The Thouless-Anderson-Palmer equation in spin glass theory

Erwin Bolthausen

Contents

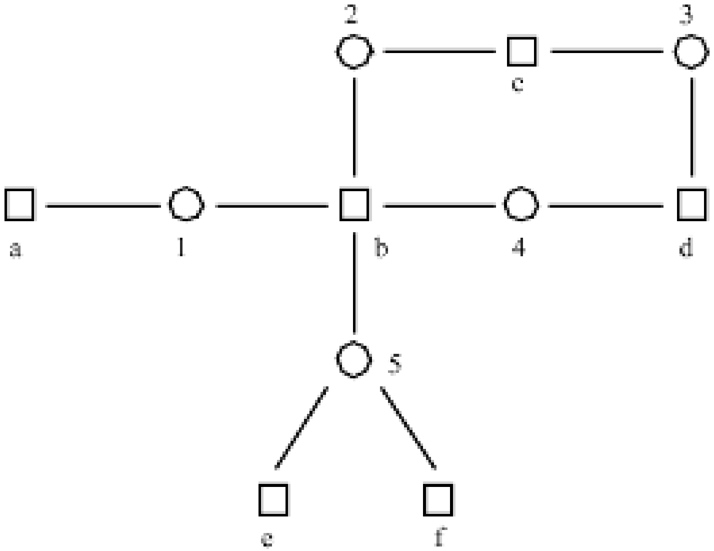
1 Lecture 1: Belief propagation

1.1 Factor graphs

We follow here largely the presentation in [20].

We consider finite bipartite graphs to define special classes of probability measures. The vertex set is divided into two sets V and F where $i \in V$ is called a "variable node", and $a \in F$ is called a "function node". An edge always connects a variable node with a function node. For $a \in F$ we write ∂a for the set of variable nodes which are connected with a, and for $i \in V$, we write ∂i for the set of function nodes connected with i. Usually, we denote variable nodes by i, j or numbers, and function nodes by $a, b, c, \ldots$. Of course, we cannot keep to this "rule" strictly.

We typically depict the function nodes by □, and the variable nodes by ○.



In the example above, we have $\partial a = \{1\}$, $\partial b = \{1, 2, 4, 5\}$, &c, and $\partial 1 = \{a, b\}$, $\partial 5 = \{b, e, f\}$ &c.

We also need a finite set $\mathcal{X}$ which we call the "alphabet". $\mathbf{x} \in \mathcal{X}^V$ is written as $\mathbf{x} = (x_i)_{i \in V}$. To each function node $a \in F$ a function $\psi_a : \mathcal{X}^{\partial a} \to \mathbb{R}^+ := [0, \infty)$ is attached. This setup is collected into an object $\Psi := (V, F, E, \psi)$, where E is the set of edges of the graph, which we call a **factor graph** with alphabet $\mathcal{X}$. For every such factor graph Ψ, we define a probability law P_Ψ on $\mathcal{X}^V$ by

$$P_\Psi(\mathbf{x}) := \frac{1}{Z_\Psi} \prod_{a \in F} \psi_a(\mathbf{x}_{\partial a}) \tag{1.1}$$

which we call the **Gibbs measure** of the factor graph. Here, $\mathbf{x} \in \mathcal{X}^V$ and $\mathbf{x}_{\partial a}$ is the restriction of $\mathbf{x}$ to the set ∂a. Z_Ψ is the normalizing constant:

$$Z_\Psi := \sum_{\mathbf{x} \in \mathcal{X}^V} \prod_{a \in F} \psi_a(\mathbf{x}_{\partial a}).$$

The above definition does not exclude that some variable node $i \in V$ is not connected with any function node, i.e. $\partial i = \emptyset$. This means that $P_\Psi(\mathbf{x})$ does not depend on x_i. It will be convenient to consider also the unnormalized measure

$$C_\Psi(\mathbf{x}) := \prod_{a \in F} \psi_a(\mathbf{x}_{\partial a}), \tag{1.2}$$

so that $P_\Psi(\mathbf{x}) = C_\Psi(\mathbf{x}) / Z_\Psi$. The one-dimensional marginals are defined by

$$C_\Psi^{[i]}(x) := \sum_{\mathbf{x}: x_i = x} C_\Psi(\mathbf{x}), \tag{1.3}$$

but we will also be interested in higher order marginals, for instance $C_\Psi^{[i,j]}$.

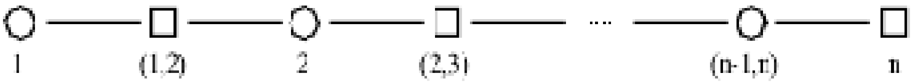
Example 1.1
The simplest case is if any function node a is connected with just one variable. Then P_Ψ is a product measure.

Example 1.2
A less trivial case is the following. Take $V := \{1, \ldots, n\}$ and

$$F := \{(1,2), (2,3), \ldots, (n-1, n)\}.$$

$(i, i+1) \in F$ is connected with i and $i+1 \in V$. Writing the variables as $x_1, \ldots, x_n$ we arrive at a measure

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{i=1}^{n-1} \psi_{(i,i+1)}(x_i, x_{i+1})$$



Exercise 1.3
Prove that Example 1.2 is nothing but a (possibly inhomogeneous) Markov chain. There exists a probability measure ν on $\mathcal{X}$ and stochastic matrices $p_i(x,y)$, $i = 1, \ldots, n-1$, $x, y \in \mathcal{X}$, i.e. $p_i(x,y) \geq 0$, $\sum_y p_i(x,y) = 1$, $\forall x$ with

$$P(\mathbf{x}) = \nu(x_1) p_1(x_1, x_2) p_2(x_2, x_3) \cdot \cdots \cdot p_{n-1}(x_{n-1}, x_n).$$

Example 1.4
A more complicated example is the Ising model. Here $V = \Lambda_N := \{-N, \ldots, N\}^d$, and F is the set of bonds of the lattice Λ_N. $\mathcal{X}$ is $\{-1, 1\}$. The variables we denote by σ_i. $a = \{i, j\} \in F$, is connected to i and to j, i.e. $\partial a = \{i, j\}$.

$$\psi_a(\sigma_i, \sigma_j) = \exp[\beta \sigma_i \sigma_j]$$

with the (inverse) temperature parameter $\beta > 0$.

Exercise 1.5 (Markov Property)
Assume that the factor graph Ψ splits at a variable node i in the sense that the graph splits into $n \geq 2$ disconnected components if we take i out. We write $V^{(1)}, \ldots, V^{(n)}$ for the variable nodes in these components, so that $V = V^{(1)} \cup \cdots \cup V^{(n)} \cup \{i\}$. We write $\pi^{(k)}$ for the projections $\mathcal{X}^V \to \mathcal{X}^{V^{(k)}}$. Prove the following Markov property

$$P_\Psi(\mathbf{x} | \pi_i = x_i) = \prod_{k=1}^n P_\Psi\left(\pi^{(k)} = \mathbf{x}^{(k)} \middle| \pi_i = x_i\right), \tag{1.4}$$

where $P(\cdot | \pi_i = x_i)$ denotes the conditional law given that the i-the component equals x_i, and $\mathbf{x}^{(k)} = (x_i)_{i \in V^{(k)}}$. A consequence is that for any k

$$P_\Psi\left(\pi^{(k)} = \mathbf{x}^{(k)} \middle| \pi_i = x_i,\ \pi^{(\ell)} = \mathbf{x}^{(\ell)},\ \ell \neq k\right) = P_\Psi\left(\pi^{(k)} = \mathbf{x}^{(k)} \middle| \pi_i = x_i\right) \tag{1.5}$$

Exercise 1.6
Prove for an arbitrary factor graph Ψ, the conditional law $P_\Psi(\mathbf{x} | \pi_i = z)$ can be realized as an unconditional one by extending the factor graph in the following way:

$$P_\Psi(\mathbf{x} | \pi_i = z) = P_{\Psi'}(\mathbf{x}),$$

where Ψ' is obtained by adding one function node, call it c, which is connected only with i, and $\psi_c(x_i) = I(x_i = z)$.

The main interest in these lectures will be in cases where the factor graph itself is random: Either the whole graph is random, or the graph is fixed, but the functions attached to the function nodes are random. We always write $(\Omega, \mathcal{F}, \mathbb{P})$ for the probability space governing this randomness. In contrast, we will occasionally write expectations under P_Ψ as $\langle \cdot \rangle_\Psi$ or simply $\langle \cdot \rangle$, of it is clear from the context what Ψ is. These will be considered to be "quenched", i.e. for fixed $\omega \in \Omega$ which governs the randomness of Ψ.

Example 1.7
The Sherrington-Kirkpatrick model has $V = \{1, \ldots, N\}$ and function nodes $\{i, j\}$ for every pair of variable nodes. The $\{i, j\}$ is connected with i and with j. The functions are

$$\psi_{\{i,j\}}(\sigma_i, \sigma_j) := \exp\left[\beta J_{ij} \sigma_i \sigma_j / \sqrt{N}\right]$$

where the J_{ij} are i.i.d. standard Gaussian variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. Then P_Ψ is defined by

$$P_\Psi(\boldsymbol{\sigma}) = \frac{1}{Z_{N,\beta}} \exp\left[\frac{\beta}{\sqrt{N}} \sum_{1 \le i < j \le N} J_{ij}\sigma_i\sigma_j\right],$$

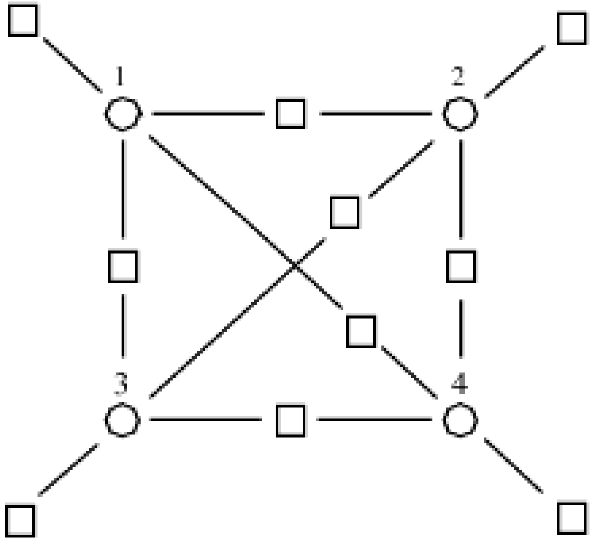
where $\beta > 0$ is the (inverse) temperature parameter. This is then a random probability measure on $\{-1,1\}^N$, formally a Markov kernel from $(\Omega, \mathcal{F})$ to $\{-1,1\}^N$.

One can also include an external field, i.e. consider the measure

$$\frac{1}{Z_{N,\beta,h}} \exp\left[\frac{\beta}{\sqrt{N}} \sum_{1 \le i < j \le N} J_{ij}\sigma_i\sigma_j + h\sum_i \sigma_i\right],$$

where $h \in \mathbb{R}$ (which may also be random, but we usually take it just fixed). This can be achieved by adding a function node a_i to every i, which is just connected with i, and $\psi_{a_i}(\sigma_i) := \mathrm{e}^{h\sigma_i}$.

Below is the drawing of the factor graph for $N = 4$ also with the function nodes for the external field.



1.2 Belief propagation for factor graphs which are trees

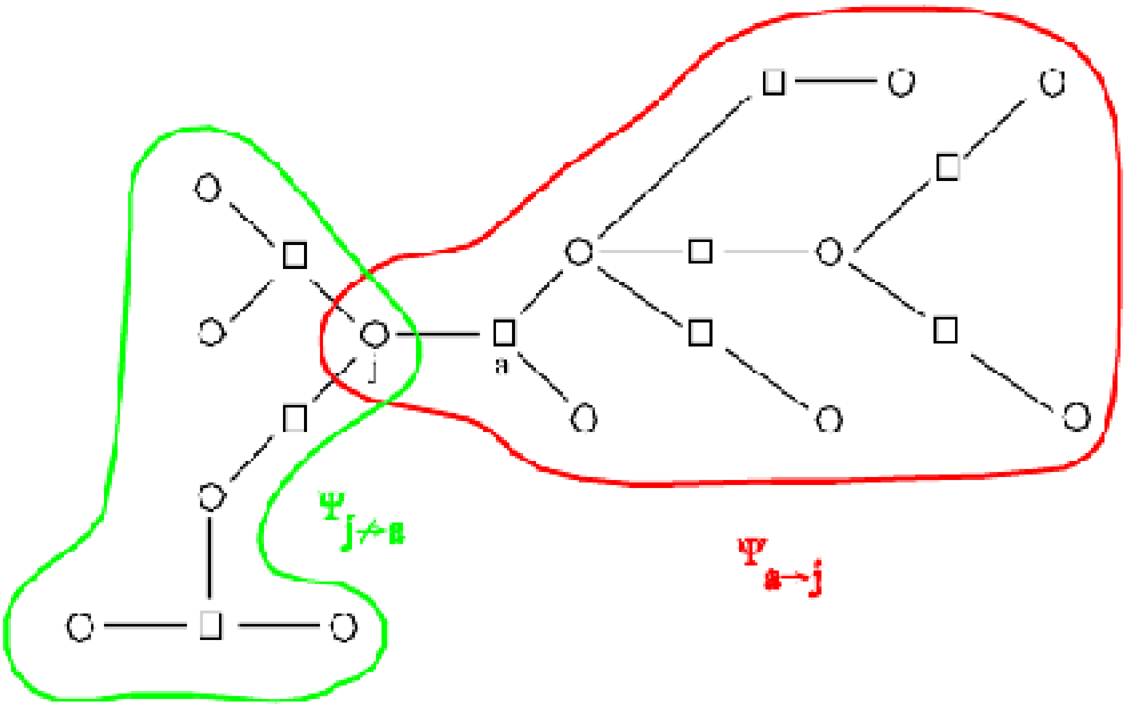
In probabilistic language, belief propagation (abbreviated as BP) is based on conditional distributions: Given a model described by a factor graph, one tries to compute the distribution of π_i at one site $i \in V$ given "the influence" which comes along the graph. We follow here the notation of [20].

