# Statistical Mechanics of Two-Way Recognition

Master's Thesis

Master's degree in Physics of Complex Systems

DISAT, Politecnico di Torino

***Candidate:***
Matteo Maria Rossi

***Supervisors:***
Andrea Pagnani
Jorge Fernandez de Cossio Diaz
Rémi Monasson
Simona Cocco

July 2025

Politecnico di Torino & Université Paris Cité

# *Abstract*

Master's degree in Physics of Complex Systems

## Statistical Mechanics of Two-Way Recognition

by Matteo Maria Rossi

Molecular recognition, the selective interaction between the elements of two different populations, plays a central role in regulating key processes in biology as, for example, immune responses where T cell receptors have the task of recognizing peptide antigens. Despite advances in sequencing technologies, predicting whether a TCR-antigen pair will bind remains a major challenge due to the experimental limitations in data collection. Data-driven approaches exploiting neural networks have shown to be promising in this regard, therefore understanding the amount and type of data on interacting partners necessary for generalisation is a question of great interest.
In this thesis we translate this problem into the language of statistical physics and study two different settings of data collection exploiting the well-known teacher-student model and replica trick. In the first model we demonstrate the interesting equivalence between the particular setting in which we have few orthogonal elements from one population tested against many elements from the other and the case of multiple independent perceptrons. The second model features many paired elements from both populations, we analyse the behaviour of the overlap parameters and observe how the volume of the space of interaction matrices decreases as a function of the amount of training data. These results offer theoretical insights into data requirements for learning in molecular recognition tasks and may inform future experimental strategies and designs of neural network models for biological applications.

# Contents

# Chapter 1

# Introduction

Molecular recognition between two populations refers to specific interactions between molecules of different types, which recognize and bind, or not, through various mechanisms amidst which we can cite hydrogen bonding, electrostatic interactions and hydrophobic effects. In such interactions, a molecule, belonging to one group, is recognized by a receptor or a binding partner in the other and, as schematized in Figure 1.1, recognition is often governed by structural and chemical complementarity similarly to how a key fits into a specific lock. This means that this process requires high specificity, so that the molecules bind only with a determined type of partner, ignoring the presence of other irrelevant molecules.

In biological systems such molecular interactions play a central role in regulating key processes, with recognition resulting from the specific binding between two distinct molecular species, each originating from different populations. A clear example is provided by the olfactory system, where each sensory neuron expresses a specific type of receptor capable of binding to particular odour molecules to trigger an electrical signal which is then transmitted to the brain. A similar mechanism is observed in the immune system, where B and T lymphocytes, two types of white blood cells with the role of defending the body, bind to

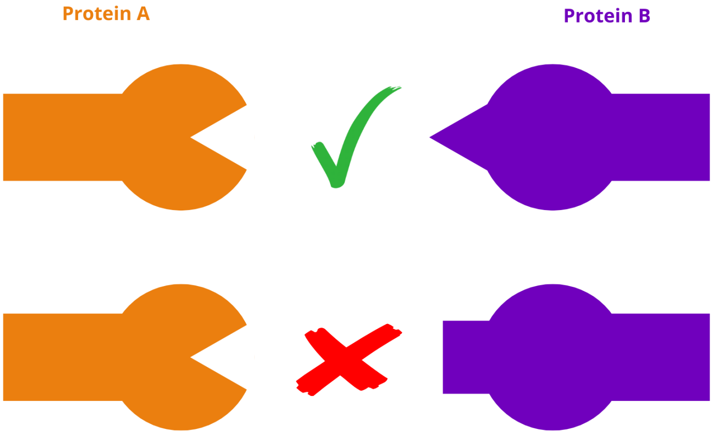


Figure 1.1: Lock-key recognition: in the top scenario the two proteins recognise each other and bind (green tick), whereas in the bottom one they mutually ignore each other (red cross).

pathogen-derived proteins to elicit an immune response and neutralize the threat.

A concrete example which serves to illustrate the general problem of selective molecular recognition involves two populations of amino acid sequences, such as proteins, that can bind to each other. As anticipated, this framework is exemplified by the interaction between T cell receptors (TCRs) and peptide antigens, where the former are proteins found on the surface of T cells and the latter are fragments of viral or bacterial proteins that appear on the surface of infected cells. When a TCR successfully binds to the target antigen, it triggers a cascade of intracellular signals to activate the T cell, ultimately leading to the elimination of the infected cell. Due to its central role in immune function, identifying which TCRs recognize and bind specific peptides constitutes a fundamental and unresolved challenge in immunology.

The main tool we have to perform this task is sequencing technology which provides access to the genetic or amino acid sequences of both the TCRs and the peptide antigens they target. Thanks to recent advances in this field, data-driven approaches, where machine learning models (particularly neural networks) are trained on datasets composed of these sequence pairs, are now a viable and valid option [4]. For instance, in [3] a specific class of neural networks, Restricted Boltzmann Machines, is used exploiting this approach to study such interactions. In general, the final goal of these models is to learn predictive patterns that generalise to unseen sequences, thereby determining whether a novel TCR-antigen pair will bind or not. Moreover, through an attentive analysis of the trained model, one hopes to deepen his understanding about the molecular determinants of the interactions, i.e. the specific features or properties of molecules that are responsible for their recognition with each other.

The ability to scale the performance to new data, in this context, to predict whether a previously unseen receptor-antigen pair will bind, is referred to as generalisation capability and is crucial for real-world applications. A schematic representation of the problem is presented in Figure 1.2. To date, very little is known about the generalisation performances of these models and the problem is further complicated by the high costs and experimental complexity involved in generating training datasets. Indeed, these datasets require extensive experimental validation, such as isolating T cells, determining receptor sequences, and measuring their binding affinities to various peptides. As a result, this constraint fuels the interest towards understanding the amount and type of data on interacting partners necessary for generalisation.

For the aforementioned reasons, in this thesis we will focus on translating this problem into statistical physics terms presenting some standard techniques, generally used to tackle problems in disordered systems and machine learning, in Chapter 2. Then, we will compute the partition functions of two different configurations of data, Chapter 3 and Chapter 4, with the aim of shedding some light on the generalisation capabilities of such models.

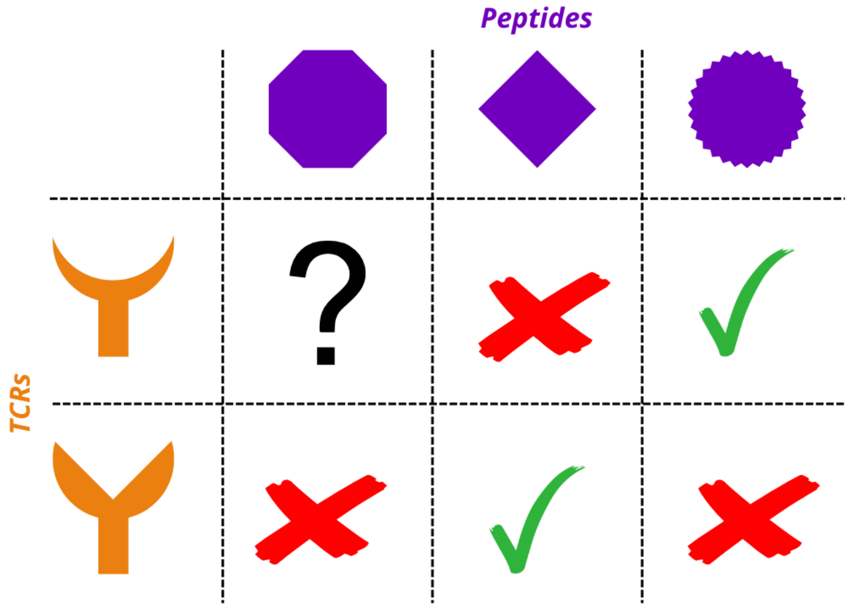


Figure 1.2: The green ticks denote situations where we know from experiments that the pair TCR-peptide binds, conversely, the red crosses indicate when they do not bind. The question mark is present whenever we do not have data concerning that particular pair and, if the model were to generalise, it would assign either a green tick or a red cross to that particular pair.

# Chapter 2

# The statistical physics framework

Neural networks are composed of many identical elementary units that interact in nontrivial ways to perform complex tasks. How such sophisticated performances emerge from the collective behaviour of individual neurons is a question that has attracted the attention of statistical physicists for several decades now. Neural networks exhibit hallmark statistical physics behaviour and statistical mechanics aims at producing exact results for their typical learning behaviour. To do so, it considers the thermodynamic limit, where both the number of degrees of freedom and the number of examples in the training set diverge but have a finite ratio. Moreover, neural networks are naturally described as a disordered system, with training data or random inter-neuron couplings playing the role of the disorder. On top of that, good performance is often separated from overfitting or unfeasibility regimes by phase transitions; hence, it is natural to use techniques and tools coming from the physics of disordered systems to tackle these problems as well. For example, pre-existing works [8, 9] treat the problem of characterizing the maximum storage capacity of a simple binary perceptron, i.e. how many patterns per input dimensions can be stored without error, in statistical physics terms also exploiting the *replica trick* which we will discuss in Section 2.3.

## 2.1 Teacher-student model

The basic scenario for learning problems in the statistical mechanics of neural networks is the so called *teacher-student model*. It considers a neural network, the *student*, which must approximate another neural network, the *teacher*, as well as possible. In principle, the architectures of the two neural networks are different, and usually the teacher's one is more complicated than the student's. The only accessible information about the target rule is contained in the training set composed of the inputs and their corresponding outputs provided by the teacher. A crucial question is how the examples of the training set are selected. The importance of this matter is given by the fact that different training sets may convey a different amount of information even when of the same size. In real-world applications the training set is determined by the experimental procedures and cannot be chosen at will; therefore, to model these situations, one assumes the examples of the training set to be selected independently at random according to some probability distribution.

The ability of the student to approximate the teacher must be quantified with some sort of similarity measure with respect to the teacher. Whether the student produces the same output as the teacher can be modelled, in the case of binary outputs, using Heaviside's step function. Let $z_T$ and $z_S$ be the output given respectively by the teacher and by the student

$$\Theta(z_T z_S) = \begin{cases} 1 & \text{student has learnt (same output as the teacher)} \\ 0 & \text{student has not learnt (different output)} \end{cases}$$

we will need to adequately adapt this expression when considering our setting, but this way of thinking about the student's performance will allow us to properly write the partition function that describes the space of interaction matrices in the following discussion.

In the scenario we are considering, we do not have two different neural networks, but what we can do is think of the real dynamics governing molecular recognition as the teacher and our neural network as the student who should learn from it. Note that this section is meant to be just a qualitative introduction to the topic in order to allow the reader to comprehend the logic and the ideas behind the work carried out in this thesis, for a more detailed discussion of this model see [1].

## 2.2 The model

To tackle the problem from a statistical physics point of view, we obviously have to translate it into mathematical terms. In the following, we will therefore present a simplified mathematical formulation of the inquiry.
Consider two sequence populations, $X$ and $Y$, that can interact with each other. For example, $X$ may refer to the space of T cell receptors and $Y$ to pathogen-derived antigens. Ideally, we want to know the binding matrix $B(\mathbf{x}, \mathbf{y})$ of any two elements $\mathbf{x} \in X$, $\mathbf{y} \in Y$, defined as:

$$B(\mathbf{x}, \mathbf{y}) = \begin{cases} +1 & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ bind} \\ -1 & \text{otherwise} \end{cases} .$$

The challenge arises when we consider the fact that the spaces $X$ and $Y$ are often huge, a detail that makes it unfeasible to measure $B(\mathbf{x}, \mathbf{y})$ for every possible pair $(\mathbf{x}, \mathbf{y})$. It is exactly in these situations that there is much interest in attempting to reconstruct this matrix from a limited number of measurements. Our goal is precisely to shed some light on when this can be done.
The usual procedure is to one-hot encode, i.e. represent using a sequence of binary inputs, the amino acid sequences as vectors of dimension proportional to the sequence length and the number of possible amino acids. Here we may choose a highly simplified model for binding between two vectors $\mathbf{x}$ and $\mathbf{y}$ where the elements $\mathbf{x} \in X$ and $\mathbf{y} \in Y$ are embedded as real vectors, $\mathbf{x} \in \mathbb{R}^N$, $\mathbf{y} \in \mathbb{R}^M$. The binding will then be described by

$$B(\mathbf{x}, \mathbf{y}) = \text{sign}\left(\sum_i \sum_j x_i W_{ij} y_j - T\right)$$

where $T$ is a real threshold, which sets the relative fractions of interacting and non-interacting partners. The interaction matrix $W$ is of size $N \times M$.

The dataset we dispose of consists of triplets $\mathcal{D} = \{(b_d, x_d, y_d)\}$ where $b_d = B(\mathbf{x}_d, \mathbf{y}_d)$ is known (from experiments) for selected pairs of vectors $(\mathbf{x}_d, \mathbf{y}_d)$. The dataset $\mathcal{D}$, due to peculiarities of the binding assay used to collect the data, often has an interesting structure itself. For example, it is often experimentally easier to obtain measurements $B(\mathbf{x}_n, \mathbf{y}_*)$ where one of the partners $\mathbf{y}_*$ is fixed and probed against many potential partners $\mathbf{x}_n$. In this scenario, the number of distinct sequences $\mathbf{y}$ and the one of distinct sequences $\mathbf{x}$ in the data are not comparable, the former being significantly smaller than the latter. Due to such peculiarities, machine learning practitioners often find that their algorithms fail to generalize to novel sequence partners not seen in training data. The question we are interested in is understanding under what conditions, i.e. kind or size of dataset $\mathcal{D}$ and dimensionalities $N$, $M$, we can draw inferences about the matrix elements $W_{ij}$, or other underlying parameters that determine binding. To try to answer this question, in this thesis we will take into consideration two particular configurations of data. The first will include a finite number of $\mathbf{y}$ vectors tested against an extensive pool of $\mathbf{x}$ vectors, whereas the second setting will feature a dataset with an extensive number of pairs $(\mathbf{x}_d, \mathbf{y}_d)$. Both problems are inspired by actual situations encountered by experimentalists on a daily basis; therefore, both are interesting and could be valuable to solve real-world challenges.

