \exp\{N\left[\frac{1}{2}(1+\ln 2\pi) + \frac{1}{2}\ln\sin^2(\pi\varepsilon) + \alpha\ln(1-\varepsilon)\right]\} \tag{8}$$

The expression in the square brackets is plotted in fig.7 as a function of $\varepsilon$ for several values of $\alpha$. Note that although it is a smooth function of $\varepsilon$, the corresponding differences in values of $\Omega_p(\varepsilon)$ are exponentially enhanced by the large prefactor $N$ in (8). We therefore conclude that by choosing a student vector *at random* from the version space we will, for large $N$, with overwhelming probability pick one with the value of $\varepsilon$ that *maximizes* the function shown in fig.7. All other values of $\varepsilon$ are realized by coupling vectors that are exponentially rare. We therefore expect that the generalization error is, in the large $N$ limit, given by:

$$\varepsilon(\alpha) = \text{argmax}[\frac{1}{2}\ln\sin^2(\pi\varepsilon) + \alpha\ln(1-\varepsilon)] \tag{9}$$

For $\alpha = 0$ the maximum of $\Omega_p$ occurs at $\varepsilon = .5$ corresponding to students "orthogonal" to the teacher. Clearly, without any further information, all the choices of $\mathbf{J}$ are equally probable, and those students which perform random guessing exponentially dominate in number over all the others. During the learning process, this effect is counterbalanced by a contribution incorporating information coming from the training set. In the present case of Gibbs learning this is the term $\alpha\ln(1-\varepsilon)$ in (8). For large training sets the generalization error becomes small and, as expected, we find $\varepsilon \to 0$ for $\alpha \to \infty$.

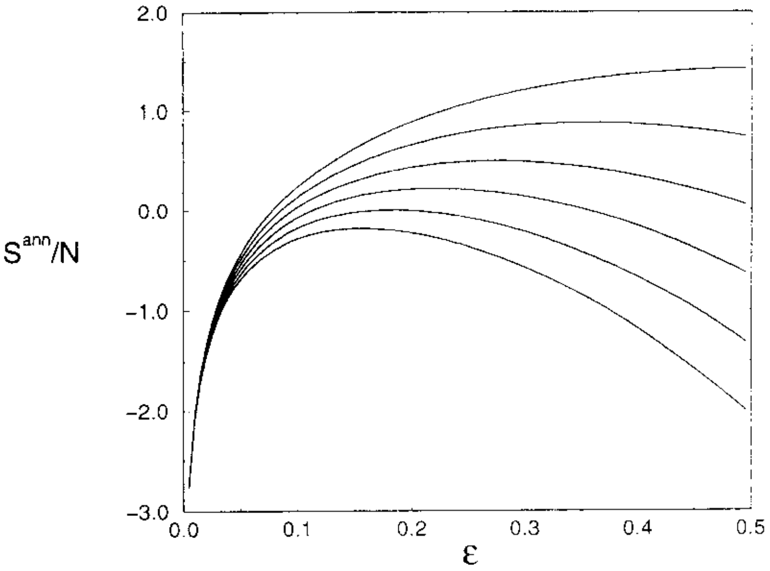

Figure 7: Expression in the square brackets of (8) as a function of $\varepsilon$ for $\alpha = 0, 1, 2, 3, 4$ and 5 (from top to bottom)

The above result also provides some qualitative insight about how artificial neural networks can learn from examples. Firstly, the specific architecture of the network implies a kind of a-priori hierarchy of implementable mappings. "Easy" input-output relations are realized by many different coupling vectors, "difficult" ones require rather detailed microscopic configurations. One could call this hierarchy the prejudices of the system. Secondly, classifications which are incompatible with the training set are eliminated. This type of learning hence implies a trade-off between accuracy and flexibility. A system with equal ability for the implementation of all mappings can in principle learn any problem, but its generalization characteristics will be equally poor for all of them. A very specialized system on the other hand will be able to learn a very restricted class of problems only, but it will do so rather fast. In fact spectacular "Eureca"-like sudden transitions to perfect generalization can occur in such systems (see lecture 5).

In order to see how accurate our simple analysis of Gibbs learning is, we have included the result (9) as the dashed line into fig.4. From the comparison with the simulation results we clearly see that although the qualitative behaviour is described correctly there are significant deviations. In particular for large $\alpha$ we find from (9)

$$\varepsilon \sim \frac{1}{\alpha} \quad . \tag{10}$$

which although giving the correct scaling is quantitatively poor due to the wrong prefactor.

So what went wrong with our analysis? The main flaw is that we did not treat the randomness in the learning examples $\boldsymbol{\xi}^\mu$ and the teacher couplings $\mathbf{T}$ correctly. Let us denote by $\Omega(\varepsilon;\boldsymbol{\xi}^\mu,\mathbf{T})$ the volume of student coupling vectors making an angle $\pi\varepsilon$ with the teacher vector $\mathbf{T}$ which remain in the version space after learning the $p$ examples $\boldsymbol{\xi}^\mu$. Due to the random choice of both $\mathbf{T}$ and the $\boldsymbol{\xi}^\mu$ also $\Omega(\varepsilon;\boldsymbol{\xi}^\mu,\mathbf{T})$ is a random quantity. However, $\Omega_p$ as defined in (6) just describes its *average*. In many situations dealing with random systems the average also gives a reasonable estimate for the *typical*, i.e. most probable value of a random quantity. This is, however, not always the case and it is in particular *not true* for $\Omega(\varepsilon;\boldsymbol{\xi}^\mu,\mathbf{T})$. As a result one finds for large $N$ with overwhelming probability a value of $\Omega(\varepsilon;\boldsymbol{\xi}^\mu,\mathbf{T})$ that is different from the average described by (8). In fact one can show that (9) always gives an *upper bound* to the true $\varepsilon$ in accordance with fig.4.

In order to demonstrate that the distribution of $\Omega(\varepsilon;\boldsymbol{\xi}^\mu,\mathbf{T})$ is indeed badly characterized by its average for large $N$ and, more importantly, to find a correct way to describe the *typical* behaviour of $\varepsilon(\alpha)$, it is useful to first reconsider the above approach within the framework of statistical mechanics. Building on this formulation we will then develop the central statistical mechanics techniques for the quantitative analysis of learning problems.

## 1.4 The annealed approximation in statistical mechanics

Let us rephrase the above learning scenario from the point of view of statistical mechanics. The microscopic variables are the components $J_i$ of the student vector. They span the so called *phase space*. The generalization error $\varepsilon$ is the relevant macrovariable. A given value of $\varepsilon$ can be achieved by many different choices of the microvariables $\mathbf{J}$. The important quantity $\Omega(\varepsilon;\boldsymbol{\xi}^\mu,\mathbf{T})$ denotes the volume in phase space occupied by all the microscopic states which for given $\boldsymbol{\xi}^\mu$ and $\mathbf{T}$ realize the macroscopic state specified by $\varepsilon$. In statistical mechanics it is quantified by the *entropy* $S$ which is nothing but the logarithm of this volume:

$$S(\varepsilon;\boldsymbol{\xi}^{\mu},\mathbf{T}) = \ln\Omega(\varepsilon;\boldsymbol{\xi}^{\mu},\mathbf{T}) \tag{11}$$

From (7) we hence find for the entropy before learning for large $N$:

$$S_0(\varepsilon) \sim \frac{N}{2}[1+\ln 2\pi + \ln\sin^2(\pi\varepsilon)] \quad . \tag{12}$$

The entropy has the appealing property of being *extensive*, i.e. it is proportional to the number of degrees of freedom $N$.

The first two terms in (12) are independent of $\varepsilon$ and just correspond to the surface of the $N$-sphere. It is convenient to normalize $\Omega(\varepsilon;\boldsymbol{\xi}^{\mu},\mathbf{T})$ with respect to this value. To this end we will use from now on the integration measure

$$d\mu(\mathbf{J}) := \frac{d\mathbf{J}\ \delta(\mathbf{J}^2-N)}{\int d\mathbf{J}\ \delta(\mathbf{J}^2-N)} \tag{13}$$

ensuring $\int d\mu(\mathbf{J}) = 1$. The renormalized entropy is then simply given by

$$S_0(\varepsilon) \sim \frac{N}{2}[\ln\sin^2(\pi\varepsilon)] \quad . \tag{14}$$

As learning proceeds, more and more couplings $\mathbf{J}$ are rejected because they are incompatible with the training examples $\boldsymbol{\xi}^{\mu}$. The function

$$\chi(\mathbf{J}) = \prod_{\mu=1}^{p}\theta((\frac{1}{\sqrt{N}}\mathbf{T}\boldsymbol{\xi}^{\mu})(\frac{1}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^{\mu})) \tag{15}$$

is one if $\mathbf{J}$ classifies *all* examples exactly like the teacher and zero otherwise. It may hence serve as an *indicator function* of the remaining couplings forming the version space.

The arguments of the $\theta$-function have been normalized such that they remain $O(1)$ for $N\to\infty$. Although this is not absolutely necessary since the Heaviside function $\theta(x)$ depends only on whether its argument is bigger or smaller than 0 it is quite helpful for keeping track of the order of different terms in the thermodynamic limit. It also facilitates the comparison with expressions for smooth, i.e. gradual neuron characteristics in which the scaling of the argument is really important.

Using the indicator function $\chi(\mathbf{J})$ the normalized volume of the whole version space after learning $p$ examples can be written as

$$\Omega_p(\boldsymbol{\xi}^\mu, \mathbf{T}) = \int d\mu(\mathbf{J}) \prod_{\mu=1}^{p} \theta((\frac{1}{\sqrt{N}}\mathbf{T}\boldsymbol{\xi}^\mu)(\frac{1}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^\mu)) \quad . \tag{16}$$

It depends on the teacher vector $\mathbf{T}$ and the particular examples $\boldsymbol{\xi}^\mu$ forming the training set all being random variables. Let us calculate the *average* $\Omega_p = \langle\langle \Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle$ of the version space volume. Here and in the following the double angle $\langle\langle \ldots \rangle\rangle$ denotes the average over the training examples and the teacher couplings. The logarithm of $\Omega_p$ is called the *annealed* entropy

$$S_p^{ann} = \ln\langle\langle \Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle \tag{17}$$

To explicitly perform the average we will use the distribution

$$P_{\mathbf{S}}(\mathbf{S}) = \prod_i \left[\frac{1}{2}\delta(S_i + 1) + \frac{1}{2}\delta(S_i - 1)\right] \tag{18}$$

for the inputs. For the teacher we assume that its coupling vector $\mathbf{T}$ is choosen with constant probability from the $N$-sphere, i.e.

$$P_{\mathbf{T}}(\mathbf{T}) = (2\pi e)^{-N/2}\delta(\mathbf{T}^2 - N) \quad . \tag{19}$$

As we will see shortly, however, the average over the teacher is trivial since, after the $\boldsymbol{\xi}^\mu$-average has been done, all choices of $\mathbf{T}$ give the same result for the generalization error.

The random variables appear in (16) in a multiplicative manner and inside a function. One can extract them by the use of delta functions. Introducing the auxiliary variables

$$\lambda_\mu = \frac{1}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^\mu \quad \text{and} \quad u_\mu = \frac{1}{\sqrt{N}}\mathbf{T}\boldsymbol{\xi}^\mu \tag{20}$$

one can rewrite (16) as follows

$$\Omega_p = \int d\mu(\mathbf{J}) \int \prod_\mu d\lambda_\mu du_\mu \prod_\mu \theta(\lambda_\mu u_\mu) \\ \prod_\mu \langle\langle \delta(\lambda_\mu - \frac{1}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^\mu)\delta(u_\mu - \frac{1}{\sqrt{N}}\mathbf{T}\boldsymbol{\xi}^\mu)\rangle\rangle \tag{21}$$

At this point, one could formally proceed to represent the delta functions by their Fourier representation, to achieve a factorization over the components of the examples $\boldsymbol{\xi}^\mu$. We will however use a shortcut. For fixed value of $\mathbf{J}$ and $\mathbf{T}$, both $\lambda_\mu$ and $u_\mu$ are sums of $N$ independent terms and by the central limit theorem they obey a Gaussian distribution. With the help of (3), (4) and (18) one finds:

$$\begin{aligned} \langle\langle \lambda_\mu \rangle\rangle &= \langle\langle u_\mu \rangle\rangle = 0 \\ \langle\langle \lambda_\mu^2 \rangle\rangle &= \langle\langle u_\mu^2 \rangle\rangle = 1 \\ \langle\langle \lambda_\mu u_\mu \rangle\rangle &= \frac{\mathbf{JT}}{N} = R \end{aligned} \tag{22}$$

To evaluate (21) we use the bi-Gaussian probability distribution determined by the moments (22) and find :

$$\Omega_p = \int d\mu(\mathbf{J}) \\ \int \prod_\mu \frac{d\lambda_\mu du_\mu}{2\pi\sqrt{1-R^2}} \prod_\mu \theta(\lambda_\mu u_\mu) \exp\{-\frac{1}{2(1-R^2)} \sum_\mu (\lambda_\mu^2 + u_\mu^2 - 2R\lambda_\mu u_\mu)\} \tag{23}$$

where $R$ is just a shorthand notation for $\mathbf{JT}/N$. To proceed, one notes that the $\mathbf{J}$- and $\lambda_\mu$-$u_\mu$-integrals are entangled through the combination $R\lambda_\mu u_\mu$. They can be decoupled by introducing an additional delta function $\delta(\frac{\mathbf{JT}}{N} - R)$ which effectively amounts to perform the $\mathbf{J}$-integral in slices of constant $R$, just as we did in the previous section.

In this way we get

$$\Omega_p = \int_{-1}^{1} dR \int d\mu(\mathbf{J}) \delta(\frac{\mathbf{JT}}{N} - R) \\ \int \prod_\mu \frac{d\lambda_\mu du_\mu}{2\pi\sqrt{1-R^2}} \prod_\mu \theta(\lambda_\mu u_\mu) \exp\{-\frac{1}{2(1-R^2)} \sum_\mu (\lambda_\mu^2 + u_\mu^2 - 2R\lambda_\mu u_\mu)\} \tag{24}$$

Now the $\lambda_\mu$-$u_\mu$-integrals factorize in $\mu$ and using

$$\frac{2}{\sqrt{1-R^2}} \int_0^\infty \frac{d\lambda}{\sqrt{2\pi}} \int_0^\infty \frac{du}{\sqrt{2\pi}} \exp\{-\frac{1}{2(1-R^2)}(\lambda^2 + u^2 - 2R\lambda u)\} = 1 - \frac{1}{\pi} \arccos R \tag{25}$$

we find with (7) and (5)

$$\Omega_p = \int_{-1}^{1} dR \exp\{N\left[\frac{1}{2}\ln(1-R^2)+\alpha\ln(1-\frac{1}{\pi}\arccos R)\right]\} \qquad (26)$$

Similarly to (8) this expression shows that the restriction of the coupling vector to the version space introduces an extra term counterbalancing the purely entropic part (14). Due to our scaling $p = \alpha N$ of the training set size this new contribution is extensive, too. It is called the *energetic* part for reasons that will be elaborated on more thoroughly in lecture 3.

The asymptotic behaviour of the remaining $R$-integral in (26) in the thermodynamic limit $N \to \infty$ can now be determined by the saddle-point method. This yields finally

$$S_p^{ann} = N \max_R \left[\frac{1}{2}\ln(1-R^2)+\alpha\ln(1-\frac{1}{\pi}\arccos R)\right] \quad . \qquad (27)$$

This result implies in particular that the average phase space volume $\Omega_p$ is dominated by couplings with overlap

$$R = \text{argmax}\left[\frac{1}{2}\ln(1-R^2)+\alpha\ln(1-\frac{1}{\pi}\arccos R)\right] \qquad (28)$$

(26) and (28) reproduce (8) and (9) respectively. This is reasonable since we have in both cases analyzed the *average* of the phase volume $\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})$. From (16), however, we infer that $\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})$ involves a *product* of many random contributions. Products of independent random numbers are known to possess distributions with long tails for which the average and the most probable value are markedly different. On the other hand, the logarithm of such a quantity is a large *sum* of independent terms and hence becomes Gaussian distributed so that its average and most probable value asymptotically coincide. Similarly, the *typical* value of $\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})$ is for large $N$ given by [2]

$$\Omega^{typ}(\boldsymbol{\xi}^\mu, \mathbf{T}) \sim \exp\{\langle\langle \ln\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle\} \quad . \qquad (29)$$

This is in accordance with the general dogma of the statistical mechanics of disordered systems that *extensive* quantities such as entropy and energy are *self-averaging*, see [7, 8]. A reliable analysis of the typical generalization behaviour therefore requires the calculation of $\langle\langle \ln\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle$ rather than $\langle\langle \Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle$. As we will see in the next lecture this indeed yields the correct result.

[2] This holds true although the different terms in $\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})$ are weakly correlated.

# 2 Lecture 2: The Gardner analysis

As became clear at the end of the last lecture the annealed entropy $S_{ann} = \ln\langle\langle\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle$ cannot be used to describe the *typical* generalization behaviour. A correct theory must instead be built on the so-called *quenched* entropy defined by

$$S = \langle\langle\ln\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle. \tag{30}$$

This involves the calculation of the average

$$\langle\langle\ln\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle = \langle\langle\ln\int d\mu(\mathbf{J})\prod_{\mu=1}^{p}\theta((\frac{1}{\sqrt{N}}\mathbf{T}\boldsymbol{\xi}^\mu)(\frac{1}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^\mu))\rangle\rangle \tag{31}$$

over the random examples of the training set and over the random teacher vector. This quenched average is technically much less straightforward than the annealed one (21). The main problem is that the integral over $\mathbf{J}$ cannot be performed analytically for every individual realization of the examples. Only *after* the average over the examples the system is "translationally invariant" in the sense that all neurons $i$ are equivalent to each other so that the integral over $\mathbf{J}$ can be reduced to an integral over a single component $J$ of $\mathbf{J}$. We hence have to interchange the average and the logarithm in (31) in some way. It is far from obvious how to accomplish this in a controlled fashion. Fortunately, quenched averages have been a central problem in the theory of disordered solids since the early seventies and we can build on the techniques developed in this field. A way which–although by no means mathematically rigorous–has turned out to be successful in many situations is the so-called *replica trick* [9], relying on the simple identity

$$\ln x = \lim_{n\to 0}\frac{d}{dn}x^n = \lim_{n\to 0}\frac{x^n - 1}{n}. \tag{32}$$

Used for our problem it yields

$$\langle\langle\ln\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle = \lim_{n\to 0}\frac{\langle\langle\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})^n\rangle\rangle - 1}{n} \tag{33}$$

and the calculation of $\langle\langle\ln\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle$ has been transformed to that of $\langle\langle\Omega^n(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle$. For general real $n$ this is, of course, as complicated as the original average.