We first assume that the factor graph Ψ is a **finite tree**.

We fix $j \in V$ and take an arbitrary $a \in \partial j$. Fixing the pair (j, a) we consider two subtrees of the original one: The first one has the nodes j, a, the edge (j, a), and all the nodes which are connected with j with a self-avoiding path which passes through a. This is evidently a bipartite graph, and to the function nodes in this graph we attach the original ψ-functions. We call this factor graph $\Psi_{a \to j}$ and define the message passage "from a to j" by

$$\hat{\nu}_{a \to j}(x) := C^{[j]}_{\Psi_{a \to j}}(x), \ x \in \mathcal{X}.$$

The second subtree has node j, and all the nodes which can be connected with j through a self-avoiding path which *avoids* a. It can of course happen that the node set consists only of j, but otherwise, the graph is again bipartite, and we denote it by $\Psi_{j \nrightarrow a}$.



Then we define

$$\nu_{j \to a}(x) := C^{[j]}_{\Psi_{j \nrightarrow a}}(x).$$

In the special case where $\Psi_{j \nrightarrow a}$ just consists of j, we define $\nu_{j \to a}(x) = 1$.[1] From

[1] I follow here the notation of [20] which is a bit misleading. $\nu_{j \to a}$ is *not* the influence of j on a, but the influence on j by the part of the graph which avoids a. It would be better to write $\nu_{j \nrightarrow a}$.

the tree structure of the factor graph, it is immediate that $\Psi_{j\nrightarrow a}$ is the union of the graphs $\Psi_{b\rightarrow j}$ for $b \in \partial j\backslash a$ with j as the only common node. In particular, the set of factor nodes in the graphs $\Psi_{b\rightarrow j},\ b \in \partial j\backslash a$, are disjoint. Therefore, the expression (1.3) just factorizes over the different $b \in \partial j\backslash a$, and one gets

$$\nu_{j\rightarrow a}\left(x\right) = \prod_{b\in\partial j\backslash a} \hat{\nu}_{b\rightarrow j}\left(x\right). \tag{1.6}$$

A similar identity expresses the $\hat{\nu}$ in terms of the ν :Writing $\partial a = \{j, k_1, \ldots, k_m\}$, $\Psi_{a\rightarrow j}$ can be presented as the nodes j, a with the edge between them, the disjoint union of the $C_{\Psi_{k_i\nrightarrow a}}$ for $i = 1, \ldots, m$, and the edges between the k_i and a. Then the expression (1.2) for $C_{\Psi_{a\rightarrow j}}$ factorizes into $\psi_a\left(x_j, x_{k_1}, \ldots, x_{k_m}\right)$ and the $C_{\Psi_{k_i\nrightarrow a}}$ and we get

$$\hat{\nu}_{a\rightarrow j}\left(x_j\right) = \sum_{x_{k_1},\ldots,x_{k_m}} \left[\psi_a\left(x_j, x_{k_1}, \ldots, x_{k_m}\right) \prod_{i=1}^{m} \nu_{k_i\rightarrow a}\left(x_{k_i}\right)\right]. \tag{1.7}$$

From the $\hat{\nu}_{a\rightarrow j}$, we can compute the marginal $C_{\Psi}^{[j]}$ by

$$C_{\Psi}^{[j]}\left(x\right) = \prod_{a\in\partial j} \hat{\nu}_{a\rightarrow j}\left(x\right) \tag{1.8}$$

from which we can get the marginal probability measure by normalization:

$$P_{\Psi}\left(\pi_j = x\right) = \frac{\prod_{a\in\partial j} \hat{\nu}_{a\rightarrow j}\left(x\right)}{\sum_{x\in\mathcal{X}} \prod_{a\in\partial j} \hat{\nu}_{a\rightarrow j}\left(x\right)}.$$

The $\nu_{j\rightarrow a}$ and $\hat{\nu}_{a\rightarrow j}$ are usually called “message passages”, and the whole procedure “belief propagation”.

Remark 1.8
Some authors, for instance [10], [11], normalize ν and $\hat{\nu}$ to probability measures. Then the above equations (1.6), (1.7) hold up to normalization.

If the factor graph is a tree, we can express also more complicated marginals through the $\hat{\nu}$. Consider a subset $F' \subset F$ of function notes, and V' the set of variable nodes which are adjacent to F'. With Ψ' we denote the corresponding factor graph and assume it is connected. Further, let ∂' be the set of function notes of Ψ which are *not* in F' but which are adjacent to a variable node $j \in V'$. Evidently, to every $a \in \partial'$ there is a unique element $j \in V' \cap \partial a$. We write $\nu\left(a\right)$ for this element j. This is defined only for $a \in \partial'$. The following expression for the multidimensional marginal on V' is evident from the tree structure.

Lemma 1.9

$$P_\Psi\left(\pi_j = x_j,\ j \in V'\right) = \frac{1}{Z'} \prod_{a \in F'} \psi_a\left(\mathbf{x}_{\partial a}\right) \prod_{a \in \partial'} \hat{\nu}_{a \to \nu(a)}\left(x_{\nu(a)}\right),$$

where Z' is the proper normalization.

We formulate a useful factorization property. We abbreviate $P_\Psi\left(\pi_i = \cdot\right)$ by μ_i. Furthermore, for any function node $a \in F$ we write μ_a for the marginal of P_Ψ on the components in ∂a.

Proposition 1.10
For a finite factor graph which is a tree, one has

$$P_\Psi\left(\mathbf{x}\right) = \prod_{a \in F} \mu_a\left(\mathbf{x}_{\partial a}\right) \prod_{i \in V} \mu_i\left(x_i\right)^{1-|\partial i|}.$$

Proof. We can assume that the graph is connected as otherwise one can prove the identity for the components separately which enter on both sides of the claimed identity in a multiplicative way.

Then we observe that we may assume that the vertices with degree 1 are all variable nodes. In fact if there is a function node with degree 1, its contribution is just that the adjacent variable comes with the the multiplication of the function from the function node, and one can change the functions at one of the other function nodes to take that into account. If the equation is true for the factor graph without this function node, then one quickly checks that it is also true with this additional single argument function.

We use induction on the number m of function nodes. If $m = 1$, then the claim is evident.

We assume $m \geq 2$. By the finiteness of the graph, and the tree property, there exists a variable node which has degree 1, and such that its unique adjacent function node a has the property that all the variable nodes in ∂a have degree one, except one, call it i. We apply the Markov property (1.5). We write π' for the projection onto $\mathcal{X}^{(V \setminus \partial a) \cup \{i\}}$, and π'' for the projection onto $\mathcal{X}^{\partial a \setminus i}$. Then, with $\mathbf{x} = (\mathbf{x}', \mathbf{x}'')$ accordingly

$$\begin{aligned} P_\Psi\left(\mathbf{x}\right) &= P_\Psi\left(\pi' = \mathbf{x}', \pi'' = \mathbf{x}''\right) = P_\Psi\left(\pi'' = \mathbf{x}'' \middle|\, \pi' = \mathbf{x}'\right) P_\Psi\left(\pi' = \mathbf{x}'\right) \\ &= P_\Psi\left(\pi'' = \mathbf{x}'' \middle|\, \pi_i = x_i\right) P_\Psi\left(\pi' = \mathbf{x}'\right) \\ &= \frac{\mu_a\left(\mathbf{x}_{\partial a}\right)}{\mu_i\left(x_i\right)} P_\Psi\left(\pi' = \mathbf{x}'\right). \end{aligned}$$

Up to normalization, the second factor can be written as

$$P_\Psi\left(\pi' = \mathbf{x}'\right) \sim \prod_{b \neq a} \psi_b\left(\mathbf{x}_{\partial b}\right) \phi_a\left(x_i\right)$$

where

$$\phi_a(x_i) := \sum_{\mathbf{x}_{\partial a \setminus \{i\}}} \psi_a(\mathbf{x}).$$

This contribution can however be included by changing the functions at one of the factor nodes $\neq a$ adjacent to i (there is at least one), call it c, by changing the function ψ_c there to

$$\psi_c'(\mathbf{x}_{\partial c}) := \psi_c(\mathbf{x}_{\partial c})\,\phi_a(x_i).$$

The new factor graph has one function node less, i has degree $|\partial i| - 1$, and so we can apply the induction hypothesis, getting

$$P_\Psi(\pi' = \mathbf{x}') = \prod_{b \neq a} \mu_b(\mathbf{x}_{\partial b}) \prod_{j \in V \setminus \partial a} \mu_j(x_j)^{1-|\partial j|}\,\mu_i(x_i)^{2-|\partial i|}.$$

So, we get

$$P_\Psi(\mathbf{x}) = \frac{\mu_a(\mathbf{x}_{\partial a})}{\mu_i(x_i)} P_\Psi(\pi' = \mathbf{x}') = \prod_b \mu_b(\mathbf{x}_{\partial b}) \prod_j \mu_j(x_j)^{1-|\partial j|}$$

as claimed. ■

We can use the above proposition the express the standard entropy

$$H(P_\Psi) := -\sum_{\mathbf{x}} P_\Psi(\mathbf{x}) \log P_\Psi(\mathbf{x})$$

in terms of the local quantities:

$$H(P_\Psi) = \sum_{a \in F} H(\mu_a) + \sum_{i \in V} (1 - |\partial i|)\, H(\mu_i).$$

Another simple consequence:

$$P_\Psi(\mathbf{x}) = \frac{1}{Z_\Psi} \prod_a \psi_a(\mathbf{x}_{\partial a}) = \prod_a \mu_a(\mathbf{x}_{\partial a}) \prod_j \mu_j(x_j)^{1-|\partial j|},$$

and therefore, for any $\mathbf{x}$

$$Z_\Psi = \frac{\prod_a \psi_a(\mathbf{x}_{\partial a})}{\prod_a \mu_a(\mathbf{x}_{\partial a}) \prod_j \mu_j(x_j)^{1-|\partial j|}},$$

$$\log Z_\Psi = \sum_a \log \frac{\psi_a(\mathbf{x}_{\partial a})}{\mu_a(\mathbf{x}_{\partial a})} - \sum_j (1 - |\partial j|) \log \mu_j(x_j).$$

We multiply this equation with $P_\Psi(\mathbf{x})$ and sum over $\mathbf{x}$. This gives

$$\log Z_\Psi = \sum_a \sum_{\mathbf{x}_{\partial a}} \mu_a(\mathbf{x}_{\partial a}) \log \frac{\psi_a(\mathbf{x}_{\partial a})}{\mu_a(\mathbf{x}_{\partial a})} - \sum_j (1 - |\partial j|) \sum_{x_j} \mu_j(x_j) \log \mu_j(x_j) \quad (1.9)$$

We express it in terms of the messages ν and $\hat{\nu}$ in the case of a tree. For that we define for $a \in F,\ i \in V$

$$\begin{aligned} \mathbf{F}_a(\Psi, \boldsymbol{\nu}) &:= \log \sum_{\mathbf{x}_{\partial a}} \psi_a(\mathbf{x}_{\partial a}) \prod_{i \in \partial a} \nu_{i \to a}(x_i), \\ \mathbf{F}_i(\boldsymbol{\nu}) &:= \log \sum_{x_i} \prod_{b \in \partial i} \hat{\nu}_{b \to i}(x_i), \\ \mathbf{F}_{(i,a)}(\boldsymbol{\nu}) &:= \log \sum_{x_i} \nu_{i \to a}(x_i) \hat{\nu}_{a \to i}(x_i). \end{aligned}$$

We define the **Bethe free entropy** in a factor graph Ψ with the collection of beliefs $\boldsymbol{\nu} = \{\nu_{i \to a}, \hat{\nu}_{a \to i}\}_{i \in V, a \in F}$ by

$$\mathbf{F}_*(\Psi, \boldsymbol{\nu}) := \sum_{a \in F} \mathbf{F}_a(\Psi, \boldsymbol{\nu}) + \sum_{i \in V} \mathbf{F}_i(\boldsymbol{\nu}) - \sum_{(i,a) \in E} \mathbf{F}_{(i,a)}(\boldsymbol{\nu}). \quad (1.10)$$

Theorem 1.11
If Ψ is a finite factor graph which is a tree, with the messages $\nu, \hat{\nu}$ given by (1.6) and (1.7). Then

$$\log Z_\Psi = \mathbf{F}_*(\Psi, \boldsymbol{\nu}). \quad (1.11)$$

Proof. We prove it by induction on M, the number of function nodes. In the case $M = 1$, there is just one function node a connected with a number of variable nodes, and the formula is easily checked.

Next assume that $M \geq 2$ and that the formula is proved for factor graphs with $M - 1$ function nodes. We can again assume that the factor graph is connected.