Getting more specific, we want to formulate the problem in statistical physics terms, and what we would like to do is to characterize the volume of matrices $W$ compatible with the given dataset. To obtain a mathematically well-defined problem, we constrain the matrix space with a $L_2$ regularisation (*a priori* Gaussian measure). We can then write a partition function,

$$\mathcal{Z} = \int dW e^{-\frac{1}{2}\sum_{ij} W_{ij}^2} \prod_{d\in\mathcal{D}} \Theta \left[ b_d \times \left( \sum_{ij} x_{d,i} W_{ij} y_{d,j} - T \right) \right] \tag{2.1}$$

where $\Theta$ denotes Heaviside's step function and $d \in \mathcal{D}$ are the measured triplets $(b_d, \mathbf{x}_d, \mathbf{y}_d)$ in the training dataset. As customary in a statistical physics framework, correctly determining the partition function of a system is crucial in studying macroscopic properties of the system and observing phase transitions and critical phenomena. In order to compute it, we will use some standard statistical physics tools, in particular, a major role will be played by the replica method.

## 2.3 The replica method

In this section we will provide the basic notions necessary to understand the following discussion, but an exhaustive description of the replica method can be found in [10].

Let us start by saying that the replica method is a powerful analytical technique, yet, it is not rigorous and it is used mainly for disordered systems and some problems in information theory and machine learning. In disordered systems, one often wants to compute the quenched entropy of a system computing the average of the logarithm of the partition function. Indeed, when treating non-self-averaging quantities one may want to compute their typical value, which does not coincide with their average; why we take the logarithm is going to be more clear once we delve into our specific problem in Chapter 3. However, directly averaging $\ln \mathcal{Z}$ over the disorder is usually challenging and the replica method allows to circumvent this issue by using the identity:

$$\langle \ln \mathcal{Z} \rangle = \lim_{n\to 0} \frac{\langle \mathcal{Z}^n \rangle - 1}{n}$$

allowing to compute $\langle \mathcal{Z}^n \rangle$ for integer $n$ and then analytically continue the result to $n \to 0$ (it is precisely this analytic continuation that makes the method not rigorous since it is not mathematically justified in general). This method introduces $n$ independent copies (*replicas*) of the original system, all subject to the same realisation of the disorder, and it is by studying how these replicas arrange themselves that we gain insight into the structure of the solution space. This trick is therefore accompanied by an ansatz regarding how the introduced replicas behave.

The first ansatz that comes to mind, which is the one we will use in the following discussion, is the Replica Symmetric (RS) one which assumes all replicas to behave the same, implying that the system has a relatively simple energy landscape with one large basin of attraction. Examples of this ansatz can be seen in [8, 9] where it is employed to study a single perceptron. The symmetry of this ansatz often breaks down, leading to more complicated ansätze which imply a more complicated energy landscape with different basins of attraction. In these cases we talk about replica symmetry breaking (RSB) which can assume different forms from 1-step RSB to full RSB. An example of how a similar ansatz can be used when studying neural networks can be found in [7].

As witnessed by the cited examples, the reason why this trick is so useful also in problems in machine learning can be easily understood thanks the aforementioned analogy they have with disordered systems. As a matter of fact, as presented in [12], one could observe the same duality of variables in both domains. The available data in machine learning can be identified with quenched interactions in disordered systems, whereas the parameters of a model can be seen as the spins adapting to the disordered background.

# Chapter 3

# Few-vs-many binding partners data

In this section we will focus on finding the partition function of the first setting mentioned in Section 2.2 where we are able to test a finite number of elements of family $Y$ against an extensive number of elements of $X$. In doing so, we will exploit the teacher-student model and the replica trick previously presented.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_B \in \mathbb{R}^N$ be random normal vectors with $\langle x_{i\xi}\rangle = 0$ and $\langle x_{i\xi}^2\rangle = 1$. Let $\mathbf{y}_1, \ldots, \mathbf{y}_K \in \mathbb{R}^M$ be some fixed vectors. Let $W, W^* \in \mathbb{R}^{N\times M}$ be respectively the interaction matrices of the student and the teacher.

$$\mathcal{Z} = \int \mu(dW) \prod_{\xi,k} \Theta(\mathbf{x}_\xi^\top W^* \mathbf{y}_k \mathbf{x}_\xi^\top W \mathbf{y}_k)$$

where $\mu(dW) \propto \exp\left(-\frac{\gamma}{2}\mathrm{tr}(W^2)\right)$ for some $\gamma > 0$, is a normalized Gaussian measure ($\int \mu(dW) = 1$), $\Theta$ is Heaviside's step function and the threshold $T$ in Eq.(2.1) is set to zero.

Let $J_{ik} = \sum_j W_{ij} y_{jk}$ (i.e. $\mathbf{J}_k = W\mathbf{y}_k$). Under $W \sim \mu(dW)$, $J$ is Gaussianly distributed with

$$\langle J_{ik}\rangle = 0, \quad \langle J_{ik}J_{jl}\rangle = \sum_{n,m} \langle W_{in}W_{jm}\rangle y_{nk} y_{ml} = \frac{\delta_{ij}}{\gamma}\mathbf{y}_k^\top \mathbf{y}_l$$

Denoting this Gaussian measure by $\mu(dJ)$

$$\mathcal{Z} = \int \mu(dJ) \prod_{\xi,k} \Theta(\mathbf{x}_\xi^\top \mathbf{J}_k^* \mathbf{x}_\xi^\top \mathbf{J}_k) \tag{3.1}$$

Where $J_{ik}^* = \sum_j W_{ij}^* y_{jk}$.

From Eq.(3.1) we infer that $\mathcal{Z}$ involves a product of many random contributions. Products of independent random numbers possess distributions with long tails for which the average and the most probable value do not coincide. On the other hand, the logarithm of such a quantity is a large sum of independent terms and therefore becomes normally distributed so that its average and its most probable value asymptotically coincide. Hence, the typical value of $\mathcal{Z}$, for large $N$, is given by

$$\mathcal{Z} \sim e^{\langle \ln \mathcal{Z}\rangle}$$

Using the replica trick, previously described in Section 2.3, we want to compute

$$\langle \ln \mathcal{Z}\rangle_X = \lim_{n\to 0} \frac{\langle \mathcal{Z}^n\rangle_X - 1}{n} \tag{3.2}$$

where the average is taken over the data $X = (\mathbf{x}_1, \dots, \mathbf{x}_B)$ with independent entries $x_{i\xi} \sim \mathcal{N}(0,1)$. Introducing replicas $a = 1, \dots, n$ and considering that the $\mathbf{x}_\xi$ are i.i.d. over $\xi$:

$$\begin{aligned}\langle \mathcal{Z}^n \rangle &= \left\langle \int \left[ \prod_a \mu(dJ^a) \right] \prod_{a,\xi,k} \Theta(\mathbf{x}_\xi^\top \mathbf{J}_k^* \mathbf{x}_\xi^\top \mathbf{J}_k^a) \right\rangle_X = \\ &= \int \left[ \prod_a \mu(dJ^a) \right] \left\langle \prod_{a,k} \Theta(\mathbf{x}^\top \mathbf{J}_k^* \mathbf{x}^\top \mathbf{J}_k^a) \right\rangle_{\mathbf{x}}^B\end{aligned}$$

where the average on the right is taken with respect to a single $\mathbf{x} \in \mathbb{R}^N$ with independent entries $x_i \sim \mathcal{N}(0,1)$. Next, we introduce the variables $u_k = \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^*$ and $\lambda_k^a = \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^a$:

$$\left\langle \prod_{a,k} \Theta(\mathbf{x}^\top \mathbf{J}_k^* \mathbf{x}^\top \mathbf{J}_k^a) \right\rangle_{\mathbf{x}} = \int \left( \prod_{ak} \Theta(u_k \lambda_k^a) \right) P(\mathbf{u}, \Lambda) d\mathbf{u} d\Lambda$$

where $\Lambda = (\lambda_k^a)$ and $\mathbf{u} = (u_k)$, and

$$P(\mathbf{u}, \Lambda) = \left\langle \prod_k \delta\left( u_k - \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^* \right) \prod_{a,\nu} \delta\left( \lambda_k^a - \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^a \right) \right\rangle_{\mathbf{x}}$$

is the distribution of $\mathbf{u}, \Lambda$ induced by $\mathbf{x}$. Since $u_k$ and $\lambda_k^a$ are linear combinations of the Gaussian variables $x$, they are themselves Gaussian. Therefore, $P(\mathbf{u}, \Lambda)$ is a multivariate normal distribution, with zero means and covariances:

$$\langle u_k u_l \rangle = R_{kl}^*, \quad \langle \lambda_k^a u_l \rangle = R_{kl}^a, \quad \langle \lambda_k^a \lambda_l^b \rangle = q_{kl}^{ab}$$

where we introduced the overlap parameters

$$R_{kl}^* = \frac{1}{N} (\mathbf{J}_k^*)^\top \mathbf{J}_l^*, \quad R_{kl}^a = \frac{1}{N} (\mathbf{J}_k^a)^\top \mathbf{J}_l^*, \quad q_{kl}^{ab} = \frac{1}{N} (\mathbf{J}_k^a)^\top \mathbf{J}_l^b.$$

Substituting and changing variables from the $dJ^a$ to these overlaps we get

$$\langle \mathcal{Z}^n \rangle = \int \exp\left\{ N \left( \alpha \ln \left\langle \prod_{a,k} \Theta(u_k \lambda_k^a) \right\rangle + \mathcal{S}(R,Q) \right) \right\} dQdR \tag{3.3}$$

where $\langle \dots \rangle$ denotes the Gaussian average under $u, \Lambda \sim P(u, \Lambda)$, $\alpha = \frac{B}{N}$, $Q = (q_{kl}^{ab})$ and $R = (R_{kl}^a)$ and $e^{N\mathcal{S}(R,Q)}$ is an entropic factor acting as the Jacobian for the change of variables. Note that we need $B \sim \mathcal{O}(N)$ to get a nontrivial regime because each measurement gives one inequality and since $W$ has $N$ degrees of freedom to get a significant constraint on $W$ it is reasonable that we need a number of inequalities comparable to $N$.
Up to a negligible multiplication factor

$$e^{N\mathcal{S}} \simeq \int \left( \prod_{a,k,l} d\hat{R}_{kl}^a \right) \left( \prod_{ak \le bl} d\hat{q}_{kl}^{ab} \right) \exp\left( -N \sum_{a,k,l} \hat{R}_{kl}^a R_{kl}^a - N \sum_{ak \le bl} \hat{q}_{kl}^{ab} q_{kl}^{ab} + \sum_i \ln \Omega_i \right)$$

where, defining $h_{ik}^a = \sum_l \hat{R}_{kl}^a J_{il}^*$ and $\mathbf{h}_i = (h_{ik}^a) \in \mathbb{R}^{nK}$,

$$\Omega_i = \det\left( \frac{1}{\gamma} Y^\top Y \right)^{-\frac{n}{2}} \det(A)^{-\frac{1}{2}} \exp\left( \frac{1}{2} \mathbf{h}_i^\top A^{-1} \mathbf{h}_i \right).$$

For large $N$ we can then use the saddle point method and we get

$$\begin{cases} \hat{R}^a_{kl} = \sum_{b,m,s} A^{ab}_{km} R^b_{ms} (R^*)^{-1}_{sl} \\ A^{-1} = Q - R(R^*)^{-1} R^\top \end{cases} .$$

Thus implying

$$\mathcal{S} = -\frac{1}{2} \ln\det A + \frac{n}{2} - \gamma \frac{n}{2} \sum_{k,l} \left(Y^\top Y\right)^{-1}_{kl} - \frac{n}{2} \ln\det \left( \frac{1}{\gamma} Y^\top Y \right) .$$

Let us now define some quantities:

$$\Sigma = \begin{pmatrix} (R^*_{kl})_{K\times K} & (R^a_{kl})^\top_{K\times nK} \\ (R^a_{kl})_{nK\times K} & \left(q^{ab}_{kl}\right)_{nK\times nK} \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} (u_k)_K \\ (\lambda^a_k)_{nK} \end{pmatrix}$$

and

$$\Delta\left(\mathbf{x}\right) = \prod_{ak} \Theta\left(u_k \lambda^a_k\right), \qquad \Gamma = \frac{\langle \mathbf{x}\mathbf{x}^\top \Delta(\mathbf{x}) \rangle}{\langle \Delta(\mathbf{x}) \rangle}$$

where the average $\langle \ldots \rangle$ is with respect to $\mathbf{u}, \Lambda \sim P(\mathbf{u}, \Lambda)$.
Using Schur's complement formula, see Appendix A, it is straightforward to find

$$-\ln\det A = \ln\det\Sigma - \ln\det R^* .$$

Once again

$$\langle \mathcal{Z}^n \rangle = \int \exp\left\{ N \left( \alpha \mathcal{E}(R,Q) + \mathcal{S}(R,Q) \right) \right\} dQdR$$

with

$$\mathcal{S}(R,Q) = \frac{1}{2} \ln\det\Sigma - \frac{1}{2} \ln\det R^* + \frac{n}{2} - \gamma \frac{n}{2} \sum_{k,l} \left(Y^\top Y\right)^{-1}_{kl} - \frac{n}{2} \ln\det \left( \frac{1}{\gamma} Y^\top Y \right) \tag{3.4}$$

$$\mathcal{E}(R,Q) = \ln\langle \Delta(\mathbf{x}) \rangle = \ln \int \prod_{a=0}^{n} \prod_{k=1}^{K} dx^a_k \Delta(\mathbf{x}) \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left( -\frac{1}{2} \mathbf{x}^\top \Sigma^{-1} \mathbf{x} \right) \tag{3.5}$$

where $x^0_k = u_k$.
Next, we can easily find the saddle point equations, see Appendix B, which read

$$\alpha \frac{2 - \delta_{ij}}{2} \left( \Sigma^{-1} - \Sigma^{-1} \Gamma \Sigma^{-1} \right)_{ij} = \frac{1}{2} \left( \Sigma^{-1} \right)_{ji} \tag{3.6}$$

where the indices $i, j$ only run through the lower part of the matrix, i.e. $R$ and $Q$ blocks, since $R^*$ is completely determined by the data and the top right corner is just the transpose of $R$.

Starting from this point we can show the interesting result of how the problem we are considering, under a specific set of assumptions, becomes equivalent to the one of $K$ independent perceptrons.