However, for *natural* numbers $n = 1, 2, 3, \ldots$ the average can be rewritten as

$$\langle\langle \Omega^n(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle = \langle\langle \left[ \int d\mu(\mathbf{J}) \prod_{\mu=1}^{p} \theta((\frac{1}{\sqrt{N}}\mathbf{T}\boldsymbol{\xi}^\mu)(\frac{1}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^\mu)) \right]^n \rangle\rangle \tag{34}$$

$$= \langle\langle \int \prod_{a=1}^{n} d\mu(\mathbf{J}^a) \prod_{\mu=1}^{p}\prod_{a=1}^{n} \theta((\frac{1}{\sqrt{N}}\mathbf{T}\boldsymbol{\xi}^\mu)(\frac{1}{\sqrt{N}}\mathbf{J}^a\boldsymbol{\xi}^\mu))\rangle\rangle$$

which looks like the combined average of $n$ different copies (*replicas*) of the original system with the *same* realization of the random examples. As we will see shortly, this average is only slightly more complicated than the annealed one calculated in the last section. The main question which remains is how to find the proper analytic continuation which allows to perform the limit $n \to 0$ from the result for natural $n$. This can be a very hard problem [3], fortunately for most of the systems considered in this book, a comparatively straightforward procedure is successful.

Let us now perform the detailed calculation of the quenched entropy

$$S = \langle\langle \ln \Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle_{\boldsymbol{\xi}^\mu, \mathbf{T}} \tag{35}$$

for a spherical perceptron $\mathbf{J}$ with $N$ inputs trained by random examples $\boldsymbol{\xi}^\mu, \mu = 1 \ldots p = \alpha N$ classified by a random spherical teacher perceptron with coupling vector $\mathbf{T}$. We will consider the probability distributions (18) for the components $\xi_i^\mu$ of the examples and (19) for the teacher's coupling vector $\mathbf{T}$.

Using the replica trick we relate $\langle\langle \ln \Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle_{\boldsymbol{\xi}^\mu, \mathbf{T}}$ to $\langle\langle \Omega^n(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle_{\boldsymbol{\xi}^\mu, \mathbf{T}}$ via

$$\langle\langle \ln \Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle = \lim_{n\to 0} \frac{\langle\langle \Omega^n(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle_{\boldsymbol{\xi}^\mu, \mathbf{T}} - 1}{n} \quad . \tag{36}$$

so the crucial quantity to calculate is

$$\Omega^{(n)} := \langle\langle \Omega^n(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle_{\boldsymbol{\xi}^\mu, \mathbf{T}} = \langle\langle \int d\mu(\mathbf{J}^a) \prod_{a,\mu} \theta(\frac{\mathbf{T}\boldsymbol{\xi}^\mu}{\sqrt{N}} \frac{\mathbf{J}^a\boldsymbol{\xi}^\mu}{\sqrt{N}})\rangle\rangle_{\boldsymbol{\xi}^\mu, \mathbf{T}} \tag{37}$$

where

$$d\mu(\mathbf{J}) := \prod_{i=1}^{N} \frac{dJ_i}{\sqrt{2\pi e}} \delta(\sum_{i=1}^{N} J_i^2 - N) \tag{38}$$

[3] An authoritative reference for the replica method is [8].

denotes the spherical measure defined in (13).

To calculate $\Omega^{(n)}$ we first introduce the variables

$$\lambda_\mu^a = \frac{\mathbf{J}^a\boldsymbol{\xi}^\mu}{\sqrt{N}} \qquad \text{and} \qquad u_\mu = \frac{\mathbf{T}\boldsymbol{\xi}^\mu}{\sqrt{N}} \tag{39}$$

by $\delta$-functions to obtain

$$\Omega^{(n)} = \int d\mu(\mathbf{J}^a) \int \prod_{a,\mu} d\lambda_\mu^a \int \prod_\mu du_\mu \prod_{a,\mu} \theta(u_\mu \lambda_\mu^a) \tag{40}$$
$$\langle\langle \delta(\lambda_\mu^a - \frac{\mathbf{J}^a\boldsymbol{\xi}^\mu)}{\sqrt{N}})\delta(u_\mu - \frac{\mathbf{T}\boldsymbol{\xi}^\mu}{\sqrt{N}}\rangle\rangle_{\boldsymbol{\xi}^\mu,\mathbf{T}}$$

and use their integral representation. We then find

$$\Omega^{(n)} = \int \prod_a d\mu(\mathbf{J}^a) \int \prod_{a,\mu} \frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi} \int \prod_\mu \frac{du_\mu d\hat{u}_\mu}{2\pi} \tag{41}$$
$$\prod_{a,\mu} \theta(u_\mu\lambda_\mu^a) \exp\{i \sum_{\mu,a} \lambda_\mu^a \hat{\lambda}_\mu^a + i \sum_\mu u_\mu \hat{u}_\mu\}$$
$$\langle\langle \exp\{-\frac{i}{\sqrt{N}} \sum_{a,\mu} \hat{\lambda}_\mu^a \mathbf{J}^a \boldsymbol{\xi}^\mu - \frac{i}{\sqrt{N}} \sum_\mu \hat{u}_\mu \mathbf{T}\boldsymbol{\xi}^\mu\}\rangle\rangle_{\boldsymbol{\xi}^\mu,\mathbf{T}}$$

With the help of (18) we find for the average in this expression because of the statistical independence of the components $\xi_i^\mu$:

$$\langle\langle \prod_{i,\mu} \langle\langle \exp\{-\frac{i}{\sqrt{N}}(\sum_a \hat{\lambda}_\mu^a J_i^a + \hat{u}_\mu T_i)\xi_i^\mu\}\rangle\rangle_{\xi^\mu}\rangle\rangle_{\mathbf{T}} \tag{42}$$
$$= \langle\langle \prod_{i,\mu} \cos(\frac{1}{\sqrt{N}} \sum_a (\hat{\lambda}_\mu^a J_i^a + \hat{u}_\mu T_i))\rangle\rangle_{\mathbf{T}}$$
$$= \langle\langle \exp\{\sum_{i,\mu} \ln\cos(\frac{1}{\sqrt{N}} \sum_a (\hat{\lambda}_\mu^a J_i^a + \hat{u}_\mu T_i))\}\rangle\rangle_{\mathbf{T}} \quad .$$

The asymptotic behaviour of this expression for large $N$ is given by [4]

$$\langle\langle \exp\{\sum_{i,\mu} \ln(1 - \frac{1}{2N}(\sum_{a} \hat{\lambda}_{\mu}{}^{a} J_i^a + \hat{u}_\mu T_i)^2)\}\rangle\rangle_{\mathbf{T}} \tag{43}$$

$$=\langle\langle \exp\{-\frac{1}{2}\sum_{\mu}\sum_{a,b} \hat{\lambda}_{\mu}{}^{a}\hat{\lambda}_{\mu}{}^{b}\frac{1}{N}\sum_{i} J_i^a J_i^b - \sum_{\mu}\sum_{a}\hat{\lambda}_{\mu}{}^{a} u_\mu \frac{1}{N}\sum_{i} J_i^a T_i$$
$$-\frac{1}{2}\sum_{\mu}\hat{u}_\mu^2 \frac{1}{N}\sum_{i} T_i^2\}\rangle\rangle_{\mathbf{T}}$$

Obviously the dominant terms for large $N$ are the same for any distribution of examples with the first two moments

$$\langle\langle \xi_i^\mu \rangle\rangle = 0 \qquad \text{and} \qquad \langle\langle \xi_i^\mu \xi_j^\nu \rangle\rangle = \delta^{\mu\nu}\delta_{ij} \quad . \tag{44}$$

Using $\mathbf{J}^2 = \mathbf{T}^2 = N$ we obtain after inserting (43) back into (41)

$$\Omega^{(n)} = \int \prod_{a} d\mu(\mathbf{J}^a) \int \prod_{a,\mu} \frac{d\lambda_\mu^a d\hat{\lambda}_{\mu}{}^{a}}{2\pi} \int \prod_{\mu} \frac{du_\mu d\hat{u}_\mu}{2\pi} \prod_{a,\mu} \theta(u_\mu \lambda_\mu^a) \tag{45}$$
$$\langle\langle \exp\{i\sum_{\mu,a} \lambda_\mu^a \hat{\lambda}_{\mu}{}^{a} + i\sum_{\mu} u_\mu \hat{u}_\mu - \frac{1}{2}\sum_{\mu,a}(\lambda_\mu^a)^2$$
$$-\frac{1}{2}\sum_{\mu}\sum_{(a,b)} \hat{\lambda}_{\mu}{}^{a}\hat{\lambda}_{\mu}{}^{b}\frac{1}{N}\sum_{i} J_i^a J_i^b - \frac{1}{2}\sum_{\mu}\hat{u}_\mu^2 - \sum_{\mu,a}\hat{\lambda}_{\mu}{}^{a}\hat{u}_\mu \frac{1}{N}\sum_{i} J_i^a T_i\}\rangle\rangle_{\mathbf{T}}$$

Here $\sum_{(a,b)}$ denotes the sum over all terms with $a \neq b$. To make further progress we introduce the auxiliary variables

$$q^{ab} = \frac{1}{N}\sum_{i} J_i^a J_i^b \qquad \text{and} \qquad R^a = \frac{1}{N}\sum_{i} T_i J_i^a \tag{46}$$

[4] Note that this expansion cannot be justified for all values of the integration variables $\hat{\lambda}_{\mu}{}^{a}$, $\hat{u}_\mu$, $J_i^a$ and $T_i$. The contributions to $\Omega^{(n)}$ from regions with $(\hat{\lambda}_{\mu}{}^{a} J_i^a + \hat{u}_\mu T_i) = 0(\sqrt{N})$ are, however, negligible for $N \to \infty$.

to decouple the $J$-from the $\lambda$-$u$-integrals. We then find

$$\Omega^{(n)} = \int \prod_{a<b} N dq^{ab} \int \prod_{a} N dR^a \tag{47}$$
$$\int \prod_{a} d\mu(\mathbf{J}^a) \langle\langle \prod_{a} \delta(\mathbf{J}^a \mathbf{T} - N R^a) \rangle\rangle_{\mathbf{T}} \prod_{a<b} \delta(\mathbf{J}^a \mathbf{J}^b - N q^{ab})$$
$$\int \prod_{a,\mu} \frac{d\lambda_\mu^a d\hat{\lambda}_\mu^a}{2\pi} \int \prod_{\mu} \frac{du_\mu d\hat{u}_\mu}{2\pi} \prod_{a,\mu} \theta(u_\mu \lambda_\mu^a) \exp\{ i \sum_{\mu,a} \lambda_\mu^a \hat{\lambda}_\mu^a + i \sum_{\mu} u_\mu \hat{u}_\mu$$
$$- \frac{1}{2} \sum_{\mu,a} (\hat{\lambda}_\mu^a)^2 - \frac{1}{2} \sum_{\mu} \sum_{a,b} \hat{\lambda}_\mu^a \hat{\lambda}_\mu^b q^{ab} - \sum_{\mu,a} \hat{\lambda}_\mu^a \hat{u}_\mu R^a - \frac{1}{2} \sum_{\mu} \hat{u}_\mu^2 \}$$

The remaining average over $\mathbf{T}$ ist trivial since the integral over $\mathbf{J}^a$ gives the same result for almost all choices of $\mathbf{T}$. Having averaged over the examples the teacher average is hence redundant. This is no surprise since for an isotropic distribution of examples no direction of $\mathbf{T}$ is in any sense special.

We finally introduce integral representations for the remaining $\delta$-functions in (47) including those contained in the integration measures $d\mu(\mathbf{J}^a)$ and perform the Gaussian $\hat{u}_\mu$-integrals. In this way we end up with

$$\Omega^{(n)} = \int \prod_{a} \frac{d\hat{k}^a}{4\pi} \int \prod_{a<b} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi/N} \int \prod_{a} \frac{dR^a d\hat{R}^a}{2\pi/N} \tag{48}$$
$$\exp\{ i \frac{N}{2} \sum_{a} \hat{k}^a + iN \sum_{a<b} q^{ab} \hat{q}^{ab} - iN \sum_{a} R^a \hat{R}^a \}$$
$$\int \prod_{i,a} \frac{dJ_i^a}{\sqrt{2\pi e}} \exp\{ -\frac{i}{2} \sum_{a} \hat{k}^a \sum_{i} (J_i^a)^2 - i \sum_{a<b} \hat{q}^{ab} \sum_{i} J_i^a J_i^b - i \sum_{a} \hat{R}_a \sum_{i} J_i^a \}$$
$$\int \prod_{\mu} \frac{du_\mu}{\sqrt{2\pi}} \int \prod_{\mu,a} d\lambda_\mu^a \int \prod_{\mu,a} \frac{d\hat{\lambda}_\mu^a}{2\pi} \prod_{a,\mu} \theta(u_\mu \lambda_\mu^a) \exp\{ -\frac{1}{2} \sum_{\mu} u_\mu^2 - \frac{1}{2} \sum_{a} (1 - (R^a)^2) \sum_{\mu} (\hat{\lambda}_\mu^a)^2$$
$$- \frac{1}{2} \sum_{\mu} \sum_{(a,b)} \hat{\lambda}_\mu^a \hat{\lambda}_\mu^b (q^{ab} - R^a R^b) + i \sum_{\mu,a} \lambda_\mu^a \hat{\lambda}_\mu^a - i \sum_{\mu} u_\mu \sum_{a} \hat{\lambda}_\mu^a R^a \}$$

Now the $J_i^a$-integrals factorize in $i$, i.e. they give rise to a single $J^a$-integral to the power $N$. Similarily the $u_\mu$-$\lambda_\mu^a$-$\hat{\lambda}_\mu^a$-integrals factorizes in $\mu$ and can hence be reduced to a single $u$-$\lambda^a$-$\hat{\lambda}^a$-integral to the power $p = \alpha N$. Altogether this yields:

$$\Omega^{(n)} = \int \prod_a \frac{d\hat{k}^a}{4\pi} \int \prod_{a<b} \frac{dq^{ab} d\hat{q}^{ab}}{2\pi/N} \int \prod_a \frac{dR^a d\hat{R}^a}{2\pi/N}$$

$$\exp\left\{ N \left[ \frac{i}{2} \sum_a \hat{k}^a + i \sum_{a<b} q^{ab}\hat{q}^{ab} + i \sum_a R^a \hat{R}^a + G_S(\hat{k}^a, \hat{q}^{ab}, \hat{R}^a) + \alpha G_E(q^{ab}, R^a) \right] \right\} \tag{49}$$

with

$$G_S(\hat{k}^a, \hat{q}^{ab}, \hat{R}^a) = \ln \int \prod_a \frac{dJ^a}{2\pi\epsilon} \exp\{ -\frac{i}{2} \sum_a \hat{k}^a (J^a)^2 - i \sum_{a<b} \hat{q}^{ab} J^a J^b - i \sum_a \hat{R}^a J^a \} \tag{50}$$

and

$$G_E(q^{ab}, R^a) = \ln \int \frac{du}{\sqrt{2\pi}} \int \prod_a d\lambda^a \int \prod_a \frac{d\hat{\lambda}^a}{2\pi} \prod_a \theta(u\lambda^a) \exp\{ -\frac{u^2}{2} - \frac{1}{2} \sum_a (1-(R^a)^2)(\hat{\lambda}^a)^2$$
$$- \frac{1}{2} \sum_{(a,b)} \hat{\lambda}^a \hat{\lambda}^b (q^{ab} - R^a R^b) + i \sum_a \lambda^a \hat{\lambda}^a - iu \sum_a \hat{\lambda}^a R^a \} \tag{51}$$

$G_S$ is called the *entropic part* since it just measures how many spherical coupling vectors **J** fulfill the constraints (46). On the other hand $G_E$ is refered to as the *energetic part* since it is specific to the cost function or learning rule respectively which is being used.

We now proceed to evaluate the asymptotic behaviour of the integrals over the auxiliary parameters $q^{ab}, \hat{q}^{ab}, R^a, \hat{R}^a$ and $\hat{k}^a$ for $N \to \infty$ by the saddle-point method.

