Lecture Notes on Complex Systems and Physical Models
Statistical Mechanics of Artificial Neural Networks

Carlo Lucibello & Enrico M. Malatesta[1⋆]

⋆ enrico.malatesta@unibocconi.it

March 31, 2025

Contents

References: A. Engel and Van den Broeck, *Statistical Mechanics of Learning*.

1 Perceptron Models

Suppose we are given a dataset $\mathcal{D} = \left\{\xi^\mu, y^\mu\right\}_{\mu=1}^P$ composed by a set of P, N-dimensional "patterns" ξ_i^μ, $i = 1, \dots, N$ and the corresponding label y^μ. The patterns can represent whatever type of data, for example an image, text or audio. We consider in the following the binary classification setting where $y^\mu = \pm 1$.

The labels y^μ will represent some particular property of the data that we want to be able to predict, e.g. if in the image there is a cat or a dog. The goal is to learn a function $f : \mathbb{R}^N \to \pm 1$ that is able to associate to each input $\xi \in \mathbb{R}^N$ the corresponding label. This function should be able to *generalize*, i.e. to *predict* the label corresponding to a pattern not in the training set.

In the following, we will focus our discussion on a basic learning architecture and a simple generative model for the data.

1.1 Architecture

In order to fit training set, we will consider the so-called *perceptron* model, as it is the simplest, yet non-trivial, one-layer neural network that we can study using statistical mechanics tools. Given an input pattern ξ^μ the perceptron predicts a label $\hat{y}^\mu$

$$\hat{y}^\mu = \text{sign}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^N w_i \xi_i^\mu\right) \tag{1}$$

where w_i are the N parameters that need to be adjusted in order to fit the training set. The model is named differently depending on the possible values that $\boldsymbol{w}$ can assume.

If $\boldsymbol{w}$ is restricted to the vertices of the hypercube, meaning $w_i = \pm 1$, $\forall i$, the model is referred to as a *binary* perceptron. On the other hand, if the weights lie on the $N-$dimensional sphere of radius $\sqrt{N}$, i.e. $\sum_{i=1}^N w_i^2 = \boldsymbol{w} \cdot \boldsymbol{w} = N$ the model is known as the *spherical* perceptron.

1.2 Generative model of the data

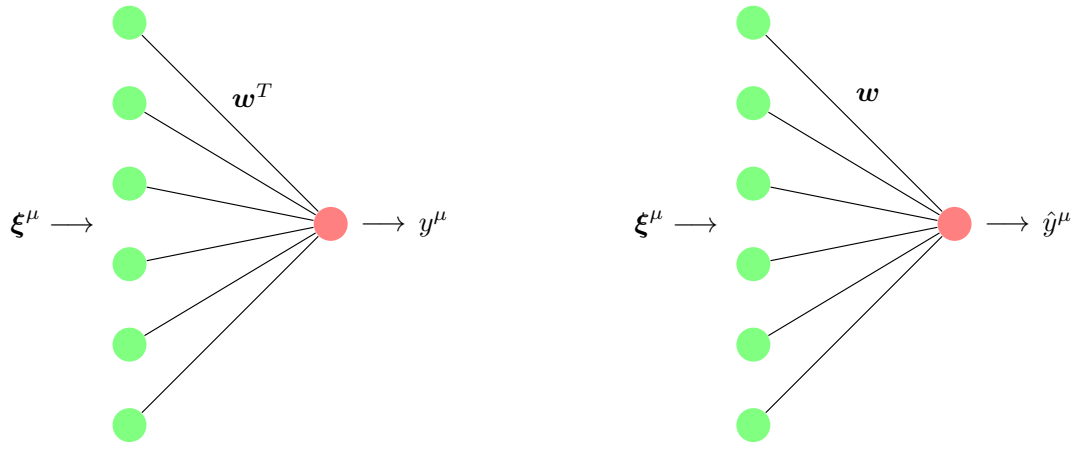

Figure 1: **Teacher-student problem**: A teacher perceptron (left) labels a set of N-dimensional inputs ξ^μ, $\mu = 1, \dots, P$. A student perceptron tries to learn the generative rule (i.e. the weights of the teacher $\boldsymbol{w}^T$) just by looking at input-output associations $\mathcal{D} = \left\{\xi^\mu, y^\mu\right\}_{\mu=1}^P$.

We consider the case of a *synthetic* dataset, where the input patterns are formed by random i.i.d. N-dimensional Gaussian $\xi_i^\mu \sim \mathcal{N}(0,1)$, $i = 1,\dots,N$. Depending on the choice of the label y^μ, we can define two different scenarios:

- in the so called *teacher-student* scenario, y^μ is generated by a network, called "teacher". The simplest case corresponds to a perceptron architecture, i.e.

$$y^\mu = \text{sign}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i^T \xi_i^\mu\right) \tag{2}$$

  with random i.i.d. Gaussian or binary weights: $w_i^T \sim \mathcal{N}(0,1)$ or $w_i^T = \pm 1$ with equal probability.

- In real scenarios, sometimes data can be corrupted or the underline rule that generates the label for a given input is noisy. This makes the problem sometimes being *unrealizable*, since, expecially for large datasets, the student architecture is not expressive enough to be able to learn the training set. The *storage problem*, describes the case of extreme noise in the dataset: y^μ is chosen to be ± 1 with equal probability. Equivalently the storage setting can be interpreted as a case in which there is an extremely noisy (and unreliable) teacher. In the storage setting, note that the label completely independent from the input data.

1.3 Large dimensional limit

We want to study the previously mentioned problems in the large dimensional limit. Namely, we will send both the number of weights N and the size of the dataset P to infinity. However, we will send them to infinity in a non-trivial fashion, by constraining the quantity

$$\alpha \equiv \frac{P}{N} \tag{3}$$

to be of $O(1)$. This means that both P and N are scaled up to infinity but their ratio is fixed to a constant value. The quantity α is referred in the following as "*constraint density*" as it basically tells us what is the ratio of constraints that our student needs to satisfy (i.e. the number of input patters to classify), compared to the number of fitting weights in the model.

2 Statistical mechanics representation

Every good statistical mechanics computation starts by writing the partition function of the model under consideration. This is the goal of this section.

Fitting the training set means that we need to satisfy the following set of constraints

$$\Delta^\mu \equiv \frac{y^\mu}{\sqrt{N}}\sum_i w_i \xi_i^\mu \geq 0\,, \qquad \mu = 1,\dots,P \tag{4}$$

indeed if the *stability* Δ^μ of example μ in the training set is positive, it means that the predicted label is equal to the true one. We will call the weights $\boldsymbol{w}$ satisfying (4) as *solutions* to the classification problem. We are now ready to write down the partition function, as

$$Z_{\mathcal{D}} = \int d\boldsymbol{w}\, \mu(\boldsymbol{w})\, \mathbb{X}_{\mathcal{D}}(\boldsymbol{w}) \tag{5}$$

where

$$\mathbb{X}_{\mathcal{D}}(\boldsymbol{w}) \equiv \prod_{\mu=1}^{P} \Theta\left[y^{\mu} \left(\frac{1}{\sqrt{N}} \sum_{i} w_i \xi_i^{\mu} \right) \right] \tag{6}$$

is an indicator function that selects a solution to the learning problem, $\Theta(x)$ being the Heaviside Theta function which is defined as

$$\Theta(x) \equiv \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{7}$$

$\mu(\boldsymbol{w})$ is a probability measure over the weights that depends on the spherical/binary nature of the problem under consideration

$$\mu(\boldsymbol{w}) = \begin{cases} \delta(\boldsymbol{w} \cdot \boldsymbol{w} - N), & \textit{spherical} \text{ case} \\ \prod_i \left[\frac{1}{2}\delta(w_i - 1) + \frac{1}{2}\delta(w_i + 1) \right], & \textit{binary} \text{ case} \end{cases} \tag{8}$$

The partition function is called also *Gardner volume* since it measures the total volume (or number in the binary case) of networks satisfying all the constraints imposed by the training set. Notice that this measure treats equally all the solutions: the Boltzmann measure (6) is indeed the flat measure among all solutions to the classification problem.