We use a simple pruning argument. Assume first that there is a function node a of degree 1, i.e. which is connected with just one variable node, call it for simplicity 1. As the graph is connected and the number of function nodes is bigger than one, 1 is connected to a number of other function nodes, call them $b_1, \ldots, b_n$. We prune now the original factor graph, call it Ψ to the following factor graph Ψ^* where a is removed. The function in Ψ^* are the same as in Ψ, except of course that ψ_a is missing, and ψ_{b_1} is replaced by

$$\psi^*_{b_1}(x_1, \ldots) := \psi_{b_1}(x_1, \ldots) \psi_a(x_1).$$

It is evident that $Z_\Psi = Z_{\Psi^*}$, and we check that the right hand side of (1.11) is not changed. For that, we check the changes in the ν : For notational simplicity, we write b for b_1.

$$\begin{aligned} \hat{\nu}^*_{b\to 1}(x_1) &= \hat{\nu}_{b\to 1}(x_1)\,\psi_a(x_1)\,, \\ \nu^*_{1\to b} &= \nu_{1\to b}/\psi_a(x_1)\,. \end{aligned}$$

All others remain unchanged. From that, we see that

$$\mathbf{F}_b(\Psi^*) = \log \sum_{\mathbf{x}_{\partial b}} \psi^*_b(\mathbf{x}_{\partial b}) \prod_{i\in\partial b} \nu^*_{i\to b}(x_i) = \mathbf{F}_b(\Psi)\,,$$

and for the other function nodes in Ψ^*, they remain the same. Therefore

$$\sum_{c\in\Psi^*} \mathbf{F}_c(\Psi^*) = \sum_{c\in\Psi} \mathbf{F}_c(\Psi) - \mathbf{F}_a(\Psi) = \sum_{c\in\Psi} \mathbf{F}_c(\Psi) - \log \sum_{x_1} \psi_a(x_1)\,\nu_{1\to a}(x_1)\,.$$

In a similar way, one checks that (the set of variable nodes stays unchanged)

$$\sum_{i\in\Psi^*} \mathbf{F}_i(\Psi^*) = \sum_{i\in\Psi} \mathbf{F}_i(\Psi)\,.$$

Finally, $\mathbf{F}_{(i,c)}(\Psi^*) = \mathbf{F}_{(i,c)}(\Psi)$ for $c \neq a$, but there is in Ψ the additional bond $(1,a)$ which gives

$$\mathbf{F}_{(1,a)}(\Psi) = \log \sum_{x_1} \hat{\nu}_{a\to 1}(x_1)\,\nu_{1\to a}(x_1)\,,$$

so this cancels with the correction in $\sum_{c\in\Psi^*} \mathbf{F}_c(\Psi^*)$. Therefore, the right hand side in (1.11) remains unchanged when switching from Ψ to Ψ^*.

It remains to consider the case when Ψ does not have a function node with degree 1. Then, there is a function node, call it again a, which is connected with a number of variable nodes $i_1, \ldots, i_k$ of degree one, and *one* variable node of higher degree, call this node i. Then prune all the variable nodes $i_1, \ldots, i_k$, but keeping a and i, and the rest. We simply change ψ_a to ψ^*_a by putting

$$\psi^*_a(x_i) := \sum_{x_{i_1},\ldots,x_{i_k}} \psi_a(x_i, x_{i_1}, \ldots, x_{i_k})\,.$$

Then again $\log Z_\Psi = \log Z_{\Psi^*}$, and the reader will easily check that also the rhs of (1.11) remains unchanged.

After this operation, the number of function nodes is unchanged, but we have created one which has degree one, and we can do the pruning of this function node in the way explained before. ■

Up to now, the definition of the ν and the $\hat{\nu}$ was based on the assumption that Ψ has a tree structure. In particular, this was also used heavily for (1.6) and (1.7) and for all the computations above. However, the main applications are for models where the factor graph is not a tree. There is then no clear way how to define message passage quantities $\nu_{i\to a}$, $\hat{\nu}_{a\to i}$ which satisfy (1.6) and (1.7). Nonetheless, it makes sense to ask if such quantities exist, and if yes, how many, and also their relation with other objects of interest like the $\log Z$. Typically, one has to be satisfied with approximately satisfied equations which are exact only in a $N \to \infty$ limit, where N measures the size of the model, usually the number of variable nodes. A serious problem is the question of uniqueness, formulated in an appropriate sense. We will come to such issues in connection with the TAP equations.

One possibility to settle this issue is to define recursively quantities which are defined for an arbitrary finite factor graph. More precisely, we are going to define $\nu_{i\to a}^{(t)}$, $\hat{\nu}_{a\to i}^{(t)}$ for $t \in \mathbb{N}$, $\{i,a\} \in E$, in the following way. The idea is in fact to use (1.6) and (1.7) for a recursive definition. The base is to set

$$\nu_{j\to a}^{(1)}(x) = 1 \tag{1.12}$$

for all $\{a,j\} \in E$, and all $x \in \mathcal{X}$, and then with the notation $\partial a = \{j, k_1, \ldots, k_m\}$

$$\hat{\nu}_{a\to j}^{(t)}(x_j) := \sum_{x_{k_1},\ldots,x_{k_m}} \left[\psi_a(x_j, x_{k_1}, \ldots, x_{k_m}) \prod_{i=1}^{m} \nu_{k_i\to a}^{(t)}(x_{k_i}) \right] \tag{1.13}$$

$$\nu_{j\to a}^{(t+1)}(x_j) := \prod_{b\in\partial j\setminus a} \hat{\nu}_{b\to j}^{(t)}(x_j), \ x_j \in \mathcal{X}. \tag{1.14}$$

The product over the empty set is understood to be 1.

This is evidently well defined for any finite factor graph, and one may ask if there exists a limit

$$\hat{\nu}_{a\to j}^{*}(x_j) := \lim_{t\to\infty} \hat{\nu}_{a\to j}^{(t)}(x_j), \ \nu_{j\to a}^{*}(x_j) := \lim_{t\to\infty} \nu_{j\to a}^{(t)}(x_j),$$

which then certainly will have to satisfy (1.6) and (1.7). Unfortunately, but fairly evidently, such a convergence is restricted again to trees, so nothing seems to be gained. There are however two points:

- Even for finite trees, the recursive construction leads to efficient algorithms to compute the quantities, as we will see in a moment. This has lead to efficient decoding algorithms for a class of codes, a topic which we will not discuss here. For that, see the Chapter 15 in [20].[2]

[2] One should mention that the first use of belief propagation was in coding theory.

- The crucial point is that for large factor graphs of size N, there can be interesting situation where the iterative scheme nearly stabilizes in the $N \to \infty$ limit, and leads to powerful tools to analyze the model. There is a huge literature about this (for instance [22] with applications to neural nets, and [28] for statistical applications, and also the monograph [20]).

For the moment, we just prove the first point, and we show that for a finite factor graph which is a tree, the iteration stabilizes at the correct objects after finitely many iterations. This in fact leads to a polynomial time algorithm for the marginals of P_Ψ in case of a tree.

For a finite factor graph Ψ, we define $\operatorname{diam}(\Psi)$ to be the maximum graph distance between any nodes.

Theorem 1.12
Let the factor graph Ψ be a finite connected tree. Then for $t \geq \operatorname{diam}(\Psi)$

$$\begin{aligned} \nu_{i\to a}^{(t)}(x) &= \nu_{i\to a}(x), \\ \hat{\nu}_{a\to i}^{(t)}(x) &= \hat{\nu}_{a\to i}(x), \end{aligned}$$

for all $\{i,a\} \in E$, and all x.

Proof. We may assume that the graph is connected. We label the *directed* bonds of the graph by natural numbers. For every directed bond, either (i,a) or (a,i), we consider the longest (directed) self-avoiding path in the graph, ending with the last directed bond the given one. Then, the bond gets the level $\ell(i,a)$ (or $\ell(a,i)$) given by the number of bonds in such a path.

From this construction, it is clear that if $\ell(i,a) = n$, then any $b \in F$ with $(b,i) \in E$ has $\ell(b,i) \leq n-1$. The same if $\ell(a,i) = n$, then any $j \in V$ with $(j,a) \in E$ satisfies $\ell(j,a) \leq n-1$.

We now claim that if $t \geq \ell(i,a)$, then $\nu_{i\to a}^{(t)} = \nu_{i\to a}$, and if $t \geq \ell(a,i)$, then $\hat{\nu}_{a\to i}^{(t)} = \hat{\nu}_{a\to i}$. This is proved by induction on ℓ.

First $\ell = 1$: If $\ell(i,a) = 1$, the i has degree 1, and $\nu_{i\to a}(x) = \nu_{i\to a}^{(t)}(x) = 1$ for all t and all x. If $\ell(a,i) = 1$ then a has degree 1, and $\hat{\nu}_{a\to i}^{(t)}(x) = \hat{\nu}_{a\to i}(x) = \psi_a(x)$.

So we assume $\ell = n > 1$. If $\ell(i,a) = n$, then for all b with $(b,i) \in E$, one has $\ell(b,i) \leq n-1$, and by the induction hypothesis, one has for $t \geq n$

$$\nu_{i\to a}^{(t)}(x) := \prod_{b\in\partial i\backslash a} \hat{\nu}_{b\to i}^{(t-1)}(x) = \prod_{b\in\partial i\backslash a} \hat{\nu}_{b\to i}(x) = \nu_{i\to a}(x).$$

Similarly, if $\ell(a,i) = n$, then for all $j \in \partial a\backslash i$ one has $\ell(j,a) \leq n-1$ and one gets for $t \geq n$

$$\begin{aligned}\hat{\nu}_{a\to i}^{(t)}(x_i) \quad : \quad &= \sum_{\mathbf{x}_{\partial a\backslash i}} \left[\psi_a(\mathbf{x}_{\partial a}) \prod_{j\in\partial a\backslash i} \nu_{j\to a}^{(t)}(x_j)\right] \\ &= \sum_{\mathbf{x}_{\partial a\backslash i}} \left[\psi_a(\mathbf{x}_{\partial a}) \prod_{j\in\partial a\backslash i} \nu_{j\to a}(x_j)\right] = \hat{\nu}_{a\to i}(x_i).\end{aligned}$$

This proves the claim. ■

If the factor graph is not a tree, one can still try to set up the above iteration scheme and discuss the question if whether the iteration converge in an approximate sense for large sizes of the graph.

Remark 1.13
Another, even simpler method is to define for any finite factor graph Ψ the measures $\nu_{i\to a}$, $\hat{\nu}_{a\to i}$ on $\mathcal{X}$ in the following way: For $\nu_{i\to a}$ consider the graph Ψ' obtained by removing the node a with all its connections, and in this graph compute $P_{\Psi'}$ as in (1.1), and its one-dimensional marginal on the variable node i. This defines $\nu_{i\to a}$, which is of course normalized to a probability measure. Similarly, for $\hat{\nu}_{a\to i}$, define Ψ'' by removing all function nodes in $\partial i\backslash a$ and compute in this factor graph the marginal at i of the Gibbs measure. One may then ask under which conditions the relations (1.6) and (1.7) are approximately true. (The equations up to normalization, as $\nu_{i\to a}$, $\hat{\nu}_{a\to i}$ are normalized). For such results, see [10], [11].

1.2.1 Simplification for binary function nodes

The scheme can be simplified if the function nodes have all degree 1 or 2. We first assume that the function nodes have all degree 2. A function node of degree 2 connects two variable nodes, say i, j, and we introduce a graph structure with vertex set V by connecting two vertices i, j if in the original factor graph, there is a function node a with $\partial a = \{i, j\}$. We write (ij) for this function node. We denote by $\hat{\partial} i$ the subset of variable nodes which are connected via a function node with i. (Later, we will of course leave out the ^. Here, we just do it to distinguish between the formerly defined ∂i). We then write ψ_{ij} for the function on $\mathcal{X}^2$ attached to the function node (ij). Furthermore, we write $\nu_{i\to j}$ instead of $\nu_{i\to(ij)}$.

Writing out the original belief equations, we get

$$\nu_{i\to j} = \nu_{i,(ij)} = \prod_{k\in\hat{\partial} i\backslash j} \hat{\nu}_{(ki)\to i}$$

$$\hat{\nu}_{(ki)\to i}(x_i) = \sum_{x_k} \psi_{ki}(x_k, x_i)\, \nu_{k\to(ki)}(x_k) = \sum_{x_k} \psi_{ki}(x_k, x_i)\, \nu_{k\to i}(x_k),$$

and so we simply get

$$\nu_{i\to j}\left(x_i\right)=\prod_{k\in\partial i\backslash\{j\}}\sum_{x_k}\psi_{ki}\left(x_k,x_i\right)\nu_{k\to i}\left(x_k\right).$$

In the presence of function nodes ψ_i of degree 1, attached to i, we get

$$\nu_{i\to j}\left(x_i\right)=\psi_i\left(x_i\right)\prod_{k\in\partial i\backslash\{j\}}\sum_{x_k}\psi_{ki}\left(x_k,x_i\right)\nu_{k\to i}\left(x_k\right),$$

and for the recursive scheme, we get

$$\nu_{i\to j}^{(t+1)}\left(x_i\right)=\psi_i\left(x\right)\prod_{k\in\partial i\backslash\{j\}}\sum_{x_k}\psi_{ki}\left(x_k,x_i\right)\nu_{k\to i}^{(t)}\left(x_k\right). \tag{1.15}$$

For the marginals, we get

$$\begin{aligned}\nu_i\left(x_i\right) &: =C_\Psi^{[i]}=\psi_i\left(x_i\right)\prod_{k\in\partial i}\sum_{x_k}\psi_{ki}\left(x_k,x_i\right)\nu_{k\to i}\left(x_k\right) &(1.16)\\ P_\Psi\left(\pi_i=x_i\right) &= \frac{\psi_i\left(x_i\right)\prod_{k\in\partial i}\sum_{x_k}\psi_{ki}\left(x_k,x_i\right)\nu_{k\to i}\left(x_k\right)}{\sum_x\psi_i\left(x_i\right)\prod_{k\in\partial i}\sum_{x_k}\psi_{ki}\left(x_k,x_i\right)\nu_{k\to i}\left(x_k\right)}.\end{aligned}$$

1.3 The general philosophy

Models where the underlying factor graph is a tree are not of great interest. However, many models have factor graphs which are "locally tree like". This means that the total size N (i.e. the number of variable nodes) is large, and one is interested in the $N\to\infty$ limit, and for any variable node i, the part of the factor graph which is at graph distance $\leq n$, is for any fixed n a tree for N sufficiently large. If the graph is random, one has a corresponding probabilistic statement.

An example is the K-SAT problem from theoretical computer science. I explain the simpler (random) XORSAT problem which in contrast to the K-SAT problem is mathematically completely solved. Here, one considers over the field $\mathbb{Z}_2$, M linear equations in N variables. We will take $M=\alpha N$, $\alpha>0$ fixed, and N large. So, the system is

$$\sum_{i=1}^N g_{aj}x_j=b_a,\ a=1,\ldots,M$$

with $g_{aj},b_a\in\{0,1\}$. In the random K-XORSAT problem, one takes for every equation a, K of coefficients g_{aj} equal to 1 and others 0, i.e.

$$g_{aj}=\begin{cases}1 & \text{if } j\in B_a\\ 0 & \text{if } j\notin B_a\end{cases},$$

where $B_a \subset \{1, \ldots, N\}$, $|B_a| = K$, is randomly chosen with probability $\binom{N}{K}^{-1}$, and independently for any of the equations independently. Furthermore, one chooses the b_a according to random coin tossing, or in another version just 0. If $b_a = 0$ for all a, then of course, there is always the 0-solution. But there one still may be interested in the number of solutions. If some of the b_i are 1, then there is the question whether "typically" a solution exists.