The extremum with respect to $\hat{k}^a, \hat{q}^{ab}$ and $\hat{R}^a$ can be found in closed form since the $J^a$-integrals are Gaussian. Introducing the $n \times n$ matrices $A$ and $B$ by

$$A_{ab} = i\hat{k}^a \delta_{ab} + i\hat{q}^{ab}(1-\delta_{ab}) \qquad \text{and} \qquad B_{ab} = \delta_{ab} + q^{ab}(1-\delta_{ab}). \tag{52}$$

the $J^a$-integral in the entropic part can be performed to yield [5]

$$G_S(\hat{k}^a, \hat{q}^{ab}, \hat{R}^a) = -\frac{n}{2} - \frac{1}{2}\ln(\det A) - \frac{1}{2}\sum_{a,b}\hat{R}^a (A^{-1})_{ab}\hat{R}^b \quad . \tag{53}$$

Using $\ln\det A = \mathrm{Tr}\ln A$ the part of the exponent in (49) which depends on $\hat{k}^a, \hat{q}^{ab}$ and $\hat{R}^a$ may be written as

$$-\frac{n}{2} - \frac{1}{2}\mathrm{Tr}\ln A - \frac{1}{2}\sum_{a,b}\hat{R}^a (A^{-1})_{ab}\hat{R}^b + \frac{1}{2}\mathrm{Tr}AB + i\sum_a \hat{R}^a R^a \tag{54}$$

To find the extremum with respect to the elements of $\hat{R}^a$ and $A$ we set the derivative of this expression with respect to $\hat{R}^c$ and $A_{cd}$ to zero.

$$0 = -\sum_a (A^{-1})_{ac}\hat{R}^a + iR^c \tag{55}$$

$$0 = -\frac{1}{2}(A^{-1})_{cd} + \frac{1}{2}\sum_{a,b}\hat{R}^a (A^{-1})_{ac}(A^{-1})_{bd}\hat{R}^b + \frac{1}{2}B_{cd} \quad . \tag{56}$$

This gives

$$\hat{R}^a = i\sum_b A_{ab}R^b \tag{57}$$

and

$$(A^{-1})_{cd} = B_{cd} - R^c R^d := C_{cd} \quad . \tag{58}$$

Using these results we find that (54) is at the saddle point simply given by $\frac{1}{2}\mathrm{Tr}\ln C$. Therefore (49) simplifies to

$$\Omega^{(n)} \sim \exp\{N \operatorname*{extr}_{Q^{ab}, R^a}\left[\frac{1}{2}\mathrm{Tr}\ln C + \alpha G_E(q^{ab}, R^a)\right]\} \tag{59}$$

Determining the remaining extremum with respect to $q^{ab}$ and $R^a$ is not straightforward, in particular in view of the analytic continuation $n \to 0$ to be performed at the end. The subtleties of the general case are discussed in

[5] The transformation $\hat{k}^a \to \hat{k}^a - i\varepsilon$ makes the integral convergent for all $\varepsilon > 0$ and does not change the asymptotic behaviour since the extremum with respect to $\hat{k}^a$ lies on the negative imaginary axis.

detail in [8]. In our case it turns out that the values of $q^{ab}$ and $R^a$ at the extremum are *replica symmetric*, i.e. they obey

$$q^{ab} = q \qquad \text{and} \qquad R^a = R \quad . \tag{60}$$

which simplifies the further analysis considerably.

First we note that in this case the matrix $C$ defined in (58) has an $(n-1)$-fold degenerated eigenvalue $(1-q)$ and an additional one $(1-q)+n(q-R^2)$. This gives

$$\frac{1}{2}\operatorname{Tr}\ln C = \frac{n}{2}\ln(1-q) + \frac{1}{2}\ln(1+n\frac{q-R^2}{1-q}) \quad . \tag{61}$$

We then use the remaining $\theta$-functions in (51) to restrict the integration range of $u$ and $\lambda^a$ and simplify the energetic part using a Hubbard-Stratonovich-transformation. With the shorthand notation $Dt = dt\exp\{-t^2/2\}/\sqrt{2\pi}$ we find

$$\begin{aligned} G_E = & \ln 2\int Dt\int_0^\infty Du\int_0^\infty \prod_a d\lambda^a \int\prod_a \frac{d\hat{\lambda}^a}{2\pi}\exp\{-\frac{1-q}{2}\sum_a(\hat{\lambda}^a)^2 \\ & +i\sum_a \hat{\lambda}^a(\lambda^a - uR - \sqrt{q-R^2}t)\} \\ = & \ln 2\int Dt\int_0^\infty Du\left[\int_0^\infty \frac{d\lambda}{\sqrt{2\pi(1-q)}}\exp\{-\frac{1}{2(1-q)}(\lambda - uR - \sqrt{q-R^2}t)^2\}\right]^n \\ = & \ln 2\int Dt\int_0^\infty Du H^n(-\frac{\sqrt{q-R^2}t+uR}{\sqrt{1-q}}) \end{aligned} \tag{62}$$

Shifting the integration variable $t \to (\sqrt{q-R^2}t + uR)/\sqrt{q}$ the $u$-integral may be performed and we get

$$G_E = \ln 2\int DtH(-\frac{Rt}{\sqrt{q-R^2}})H^n(-\sqrt{\frac{q}{1-q}}t) \tag{63}$$

Extracting the dominant terms in (61) and (63) for $n \to 0$ is now a simple matter and we find

$$\begin{aligned} \Omega^{(n)} \sim \exp\{Nn\operatorname*{extr}_{q,R}\Bigg[ & \frac{1}{2}\ln(1-q) + \frac{q-R^2}{2(1-q)} \\ & +2\alpha\int DtH(-\frac{Rt}{\sqrt{q-R^2}})\ln H(-\sqrt{\frac{q}{1-q}}t)\Bigg]\} \end{aligned} \tag{64}$$

Using (36) this finally gives

$$\frac{1}{N}\langle\langle \ln \Omega(\boldsymbol{\xi}^{\mu}, \mathbf{T})\rangle\rangle = \operatorname*{extr}_{q,R}\left[\frac{1}{2}\ln(1-q) + \frac{q-R^2}{2(1-q)} + 2\alpha \int Dt H(-\frac{Rt}{\sqrt{q-R^2}}) \ln H(-\sqrt{\frac{q}{1-q}}t)\right] \qquad (65)$$

Setting the derivatives with respect to $q$ and $R$ of the right hand side of this expression to zero we find after some partial integration

$$\frac{q-R^2}{1-q} = \frac{\alpha}{\pi}\int DtH(-\frac{Rt}{\sqrt{q-R^2}})\frac{\exp\{-\frac{q}{1-q}t^2\}}{H^2(-\sqrt{\frac{q}{1-q}}t)} \qquad (66)$$

and

$$\frac{R\sqrt{q-R^2}}{\sqrt{q}\sqrt{1-q}} = \frac{\alpha}{\pi}\int Dt\frac{\exp\{-\frac{t^2}{2}(\frac{R^2}{q-R^2}+\frac{q}{1-q})\}}{H(-\sqrt{\frac{q}{1-q}}t)} \quad . \qquad (67)$$

These two equations coincide for $q = R$. In fact our special learning situation is characterized by an interesting symmetry. The teacher $\mathbf{T}$ is chosen at random from a uniform distribution on the whole sphere (cf.(19)) and of course, by definition, lies within the version space. On the other hand our learning scenario consists of sampling student vectors at random with equal probability from the version space. Therefore the typical teacher-student overlap $R$ should coincide with the typical student-student overlap $q$ and we find

$$q = R = \frac{\alpha}{\pi}\sqrt{1-R}\int Dt\frac{\exp\{-\frac{Rt^2}{2}\}}{H(\sqrt{R}t)} \qquad (68)$$

For the entropy we get for $q = R$ from (65)

$$\frac{1}{N}\langle\langle \ln \Omega(\boldsymbol{\xi}^{\mu}, \mathbf{T})\rangle\rangle = \operatorname*{extr}_{R}\left[\frac{1}{2}\ln(1-R) + \frac{R}{2} + 2\alpha\int DtH(-\sqrt{\frac{R}{1-R}}t)\ln H(-\sqrt{\frac{R}{1-R}}t)\right] \qquad (69)$$

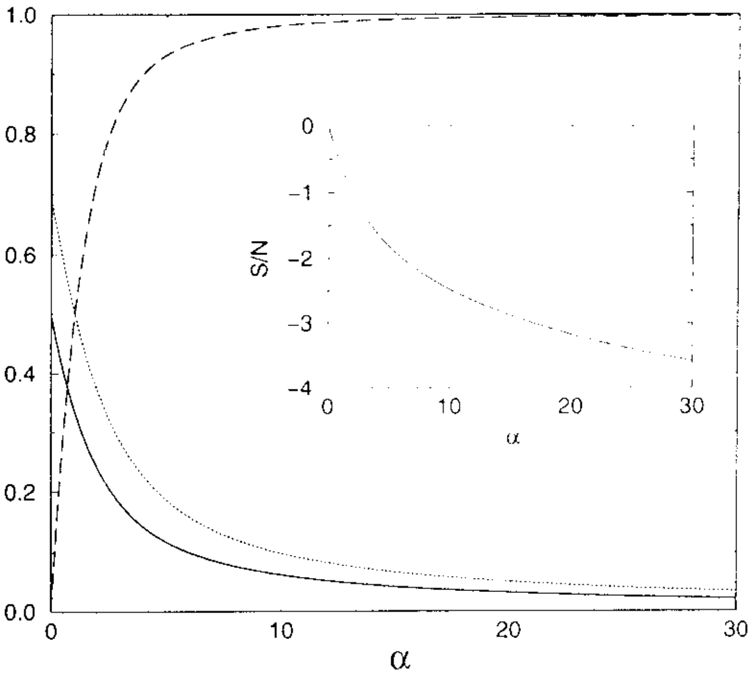

Figure 8: Results for the teacher-student overlap $R$ (dashed) and the generalization error $\varepsilon$ (full) as functions of the training set size $\alpha$. The inset shows the decrease of the quenched entropy per coupling with increasing $\alpha$. The dotted line gives the behaviour of the information gain $\Delta I$.

Fig.8 shows the behaviour of the relevant quantities as resulting from the numerical solution of (68) and subsequent use in (69) and (5). The qualitative behaviour is as in the annealed calculation. The teacher-student overlap $R$ starts at $R = 0$ for $\alpha = 0$ and monotonically increases with $\alpha$ tending asymptotically to $R = 1$ for $\alpha \to \infty$. Correspondingly the generalization error $\varepsilon$ monotonically decreases from its "pure guessing" value $\varepsilon = .5$ to zero for $\alpha \to \infty$. The shrinkage of the version space with increasing number of training examples is quantified by the decrease of the quenched entropy with increasing $\alpha$.

To test the quantitative accuracy of the Gardner approach we have included the result for $\varepsilon(\alpha)$ as full line in fig.4. As can be seen, the agreement with the simulation results is very good. The remaining differences for intermediate values of $\alpha$ are due to the fact that the simulations were done for $N = 20$ whereas the analytical calculations are strictly valid for $N \to \infty$ only. From the nevertheless good agreement between theory and simulation we may conclude that the statistical mechanics analysis, which relies on the thermodynamic limit $N \to \infty$, often yields results that also describe finite systems quite accurately.

We have hence seen that the volume of the version space is *not* self-averaging. Although its probability distribution becomes sharply peaked for $N \to \infty$, the most probable and average values do not coincide! The annealed calculation, resting on the average of the version space volume, therefore fails to describe the typical behaviour. Being easy to calculate it is nevertheless often useful since it may give insight into the qualitative behaviour of the system and can also yield fairly good quantitative approximations.

The correct self-averaging quantity in the generalization problem turns out to be the *entropy* in phase space, i.e. the *logarithm* of the version space volume. Its analytical determination is possible for $N \to \infty$ using the replica trick borrowed from the statistical mechanics of disordered systems. The calculation is fairly straightforward within the replica symmetric ansatz, which we will always use in the first place for the investigation of a learning problem. One has to keep in mind, however, that a consistent treatment always requires checking the reliability of this ansatz. For the simple learning scenario discussed above the replica symmetric assumption yields correct results. This is related to the fact that the version space is connected (it is even convex) meaning that for any two elements there is always a line connecting them that lies entirely inside the version space. A connected version space is the equivalent to an ergodic dynamics which in the theory of disordered systems is known to be correctly described by replica symmetry.

# 3 Lecture 3: Learning by minimizing cost functions

## 3.1 Gibbs learning at non-zero temperature

Gibbs learning as introduced in the previous lecture characterizes the generalization performance of a typical compatible student by averaging over the version space comprising all vectors $\mathbf{J}$ that realize zero training error. On the other hand appreciable generalization is sometimes possible also with non-zero training error. There are interesting situations in which perfect learning is either *impossible* or *too expensive* in the sense that practically the same quality of generalization can be achieved before zero training error has actually been reached. Under these circumstances it is interesting to know the typical generalization behaviour of a student with given *non-zero* training error $\varepsilon_t$. Let us see how the approach of statistical mechanics can be adapted to this more general situation.

For a particular learning situation specified by a teacher vector $\mathbf{T}$ and a set of inputs $\boldsymbol{\xi}^\mu$ the training error $\varepsilon_t$ is defined as the fraction of disagreements between teacher and student on the examples $\boldsymbol{\xi}^\mu$, i.e.

$$\varepsilon_t(\mathbf{J}) = \frac{1}{p}\sum_{\mu=1}^{p}\theta(-\Delta^\mu) \tag{70}$$

with the stability $\Delta^\mu$ defined by

$$\Delta^\nu = \frac{1}{\sqrt{N}}\mathbf{J}\boldsymbol{\xi}^\nu\sigma_T^\nu \quad . \tag{71}$$

The central quantity in the analysis of compatible students was the version space volume

$$\Omega(\boldsymbol{\xi}^\mu, \mathbf{T}) = \int d\mu(\mathbf{J})\prod_{\mu=1}^{p}\theta(\Delta^\mu) \quad . \tag{72}$$

It is clear that in order to include coupling vectors $\mathbf{J}$ with non-zero training error into the analysis we have to replace the indicator function $\chi(\mathbf{J})$ as defined in (15) by a function weighing the different $\mathbf{J}$-vectors according to their training error $\varepsilon_t$ in a smooth manner. The choice advocated by statistical mechanics [6] is

$$\chi(\mathbf{J}) \mapsto \exp(-\beta E(\mathbf{J})) \tag{73}$$

where

$$E(\mathbf{J}) = \sum_{\mu} \theta(-\Delta^{\mu}) = N\alpha\varepsilon_t(\mathbf{J}) \tag{74}$$

and $\beta$ is a free parameter. It implies that the probability of a student vector $\mathbf{J}$ to be chosen on the basis of his performance on the training examples is given by

$$P(\mathbf{J}) \sim \exp(-\beta E(\mathbf{J})) \quad . \tag{75}$$

hence also couplings with $\varepsilon_t > 0$ are taken into account. Surely, their chance to be realized quickly decreases with increasing $\varepsilon_t$.

Let us shortly illustrate why the replacement (73) serves the required purpose. Consider the average of some function $g(\mathbf{J})$ that depends on $\mathbf{J}$ only through the training error $\varepsilon_t(\mathbf{J})$. This average can be written as

$$\int d\mu(\mathbf{J}) \exp(-\beta E(\mathbf{J}))\, g(\mathbf{J}) = \int_0^1 d\varepsilon_t\, \Omega(\varepsilon_t) \exp(-\beta\alpha N\varepsilon_t)\, g(\varepsilon_t) \tag{76}$$

where

$$\Omega(\varepsilon_t) = \int d\mu(\mathbf{J})\, \delta(\varepsilon_t - \frac{1}{p}\sum_{\mu=1}^{p} \theta(-\Delta^{\mu})) \tag{77}$$

describes the part of $\mathbf{J}$-space giving rise to training error $\varepsilon_t$. Similar to the previous lecture it can be shown that the entropy $S(\varepsilon_t) = \ln\Omega(\varepsilon_t)$ is extensive and can hence be written as $S(\varepsilon_t) = Ns(\varepsilon_t)$ with $s(\varepsilon_t) = O(1)$ for $N \to \infty$. The integral (76) therefore assumes the form

$$\int_0^1 d\varepsilon_t \exp(N[s(\varepsilon_t) - \beta\alpha\varepsilon_t])\, g(\varepsilon_t) \tag{78}$$

and for large $N$ the average of $g(\mathbf{J})$ is dominated by coupling vectors $\mathbf{J}$ with the *typical* training error $\varepsilon_t$ that maximizes the exponent in (78), and which

[6] For an argument based on statistics, see the lectures by Sara Solla.

is hence the solution of $ds/d\varepsilon_t = \beta\alpha$. Consequently, by appropriately choosing the free parameter $\beta$ one can concentrate the whole average on student vectors with a particular training error $\varepsilon_t$. As $\beta$ increases, the corresponding typical training error decreases. The version space analysis of the previous lecture is eventually recovered for $\beta \to \infty$.

The observation of the temperature selecting a specific training error, hence a specific value of $E$ is, of course, just the illustration of the equivalence between a canonical and a microcanonical description, well known from statistical mechanics. $E(\mathbf{J})$ as defined in (74) plays the role of *energy* and $\beta$ that of the inverse temperature $T = 1/\beta$. The microcanonical approach used in lecture 2 based on the evaluation of the phase space volume $\Omega(\varepsilon; \boldsymbol{\xi}^\mu, \mathbf{T})$ is thus replaced by the calculation of the so-called *partition function*

$$Z(\beta, \boldsymbol{\xi}^\mu, \mathbf{T}) = \int d\mu(\mathbf{J}) \exp(-\beta E(\mathbf{J})) \tag{79}$$

normalizing the canonical distribution (75). The central quantity of the canonical approach becomes the so-called *free energy*

$$F(\beta, \boldsymbol{\xi}^\mu, \mathbf{T}) = -T \ln Z(\beta, \boldsymbol{\xi}^\mu, \mathbf{T}) \tag{80}$$

which is extensive and therefore assumed to be self-averaging for large $N$. The corresponding free-energy density

$$f(\beta, \alpha) = \lim_{N\to\infty} \frac{1}{N} \langle\langle F(\beta, \boldsymbol{\xi}^\mu, \mathbf{T}) \rangle\rangle = -T \lim_{N\to\infty} \frac{1}{N} \langle\langle \ln Z(\beta, \boldsymbol{\xi}^\mu, \mathbf{T}) \rangle\rangle \tag{81}$$

can be calculated by a minor modification of the formalism discussed in the previous lecture. Since

$$\begin{aligned} \exp(-\beta E(\mathbf{J})) &= \exp(-\beta \sum_\mu \theta(-\Delta^\mu)) \\ &= \prod_\mu [e^{-\beta} + (1 - e^{-\beta})\theta(\Delta^\mu)] \end{aligned} \tag{82}$$

this modification reduces to the replacement

$$\theta(\Delta) \mapsto [e^{-\beta} + (1 - e^{-\beta})\theta(\Delta)] \tag{83}$$

in (51). The final results can then be read off (65)-(67):

$$f(\beta,\alpha) = -\operatorname*{extr}_{q,R}\left[\frac{q-R^2}{2\beta(1-q)} + \frac{1}{2\beta}\ln(1-q)\right.$$
$$\left. +\frac{2\alpha}{\beta}\int Dt\, H(-\frac{Rt}{\sqrt{q-R^2}})\ln\left[e^{-\beta} + (1-e^{-\beta})H(-\sqrt{\frac{q}{1-q}}t)\right]\right] \tag{84}$$

with the corresponding saddle-point equations

$$\frac{q-R^2}{1-q} = \frac{\alpha}{\pi}\int Dt\; H(-\frac{Rt}{\sqrt{q-R^2}})\frac{\exp\{-\frac{q}{1-q}t^2\}}{\left[\frac{1}{e^{\beta}-1} + H(-\sqrt{\frac{q}{1-q}}t)\right]^2} \tag{85}$$

and

$$\frac{R\sqrt{q-R^2}}{\sqrt{q}\sqrt{1-q}} = \frac{\alpha}{\pi}\int Dt\; \frac{\exp\{-\frac{t^2}{2}(\frac{R^2}{q-R^2}+\frac{q}{1-q})\}}{\frac{1}{e^{\beta}-1} + H(-\sqrt{\frac{q}{1-q}}t)} \tag{86}$$

fixing the order parameters $R$ and $q$. Again $R$ denotes the typical teacher-student overlap, determining the corresponding generalization error $\varepsilon$ through (5), while $q$ corresponds to the typical overlap between two students. Note that $R \neq q$ due to the absence of a teacher-student symmetry.