2.1 Bayesian interpretation

It is useful to interpret the partition function that we have introduced in a Bayesian way. In full generality, denote by $P_{tp}(\boldsymbol{w}^T)$ and $P_{tl}(\mathcal{D}|\boldsymbol{w}^T)$ respectively the teacher prior and likelihood which are used to generate the dataset $\mathcal{D}$. In our teacher student classification setting with the teacher being a binary perceptron we have

$$P_{tp}(\boldsymbol{w}^T) = \prod_i \left[\frac{1}{2}\delta\left(w_i^T - 1\right) + \frac{1}{2}\delta\left(w_i^T + 1\right) \right] \tag{9a}$$

$$P_{tl}(\mathcal{D}|\boldsymbol{w}^T) = \prod_{\mu=1}^{P} \delta\left(y^{\mu} - \text{sign}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i^T \xi_i^{\mu} \right) \right) \tag{9b}$$

whereas in the storage case

$$P_{tl}(\mathcal{D}|\boldsymbol{w}^T) = \prod_{\mu=1}^{P} \prod_{i=1}^{N} \left[\frac{1}{2}\delta(y^{\mu} - 1) + \frac{1}{2}\delta(y^{\mu} + 1) \right] \tag{10a}$$

Similarly denote the prior information that the student has about the teacher's weights as $P(\boldsymbol{w})$ and by $P(\mathcal{D}|\boldsymbol{w})$ the corresponding likelihood, i.e. the statistical model that the student will use to infer the labels. In both the storage and the teacher student perceptron classification settings for example we have

$$P(\boldsymbol{w}) = \prod_i \left[\frac{1}{2}\delta(w_i - 1) + \frac{1}{2}\delta(w_i + 1) \right] \tag{11a}$$

$$P(\mathcal{D}|\boldsymbol{w}) = \prod_{\mu=1}^{P} \delta\left(y^{\mu} - \text{sign}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^{\mu} \right) \right) \tag{11b}$$

Note that In our teacher student setting the student has a prior and a likelihood that *exactly* match the teacher ones, i.e.

$$P(\boldsymbol{w}) = P_{tp}(\boldsymbol{w}) \tag{12a}$$

$$P(\mathcal{D}|\boldsymbol{w}) = P_{tl}(\mathcal{D}|\boldsymbol{w}) \tag{12b}$$

but this is not true in the storage setting, as the labels are completely random.

Given the information hidden in the dataset $\mathcal{D}$, the student follows the classical strategy prescribed by Bayesian statistics, considering the posterior distribution $P(\boldsymbol{w}|\mathcal{D})$ which by Bayes theorem can be written in terms of the prior and the likelihood as

$$P(\boldsymbol{w}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{w})P(\boldsymbol{w})}{Z_{\mathcal{D}}} \tag{13}$$

$Z_{\mathcal{D}}$ is a normalization factor that is called *evidence* in Bayesian statistics and plays the same role as the partition function introduced in the previous section. Its expression is

$$Z_{\mathcal{D}} \equiv \int d\boldsymbol{w}\, P(\mathcal{D}|\boldsymbol{w})P(\boldsymbol{w}) \tag{14}$$

Indeed the prior distribution $P(\boldsymbol{w})$ plays the same role as $\mu(\boldsymbol{w})$ and $\mathbb{X}_{\mathcal{D}}(\boldsymbol{w}) = P(\mathcal{D}|\boldsymbol{w})$[1]. This tells us that the Bayesian approach is completely equivalent in spirit to the statistical mechanics one, which focuses on the analytical evaluation of partition function.

2.2 Simple geometric properties of the solution space

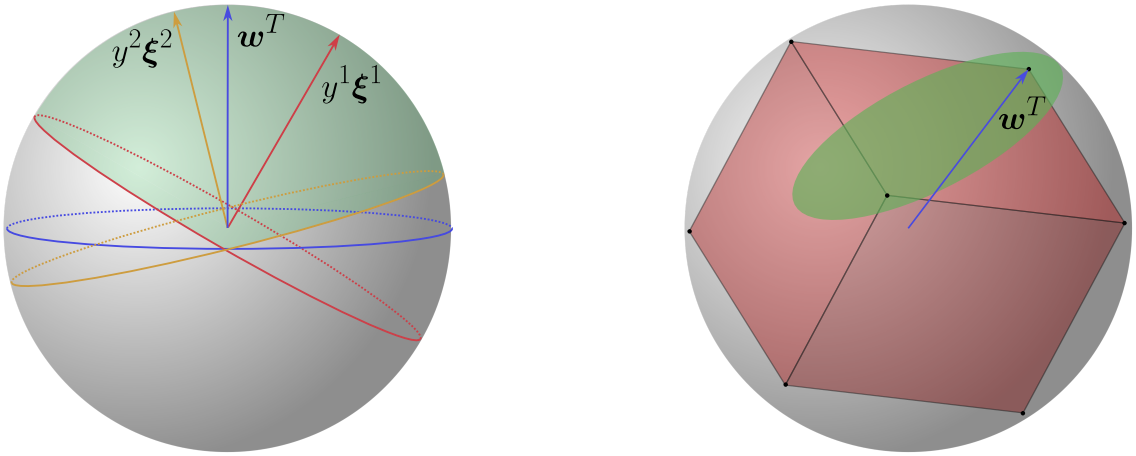

Figure 2: **Left panel**: space of solutions (green shaded area) in the spherical teacher-student perceptron. The green shaded area is obtained by taking the intersection of two spherical caps with positive overlap with the vectors $y^1\xi^1$ and $y^2\xi^2$ where y^1 and y^2 are the labels given by the teacher to the inputs ξ^1 and ξ^2. **Right panel**: binary weight case. The cube (red) is inscribed in the sphere. The green shaded area represents the (convex) space of solutions of the corresponding spherical problem. In this simple example only two vertices of the cube (among which there is of course the teacher) are inside the green region. The space of solution is therefore non-convex, since in order to go from one solution to the other, one should leave the solution space, since one is forced to pass through a vertex that is not a solution.

We want here to give a simple geometrical interpretation of the space of solutions corresponding to imposing constraints of equation (4). By those simple arguments, we will be able to unravel the convexity/non-convexity properties of the space of solutions. Without no loss of generality consider the teacher to be the vector of all ones $\boldsymbol{w}_i^T = 1$, $\forall i \in [N]$.

[1]This follows from the fact that we are considering binary labels so that $\Theta(y\hat{y}) = \delta(y - \hat{y})$.

Let's start by analyzing the *spherical case*. Initially, when no pattern has been presented, the whole volume of the N-dimensional sphere is a solution. We then extract the first pattern ξ^1 uniformly on the N-dimensional sphere; the label y^1 will be 1 if $\sum_i \xi_i^1 > 0$ and -1 otherwise. Correspondingly, the allowed solutions lie on the half-sphere with a positive dot product with $y^1\xi^1$. Of course, the same would happen if we had presented any other pattern ξ^μ of the training set. The intersection of those half-spheres form the space of solutions. Since the intersection of half spheres is a convex set, it turns out that the manifold of solutions is always convex, see left panel of Fig. 2 for an example. This tells us right away that this problem should be easy from the algorithmic point of view: since the problem is convex in nature, every non-trivial algorithms should converge in polynomial time on a weight that perfectly fits the training set.

In the right panel of Fig. 2 we show the binary case. Since $w_i = \pm 1$, the hypercube (in red) is inscribed in the sphere in N-dimension, i.e. the vertices of the hypercube are contained in the space of solutions of the corresponding spherical problem (green region). As can be readily seen from the example in the figure, the binary weight case is a non-convex problem, since in order to move from one solution to another, it is possible that one has to pass through a set of vertices not in the solution space.

In the following we will focus on the case of the binary weight problem since the non-convex nature of the problem will induce a non-trivial behavior on algorithms depending on the ratio of the dataset size P and N.

2.3 Generalization error

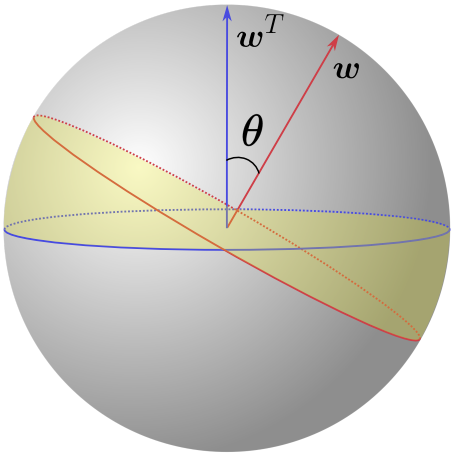

Figure 3: The decision boundaries of the teacher (blue) and of the student (red) are showed. Patterns landing in the yellow shaded area are the ones that are misclassified by the student vector $\boldsymbol{w}$.