The problem can be formulated through a factor graph: The variable nodes are $i \in \{1, \ldots, N\}$, and the function nodes are indexed by the equations. Every function node $a \in \{1, \ldots, M\}$ is connected with the K variable nodes j which satisfy $g_{aj} = 1$, i.e. $\partial a = B_a$. The functions ψ_a are defined by

$$\psi_a \left(\{x_j\}_{j \in B_a} \right) := I \left(\sum\nolimits_{j \in B_a} x_j - b_a = 0 \right),$$

sums computed in $\mathbb{Z}_2$.

In that case, Z_Ψ is the number of solutions of the linear system. Of course, the set of solutions can be determined by Gauss elimination in essentially N^3 steps. An interesting question is however the large N behavior for typical realizations of the system under the above random choice. Z_Ψ is just the number of solutions.

If $M = \alpha N$, K fixed (for instance $K = 3$), then the factor graph is locally tree like in the sense that if one is fixing a node, for instance a function node a, i.e. an equation, then for any n, the probability that the factor graph, restricted to a graph neighborhood of radius n of a, is a tree, converges to 1 as $N \to \infty$.

The model has in fact two critical values:

$$\alpha_d(K) := \sup \left\{ \alpha : x > 1 - \exp\left[-K\alpha x^{K-1} \right], \ \forall x \in (0, 1] \right\}.$$

For instance, $\alpha_d(3) = 0.818469...$. For $\alpha < \alpha_d(K)$, the believe propagation equations have a unique fixed point, and then

$$\log Z_\Psi \approx \mathbf{F}_* (\Psi, \boldsymbol{\nu})$$

for this fixed point $\boldsymbol{\nu}$.

For $\alpha > \alpha_d(K)$, the situation changes drastically, and the believe propagation has many fixed points, despite the fact that the factor graph is still locally tree like. However, in this regime, the loops start the play a considerable role. There are however still methods based on belief propagation which leads to detailed analysis: Essentially, one is then interested in the number of solutions of the belief propagation equations. This can be done by constructing an auxiliary model which has a unique BP fixed point and whose Bethe entropy counts the number of BP fixed points of the original model. This is discussed in Chapter 19 of [20], but would lead too far here. In fact, the mathematical foundation of these ideas is very rudimentary. The random XORSAT is mathematically completely understood, but in related models, for instance in the K-SAT problem, there are many mathematically open problems.

The fact is that the mathematical proofs, if they exist, typically use the intuition coming from the these ideas, but then use hard combinatorial considerations to achieve the result.

2 Lecture 2: The TAP equations

For certain models, there is a possible simplification of the belief propagation equations. This requires that the graph is essentially (nearly) fully connected, but with weak interactions which of course means that the graph is *not* locally tree like. However, the weakness of the interactions suggests that the flux of information through the network is not influenced by local loops in the network. These equations are usually called TAP equations, after Thouless, Anderson, and Palmer, who introduced them first for the SK-model [27]. Variants of these equations have recently found wide applications, also in the analysis of neural nets, and in statistics and compressed sensing. The field is too large to be presented here in any details, and I focus on the SK-model and the perceptron, giving some sketchy information also for the applications to compressed sensing. The chapter here is essentially non-rigorous, giving a heuristic derivation of the equations from the belief propagation equations.

For the SK-model, the validity of the TAP equations has been proved by Talagrand for high temperature in [25], and recently by [1] in the full temperature regime. However, the focus of this minicourse is *not* to use established spin glass theory to prove the TAP equations, but to use the TAP equations to prove properties of the models. For the SK model, there are the interpolation methods first used by Guerra, and very special inequalities, together with the theory developed by Panchenko (see [23]), but up to this date, they seem to be powerless for many models, for instance the perceptron. Also, these methods are somewhat indirect, and don't give much insight what is really happening on the level of Gibbs distribution.

2.1 The Sherrington-Kirkpatrick model

As an example, we write out the belief propagation for the SK-model. We take the model also with an external field. The alphabet is $\mathcal{X} = \{-1, 1\}$ and the set of spin variables is $\{\sigma_i\}_{i=1,\ldots,N}$. The Hamiltonian is given as

$$\beta \sum_{1\leq i<j\leq N}^{N} \frac{J_{ij}}{\sqrt{N}} \sigma_i \sigma_j + h \sum_{i=1}^{N} \sigma_i \tag{2.1}$$

Here h is a real parameter and $\beta > 0$. The J_{ij} are usually assumed to be standard Gaussians, so that we get a factor graphs with random functions. For the moment, we may assume that the J_{ij} are just fixed numbers.

Assuming that we can use (1.16) approximately, for large enough N, we get

$$P(\sigma_i = \pm 1) \approx \frac{\mathrm{e}^{\pm h} \prod_{k:k\neq i} \sum_{\sigma_k} \exp\left[\pm \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k\right] \nu_{k\to i}(\sigma_k)}{\sum_{\tau=\pm 1} \mathrm{e}^{\tau h} \prod_{k:k\neq i} \sum_{\sigma_k} \exp\left[\tau \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k\right] \nu_{k\to i}(\sigma_k)}.$$

We denote by $\nu_{\to i}$ the product measure of the $\nu_{k\to i}$ on $\{\sigma_k\}_{k\neq i}$. As it appears in the numerator and denominator, we may assume that it is normalized to a probability measure. With this notation, we have

$$P(\sigma_i = \pm 1) \approx \frac{\mathrm{e}^{\pm h} E_{\nu_{\to i}} \exp\left[\pm \sum_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k\right]}{\sum_{\tau=\pm 1} \mathrm{e}^{\tau h} E_{\nu_{\to i}} \exp\left[\tau \sum_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k\right]}.$$

Setting $m_i := \langle \sigma_i \rangle = \sum_{\tau=\pm 1} \tau P(\sigma_i = \tau)$, we get

$$m_i \approx \frac{E_{\nu_{\to i}} \sum_{\tau=\pm 1} \tau \mathrm{e}^{\tau h} \exp\left[\tau \sum\limits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k\right]}{E_{\nu_{\to i}} \sum_{\tau=\pm 1} \mathrm{e}^{\tau h} \exp\left[\tau \sum\limits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k\right]} = \frac{E_{\nu_{\to i}} \sinh\left[\sum\limits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k + h\right]}{E_{\nu_{\to i}} \cosh\left[\sum\limits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k + h\right]}. \quad (2.2)$$

We use now the fact that $\nu_{\to i}$ is a product measure. We write $m_{k\to i} := E_{\nu_{\to i}} \sigma_k$. Under the product measure $P_{\nu_{\to i}}$, $\sum_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} (\sigma_k - m_{k\to i})$ is approximately normally distributed with a certain variance, call it $\gamma^2 > 0$ which is of no interest to us. Using this, we get

$$\begin{aligned}
& E_{\nu_{\to i}} \sinh\left[\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} \sigma_k + h\right] \\
= \; & E_{\nu_{\to i}} \sinh\left[\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i} + h + \sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} (\sigma_k - m_{k\to i})\right] \\
= \; & \frac{1}{2}\Big[\exp\left[\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i} + h\right] E_{\nu_{\to i}} \exp\left[\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} (\sigma_k - m_{k\to i})\right] \\
& - \exp\left[-\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i} - h\right] E_{\nu_{\to i}} \exp\left[-\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} (\sigma_k - m_{k\to i})\right] \Big] \\
\approx \; & \frac{1}{2}\left[\exp\left[\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i} + h\right] \mathrm{e}^{\gamma^2/2} - \exp\left[-\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i} + h\right] \mathrm{e}^{\gamma^2/2} \right] \\
= \; & \mathrm{e}^{\gamma^2/2} \sinh\left(\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i} + h\right).
\end{aligned}$$

Here we have used the approximation

$$E_{\nu_{\to i}} \exp\left[\sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}}\left(\sigma_k - m_{k\to i}\right)\right] \approx \frac{1}{\sqrt{2\pi}\gamma}\int \mathrm{e}^{x}\mathrm{e}^{-x^2/2\gamma^2}dx = \mathrm{e}^{\gamma^2/2},$$

and the same for the expression with the minus sign before $\sum_{k:k\neq i}$. Similarly, we get for the denominator in (2.2):

$$\mathrm{e}^{\gamma^2/2}\cosh\left(\sum\nolimits_{k:k\neq i}\frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i} + h\right),$$

and so, $\mathrm{e}^{\gamma^2/2}$ cancels out, and we get

$$m_i \approx \tanh\left(h + \sum\nolimits_{k:k\neq i}\frac{\beta J_{ik}}{\sqrt{N}} m_{k\to i}\right). \tag{2.3}$$

The somewhat fishy point with this pseudoequation is that we don't really know what $m_{k\to i}$ is. A reasonable procedure is to define $\nu_{k\to i}$ with the recursive procedure we have discussed in the last lecture, or the one in Remark 1.13. There is in fact convergence in the high temperature regime, but we will not discuss this here, as we will analyze it for the TAP equations.

The TAP equations are obtained by approximating $m_{k\to i}$ by m_k plus a correction which in fact is of crucial relevance also in the $N \to \infty$ limit. If the graph would be a tree, then we obtain $\nu_{k\to i}$ by putting simply $J_{ki} = 0$. Therefore, using the above "equation" for m_k instead of m_i :

$$\begin{aligned}
m_k &\approx \tanh\left(h + \sum\nolimits_{j:j\neq k}\frac{\beta J_{jk}}{\sqrt{N}} m_{j\to k}\right)\\
&= \tanh\left(h + \sum\nolimits_{j:j\neq i,k}\frac{\beta J_{jk}}{\sqrt{N}} m_{j\to k} + \frac{\beta J_{ik}}{\sqrt{N}} m_{i\to k}\right)\\
&\approx \tanh\left(h + \sum\nolimits_{j:j\neq i,k}\frac{\beta J_{jk}}{\sqrt{N}} m_{j\to k}\right)\\
&\quad + \frac{\beta J_{ik}}{\sqrt{N}} m_{i\to k}\left(1 - \tanh^2\left(h + \sum\nolimits_{j:j\neq i,k}\frac{\beta J_{jk}}{\sqrt{N}} m_{j\to k}\right)\right)\\
&\approx m_{k\to i} + \frac{\beta J_{ik}}{\sqrt{N}} m_{i\to k}\left(1 - m_{k\to i}^2\right).
\end{aligned}$$

The crucial point is that $m_k - m_{k\to i}$ is of order $1/\sqrt{N}$ only, and higher order corrections don't matter in the $N \to \infty$ limit. Also, for the same reason, the $m_{k\to i}$ can be replaced in the correction term by $m_i\left(1 - m_k^2\right)$. Implementing that into (2.3), one arrives at

$$m_i \approx \tanh\left(h + \sum\nolimits_{k:k\neq i}\frac{\beta J_{ik}}{\sqrt{N}} m_k - \beta^2 m_i \sum\nolimits_{k:k\neq i}\frac{J_{ik}^2}{N}\left(1 - m_k^2\right)\right).$$

The next heuristic step is that the direct influence of one single J_{ik} on m_k is again of order $1/\sqrt{N}$, and can be neglected on the correction term, therefore, by a law of large numbers, we can replace the correction term by

$$\beta^2 m_i \left(1 - \frac{1}{N}\sum_{k=1}^{N} m_k^2\right),$$

so we finally arrive at the TAP equation

$$m_i \approx \tanh\left(h + \sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_k - \beta^2 m_i \left(1 - \frac{1}{N}\sum_{k=1}^{N} m_k^2\right)\right).$$

From standard mean-field type models, like the Curie-Weiss model or variants of it, one wouldn't expect the third term inside $\tanh(\cdot)$, but it turns out to be absolutely crucial. It is called the **Onsager correction term.**[3]

Usually one does one further approximation step which is however valid only in the high-temperature regime, namely to apply a law of large number for the average over the m_k^2.[4] Defining

$$q := \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} m_k^2$$

assuming that this q is just a non-random number, we can find an equation for it: Using again (2.3):

$$m_k = \tanh\left(h + \sum\nolimits_{j:j\neq k} \frac{\beta J_{jk}}{\sqrt{N}} m_{j\to k}\right),$$

and using that $m_{j\to k}$ should be independent of J_{jk} and therefore

$$\sum\nolimits_{j:j\neq k} \frac{J_{jk}}{\sqrt{N}} m_{j\to k}$$

should be Gaussian with variance

$$\frac{1}{N}\sum_{j:j\neq k} m_{j\to k}^2 \approx \frac{1}{N}\sum_{j} m_j^2 \approx q.$$

Therefore, q should satisfy

$$q = \int \tanh^2\left(h + \beta\sqrt{q}z\right)\phi\left(dz\right) \tag{2.4}$$

with ϕ the standard normal distribution.

[3] This equations is the only equation I know where three Nobel laureates have been involved: Thouless, Anderson, and Onsager.

[4] The reader should not be overly concerned about the absent of mathematical rigour in these approximations. We will have rigorous versions of these arguments in the next lecture.

Lemma 2.1 (Guerra-Latala)
For any $\beta \geq 0,\ h \neq 0$, the equation (2.4) has a unique solution $q = q(\beta, h) > 0$. For $h = 0$, $q = 0$ is always a solution which is the unique one for $\beta \leq 1$. For $h = 0,\ \beta > 1$, there is one other solution $q(\beta, 0) > 0$.