## 3.1 Orthogonal data and independent perceptrons

First, let us state that, as it can be seen in Appendix D.1, in the case of $K = 1$ and under specific assumptions, i.e. $\gamma = 1$, $q^{aa} = 1$, the problem we are considering is equivalent to the one of a single perceptron. Let us now consider a very specific setting for our problem and take the $K$ vectors from family $Y$ orthogonal to each other, i.e. $\mathbf{y}_k \cdot \mathbf{y}_l = 0$ when $k \neq l$. Where of course we have the constraint $K < M$ on the number of $\mathbf{y}$ vectors otherwise they could not be all orthogonal. In such a configuration, given the orthogonality of the data, one expects a result similar to the one describing $K$ independent perceptrons. Similarly to the case of $K = 1$, to allow for a precise comparison, we assume $\gamma = 1$ and $q^{aa}_{kk} = 1$.
Let $S_1, \dots, S_K$ be the entropies of $K$ independent perceptrons. We can compare the entropies

$$\mathcal{S} - \sum_{k=1}^{K} S_k = -\frac{1}{2}\left(\ln\det A - \sum_{K=1}^{K} \ln\det A_k\right)$$

In order for it to simplify we would like $A$ to be a block-diagonal matrix

$$A = \begin{pmatrix} A_{11} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & A_{22} \end{pmatrix}$$

with $A_{kk} = A_k \; \forall\, k = 1, \dots, K$.
This automatically implies the need to have $\left(A^{-1}\right)_{kl} = 0$ and the simplest way to get this is to use an ansatz where all terms mixing $\mathbf{y}_k, \mathbf{y}_l$ for $k \neq l$ are equal to zero.
In this way also the energetic term will split into independent terms since the integration variables coming from different $\mathbf{y}$ vectors will be uncorrelated:

$$\mathcal{E} = \sum_{k=1}^{K} \mathcal{E}_k \,.$$

The energetic term of the $K$ independent perceptrons will consist of

$$E_{tot} = \sum_{k=1}^{K} E_k \,.$$

From the equivalence between the case with $K = 1$ and the single perceptron it automatically follows that $\mathcal{E}_k = E_k$ thus proving that these two terms are equivalent.
Before deeming this result to be reliable we need to test the assumption we made to see whether it is consistent with the saddle point equations in Eq.(3.6) that we use to compute the partition function.
In the case of our assumptions, for $\mathbf{y}_k, \mathbf{y}_l$ orthogonal when $k \neq l$, $\Sigma$ will have a particular structure, here we only present $\Sigma$ in the case where $K = 2$ but it should be easy for the

reader to picture the general $K$ case

$$\Sigma = \begin{pmatrix} R_1^* & 0 & R_1^\top & 0 \\ 0 & R_2^* & 0 & R_2^\top \\ R_1 & 0 & Q_1 & 0 \\ 0 & R_2 & 0 & Q_2 \end{pmatrix} .$$

The inverse of such a matrix will have the same structure, i.e. zero elements in the same spots. It follows that, for the terms we set to zero, the saddle point equations reduce to

$$\left(\Sigma^{-1}\Gamma\Sigma^{-1}\right)_{ij} = 0 .$$

Regarding $\Gamma$, under our assumptions we have

$$\left\langle u_k \lambda_l^a \prod_{a,m} \Theta\left(u_m \lambda_m^a\right) \right\rangle = \left\langle u_k \prod_a \Theta\left(u_k \lambda_k^a\right) \right\rangle \left\langle \lambda_l^a \prod_a \Theta\left(u_l \lambda_l^a\right) \right\rangle \left\langle \prod_a \prod_{m \neq k,l} \Theta\left(u_m \lambda_m^a\right) \right\rangle ,$$

then

$$\Gamma_{kl}^a = \frac{\left\langle u_k \prod_a \Theta\left(u_k \lambda_k^a\right)\right\rangle \left\langle \lambda_l^a \prod_a \Theta\left(u_l \lambda_l^a\right)\right\rangle}{\left\langle \prod_a \Theta\left(u_k \lambda_k^a\right)\right\rangle \left\langle \prod_a \Theta\left(u_l \lambda_l^a\right)\right\rangle} .$$

We observe that

$$\left\langle u_k \prod_a \Theta\left(u_k \lambda_k^a\right) \right\rangle = -\left\langle u_k \prod_a \Theta\left(-u_k \lambda_k^a\right) \right\rangle = -\left\langle u_k \prod_a \Theta\left(u_k \lambda_k^a\right) \right\rangle = 0 .$$

Note that we can do the same for every other term of $\Gamma$ mixing orthogonal vectors. We then conclude that $\Gamma$ has the same structure as $\Sigma^{-1}$. The same structure is preserved during the multiplication $\Sigma^{-1}\Gamma\Sigma^{-1}$, thus implying that $\forall (i,j)$ s.t. $\Sigma_{ij}$ is set to 0 then

$$\left(\Sigma^{-1}\Gamma\Sigma^{-1}\right)_{ij} = 0$$

thus proving that our assumptions are compatible with the saddle point equations.
Now that we proved this consistency, we can state that under the assumptions we made we recover the case of $K$ independent perceptrons from our model and this means that, in this particular setting, we could exploit the already well-known results for the individual perceptron to characterise our problem.

In general, to proceed with the calculations for the partition function, we should assume a particular behaviour of the replicas and we will consider the following RS ansatz

$$R_{kl}^a = R_{kl} , \qquad q_{kl}^{ab} = q_{kl} + p_{kl}\delta_{ab} .$$

Such an ansatz assumes all replicas to be statistically equivalent and implies the energy landscape of the system to be relatively simple having one large basin of attraction for each $R_{kl}$ and $q_{kl}$. In spite of its apparent simplicity, the saddle point equations are rather involved and feature three different $K$-dimensional integrals which make solving them an expensive task which we keep for future work, but the details of the calculations can be found in Appendix D.2.

To conclude the discussion, let us briefly summarise the results we found. We started writing the partition function of a system in which we are able to test a small number $K$ of vectors from family $Y$ against an extensive pool of $\mathbf{x}$ vectors, from there we proceeded with replica calculations introducing overlap parameters. Thanks to these parameters we were able to prove the equivalence between our model, when the $\mathbf{y}$ vectors are taken orthogonal to each other, and the case of $K$ individual perceptrons. This parallelism obviously allows to exploit the well-known results for a single perceptron to draw conclusions on our model in that particular setting, hence it proves to be highly valuable. Expensive numerical solutions of the saddle point equations under RS ansatz are instead kept for future work, but what we can say is that the neural network will not be able, in any case, to generalise and learn about vectors orthogonal to the hyperplane identified by the $K$ vectors tested from family $Y$.

# Chapter 4

# Many-vs-many binding partners data

We now take into consideration the model where each element of $X$ is tested against one element of $Y$ creating a dataset containing a number $B$ of triplets $(b_\xi, \mathbf{x}_\xi, \mathbf{y}_\xi)$.
Let $X = (\mathbf{x}_1, \dots, \mathbf{x}_B) \in \mathbb{R}^{N\times B}$ and $Y = (\mathbf{y}_1, \dots, \mathbf{y}_B) \in \mathbb{R}^{M\times B}$ be data points, with independent components drawn from $x_i, y_j \sim \mathcal{N}(0,1)$. Let $W, W^* \in \mathbb{R}^{N\times M}$ be the interaction matrices of the student and the teacher, respectively. We are interested in

$$\mathcal{Z} = \int \mu\left(dW\right) \prod_\xi \Theta\left(z^*_\xi \mathbf{x}_\xi^\top W \mathbf{y}_\xi\right), \quad z^*_\xi = \frac{1}{\sqrt{MN}} \mathbf{x}_\xi^\top W^* \mathbf{y}_\xi$$

where $\mu\left(dW\right) \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(W^\top W\right)\right\}$ is a normalised Gaussian measure. Note that $\mathbb{E}[W_{ij}] = 0$ and $\mathbb{E}[W_{ij}^2] = 1$ under $\mu\left(dW\right)$. In particular $\mathbb{E}\left[\mathrm{tr}\left(W^\top W\right)\right] = NM$, and typically $W_{ij} \sim \mathcal{O}(1)$. We assume $\mathrm{tr}\big(W^* W^{*\top}\big) = NM$. In a similar fashion as before, we want to exploit the replica trick to find the typical value of the partition function, therefore, Assuming that the data $(z^*_\xi, \mathbf{x}_\xi, \mathbf{y}_\xi)$ are i.i.d. over $\xi$ we want to compute

$$\begin{aligned}\langle \mathcal{Z}^n \rangle &= \int \left\langle \prod_{a,\xi} \Theta\left(z^*_\xi \mathbf{x}_\xi^\top W^a \mathbf{y}_\xi\right) \right\rangle_{\mathcal{D}} \prod_a \mu\left(dW^a\right) = \\ &= \int \exp\left\{ B \ln \left\langle \prod_a \Theta\left(z^* \mathbf{x}^\top W^a \mathbf{y}\right) \right\rangle_{\mathcal{D}} \right\} \prod_a \mu\left(dW^a\right) .\end{aligned}$$

This time we define the quantities $z^a = \frac{1}{\sqrt{MN}} \mathbf{x}^\top W^a \mathbf{y}$ and $\mathbf{z} = (z^*, z^1, \dots, z^n)$ so as to have

$$\left\langle \prod_a \Theta(z^* \mathbf{x}^\top W^a \mathbf{y}) \right\rangle_{\mathcal{D}} = \left\langle \prod_a \Theta(z^* z^a) \right\rangle_{\mathbf{z}\sim P(\mathbf{z})} = 2 \int_0^{+\infty} P(\mathbf{z}) d\mathbf{z}$$

where $P(\mathbf{z})$ is the distribution induced by $\mathbf{x}, \mathbf{y}$ which are assumed to be normally distributed with independent components of zero mean and unit variance. Non-zero contributions come from points $\mathbf{z}$ where all entries have the same sign. The last equality follows from $P(\mathbf{z}) = P(-\mathbf{z})$.
Note that the exact form of $P(\mathbf{z})$ is a complex distribution, but still convex, hence, for simplicity, we will assume it to be Gaussian with $\langle z^a \rangle = \langle z^* \rangle = 0$ and

$$\langle z^a z^b \rangle = \frac{1}{NM} \sum_{i,j} W^a_{ij} W^b_{ij} = q_{ab}\,, \quad \langle z^a z^* \rangle = \frac{1}{NM} \sum_{i,j} W^a_{ij} W^*_{ij} = R_a\,, \quad \langle z^{*2} \rangle = 1$$

Thus $P(\mathbf{z})$ only depends on $W^1, \dots, W^n$ through $Q = (q^{ab})$ and $R = (R^a)$:

$$P(\mathbf{z}) \propto \exp\left\{-\frac{1}{2}\mathbf{z}^\top \Sigma^{-1}\mathbf{z}\right\}, \quad \Sigma = \begin{pmatrix} 1 & R^\top \\ R & Q \end{pmatrix}.$$

Note that $P(\mathbf{z}) = P(z^*)P(z^1, \dots, z^n \mid z^*)$, where $P(z^*) \sim \mathcal{N}(0,1)$, while $P(z^1, \dots, z^n \mid z^*)$ is Gaussian with $\langle z^a \mid z^* \rangle = R^a z^*$ and $\langle z^a z^b \mid z^* \rangle_c = q^{ab} - R^a R^b$.
Next, we proceed in a similar way as before and we write the integrand as the exponential of an energetic and an entropic part:

$$\langle \mathcal{Z}^n \rangle = \int \exp\left\{MN\left(\alpha \mathcal{E}(R,Q) + \mathcal{S}(R,Q)\right)\right\} dQdR$$

where $\alpha = \frac{B}{MN}$, $\mathcal{E}(R,Q) = \ln\left(2\int_0^{+\infty} P(\mathbf{z})d\mathbf{z}\right)$ and $\mathcal{S}$ is such that

$$e^{NM\mathcal{S}(R,Q)} = \int \prod_a \delta\left(R^a - \frac{1}{NM}\operatorname{tr}\left(W^a W^{*\top}\right)\right) \prod_{a \le b} \delta\left(Q^{ab} - \frac{1}{NM}\operatorname{tr}\left(W^a W^{b\top}\right)\right) \prod_a \mu\left(dW^a\right)$$

Note that we need $B \sim \mathcal{O}(MN)$ to get a nontrivial regime. Each measurement gives one inequality $z_t \mathbf{x}_t^\top W \mathbf{y}_t \ge 0$ and since $W$ has $M \times N$ degrees of freedom it is reasonable that we need $\mathcal{O}(MN)$ inequalities to get a significant constraint on $W$.
Using Fourier transforms $\delta(x) = \frac{N}{2\pi}\int e^{iN\xi x}d\xi$, we get

$$e^{NMS(Q,R)} = \int_{-i\infty}^{i\infty} \prod_{a \le b} \mathrm{d}\hat{q}^{ab} \prod_a \mathrm{d}\hat{R}^a \exp\Bigg\{NM\sum_{a \le b}\hat{q}^{ab}q^{ab} + NM\sum_a \hat{R}^a R^a + \\ + \frac{NM}{2}\sum_{ab}(L^{-1})^{ab}\hat{R}^a\hat{R}^b - \frac{NM}{2}\ln\det(L)\Bigg\}$$

where $L^{ab} = \hat{q}^{ab} + \delta_{ab}\hat{q}^{ab} + \delta_{ab}$. Note that we omit irrelevant constants and exploit $\sum_{i,j} W_{ij}^{*2} = MN$. Then, we extremise in $\hat{R}^a$ and $\hat{q}^{ab}$ for large $MN$ finding

$$R^a + \sum_b (L^{-1})^{ab}\hat{R}^b = 0, \quad (L^{-1})^{ab} = q^{ab} - R^a R^b = \langle z^a z^b \mid z^* \rangle_c$$

and $\hat{q}^{ab} = \left(1 - \frac{\delta_{ab}}{2}\right)\left(L^{ab} - \delta_{ab}\right)$. Substituting we get

$$\mathcal{S}(Q,R) = \frac{n}{2} - \frac{1}{2}\operatorname{tr} Q + \frac{1}{2}\ln\det\left(Q - RR^\top\right).$$

## 4.1 Replica Symmetric ansatz

We shall now choose an ansatz to proceed with the calculations and we opt for a Replica Symmetric ansatz of the form

$$R^a = r, \quad q^{ab} = (1 - \delta_{ab})q_0 + \delta_{ab}q_1.$$

Then $P(z^1, \dots, z^n \mid z^*)$ becomes a Gaussian distribution with

$$\langle z^a \mid z^* \rangle = rz^* \quad \text{and} \quad \langle z^a z^b \mid z^* \rangle_c = (1 - \delta_{ab})q_0 + \delta_{ab}q_1 - r^2.$$