The relation between the free energy and the internal energy $\alpha\epsilon_t$ and the entropy $s$ can of course be written as usual: $f = \min_{\varepsilon_t}[\alpha\varepsilon_t - Ts(\varepsilon_t)]$. The *typical* training error then follows from the standard relation:

$$\varepsilon_t = \frac{1}{\alpha}\frac{\partial(\beta f(\beta,\alpha))}{\partial\beta} \quad . \tag{87}$$

In fig.9 the training and generalization error are shown for different values of $T$ as a function of $\alpha$. One observes that the generalization error is almost as small for moderate $T > 0$ as it is for zero-temperature learning. Asymptotically the behaviour remains $\varepsilon \sim 1/\alpha$ and the whole influence of the non-zero training error boils down to a temperature dependent prefactor which increases with increasing $T$.

## 3.2 General statistical mechanics formulation

We are now ready to demonstrate that for most of the learning rules discussed in the literature an analytical calculation of the learning and generalization

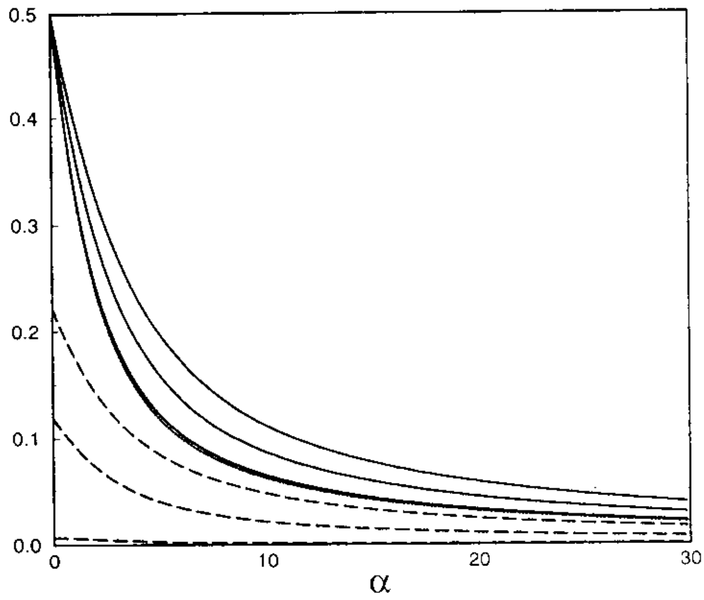

Figure 9: Generalization error (full) and training error (dashed) as functions of the training set size $\alpha$ for Gibbs-learning at temperature $T = .2, .5$ and $.8$ (from bottom to top). For comparison the generalization error for $T = 0$ is shown as dotted line.

error is possible within the statistical mechanics framework. The main point is the observation that the learning rules are equivalent to selecting a student vector which gives the *minimum* value of an appropriately chosen *cost function* which we will call the *learning error* $E(\mathbf{J})$.

For random examples chosen independently of each other, it is most natural to use a learning error of the following additive form :

$$E(\mathbf{J}) = \sum_{\mu} V(\Delta^{\mu}) \tag{88}$$

with the stabilities $\Delta^{\mu}$ defined in (71). The cost function (74) used in Gibbs learning forms a special case of (88).

The learning error introduced above is extensive in the number $p$ of examples, which itself is chosen proportional to the dimensionality $N$ of the input space. We saw in finite temperature Gibbs learning, that within the statistical mechanics analysis this error plays the role of the *energy* with which by happy coincidence it shares the same abbreviation $E$. The generalization error will be determined by the competition of this energy with the entropic term, namely the number of **J**-vectors that correspond to a given value of the learning error.

Our aim is to calculate the generalization performance as a function of the training set size $\alpha$ for different choices of the *potential* $V$ occuring in (88). Similar to the analysis of general Gibbs learning above this can be achieved by introducing the partition function associated with the error $E$:

$$Z = \int d\mu(\mathbf{J})\, e^{-\beta E(\mathbf{J})} \quad . \tag{89}$$

The partition function is a random variable through its dependence on the randomly chosen training examples and teacher couplings. Again the corresponding free energy $F = Nf = -T \ln Z$ is an extensive quantity, which is expected to be self-averaging in the thermodynamic limit. Hence $F$ can be evaluated by performing the average over the examples and the teacher. The required replica technique follows very closely the line of calculation for the finite temperature Gibbs case, with the only modification that $\theta(-\Delta)$ is substituted by $V(\Delta)$. Under the assumption of replica symmetry one finds that :

$$f(\beta,\alpha) = -\operatorname*{extr}_{q,R}\left[\frac{q-R^2}{2\beta(1-q)} + \frac{1}{2\beta}\ln(1-q)\right.$$
$$\left. + \frac{2\alpha}{\beta}\int Dt\; H(-\frac{Rt}{\sqrt{q-R^2}})\ln\int \frac{d\Delta}{\sqrt{2\pi(1-q)}}\exp(-\beta V(\Delta) - \frac{(\Delta-\sqrt{q}t)^2}{2(1-q)})\right] \quad . \tag{90}$$

The order parameters $q$ and $R$ have their usual meaning and are determined from the extremum conditions corresponding to (90). The generalization error is then obtained by inserting the value of $R$ into (5).

Only those coupling vectors **J** contribute substantially to the partition function $Z$ defined in (89), for which the learning error $E(\mathbf{J})$ is only of order $1/\beta$ larger than the minimal possible value $E_{min}$. It is hence clear that with increasing $\beta$ both the partition function and the free energy will be dominated by vectors **J** with smaller and smaller values of $E(\mathbf{J})$. Eventually, i.e. in the limit $\beta \to \infty$, it will therefore be possible to extract the properties of the optimal coupling vector minimizing the cost function from (90).

From an algorithmic point of view, those cost functions $E(\mathbf{J})$ which possess a *unique minimum* are of particular interest. Indeed, if this is the case, the minimum may be found using a standard gradient descent algorithm. In such a case, the limit $\beta \to \infty$ will be accompanied by $q \to 1$ since the different solutions $\mathbf{J}$ all have to converge to the same minimum. One observes from the saddle point equations corresponding to (90) that $\beta \to \infty$ and $q \to 1$ are compatible if we take $x = \beta(1-q)$ to be finite. Replacing the extremum condition for $q$ by one for $x$ and after applying a saddle-point argument to the $\Delta$-integral, the free energy density in this limit reduces to

$$\begin{aligned} & f(T=0,\alpha) = e_{min}(\alpha) \\ & = -\operatorname*{extr}_{x,R}\left[\frac{1-R^2}{2x} - 2\alpha\int Dt\ H(-\frac{Rt}{\sqrt{1-R^2}})\min_{\Delta}\left(V(\Delta)+\frac{(\Delta-t)^2}{2x}\right)\right] \end{aligned} \tag{91}$$

where $E_{min} = Ne_{min}$ is the minimum of the learning error $E(\mathbf{J})$.

For a given potential $V(\Delta)$ the generalization performance of the $\mathbf{J}$ vector minimizing $V$ can hence be obtained by the following two steps:

(1) Find the function $\Delta_0(t,x)$ which minimizes

$$V(\Delta) + \frac{(\Delta-t)^2}{2x} \quad . \tag{92}$$

(2) Determine the values of $R$ and $x$ as a function of $\alpha$ from the saddle-point equations corresponding to (91) which are of the form:

$$2\alpha\int Dt\ (\Delta_0(t,x)-t)^2\ H(-\frac{Rt}{\sqrt{1-R^2}}) \ = \ 1-R^2 \tag{93}$$

$$\frac{2\alpha}{\sqrt{2\pi(1-R^2)}}\int Dt\ \Delta_0(t,x)\ \exp(-\frac{R^2t^2}{2(1-R^2)}) \ = \ R \quad . \tag{94}$$

## 3.3 A choice of learning rules

The above formalism can be applied to various learning rules. As a particular simple example Hebb-learning gives a useful illustration. It corresponds to the specific choice

$$V(\Delta) = -\Delta \tag{95}$$

for which the minimum of the cost function $E(\mathbf{J})$ can be found explicitly, namely

$$\mathbf{J} \sim \sum_{\mu} \boldsymbol{\xi}^{\mu} \sigma_T^{\mu} \quad . \tag{96}$$

From (92) - (94) one obtaines

$$\Delta_0(t, x) = t + x \tag{97}$$

and

$$R = \sqrt{\frac{2\alpha}{2\alpha + \pi}} \tag{98}$$

$$x = \sqrt{\frac{\pi}{2}\frac{R}{\alpha}} \tag{99}$$

which results in

$$\varepsilon = \frac{1}{\pi} \arccos \sqrt{\frac{2\alpha}{2\alpha + \pi}} \quad . \tag{100}$$

This result can also be obtained by a simple statistical analysis [11].

Gibbs learning uses the training error as cost function, i.e.

$$V(\Delta) = \theta(-\Delta) \tag{101}$$

such that the corresponding error function $E$ just counts the total number of misclassifications. Obviously (90) reduces to (84) in this case.

Adaline learning is defined by a least square error goal of the form

$$V(\Delta) = \frac{1}{2\kappa^2}(\Delta - \kappa)^2 \quad . \tag{102}$$

For $\alpha < 1$ this rule gives rise to the coupling vector

$$\mathbf{J}^{PI} = \frac{1}{\sqrt{N}} \sum_{\mu,\nu} (C^{-1})_{\mu\nu} \boldsymbol{\xi}^{\mu} \sigma_T^{\nu} \tag{103}$$

with the correlation matrix

$$C_{\mu\nu} = \frac{1}{N} \boldsymbol{\xi}^{\mu} \boldsymbol{\xi}^{\nu} \sigma_T^{\mu} \sigma_T^{\nu} \quad . \tag{104}$$

If the vectors $\boldsymbol{\xi}^\mu \sigma_T^\mu$ are linearly independent as happens with probability 1 for large $N$ and independent $\boldsymbol{\xi}^\mu$ if $\alpha < 1$ the matrix $C_{\mu\nu}$ is non-singular. Since the student couplings are then explicitly known the generalization behaviour can be analyzed using methods from statistics [12]. The application of adaline for $\alpha > 1$ involves an additional minimization in $\kappa$. The analysis of adaline learning for $\alpha > 1$ can hence be easily accomplished as follows. From (102) we get

$$\Delta_0(t, x) = \frac{t + \frac{x}{\kappa}}{1 + \frac{x}{\kappa^2}} \tag{105}$$

so that the integral in the expression (91) for the ground state energy can be performed analytically and we find including the minimization in $\kappa$:

$$e_{min}(\alpha) = -\operatorname*{extr}_{x,R,\kappa} \left[ \frac{1 - R^2}{2x} - \frac{\alpha}{2(\kappa^2 + x)}(\kappa^2 + 1 - 2\sqrt{\frac{2}{\pi}}\kappa R) \right] \tag{106}$$

The extremum conditions give rise to the following three equations

$$1 - R^2 = \alpha \left( \frac{x}{\kappa^2 + x} \right)^2 (\kappa^2 + 1 - 2\sqrt{\frac{2}{\pi}}\kappa R) \tag{107}$$

$$R = \sqrt{\frac{2}{\pi}} \frac{x}{\kappa^2 + x} \alpha \kappa \tag{108}$$

$$(1 - R^2) = \alpha \frac{x^2}{\kappa^2 + x} (1 - \sqrt{\frac{2}{\pi}} \frac{R}{\kappa}) \tag{109}$$

For $\alpha > 1$ these equations admit the unique solution

$$\kappa = \sqrt{\frac{\pi(\alpha - 1)}{\pi + 2\alpha - 4}} \qquad R = \sqrt{\frac{2(\alpha - 1)}{\pi + 2\alpha - 4}} \qquad x = \frac{\pi}{\pi + 2\alpha - 4} \quad . \tag{110}$$

The result for $R(\alpha)$ specifies the decay of the generalization error and is included in fig.10. The asymptotic behaviour is given by

$$\varepsilon(\alpha) \sim \sqrt{\frac{\pi - 2}{2\pi^2 \alpha}} \tag{111}$$

which obeys the same power law as for the Hebb rule. Note the remarkable behavior of the generalization error at intermediate values of $\alpha$. It goes through a local minimum at $\alpha = .62$ and *increases back* to the random guessing value $\epsilon = .5$ for $\alpha = 1$. This non-monotonic behaviour is termed *overfitting* because of its similarity with the analogous phenomenon observed in polynomial interpolation. In the present case it results from the insistance on a "wrong concept" when training the system. Indeed the real stabilities of the examples with respect to the teacher are not all identical and forcing the student to obey this prescription completely compromizes his generalization ability.

A similar analysis is possible for many other learning rules [10]. Some results are summarized in table 1 and fig.10.

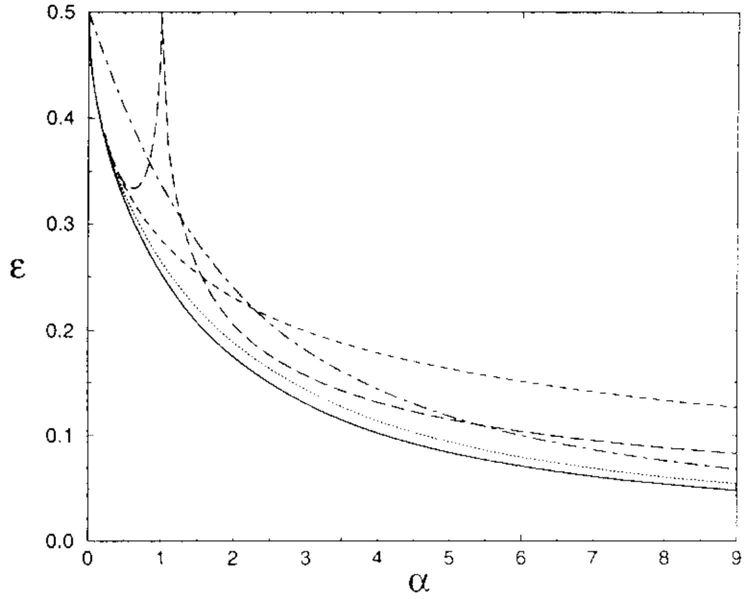

Figure 10: Generalization error $\varepsilon$ as a function of the training set size $\alpha$ for the Hebb-rule (dashed). Gibbs learning at zero temperature (dashed-dotted), the adaline / pseudo-inverse rule (long dashed). the maximal stable vector resulting e.g. from the adaline rule (dotted) and the Bayes rule (full). The asymptotic behaviour for $\alpha \to \infty$ is listed in table 1.

| Learning rule | potential $V(\Delta)$ | large $\alpha$-asymptotics of $\varepsilon$ |
|---|---|---|
| Hebb | $-\Delta$ | $\frac{1}{\sqrt{2\pi\alpha}}$ |
| Gibbs ($T = 0$) | $\theta(-\Delta)$ | $\frac{.625}{\alpha}$ |
| adaline | $\frac{1}{2\kappa^2}(\Delta - \kappa)^2$ | $\sqrt{\frac{\pi - 2}{2\pi^2\alpha}}$ |
| adatron | $\frac{1}{2}(\Delta - \kappa)^2\theta(\kappa - \Delta)$ | $\frac{.5005}{\alpha}$ |
| Bayes | | $\frac{.442}{\alpha}$ |

Table 1: Potential of the cost function and asymptotic behaviour of the generalization error $\varepsilon$ for large values of the training set size $\alpha$ for various learning rules. The generalization performance for smaller values of $\alpha$ are shown in fig.10.

### 3.4 The optimal potential $V(\Delta)$

Two natural questions have remained unanswered so far: first, is there an optimal choice of the potential $V(\Delta)$ and second, does the resulting generalization error saturate the lower bound resulting from the Bayes rule? A variational calculation allows to answer both questions affirmatively [15]. Let us define the quantities

$$\begin{aligned} F(t,x) &= \Delta^0(t,x) - t = -xV'(\Delta^0(t,x)) && (112) \\ X &= 2H(-\frac{Rt}{\sqrt{1-R^2}}) && (113) \\ Y &= \frac{2\exp(-\frac{R^2t^2}{2(1-R^2)})}{\sqrt{2\pi(1-R^2)}} && (114) \end{aligned}$$

where $\Delta^0$ is defined through (92). From (93) and (94), one obtaines:

$$\frac{R^2}{1-R^2} = \alpha \frac{\langle YF\rangle^2}{\langle XF^2\rangle} = \alpha \frac{\langle \frac{Y^2}{X} \frac{XF}{Y}\rangle^2}{\langle \frac{Y^2}{X}(\frac{XF}{Y})^2\rangle} \tag{115}$$

where the average is with respect to the Gaussian measure $Dt$. It is easy to verify, e.g. on the basis of the Schwartz inequality, that $\langle u\, v\rangle^2 \leq \langle u\rangle\langle u\, v^2\rangle$ for any function $u > 0$, with the equality sign attained only for $v$ constant with respect to the variables over which the average is being taken. It follows that the r.h.s. of (115), and hence also the value of $R$, is maximal for the choice $F^* = CY/X$, where $C$ is a constant independent of $t$. The r.h.s. of (115) then simplifies to $\alpha\langle Y^2/X\rangle$. By a change of variables $t = -u\sqrt{1-R^2}$, one thus finds that (115) is equivalent to

$$\frac{R_B^2}{\sqrt{1-R_B^2}} = \frac{\alpha}{\pi}\int Dt \frac{\exp\{-\frac{R_B^2}{2}t^2\}}{H(-R_B t)} \quad , \tag{116}$$

the equation for the Bayes rule [13, 14]. Hence, we have identified a cost function through the explicit form of $F^*$, whose unique minimum reproduces the Bayes generalization behaviour.