Proof. [25] Proposition 1.3.8 ■

Using this, one arrives at the TAP equations for the high temperature regime, with a simpler Onsager-correction.

$$m_i \approx \tanh\left(h + \sum\nolimits_{k:k\neq i} \frac{\beta J_{ik}}{\sqrt{N}} m_k - \beta^2 (1-q)\, m_i\right).$$

It is implicitly understood that these approximate equations become true equations, in a way to be made precise, in the $N \to \infty$ limit.

The replacement of $N^{-1}\sum_k m_k^2$ by q, solution of (2.4) is certainly only possible in an appropriately defined high-temperature regime. In physics literature, this high temperature regime is claimed to be the region (β, h) which satisfies

$$\beta^2 \int \frac{\phi(dx)}{\cosh^4\left(h + \beta\sqrt{q}x\right)} \leq 1, \tag{2.5}$$

which is the celebrated **de Almeida - Thouless condition**, **(AT)** for short, which is generally believed (but not completely proved) to be the region where the replica symmetric formula for the free energy holds:

$$\begin{aligned} f(\beta, h) \quad &: \quad = \lim \frac{1}{N} \log Z_N \\ &= \quad \int \log\cosh\left(h + \beta\sqrt{q}z\right)\phi(dz) + \frac{\beta^2(1-q)^2}{4}. \end{aligned}$$

2.2 The perceptron

The perceptron was proposed by Frank Rosenblatt in 1957 as a simple binary classifier. Originally, it was considered as a single layer network. After it was realized that this has severe limitations, it was more or less forgotten. Later, it was realized that multi-layer versions don't have these limitations.

Consider two sets disjoint sets $C_1, C_2 \subset \mathbb{R}^n$ (or $\subset \mathbb{Z}^n$, $\{-1,1\}^n$). Given a point $\mathbf{x} \in C_1 \cup C_2$, the network should be able to "decide" whether $\mathbf{x} \in C_1$ or $\in C_2$. The network does that in a very simple way. For $\mathbf{w} \in \mathbb{R}^n,\ b \in \mathbb{R}$ it decides that $\mathbf{x} \in C_1$ provided $\langle \mathbf{x}, \mathbf{w}\rangle - b \geq 0$ and that $\mathbf{x} \in C_2$ if $\langle \mathbf{x}, \mathbf{w}\rangle - b < 0$, always assuming of course that one knows that $\mathbf{x}$ has to be in one of the two sets. $\langle\cdot,\cdot\rangle$ is the usual inner product. Formally, it is described in the following way: The "input signal" is $\mathbf{x}$ and the output signal is

$$y := 1_{[0.\infty)}\left(\langle \mathbf{x}, \mathbf{w}\rangle - b\right).$$

$\mathbf{w}$ and b are the parameters of the model. It is evident that this is possible only if C_1, C_2 can be separated by a hyperplane. For instance, logical operations AND and OR on $\{0,1\}^2$ can be separated in this way, but XOR not.

However, it can be proved that this drawback disappears with **multi-layer perceptrons**. There, one typically allows in a first layer to compute a sequence

$$y_i := \psi\left(\left\langle \mathbf{x}, \mathbf{w}^{(i)} \right\rangle - b_i\right),\ i = 1, \ldots, m$$

where ψ is a monotone function $\mathbb{R} \to \mathbb{R}$, not necessarily $1_{[0.\infty)}$. $\psi(x) = \tanh(x)$ is often taken, so that the output is again real valued. The $\mathbf{w}^{(i)}$ are $\in \mathbb{R}^n$, and $b_i \in \mathbb{R}$. The output of the first layer is then a vector in $\mathbb{R}^m$. Then, this output $\mathbf{y}$ is taken as the input in another layer which for the simple classification problem gives output $\in \{0,1\}$ through

$$z := 1_{[0.\infty)}\left(\langle \mathbf{y}, \mathbf{v} \rangle - c\right),$$

with $\mathbf{v} \in \mathbb{R}^m, c \in \mathbb{R}$ the parameters for the second level. It can be proved that by taking the parameters of the network appropriately, one can separate arbitrary "nice" sets C_1, C_2.

For reasons which are mathematically not properly understood, networks with many such layers behave algorithmically much better than networks with just two or three layers. Today networks in AI often have dozens of layers.

The main problem is of course to device algorithms which determine the parameters $\mathbf{w}^{(i)}, \mathbf{v}, b_i, c$ of the network. Without going into any details, this is done by a "training phase" where one presents a sequence of input signals $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ to the network together with the information to which of the sets they belong. There are algorithms ("back propagation" and steepest decent) which modify the parameters. Ideally, the network is able to adjust its parameters after mistakes.

We stick here with the single layer perceptron, and ask the following "simple" question: Given are M patterns $(\xi_i^\mu)_{1 \le i \le N}$, $1 \le \mu \le M$, say $\xi^\mu \in \{-1,1\}^N$, one has to find "neural net parameters" w_i, say in $\mathbb{R}$, $i = 1, \ldots, N$, and $b \in \mathbb{R}$, which produce "true" for every pattern ξ^μ, i.e. such that

$$1_{[0.\infty)}\left(\langle \xi^\mu, \mathbf{w} \rangle - b\right) = 1,\ \forall \mu = 1, \ldots, M.$$

In such a case one would say that the one layer perceptron with parameters $\mathbf{w}$ and b has been able to store the M patterns. The question is: Given N, large of course, what is the maximal number of patterns which can be stored. A widely used (theoretical) benchmark for the behavior of such networks is to assume that the NM elements $\xi_i^\mu \in \{-1,1\}$, $1 \le i \le N$, $1 \le \mu \le M$, are just i.i.d. symmetric random variables. Then one can ask about properties which hold with probability $\mathbb{P}$-probability close to 1, where $\mathbb{P}$ governs these random variables. It turns out that the proper dependence of M on N is that $M = \alpha N$ for some parameter, and then it turns out that there is a critical value for α.

The first results were obtained by Elizabeth Gardner (partly together with Derrida) in the late eighties ([15]). Her results are based on replica computations. Later, there was also an approach by Mézard [18] using the so-called cavity method which is close to "belief propagation". However, neither the Gardner approach nor the one by Mézard is rigorous. The only rigorous results obtained so far are the ones by Talagrand (and in a special case by Shcherbina and Tirozzi) which fill about 4 chapters in his two-volume monograph [25] on spin glasses, and which cover part of the claims by the physicists.

Fixing b, and one pattern ξ^μ, the set of $\mathbf{w}$'s which do the job, are the elements of a half space

$$H_{b,\xi^\mu} := \{\mathbf{w} : \langle \xi^\mu, \mathbf{w} \rangle \geq b\},$$

and the question is whether

$$\bigcap_{\mu=1}^{M} H_{b,\xi^\mu} \neq \emptyset,$$

or more generally about the volume of H_{b,ξ^μ}. To speak about the volume one should restrict the set of $\mathbf{w}$'s for instance require $\|\mathbf{w}\| = N$, or $w_i \in \{-1, 1\}$ which are the two cases investigated by Gardner-Derrida.

Mathematically easiest are patterns with continuous instead of discrete components, typically assumed to be standard normal, and the $\mathbf{w} \in \{-1,1\}^N$, but even in this case, a rigorous analysis is restricted to $M = \alpha N$, small α, and I take also $b = 0$. This is not so natural for the neural nets, but makes the analysis easier. (Talagrand also investigated to ± 1 case for the patterns)

Gaussian patterns have the advantage that they lead to a "very simple" geometric question, which is actually in spirit somewhat close to the questions popping up in the problems on compressed sensing. Then the M "Gaussian patterns" define M half spaces $H_\mu := H_{0,\xi^\mu}$ of $\mathbb{R}^N$ which are randomly chosen, independent, and with the uniform distribution.

We have $M = \alpha N$ independent such half spaces and ask if $\bigcap_{\mu=1}^{M} H_i$ contains a point in $\Sigma_N = \{-1,1\}^N$. It turns out that there is a critical value α_{cr} such that for $\alpha < \alpha_{\mathrm{cr}}$, $\mathcal{N}_{N,\alpha} := \left|\bigcap_{\mu=1}^{\alpha N} H_\mu \cap \Sigma_N\right|$ is exponentially growing in N, and if $\alpha > \alpha_{\mathrm{cr}}$, the intersection is $\emptyset$ with high probability (as $N \to \infty$).

The most interesting aspect is the precise evaluation of α_{cr}. Some information can be obtained through a first moment computation.

$$\mathbb{E}\mathcal{N}_{N,\alpha} = \mathbb{E} \sum_{\sigma \in \Sigma_N} I\left(\sigma \in \bigcap\nolimits_{\mu=1}^{[\alpha N]} H_\mu\right) = \sum_{\sigma \in \Sigma_N} \mathbb{P}\left(\sigma \in \bigcap\nolimits_{\mu=1}^{[\alpha N]} H_\mu\right).$$

For any σ, and any μ, the probability is $1/2$ that $\sigma \in H_\mu$. By the independence of the choice of the H_μ, one has

$$\mathbb{P}\left(\sigma \in \bigcap\nolimits_{\mu=1}^{[\alpha N]} H_\mu\right) = 2^{-[\alpha N]},$$

and therefore

$$\mathbb{E}\mathcal{N}_{N,\alpha} = 2^{N-[\alpha N]}.$$

Therefore, if $\alpha > 1$, one has

$$\mathbb{P}\left(\bigcap\nolimits_{\mu=1}^{[\alpha N]} H_\mu \cap \Sigma_N \neq \emptyset\right) = \mathbb{P}\left(\mathcal{N}_{N,\alpha} \geq 1\right) \leq \mathbb{E}\mathcal{N}_{N,\alpha} \to 0.$$

As for $\alpha < 1$, $\mathbb{E}\mathcal{N}_{N,\alpha} \to \infty$ exponentially fast, one may conclude that in this case $\mathbb{P}\left(\bigcap\nolimits_{\mu=1}^{[\alpha N]} H_\mu \cap \Sigma_N \neq \emptyset\right) \to 1$, but this is not the case. In fact

$$\alpha_{\mathrm{cr}} < 1,$$

and so there are $\alpha < 1$ with $\mathbb{E}\mathcal{N}_{N,\alpha} \to \infty$ but $\mathcal{N}_{N,\alpha} \to 0$ in probability.

The approach by Gardner and Derrida [15] was through a mathematically non-rigorous replica computation. Shortly later, Mézard in [18] used a (non-rigorous) version of his cavity to reproof the results. Talagrand was able to give rigorous versions of cavity method in [25], particularly in Chapters 2,3,8,9.

To see the relations with spin glass theory: Let u be a function $\mathbb{R} \to [-\infty, \infty)$, and i.i.d. Gaussians $J_{ij},\ 1 \leq i \leq M = \alpha N,\ 1 \leq j \leq N$, we define

$$\mathcal{N}_{N,u,\alpha} = \sum_\sigma \exp\left[\sum\nolimits_{i=1}^{[\alpha N]} u\left(N^{-1/2} \sum\nolimits_{j=1}^{N} J_{ij}\sigma_j\right)\right],$$

and

$$f\left(\alpha, u\right) = \lim_{N\to\infty} \frac{1}{N} \log \mathcal{N}_{N,u,\alpha}.$$

Evidently, our half-space problem is just the case $u\left(x\right) := -\infty 1_{x<0}$. The correct critical value α_{cr} is then determined by

$$\alpha_{\mathrm{cr}} := \sup\left\{\alpha : f\left(\alpha\right) > 0\right\}.$$

The case $u\left(x\right) = -\infty 1_{x<0}$ makes the analysis difficult, and Talagrand first chose u bounded and smooth, and later used a very delicate approximation of the u by smooth functions. At the end of a rather complicated analysis, he was able to prove the following result, essentially for rather general u including $u\left(x\right) = -\infty 1_{x<0}$:

Theorem 2.2
Under some smallness condition (e.g. for $u = -\infty 1_{x<0}$ that α is small enough, or that $u(x) = v(\beta x)$, v "nice" and β small enough) one has the "replica symmetric" situation, meaning

a) *The fixed point equation for the pair $(q, r) \in [0, 1) \times \mathbb{R}^+$*

$$q = \int \tanh^2\left(\sqrt{\alpha r} z\right) \phi(dz), \; r = \int \psi_q^2\left(\sqrt{q} z\right) \phi(dz) \tag{2.6}$$

with

$$\psi_q(x) := \frac{1}{\sqrt{1-q}} \frac{\int z \exp\left[u\left(x + \sqrt{1-q} z\right)\right] \phi(dz)}{\int \exp\left[u\left(x + \sqrt{1-q} z\right)\right] \phi(dz)} \tag{2.7}$$

has a unique solution.

b)

$$\begin{aligned} f(\alpha, u) &= -\frac{\alpha r(1-q)}{2} + \int \log\cosh\left(\sqrt{\alpha r} z\right) \phi(dz) \\ &\quad + \alpha \int \phi(dz') \log \int \phi(dz) \exp\left[\sqrt{q} z' + \sqrt{1-q} z\right] + \log 2 \end{aligned}$$

c) *For $u = -\infty 1_{x<0}$, $\alpha_{\text{cr}} < 1$.*

Remark 2.3

a) The above formula is the so-called "replica symmetric" formula which can be obtained (and has been obtained by Gardner-Derrida) by a replica computation assuming a replica symmetric ansatz. Talagrand doesn't prove the correctness of the formula up to α_{cr}, but nevertheless he was able to prove c) by an additional argument. Gardner-Derrida argue that the formula is correct up to α_{cr} so that α_{cr} can be determined by the formula. It is understood in the physics literature that the replica symmetric case is the one where the BP equations have a unique solution, and the free energy is given through the Bethe entropy.

b) There exist no results about replica symmetry breaking for this model. There are some results (or considerations) that a proper Parisi-type formula should involve minimax formulae. (see [2])

c) For $u = -\infty 1_{x<0}$ one has

$$\psi_q(x) = \frac{1}{\sqrt{1-q}} \frac{\varphi\left(x/\sqrt{1-q}\right)}{\Phi\left(x/\sqrt{1-q}\right)}$$

with $\varphi,\ \Phi$ the standard normal density, and distribution function. Remark that this function is smooth even though u is of course not. In fact, ψ_q is always smooth provided u satisfies only a growth condition $u(x) \leq C(1+|x|)$.