Given a student weight $\boldsymbol{w}$, possibly fitting the training set, we want to know what would it be the corresponding generalization error, i.e. the probability that the students classifies incorrectly a *new* input pattern $\xi^\star$. Denoting by $y^\star$ the corresponding label assigned by the teacher the generalization error is

$$\epsilon_g \equiv \mathbb{E}_{\xi^\star} \Theta\left[-y^\star \hat{y}^\star\right]. \tag{15}$$

Computing the previous expectation seems an hard task. I'll leave it as an exercise for you.

However, we can get to an answer without doing so many computations by using the following argument. We fill focus for simplicity on the case of spherical weights, but the

argument can be trivially applied on the binary weight case and does not change the final answer.

Imagine to generate a new pattern $\xi^\star$; how the teacher and the student will classify it? Both of them will classify as 1 the patterns having a positive overlap with their weight vector and -1 otherwise, see Figure 3. It is straightforward to see that the test patterns $\xi^\star$ falling within the yellow shaded region correspond to inputs that are classified differently by the teacher and student. This is constructed by considering the weights in between the decision boundaries (i.e. the set of points where the classification switches) of the teacher and the student. Since test inputs are chosen randomly, the probability of such misclassification is given by the probability that $\xi^\star$ falls into this region.

Since the decision boundaries are orthogonal to the vectors $\boldsymbol{w}^T$ and $\boldsymbol{w}$, it follows that $\epsilon_g = \theta/\pi$ where θ is the angle between $\boldsymbol{w}^T$ and $\boldsymbol{w}$. This implies that the generalization error is proportional to the geodesic distance between the two points on the N-dimensional sphere that correspond to the teacher and student weights. To quantify their alignment, it is useful to introduce the so-called teacher–student overlap

$$r \equiv \frac{1}{N}\sum_{i=1}^{N} w_i w_i^T . \tag{16}$$

Since $\boldsymbol{w}$ and $\boldsymbol{w}^T$ have norm $\sqrt{N}$, we have

$$r = \frac{1}{N}\|\boldsymbol{w}\|\|\boldsymbol{w}^T\|\cos\theta = \cos\theta . \tag{17}$$

Therefore the generalization error can be written as

$$\epsilon_g = \frac{1}{\pi}\arccos r \tag{18}$$

Exercise 2.1

Show that equation (15) is equivalent to (18). Do it by explicitly averaging over standard normal test patterns $\xi^\star$.

Hint: use the central limit theorem, or extract the variables $y^\star = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} w_i^T \xi_i^\star$ *and* $\hat{y}^\star = \frac{1}{\sqrt{N}}\sum_i w_i \xi_i^\star$ *using delta functions in order to perform the average. Also use the identity*

$$2\int_0^\infty Dx\, H\left(\frac{rx}{\sqrt{1-r^2}}\right) = \frac{1}{\pi}\arccos r$$

What about the storage setting? In this case there is no notion of generalization, since there is no rule to infer. We therefore should get that $\epsilon_g = 0.5$ always. This is trivial to check mathematically. By using (18) this implies that in the storage setting $r = 0$.

2.4 The typical Gardner volume

The Gardner volume $Z_{\mathcal{D}}$ introduced in equation (5) is random variable since it explicitly depends on the dataset. The goal of statistical mechanics is to characterize the *typical*, i.e. the most probable value of this random quantity. A first guess would be to compute the averaged volume $\langle Z_{\mathcal{D}}\rangle_{\mathcal{D}}$ where we have introduced the "disorder" average notation

$$\langle \bullet \rangle_{\mathcal{D}} \equiv \mathbb{E}_{\boldsymbol{w}^T}\mathbb{E}_{\xi}\bullet, . \tag{19}$$

However, since (5) involves a product of many random contributions, the probability distribution for large N tends to be log-normal, and the most probable value of $Z_{\mathcal{D}}$ and its average do not coincide.

In probabilistic terms, the random variable $Z_{\mathcal{D}}$ does not concentrate, i.e. it does not become sharply peaked around its expected value as N grows large[2]. On the other hand, the log of the product of independent random variables is equivalent to a large sum of independent terms, that, because of the central limit theorem, is Gaussian distributed; in that case we expect that the most probable value coincides with the average. Therefore we expect that for large N

$$Z_{\mathcal{D}} \sim e^{N\phi} \tag{20}$$

where

$$\phi = \lim_{N\to\infty} \frac{1}{N} \langle \ln Z_{\mathcal{D}} \rangle_{\mathcal{D}} \,. \tag{21}$$

is the averaged log-volume. Since we are at zero training error ϕ coincides with the *entropy* of solutions. Performing the average over the log is usually called as *quenched* average in spin glass theory, to distinguish from the log of the average, which is instead called *annealed*. Annealed averages are much easier than quenched ones; even if they do not give information to the typical value of a random variable, they can still be useful, since they can give an upper bound to the quenched entropy. Indeed due to Jensen's inequality

$$\phi \leq \phi_{\text{ann}} = \lim_{N\to\infty} \frac{1}{N} \ln \langle Z_{\mathcal{D}} \rangle_{\mathcal{D}} \,. \tag{22}$$

Exercise 2.2

Generate numerically M random numbers x each being the product of N independent random numbers equally distributed between 1 and 2 with M between 10^3 and 10^6 and N between 5 and 50. Approximate the distribution of x using a histogram and analyze how the most probable value x_{mp} (the peak of the histogram) evolves as N increases comparing it with the mean value $\langle x \rangle$ and the typical value $x_{\text{typ}} \equiv e^{\langle \ln x \rangle}$. Explain why, in the limit $N \to \infty$, the most probable value x_{mp} and the typical value x_{typ} should coincide. Additionally, discuss why in numerical simulations the most probable value is always smaller than the typical value.
Hint: Compare the equations determining the most probable value of x and of $\ln x$.

3 Gardner's computation via the Replica Method

Performing the average of the log as in (21) is a *non-trivial* task! In order achieve this task, it is very useful to use the so called *replica trick*, which is based on the identity

$$\langle \ln Z_{\mathcal{D}} \rangle_{\mathcal{D}} = \lim_{n\to 0} \frac{\langle Z_{\mathcal{D}}^n \rangle_{\mathcal{D}} - 1}{n} = \lim_{n\to 0} \frac{1}{n} \ln \langle Z_{\mathcal{D}}^n \rangle_{\mathcal{D}} \,. \tag{23}$$

The *replica method* consists in performing the average over the disorder of $Z_{\mathcal{D}}^n$ considering n integer (a much easier task with respect to averaging the log of $Z_{\mathcal{D}}$), and then performing an analytic continuation of the result to $n \to 0$. It has been used firstly in spin glasses models such as the Sherrington-Kirkpatrick model (basically a Curie-Weiss model with random Gaussian interactions) and then applied by Elizabeth Gardner to our setting.

Every good replica calculation consists in 4 steps

[2] In the spin glass literature one says that $Z_{\mathcal{D}}$ is *not* a *self-averaging* quantity for large N.

1. Replicate the partition function and average over the quenched disorder;
2. introduce the order parameters;
3. write the replicated partition function in a saddle point fashion;
4. perform an ansatz over the structure of the order parameter and explicitly perform the limit $n \to 0$.

In the following subsections we will show and discuss all these steps in detail. We will consider the case of the binary teacher student model.