This result sounds exciting, since we are able to reproduce the lowest possible generalization error through gradient descent. There are however some reasons for caution. First, the potential is specific to the teacher student perceptron scenario with random examples. Second, even though one can construct the optimal potential explicitly, it does not have a simple form. In particular, a more detailed analysis [15] reveals that the potential is infinite for negative stabilities, hence one has to start the gradient descent procedure while already inside the version space (which can be achieved by a prior application of the adatron algorithm). Third, the optimal potential depends on $\alpha$, i.e. its form changes with the size of the training set. This feature is expected to be rather generic since the optimal strategy to be followed will in general depend on the amount of information already obtained. Note also that the optimal cost function depends on $R$, hence in order to use the optimal potential we have first to calculate $R(\alpha)$ analytically and to rely on the self-averaging character of $R$.

# 4 Lecture 4: Noisy Teachers

## 4.1 Motivation

As a rule teachers are unreliable. From time to time they mix up questions or answer absentmindedly. How much can a student network learn about a target rule if some of the examples in the training set are corrupted by random noise? What is the optimal strategy for the student in this more complicated situation?

Let us analyze these questions in detail for the two perceptron scenario. It should be emphasized that quite generally a certain robustness with respect to random influences is an indispensable requirement for any information processing system, both in biological and in technical context. If learning from examples were possible only for perfectly error-free training sets it would be of no practical interest. In fact, since the noise blurring the correct classifications of the teacher can usually be assumed to be independent of the examples one expects that it remains possible to infer the rule, probably at the expense of a larger training set.

A general feature of noisy generalization tasks is that the training set is no longer generated by a rule that can be implemented by the student. The problem is said to be *unrealizable*. A simple example is a training set containing the same input with different outputs which is well possible for noisy teachers. This means that for large enough training sets no student exists that can reproduce all classifications and the version space becomes *empty*. For networks with many inputs this transition occurs at a sharp threshold $\alpha_c$ of the training set size. Above this threshold the training error $\varepsilon_t$ of the student is always larger than zero and the question arises whether it is really necessary or advantegeous to insist on zero training error below the threshold.

## 4.2 Sources of noise

We will now investigate these problems in detail for the paradigmatic setup of a teacher and a student perceptron. As before $\mathbf{T}$ and $\mathbf{J}$ are the coupling vectors of teacher and student perceptron respectively, and $\boldsymbol{\xi}^\mu$ denotes the inputs of the training set chosen at random according to the distribution

(18). However, the corresponding outputs are now given by

$$\sigma'^{\mu}_{T} = \eta^{\mu}\mathrm{sgn}(\frac{1}{\sqrt{N}}\mathbf{T}'^{\mu}\boldsymbol{\xi}'^{\mu}) \quad . \tag{117}$$

Several sources of noise have been incorporated into this expression. The possibility that the couplings of the teacher herself are fluctuating from example to example around the pure values T has been accounted for by the replacement $\mathbf{T} \mapsto \mathbf{T}'^{\mu}$. An appropriate distribution for the $T'^{\mu}_{i}$ is

$$P(T'^{\mu}_{i}) = \frac{1}{\sqrt{2\pi\sigma^2_w}}\exp(-\frac{(T'^{\mu}_{i} - T_i)^2}{2\sigma^2_w}) \tag{118}$$

where $\sigma_w$ denotes the strength of the fluctuations. This type of noise is called *weight noise.*

Alternatively the errors in the training set may arise because the teacher receives corrupted inputs $\boldsymbol{\xi}'^{\mu}$ instead of the original $\boldsymbol{\xi}^{\mu}$. The generic probability distribution for this kind of *input noise* is given by

$$P(\xi'^{\mu}_{i}) = \frac{1}{\sqrt{2\pi\sigma^2_{in}}}\exp(-\frac{(\xi'^{\mu}_{i} - \xi^{\mu}_{i})^2}{2\sigma^2_{in}}) \quad . \tag{119}$$

where again $\sigma_{in}$ characterizes the noise strength. The impact of input noise is for large input dimension $N$ completely identical to that of weight noise because in both cases the local field of the teacher is Gaussian distributed around the error-free value $\mathbf{T}\boldsymbol{\xi}^{\mu}/\sqrt{N}$. Both types of noise therefore affect in particular examples that are near to the decision boundary of the teacher.

A qualitatively different mechanism to blur the examples of the training set is described by the parameters $\eta^{\mu}$ with distribution

$$P(\eta^{\mu}) = \frac{1+a}{2}\delta(\eta^{\mu} - 1) + \frac{1-a}{2}\delta(\eta^{\mu} + 1) \quad . \tag{120}$$

In this case a fraction $(1-a)/2$ of all classifications is inverted, irrespective of how "confident" the teacher is. It will turn out that this type of *output noise* has a quite different influence on the generalization behaviour of the student.

At this point we have to note that in the case of noisy examples there are two possibilities to characterize the degree to which the student is able to approximate the teacher. We may either ask for the probability that the student gives for a randomly chosen input a different output than the teacher. Or we may be interested in the probability that there is a difference in the *error-free* classifications of the teacher and those of the student. The first quantity is called the *prediction error* $\varepsilon_p$ the latter is the generalization error $\varepsilon$. Only in the case of noiseless learning do both error measures coincide. Clearly, if the student is able to decipher the target rule the generalization error $\varepsilon$ will tend to zero for large training set sizes. The prediction error $\varepsilon_p$, however, will always be larger than a certain residual error $\varepsilon_r$ characterizing the noise corrupting the teacher classifications.

Both the generalization and the prediction error are simple functions of the teacher-student overlap $R$ which, also in the presence of noise, remains a self-averaging quantity. The generalization error is again given by $\varepsilon = \arccos R/\pi$. The prediction error can be obtained by a similar argument. For the case of input and weight noise we get from its definition:

$$\varepsilon_p = \langle\langle\theta(-\sigma'_T\sigma)\rangle\rangle_{\mathbf{S},noise} = \langle\langle\theta(-\frac{1}{\sqrt{N}}\sum_i T'_i\xi'_i \frac{1}{\sqrt{N}}\sum_i J_i\xi_i))\rangle\rangle_{\mathbf{S},noise} \qquad (121)$$

where the average is now over the test example $\mathbf{S}$ and the noise. Similar to (20) we introduce the auxiliary variables

$$\lambda = \frac{1}{\sqrt{N}}\mathbf{J}\cdot\mathbf{S} \quad \text{and} \quad u' = \frac{1}{\sqrt{N}}\mathbf{T}'\cdot\mathbf{S}' \qquad (122)$$

that for large $N$ are Gaussian random variables with moments

$$\begin{aligned} \langle\langle\lambda\rangle\rangle &= \langle\langle u'\rangle\rangle = 0 \qquad (123)\\ \langle\langle\lambda^2\rangle\rangle &= 1 \\ \langle\langle u'^2\rangle\rangle &= (1+\sigma_{in}^2)(1+\sigma_w^2) \\ \langle\langle\lambda u'\rangle\rangle &= R \end{aligned}$$

Replacing the disorder average in (121) by the average over $u'$ and $\lambda$ and using (25) we find for input and weight noise

$$\varepsilon_p = \frac{1}{\pi}\arccos(\gamma R) \qquad (124)$$

where

$$\gamma = \frac{1}{\sqrt{(1+\sigma_{in}^2)(1+\sigma_w^2)}} \qquad (125)$$

is an appropriate parameter describing the noise strength. The residual error $\varepsilon_r$ that remains even if the student has learned the error-free classification of the teacher perfectly results from this expression for $R = 1$, i.e.

$$\varepsilon_r = \frac{1}{\pi} \arccos \gamma \tag{126}$$

In the case of output noise we get similarily

$$\begin{aligned} \varepsilon_p &= \langle\langle \theta(-\sigma'_T \sigma) \rangle\rangle_{\mathbf{S},noise} \\ &= \frac{1+a}{2} \langle\langle \theta(-\sigma_T \sigma) \rangle\rangle_{\mathbf{S}} + \frac{1-a}{2} \langle\langle \theta(\sigma_T \sigma) \rangle\rangle_{\mathbf{S}} \\ &= \frac{1}{\pi} (\frac{1+a}{2} \arccos R + \frac{1-a}{2} \arccos(-R)) \\ &= \frac{1-a}{2} + \frac{a}{\pi} \arccos R \end{aligned} \tag{127}$$

The residual error is of course given by

$$\varepsilon_r = \frac{1-a}{2} \quad . \tag{128}$$

Hence, also in the presence of noise the central quantity to characterize the generalization performance is the teacher-student overlap $R$. In order to calculate it as a function of the training set size $\alpha$ and the parameters of the noise we can use a straightforward generalization of the methods introduced in lecture 2.

## 4.3 Trying perfect learning

To begin with we investigate the generalization behaviour of a student perceptron that irrespective of the noise present in the training set tries to reproduce all example classifications exactly. A suitable learning rule to do that is zero-temperature Gibbs-learning characterizing the performance of a typical student from the version space.

Let us first consider the case of weight or input noise. Including the average over the noise as specified by the probability distributions (118) and (119) into the pattern average (42) we find that all that changes in the subsequent calculations is the replacement $R \mapsto \gamma R$ in the energetic part

(51). Consequently the saddle point equations (66),(67) for the two order parameters $R$ and $q$ are modified to

$$\frac{q-R^2}{1-q} = \frac{\alpha}{\pi}\int DtH(-\frac{\gamma Rt}{\sqrt{q-\gamma^2R^2}})\frac{\exp\{-\frac{q}{1-q}t^2\}}{H^2(-\sqrt{\frac{q}{1-q}}t)} \tag{129}$$

and

$$\frac{R\sqrt{q-\gamma^2R^2}}{\sqrt{q}\sqrt{1-q}} = \gamma\frac{\alpha}{\pi}\int Dt\frac{\exp\{-\frac{t^2}{2}(\frac{\gamma^2R^2}{q-\gamma^2R^2}+\frac{q}{1-q})\}}{H(-\sqrt{\frac{q}{1-q}}t)} \tag{130}$$

Clearly the solution $q = R$ is lost reflecting the absence of symmetry between teacher and student. In fact for $\gamma < 1$ one has always $q > R$ and there is a critical training set size $\alpha_c$ at which $q \to 1$ with $R = R_c < 1$. Performing the limit $q \to 1$ in the above saddle point equations by using the asymptotic behaviour of the $H$-function one finds after some integration

$$\frac{\gamma}{R_c}\sqrt{1-\gamma^2R_c^2} = \arccos(\gamma R_c) \tag{131}$$

and

$$\frac{1}{\alpha_c} = \frac{1}{\pi}\arccos(\gamma R_c) = \varepsilon_p(R_c) \tag{132}$$

from which $R_c$ and $\alpha_c$ can be easily determined numerically.

A qualitatively similar behaviour results for the case of output noise. Now the presence of the $\eta$-variables gives rise to the replacement $\prod_a \theta(u\lambda^a) \mapsto \langle\langle\prod_a \theta(\eta u\lambda^a)\rangle\rangle_\eta$ in the expression (51) for the energetic part and the saddle-point equations assume the form

$$\frac{q-R^2}{1-q} = \frac{\alpha}{\pi}\int Dt\left[\frac{1-a}{2}+aH(-\frac{Rt}{\sqrt{q-R^2}})\right]\frac{\exp\{-\frac{q}{1-q}t^2\}}{H^2(-\sqrt{\frac{q}{1-q}}t)} \tag{133}$$

and

$$\frac{R\sqrt{q-R^2}}{\sqrt{q}\sqrt{1-q}} = a\frac{\alpha}{\pi}\int Dt\frac{\exp\{-\frac{t^2}{2}(\frac{R^2}{q-R^2}+\frac{q}{1-q})\}}{H(-\sqrt{\frac{q}{1-q}}t)} \tag{134}$$

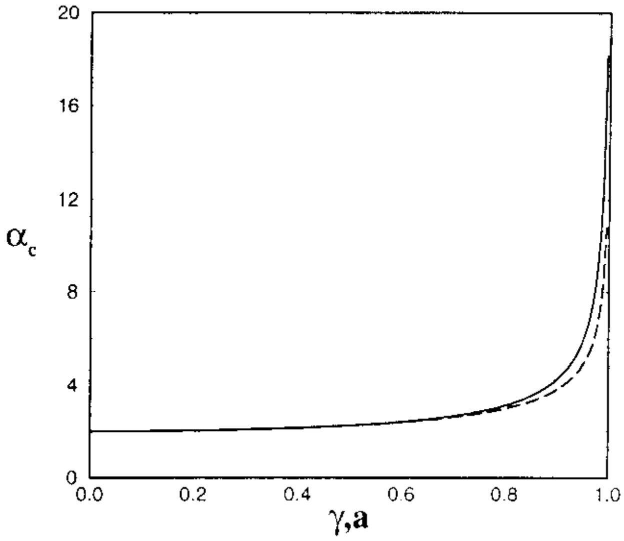

Figure 11: Critical training set size $\alpha_c$ above which no student can be found who reproduces all examples perfectly. Results for input or weight noise of intensity $\gamma$ defined in (125) (full line) and output noise characterized by strength $a$ (cf. (120)) (dashed line) respectively.

Again $q > R$ for $a < 1$ and performing the limit $q \to 1$ as above we arrive at the following equations determining $\alpha_c$ and $R_c$:

$$\frac{\sqrt{1-R_c^2}}{R_c} = \pi\frac{1-a}{2a} + \arccos R_c \tag{135}$$

and

$$\frac{1}{\alpha_c} = \frac{a}{\pi}\frac{\sqrt{1-R_c^2}}{R_c} = \varepsilon_p(R_c) \quad . \tag{136}$$

It is intuitive that the critical training set size $\alpha_c$ is reached when the product of prediction error and training set size approaches 1. The dependence of $\alpha_c$ on $\gamma$ is shown in fig.11 for both types of noise.

For $\gamma \to 1$ and $a \to 1$ respectively the intensity of the noise tends to zero (cf.(125)) and accordingly $\alpha_c \to \infty$. The value of $\alpha_c$ for output noise is always smaller than that for weight or input noise. This is in accordance with the fact that output noise introduces some "gross errors" into the training set whereas weight and input noise only give rise to misclassifications at the decision boundary of the teacher resulting in an "almost realizable" problem for the student.

In order to investigate the generalization behaviour for $\alpha > \alpha_c$ we cannot use the above methods resting on averages over the version space because the latter is empty. Nevertheless it is still possible to study the performance of the student that *minimizes* the training error $\varepsilon_t$ by using the canonical phase space analysis introduced in the previous lecture with the training error as cost function. The minimum of $\varepsilon_t$ is found by performing the zero-temperature limit $\beta \to \infty$ that is accompanied by the limit $q \to 1$ giving rise to the new order parameter $x = \beta(1-q)$. The replica symmetric calculations are again straightforward modifications of those performed before. They result in the case of input and weight noise in the saddle point equations

$$1 - R^2 = 2\alpha \int\limits_{-\sqrt{2x}}^{0} Dt\, t^2 H(-\frac{\gamma R t}{\sqrt{q - \gamma^2 R^2}}) \qquad (137)$$

and

$$\frac{R}{\sqrt{1-\gamma^2 R^2}} = \gamma \frac{\alpha}{\pi} \left( 1 - \exp(-\frac{x}{1-\gamma^2 R^2}) \right) \qquad (138)$$

which fix the order parameters $R$ and $x$ and the expression

$$\varepsilon_t = \frac{1}{\pi} \arccos(\gamma R) - \frac{1-R^2}{2\alpha x} + 2 \int\limits_{-\sqrt{2x}}^{0} Dt\, H(-\frac{\gamma R t}{\sqrt{q - \gamma^2 R^2}})(\frac{t^2}{2x} - 1) \qquad (139)$$

which gives the typical training error $\varepsilon_t$ as a function of $R$ and $x$. Note that for $x = \infty$ equations (131) and (132) for $R_c$ and $\alpha_c$ are reproduced as it should be since for $\alpha < \alpha_c$ the minimum of $\varepsilon_t$ is not unique and $q$ remains smaller than 1 even if $\beta \to \infty$. The results of the numerical solution of (129),(130),(137), and (138) are shown in fig.12 and give a complete picture of the generalization behaviour in the presence of weight and input noise.