Next, we compute

$$\begin{aligned}\int_0^{+\infty} P(\mathbf{z})d\mathbf{z} &= \int_0^{+\infty} dz^* P(z^*) \int_0^{+\infty} P(z^1,\dots,z^n \mid z^*)dz^1\dots dz^n = \\ &= \int_0^{+\infty} dz^* \varphi(z^*) \int_{-\infty}^{+\infty} d\zeta \varphi(\zeta) \Phi^n(A\zeta + Bz^*)\end{aligned}$$

where $A = A(r, q_0, q_1) = \sqrt{\frac{q_0 - r^2}{q_1 - q_0}}$, $B = B(r, q_0, q_1) = \frac{r}{\sqrt{q_1 - q_0}}$ and

$$\varphi(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}\,, \quad \Phi(c) = \int_{-\infty}^{c} \varphi(w)dw = \frac{1}{2}\left\{1 - \mathrm{erf}\left(\frac{c}{\sqrt{2}}\right)\right\}$$

where we exploited the analogy with the problem of the equicorrelated Gaussian orthant [11], see Appendix C for more details.
For $n \to 0$ we can write

$$\mathcal{E}(R, Q) = \ln\left\{2\int_0^{+\infty} P(\mathbf{z})d\mathbf{z}\right\} \simeq 2n \int_0^{+\infty} dz^* \varphi(z^*) \int_{-\infty}^{+\infty} d\zeta \varphi(\zeta) \ln \Phi(A\zeta + Bz^*)\,.$$

For the entropy we have $\det\left(Q - RR^\top\right) = \det \Sigma = (q_1 - q_0)^{n-1} d$ with $d = q_1 + (n-1)q_0 - nr^2$. Therefore

$$\begin{aligned}\mathcal{S}(Q, R) =& \frac{n}{2}(1 - q_1) + \frac{1}{2}\ln\left(q_1 + (n-1)q_0 - nr^2\right) + \frac{n-1}{2}\ln(q_1 - q_0) \simeq \\ \simeq& \frac{n}{2}\left\{1 - q_1 + \ln(q_1 - q_0) + \frac{q_0 - r^2}{q_1 - q_0}\right\}\end{aligned}$$

where we considered the $n \to 0$ limit. Then, we can write $\mathcal{E} = n\tilde{\mathcal{E}}$ and $\mathcal{S} = n\tilde{\mathcal{S}}$ where

$$\tilde{\mathcal{E}}(r, q_0, q_1) = 2\int_0^{+\infty} dz^* \varphi(z^*) \int_{-\infty}^{+\infty} d\zeta \varphi(\zeta) \ln \Phi(A\zeta + Bz^*)$$

and

$$\tilde{\mathcal{S}}(r, q_0, q_1) = \frac{1}{2}\left\{1 - q_1 + \ln(q_1 - q_0) + \frac{q_0 - r^2}{q_1 - q_0}\right\}\,. \tag{4.1}$$

Substituting we get

$$\langle \mathcal{Z}^n \rangle = \exp\left\{nMN \underset{Q,R}{\mathrm{extr}} \left(\alpha\tilde{\mathcal{E}}(R, Q) + \tilde{\mathcal{S}}(R, Q)\right)\right\}\,.$$

Using the definitions of $A(r, q_0, q_1)$ and $B(r, q_0, q_1)$ we can rewrite the entropy in Eq.(4.1) as

$$\tilde{\mathcal{S}}(q_1, A, B) = \frac{1}{2}\left\{1 - q_1 + \ln q_1 - \ln\left(1 + A^2 + B^2\right) + A^2\right\}$$

from which, since $\tilde{\mathcal{E}}$ depends on $r, q_0, q_1$ only through $A$ and $B$, we can determine the optimal value of $q_1$. Extremizing $\tilde{\mathcal{S}}$ with respect to $q_1$ at fixed $A, B$ we get $q_1 = 1$.

### 4.1.1 Saddle point equations

Now we extremize with respect to $A$ and $B$ to find the saddle point equations. Starting from the energetic term we get

$$\begin{aligned}\frac{\partial\tilde{\mathcal{E}}}{\partial A} =&2\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^{\infty}\mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta+Bz^*\right)}{\Phi\left(A\zeta+Bz^*\right)}\zeta\,,\\ \frac{\partial\tilde{\mathcal{E}}}{\partial B} =&2\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^{\infty}\mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta+Bz^*\right)}{\Phi\left(A\zeta+Bz^*\right)}z^*\,.\end{aligned}$$

Extremizing the entropic term

$$\frac{\partial\tilde{\mathcal{S}}}{\partial A}=A-\frac{A}{1+A^2+B^2}\,,\quad \frac{\partial\tilde{\mathcal{S}}}{\partial B}=-\frac{B}{1+A^2+B^2}\,.$$

We then write the saddle point equations as follows:

$$2\alpha\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^{\infty}\mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta+Bz^*\right)}{\Phi\left(A\zeta+Bz^*\right)}\zeta+A-\frac{A}{1+A^2+B^2}=0\,,$$

$$2\alpha\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^{\infty}\mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta+Bz^*\right)}{\Phi\left(A\zeta+Bz^*\right)}z^*-\frac{B}{1+A^2+B^2}=0\,.$$

These equations admit solutions with $q_0=r$ which means considering the teacher to be equivalent to just another student, which also intuitively should be reasonable, but a more detailed discussion of this argument can be found in [6]. Thanks to this observation we can write $A=\sqrt{r}$ and $B=\frac{r}{\sqrt{1-r}}$. Then, we can focus on one of the two equations and solve parametrically:

$$\alpha=\frac{1}{2}\frac{\frac{A}{1+A^2+B^2}-A}{\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^{\infty}\mathrm{d}\zeta\varphi(\zeta)\frac{\varphi(A\zeta+Bz^*)}{\Phi(A\zeta+Bz^*)}\zeta}$$

for $\alpha$ as a function of $r$. Given $q_0=r$ we can rewrite the energetic and the entropic terms as follows

$$\tilde{\mathcal{E}}(r)=2\int_0^{+\infty}dz^*\varphi(z^*)\int_{-\infty}^{+\infty}d\zeta\varphi(\zeta)\ln\Phi(\sqrt{r}\zeta+\frac{r}{\sqrt{1-r}}z^*)\,,$$

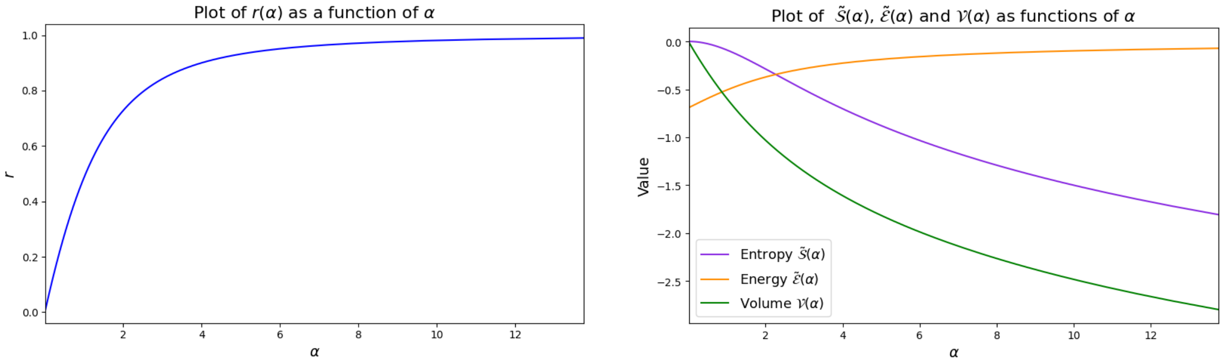
$$\tilde{\mathcal{S}}(r)=\frac{1}{2}\left\{r+\ln(1-r)\right\}\,,$$

and we can define another interesting quantity which we call $\mathcal{V}$ as

$$\mathcal{V}(\alpha,r)=\alpha\tilde{\mathcal{E}}(r)+\tilde{\mathcal{S}}(r)$$

which is nothing less than the volume of the space of the interaction matrices.

Since the only parameter we can control is $\alpha$ it makes sense to look at the behaviour of the relevant quantities of this problem as a function of the latter as presented in Figure 4.1. In particular, from Figure 4.1a we can see how the overlap between the student and the teacher starts at $r=0$ for $\alpha=0$ and increases with $\alpha$ as expected. In particular we notice a steep growth for small values of $\alpha$, whereas the curve starts saturating for $\alpha\simeq 2$ and, as expected, for large values of $\alpha$ we have $r\to 1$. This result is exactly what we expect and is consistent with what we would expect knowing the results for the perceptron [6].



(a) (b)

Figure 4.1: Plots of the overlap parameter $r$, the entropy $\tilde{\mathcal{S}}$, the energy $\tilde{\mathcal{E}}$ and the volume $\mathcal{V}$ as functions of $\alpha$.

In Figure 4.1b the behaviour of the entropic term, the energetic term, and the volume with respect to $\alpha$ are shown in the same plot to allow for a more immediate comparison of the three curves. We can see that $\tilde{\mathcal{S}}$ and $\tilde{\mathcal{E}}$ are both monotonic but their trends are one opposite to the other. As a matter of fact, they respectively increase and decrease their values with $\alpha$, with $\tilde{\mathcal{E}}$ that saturates to 0 for $\alpha \to +\infty$. Moreover, the entropic term presents an inflection point roughly at the same value of $\alpha \simeq 2$ for which the plot of $r$ begins to saturate. The general behaviour of $\mathcal{V}$, as expected, is still monotonic and decreases with $\alpha$, implying that the more data are fed into the neural network, the more the space of possible interaction matrices shrinks.

Let us now briefly summarise the results obtained in this section. We started writing the partition function of a system in which we are able to test an extensive number of $\mathbf{x}, \mathbf{y}$ pairs, thus building a dataset with a number $B$ of triplets $(b_d, \mathbf{x}_d, \mathbf{y}_d)$ describing every time the interaction between two new vectors. After some replica calculations, again under the RS ansatz, we were able to find the saddle point equations and to plot the behaviour of the overlap parameters, as well as of the entropic term, the energetic term and the volume of the space of interaction matrices, as a function of the amount of data fed into the network $\alpha$. These quantities behave as expected with the overlap between student and teacher beginning to saturate towards one at $\alpha \simeq 2$, thus meaning that the neural network actually learns. The same can be said for the overlap among the different students (replicas) and in general the entropic term decreases whereas the energetic term grows and saturates to zero. As expected, the volume of the space of the interaction matrices keeps decreasing with $\alpha$.

# Chapter 5

# Conclusion

In order to bring together the central arguments of this thesis, we now summarize the key results and their implications. The problem posed was to characterize the parameter space of a neural network with the task of learning the binding matrix between two families of data. We presented two different models for which we computed the partition functions exploiting the replica trick with Replica Symmetric ansatz.

Studying the first model we found the equivalence between the case of $K$ independent perceptrons and the one in which we can test $K$ orthogonal vectors of one family against an extensive number of vectors from the other family. Then, under RS ansatz, we were able to derive a simplified expression for the partition function of the model which in future work will permit to write and solve a set of saddle point equations allowing to study the behaviour of the overlap parameters.
Studying the second model, where we considered an extensive number of $(\mathbf{x}, \mathbf{y})$ pairs, we found the behaviour of the overlap parameters, coming from replica calculations with RS ansatz, which proved to align with our expectations showing a decreasing volume of the space of interaction matrices with the amount of data fed into the neural network.

We deem both models to be of great importance both from a theoretical and practical point of view, since the characterisation of these and similar settings might lead to a deeper understanding of how to collect actual data for an optimal deployment of experimental resources. Because of this, in the future, it would certainly be interesting to compare the results obtained in this thesis with the performances of actual neural networks trained on real-world data to see how accurate the predictions of these models are.

Ongoing continuation of this work is to derive the saddle point equations of the model in Section 3 and to solve them numerically for different values of the amount of data fed into the neural network. With this procedure, hopefully, interesting behaviours of the generalization capabilities of the model, such as phase transitions, might arise. Further changing the structure of the dataset could also lead to interesting results. A possible option for experimentalists would be to test few elements of family $X$ against many elements of family $Y$ and the other way around; under these assumptions nontrivial results could arise and it is definitely something that we envisage to analyse in the future. Moreover, in the case of TCRs and peptides the interaction matrix $W$ is generally sparse; therefore, future perspectives might include taking this feature into account to look at the changes induced by this additional constraint. In general making precise assumptions on the structure of the interaction matrix might lead to interesting behaviours.

Always with the aim of reducing as much as possible the need for collection of new data, an interesting question certainly is if there are optimal choices for the dataset $\mathcal{D}$ in terms of the type of data and not just the quantity. To pose the question in a statistical physics language we would want to find the best way to choose the vectors $\mathbf{x}_d$, $\mathbf{y}_d$ such that $\mathcal{Z}$ is much smaller than its typical value, i.e. the possible $W$ matrices are highly constrained. Answering this question would require to look at large deviations and non-zero number of replicas, in a similar fashion to what is done in [5].

Other cases that could be considered are those of more complex binding models involving higher-order interactions between the binding partners or when the threshold in Eq.(2.1) is set to be non-zero, thus describing a situation in which a different fraction of pairs actually bind. This last setting would be an important step towards a more realistic model of what happens in the actual interactions; setting the threshold to zero means considering that on average half of the considered molecules bind, whereas in real life the fraction is obviously smaller. Another assumption that could be changed to make the model more realistic is changing the distributions from which the $\mathbf{x}$ and $\mathbf{y}$ vectors are sampled, as a matter of fact choosing a spin-like distribution would mean studying a more realistic model and could show some interesting dependence on the form of the interaction matrix $W$. Always with a glance at the practical implications of this research, experiments can produce noisy data, meaning that while collecting a data point $(b_d, \mathbf{x}_d, \mathbf{y}_d)$ there is some probability that $\mathbf{x}$ or $\mathbf{y}$ are mildly corrupted versions of the true sequences, or that $b_d$ has the wrong sign. We can then modify the partition function to take into account such sources of noise trying to understand their effect on the inference of the binding matrix $W$.

As it should be clear by now, the possibilities are numerous and all of them can be of great value not only from a mere theoretical point of view, but with many direct implications in more practical tasks especially if more realistic features are included in the studied models. Hence, the hope is to be able to explore as many of them as possible, further shedding light on the intricate mechanisms governing complex networks.

# Bibliography

[1] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[3] Barbara Bravi, Andrea Di Gioacchino, Jorge Fernandez-de Cossio-Diaz, Aleksandra M Walczak, Thierry Mora, Simona Cocco, and Rémi Monasson. A transfer-learning approach to predict antigen immunogenicity and t-cell receptor specificity. *eLife*, 12:e85126, sep 2023.