3.1 Step 1: disorder average

Replicating n times the partition function we have

$$Z_{\mathcal{D}}^{n} = \sum_{\{\boldsymbol{w}^a\}_{a=1}^{n}} \prod_{a=1}^{n} \prod_{\mu=1}^{P} \Theta\left[\left(\frac{1}{\sqrt{N}} \sum_i w_i^T \xi_i^\mu\right)\left(\frac{1}{\sqrt{N}} \sum_i w_i^a \xi_i^\mu\right)\right] \tag{24}$$

We now introduce the two auxiliary variables v_a^μ and u^μ which are defined as $v_a^\mu \equiv \frac{1}{\sqrt{N}} \sum_i w_i^a \xi_i^\mu$ and $u^\mu \equiv \frac{1}{\sqrt{N}} \sum_i w_i^T \xi_i^\mu$. We introduce them in the previous equation by using Dirac delta functions[3]

$$\begin{aligned} Z_{\mathcal{D}}^{n} = & \int \prod_\mu \frac{du^\mu d\hat{u}^\mu}{2\pi} e^{iu^\mu \hat{u}^\mu} \int \prod_{a\mu} \frac{dv_a^\mu d\hat{v}_a^\mu}{2\pi} e^{iv_a^\mu \hat{v}_a^\mu} \prod_{\mu a} \Theta\left(u^\mu v_a^\mu\right) \\ & \times \sum_{\{\boldsymbol{w}^a\}_{a=1}^{n}} e^{-i\sum_{\mu,a} \hat{v}_a^\mu \frac{1}{\sqrt{N}} \sum_i w_i^a \xi_i^\mu - i \sum_\mu \hat{u}^\mu \frac{1}{\sqrt{N}} \sum_i w_i^T \xi_i^\mu} \end{aligned} \tag{25}$$

where we have used the Fourier integral representation of the Dirac delta function

$$\delta(v) = \int_{-\infty}^{+\infty} \frac{d\hat{v}}{2\pi} e^{iv\hat{v}}. \tag{26}$$

Now we can perform the average over the input patters that are distributed with a Gaussian distribution with mean zero and unit variance. In the limit of large N obtaining

$$\begin{aligned} \prod_{\mu i} \left\langle e^{-i \frac{\xi_i^\mu}{\sqrt{N}} \left(\sum_a w_i^a \hat{v}_a^\mu + w_i^T \hat{u}^\mu\right)} \right\rangle_{\xi_i^\mu} &= \prod_{\mu i} e^{-\frac{1}{2N}\left(\sum_a w_i^a \hat{v}_a^\mu + w_i^T \hat{u}^\mu\right)^2} \\ &= \prod_\mu e^{-\sum_{a<b} \hat{v}_a^\mu \hat{v}_b^\mu \left(\frac{1}{N}\sum_i w_i^a w_i^b\right) - \frac{1}{2}\sum_a \left(\hat{v}_a^\mu\right)^2 - \frac{1}{2}(\hat{u}^\mu)^2 - \sum_a \hat{v}_a^\mu \hat{u}^\mu \left(\frac{1}{N}\sum_i w_i^a w_i^T\right)}. \end{aligned} \tag{27}$$

Exercise 3.1

Show that that the disorder average (27) would be unaffected in the large N limit if one considers a Rademacher distribution over input patterns. Finally, considering a generic factorized distribution over inputs $P(\xi_i^\mu)$, show that the disorder average (27) depends only on the first two moments of $P(\xi_i^\mu)$.

[3] We are basically using the property of the delta function $f(u) = \int dx\, f(x)\, \delta(x-u)$ where f is a generic scalar function.

3.2 Step 2: introduction of the order parameters

After having performed the disorder average in (27), two new quantities naturally appear

$$q_{ab} \equiv \frac{1}{N}\sum_{i=1}^{N} w_i^a w_i^b \tag{28a}$$

$$r_a \equiv \frac{1}{N}\sum_{i} w_i^a w_i^T \tag{28b}$$

Notice that even if we managed to perform the average over the input disorder, we have payed a price: now we have a coupling the replicas of the system, and this is reflected in the matrix q_{ab}. Those quantities are called *order parameters* and play the same role of the magnetization in the Ising model. As we will see later, even if q_{ab} and r_a popped out of the calculation, they actually have a deep physical meaning: q_{ab} represents the typical overlap between two replicas a and b, sampled from the Boltzmann measure corresponding to (5) and having the same realization of the training set; similarly, r_a represent the overlap of a typical replica sampled from the Boltzmann measure with the teacher $\boldsymbol{w}^T$. Due to the binary nature of the weights both those quantities are bounded in the interval $[-1, 1]$.

The next step is to enforce the definition of the q_{ab} and r_a by using delta functions and their integral representation (26). We have

$$\begin{aligned}\langle Z_{\mathcal{D}}^n \rangle_{\mathcal{D}} = \mathbb{E}_{\boldsymbol{w}^T} &\int \prod_{a<b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi/N} \prod_a \frac{dr_a d\hat{r}_a}{2\pi/N} e^{-N\sum_{a<b} q_{ab}\hat{q}_{ab} - N\sum_a r_a \hat{r}_a} \\ &\times \sum_{\{\boldsymbol{w}^a\}_{a=1}^n} e^{\sum_{a<b}\hat{q}_{ab}\sum_i w_i^a w_i^b + \sum_a \hat{r}_a \sum_i w_i^a w_i^T} \\ &\times \int \prod_a \frac{du^\mu d\hat{u}^\mu}{2\pi} \int \prod_{a\mu} \frac{dv_a^\mu d\hat{v}_a^\mu}{2\pi} \prod_{\mu a} \Theta\left(u_a^\mu v_a^\mu\right) \\ &\times e^{i\sum_{a\mu} v_a^\mu \hat{v}_a^\mu + i\sum_\mu u^\mu \hat{u}^\mu - \frac{1}{2}\sum_{ab,\mu} q_{ab}\hat{v}_a^\mu \hat{v}_\mu^b - \frac{1}{2}(\hat{u}^\mu)^2 - \sum_{a\mu} r_a \hat{v}_a^\mu \hat{u}^\mu}.\end{aligned} \tag{29}$$

Notice that we have implicitly defined $q_{aa} = 1$, $a \in [n]$.

3.3 Step 3: decoupling and saddle point method

The next step of every replica computation is to notice that terms depending on the pattern indices $\mu = 1, \dots, P$ and on the index of the weights $i = 1, \dots, N$ ("site" indices) have been decoupled (initially they were not!). We started with a partition function with coupling between site and input patterns indices but uncoupled replicas, but after averaging over the input disorder and introducing the order parameter we ended up in having no coupling over pattern and input indices but a coupling between the replicas that is mediated by the order parameter q_{ab} itself.

Setting $P = \alpha N$ we reach the following integral representation of the averaged replicated partition function

$$\left\langle Z_{\mathcal{D}}^n \right\rangle_{\mathcal{D}} \propto \int \prod_{a<b} \frac{dq_{ab}d\hat{q}_{ab}}{2\pi} \prod_a \frac{dr_a d\hat{r}_a}{2\pi} e^{NS(q,\hat{q},r,\hat{r})} \tag{30}$$

where we have defined

$$S(q,\hat{q},r,\hat{r}) = G_S(q,\hat{q},r,\hat{r}) + \alpha G_E(q,r) \tag{31a}$$

$$G_S = -\frac{1}{2}\sum_{a\neq b} q_{ab}\hat{q}_{ab} - \sum_a r_a \hat{r}_a + \ln \mathbb{E}_{w^T} \sum_{\{w^a\}} e^{\frac{1}{2}\sum_{a\neq b}\hat{q}_{ab}w^a w^b + w^T\sum_a \hat{r}_a w^a} \tag{31b}$$

$$G_E = \ln \int \frac{du d\hat{u}}{2\pi} \prod_a \frac{dv_a d\hat{v}_a}{2\pi} \prod_a \Theta(uv_a)\, e^{i\sum_a v_a\hat{v}_a + iu\hat{u} - \frac{1}{2}\sum_{ab} q_{ab}\hat{v}_a\hat{v}_b - \frac{\hat{u}^2}{2} - \hat{u}\sum_a \hat{v}_a r_a}\,. \tag{31c}$$

G_S is the so called "entropic" term, since it represents the logarithm of the volume at $\alpha = 0$, where there are no constraints induced by the training set. G_E is instead called the "energetic" term and it represents the logarithm of the fraction of solutions. In the energetic term we have also used the fact that, because of equation (28a), $q_{aa} = 1$.

Expression (30) can be evaluated by using a saddle point approximation, since we are interested in a regime where N is large. The saddle point have to be found by finding two $n \times n$ matrices q_{ab} and $\hat{q}_{ab}$ that maximize the action S. This gives access to the entropy ϕ of solutions

$$\phi = \lim_{N\to\infty} \frac{1}{N}\langle \ln Z_{\mathcal{D}}\rangle_{\mathcal{D}} = \lim_{n\to 0}\frac{1}{n}\max_{\{q,\hat{q},r,\hat{r}\}} S(q,\hat{q},r,\hat{r})\,. \tag{32}$$

3.4 Step 4: ansatz on the structure of the order parameters

Finding the solution to the maximization procedure is not a trivial task. One proceeds by formulating a simple guess or an *ansatz* on the structure of the saddle points. The simplest ansatz one can formulate is the Replica-Symmetric (RS) one. It corresponds to imposing the following parametrization of the order parameters

$$q_{ab} = \delta_{ab} + q(1-\delta_{ab}), \tag{33a}$$

$$\hat{q}_{ab} = \hat{q}(1-\delta_{ab}), \tag{33b}$$

$$r_a = r\,, \tag{33c}$$

$$\hat{r}_a = \hat{r} \tag{33d}$$

This ansatz is the easiest and the most natural to impose, since it corresponds to treat in the same way the replica indices. Indeed replicas have been introduced artificially in order to deal with the disorder average so one expects that the dependence on those replica indices should not affect the behavior of the system. Imposing this ansatz has also two benefits: firstly it allows us to perform the small n limit, and secondly it permits to write a set of coupled equations that allow us to find the order parameters explicitly.