TAP-type equations had first been given (non-rigorously) by Mézard [18]. There exists no mathematical proof of their validity in the perceptron model, but nonetheless, we will see how to construct solutions by iterations.

We first assume that u is smooth, but it will quickly turn out that this is not of importance. The key point is that one best introduces additional "virtual" variables

$$S_k := \sum_{\mu=1}^{N} \sigma_\mu \frac{J_{\mu k}}{\sqrt{N}},\ k = 1, \ldots, M,$$

so that the Hamiltonian is just $\sum_k u(S_k)$. A somewhat artificial factor graph can be constructed in the following way: The variable nodes are $\mu = 1, \ldots, N$, and $k = 1, \ldots, M$. The variables are $\sigma_i \in \{-1, 1\}$ and $S_k \in \mathbb{R}$. This is not quite covered by the framework of Lecture 1, as the alphabet for the S_k is not finite, but we disregard this point. There is a function node a_k which is connected with k and all of the i, and the corresponding function is

$$\psi_{a_k}(S_k, (\sigma_i)) = I\left(S_k = N^{-1/2} \sum\nolimits_i J_{ik}\sigma_i\right),$$

I denoting the indicator function. Then, there are functions nodes b_k which are just connected with k, with the function $\psi_{b_k}(S_k) = \exp[u(S_k)]$. I am not going to give the heuristic derivation of the TAP equations from the corresponding BP equations. The interested reader may consult the paper by Mézard [22] where the Hopfield net is considered which has a somewhat similar bipartite structure, and where the TAP equations are derived from the BP equations. In principle, we are of course only interested in $m_i = E\sigma_i$, but it turns out that one better relates them to S_k. Fortunately, it suffices to relate them to

$$n_k := Eu'(S_k).$$

The TAP equations then are

$$\begin{aligned} m_i &= \tanh\left(\sum\nolimits_{k=1}^{M} \frac{J_{ik}}{\sqrt{N}} n_k - \alpha m_i \int \psi_q'(\sqrt{q}z)\,\phi(dz)\right), \\ n_k &= \psi_q\left(\sum\nolimits_{i=1}^{N} \frac{J_{ik}}{\sqrt{N}} m_i - n_k(1-q)\right). \end{aligned} \tag{2.8}$$

As usual, they have to be understood to be valid (if at all) in the $N \to \infty$ limit.

An interesting point is that these equations make perfect sense, even if u is not differentiable. Therefore, one can discuss such solutions even for non-differentiable

u disregarding that then, one does know what the n_k really are. This is important, as the equations make perfect sense also in the case $u(x) = -\infty 1_{x<0}$.

One remark: The Onsager term $n_k(1-q)$ in the second TAP equation is just

$$n_k \int \tanh'\left(\sqrt{\alpha r} z\right) \phi(dz).$$

This suggests that we could consider pairs of equations, which come from two arbitrary smooth functions $f, h : \mathbb{R} \to \mathbb{R}$ satisfying growth conditions

$$|f(x)|, \ |h(x)| \leq C(1+|x|)$$

for some $C > 0$, which play the rôle of tanh and ψ_q. The pair (q, r) should satisfy

$$q = \int f^2\left(\sqrt{\alpha r} z\right) \phi(dz), \ r = \int h^2\left(\sqrt{q} z\right) \phi(dz),$$

and the coupled TAP-type equations would look as

$$\begin{aligned} m_i &= f\left(\sum\nolimits_{k=1}^{\alpha N} \frac{J_{ik}}{\sqrt{N}} n_k - \alpha m_i \int h'\left(\sqrt{q} z\right) \phi(dz)\right) \\ n_k &= h\left(\sum\nolimits_{i=1}^{N} \frac{J_{ik}}{\sqrt{N}} m_i - n_k \int f'\left(\sqrt{\alpha r} z\right) \phi(dz)\right). \end{aligned} \tag{2.9}$$

In the next lecture, I will present the scheme how to construct solutions of these equations (in the $N \to \infty$ limit), and in the last lecture I will indicate how to derive the Gardner formula from these equations.

3 Lecture 3: The iterative construction of solutions of the TAP equations

3.1 The Sherrington-Kirkpatrick model

It is quite suggestive to solve the TAP equations with an iterative procedure. There had been a number of suggestions in the literature, but except for very high temperature (i.e. small β), they had not been successful, and therefore, it had been considered to be difficult to construct solutions of the TAP equations directly, even in the replica symmetric regime (see the comments about this point in [21]). The first procedure with convergence in the full high temperature regime was given in [5], which reveals a quite interesting structure, and uses for the mathematical analysis a conditioning technique which has now found other applications, see for instance [3], [28], [4].

The iteration constructs a sequence $\mathbf{m}^{[t]} = \left(m_i^{[t]} \right)_{i=1,\ldots,N}$, $t \in \mathbb{N}$, in the following way:

$$m_i^{[t+1]} := \tanh \left(h + \beta \sum\nolimits_{j=1}^{N} \frac{J_{ij}}{\sqrt{N}} m_j^{[t]} - \beta^2 (1-q) \, m_i^{[t-1]} \right),$$

with the starting condition $m_i^{[0]} = 0$, $m_i^{[1]} = \sqrt{q}$. Here $q = q(\beta, h)$ is always the solution of the fixed point equation (2.4). The fact that one takes the Onsager term one time index back is very crucial for the convergence property.

The difficulty in the analysis of the convergence properties comes from the fact that the $m_j^{[t]}$ depend in a complicated and non-linear way on the matrix (J_{ij}), and so, it is not clear how to proceed. The surprising point is that the "dangerous" part of this dependence is cancelled by the Onsager term, but only if one takes it back one time index.

There is one modification we want to make which is not very important, but which simplifies the computations (unfortunately it complicates the notation). Up to now, we have assumed that $J_{ij} = J_{ji}$, and $J_{ii} = 0$. First of all, it is not important to take $J_{ii} = 0$, as this leads only to a correction of order $1/\sqrt{N}$ which can be neglected. Secondly, it is more convenient to take the N^2 random variables J_{ij} as completely independent standard Gaussians. Then

$$J_{ij}^{\mathrm{sym}} = \frac{J_{ij} + J_{ji}}{\sqrt{2}}$$

is the matrix which enters in the expression for the TAP iteration, so we take it in this form:

$$m_i^{[t+1]} := \tanh \left(h + \beta \sum\nolimits_{j=1}^{N} \frac{J_{ij}^{\mathrm{sym}}}{\sqrt{N}} m_j^{[t]} - \beta^2 (1-q) \, m_i^{[t-1]} \right),$$

where it is now understood that the J_{ij} are i.i.d. This looks to be rather irrelevant, but we will now do the construction based on J_{ij} which leads to slightly different objects than if we would do it directly on J_{ij}^{sym}. The latter was actually done in [5] which contributed to the heavy technicalities there.

As $\mathbf{m}^{[1]} = \sqrt{q}\mathbf{1}$, and $\mathbf{m}^{[0]} = 0$, we have for the $t = 2$ the expression

$$m_i^{[2]} := \tanh \left(h + \beta \sqrt{q} \sum\nolimits_{j=1}^{N} \frac{J_{ij}}{\sqrt{2N}} + \beta \sqrt{q} \sum\nolimits_{j=1}^{N} \frac{J_{ji}}{\sqrt{2N}} \right).$$

We write

$$\xi_i := \sum\nolimits_{j=1}^{N} \frac{J_{ij}}{\sqrt{N}}, \; \eta_i := \sum\nolimits_{j=1}^{N} \frac{J_{ji}}{\sqrt{N}}, \; \zeta_i := \frac{\xi_i + \eta_i}{\sqrt{2}},$$

so that

$$m_i^{[2]} := \tanh\left(h + \beta\sqrt{q}\zeta_i\right).$$

The ξ_i are of course i.i.d. standard Gaussians, as well as the η_i, but there is a slight correlation between ξ_i and $\eta_j : \mathbb{E}\xi_i\eta_j = N^{-1}$. Therefore, also the ζ_i are not totally independent, but the correlations $\mathbb{E}\zeta_i\zeta_j = 1/N,\ i \neq j$, become small for large N.

So far, there is nothing exciting going on, and we go to the next step: There one also has the Onsager term appearing but as we go two time indices back for it, it is still deterministic:

$$m_i^{[3]} := \tanh\left(h + \beta \sum\nolimits_{j=1}^{N} \frac{J_{ij} + J_{ji}}{\sqrt{2N}} m_j^{[2]} - \beta^2 (1-q)\sqrt{q}\right). \tag{3.1}$$

We are now faced with the problem that the $m_j^{[2]}$ already depend in a non-linear way on the J-matrix, but it turns out that it does so in a rather controllable way. As seen from the expression for $\mathbf{m}^{[2]}$, this depends on (J_{ij}) only through the ξ's and η's. By standard linear algebra, we can correct (J_{ij}) by a linear combination of these variables to get it independent of them. In fact, an elementary computation reveals that

$$J_{ij}^{[2]} = J_{ij} - \frac{\xi_i}{\sqrt{N}} - \frac{\eta_j}{\sqrt{N}} + \frac{1}{N^{3/2}} \sum_{i=1}^{N} \xi_i.$$

is independent of $\left(\xi_i, \eta_j\right)$. Actually, it would suffice to get $J_{ij}^{[2]}$ independent of the ζ_i, but it is more convenient to do it in the above form. (At least the formulae are simpler, but I am not completely sure that it doesn't have drawbacks for the later analysis). For the purpose here, the last summand on the rhs is not relevant as it gives a negligible contribution. (It will however be important in the next lecture). Replacing J by $J^{[2]}$, and neglecting the last summand, we get

$$\sum\nolimits_{j=1}^{N} \frac{J_{ij}^{\mathrm{sym}}}{\sqrt{N}} m_j^{[2]} \approx \sum\nolimits_{j=1}^{N} \frac{J_{ij}^{[2]\mathrm{sym}}}{\sqrt{N}} m_j^{[2]} + \sum\nolimits_{j=1}^{N} \frac{\zeta_i + \zeta_j}{N} m_j^{[2]}.$$

The ζ_j are centered Gaussian, but not quite independent:

$$\mathbb{E}\zeta_i^2 = 1 + \frac{1}{N},\ \mathbb{E}\zeta_i\zeta_j = \frac{1}{N}.$$

Despite of these correlations, we can apply a law of large numbers, giving

$$\frac{1}{N}\sum_{j=1}^{N} m_j^{[2]} = \frac{1}{N}\sum_{j=1}^{N} \tanh\left(h + \beta\sqrt{q}\zeta_j\right) \to \int \tanh\left(h + \beta\sqrt{q}z\right)\phi\left(dz\right) =: \gamma_1,$$

$$
\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N}\zeta_j m_j^{[2]} &= \frac{1}{N}\sum_{j=1}^{N}\zeta_j \tanh\left(h+\beta\sqrt{q}\zeta_j\right) \to \int z\tanh\left(h+\beta\sqrt{q}z\right)\phi\left(dz\right) \\
&= \beta\sqrt{q}\int \tanh'\left(h+\beta\sqrt{q}z\right)\phi\left(dz\right) \\
&= \beta\sqrt{q}\int \left[1-\tanh^2\left(h+\beta\sqrt{q}z\right)\right]\phi\left(dz\right) \\
&= \beta\sqrt{q}\left(1-q\right),
\end{aligned}
$$

so this last part miraculously cancels the Onsager part in (3.1). In the end, we get

$$
m_i^{[3]} \approx \tanh\left(h+\beta\sum\nolimits_{j=1}^{N}\frac{J_{ij}^{[2]\mathrm{sym}}}{\sqrt{N}}m_j^{[2]}+\beta\gamma_1\zeta_i\right). \tag{3.2}
$$

I am not precise with the approximations which indeed needs considerable care (not here but in the later iterations), and just pretend that for all the ensuing computations, we can replace $m^{[3]}$ with the right hand side of (3.2). What have we really gained? The fact is: quite a lot. $J_{ij}^{[2]\mathrm{sym}}$ and $m_j^{[2]}$ are now independent. Therefore, conditionally on $\mathcal{F}_1 := \sigma\left(\xi_i,\eta_i,\ 1\le i\le N\right),\ X_i := N^{-1/2}\sum\nolimits_{j=1}^{N}J_{ij}^{[2]\mathrm{sym}}m_j^{[2]}$ is Gaussian. It is not difficult to check that the conditional covariances of the X_i are negligible (or order N^{-1}), so we compute the conditional variance. Fortunately, the covariances of $J_{ij}^{[2]}$ have a simple form

$$
\mathbb{E}\left(J_{ij}^{[2]}J_{st}^{[2]}\right) = \left[\delta_{is}-\frac{1}{N}\right]\left[\delta_{jt}-\frac{1}{N}\right]. \tag{3.3}
$$

The $1/N$-corrections are crucial here. Using the independence of $J^{[2]}$ and $m^{[2]}$, we get

$$
\begin{aligned}
&\mathbb{E}\left(\left(\sum\nolimits_{j=1}^{N}\frac{J_{ij}^{[2]\mathrm{sym}}}{\sqrt{N}}m_j^{[2]}\right)^2\middle|\mathcal{F}_1\right) \\
&= \frac{1}{N}\sum_{j,s}\left[\frac{1}{2}\mathbb{E}\left(J_{ij}^{[2]}J_{is}^{[2]}\right)+\mathbb{E}\left(J_{ij}^{[2]}J_{si}^{[2]}\right)+\frac{1}{2}\mathbb{E}\left(J_{ji}^{[2]}J_{si}^{[2]}\right)\right]m_j^{[2]}m_s^{[2]}.
\end{aligned}
$$

Using (3.3), one sees that the second summand gives only a $1/N$ contribution, and the first and third give the same. So we get

$$
\frac{1}{N}\sum_{j,s}\left[1-\frac{1}{N}\right]\left[\delta_{js}-\frac{1}{N}\right]m_j^{[2]}m_s^{[2]} = \frac{1}{N}\sum_j m_j^{[2]2}-\left(\frac{1}{N}\sum_j m_j^{[2]}\right)^2+O\left(\frac{1}{N}\right).
$$

The crucial point is that although this conditional variance is of course random (and $\mathcal{F}_1$-m.b.), it is not so in the limit, as we can apply the LLN for the expression on the rhs:

$$\frac{1}{N}\sum_j m_j^{[2]2} \approx q,\ \frac{1}{N}\sum_j m_j^{[2]} \approx \gamma_1.$$

Therefore, one has for large N:

$$m_i^{[3]} \approx \tanh\left(h + \beta\sqrt{q-\gamma_1^2}Z_i^{[2]} + \beta\gamma_1\zeta_i\right) \tag{3.4}$$

with independent Gaussians $Z_i^{[2]}$, which are also independent of the ζ_i, the latter, although being slightly correlated, also just behave like i.i.d. standard Gaussians. Using that, one can approximately compute the inner products in $\mathbb{R}^N$:

$$\frac{1}{N}\sum_{i=1}^N m_i^{[3]} \approx \int \tanh\left(h+\beta\sqrt{q}z\right)\phi\left(dz\right) = \gamma_1, \tag{3.5}$$

$$\frac{1}{N}\sum_{i=1}^N m_i^{[3]2} \approx \int \tanh^2\left(h+\beta\sqrt{q}z\right)\phi\left(dz\right) = q. \tag{3.6}$$

Slightly more interesting is

$$\frac{1}{N}\sum_{i=1}^N m_i^{[3]}m_i^{[2]} \approx \int \tanh\left(h+\beta\sqrt{q-\gamma_1^2}z' + \beta\gamma_1 z\right) \\ \times \tanh\left(h+\beta\sqrt{q}z\right)\phi\left(dz\right)\phi\left(dz'\right) =: \rho_2.$$

For later purposes, we write $\zeta^{[1]},\ \xi^{[1]},\ \eta^{[1]}$ for ζ,ξ,η, respectively.