[4] Simona Cocco, Rémi Monasson, and Francesco Zamponi. *From Statistical Physics to Data-Driven Modelling: with Applications to Quantitative Biology.* Oxford University Press, 09 2022.

[5] Hugo Cui, Luca Saglietti, and Lenka Zdeborova. Large deviations for the perceptron model and consequences for active learning. In Jianfeng Lu and Rachel Ward, editors, *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pages 390–430. PMLR, 20–24 Jul 2020.

[6] A. Engel and C. Van den Broeck. *Statistical Mechanics of Learning.* Cambridge University Press, 2001.

[7] Jorge Fernandez-De-Cossio-Diaz, Thomas Tulinski, Simona Cocco, and Rémi Monasson. Replica symmetry breaking and clustering phase transitions in undersampled restricted Boltzmann machines. working paper or preprint, February 2024.

[8] E. Gardner. Maximum storage capacity in neural networks. *Europhysics Letters*, 4(4):481, aug 1987.

[9] E Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257, jan 1988.

[10] M. Mezard, G. Parisi, and M.A. Virasoro. *Spin Glass Theory and Beyond.* World Scientific lecture notes in physics. World Scientific, 1987.

[11] G. P. STECK. Orthant probabilities for the equicorrelated multivariate normal distribution. *Biometrika*, 49(3-4):433–445, 12 1962.

[12] J Steinberg, U Adomaitytė, A Fachechi, P Mergny, D Barbier, and R Monasson. Replica method for computational problems with randomness: principles and illustrations. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104002, oct 2024.

# Appendix A

# Schur's complement

The Schur complement is defined for a block matrix. Let matrix $M$ be defined as

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{A.1}$$

If respectively $D$ (or $A$) is invertible, then the Schur complement of the block $D$ (or $A$) of $M$ is the matrix defined by

$$M/D = A - BD^{-1}C \tag{A.2}$$

or

$$M/A = D - CA^{-1}B \tag{A.3}$$

This complement allows to write very useful properties of matrix $M$. For instance its inverse can be written as

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(M/A)^{-1}CA^{-1} & -A^{-1}B(M/A)^{-1} \\ -(M/A)^{-1}CA^{-1} & (M/A)^{-1} \end{pmatrix} \tag{A.4}$$

or in a symmetric way for $M/D$.

$$M^{-1} = \begin{pmatrix} (M/D)^{-1} & -(M/D)^{-1}BD^{-1} \\ -D^{-1}C(M/D)^{-1} & D^{-1} + D^{-1}C(M/D)^{-1}BsD^{-1} \end{pmatrix} \tag{A.5}$$

It also allows to write the determinant of $M$, when $A$, respectively $D$, is invertible, as follows

$$\det M = \det A \det \left(D - CA^{-1}B\right) \tag{A.6}$$

respectively

$$\det M = \det D \det \left(A - BD^{-1}C\right) \tag{A.7}$$

# Appendix B

# Gradient of multivariate normal

Consider a centred multivariate normal density

$$P(\mathbf{x}) = (\mathbf{2}\pi)^{-\mathbf{N/2}} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1} \mathbf{x}\right) \tag{B.1}$$

with covariance matrix $\left\langle \mathbf{x}\mathbf{x}^\top \right\rangle = \Sigma$. Then

$$\frac{\partial}{\partial \Sigma_{ij}} \ln P(\mathbf{x}) = \frac{2-\delta_{ij}}{2} \left[-\Sigma^{-1} + \Sigma^{-1}\mathbf{x}\mathbf{x}^\top \Sigma^{-1}\right]_{ij} \qquad (i \leq j) \tag{B.2}$$

where the derivative is taken with respect to a triangle $i \leq j$ of the symmetric matrix $\Sigma$. That is, the variation of $\Sigma_{ij}$ for $i < j$ implies the same variation of $\Sigma_{ji}$, which is why we need the factor $1 - \frac{1}{2}\delta_{ij}$.

For a proof of this formula, one needs the gradient of $\ln\det(\ldots)$. This is given in Appendix A.4.1 of Boyd & Vandenberghe's book [2].

# Appendix C

# Equicorrelated Gaussian orthant

Let $x_1, \dots, x_n$ be normally distributed and *equicorrelated* with $\langle x_i \rangle = 0$ and $\langle x_i x_j \rangle = \delta_{ij} + (1 - \delta_{ij})\rho$. Their joint density reads

$$P(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \det \Sigma^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{x}^\top \Sigma^{-1} \mathbf{x}\right\} \tag{C.1}$$

where $\Sigma_{ij} = \delta_{ij} + (1-\delta_{ij})\rho$ and has a non-degenerate eigenvalue $1+(n-1)\rho$ and a $(n-1)$-fold degenerate eigenvalue $1 - \rho$ which must be positive for $\Sigma$ to be positive definite.
We want to compute the orthant probability

$$\mathcal{P} = P(x_1 > a, \dots, x_n > a) = \int_a^{+\infty} P(x_1, \dots, x_n) dx_1 \dots dx_n \tag{C.2}$$

where $a$ is a threshold.
Assuming $\rho > 0$, we can write $x_i = \sqrt{\rho} Z_0 + \sqrt{1-\rho} Z_i$, where $Z_0, Z_1, \dots, Z_n$ are independent standard normals. Indeed, from $\langle Z_i Z_j \rangle = \delta_{ij}$, we see that this representation implies $\langle x_i x_j \rangle = \rho + (1 - \rho)\delta_{ij}$, which is consistent with the previous definition. It follows that $P(x_1, \dots, x_n \mid Z_0) = \prod_i P(x_i \mid Z_0)$ where $P(x_i \mid Z_0)$ is Gaussian with $\langle x_i \mid Z_0 \rangle = \sqrt{\rho} Z_0$ and $\langle x_i^2 \mid Z_0 \rangle = 1 - \rho$. Clearly,

$$\mathcal{P} = \int P(x_1 > a, \dots, x_n > a \mid Z_0 = \zeta) P(\zeta) d\zeta = \tag{C.3}$$

$$= \int \left[P(x_i > a \mid Z_0 = \zeta)\right]^n P(\zeta) d\zeta \tag{C.4}$$

Let $\Phi(c)$ be the CDF of a standard normal

$$\varphi(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}, \quad \Phi(c) = \int_{-\infty}^{c} \varphi(w) dw = \frac{1}{2}\left\{1 - \operatorname{erf}\left(\frac{c}{\sqrt{2}}\right)\right\}. \tag{C.5}$$

Using

$$P(x_i > a \mid Z_0 = \zeta) = \Phi\left(\frac{\zeta\sqrt{\rho} - a}{\sqrt{1-\rho}}\right) \tag{C.6}$$

we obtain

$$\mathcal{P} = \int_{-\infty}^{+\infty} \Phi^n \left(\frac{\zeta\sqrt{\rho} - a}{\sqrt{1-\rho}}\right) \varphi(\zeta) d\zeta \tag{C.7}$$

Note that this gives the correct real result even if $\rho < 0$. For more details refer to [11].

# Appendix D

# Few-vs-many binding partners data: detailed calculations

Let $\mathbf{x}_1, \ldots, \mathbf{x}_B \in \mathbb{R}^N$ be random normal vectors with $\langle x_{i\xi} \rangle = 0$ and $\langle x_{i\xi}^2 \rangle = 1$. Let $\mathbf{y}_1, \ldots, \mathbf{y}_K \in \mathbb{R}^M$ be some fixed vectors. Let $W, W^* \in \mathbb{R}^{N \times M}$ be respectively the interaction matrices of the student and the teacher.

$$\mathcal{Z} = \int \mu(dW) \prod_{\xi,k} \Theta(\mathbf{x}_\xi^\top W^* \mathbf{y}_k \mathbf{x}_\xi^\top W \mathbf{y}_k) \tag{D.1}$$

where $\mu(dW) \propto \exp\left(-\frac{\gamma}{2}\mathrm{tr}(W^2)\right)$ for some $\gamma > 0$, is a normalized Gaussian measure ($\int \mu(dW) = 1$), $\Theta$ is Heaviside's step function and the threshold $T$ in Equation (2.1) is set to zero.
Let $J_{ik} = \sum_j W_{ij} y_{jk}$ (i.e. $\mathbf{J}_k = W\mathbf{y}_k$). Under $W \sim \mu(dW)$, $J$ is Gaussianly distributed with

$$\langle J_{ik} \rangle = 0, \quad \langle J_{ik} J_{jl} \rangle = \sum_{n,m} \langle W_{in} W_{jm} \rangle y_{nk} y_{ml} = \frac{\delta_{ij}}{\gamma} y_k^\top y_l \tag{D.2}$$

Denoting this Gaussian measure by $\mu(dJ)$

$$\mathcal{Z} = \int \mu(dJ) \prod_{\xi,k} \Theta(\mathbf{x}_\xi^\top \mathbf{J}_k^* \mathbf{x}_\xi^\top \mathbf{J}_k) \tag{D.3}$$

Where $J_{ik}^* = \sum_j W_{ij}^* y_{jk}$.
From Eq.(D.3) we infer that $\mathcal{Z}$ involves a product of many random contributions. Products of independent random numbers possess distributions with long tails for which the average and the most probable value do not coincide. On the other hand, the logarithm of such a quantity is a large sum of independent terms and therefore becomes normally distributed so that its average and its most probable value asymptotically coincide. Hence, the typical value of $\mathcal{Z}$, for large $N$, is given by

$$\mathcal{Z} \sim e^{\langle \ln \mathcal{Z} \rangle} \tag{D.4}$$

Using the replica trick previously described in Section 2.3, we want to compute

$$\langle \ln \mathcal{Z} \rangle_X = \lim_{n \to 0} \frac{\langle \mathcal{Z}^n \rangle_X - 1}{n} \tag{D.5}$$

where the average is taken over the data $X = (\mathbf{x}_1, \ldots, \mathbf{x}_B)$ with independent entries $x_{i\xi} \sim \mathcal{N}(0,1)$. Introducing replicas $a = 1, \ldots, n$:

$$\begin{aligned}\langle \mathcal{Z}^n \rangle &= \left\langle \int \left[ \prod_a \mu(dJ^a) \right] \prod_{a,\xi,k} \Theta(\mathbf{x}_\xi^\top \mathbf{J}_k^* \mathbf{x}_\xi^\top \mathbf{J}_k^a) \right\rangle_X = \\ &= \int \left[ \prod_a \mu(dJ^a) \right] \left\langle \prod_{a,\xi,k} \Theta(\mathbf{x}_\xi^\top \mathbf{J}_k^* \mathbf{x}_\xi^\top \mathbf{J}_k^a) \right\rangle_X \end{aligned} \tag{D.6}$$

Since the $\mathbf{x}_\xi$ are i.i.d. over $\xi$

$$\left\langle \prod_{a,\xi,k} \Theta(\mathbf{x}_\xi^\top \mathbf{J}_k^* \mathbf{x}_\xi^\top \mathbf{J}_k^a) \right\rangle_X = \prod_\xi \left\langle \prod_{a,k} \Theta(\mathbf{x}_\xi^\top \mathbf{J}_k^* \mathbf{x}_\xi^\top \mathbf{J}_k^a) \right\rangle_{\mathbf{x}_\xi} = \left\langle \prod_{a,k} \Theta(\mathbf{x}^\top \mathbf{J}_k^* \mathbf{x}^\top \mathbf{J}_k^a) \right\rangle_{\mathbf{x}}^B$$

where the average on the right is taken with respect to a single $\mathbf{x} \in \mathbb{R}^N$ with independent entries $x_i \sim \mathcal{N}(0,1)$. Next, we introduce the variables $u_k = \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^*$ and $\lambda_k^a = \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^a$:

$$\left\langle \prod_{a,k} \Theta(\mathbf{x}^\top \mathbf{J}_k^* \mathbf{x}^\top \mathbf{J}_k^a) \right\rangle_{\mathbf{x}} = \int \left( \prod_{ak} \Theta(u_k \lambda_k^a) \right) P(\mathbf{u}, \Lambda) d\mathbf{u} d\Lambda \tag{D.7}$$

where $\Lambda = (\lambda_k^a)$ and $\mathbf{u} = (u_k)$, and

$$P(\mathbf{u}, \Lambda) = \left\langle \prod_k \delta\left( u_k - \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^* \right) \prod_{a,\nu} \delta\left( \lambda_k^a - \frac{1}{\sqrt{N}} \mathbf{x}^\top \mathbf{J}_k^a \right) \right\rangle_{\mathbf{x}} \tag{D.8}$$

is the distribution of $\mathbf{u}, \Lambda$ induced by $\mathbf{x}$. Since $u_k$ and $\lambda_k^a$ are linear combinations of the Gaussian variables $x$, they are themselves Gaussian. Therefore, $P(\mathbf{u}, \Lambda)$ is a multivariate normal distribution, with zero means and covariances:

$$\langle u_k u_l \rangle = R_{kl}^*, \quad \langle \lambda_k^a u_l \rangle = R_{kl}^a, \quad \langle \lambda_k^a \lambda_l^b \rangle = q_{kl}^{ab} \tag{D.9}$$

where we introduced the overlap parameters

$$R_{kl}^* = \frac{1}{N} (\mathbf{J}_k^*)^\top \mathbf{J}_l^*, \quad R_{kl}^a = \frac{1}{N} (\mathbf{J}_k^a)^\top \mathbf{J}_l^*, \quad q_{kl}^{ab} = \frac{1}{N} (\mathbf{J}_k^a)^\top \mathbf{J}_l^b \tag{D.10}$$