Performing the small n limit, the entropy can be written as

$$\phi = \mathcal{G}_S + \alpha \mathcal{G}_E \tag{34}$$

where

$$\mathcal{G}_S \equiv \lim_{n\to 0}\frac{G_S}{n} \tag{35a}$$

$$\mathcal{G}_E \equiv \lim_{n\to 0}\frac{G_E}{n} \tag{35b}$$

In the following we are going to derive explicitly the expression of $\mathcal{G}_S$ and $\mathcal{G}_E$.

3.4.1 Entropic term

Let's analyze the entropic term (31b). The hard term to treat is the one inside the logarithm

$$e^{\frac{1}{2}\sum_{a\neq b}\hat{q}_{ab}w^a w^b + w^T\sum_a \hat{r}_a w^a} = e^{\frac{\hat{q}}{2}\sum_{a\neq b} w^a w^b + w^T\hat{r}\sum_a \hat{w}^a} = e^{-\frac{n\hat{q}}{2}+\frac{\hat{q}}{2}\left(\sum_a w^a\right)^2 + w^T\hat{r}\sum_a \hat{w}^a} \tag{36}$$

Next as done in the Ising model case, we linearize the term quadratic in the weights by using an Hubbard-Stratonovich transformation

$$e^{\frac{b^2}{2}} = \int Dx\, e^{bx} \tag{37}$$

where $Dx \equiv \frac{e^{-x^2/2}}{\sqrt{2\pi}}$. We get

$$\begin{aligned}
\ln \mathbb{E}_{w^T}\sum_{\{w^a\}} e^{\frac{1}{2}\sum_{a\neq b}\hat{q}_{ab}w^a w^b + w^T\sum_a \hat{r}_a w^a} &= -\frac{n\hat{q}}{2} + \ln \mathbb{E}_{w^T}\int Dx \sum_{\{w^a\}} e^{\left(\sqrt{\hat{q}}x + w^T\hat{r}\right)\sum_a w^a} \\
&= -\frac{n\hat{q}}{2} + \ln \mathbb{E}_{w^T}\int Dx \left[2\cosh\left(\sqrt{\hat{q}}x + w^T\hat{r}\right)\right]^n \\
&\simeq -\frac{n\hat{q}}{2} + n\int Dx\, \ln\left[2\cosh\left(\sqrt{\hat{q}}x + \hat{r}\right)\right]
\end{aligned} \tag{38}$$

where in the last step we have Taylor expanded in $n \to 0$ at first order and we have also used the change of variables $x \to w^T x$ which does not change the Gaussian integration measure since w^T is a binary variable. Notably, the dependence on the teacher disappears because the cosh is an even function. We therefore find that the entropic term is

$$\mathcal{G}_S \equiv \lim_{n\to 0}\frac{G_S}{n} = -\frac{\hat{q}}{2}(1-q) - r\hat{r} + \int Dx \ln 2\cosh\left(\sqrt{\hat{q}}x + \hat{r}\right). \tag{39a}$$

Exercise 3.2

In the last step of (38) we have obtained that the average over the teacher is completely trivial; namely, we could have used from the beginning the all ones vector as teacher and nothing would have changed. Can you understand the deep reason why, just by looking to equation (6)?

3.4.2 Energetic term

Before imposing the ansatz, we can integrate over the variables $\hat{u}$ in (31c), obtaining

$$G_E = \ln\int Du \int\prod_a \frac{dv_a d\hat{v}_a}{2\pi}\prod_a \Theta(uv_a)\, e^{i\sum_a \hat{v}_a(v_a - r_a u) - \frac{1}{2}\sum_{ab}(q_{ab} - r_a r_b)\hat{v}_a\hat{v}_b}. \tag{40}$$

Imposing the RS ansatz we have

$$\begin{aligned}
G_E &= \ln\int Du \int\prod_a \frac{dv_a d\hat{v}_a}{2\pi}\prod_a \Theta(uv_a)\, e^{i\sum_a \hat{v}_a(v_a - ru) - \frac{1-r^2}{2}\sum_a \hat{v}_a^2 - \frac{q-r^2}{2}\sum_{a\neq b}\hat{v}_a\hat{v}_b} \\
&= \ln\int Du \int\prod_a \frac{dv_a d\hat{v}_a}{2\pi}\prod_a \Theta(uv_a)\, e^{i\sum_a \hat{v}_a(v_a - ru) - \frac{1-q}{2}\sum_a \hat{v}_a^2 - \frac{q-r^2}{2}\left(\sum_a \hat{v}_a\right)^2}
\end{aligned} \tag{41}$$

Next we use an Hubbard-Stratonovich transformation in order to linearize the $\hat{v}_a$ variables and decouple over the replica indices

$$\begin{aligned} G_E &= \ln \int DuDx \prod_a \left[\int \frac{dv_a d\hat{v}_a}{2\pi} \Theta(uv_a) e^{i\hat{v}_a(v_a - ru - \sqrt{q-r^2}x) - \frac{1-q}{2}\hat{v}_a^2} \right] \\ &= \ln \int DuDx \left[\int Dv\, \Theta\left[u\left(\sqrt{1-q}v + ru + \sqrt{q-r^2}x \right) \right] \right]^n \\ &\simeq n \int DuDx \ln \int Dv\, \Theta\left[u\left(\sqrt{1-q}v + ru + \sqrt{q-r^2}x \right) \right]. \end{aligned} \tag{42}$$

In the second equality we have integrated over the variables $\hat{v}_a$ and performed a change of variable over v_a so that to have a standard normal integration measure Dv. Next we perform a two dimensional change of variables over u and x

$$\begin{pmatrix} x' \\ u' \end{pmatrix} = \begin{pmatrix} \sqrt{1-\frac{r^2}{q}} & \frac{r}{\sqrt{q}} \\ -\frac{r}{\sqrt{q}} & \sqrt{1-\frac{r^2}{q}} \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \tag{43}$$

This transformation is a rotation in the $2D$ plane, since its determinant is 1. We have

$$\mathcal{G}_E = \int DuDx \ln \int Dv\, \Theta\left[\left(\frac{r}{\sqrt{q}}x + \sqrt{1-\frac{r^2}{q}}u \right) \left(\sqrt{1-q}v + \sqrt{q}x \right) \right] \tag{44}$$

Finally we can notice that in order for the argument of the Heaviside function to be non zero the quantity its argument should be positive. If for example $\frac{r}{\sqrt{q}}x + \sqrt{1-\frac{r^2}{q}}u > 0$ then $\sqrt{1-q}v + \sqrt{q}x > 0$. The other case is symmetric since the variables v, u and x are all Gaussian random variables with zero mean. The two cases have therefore same probability; this allows us to focus on the case $\frac{r}{\sqrt{q}}x + \sqrt{1-\frac{r^2}{q}}u > 0$ and then multiply the result by 2, i.e.

$$\mathcal{G}_E = 2 \int DuDx\, \Theta\left(\frac{r}{\sqrt{q}}x + \sqrt{1-\frac{r^2}{q}}u \right) \ln \int Dv\, \Theta\left(\sqrt{1-q}v + \sqrt{q}x \right). \tag{45}$$

Introducing the function

$$H(x) \equiv \int_x^\infty Dy = \frac{1}{2}\mathrm{Erfc}\left(\frac{x}{\sqrt{2}} \right) \tag{46}$$

we can write the energetic term as

$$\mathcal{G}_E = 2 \int Dx\, H\left(\frac{rx}{\sqrt{q-r^2}} \right) \ln H\left(\sqrt{\frac{q}{1-q}}x \right). \tag{47}$$

3.4.3 Recap and saddle point equations

Summarizing, after this long computation we found

$$\phi(\alpha) = \underset{q,r,\hat{q},\hat{r}}{\mathrm{extr}} \left[\mathcal{G}_S(q,r,\hat{q},\hat{r}) + \alpha \mathcal{G}_E(q,r) \right] \tag{48a}$$

$$\mathcal{G}_S(q,r,\hat{q},\hat{r}) = -\frac{\hat{q}}{2}(1-q) - r\hat{r} + \int Dx \ln 2\cosh\left(\sqrt{\hat{q}}x + \hat{r} \right) \tag{48b}$$

$$\mathcal{G}_E(q,r) = 2 \int Dx\, H\left(\frac{rx}{\sqrt{q-r^2}} \right) \ln H\left(\sqrt{\frac{q}{1-q}}x \right) \tag{48c}$$

The value of the order parameters $\hat{q}$, q, r, $\hat{r}$ can be obtained by differentiating the entropy and equating it to zero.