The computations of the inner products can now be used to handle

$$m_i^{[4]} = \tanh\left(h + \beta\sum\nolimits_j \frac{J_{ij}^{\mathrm{sym}}}{\sqrt{N}}m_j^{[3]} - \beta^2\left(1-q\right)m_i^{[2]}\right).$$

which is more interesting as there the Onsager term in the iterative scheme is random. That we take $m^{[2]}$ in this Onsager correction of $m^{[4]}$ is absolutely crucial for the cancellation which would not happen with $m^{[3]}$, and in fact, the convergence property would be considerably worse. I sketch the computation for $m^{[4]}$, as one sees only here what is really happening.

We first replace J by $J^{[2]}$. The corrections are similar as before:

$$\begin{aligned}\sum\nolimits_{j=1}^N \frac{J_{ij}^{\mathrm{sym}}}{\sqrt{N}}m_j^{[3]} &\approx \sum\nolimits_{j=1}^N \frac{J_{ij}^{[2]\mathrm{sym}}}{\sqrt{N}}m_j^{[3]} + \sum\nolimits_{j=1}^N \frac{\zeta_i^{[1]}+\zeta_j^{[1]}}{N}m_j^{[3]} \\ &\approx \sum\nolimits_{j=1}^N \frac{J_{ij}^{[2]\mathrm{sym}}}{\sqrt{N}}m_j^{[3]} + \gamma_1\zeta_i^{[1]} + \frac{1}{N}\sum\nolimits_{j=1}^N \zeta_j^{[1]}m_j^{[3]},\end{aligned}$$

using (3.5), but the last summand is different from the one obtained before. We can however get it again by the type of computation before, using (3.4):

$$\begin{aligned}\frac{1}{N}\sum\nolimits_{j=1}^{N}\zeta_j m_j^{[3]} &\approx \int z\tanh\left(h+\beta\sqrt{q-\gamma_1^2}z'+\beta\gamma_1 z\right)\phi^{\otimes 2}\left(dz,dz'\right)\\ &= \beta\gamma_1\int\left(1-\tanh^2\left(h+\beta\sqrt{q}z\right)\right)\phi\left(dz\right)=\beta\gamma_1\left(1-q\right).\end{aligned}$$

Plugging that in, we arrive at

$$m_i^{[4]}\approx\tanh\left(h+\beta\sum\nolimits_{j=1}^{N}\frac{J_{ij}^{[2]\mathrm{sym}}}{\sqrt{N}}m_j^{[3]}+\gamma_1\zeta_i^{[1]}-\beta^2\left(1-q\right)\left(m_i^{[2]}-\gamma_1\right)\right). \tag{3.7}$$

In this form, we still have a non-trivial dependence between $J^{[2]}$ and $m^{[3]}$ with which it is difficult to see what the behavior is. We cannot argue in exactly the same way as we did before: Although $J^{[2]}$ is still Gaussian ($J^{[3]}$ will no longer be), $m^{[3]}$ is not a simple function (i.e. tanh) of a linear combination of the $J^{[2]}$. The way out is to condition of $\mathcal{F}_1$, i.e. keep the ξ,η,ζ "fixed". Looking at (3.2), one sees that conditionally on $\mathcal{F}_1$, inside tanh for $m^{[3]}$ one has simply a linear combination of the $J^{[2]}$ with coefficients which are $\mathcal{F}_1$-measurable. There is the technical difficulty that the expression on the rhs of (3.2) is only an approximation of $m^{[3]}$, but we neglect this, although it turns out to be quite a nasty difficulty. Anyway, the natural procedure is to correct $J^{[2]}$ to make it *conditionally* independent of $\sum_j J^{[2]\mathrm{sym}}m_j^{[2]}$, conditionally on $\mathcal{F}_1$. This can now be done just by linear algebra again, and the outcome is the matrix $J_{ij}^{[3]}$ which is fairly simple.

$$J_{ij}^{[3]}=J_{ij}^{[2]}-\frac{\xi_i^{[2]}\phi_j^{[2]}}{\sqrt{N}}-\frac{\phi_i^{[2]}\eta_j^{[2]}}{\sqrt{N}}+\frac{\sum_i\xi_i^{[2]}\phi_i^{[2]}}{N^{3/2}}\phi_i^{[2]}\phi_j^{[2]} \tag{3.8}$$

$$\phi_i^{(2)}:=\frac{m_i^{(2)}-\gamma_1}{\sqrt{q-\gamma_1^2}},\ \xi_i^{[2]}:=\sum_j\frac{J_{ij}^{[2]}}{\sqrt{N}}\phi_j^{[2]},\ \eta_j^{[2]}:=\sum_i\phi_i^{[2]}\frac{J_{ij}^{[2]}}{\sqrt{N}}.$$

The matrix $J^{[3]}$ is conditionally independent of $\xi^{[2]},\eta^{[2]}$, conditioned on $\mathcal{F}_1$, and is conditionally Gaussian. However, it is evidently no longer unconditionally Gaussian. The conditional covariances have an explicit expression:

$$\mathbb{E}\left(J_{ij}^{[3]}J_{st}^{[3]}\middle|\mathcal{F}_1\right)=\left[\delta_{is}-\frac{1}{N}-\frac{\phi_i^{[2]}\phi_s^{[2]}}{N}\right]\left[\delta_{jt}-\frac{1}{N}-\frac{\phi_j^{[2]}\phi_t^{[2]}}{N}\right]. \tag{3.9}$$

Remark here that the $\phi^{[2]}$ are functions of $m^{[2]}$, the latter being $\mathcal{F}_1$-m.b.

Replacing now $J_{ij}^{[2]\mathrm{sym}}$ by $J_{ij}^{[3]\mathrm{sym}}$, a simple computation leads to

$$m_i^{[4]} \approx \tanh\left(h + \beta \sum\nolimits_{j=1}^{N} \frac{J_{ij}^{[3]\mathrm{sym}}}{\sqrt{N}} m_j^{[3]} + \beta\gamma_1\zeta_i^{[1]} + \beta\gamma_2\zeta_i^{[2]}\right). \tag{3.10}$$

In fact, the last correction in (3.8) is again negligible, and the first give the additional $\gamma_2\zeta_i^{[2]}$ with

$$\zeta_i^{[2]} = \frac{1}{\sqrt{2}}\left(\xi_i^{[2]} + \eta_i^{[2]}\right),$$

with some (non-random) $\gamma_2 > 0$ satisfying $\gamma_1^2 + \gamma_2^2 < q$, and the second part in (3.8) cancels (miraculously) the rest of the Onsager correction.

The crucial advantage of the expression in (3.10) is the following: Conditionally on $\mathcal{F}_2 := \sigma\left(\xi^{[s]}, \eta^{[s]} : s \leq 2\right)$, the summand $N^{-1/2}\sum_{j=1}^{N} J_{ij}^{[3]\mathrm{sym}} m_j^{[3]}$ is Gaussian. One can compute its variance, using (3.9), which for finite N depends on $\mathcal{F}_2$ in a non-trivial way, but by a LLN as the one discussed before, the variance is non-random in the $N \to \infty$ limit, namely just

$$q - \gamma_1^2 - \gamma_2^2.$$

So, this part becomes in the $N \to \infty$ limit independent of $\mathcal{F}_2$. In a similar way, $\zeta^{[2]}$ is conditionally Gaussian, conditioned on $\mathcal{F}_1$, which in the limit $N \to \infty$ becomes independent (and standard normal) of $\mathcal{F}_1$. In the end, one has

$$m_i^{[4]} \approx \tanh\left(h + \beta\sqrt{q - \gamma_1^2 - \gamma_2^2} Z_i^{[3]} + \beta\gamma_1\zeta_i^{[2]} + \beta\gamma_2\zeta_i^{[1]}\right) \tag{3.11}$$

with (asymptotically) independent standard normal $\zeta_i^{[1]}, \zeta_i^{[2]}, Z_i^{[3]}$. This in turn makes it possible the compute

$$\lim_{N\to\infty} \frac{1}{N} \sum_i m_i^{[s]} m_i^{[t]}$$

for $s, t \leq 4$, which is needed for the next iteration.

Here now is the general scheme: Define sequences $\{\rho_k\}$, $\{\gamma_k\}$ by

$$\gamma_1 = \int \tanh\left(h + \beta\sqrt{q}z\right)\phi\left(dz\right),\ \rho_1 := \sqrt{q}\gamma_1,$$

and recursively

$$\rho_k := \psi\left(\rho_{k-1}\right),\ \gamma_k := \frac{\rho_k - \sum_{j=1}^{k-1}\gamma_j^2}{\sqrt{q - \sum_{j=1}^{k-1}\gamma_j^2}}$$

where $\psi : [0, q] \to (0, q]$ is defined by

$$\begin{aligned}\psi(t) := \int & \tanh\left(h + \beta\sqrt{t}z + \beta\sqrt{q-t}z'\right) \\ \times \tanh & \left(h + \beta\sqrt{t}z + \beta\sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'') .\end{aligned}$$

Remark that $\psi(q) = q$, and $\psi(0) = \gamma_1^2$.

Of crucial importance is the behavior of this function: For any bounded (smooth) function $f : \mathbb{R} \to \mathbb{R}$, one gets, using $\int xg(x)\phi(dx) = \int g'(x)\phi(dx)$

$$\begin{aligned} & \frac{d}{dt}\int f\left(\sqrt{t}z + \sqrt{q-t}z'\right) f\left(\sqrt{t}z + \sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'') \\ = & \int \left(\frac{z}{\sqrt{t}} - \frac{z'}{\sqrt{q-t}}\right) f'\left(\sqrt{t}z + \sqrt{q-t}z'\right) f\left(\sqrt{t}z + \sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'') \\ = & \int f''\left(\sqrt{t}z + \sqrt{q-t}z'\right) f\left(\sqrt{t}z + \sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'') \\ & + \int f'\left(\sqrt{t}z + \sqrt{q-t}z'\right) f'\left(\sqrt{t}z + \sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'') \\ & - \int f''\left(\sqrt{t}z + \sqrt{q-t}z'\right) f\left(\sqrt{t}z + \sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'') \\ = & \int f'\left(\sqrt{t}z + \sqrt{q-t}z'\right) f'\left(\sqrt{t}z + \sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'') \geq 0. \end{aligned}$$

This is strictly positive, unless f is constant. This implies that ψ above is strictly increasing, and applying the same computation to f' instead of f, that f is strictly convex on $[0, q]$. Furthermore

$$\begin{aligned} \left.\frac{d\psi(t)}{dt}\right|_{t=q} &= \beta^2 \int \tanh'\left(h + \beta\sqrt{q}x\right)^2 \phi(dx) \\ &= \beta^2 \int \frac{\phi(dx)}{\cosh^4\left(h + \beta\sqrt{q}x\right)}. \end{aligned}$$

Evidently, this derivative is ≤ 1, i.e. the de Almeida-Thouless condition is satisfied, if and only if $\psi(t) = t$ has the only solution q. It is easy to see that this implies the following result

Lemma 3.1
(2.5) is satisfied, if and only if

$$\lim_{n\to\infty} \rho_n = q$$

which holds, if and only if

$$\sum_{n=1}^{\infty} \gamma_n^2 = q.$$

Similarly, as in (3.11), one gets a representation (asymptotically with $N \to \infty$), for any k

$$m_i^{[k]} \approx \tanh\left(h + \beta\sqrt{q - \sum\nolimits_{s=1}^{k-2} \gamma_s^2} Z_i^{[k-1]} + \beta \sum\nolimits_{s=1}^{k-2} \gamma_s \zeta_i^{[s]}\right), \qquad (3.12)$$

where for $k > \ell$, $\zeta^{[1]}, \ldots, \zeta^{[\ell-1]}$ in the expression for $m_i^{[\ell]}$ are the same as in the expression for $m_i^{[k]}$.

Theorem 3.2
For any β, h, and any k, $Z_i^{[k-1]}, \zeta_i^{[s]}$, $s \leq k-2$ in the representation on the rhs of (3.12) become independent standard normal in the $N \to \infty$ limit.