Substituting,

$$\langle \mathcal{Z}^n \rangle = \int \exp\left\{ B \ln \left\langle \prod_{a,k} \Theta(u_k \lambda_k^a) \right\rangle \right\} \prod_a \mu(dJ^a) \tag{D.11}$$

where $\langle \ldots \rangle$ denotes the Gaussian average under $u, \Lambda \sim P(u, \Lambda)$. We change variables from the $J^a$ to these overlaps, writing:

$$\langle \mathcal{Z}^n \rangle = \int \exp\left\{ N \left( \alpha \ln \left\langle \prod_{a,k} \Theta(u_k \lambda_k^a) \right\rangle + S(R, Q) \right) \right\} dQ dR \tag{D.12}$$

where $\alpha = \frac{B}{N}$, $Q = (q_{kl}^{ab})$ and $R = (R_{kl}^a)$ and $e^{NS(R,Q)}$ is an entropic factor acting as the Jacobian for the change of variables.Note that we need $B \sim \mathcal{O}(N)$ to get a nontrivial regime

because each measurement gives one inequality and since $W$ has $N$ degrees of freedom to get a significant constraint on $W$ it is reasonable that we need a number of inequalities comparable to $N$.
Up to a negligible multiplication factor

$$e^{NS} \simeq \int \left(\prod_{a,k,l} d\hat{R}^a_{kl}\right)\left(\prod_{ak\leq bl} d\hat{q}^{ab}_{kl}\right) \exp\left(-N\sum_{a,k,l}\hat{R}^a_{kl}R^a_{kl} - N\sum_{ak\leq bl}\hat{q}^{ab}_{kl}q^{ab}_{kl} + \sum_i \ln\Omega_i\right) \tag{D.13}$$

where, defining $h^a_{ik} = \sum_l \hat{R}^a_{kl}J^*_{il}$ and $\mathbf{h}_i = (h^a_{ik}) \in \mathbb{R}^{nK}$,

$$\Omega_i = \det\left(\frac{1}{\gamma}Y^\top Y\right)^{-\frac{n}{2}} \det(A)^{-\frac{1}{2}} \exp\left(\frac{1}{2}\mathbf{h}_i^\top A^{-1}\mathbf{h}_i\right) \tag{D.14}$$

For large $N$ we can then use the saddle point method and we know

$$R^a_{kl} = \frac{1}{N}\sum_i \langle J^a_{ik}\rangle J^*_{il} \qquad q^{ab}_{kl} = \frac{1}{N}\sum_i \langle J^a_{ik}J^b_{il}\rangle \tag{D.15}$$

and

$$\langle J^a_{ik}\rangle J^*_{il} = \frac{\partial \ln\Omega_i}{\partial \hat{R}^a_{kl}} \qquad \langle J^a_{ik}J^b_{il}\rangle = \frac{\partial \ln\Omega_i}{\partial \hat{q}^{ab}_{kl}}$$

Here $\langle J^a_{ik}\rangle$ and $\langle J^a_{ik}J^b_{il}\rangle$ are moments to be computed under the Gaussian measure with $\langle J^a_{ik}J^b_{il}\rangle_c = (A^{-1})^{ab}_{kl}$ and $\langle J^a_{ik}\rangle = (A^{-1}\mathbf{h}_i)^a_k$. Then at saddle point

$$\begin{cases} \hat{R}^a_{kl} = \sum_{b,m,s} A^{ab}_{km}R^b_{ms}(R^*)^{-1}_{sl} \\ A^{-1} = Q - R(R^*)^{-1}R^\top \end{cases} \tag{D.16}$$

Thus implying

$$\begin{aligned} S = &-\frac{1}{2}\ln\det A + \frac{1}{2}\left[\mathrm{Tr}(AQ) - \mathrm{Tr}\left(AR(R^*)^{-1}R^\top\right)\right] + \\ &-\frac{\gamma}{2}\sum_{a,k,l}\left(Y^\top Y\right)^{-1}_{kl} - \frac{n}{2}\ln\det\left(\frac{1}{\gamma}Y^\top Y\right) = \\ = &-\frac{1}{2}\ln\det A + \frac{n}{2} - \gamma\frac{n}{2}\sum_{k,l}\left(Y^\top Y\right)^{-1}_{kl} - \frac{n}{2}\ln\det\left(\frac{1}{\gamma}Y^\top Y\right) \end{aligned} \tag{D.17}$$

Let us now define some quantities:

$$\Sigma = \begin{pmatrix} (R^*_{kl})_{K\times K} & (R^a_{kl})^\top_{K\times nK} \\ (R^a_{kl})_{nK\times K} & \left(q^{ab}_{kl}\right)_{nK\times nK} \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} (u_k)_K \\ (\lambda^a_k)_{nK} \end{pmatrix} \tag{D.18}$$

and

$$\Delta\left(\mathbf{x}\right) = \prod_{ak}\Theta\left(u_k\lambda^a_k\right), \qquad \Gamma = \frac{\langle \mathbf{x}\mathbf{x}^\top\Delta(\mathbf{x})\rangle}{\langle\Delta(\mathbf{x})\rangle} \tag{D.19}$$

where the average $\langle\ldots\rangle$ is with respect to $\mathbf{u}, \Lambda \sim P(\mathbf{u}, \Lambda)$.

Using Schur's complement formula, see Appendix A, it is straightforward to show $\det A^{-1} = \left(\det R^*\right)^{-1} \det \Sigma$, therefore

$$-\ln \det A = \ln \det \Sigma - \ln \det R^* . \tag{D.20}$$

Once again

$$\langle \mathcal{Z}^n \rangle = \int \exp\left\{N\left(\alpha \mathcal{E}(R,Q) + \mathcal{S}(R,Q)\right)\right\} dQdR \tag{D.21}$$

with

$$\mathcal{S}(R,Q) = \frac{1}{2}\ln\det\Sigma - \frac{1}{2}\ln\det R^* + \frac{n}{2} - \gamma\frac{n}{2}\sum_{k,l}\left(Y^\top Y\right)^{-1}_{kl} - \frac{n}{2}\ln\det\left(\frac{1}{\gamma}Y^\top Y\right) , \tag{D.22}$$

$$\mathcal{E}(R,Q) = \ln\langle\Delta(\mathbf{x})\rangle = \ln\int\prod_{a=0}^{n}\prod_{k=1}^{K} dx_k^a \Delta(\mathbf{x})\frac{1}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{1}{2}\mathbf{x}^\top\Sigma^{-1}\mathbf{x}\right) , \tag{D.23}$$

where $x_k^0 = u_k$ We can easily find the saddle point equations, see Appendix B, which read

$$\alpha\frac{2-\delta_{ij}}{2}\left(\Sigma^{-1} - \Sigma^{-1}\Gamma\Sigma^{-1}\right)_{ij} = \frac{1}{2}\left(\Sigma^{-1}\right)_{ji} \tag{D.24}$$

where the indices $i, j$ only run through the lower part of the matrix, i.e. $R$ and $Q$ blocks, since $R^*$ is completely determined by the data and the top right corner is just the transpose of $R$.

In the following we will first prove the equivalence between our problem when $K = 1$, under certain specific assumptions, and the one of an individual perceptron. Then, we will proceed with the calculations for the RS ansatz.

## D.1 Case $K = 1$ and single perceptron

First we consider the case $K = 1$ which, under some specific assumptions on the parameters of our model, we expect to behave as the single perceptron [6].
First, let us look at what happens in our calculations when $K = 1$. The overlap parameters can be defined as follows:

$$R^a = \frac{1}{N}\sum_i \langle J_i^a\rangle J_i^* \qquad q^{ab} = \frac{1}{N}\sum_i\langle J_i^a J_i^b\rangle \tag{D.25}$$

and, similarly as before, at saddle point we have

$$\begin{cases}\hat{\mathbf{R}} = A\mathbf{R}\\ A^{-1} = Q - \mathbf{R}\mathbf{R}^\top\end{cases} \tag{D.26}$$

### Entropic term

Following the previous result at saddle point, the entropy becomes

$$\mathcal{S}(R,Q) = -\frac{1}{2}\ln(\det A) + \frac{n}{2}\left(1 - \gamma + \ln\gamma\right) \tag{D.27}$$

Now, in order to do a proper comparison with the results in [6] we need to set $\gamma = 1$ and we also have $q^{aa} = 1$. Regarding the results of the entropic part in [6]

$$\begin{aligned} S_P(R, \hat{R}, Q, \hat{Q}, \hat{k}^a) = &-\frac{n}{2} - \frac{1}{2}\ln\Big(\det\tilde{A}\Big) - \frac{1}{2}\sum_{a,b}\hat{R}^a(\tilde{A}^{-1})_{ab}\hat{R}^b + \\ &+ \frac{i}{2}\sum_a \hat{k}^a + i\sum_{a<b} q^{ab}\hat{q}^{ab} + i\sum_a R^a\hat{R}^a \end{aligned} \tag{D.28}$$

we only have to change the variable $\hat{k}^a \mapsto 2\hat{q}^{aa}$ to find

$$\begin{aligned} S_P(R, \hat{R}, Q, \hat{Q}) &= \ln\left[(2\pi e)^{-\frac{n}{2}}\int d\mathbf{J}\exp\left(-\frac{1}{2}\mathbf{J}^\top\tilde{A}\mathbf{J} - i\hat{\mathbf{R}}\mathbf{J}\right)\right] + i\sum_a \hat{R}^a R^a + i\sum_{a\le b}\hat{q}^{ab}q^{ab} = \\ &= -\frac{n}{2} - \frac{1}{2}\ln\det\tilde{A} - \frac{1}{2}\hat{\mathbf{R}}^\top\tilde{A}^{-1}\hat{\mathbf{R}} + i\sum_a \hat{R}^a R^a + i\sum_{a\le b}\hat{q}^{ab}q^{ab} \end{aligned} \tag{D.29}$$

Rewriting $\sum_{a\le b}$ as $\frac{1}{2}\sum_{a,b} + \frac{1}{2}\sum_a$ and changing variables $-i\hat{q}^{ab} \mapsto \hat{q}^{ab}$ and $-i\hat{R}^a \mapsto \hat{R}^a$

$$S_P(R, \hat{R}, Q, \hat{Q}) = -\frac{n}{2} - \frac{1}{2}\ln\det\tilde{A} + \frac{1}{2}\mathrm{Tr}(\hat{\mathbf{R}}^\top\tilde{A}^{-1}\hat{\mathbf{R}}) + \frac{1}{2}\mathrm{Tr}(\tilde{A}Q) - \sum_a R^a\hat{R}^a \tag{D.30}$$

Using once more the saddle point approximation

$$\begin{cases} \hat{R}^a = \sum_b \tilde{A}^{ab}R^b \\ (\tilde{A}^{-1})^{ab} = Q^{ab} - R^aR^b \end{cases} \tag{D.31}$$

Therefore we have that at saddle point $\tilde{A} \equiv A$ and

$$\begin{aligned} S_P &= -\frac{1}{2}\ln\det\tilde{A} && \text{(D.32)} \\ &= -\frac{1}{2}\ln\det A && \text{(D.33)} \end{aligned}$$

Which is exactly what we find setting $\gamma = 1$ in our calculations.

## Energetic term

The energetic term in our calculation so far is "untouched" so in principle it should coincide with the one in [6] which reads

$$\begin{aligned} E\left(q^{ab}, R^a\right) = \ln\int\frac{du}{\sqrt{2\pi}}\int\prod_a\frac{d\lambda^a}{2\pi}\int\prod_a d\hat{\lambda}^a\prod_a\Theta(u\lambda^a)\times \\ \times\exp\left(-\frac{u^2}{2} - \frac{1}{2}\sum_{a,b}(q^{ab} - R^aR^b)\hat{\lambda}^a\hat{\lambda}^b + i\sum_a\hat{\lambda}^a\lambda^a - iu\sum_a\hat{\lambda}^aR^a\right) \end{aligned} \tag{D.34}$$

Which after integration becomes

$$\begin{aligned}E\left(q^{ab},R^{a}\right) =& \ln\int du\int\prod_{a}d\lambda^{a}\prod_{a}\Theta(u\lambda^{a})\frac{\exp\left(-\frac{u^{2}}{2}-\frac{1}{2}\left(\lambda^{\top}-u\mathbf{R}^{\top}\right)A\left(\lambda-u\mathbf{R}\right)\right)}{\sqrt{\det(2\pi A^{-1})}} = \\ =& \ln\int\prod_{a=0}^{n}dx^{a}\prod_{a}\Theta(x^{0}x^{a})\frac{1}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{1}{2}\mathbf{x}^{\top}\mathcal{A}\mathbf{x}\right)\end{aligned} \tag{D.35}$$

where

$$\mathcal{A}:=\begin{pmatrix}\mathbf{R}^{\top}A\mathbf{R}+1 & -\mathbf{R}^{\top}A\\ -A\mathbf{R} & A\end{pmatrix} \tag{D.36}$$

The energetic part in our calculations presents a similar structure

$$\mathcal{E}=\ln\langle\Delta(\mathbf{x})\rangle=\ln\int\prod_{a=0}^{n}dx^{a}\Delta(\mathbf{x})\frac{1}{\sqrt{\det(2\pi\Sigma)}}\exp\left(-\frac{1}{2}\mathbf{x}^{\top}\Sigma^{-1}\mathbf{x}\right) \tag{D.37}$$

And, thanks to Schur's complement formula, it can be shown that $\mathcal{A}\equiv\Sigma^{-1}$, thus proving the consistency between what we found and the results in [6].

## D.2 Replica Symmetric ansatz

In order to get to the saddle point equations, we will work on equations (D.22) and (D.23). To proceed, it is now necessary to assume a particular behaviour of the replicas and we will therefore consider the following RS ansatz.

$$R_{kl}^{a}=R_{kl}\,,\qquad q_{kl}^{ab}=q_{kl}+p_{kl}\delta_{ab} \tag{D.38}$$

Such an ansatz assumes all replicas to be statistically equivalent and implies that the system has a relatively simple energy landscape with one large basin of attraction for each $R_{kl}$ and $q_{kl}$.
Our goal is to find an expression that makes their dependence on the number of replicas $n$ explicit. This will allow us to group everything by $n$ and proceed with the replica trick.