Notice how, in the large-dimensional limit $N, P \to \infty$ with fixed $\alpha = P/N$, we have transformed the challenging problem of estimating the entropy, originally involving a sum over 2^N configurations, into a much simpler problem expressed in terms of a few macroscopic order parameters. These order parameters summarize the essential statistical properties of the microscopic degrees of freedom (i.e. the weights $\boldsymbol{w}$), allowing for a tractable analysis without explicitly handling the full complexity of the system.

Exercise 3.3

Solve analytically the extremization problem for $\alpha = 0$ and compute the corresponding entropy. Can you geometrically interpret the result?

Exercise 3.4

Compute the annealed entropy. Show that

$$\phi_{\text{ann}}(\alpha) = H_2(r) + \alpha \ln\left(1 - \frac{1}{\pi}\arccos r\right)$$

where $H_2(r) \equiv -\frac{1+r}{2}\ln\left(\frac{1+r}{2}\right) - \frac{1-r}{2}\ln\left(\frac{1-r}{2}\right)$ is the binary entropy function and r is obtained by solving the following self-consistent equation

$$r = \tanh\left(\frac{\alpha}{\pi - \arccos r}\frac{1}{\sqrt{1-r^2}}\right)$$

Hint: In order to match the results, you should use the identities

$$2\int_0^\infty Dx\, H\left(\frac{rx}{\sqrt{1-r^2}}\right) = \frac{1}{\pi}\arccos r$$

$$\operatorname{arctanh}(x) = \frac{1}{2}\ln\left(\frac{1+x}{1-x}\right)$$

3.5 Order parameters and geometrical informations

The order parameters r^a and q^{ab} popped out in the calculation naturally, right after we performed the disorder average; we have introduced them at the start of the step 2 of the replica calculation, by using delta functions in order to enforce their definitions as in (28). Intuitively they play the same role as the magnetization in the Ising model. We should therefore expect that their corresponding replica symmetric values have a definite and clear physical and geometrical meaning.

Let's start by interpreting what is the physical meaning of r obtained by solving the saddle point equations. r represents the typical overlap between one solution extracted from the Boltzmann measure (6) and the teacher. This can be shown by establishing the following relationship

$$r = \left\langle \frac{\sum_{\boldsymbol{w}=\pm 1}\left(\frac{1}{N}\sum_i w_i w_i^T\right)\mathbb{X}_{\mathcal{D}}(\boldsymbol{w})}{Z_{\mathcal{D}}} \right\rangle_{\mathcal{D}} \tag{49}$$

The physical meaning of q is even more intriguing. Indeed one can show that q corresponds

to

$$q = \left\langle \frac{\sum_{\boldsymbol{w}^1=\pm1}\sum_{\boldsymbol{w}^2=\pm1}\left(\frac{1}{N}\sum_i w_i^1 w_i^2\right)\mathbb{X}_{\mathcal{D}}(\boldsymbol{w}^1)\mathbb{X}_{\mathcal{D}}(\boldsymbol{w}^2)}{Z_{\mathcal{D}}^2}\right\rangle_{\mathcal{D}} \tag{50}$$

i.e. it represents the *typical* (the most probable) overlap between two solutions $\boldsymbol{w}^1$ and $\boldsymbol{w}^2$ extracted from the Boltzmann measure (6).

Solving the saddle point equations for r and q therefore not only gives access to important quantities as the generalization error (which can be computed by using (18)) or the entropy of solutions (using (48a)), but it also gives access to interesting geometrical information: it suggests how distant the solutions extracted from the Boltzmann measure (6) are from each other and with to the teacher. The distance between solutions can be obtained from the overlap using the relation

$$d = \frac{1-q}{2} \in [0,1] \tag{51}$$

This definition coincides with the *Hamming distance* between $\boldsymbol{w}^1$ and $\boldsymbol{w}^2$, since in that case d is equal to the fraction of indexes i at which the corresponding w_i^1 and w_i^2 are different.

Notice that both order parameters depend on α, so this geometrical information can be accessed for any value of the constraint density.

Exercise 3.5

Show equation (49) by using replica method.
Hint: insert the n replicas by using the relation

$$\frac{1}{Z_{\mathcal{D}}} = \lim_{n\to 0} Z_{\mathcal{D}}^{n-1}.$$

In order to prove equation (49), show that

$$\left\langle \frac{\sum_{\boldsymbol{w}=\pm1}\left(\frac{1}{N}\sum_i w_i w_i^T\right)\mathbb{X}_{\mathcal{D}}(\boldsymbol{w})}{Z_{\mathcal{D}}}\right\rangle_{\mathcal{D}} = \int \prod_{a<b}\frac{dq_{ab}d\hat{q}_{ab}}{2\pi}\prod_a \frac{dr_a d\hat{r}_a}{2\pi}\, r_1\, e^{NS(q,\hat{q},r,\hat{r})} \tag{52}$$

with $S(q,\hat{q},r,\hat{r})$ being the same as in equation (31). Apply the RS ansatz and use the saddle point method in order to arrive at the final result (49).

4 Storage case

We start here studying the storage case, where the labels are extracted randomly. How do we get this case from our teacher student formulas? If there is no teacher, there cannot be any hope that we generalize. If the labels are totally uncorrelated to the inputs, for a fresh new, test input, there is no hope to perform better than random chance: $\epsilon_g = 0.5$. This implies in this case that $r = 0$. Furthermore, the saddle point equation for $\hat{r}$ implies $\hat{r} = 0$. Indeed if we examined from the beginning the random label case the order parameter r^a would have not popped out and so there would have been no need to introduce them by using delta functions.

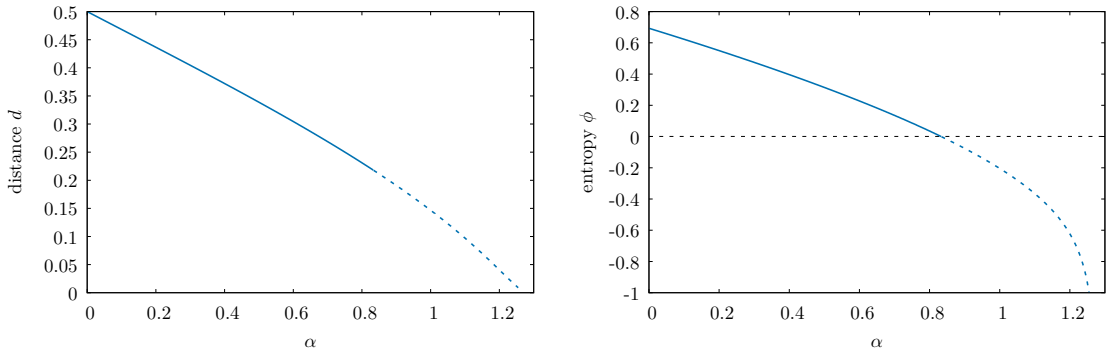

Figure 4: **Binary perceptron, random label case**. **Left**: Typical distance between solutions as a function of the constraint density α. The lines change from solid to dashed when the entropy of solutions becomes negative. **Right**: Replica Symmetric entropy as a function of α. Dashed lines show the nonphysical parts of the curves, where entropy is negative. The value of α where the entropy vanishes corresponds to the SAT/UNSAT transition $\alpha_c \simeq 0.833$.