Equivalent calculations for output noise give rise to

$$1 - R^2 = 2\alpha \int\limits_{-\sqrt{2x}}^{0} Dt\, t^2 \left[ \frac{1-a}{2} + a\, H(-\frac{Rt}{\sqrt{q - R^2}}) \right] \qquad (140)$$

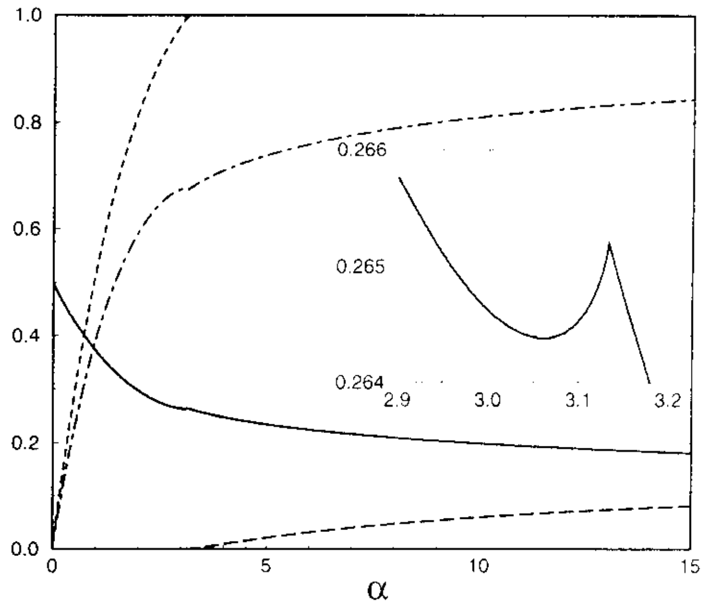

Figure 12: Generalization (full) and training error (long dashed) as well as the order parameters $R$ (dashed dotted) and $q$ (dashed) as functions of the training set size $\alpha$ for $T = 0$ Gibbs-learning in the presence of input or weight noise of intensity $\gamma = .8$. The vertical dotted line marks the critical training set size $\alpha_c$. The inset shows an enlarged part of the $\varepsilon(\alpha)$-curve around $\alpha_c$ displaying the overfitting preceeding criticality.

and

$$\frac{R}{\sqrt{1-R^2}} = a\,\frac{\alpha}{\pi}\left(1 - \exp(-\frac{x}{1-R^2})\right) \tag{141}$$

as saddle point equations for the order parameters $R$ and $x$, and

$$\varepsilon_t = \frac{1-a}{2} + \frac{a}{\pi}\arccos R - \frac{1-R^2}{2\alpha x} + 2\int_{-\sqrt{2x}}^{0} Dt\left[\frac{1-a}{2} + aH(-\frac{Rt}{\sqrt{q-R^2}})\right](\frac{t^2}{2x} - 1) \tag{142}$$

as expression for the typical training error. Again we recover for $x = \infty$ (135) and (136) determining $R_c$ and $\alpha_c$. Fig. 13 gives the complete information on the generalization performance of zero temperature Gibbs-learning in the presence of output noise. Qualitatively the behaviour is similar to the case of input or weight noise.

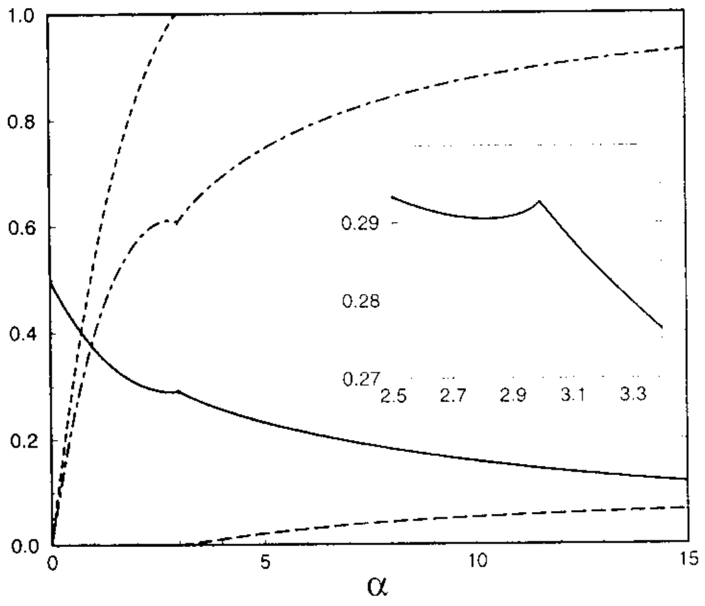

Figure 13: Generalization (full) and training error (long dashed) as well as the order parameters $R$ (dashed dotted) and $q$ (dashed) as functions of the training set size $\alpha$ for $T = 0$ Gibbs-learning in the presence of output noise of strength $a = .8$. The vertical dotted line marks the critical training set size $\alpha_c$. The inset shows an enlarged part of the $\varepsilon(\alpha)$-curve around $\alpha_c$ displaying the overfitting preceeding criticality.

It is finally interesting to investigate the asymptotic behaviour of the error measures for $\alpha \to \infty$. Remarkably for all kinds of noise considered an asymptotic analysis of the saddle point equations gives $R \to 1$ implying $\varepsilon \to 0$ for $\alpha \to \infty$. This means that the student is able to perfectly decipher the rule behind the examples even if these are corrupted by random noise. Using the asymptotic behaviour of the order parameters in the expression for the training and prediction error one finds that both converge to the respective residual error for large training set sizes as expected.

The detailed dependence of the generalization error for large $\alpha$ is, however, different for weight or input noise on the one hand and output noise on the other hand. In the former case one finds from (137) and (138)

$$\varepsilon \sim \sqrt{\frac{2}{3\pi}} \left( \frac{1-\gamma^2}{\gamma^2} \right)^{\frac{1}{4}} \alpha^{-\frac{1}{4}} \tag{143}$$

which is much slower than the decay $\varepsilon \sim .625/\alpha$ found for Gibbs learning from an errorfree example set. On the other hand for output noise (140),(141) yield the asymptotic result

$$\varepsilon \sim \frac{C(a)}{\alpha} \tag{144}$$

where the dependence of the prefactor $C$ on the noise parameter $a$ has to be determined numerically. One finds $C \to .625$ in the zero noise limit $a \to 1$ and $C \to \infty$ for $a \to 0$. It is remarkable that, contrary to weight or input noise, output noise does not change the qualitative character of the asymptotic decay of the generalization error but just increases the prefactor. The reason for that is the same that makes $\alpha_c$ rather small for output noise: For large training sets the corrupted examples are typically so evidently inconsistent with the rest of the set that it is easy for the student to detect them as being "obvious nonsense". In the case of weight or input noise the situation is quite different. If $\varepsilon$ is already rather small the student needs in particular reliable classifications of examples near the decision boundary of the teacher to make further progress and these are notoriously difficult to get in the presence of input or weight noise.

From both fig.12 and 13 we can moreover realize a typical feature of zero temperature learning. Slightly below the critical training set size $\alpha_c$ the generalization error stops to decrease and even increases as can be clearly seen in the insets. This is another example of overfitting where the student tries to imitate the teacher perfectly by using, however, the wrong concept. In trying to reproduce the noisy training set exactly the student spoils his understanding of the rule. One may therefore suspect that a noisy student that also below $\alpha_c$ uses a learning rule with non-zero training error may be able to avoid overfitting and could consequently be more successful in grasping the rule hidden in the examples. Whether this is really true will be investigated in the next section.

To complete the $T = 0$ analysis we have finally to note that for $\alpha > \alpha_c$ a technical complication shows up that we have ignored so far. As turns out the replica symmetric saddle-point (60) we have been using throughout becomes unstable [16] and the results (137)-(144) have to be considered as mere approximations. From related problems it is, however, known that these approximations are quite good so we will not discuss here which modifications occur if replica symmetry breaking is included.

## 4.4 Learning with errors

In a situation where the trainings set is free of noise it is reasonable to try to reproduce *all* classifications exactly in order to approximate the target rule as well as possible. On the other hand if the training set contains mistakes we have seen that insisting on zero training error can be misleading. In this case it may be more advisable for the student to "imitate" the teacher also with respect to her errors, i.e. to use a learning rule that gives non-zero training error from the start and hence does not reproduce the training set exactly. A simple possibility to do this is given by Gibbs learning at non-zero temperature $T > 0$ as introduced in the previous lecture that characterizes the typical performance of a student perceptron with given training error $\varepsilon_t$.

Let us briefly discuss what happens in this case by using again variants of the statistical mechanics treatment. We will only consider the case of output noise explicitly. Using a straightforward generalization of (85) and (86) the order parameters follow from

$$\frac{q-R^2}{1-q} = \frac{\alpha}{\pi}\int Dt\left[\frac{1-a}{2} + a\,H(-\frac{Rt}{\sqrt{q-R^2}})\right]\frac{\exp\{-\frac{q}{1-q}t^2\}}{\left[\frac{1}{e^\beta-1} + H(-\sqrt{\frac{q}{1-q}}t)\right]^2} \tag{145}$$

and

$$\frac{R\sqrt{q-R^2}}{\sqrt{q}\sqrt{1-q}} = a\frac{\alpha}{\pi}\int Dt\,\frac{\exp\{-\frac{t^2}{2}(\frac{R^2}{q-R^2}+\frac{q}{1-q})\}}{\left[\frac{1}{e^\beta-1} + H(-\sqrt{\frac{q}{1-q}}t)\right]} \tag{146}$$

whereas the typical training error results from (87):

$$\varepsilon_t = 2\int Dt\left[\frac{1-a}{2} + a\,H(-\frac{Rt}{\sqrt{q-R^2}})\right]\frac{H(\sqrt{\frac{q}{1-q}}t)}{1+(e^\beta-1)H(-\sqrt{\frac{q}{1-q}}t)} \quad . \tag{147}$$

The results of a numerical solution of these equations are shown in fig.14. The main differences to the corresponding behaviour for $T = 0$ as displayed in fig.13 are that $q < 1$ for all values of $\alpha$ and $\varepsilon_t > 0$ from the start. Note also that there is no overfitting since $R$ and consequently $\varepsilon$ are monotonous functions of $\alpha$.

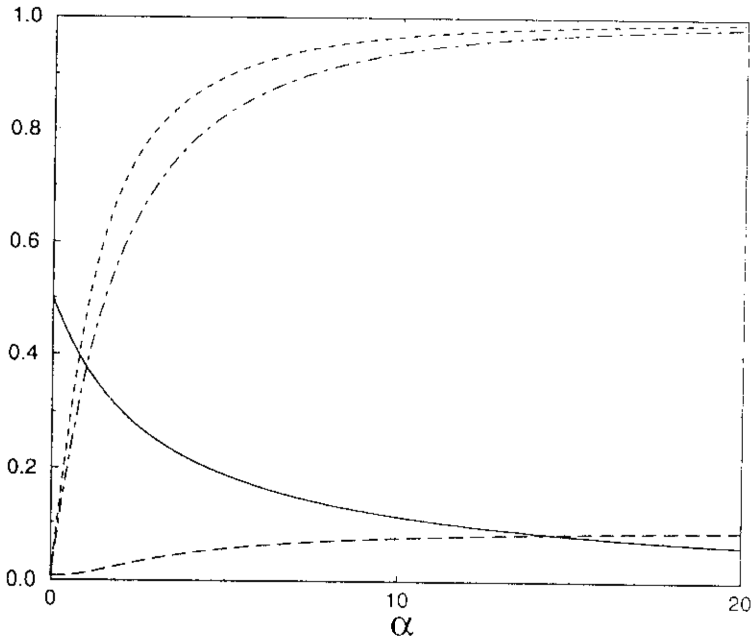

Figure 14: Generalization (full) and training error (long dashed) as well as the order parameters $R$ (dashed dotted) and $q$ (dashed) as functions of the training set size $\alpha$ for Gibbs-learning with temperature $T = .2$ in the presence of output noise with parameter $a = .8$.

The generalization behaviour for $T > 0$ is compared with that at $T = 0$ in fig.15. It is clearly seen that in the presence of noise in the training data a non-zero temperature in the learning rule is advantageous for learning from examples. A larger training error allows a smaller generalization error since it enables the student to ignore those classifications of the training set which are probably due to the noise and therefore misleading in his task to reproduce the pure target rule. The difference between $\varepsilon$ for $T = 0$ and $T > 0$ displayed in fig.15 is asymptotically due to different prefactors of the $1/\alpha$-behaviour since even in the absence of noise the same asymptotics holds.

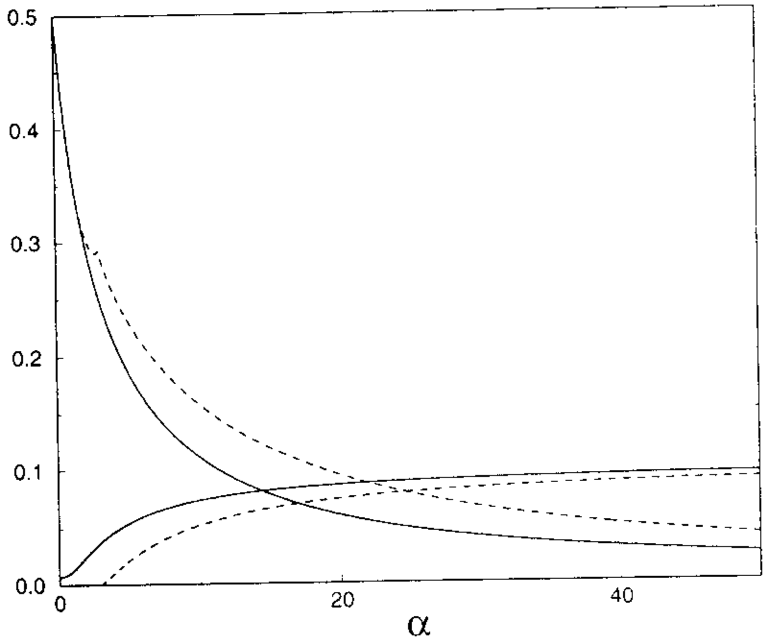

Figure 15: Generalization and training error as functions of the training set size $\alpha$ for $T = 0$ (dashed) and $T = .2$ (full) Gibbs-learning in the presence of output noise with parameter $a = .8$.

Remarkably also in the case of input and weight noise Gibbs learning with general temperature $T > 0$ does not qualitatively alter the, in this case very slow, asymptotic decrease of the generalization error. Here a more substantial improvement would have been desirable. In the next section we will investigate whether more can be gained by *tuning* the training error or equivalently the learning temperature to the intensity of the noise present in the training set.

## 4.5 Refinements

The optimal performance of a student in learning from examples clearly depends on the overall information he gets. In the previous section we have seen that if this information includes besides the training set per se also the hint that some of the teachers classifications are unreliable he can improve by using a learning rule with non-zero training error. If in addition he also happens to know the *type* and *intensity* of the noise that corrupted the training set a further improvement is possible. The question what he can optimally gain from the available information is a typical problem of mathematical statistics and the answer is most elegantly derived in the framework of Bayesian density estimation [18, 5], cf. also the lectures of Sara Solla. Although the improvements that can be obtained are often too small to matter in practical applications the lower bounds for the generalization error that follow are of principal interest.

We will discuss the question of optimal performance here only briefly and in a rather qualitative manner. Let us consider the case of output noise. The clue lies in the relation between the order parameters $R$ and $q$ describing the similarity between teacher and student and between two students respectively. If there is no noise at all we found in lecture 2 that Gibbs learning is characterized by $q = R$. This symmetry between teacher and students is in general lost in the presence of noise. If the learning temperature of the student is very high we find that $q < R$ and the student overestimates the error rate of the teacher. On the other hand for rather low temperature we have $q > R$, the student takes the teacher too seriously and risks overfitting. It turns out that the optimal $T$ is characterized by $q = R$ in which case the training error of the student is exactly equal to the fraction of wrong clasifications in the training set. From (145) and (146) we infer that the symmetry between teacher and student is restored if

$$\beta = \ln \frac{1+a}{1-a} \quad . \tag{148}$$

This gives indeed the optimal temperature the student should choose if he knows that the training set is blurred by output noise with parameter $a$ [18]. For $a = .8$ the curve for $\varepsilon(\alpha)$ that results for the corresponding value $\beta \cong 2.197$ from the numerical solution of the saddle point equation is hardly distinguishable from the curve for $\beta = 5$ given in fig.14. Asymptotically again only the prefactor is slightly reduced.

More can be gained in the case of input or weight noise. If the student knows that the examples are corrupted by such a kind of noise he may use a modified cost function that changes the asymptotic decay of the generalization error from $\varepsilon \sim \alpha^{-1/4}$ to $\varepsilon \sim \alpha^{-1/2}$, see [5] for details.

Learning a rule from examples is hence possible also in the situation where the training set classifications are corrupted by various types of random noise. For any noise type and intensity there is a critical size $\alpha_c$ of the training set beyond which the version space becomes empty and any learning rule has a non-zero training error. Nevertheless, since the noise is uncorreletad with the examples the student is able to filter it out and hence succeeds in approximating the pure teacher to any desired accuracy.

The detailed generalization behaviour can be analyzed in the two perceptron scenario using statistical mechanics techniques. Two general types of noise impact can be distinguished. Either the local field of the teacher is disturbed (resulting from input or weight noise) giving rise to erroneous classifications of inputs near her decision boundary or the teacher outputs are simply flipped at random with a given rate (output noise) resulting in "gross errors" completely at variance with the underlying rule.

At the beginning of the generalization process, i.e. for small $\alpha$, inputs with large local field of the teacher are important and correpondingly output noise is more deteriorating. This shows up in a smaller value of $\alpha_c$ for comparable noise strengths. The asymptotic decay of the generalization error for large values of $\alpha$, on the other hand, is determined by inputs near the decision boundary of the teacher. Now input or weight noise is harder and the $1/\alpha$-decay of noise-free learning is changed into the much slower $1/\alpha^{1/4}$-behaviour in this case. Remarkably, for output noise the $1/\alpha$- law persists with just a larger prefactor.

The generalization behaviour can be improved by using learning rules with non-zero training error from the start. In fact, trying perfect learning for $\alpha < \alpha_c$ results in overfitting that can be overcome by a non-zero training error as for instance in $T$>0-Gibbs learning. The training errors allow the student to ignore those classifications of the training set that are most probably due to the noise and hence misleading in the attemp to reproduce the pure target rule. For output noise the optimal choice of the training error further reduces the prefactor of the asymptotic $\varepsilon(\alpha)$-dependence. In the case of input or weight noise even a crossover to a $1/\alpha^{1/2}$-decay is possible.