### Entropic term

When we take the derivative of the entropic term with respect to the overlap parameters, the only non-zero contribution will be $\ln\det\Sigma-\ln\det R^{*}=\ln\det A^{-1}$, therefore in the following we will only consider this term.
First we observe

$$\left(A^{-1}\right)_{kl}^{ab}=q_{kl}+p_{kl}\delta_{ab}-\sum_{m,s}R_{km}(R^{*})_{ms}^{-1}R_{ls}=\tilde{q}_{kl}+p_{kl}\delta_{ab} \tag{D.39}$$

Our goal, in order to find the determinant of the matrix, is to find its eigenvalues and we will do so distinguishing between two different sets of eigenvectors: we will have $K$ eigenvectors not depending on the replica index which we will call $v_k$ and $(n-1)K$ eigenvectors $v_k^a$, which

must satisfy the condition $\sum_a v_k^a = 0$.
Considering the first set

$$\sum_{b,l}(A^{-1})^{ab}_{kl} v_l = \sum_{b,l} \tilde{q}_{kl} v_l + \sum_{b,l} p_k \delta_{ab} v_l = n\sum_{l} \tilde{q}_{kl} v_l + \sum_{l} p_{kl} v_l = \sum_{l}(n\tilde{q}_{kl} + p_{kl})v_l \tag{D.40}$$

we get $K$ different eigenvalues The equation for the second set is

$$\sum_{b,l}(A^{-1})^{ab}_{kl} v_l^b = \sum_{b,l} \tilde{q}_{kl} v_l^b + \sum_{b,l} p_{kl} \delta_{ab} v_l^b = \sum_{l} p_{kl} v_l^a \tag{D.41}$$

which leads to $K$ different eigenvalues each of which is $(n-1)$-fold degenerate. Therefore, the determinant we are looking for will be of the form

$$\det A^{-1} = (\det P)^{n-1} \det\Big(P + n\tilde{Q}\Big) \tag{D.42}$$

which can be expanded for $n \to 0$ thanks to Jacobi's formula and becomes

$$\det A^{-1} \simeq (\det P)^n \left(1 + n\sum_{k,l}(P^{-1})_{kl}\tilde{q}_{kl}\right) \tag{D.43}$$

thus leading to

$$\ln\det A^{-1} \simeq n\left[\ln(\det P) + \sum_{k,l}(P^{-1})_{kl}\tilde{q}_{kl}\right] \tag{D.44}$$

## Energetic term

In order to get to the saddle-point equations, let us work on the energetic term of the partition function.

$$\langle\Delta(\mathbf{x})\rangle = \left\langle \prod_{a,k}\Theta\left(u_k\lambda_k^a\right)\right\rangle = \frac{1}{\int_0^{+\infty} d\mathbf{u} P_{R^*}(\mathbf{u})}\int_0^{+\infty}\left\langle\prod_{a,k}\Theta\left(\lambda_k^a \mid \mathbf{u}\right)\right\rangle P_{R^*}(\mathbf{u})d\mathbf{u} \tag{D.45}$$

where

$$\left\langle\prod_{a,k}\Theta\left(\lambda_k^a \mid \mathbf{u}\right)\right\rangle = \int_0^{+\infty} P(\Lambda \mid \mathbf{u}) d\Lambda = \tag{D.46}$$

$$\begin{aligned}&=(2\pi)^{-\frac{nK}{2}}(\det C)^{-\frac{1}{2}}\times\\ &\times\int_0^{+\infty}\exp\left\{-\frac{1}{2}\sum_{a,b}\sum_{k,l}\left(\lambda_k^a - \mu_k^a\right)\left(C^{-1}\right)^{ab}_{kl}\left(\lambda_l^b - \mu_l^b\right)\right\}\prod_{a,k} d\lambda_k^a\end{aligned} \tag{D.47}$$

Using the Hubbard Stratonovich transformation we are able to write it as

$$\left\langle \prod_{a,k} \Theta\ \ (\lambda_k^a \mid \mathbf{u}) \right\rangle =$$
$$= (2\pi)^{-nK} \int_0^{+\infty} \prod_{a,k} d\lambda_k^a \int \prod_{a,k} d\hat{\lambda}_k^a \exp\left\{ i\sum_{a,k}\hat{\lambda}_k^a(\lambda_k^a-\mu_k^a) - \frac{1}{2} \sum_{a,b}\sum_{k,l}\hat{\lambda}_k^a C_{kl}^{ab}\hat{\lambda}_l^b\right\} \tag{D.48}$$

Using the Replica Symmetric ansatz in Eq.(D.38) we can write the conditional probability $P\left(\Lambda \mid \mathbf{u}\right) \sim \mathcal{N}\left(\mu, C\right)$ where

$$\mu_k^a = \langle \lambda_k^a \mid \mathbf{u}\rangle = \sum_{l,m} R_{kl}^a (R^*)^{-1}_{lm} u_m = \sum_{l,m} R_{kl}(R^*)^{-1}_{lm} u_m = \mu_k \tag{D.49}$$

$$C_{kl}^{ab} = q_{kl}^{ab} - \sum_{m,s} R_{km}^a (R^*)^{-1}_{ms} R_{sl}^b = q_{kl} + p_{kl}\delta_{ab} - \sum_{m,s} R_{km}(R^*)^{-1}_{ms} R_{sl} = \tilde{C}_{kl} + p_{kl}\delta_{ab} \tag{D.50}$$

it becomes

$$\begin{aligned}\left\langle \prod_{a,k} \Theta\left(\lambda_k^a \mid \mathbf{u}\right) \right\rangle =& (2\pi)^{-nK} \int_0^{+\infty} \prod_{a,k} d\lambda_k^a \int \prod_{a,k} d\hat{\lambda}_k^a \exp\left\{ i\sum_{a,k}\hat{\lambda}_k^a(\lambda_k^a-\mu_k^a) + \right. \\ & \left. -\frac{1}{2}\sum_{k,l}\tilde{C}_{kl}\left(\sum_a \hat{\lambda}_k^a\right)\left(\sum_b \hat{\lambda}_l^b\right) - \frac{1}{2}\sum_{a,k,l} p_{kl}\hat{\lambda}_k^a\hat{\lambda}_l^a \right\}\end{aligned} \tag{D.51}$$

Once again exploiting the Hubbard-Stratonovich transformation we get

$$\begin{aligned}\left\langle \prod_{a,k} \Theta\ \ (\lambda_k^a \mid \mathbf{u}) \right\rangle &= (2\pi)^{-nK-\frac{K}{2}} (\det\tilde{C})^{-\frac{1}{2}} \int_0^{+\infty} \prod_{a,k} d\lambda_k^a \int \prod_{a,k} d\hat{\lambda}_k^a \int \prod_k \frac{t_k}{\sqrt{2\pi}} \times \\ & \times \exp\left\{ i\sum_{a,k}\hat{\lambda}_k^a(\lambda_k^a-\mu_k^a) + i\sum_{a,k} t_k\hat{\lambda}_k^a - \frac{1}{2}\sum_{k,l}\left(\tilde{C}^{-1}\right)_{kl} t_k t_l - \frac{1}{2}\sum_{a,k,l} p_{k,l}\hat{\lambda}_k^a\hat{\lambda}_l^a \right\}\end{aligned} \tag{D.52}$$

We can now take outside of the integral the sum over the replica index and perform the $K$-dimensional integration over $\hat{\lambda}_k^a$ thus obtaining

$$\begin{aligned}\left\langle \prod_{a,k} \Theta\left(\lambda_k^a \mid \mathbf{u}\right) \right\rangle =& (2\pi)^{-\frac{(n+1)K}{2}} (\det\tilde{C})^{-\frac{1}{2}} (\det P)^{-\frac{n}{2}} \int \prod_k \frac{dt_k}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\sum_{k,l} t_k \left(\tilde{C}^{-1}\right)_{kl} t_l \right\} \times \\ & \times \left( \int_0^{+\infty} \prod_k d\lambda_k \exp\left\{ -\frac{1}{2}\sum_{k,l} (\lambda_k - \mu_k + t_k)(P^{-1})_{kl}(\lambda_l - \mu_l + t_l) \right\}\right)^n\end{aligned} \tag{D.53}$$

Which allows to write the energetic part as

$$\langle \Delta(\mathbf{x})\rangle = \frac{1}{\int_0^{+\infty} d\mathbf{u} P_{R^*}(\mathbf{u})} \int_0^{+\infty} d\mathbf{u} P_{R^*}(\mathbf{u}) \int d\mathbf{t} P_{\tilde{C}}(\mathbf{t}) H^n\left(\mathbf{t} - \boldsymbol{\mu}, P\right) \tag{D.54}$$

where

$$H\left(\mathbf{t}-\boldsymbol{\mu},P\right):=\int_0^{+\infty}\prod_k\frac{d\lambda_k}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\sum_k(\lambda_k-\mu_k+t_k)(P^{-1})_{kl}(\lambda_l-\mu_l+t_l)\right\}\tag{D.55}$$

and the two probabilities are multivariate normal distributions, $P_{R^*}(\mathbf{u})\sim\mathcal{N}\left(0,R^*\right)$ and $P_{\tilde{C}}(\mathbf{t})\sim\mathcal{N}\left(0,\tilde{C}\right)$.

For $n\to 0$

$$H^n\left(\mathbf{t}-\boldsymbol{\mu},P\right)\simeq 1+n\ln H\left(\mathbf{t}-\boldsymbol{\mu},P\right)\simeq 1+n\ln H\left(\mathbf{t}-\boldsymbol{\mu},P\right)\tag{D.56}$$

which means that we can write

$$\left\langle\prod_{a,k}\Theta\left(\lambda_k^a\mid\mathbf{u}\right)\right\rangle=1+n\int d\mathbf{t}P_{\tilde{C}}(\mathbf{t})\ln H\left(\mathbf{t}-\boldsymbol{\mu},P\right)$$

Then we are left with

$$\begin{aligned}\ln\left\langle\Delta(\mathbf{x})\right\rangle=&\ln\left[1+\frac{n}{\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})}\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})\int d\mathbf{t}P_{\tilde{C}}(\mathbf{t})\ln H\left(\mathbf{t}-\boldsymbol{\mu},P\right)\right]\simeq\\ \simeq& n\frac{\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})\int d\mathbf{t}P_{\tilde{C}}(\mathbf{t})\ln H\left(\mathbf{t}-\boldsymbol{\mu},P\right)}{\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})}\end{aligned}\tag{D.57}$$

Using the results obtained for the entropic and the energetic terms, we can now write

$$\begin{aligned}\left\langle\mathcal{Z}^n\right\rangle\sim\exp\Bigg\{nN\operatorname*{extr}_{Q,R}\Bigg(&\alpha\frac{\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})\int d\mathbf{t}P_{\tilde{C}}(\mathbf{t})\ln H\left(\mathbf{t}-\boldsymbol{\mu},P\right)}{\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})}+\\ &-\ln(\det P)-\sum_{k,l}(P^{-1})_{kl}\tilde{q}_{kl}+\frac{1}{2}-\gamma\frac{1}{2}\sum_{\nu,\rho}\left(Y^\top Y\right)^{-1}_{\nu\rho}-\frac{1}{2}\ln\det\left(\frac{1}{\gamma}Y^\top Y\right)\Bigg)\Bigg\}\end{aligned}\tag{D.58}$$

Using Eq.(D.5) this finally gives

$$\begin{aligned}\frac{1}{N}\left\langle\ln\mathcal{Z}\right\rangle\sim\operatorname*{extr}_{Q,R}\Bigg(&\alpha\frac{\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})\int d\mathbf{t}P_{\tilde{C}}(\mathbf{t})\ln H\left(\mathbf{t}-\boldsymbol{\mu},P\right)}{\int_0^{+\infty}d\mathbf{u}P_{R^*}(\mathbf{u})}+\\ &-\ln(\det P)-\sum_{k,l}(P^{-1})_{kl}\tilde{q}_{kl}+\frac{1}{2}-\gamma\frac{1}{2}\sum_{\nu,\rho}\left(Y^\top Y\right)^{-1}_{\nu\rho}-\frac{1}{2}\ln\det\left(\frac{1}{\gamma}Y^\top Y\right)\Bigg)\end{aligned}\tag{D.59}$$

# Appendix E

# Many-vs-many binding partners data: detailed calculations

Let $X = (\mathbf{x}_1, \ldots, \mathbf{x}_B) \in \mathbb{R}^{N\times B}$ and $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_B) \in \mathbb{R}^{M\times B}$ be data points, with independent components drawn from $x_i, y_j \sim \mathcal{N}(0,1)$. Let $W, W^* \in \mathbb{R}^{N\times M}$ be the interaction matrices of the student and the teacher, respectively. We are interested in

$$\mathcal{Z} = \int \mu\left(dW\right) \prod_{t=1}^{B} \Theta\left(z_t^* \mathbf{x}_t^\top W \mathbf{y}_t\right), \quad z_t^* = \frac{1}{\sqrt{MN}} \mathbf{x}_t^\top W^* \mathbf{y}_t \tag{E.1}$$

where $\mu\left(dW\right) \propto \exp\left\{-\frac{1}{2}\mathrm{tr}\left(W^\top W\right)\right\}$ is a normalised Gaussian measure. Note that $\mathbb{E}[W_{ij}] = 0$ and $\mathbb{E}[W_{ij}^2] = 1$ under $\mu\left(dW\right)$. In particular $\mathbb{E}\left[\mathrm{tr}\left(W^\top W\right)\right] = NM$, and typically $W_{ij} \sim \mathcal{O}(1)$. We assume $\mathrm{tr}\big(W^* W^{*\top}\big) = NM$. In a similar fashion as before, we want to exploit the replica trick to find the typical value of the partition function, therefore, Assuming that the data $(z_t^*, \mathbf{x}_t, \mathbf{y}_t)$ are i.i.d. over $t$

$$\langle \mathcal{Z}^n \rangle_{\mathcal{D}} = \int \left\langle \prod_{a,t} \Theta(z_t^* \mathbf{x}_t^\top W^a \mathbf{y}_t) \right\rangle_{\mathcal{D}} \prod_a \mu\left(dW^a\right) = \tag{E.2}$$

$$= \int \exp\left\{ B \ln \left\langle \prod_a \Theta(z^* \mathbf{x}^\top W^a \mathbf{y}) \right\rangle_{\mathcal{D}} \right\} \prod_a \mu\left(dW^a\right) \tag{E.3}$$

where now $\langle \ldots \rangle_{\mathcal{D}}$ is the average over a single data point triple $(z^*, \mathbf{x}, \mathbf{y})$. we assume here that $\mathbf{x}, \mathbf{y} \sim P(\mathbf{x}, \mathbf{y})$ are normally distributed with independent components of zero mean and unit variance.
Introducing $z^a = \frac{1}{\sqrt{MN}} \mathbf{x}_t^\top W^a \mathbf{y}_t$, and $\mathbf{z} = (z^*, z^1, \ldots, z^n)$, we have

$$\left\langle \prod_a \Theta(z^* \mathbf{x}^\top W^a \mathbf{y}) \right\rangle_{\mathcal{D}} = \left\langle \prod_a \Theta(z^* z^a) \right\rangle_{\mathbf{z}\sim P(\mathbf{z})} = 2 \int_0^{+\infty} P(\mathbf{z}) d\mathbf{z} \tag{E.4}$$

where $P(\mathbf{z})$ is the distribution induced by $\mathbf{x}, \mathbf{y}$. Non-zero contributions come from points $\mathbf{z}$ where all entries have the same sign. The last equality follows from $P(\mathbf{z}) = P(-\mathbf{z})$.
Note that the exact form of $P(\mathbf{z})$ is a complicated distribution, but still convex, hence, for simplicity, we will assume it to be Gaussian with $\langle z^a \rangle = \langle z^* \rangle = 0$ and

$$\langle z^a z^b \rangle = \frac{1}{NM} \sum_{i,j} W_{ij}^a W_{ij}^b, \quad \langle z^a z^* \rangle = \frac{1}{NM} \sum_{i,j} W_{ij}^a W_{ij}^*, \quad \langle z^{*2} \rangle = 1 \tag{E.5}$$

where we can define the overlap parameters as

$$q_{ab} = \frac{1}{NM}\sum_{i,j} W^a_{ij}W^b_{ij}\,, \quad R_a = \frac{1}{NM}\sum_{i,j} W^a_{ij}W^*_{ij} \tag{E.6}$$