The entropy therefore simplifies to

$$\phi(\alpha) = \underset{q,\hat{q}}{\text{extr}} \left[\mathcal{G}_S(q,\hat{q}) + \alpha \mathcal{G}_E(q)\right] \tag{53a}$$

$$\mathcal{G}_S(q,\hat{q}) = -\frac{\hat{q}}{2}(1-q) + \int Dx \ln 2\cosh\left(\sqrt{\hat{q}}x\right) \tag{53b}$$

$$\mathcal{G}_E(q,r) = \int Dx \ \ln H\left(\sqrt{\frac{q}{1-q}}x\right) \tag{53c}$$

In the binary case only the equation involving derivatives of the entropic term changes. The saddle point equations for the remaining order parameters q and $\hat{q}$ are fairly easy to write in this case and are

$$q = \int Dx \ \tanh^2\left(\sqrt{\hat{q}}x\right) \tag{54a}$$

$$\hat{q} = -2\alpha\frac{\partial \mathcal{G}_E}{\partial q} = \frac{\alpha}{1-q}\int Dx \left[GH\left(\sqrt{\frac{q}{1-q}}x\right)\right]^2 \tag{54b}$$

where $G(x) \equiv e^{-x^2/2}/\sqrt{2\pi}$ and $GH(x) \equiv G(x)/H(x)$.

The saddle point equations (54), can be easily solved numerically by simple recursion with α being an external parameter. Once they are solved numerically, we can compute the value of the entropy of solutions for a given α. The typical Hamming distance between solutions $d = (1-q)/2$ and the corresponding entropy is plotted respectively in the left and the right panel of Fig. 4 as a function of α. Note that both are decreasing functions of α: as intuition suggests, as we increase the number of constraints relative to the number of fitting parameters the total number of solutions decrease as well.

4.1 SAT/UNSAT transition

The main question we want to ask in this subsection is how many input output pairs can be typically be implemented by an appropriate choice of the weights $\boldsymbol{w}$? Indeed we should expect that if there are too many input examples, there is no way of finding the value of the $\boldsymbol{w}$ that is

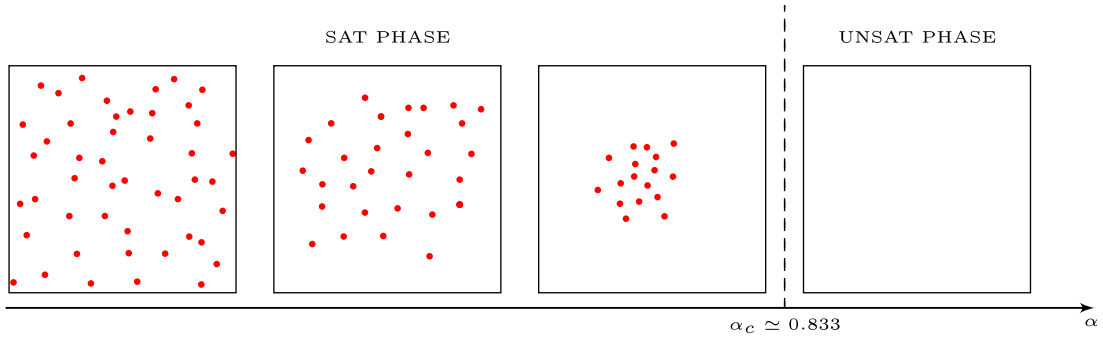

Figure 5: Pictorial view of the space of typical solutions of the binary perceptron problem, as a function of α. The squares represents an intuitive, rough picture of the space of allowed weights $\boldsymbol{w}$ for four representative values of α. Within each square, the red dots indicate typical solutions to the problem. As a function of α the typical distance between solutions decreases. Above the SAT/UNSAT transition α_c, the solutions suddenly disappear so that it is impossible to find a $\boldsymbol{w}$ that fits all the input patterns.

able to fit all the corresponding random labeling, namely the dataset will no longer be linearly separable anymore. In the high dimensional limit this transition is *sharp*, meaning that it exists a precise value of α that we will call α_c, such that for $\alpha > \alpha_c$ the probability that a solution exists is 0, whereas for $\alpha < \alpha_c$ is 1. This transition is also called "SAT/UNSAT" transition, since it separates a "satisfiable" phase $\alpha < \alpha_c$ where there are (exponentially) many possible weights that are solutions, form an "unsatisfiable" phase where none of them exists.

What is a good candidate for α_c? It is interesting to notice by looking to Figure 4 that at $\alpha = \frac{4}{\pi} \simeq 1.27$ the typical distance between solutions goes to zero, i.e. the solution space shrinks to a point as we approach it. The corresponding entropy diverges to $-\infty$ at this value of α.

However this *cannot* be the true value of $\alpha_c(\kappa)$ in the binary case! Why? If we can store 1.27 binary outputs, this suggests that we are encoding $1.27N$ bits of information within the structure of the perceptron which is specified by a set of N binary weights w_i, that is by just N bits. If we could reliably store more than N bits in the network (i.e., if $\alpha_c > 1$ were possible), we would be storing more information than the network physically has capacity for, which is not possible. Moreover, by looking at Fig. 4 in binary models the entropy cannot be negative, since we are *counting* solutions. This means that the analytical results obtained are wrong whenever $\phi < 0$[4]; for this reason in Fig. 4 the nonphysical parts of the curves are dashed.

As shown by Krauth and Mézard in 1989, in order to compute the SAT/UNSAT transition one should compute the value of α for which the *RS entropy* vanishes

$$\phi(\alpha_c) = 0\,. \tag{55}$$

This is known as the *zero entropy condition*; at this value of α_c, the distance between solutions does not go to 0 ($d \simeq 0.218$). One numerically obtains $\alpha_c = 0.833\ldots$. This value has been rigorously proved to be the correct one by mathematicians only in 2024.

[4]the problem arises because the Replica-Symmetric ansatz that we have used on the order parameters in equation (33) is not valid, but it requires the Replica Symmetry Breaking scheme introduced by G. Parisi which earned him the Nobel Prize.

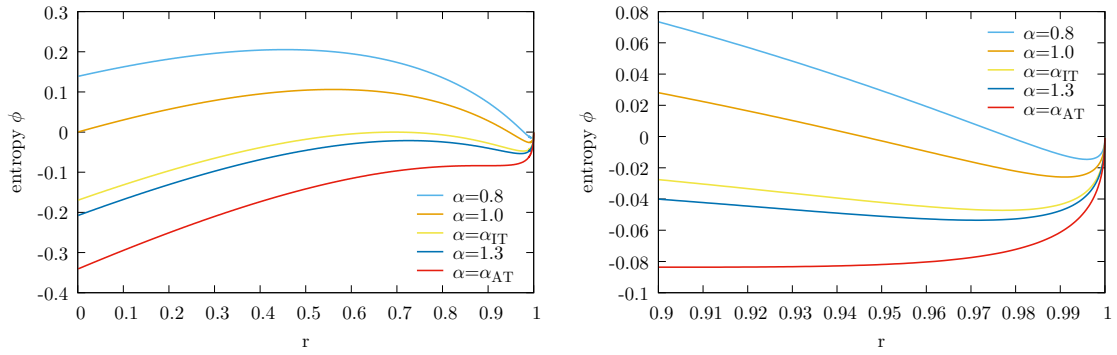

Figure 6: **Teacher-Student Binary perceptron**. **Left**: entropy as a function of the teacher student order parameter r, for several values of α. **Right**: same as left panel, but zoomed around the region $r \lesssim 1$. The global maximum of those curves represents the equilibrium value of r. Note that another point that extremizes the entropy exists near $r \sim 1$, but it is unstable because it corresponds to a minimum. $r = 1$ is always locally stable instead, as it corresponds to maximum. For $\alpha_{\text{IT}} < 1.245$ the global maximum of the free energy is at $r < 1$. For $\alpha = \alpha_{\text{IT}} \simeq 1.245$ the perfect generalization saddle point $r = 1$ becomes the global maximum of the entropy and recovering the teacher is information-theoretically possible. When $\alpha = \alpha_{\text{AT}} \simeq 1.492$ the poor generalization saddle point becomes unstable.