Learning from noisy data is a prototype of what is called an *unrealizable* learning problem since no student vector **J** exists that is able to reproduce all classifications. Several other unrealizable situations have been studied including a teacher with a threshold $\theta \neq 0$ or with a non-monotonous activation function (e.g. the "reversed wedge" perceptron [29]) as well as situations in which teacher and student differ in their architecture as for instance an Ising student $J_i = \pm 1$ trying to learn from a spherical teacher [17]. All these scenarios share some important features. First, merely from the definition of an unrealizable problem, there is a critical size $\alpha_c$ of the training set above which the version space is empty, i.e. $\varepsilon_t(\alpha) > 0$ if $\alpha > \alpha_c$. Then, in all these cases one observes that trying perfect learning sooner or later fails by resulting in $\partial\varepsilon/\partial\alpha > 0$ meaning overfitting. Therefore in unrealizable situations it is usually advantegeous to use learning rules with non-zero training error. Finally, when analyzing an unrealizable learning problem in the statistical mechanics framework one should be aware of the fact that replica symmetry breaking is likely to occur.

Besides trying to imitate a non-realizable target functions as well as possible it may also be desirable to simply *detect* that a problem is non-realizable. For a perceptron this is done by an algorithm introduced in [20] which learns a target function if possible and indicates non-linear-seperability otherwise.

There is an important extreme case of learning from a noisy source as discussed above. It concerns the situation of an *extremely noisy* teacher in which the added noise is so strong that it completely dominates the teacher output. The task for the student is then to reproduce a mapping with no correlations between input and output so that the notion of a teacher actually becomes obsolete. The central question is *how many* input-output pairs can typically be implemented by an appropriate choice of the couplings **J**. This is the so-called *storage problem*. Its investigation yields a measure for the *flexibility* of the network under consideration with respect to the implementation of different mappings between input and output.

The storage problem is interesting for several reasons. Firstly, there is a historical point: in the physics community the storage properties of neural networks were discussed before emphasis was on their ability to learn from examples and several important concepts have been introduced in connection with these earlier investigations [21, 22]. Secondly, in several situations the storage problem is somewhat simpler to analyse and therefore forms a suitable starting point for the more complicated investigation of the generalization performance. Thirdly, the flexibility of a network architecture to the implementation of different input-output relations also gives useful information on its generalization behaviour [23, 24].

# 5 Lecture 5: Variations of perceptron learning

## 5.1 Discontinuous learning

The learning scenarios discussed so far were described in the framework of statistical mechanics as a continuous transformation of the balance between energy and entropy. The energy is most naturally given by the sum of the training errors of the individual inputs. For large training set sizes the training and generalization errors are rather similar and we hence get the scaling $e \sim \alpha\varepsilon$ for the energy per coupling in the limit of large $\alpha$. Being defined as the logarithm of the available phase space volume $\Omega$ the entropy measures the diversity of different couplings that realize the same training error. For a system with *continuous* couplings $\Omega$ is a $N$-dimensional volume with linear extension $\epsilon$ and the scaling of the entropy per coupling is hence $s \sim \ln\varepsilon$ for large $\alpha$. The balance between energy and entropy is mathematically described by the minimum condition for the free energy $f = s/\beta - e$. For small $\alpha$ those couplings with large a-priori measure dominate the version space and the entropic part in the free energy $f = s/\beta - e$ is the decisive one. With increasing $\alpha$ this balance is shifted more and more to the energetic part and for large $\alpha$ most couplings have been eliminated from the version space and only those with small values of $\epsilon$ remain. Using the above stated asymptotics of energy and entropy we find by minimizing the free energy:

$$\begin{aligned} 0 =& \frac{\partial(\beta f)}{\partial\varepsilon} = \frac{\partial}{\partial\varepsilon}(\ln\varepsilon - \alpha\beta\varepsilon) \\ =& \frac{1}{\varepsilon} - \alpha\beta \end{aligned} \tag{149}$$

resulting in the ubiquitous $1/\alpha$-decay of the generalization error $\varepsilon$ for large training set size $\alpha$. Different learning rules just give different prefactors to this aymptotic law.

If the coupling space of the network under consideration is, however, not continuous, the behaviour of the entropy for small generalization error can be quite different. An instructive example is given by the so-called *Ising perceptron* the coupling components $J_i$ of which can only assume the values $\pm 1$. Simple combinatorical arguments show that the entropy of couplings $\mathbf{J}$ that realize a certain overlap $R$ with a teacher vector $\mathbf{T}$ of the same type is given by:

$$s(R) = -\frac{1+R}{2}\ln\frac{1+R}{2} - \frac{1-R}{2}\ln\frac{1-R}{2} \tag{150}$$

which by using $\varepsilon = \arccos R/\pi$ gives rise to

$$s(\varepsilon) \cong -\frac{\pi^2}{2}\varepsilon^2 \ln\varepsilon \tag{151}$$

for $\varepsilon \to 0$. The minimization of the free energy now results in

$$0 = -\pi^2\varepsilon\ln\varepsilon - \alpha \tag{152}$$

to be contrasted with (149). For large $\alpha$ this equation has no solution indicating that the minimum of $f$ does not lie inside the interval of allowed values of $\varepsilon$ but at the boundary. In this case a continuous decrease of the generalization error with increasing training set size is not to be expected.

It is hence interesting to investigate the teacher-student perceptron scenario for the case in which both teacher and student are Ising perceptrons. The statistical mechanics analysis is a slight variation of the procedure outlined in lecture 2 and we only quote here the main differences and the final results. The energetic part $G_E(q^{ab}, R^a)$ is the same as in the case of continuous couplings and hence again given by (51). For the entropic part one finds by replacing the integral by a sum and omitting the spherical constraint in (50)

$$G_S(\hat{q}^{ab}, \hat{R}^a) = \ln \sum_{\{J^a=\pm 1\}} \exp\{-i\sum_{a<b}\hat{q}^{ab}J^aJ^b - i\sum_a \hat{R}^a J^a\} \quad . \tag{153}$$

Assuming replica symmetry this gives rise to the following result for the quenced average of the phase space volume [25]:

$$\begin{aligned}\frac{1}{N}\langle\langle\ln\Omega(\boldsymbol{\xi}^\mu, \mathbf{T})\rangle\rangle = \operatorname*{extr}_{q,\hat{q},R,\hat{R}} \Bigg[ &\frac{\hat{q}}{2}(q-1) - \hat{R}R + \int Dz\,\ln 2\cosh(z\sqrt{\hat{q}} + \hat{R}) \\ &+2\alpha\int Dt\, H(-\frac{Rt}{\sqrt{q-R^2}})\ln H(-\sqrt{\frac{q}{1-q}}t)\Bigg] \quad .\end{aligned} \tag{154}$$

Note that contrary to the case of the spherical perceptron the conjugated order parameters cannot be eliminated analytically. Nevertheless, due to the teacher student symmetry there is again the solution $q = R$ and $\hat{q} = \hat{R}$ so that we are finally left with the two equations

$$R = \int Dz \tanh(z\sqrt{\hat{R}} + \hat{R}) \tag{155}$$

$$\hat{R}\sqrt{1-R} = \frac{\alpha}{\pi}\int Dt \frac{\exp(-\frac{Rt^2}{2})}{H(\sqrt{R}t)} \quad . \tag{156}$$

The results of the numerical solution of these equations are shown in figure 16.

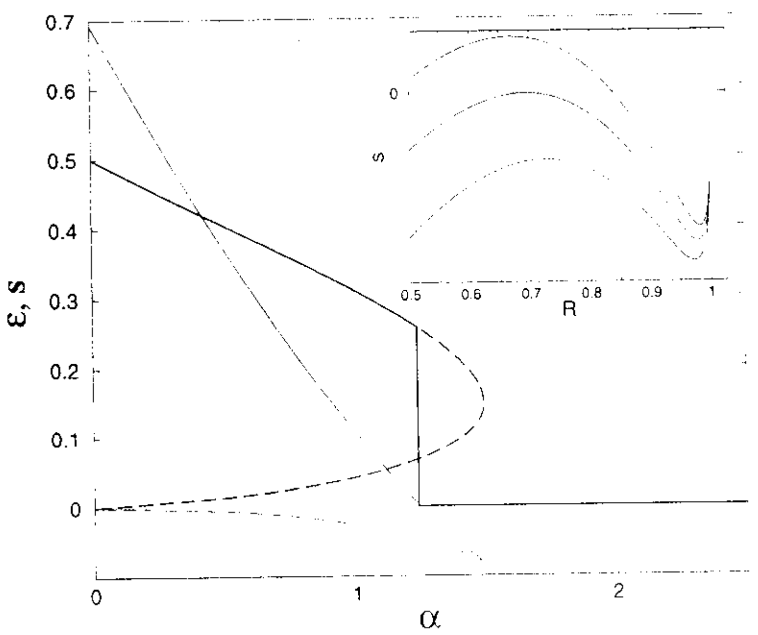

Figure 16: Generalization error (thick line) and quenched entropy (thin line) as a function of the training set size $\alpha$ for an Ising student perceptron learning a classification provided by an Ising teacher. The dashed parts of the lines correspond to unstable solutions. The inset shows the quenched entropy as a function of the teacher-student overlap $R$ for $\alpha = 1.2, \alpha_c = 1.245$, and $\alpha = 1.3$.

The most remarkable feature is that at a finite value $\alpha_p \cong 1.245$ of $\alpha$ there is a *discontinuous* transition to the state of *perfect generalization* characterized by $R = 1, \varepsilon = 0$. Hence, if a large Ising perceptron classifies $1.245N$ random inputs in the same way as a teacher perceptron of the same architecture it will do so (with probability 1) for *all* the $2^N$ possible inputs. Technically speaking this is due to a *first order phase transition* as is apparent from the behaviour of the quenched entropy as a function of the overlap shown in the inset of fig.16. The couplings dominating the version space are those with overlap $R$ maximizing the entropy. As is clearly seen this optimal value of $R$ jumps at $\alpha = \alpha_p$ from $R \cong .0695$ to $R = 1$. Note that this transition occurs exactly at the point where the entropy becomes zero. In fact contrary to a sytem with continuous phase space a system with discrete degrees of freedom can never have a negative entropy (cf. (151)). At the point where the entropy of the replica symmetric solution turns to become negative there must hence be a transition to another state for which the entropy is equal or larger than zero.

Discontinuous learning of the described type is as a rule always present when at least some of the adjustable parameters have a discrete nature. Another interesting example is given by multilayer networks where the binary internal representations serve as discrete degrees of freedom and may result in a cascade of first order transitions in the generalization behaviour [26]. Note, however, that these discontinuous transitions are not a trivial consequence of the discrete nature of the phase space: The minimal non-zero overlap occuring in the Ising scenario is $R \cong .0695$ and therefore much larger than the minimal possible non-zero overlap $R_{min} = 1 - 2/N$ of two Ising vectors $\mathbf{J}$ and $\mathbf{T}$.

In view of the above results discontinuous learning to perfect generalization looks extremely attractive. However, it turns out that the learning process itself, i.e. determining the appropriate coupling vector $\mathbf{J}$ on the basis of examples is an exteremely hard problem. In fact it has been shown to be a NP-complete problem of algorithmic complexity [27] which roughly means that no numerical procedure is known at present that could solve the problem in a time that only grows like a power of its size $N$. From the point of view of learning theory it is nevertheless very interesting, also because discontinuous transitions can natuarally be described as first order phase transitions in a statistical mechanics treatment but seem to be inaccessible to many of the alternative approaches to learning problems.

## 5.2 Queries

In the simple example of the first lecture we have realized that the initial efficiency of learning from examples is surprisingly high. Practically every new example cuts the version space in two almost equal halfs so that its volume decreases very quickly. On the other hand, in the later stages of the learning process when the student is already fairly aligned with the teacher most of the new examples do not convey any new information since its perpendicular hyperplane rarely cuts the, by then already rather small, version space. This can be explicitely seen from the behaviour of the information gain $I(\alpha)$ shown in fig.8: For $\alpha \to \infty$ it converges to zero. As a result the final approach to perfect generalization is only algebraic with $\varepsilon \sim 1/\alpha$. It is hence tempting to modify the learning scenario in a way that avoids the presentation of many redundant examples at the later stages and to realize in this way a more constant flow of information to the student. In fact, if every input label received from the teacher would convey the maximal possible information of one bit about the teacher the version space would decrease in size like $\sim e^{-C\alpha}$ giving rise to a generalization error decreasing likewise *exponentially* with the size of the training set.

Clearly, what has to be done is a modification of the distribution of inputs by introducing correlations with the actual state of the learning process. This on the other hand means nothing but allowing the student to *pose* special questions (*queries*) that he feels are most helpful at the particular moment. A simple idea is to choose a new example $\boldsymbol{\xi}^\mu$ at random from the subspace perpendicular to the present student vector $\mathbf{J}^\mu$ [28], i.e. the student asks for the classification of inputs he is at the moment most uncertain about.

A mathematical analysis of this idea is relatively simple for the Hebb-rule. Then

$$\mathbf{J}^\mu = \mathbf{J}^{\mu-1} + \frac{1}{\sqrt{N}}\sigma_T^\mu \boldsymbol{\xi}^\mu \tag{157}$$

and

$$\mathbf{J}^{\mu-1}\boldsymbol{\xi}^\mu = 0 \quad . \tag{158}$$

We are interested in the evolution of the overlap $\rho^\mu = \mathbf{T}\mathbf{J}^\mu/N$. Multiplying (157) with $\mathbf{T}$ we find

$$N(\rho^\mu - \rho^{\mu-1}) = \frac{1}{\sqrt{N}}|\mathbf{T}\boldsymbol{\xi}^\mu| \tag{159}$$

and by iteration

$$\frac{(\rho^\mu - \rho^{\mu-l})}{l} = \frac{1}{Nl}\sum_{i=1}^{l}\frac{1}{\sqrt{N}}|\mathbf{T}\boldsymbol{\xi}^{\mu-i}| \quad . \tag{160}$$

With the thermodynamic limit $N \to \infty$, $p \to \infty$ with $\alpha = p/N$ finite in mind, we now consider the limit $l \to \infty$, but with $l/N = d\alpha \to 0$. In this limit, $\alpha$ plays the role of a continuous time variable, with (160) prescribing the small change of $\rho^\mu = \rho(\alpha)$ during a small increment $d\alpha$ of time [7]. The r.h.s. of (160) is then a sum of a large number of independent random variables and we can replace it by the average to be taken with respect to the training examples. On the other hand, the l.h.s. is clearly equal to the derivative of $\rho$ versus time $\alpha$. We can thus write

$$\frac{d\rho}{d\alpha} = \langle\langle \frac{1}{\sqrt{N}} |\mathbf{T}\boldsymbol{\xi}^\mu| \rangle\rangle \tag{161}$$

and have finally to calculate the average on the r.h.s.

To this end we first note that in complete analogy with the above reasoning we can get an equation describing the evolution of the norm $Q = \sqrt{\mathbf{J}^2/N}$ by observing that in view of (158)

$$(\mathbf{J}^\mu)^2 = (\mathbf{J}^{\mu-1})^2 + 1 \tag{162}$$

and hence

$$\frac{d}{d\alpha} Q^2 = 1 \quad . \tag{163}$$

Using $Q(\alpha = 0) = 0$ this gives $Q = \sqrt{\alpha}$. Therefore the angle between teacher and student vector is given by $\theta^\mu = \arccos(\rho^\mu/\sqrt{\alpha})$. Since the examples are perpendicular to the student the angle between teacher and examples is $(\pi - \theta^\mu)$ and using the central limit theorem it is easy to show that $\mathbf{T}\boldsymbol{\xi}^\mu/\sqrt{N}$ is a Gaussian random variable with zero mean and variance $\sin^2\theta^\mu$. This in turn entails

$$\langle\langle \frac{1}{\sqrt{N}} |\mathbf{T}\boldsymbol{\xi}^\mu| \rangle\rangle = \sqrt{\frac{2}{\pi}} \sin\theta^\mu = \sqrt{\frac{2}{\pi}} \sqrt{1 - \frac{(\rho^\mu)^2}{\alpha}} \tag{164}$$

and we finally find the evolution equation for the overlap as

$$\frac{d\rho}{d\alpha} = \sqrt{\frac{2}{\pi}} \sqrt{1 - \frac{(\rho)^2}{\alpha}} \quad . \tag{165}$$

[7] This procedure will be discussed in detail in the lectures of Michael Biehl.

This differential equation has to be solved numerically and from the solution $\rho(\alpha)$ we can determine the generalization error as usual via $\varepsilon = \theta/\pi$. The solution shows that the generalization error is for all values of $\alpha$ smaller than for randomly choosen examples. For small $\alpha$ the asymptotic behaviour is the same in both cases, which is reasonable, for large $\alpha$ one finds that the generalization error for queries is only half that of learning from unselected examples.

It is reassuring that also in the special case of the Hebb-rule learning with queries is superiour to learning from random examples, however, the improvement found is rather modest. As discussed in the beginning of this section one would expect for a constant flow of information that the generalization error would decrease exponentially with the size of the training set. It turns out that the Hebb-rule is not able to use the advantages of query learning adaquately. Using more sophisticated learning rules results indeed in an exponential decrease of $\varepsilon$ for large $\alpha$ [29].

The above investigated query algorithm relies on the simple geometrical interpretation of perceptron classifications. In the case of more complicated learning machines it might be non-trivial to find a criterion according to which queries should be generated. In this case one can use the properties of the system itself. Consider an *ensemble* of students learning the same teacher and receiving the same inputs. Then use as queries those inputs for which there is *maximal disagreement* between the students. If the number of students gets very large this clearly produces queries that always bisect the version space. But even with only two students this algorithm of *query by committee* works already rather well and gives rise to an exponential decrease of the generalization error [30]. The price to pay is that it becomes increasingly hard to find a new query. After all the probability to find an input on which the two students disagree is for Gibbs learning proportional to the generalization error itself and hence decreases quickly with $\alpha$. Query by committee is hence suitable if it is cheap to generate new inputs but costly to receive their respective teacher classifications.

## 5.3 Easy questions first!

Complimentary to the previous section it may also be interesting to investigate the implications of correlations between the training examples and the *teacher* instead of those with the student. In particular we may ask whether it might be advantegeous for the learning process if the examples are selected by the teacher. An intuitive suggestion is that the teacher avoids examples that are "difficult" in the sense that they lie very near to her own decision boundary [31].