Thus $P(\mathbf{z})$ depends on $W^1, \ldots, W^n$ only through $Q = (q^{ab})$ and $R = (R^a)$

$$P(\mathbf{z}) \propto \exp\left\{-\frac{1}{2}\mathbf{z}^\top \Sigma^{-1}\mathbf{z}\right\}\,, \quad \Sigma = \begin{pmatrix} 1 & R^\top \\ R & Q \end{pmatrix} \tag{E.7}$$

Note that $P(\mathbf{z}) = P(z^*)P(z^1, \ldots, z^n \mid z^*)$, where $P(z^*)$ is Gaussian with zero mean and unit variance, while $P(z^1, \ldots, z^n \mid z^*)$ is Gaussian with mean $\langle z^a \mid z^*\rangle = R^a z^*$ and covariance $\left\langle z^a z^b \mid z^*\right\rangle_c = q^{ab} - R^a R^b$.
Next, we write

$$\langle \mathcal{Z}^n\rangle = \int \exp\left\{B\mathcal{E}(R,Q) + NM\mathcal{S}(R,Q)\right\} dR dQ \tag{E.8}$$

where $\mathcal{E}(R,Q) = \ln\left\{2\int_0^{+\infty} P(\mathbf{z})d\mathbf{z}\right\}$ and $\mathcal{S}(R,Q)$ is the entropy

$$\begin{aligned} &e^{NM\mathcal{S}(R,Q)} = \\ &= \int \prod_a \delta\left(R^a - \frac{1}{NM}\operatorname{tr}\left(W^a W^{*\top}\right)\right)\prod_{a\le b}\delta\left(Q^{ab} - \frac{1}{NM}\operatorname{tr}\left(W^a W^{b\top}\right)\right)\prod_a \mu(\,\mathrm{d}W^a) \end{aligned} \tag{E.9}$$

Using Fourier transforms, $\delta(x) = \frac{N}{2\pi}\int e^{iN\xi x}d\xi$

$$\begin{aligned} e^{NMS(Q,R)} &= \int \prod_a \mu(\,\mathrm{d}W^a)\int_{-i\infty}^{i\infty}\prod_{a\le b}\mathrm{d}\hat{q}^{ab}\prod_a \mathrm{d}\hat{R}^a \exp\left\{NM\sum_{a\le b}\hat{q}^{ab}q^{ab} + \right. \\ &\quad \left. + NM\sum_a \hat{R}^a R^a - \sum_{a\le b}\hat{q}^{ab}\operatorname{tr}\left(W^a W^{b\top}\right) - \sum_a \hat{R}^a\operatorname{tr}\left(W^a W^{*\top}\right)\right\} = \\ &= \int_{-i\infty}^{i\infty}\prod_{a\le b}\mathrm{d}\hat{q}^{ab}\prod_a \mathrm{d}\hat{R}_a \exp\left\{NM\sum_{a\le b}\hat{q}^{ab}q^{ab} + NM\sum_a \hat{R}^a R^a + \sum_{ij}\ln Y_{ij}\right\} = \\ &= \int_{-i\infty}^{i\infty}\prod_{a\le b}\mathrm{d}\hat{q}^{ab}\prod_a \mathrm{d}\hat{R}^a \exp\left\{NM\sum_{a\le b}\hat{q}^{ab}q^{ab} + NM\sum_a \hat{R}^a R^a + \right. \\ &\quad \left. + \frac{NM}{2}\sum_{ab}(L^{-1})^{ab}\hat{R}^a\hat{R}^b - \frac{NM}{2}\ln\det(L)\right\} \end{aligned} \tag{E.10}$$

where $L^{ab} = \hat{q}^{ab} + \delta_{ab}\hat{q}^{ab} + \delta_{ab}$ and

$$\begin{aligned} Y_{ij} &= \int \exp\left\{-\frac{1}{2}\sum_a W^{a2} - \sum_{a\le b}\hat{q}_{ab}W^a W^b - \sum_a \hat{R}_a W^a W^*_{ij}\right\}\prod_a \mathrm{d}W^a \\ &\propto \frac{1}{\sqrt{\det(L)}}\exp\left\{\frac{1}{2}W^{*2}_{ij}\sum_{ab}L^{-1}_{ab}\hat{R}_a\hat{R}_b\right\} \end{aligned} \tag{E.11}$$

Note that we omit irrelevant constants and exploit $\sum_{i,j} W_{ij}^{*2} = MN$. Here $Y_{ij}$ is the normalization constant of a Gaussian distribution over $W^1, \ldots, W^n$ with covariance matrix $\langle W^a W^b \rangle_c = (L^{-1})^{ab}$ and averages $\langle W^a \rangle = W_{ij}^* \sum_b (L^{-1})^{ab} \hat{R}_b$. For large $MN$, we get the stationarity equations by extremizing in $\hat{R}^a$, $\hat{q}^{ab}$

$$R^a + \sum_b (L^{-1})^{ab} \hat{R}^b = 0\,, \quad (L^{-1})^{ab} = q^{ab} - R^a R^b \tag{E.12}$$

and $\hat{q}^{ab} = \left(1 - \frac{\delta_{ab}}{2}\right)\left(L^{ab} - \delta_{ab}\right)$. Note that $(L^{-1})^{ab} = \langle z^a z^b \mid z^* \rangle_c$ as we wrote above. Substituting,

$$\mathcal{S}(R, Q) = \frac{n}{2} - \frac{1}{2} \operatorname{tr} Q + \frac{1}{2} \ln \det\left(Q - RR^\top\right) \tag{E.13}$$

Note that we need $B \sim \mathcal{O}(NM)$ to get a nontrivial regime. Each measurement $(z_t^*, \mathbf{x}_t, \mathbf{y}_t)$ gives one inequality $z_t \mathbf{x}_t^\top W \mathbf{y}_t \geq 0$. Since $W$ has $MN$ degrees of freedom, it is reasonable that we need $\mathcal{O}(MN)$ inequalities to get a significant constraint on $W$; then let $\alpha = \frac{B}{MN}$.

## E.1 Replica Symmetric ansatz

We assume the following Replica Symmetric ansatz

$$R_a = r\,, \quad q_{ab} = (1 - \delta_{ab}) q_0 + \delta_{ab} q_1 \tag{E.14}$$

then $P(z^1, \ldots, z^n \mid z^*)$ is a Gaussian with $\langle z^a \mid z^* \rangle = r z^*$ and $\langle z^a, z^b \mid z^* \rangle_c = (1 - \delta_{ab}) q_0 + \delta_{ab} q_1 - r^2$. Next, exploiting the calculations presented in Appendix C, we compute

$$\begin{aligned} \int_0^{+\infty} P(\mathbf{z}) d\mathbf{z} &= \int_0^{+\infty} dz^* P(z^*) \int_0^{+\infty} P(z^1, \ldots, z^n \mid z^*) dz^1 \ldots dz^n = \\ &= \int_0^{+\infty} dz^* \varphi(z^*) \int_{-\infty}^{+\infty} d\zeta \varphi(\zeta) \Phi^n(A\zeta + Bz^*) \end{aligned} \tag{E.15}$$

where $A = A(r, q_0, q_1) = \sqrt{\frac{q_0 - r^2}{q_1 - q_0}}$, $B = B(r, q_0, q_1) = \frac{r}{\sqrt{q_1 - q_0}}$ and

$$\varphi(w) = \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}}\,, \quad \Phi(c) = \int_{-\infty}^{c} \varphi(w) dw = \frac{1}{2}\left\{1 - \operatorname{erf}\left(\frac{c}{\sqrt{2}}\right)\right\} \tag{E.16}$$

For $n \to 0$ we can write

$$\begin{aligned} \mathcal{E}(R, Q) &= \ln\left\{2 \int_0^{+\infty} P(\mathbf{z}) d\mathbf{z}\right\} \simeq \\ &\simeq \ln 2 + \ln \int_0^{+\infty} dz^* \varphi(z^*) \int_{-\infty}^{+\infty} d\zeta \varphi(\zeta) \left\{1 + n \ln \Phi(A\zeta + Bz^*)\right\} \simeq && \text{(E.17)} \\ &\simeq \ln 2 + \ln\left\{\frac{1}{2} + n \int_0^{+\infty} dz^* \varphi(z^*) \int_{-\infty}^{+\infty} d\zeta \varphi(\zeta) \ln \Phi(A\zeta + Bz^*)\right\} = && \text{(E.18)} \\ &= n \tilde{\mathcal{E}}(r, q_0, q_1) && \text{(E.19)} \end{aligned}$$

where

$$\tilde{\mathcal{E}}(r, q_0, q_1) = 2 \int_0^{+\infty} dz^* \varphi(z^*) \int_{-\infty}^{+\infty} d\zeta \varphi(\zeta) \ln \Phi(A\zeta + Bz^*) \tag{E.20}$$

Considering that $\Sigma$ under the RS ansatz has a non-degenerate eigenvalue $d = q_1 + (n-1)q_0 - nr^2$ and a $(n-1)$-degenerate eigenvalue $q_1 - q_0$, for the entropic term, exploiting the properties of Schur's complement, we have $\det\bigl(Q - RR^\top\bigr) = \det\Sigma = (q_1-q_0)^{n-1}d$. Hence

$$\mathcal{S}(Q,R) = \frac{n}{2}(1-q_1) + \frac{1}{2}\ln\bigl(q_1 + (n-1)q_0 - nr^2\bigr) + \frac{n-1}{2}\ln(q_1-q_0)\,.$$

Then, taking the limit for $n \to 0$ we can write

$$\mathcal{S}(R,Q) \simeq n\tilde{\mathcal{S}}\,, \quad \tilde{\mathcal{S}} = \frac{n}{2}\left\{1 - q_1 + \ln(q_1-q_0) + \frac{q_0-r^2}{q_1-q_0}\right\}\,. \tag{E.21}$$

Substituting we get

$$\langle \mathcal{Z}^n\rangle = \exp\left\{nMN\operatorname*{extr}_{Q,R}\left(\alpha\tilde{\mathcal{E}}(R,Q) + \tilde{\mathcal{S}}(R,Q)\right)\right\} \tag{E.22}$$

$$= \exp\left\{nMN\operatorname*{extr}_{r,q_0,q_1}\mathcal{V}(r,q_0,q_1)\right\} \tag{E.23}$$

with $\mathcal{V}(r,q_0,q_1) = \alpha\tilde{\mathcal{E}}(r,q_0,q_1) + \tilde{\mathcal{S}}(r,q_0,q_1)$ for small $n$. Using the definitions of $A(r,q_0,q_1)$ and $B(r,q_0,q_1)$ we can rewrite the entropy in Eq.(E.21) as

$$\tilde{\mathcal{S}}(q_1,A,B) = \frac{1}{2}\left\{1 - q_1 + \ln q_1 - \ln\bigl(1 + A^2 + B^2\bigr) + A^2\right\} \tag{E.24}$$

from which, since $\tilde{\mathcal{E}}$ depends on $r, q_0, q_1$ only through $A$ and $B$, we can determine the optimal value of $q_1$. Extremizing $\tilde{\mathcal{S}}$ with respect to $q_1$ at fixed $A, B$ we get $q_1 = 1$, which implies

$$\tilde{\mathcal{S}}(A,B) = \frac{1}{2}\left\{A^2 - \ln\left(1 + A^2 + B^2\right)\right\} \tag{E.25}$$

We now look at $\mathcal{V}$ as a function of $A$ and $B$. Extremizing, starting from the enregetic term, we get

$$\frac{\partial\tilde{\mathcal{E}}}{\partial A} = 2\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^\infty \mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta + Bz^*\right)}{\Phi\left(A\zeta + Bz^*\right)}\zeta\,, \tag{E.26}$$

$$\frac{\partial\tilde{\mathcal{E}}}{\partial B} = 2\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^\infty \mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta + Bz^*\right)}{\Phi\left(A\zeta + Bz^*\right)}z^*\,. \tag{E.27}$$

Extremizing the entropic term

$$\frac{\partial\tilde{\mathcal{S}}}{\partial A} = A - \frac{A}{1+A^2+B^2}\,, \quad \frac{\partial\tilde{\mathcal{S}}}{\partial B} = -\frac{B}{1+A^2+B^2}\,. \tag{E.28}$$

We then write the saddle point equations as follows:

$$2\alpha\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^\infty \mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta + Bz^*\right)}{\Phi\left(A\zeta + Bz^*\right)}\zeta + A - \frac{A}{1+A^2+B^2} = 0\,, \tag{E.29}$$

$$2\alpha\int_0^\infty \mathrm{d}z^*\varphi\left(z^*\right)\int_{-\infty}^\infty \mathrm{d}\zeta\varphi(\zeta)\frac{\varphi\left(A\zeta + Bz^*\right)}{\Phi\left(A\zeta + Bz^*\right)}z^* - \frac{B}{1+A^2+B^2} = 0\,. \tag{E.30}$$

These equations admit solutions with $q_0 = r$ which means considering the teacher to be equivalent to just another student, which also intuitively should be reasonable, but a more detailed discussion of this argument can be found in [6]. Thanks to this observation we can write $A = \sqrt{r}$ and $B = \frac{r}{\sqrt{1-r}}$. Then, we can focus on one of the two equations and solve parametrically:

$$\alpha = \frac{1}{2} \frac{\frac{A}{1+A^2+B^2} - A}{\int_0^\infty \mathrm{d}z^* \varphi\left(z^*\right) \int_{-\infty}^\infty \mathrm{d}\zeta \varphi(\zeta) \frac{\varphi(A\zeta + Bz^*)}{\Phi(A\zeta+Bz^*)} \zeta} \tag{E.31}$$

for $\alpha$ as a function of $r$. From here we can plot parametrically all the quantities of interest that we have considered.