Exercise 4.1

By using the annealed computation (see Exercise 3.4) show that that the SAT/UNSAT transition satisfies

$$\alpha_c \leq 1$$

5 Teacher student perceptron

5.1 Bayes optimality and Nishimori condition

We here focus back on the teacher student model, where one has to solve the set of saddle point equations obtained by deriving (48a) with respect to r, $\hat{r}$, q and $\hat{q}$ in order to describe the physics of the model. Interestingly, if one tries to solve those equations, one always finds the relations

$$q = r \tag{56a}$$

$$\hat{q} = \hat{r} \tag{56b}$$

In order to understand why this simplification happens, lets go back to our Bayesian formulation of the problem, see section 2.1. Imagine to have, besides the ground truth $\boldsymbol{w}^T$, two independent samples from the posterior distribution $\boldsymbol{w}_1$, $\boldsymbol{w}_2 \sim P(\boldsymbol{w}|\mathcal{D})$. Suppose that we

want to evaluate the expectation of two functions $f(\boldsymbol{w}_1, \boldsymbol{w}_2)$ and $f(\boldsymbol{w}_1, \boldsymbol{w}^T)$. By definition

$$\begin{aligned}\mathbb{E}[f(\boldsymbol{w}_1, \boldsymbol{w}_2)] &= \mathbb{E}_{\xi,y} \int d\boldsymbol{w}_1 d\boldsymbol{w}_2 f(\boldsymbol{w}_1, \boldsymbol{w}_2) P(\boldsymbol{w}_1|\mathcal{D}) P(\boldsymbol{w}_2|\mathcal{D}) \\ &= \mathbb{E}_{\xi,y} \frac{\int d\boldsymbol{w}_1 d\boldsymbol{w}_2 f(\boldsymbol{w}_1, \boldsymbol{w}_2) P(\mathcal{D}|\boldsymbol{w}_1) P(\mathcal{D}|\boldsymbol{w}_2) P(\boldsymbol{w}_1) P(\boldsymbol{w}_2)}{\left[\int d\boldsymbol{w}\, P(\mathcal{D}|\boldsymbol{w}) P(\boldsymbol{w})\right]^2}\end{aligned} \tag{57}$$

where we have used Bayes theorem. Similarly, the expectation of $f(\boldsymbol{w}_1, \boldsymbol{w}^T)$ reads

$$\mathbb{E}\left[f(\boldsymbol{w}_1, \boldsymbol{w}^T)\right] = \mathbb{E}_{\xi,y} \frac{\int d\boldsymbol{w}_1 d\boldsymbol{w}^T f(\boldsymbol{w}_1, \boldsymbol{w}^T) P(\mathcal{D}|\boldsymbol{w}_1) P_{tl}(\mathcal{D}|\boldsymbol{w}^T) P(\boldsymbol{w}_1) P_{tp}(\boldsymbol{w}^T)}{\left[\int d\boldsymbol{w}\, P(\mathcal{D}|\boldsymbol{w}) P(\boldsymbol{w})\right]\left[\int d\boldsymbol{w}^T\, P_{tl}(\mathcal{D}|\boldsymbol{w}^T) P_{tp}(\boldsymbol{w}^T)\right]} \tag{58}$$

If we are in the so-called *Bayes optimal* case

$$P(\boldsymbol{w}) = P_{tp}(\boldsymbol{w}) \tag{59a}$$

$$P(\mathcal{D}|\boldsymbol{w}) = P_{tl}(\mathcal{D}|\boldsymbol{w}) \tag{59b}$$

i.e. the teacher gives access to the student both the prior and the likelihood distributions, then

$$\mathbb{E}[f(\boldsymbol{w}_1, \boldsymbol{w}_2)] = \mathbb{E}\left[f(\boldsymbol{w}_1, \boldsymbol{w}^T)\right] \tag{60}$$

This practically means that the students knows everything about how the data was generated, except for the teacher weights $\boldsymbol{w}^T$. As we have observed in section 2.1, this is exactly the situation in the teacher student setting we have been used so far. In this case, equation (60) implies that the teacher is entirely indistinguishable from a student drawn from the posterior distribution. As a result, any observable that depends on $\boldsymbol{w}^T$ will, on average, yield the same outcome when evaluated within a sample from the student posterior distribution. Equation (60) is called *Nishimori condition*. Note that the Nishimori condition do not apply in the storage setting: there clearly $r = 0$ but $q > 0$ for $\alpha > 0$.

Identities (56) therefore follow by a trivial application of the Nishimori condition to the function $f(\boldsymbol{w}_1, \boldsymbol{w}_2) = \frac{1}{N} \boldsymbol{w}_1 \cdot \boldsymbol{w}_2$.

5.2 Phase diagram

With the simplification given by the Nishimori conditions (56) the entropy becomes

$$\phi(\alpha) = -\frac{\hat{r}}{2}(1+r) + \int Dx \ln 2\cosh\left(\sqrt{\hat{r}}x + \hat{r}\right) + 2\alpha \int Dx\, H\left(\sqrt{\frac{r}{1-r}}x\right) \ln H\left(\sqrt{\frac{r}{1-r}}x\right) \tag{61}$$

with r and $\hat{r}$ satisfying the saddle point equations

$$r = \int Dz\, \tanh\left(\sqrt{\hat{r}}z + \hat{r}\right) \tag{62a}$$

$$\hat{r} = \frac{\alpha}{\pi\sqrt{1-r^2}} \int Dz\, H\left(\sqrt{\frac{r}{1+r}}z\right)^{-1} . \tag{62b}$$

respectively obtained by differentiating the entropy with respect to $\hat{r}$ and r. To find the global maximum value of the entropy, these equations needs to be solved numerically and the result has to be compared with the boundary $r = 1$. Indeed one can easily check that $r = 1$ and $\hat{r} = \infty$ is always a solution to (62) having zero entropy. In order to understand what is the correct maxima we use the following procedure. We *fix* r as an external parameter and we solve the second of equation (62) for $\hat{r}$. Then we plot entropy as a function of r for several

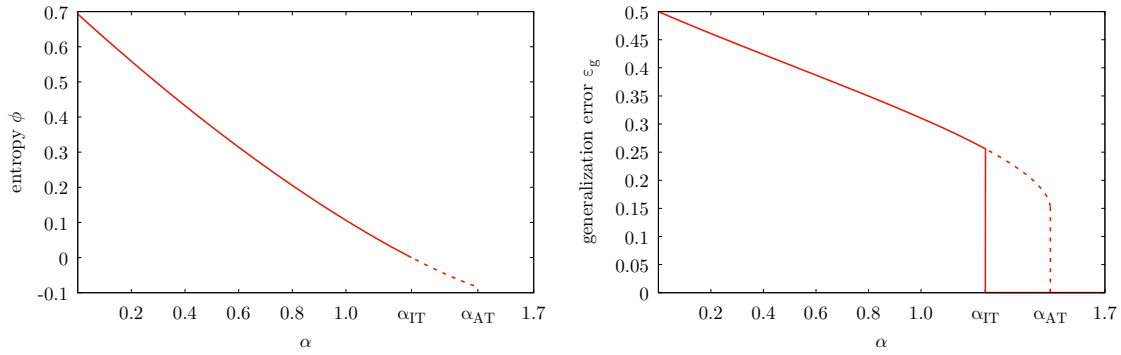

Figure 7: **Teacher-Student Binary perceptron**. Entropy and generalization as a function of α (respectively left and right panel). The dashed lines represents non-physical continuation of the curves with negative entropy. At $\alpha_{IT} \simeq 1.245$ the teacher becomes the global maximum of the entropy and the generalization jumps discontinuously to zero.

values of α. This is shown in Figure 6. As one can see a maximum of the entropy with $r < 1$ exists, but it is the global maximum only for $\alpha < \alpha_{IT}$. This tells us that for $\alpha < \alpha_{IT}$ it is *impossible* to find the teacher. This is because we have an exponential number of student weights that perfectly fit the data, and finding the teacher among them is as trying to find a needle in a haystack. For $\alpha > \alpha_{IT}$, the teacher becomes the global maximum of the entropy and therefore it is *information theoretically* possible to infer it. This change of maxima happens when the entropy of the poor generalization $r < 1$ saddle point becomes lower that the entropy of the perfect generalization maxima $r = 1$, i.e. when its entropy becomes lower than 0. Note that the saddle point $r < 1$ continue to exist still for $\alpha > \alpha_{IT}$ and can be recovered by solving *both* of (62). At $\alpha = \alpha_{AT}$ this saddle point becomes unstable as it becomes an inflection point. For $\alpha > \alpha_{AT}$ the poor generalization saddle point disappears completely. This point is called *spinodal* point or *Almeida-Thouless* transition point in the statistical mechanics literature.

Once the picture of what is the correct maximum to select, one can then easily compute the equilibrium entropy and the generalization error via (18). We show them in Figures 7. As in the storage case the entropy decreases as a function of α starting from $\phi = \ln 2$ at $\alpha = 0$ (all the vertices of the hypercube satisfy the data constraints). When $\alpha = \alpha_{IT}$ there is enough signal in the data to infer the teacher, and the generalization error drops to zero with a first order phase transition.