In this case the distribution of the examples would again deviate from the uniform one. Introducing a normalized weight function $f$ it can in general be written as

$$P(\boldsymbol{\xi}) = \frac{1}{(2\pi)^{\frac{N}{2}}}\exp(-\frac{\boldsymbol{\xi}^2}{2})f(\frac{\mathbf{T}\boldsymbol{\xi}}{\sqrt{N}}) \tag{166}$$

and the only modification of the calculations performed in lecture 2 is that the distribution of the teacher alignment $u = \mathbf{T}\boldsymbol{\xi}/\sqrt{N}$ is now of the form

$$P(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{u^2}{2})f(u) \tag{167}$$

instead of the simple Gaussian for $f = 1$. Proceeding as there we hence end up with the following expression for the quenched entropy:

$$s = \operatorname*{extr}_{q,R}\left[\frac{1}{2}\ln(1-q) + \frac{q-R^2}{2(1-q)} + 2\alpha\int_0^\infty Du\, f(u)\int Dt\ln H(-\sqrt{\frac{q}{1-q}}t)\right] \quad . \tag{168}$$

Assume now that the teacher avoids examples in the training that seem too difficult to her, i.e. for which $u \leq u_{min}$. This correspondes to avoiding a fraction $1 - 2H(u_{min})$ of randomly generated examples. Then

$$f(u) = \frac{\theta(|u| - u_{min})}{2H(u_{min})} \tag{169}$$

and solving the saddle-point equations for the order parameters $R$ and $q$ corresponding to (168) the performance of such a learning scenario can be analyzed. The results show that for small $\alpha$ the obverlap $R$ increases more rapidly than in the general case and hence learning gets more effective. The reason is easy to understand. For the unrestricted ensemble most examples lie in the hyperplace perpendicular to the teacher vector. Their classification cuts the version space but does not improve the alignment with the teacher. If these inputs are avoided every example carries more information on the direction of the teacher and increases the overlap $R$ with the student. On the other hand for larger values of $\alpha$ the improvement of the generalization ability is slowed down by such a kind of input selection and asymptotically one finds the extremely slow decrease $\varepsilon \sim 1/\sqrt{\ln \alpha}$. In view of our investigation of query learning this is also no surprise: At later stages of the generalization process the student needs in particular information on the classification of inputs that are near the decision boundary. i.e. exactly of those difficult questions the teacher suppresses. Note, however, that it is possible to learn the teacher perfectly also on the basis of the restricted input set only! The reason is the different scaling of $R$ and $u$. Despite the fact that $u \leq u_{min}$ for all examples there are for large $N$ always some with arbitrarily small angle $\theta = \arcsin \mathbf{T}\boldsymbol{\xi}/N$ with the decision hyperplane of the teacher.

Teaching easy questions first is hence a sensible concept in order to accelerate the learning process [8]. It helps the student in the early stages to grasp the concepts of the teacher. Later it should be abandoned in order not to mask the fine points the student needs for final improvement. This concept can, e.g. be implemented by using a varying threshold $u_{min}(\alpha)$ in (169) the optimal dependence on $\alpha$ can be determined variationally [31].

## 5.4 The reversed wedge perceptron

The reversed wedge perceptron classifies inputs $\boldsymbol{\xi}^{\mu}$ according to their projection on the coupling vector $\mathbf{J}$ just as the usual perceptron, however the simple sgn-function between local field and output is replaced by

$$\sigma = \operatorname{sgn}((\lambda - \gamma)\lambda(\lambda + \gamma)) \quad . \tag{170}$$

The most remarkable feature of this activation function is that it is *non-monotonous*, see fig.17. As a result inputs that are classified by $\sigma = +1$ may have either stabilities larger than $\gamma$ or in the interval $(-\gamma, 0)$ and a similar ambiguity holds, of course, for inputs classified as $-1$. There is hence an additional internal degree of freedom, which may be called *internal representation* for each input, specifying the internal realization of the classification.

The reversed wedge perceptron looks rather artificial and in fact there seems to be no biological indication that neurons with non-monotonous activation functions exist. However the occurrence of an internal representation, which are the trademark of multilayer networks, gives rise to some new and interesting storage and generalization properties that are worthwhile to study in some detail. Because the reversed wedge perceptron is on the other hand

[8] cf. also the organization of these lectures!

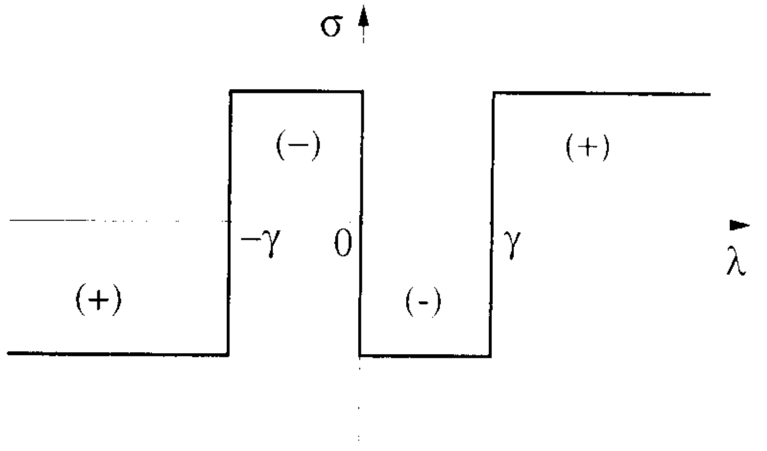

Figure 17: Non-monotonous activation function of the reversed wedge perceptron. There are two types of inputs that realize positive output, those with local field larger than $\gamma$ (internal representation "+") and those with local field between $-\gamma$ and zero (internal representation "-")

not much more difficult to analyze than the simple perceptron it may serve as a nice toy model for the more complex multilayer nets.

A straightforward investigation of the storage problem for a reversed wedge perceptron reveals that due to the possibility of different internal representations the Gardner volume is not necessarily connected and hence replica symmetry breaking is necessary for a reliable calculation of the storage capacity [32].

As for the generalization abilities we will only consider the simple realizable case where a reversed wedge perceptron with coupling vector **J** is trying to infer the coupling vector **T** of another reversed wedge perceptron with the same wedge parameter $\gamma$ from labeled examples. Using the statistical properties (22) of the local fields of teacher and student we find for the dependence of the generalization error $\varepsilon$ on the teacher-student overlap $R$ now

$$\varepsilon = 2\left(\int_{-\gamma}^{0} Du + \int_{\gamma}^{\infty} Du\right)\left[H(\frac{uR+\gamma}{\sqrt{1-R^2}}) - H(\frac{uR}{\sqrt{1-R^2}}) + H(\frac{uR-\gamma}{\sqrt{1-R^2}})\right] \tag{171}$$

generalizing (5) which is recovered for both $\gamma = 0$ and $\gamma \to \infty$. Performing the statistical mechanics analysis as detailed in lecture 2 for the present case, assuming replica symmetry, which, for the generalization problem, can be shown to be reliable [33], and using the teacher-student symmetry giving rise to $R = q$ one ends up with the following expression for the quenched entropy:

$$s = \max_R \left[ \frac{1}{2} \ln(1 - R) + \frac{R}{2} + 2\alpha \int Dt H_\gamma(t) \ln H_\gamma(t) \right] \tag{172}$$

with

$$H_\gamma(t) = H(\frac{\sqrt{R}t + \gamma}{\sqrt{1 - R}}) - H(\frac{\sqrt{R}t}{\sqrt{1 - R}}) + H(\frac{\sqrt{R}t - \gamma}{\sqrt{1 - R}}) \quad . \tag{173}$$

Solving the self-consistent equation corresponding to (172) numerically one finds that for typical values of $\gamma$ there is an interval of $\alpha$-values for which two local maxima of the entropy (172) exist . This gives rise to the following

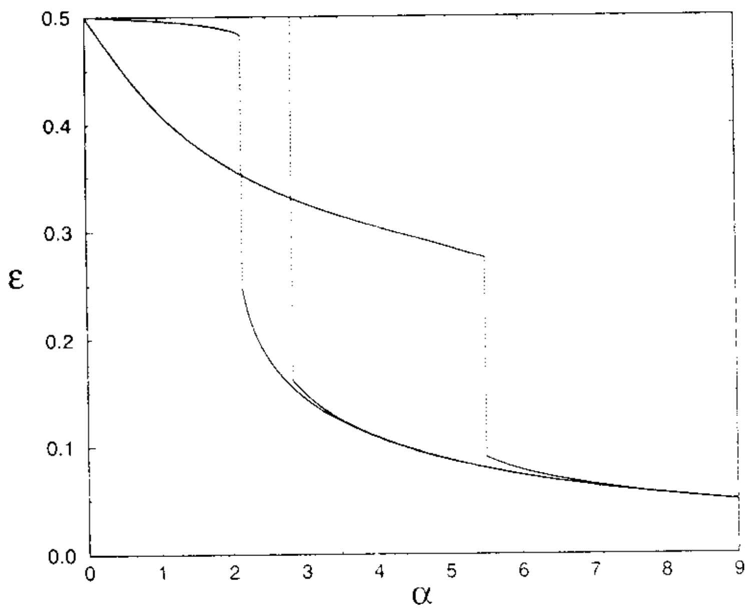

Figure 18: Generalization error of a reversed wedge perceptron learning from examples labeled by another reversed wedge perceptron with the same wedge size $\gamma$ as a function of $\alpha$ for $\gamma = 1.5$, $\gamma = \gamma_c = \sqrt{2 \ln 2} \sim 1.177$ and $\gamma = .4$ (from left to right).

generic scenario [33] cf. fig.18. For small values of $\alpha$ the system always starts in a phase with relatively poor generalizion ability in which the generalization error decays only slowly from its pure guessing value $\varepsilon = .5$. This phase is characterized by a large misfit between the internal representations of teacher and student that holds despite the agreement on the final classification of the inputs. With increasing $\alpha$ the entropy of this phase decreases more rapidly than the one of the other phase and correspondingly there is at a certain value $\alpha_d$ of $\alpha$ a *discontinuous* transition to the well generalizing phase with large similarity between the internal representations of teacher and student. The final decay of the generalization error follows than the usual $1/\alpha$-behaviour.

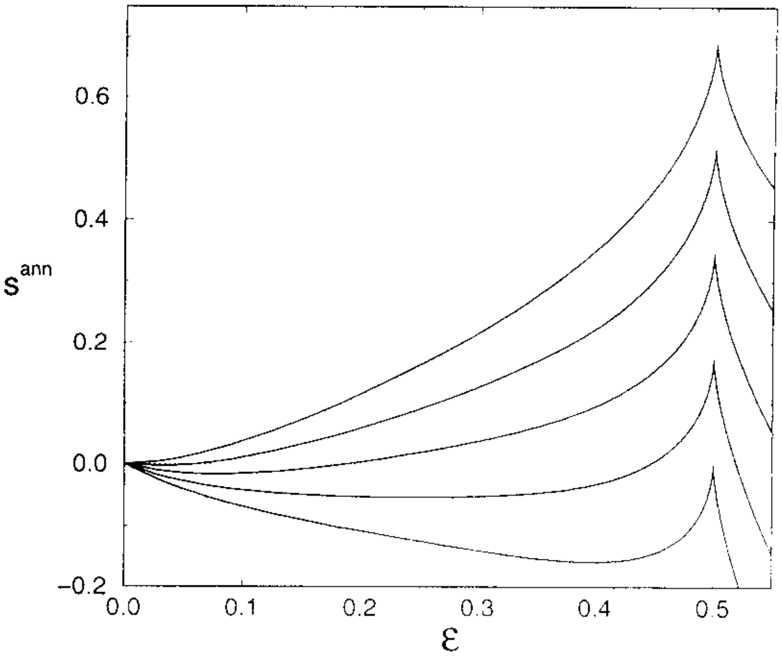

Figure 19: Annealed entropy of an Ising reversed wedge perceptron as a function of the generalization error $\varepsilon$ for different values of the training set size $\alpha$. The cusp at $\varepsilon = .5$ is specific for the critical wedge parameter $\gamma_c = \sqrt{2 \ln 2}$.

A special situation occurs if $\gamma = \gamma_c = \sqrt{2 \ln 2}$. In this case one finds that $R = 0$ is a solution of the saddle-point equation corresponding to (172) *for all* $\alpha$. Consequently the poorly generalizing phase is characterized by a constant value $\varepsilon = .5$ of the generalization error. On the other hand one finds from (172) that this phase is characterized by $s = -\alpha \ln 2$! In the initial learning phase the version space is hence bisected by every new example which correspondes to an optimal reduction of its volume, nevertheless the student gains no information at all about the teacher. This is again due to the large misfit in internal representations. Only later there is a sudden "Eureca" phenomenon that puts the student on the right track and starts a final stage of continuous learning.

This transition is even more dramatic in the case of reversed wedge perceptrons with Ising couplings, i.e. $T_i = \pm 1$ and $J_i = \pm 1$. One then finds that for $\gamma = \gamma_c$ the generalization error remains at its initial value $\varepsilon = .5$ for all $\alpha < 1$ and then jumps discontinuously to $\varepsilon = 0$ at $\alpha_d = 1$ [34]. There is hence a transition from learning nothing at all to perfect generalization! In fact this can already be anticipated from the annealed approximation. Fig.19 shows the annealed entropy as a function of $\varepsilon$ for this situation (cf. also fig.7). The cusp at $\varepsilon = .5$ arises because of $\partial\varepsilon/\partial R(R = 0, \gamma = \gamma_c) = 0$ as can be easily verified from (171). This ensures that $s$ has always a local maximum at $\varepsilon = .5$. Hence student vectors with $R = 0$ dominate the phase space up to the point $\alpha_d = 1$ at which $s(R = 0)$ becomes negative and the transisiton to perfect generalization ($\varepsilon = 0$) occurs. Note that $\alpha_d = 1$ is in fact the smallest possible threshold for which a transition to perfect generalization is possible: the teacher has $N$ unknown binary couplings and hence at least $N$ bits are necessary to pin them down. It is remarkable that the reversed wedge Ising perceptron can saturate this information theoretical bound.

# References

[1] S. Patarnello. and P. Carnevali. Europhys. Lett. **4**. 503 (1987)

[2] T. J. Sejnowski and C. R. Rosenberg. Complex Systems **1**. 145 (1987)

[3] D. E. Rummelhart and J. E. McClelland (eds.) *Parallel Distributed Processing*. (MIT Press. Cambridge MA. 1986)

[4] E. Levin. N. Tishby. and S. Solla. in *Proceedings of the 2nd workshop on Computational Learning Theory* (Morgan Kaufmann. San Mateo, 1989)

[5] M. Opper and W. Kinzel. in *Models of Neural Networks III*. E. Domany, J. L. van Hemmen, K. Schulten (eds.). (Springer. New York, 1996)

[6] E. W. Montroll. and M. F. Shlesinger. J. Stat. Phys. **32**. 209 (1983)

[7] K. Binder. and A. P. Young. Rev. Mod. Phys. **58**. 801 (1986)

[8] M. Mezard. G. Parisi. M. A. Virasoro *Spin glass theory and beyond*, (World Scientific. Singapore. 1987)

[9] M. Kac. Ark. Det. Fys. Seminar i Trondheim. **11**. 1 (1968),S. F. Edwards. in *Proc. Third Int. Conf. on Amorphous Materials 1970*, edited by R. W. Douglass and B. Ellis. (Wiley. New York. 1972), S. F. Edwards and P. W. Anderson. J. Phys. **F5**. 965 (1975)

[10] M. Bouten. J. Schietse and C. Van den Broeck. Phys. Rev. E **52**. 1958 (1995)

[11] F. Vallet. Europhys. Lett. **8**. 747 (1989)

[12] F. Vallet. J.-G. Cailton. and P. Refregier. Europhys. Lett. **9**. 315 (1989)

[13] T. L. H. Watkin. Europhys. Lett. **21**. 871 (1993)

[14] M. Opper and D. Haussler. Phys. Rev. Lett. **66** . 2677 (1991)

[15] O. Kinouchi and N. Caticha. Phys. Rev. E **54**. 8874 (1996)

[16] G. Györgyi and N. Tishby, *Workshop on Neural Networks and Spin Glasses*, K. Theumann and W. K. Koeberle (eds.), (World Scientific, Singapore, 1990)

[17] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. **A45**, 6056 (1992)

[18] M. Opper and D. Haussler, in *IVth Annual Workshop on Computational Learning Theory (COLT91), Santa Cruz, 1991* (Morgan Kaufmann, San Mateo, 1992)

[19] T. L. H. Watkin and A. Rau, Phys. Rev. **A45**, 4102 (1992)

[20] D. Nabutovsky and E. Domany, Neural Comp. **3**, 604 (1991)

[21] E. Gardner, J. Phys. **A21**, 257 (1988)

[22] E. Gardner and B. Derrida, J. Phys. **A21**, 271 (1988)

[23] A. Engel and W. Fink, J. Phys. **A26**, 6893 (1993)

[24] A. Engel, Mod. Phys. Lett. **B8**, 1683 (1994)

[25] G. Györgyi, Phys. Rev. **A41**, 7097 (1990)

[26] H. Schwarze, J. Phys. **A26**, 5781 (1993)

[27] L. Pitt and L. G. Valiant, J.ACM **35**, 965 (1988)

[28] W. Kinzel and P. Rujan, Europhys. Lett. **13**, 473 (1990)

[29] T. L. H. Watkin and A. Rau, J. Phys. **A25**, 113 (1992)

[30] S. Seung, M. Opper, and H. Sompolinsky, in *Vth Annual Workshop on Computational Learning Theory (COLT92)* (Morgan Kaufmann, San Mateo, 1993)

[31] I. Derényi, T. Geszti, and G. Györgyi, Phys. Rev. **E50**, 3192 (1994)

[32] G. Bofetta, R. Monasson, and R. Zecchina, J. Phys. **A26**, L507 (1993)

[33] A. Engel and L. Reimers, Europhys. Lett. **28**, 531 (1994)

[34] G. J. Bex, R. Serneels, and C. van den Broeck, Phys. Rev. **E51**, 6309 (1995)