These results are correct for any β, h. The above Lemma 3.1, then implies that the iteration converges if and only if (2.5) holds, in the form of the following result:

Theorem 3.3
(2.5) is satisfied if and only if

$$\lim_{k,\ell\to\infty} \limsup_{N\to\infty} \mathbb{E}\frac{1}{N}\sum_{i=1}^{N}\left(m_i^{[k]} - m_i^{[\ell]}\right)^2 = 0.$$

Remark 3.4
Although Theorem 3.2 is valid for all β, h, the approach to construct (asymptotically) solutions of the TAP equation is unfortunately still strictly restricted to the high-temperature region. For the belief propagation equations in locally tree like models, there is complicated non-rigorous approach explained in Ch 19 of [20] for the behavior in a "replica symmetry breaking" regime, but this non-rigorous approach is for the time restricted to 1RSB situations. I also don't know if it can be adapted for the TAP equations.

For the discussion in the next lecture, we need the exact scheme how to construct the $\cdot^{[s]}$-objects for general s. First, $\phi^{[s]} \in \mathbb{R}^N$ comes from a Gram-Schmidt orthogonalization out of the $m^{[s]}$:

$$\phi^{[s]} := \frac{m^{[s]} - \sum_{k=1}^{s-1}\left\langle m^{[s]}, \phi^{[k]}\right\rangle \phi^{[k]}}{\left\| m^{[s]} - \sum_{k=1}^{s-1}\left\langle m^{[s]}, \phi^{[k]}\right\rangle \phi^{[k]}\right\|}$$

with $\langle x, y\rangle := N^{-1}\sum_{i=1}^{N} x_i y_i$, $\|x\| := \sqrt{\langle x, x\rangle}$. Assuming that the matrices $J^{[k]} = \left(J_{i,j}^{[k]}\right)$ have been constructed for $k \leq s$, one has

$$\xi_i^{[s]} := \sum_j \frac{J_{ij}^{[s]}}{\sqrt{N}}\phi_j^{[s]},\ \eta_j^{[s]} := \sum_i \phi_i^{[s]}\frac{J_{ij}^{[s]}}{\sqrt{N}},\ \zeta_i^{[s]} := \frac{\xi_i^{[s]} + \eta_i^{[s]}}{\sqrt{2}},$$

and then $J^{[s+1]}$ is conditionally Gaussian, given

$$\mathcal{F}_{s-1} := \sigma\left(\xi^{[k]}, \eta^{[k]} : k \leq s-1\right)$$

with conditional covariances

$$\mathbb{E}\left(J_{ij}^{[s+1]} J_{uv}^{[s+1]} \middle| \mathcal{F}_{s-1} \right) = \left(\delta_{iu} - \frac{1}{N} \sum_{k=1}^{s} \phi_i^{[k]} \phi_u^{[k]} \right) \left(\delta_{jv} - \frac{1}{N} \sum_{k=1}^{s} \phi_j^{[k]} \phi_v^{[k]} \right) \tag{3.13}$$

(The $\phi^{[k]}$ are $\mathcal{F}_{k-1}$-m.b.). The $J^{[s]}$ are concretely constructed from J and the ϕ, ξ, η

$$J_{ij}^{[s]} = J_{ij} - \sum_{k=1}^{s-1} \rho_{ij}^{[k]} \tag{3.14}$$

with

$$\rho_{ij}^{[k]} = \frac{\xi_i^{[k]} \phi_j^{[k]} + \phi_i^{[k]} \eta_j^{[k]}}{\sqrt{N}} - \frac{\phi_i^{[k]} \phi_j^{[k]}}{\sqrt{N}} \left\langle \phi^{[k]}, \xi^{[k]} \right\rangle. \tag{3.15}$$

An important point is that $\phi^{[k]}$ and $m^{[k]}$ are $\mathcal{F}_{k-1}$-m.b. Therefore, the rhs above is conditionally on $\mathcal{F}_{k-1}$ a linear combination of the $\xi^{[k]}, \eta^{[k]}$.

Lemma 3.5
With this construction, $J^{[s+1]}$ is conditionally independent of $\mathcal{F}_s$, given $\mathcal{F}_{s-1}$.

3.2 The perceptron

Surprisingly, the above scheme applies, with modifications of course, to many other situations, including the perceptron [7], the Hopfield net [22], compressed sensing [3], and probably others.

I sketch the approach for the perceptron, but I just take the general situation in (2.9). Then the iteration is

$$\begin{aligned} m_i^{[t+1]} &= f\left(\textstyle\sum_{k=1}^{\alpha N} \frac{J_{ik}}{\sqrt{N}} n_k^{[t]} - \alpha m_i^{[t-1]} \int h'\left(\sqrt{q} z\right) \phi\left(dz\right) \right) \\ n_k^{[t+1]} &= h\left(\textstyle\sum_{i=1}^{N} \frac{J_{ik}}{\sqrt{N}} m_i^{[t]} - n_k^{[t-1]} \int f'\left(\sqrt{\alpha r} z\right) \phi\left(dz\right) \right). \end{aligned}$$

For the starting one takes

$$m_i^{[1]} = \sqrt{q}, \; n_k^{[t]} = \sqrt{r}$$

where q, r satisfy

$$q = Ef^2\left(\sqrt{\alpha r} Z\right), \; r = Eh^2\left(\sqrt{q} Z\right).$$

We don't even have to assume that q, r are uniquely determined by f, h. We just have to assume that we have such q, r. Then, the scheme converges if and only if

$$(\Psi_f \circ \Psi_h)'(r) \leq 1$$

where

$$\begin{aligned}\Psi_f(t) &:= \int f\left(\sqrt{t}z + \sqrt{q-t}z'\right) f\left(\sqrt{t}z + \sqrt{q-t}z''\right) \phi^{\otimes 3}(dz, dz', dz'')\\ \Psi_h(t) &:= \int h\left(\sqrt{t}z + \sqrt{r-t}z'\right) h\left(\sqrt{t}z + \sqrt{r-t}z''\right) \phi^{\otimes 3}(dz, dz', dz''),\end{aligned}$$

and there are completely analogous representations as in the SK-case ([7]).

3.3 Compressed sensing

This is the "classical" regression problem to find estimates for an unknown parameter $x \in \mathbb{R}^N$ based on observations y given as

$$y = Ax + w$$

where A is an $n \times N$ matrix, and w possible noise. The standard (for instance least square) regression procedure requires of course $n > N$. In many modern applications, one however has $n \ll N$, assuming however that the number of relevant parameters x_i is small. More precisely, one assumes that

$$m := \#\{i : x_i \neq 0\}$$

is much smaller than N, but one does not assume any knowledge about $\{i : x_i \neq 0\}$ except its smallness.

There is a tremendous amount of practical, algorithmic, and theoretical work on such problems, and it has been one of the main research topics in statistics over the past 20 years. There are many very sophisticated methods to achieve good estimates of x mainly developed by Dave Donoho, Emmanuel Candès and others. For a survey, see [8]. One of the favorite methods is the "lasso" which is based on minimizing L_1-norms.

An algorithmically fast procedure is a simple iterative scheme, namely to start with an original "guess", $x^{[0]} \in \mathbb{R}^N$ and iteratively define

$$x^{[t+1]} := \eta^{[t]}\left(A^T z^{[t]} + x^{[t]}\right)$$

with A^T the transpose of A, a properly chosen threshold function $\eta^{[t]} : \mathbb{R}^N \to \mathbb{R}$, and the residual

$$z^{[t]} := y - Ax^{[t]}. \tag{3.16}$$

The threshold function is chosen to kick parameters out which are small, and therefore should not have any influence in the end. Evidently, one has to tune this threshold function carefully. Although algorithmically very fast, this iteration proved to be not very satisfactory, both from practical and theoretical viewpoints. For the theoretical sides, one usually assumes a probabilistic structure of A, for instance taking i.i.d. or even Gaussian components, and one lets $N \to \infty$ with n proportional to n, say $n = \delta N$. Furthermore, one assumes that $m = \rho n$, and one wants to decide for which range of ρ, it is possible to identify the parameters in the $N \to \infty$ limit. Furthermore, a "good" algorithm should be able to converge to the right parameters in this region. It turns out that there is a critical value $\rho_{\mathrm{cr}}(\delta)$, such that an identification is possible for $\rho < \rho_{\mathrm{cr}}(\delta)$. However, the above iterative procedure is not able to catch that regardless how the threshold is chosen.

Using BP arguments, Donoho, Maleki, and Montanari [14] developed a "correction" to the above procedure, which is closely related to the Onsager correction: Instead of (3.16), one takes

$$z^{[t]} := y - Ax^{[t]} + \delta^{-1} z^{[t-1]} \left(\eta^{[t]} - \eta^{[t-1]}\right) \left\langle A^T z^{[t-1]} + x^{[t-1]} \right\rangle$$

which numerically lead to much better results. The argument was based on a careful investigation of BP equations, and later, in a number of papers, for instance in [3], [4], a theoretic foundation was found, based on the conditioning method explained above.

4 Lecture 4: Applications of the iterative scheme to evaluate the free energy

I will concentrate on showing how the iterative scheme can be used to rederive the free energy for the SK model in the replica symmetric regime [6]. This is of course far from a new result, and for the case I present, there are quite simple proofs, the simplest one is that by Latala (see [25], Sect 1.4). The method however also probably works for the perceptron [7], where interpolation methods are rather cumbersome, and probably also for other cases, like the Hopfield net or the assignment problem, but this has not been worked out, yet. The method could also give some new results about finite N corrections, but also this is not worked out, yet.

So, we consider the standard **Sherrington-Kirkpatrick model** with an external field having the random Hamiltonian

$$H_{\beta,h}(\boldsymbol{\sigma}) := \frac{\beta}{\sqrt{2}} \sum_{i,j=1}^{N} \frac{J_{ij}}{\sqrt{N}} \sigma_i \sigma_j + h \sum_{i=1}^{N} \sigma_i$$

where $\beta > 0$ and $h \in \mathbb{R}$ are real parameters, $\boldsymbol{\sigma} = (\sigma_i) \in \Sigma_N := \{-1, 1\}^N$, and J_{ij} for i, j are i.i.d. centered Gaussians with variance $1/N$, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The random partition function is

$$Z_{N,\beta,h} := 2^{-N} \sum_{\boldsymbol{\sigma}} \exp\left[H_{\beta,h}(\boldsymbol{\sigma})\right],$$

and the Gibbs distribution is

$$\mathrm{GIBBS}_{N,\beta,h}(\boldsymbol{\sigma}) := \frac{2^{-N}}{Z_{N,\beta,h}} \exp\left[H_{\beta,h}(\boldsymbol{\sigma})\right]. \tag{4.1}$$

It is known that

$$f(\beta, h) := \lim_{N\to\infty} \frac{1}{N} \log Z_{N,\beta,h} = \lim_{N\to\infty} \frac{1}{N} \mathbb{E} \log Z_{N,\beta,h}$$

exists, is non-random, and is given by the Parisi variational formula (see [16], [25], [23]). Furthermore, it is known that for small β, $f(\beta, h)$ is given by the replica-symmetric formula, originally proposed by Sherrington and Kirkpatrick ([26]):

Theorem 4.1
There exists $\beta_0 > 0$ such that for all h, β with $\beta \leq \beta_0$

$$f(\beta, h) = \mathrm{RS}(\beta, h) := \inf_{q \geq 0} \left[\int \log\cosh\left(h + \beta\sqrt{q}x\right) \phi(dx) + \frac{\beta^2 (1-q)^2}{4}\right]. \tag{4.2}$$

Here, ϕ is the standard Gaussian distribution.

For $h \neq 0$, the infimum is uniquely attained at $q = q(\beta, h)$ which satisfies

$$q = \int \tanh^2\left(h + \beta\sqrt{q}x\right) \phi(dx).$$

As stated before, this equation has a unique solution for $h \neq 0$, and for $h = 0$ if $\beta \leq 1$.

$f(\beta, h) = \mathrm{RS}(\beta, h)$ is believed to be true under the de Almeida-Thouless condition (2.5), but this is still an open problem, despite of the fact that it has been proved to be the right condition for the convergence of the TAP iterates in Lecture 3.

At $h = 0$, the AT-condition is $\beta \leq 1$, and in this regime, $f(\beta, 0) = \mathrm{RS}(\beta, 0) = \beta^2/4$ is known since long and can easily be proved by a second moment method. In fact, in this case, the free energy equals the annealed free energy

$$f(\beta, 0) = f_{\mathrm{ann}}(\beta, 0) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{E} Z_{N,\beta,0} = \beta^2/4.$$

This last equation is correct for $h = 0$ and all β. However, one needs $\beta < 1$ to prove that

$$\mathbb{E}Z^2 \leq C(\beta)(\mathbb{E}Z)^2$$

where $C(\beta)$ does not depend on N. Together with a concentration of measure result, this proves $f(\beta, 0) = \beta^2/4$.

It is however easy to see that for $h \neq 0$, and any $\beta > 0$, neither $f(\beta, h)$ nor $\mathrm{RS}(\beta, h)$ equals $f_{\mathrm{ann}}(\beta, h)$.

Our aim is to prove that $f(\beta, h) = \mathrm{RS}(\beta, h)$ can, for small β, be proved by a *conditional* "quenched=annealed" argument, via a second moment method. Roughly speaking, we prove that there is a sub-σ-field $\hat{\mathcal{F}} \subset \mathcal{F}$ such that

$$f(\beta, h) = \lim_{N\to\infty} N^{-1} \log \mathbb{E}\left(Z_N \middle| \hat{\mathcal{F}}\right) = \mathrm{RS}(\beta, h)$$

almost surely, and where we can estimate the conditional second moment by the square of the first one. A key point is the connection with TAP iteration of the last lecture. The reason the method works is that the conditionally annealed Gibbs measure is essentially a Curie-Weiss type model, centered at the solution of the TAP equation, and as such it can be analyzed as a classical mean-field model.

The method is closely related to arguments used for the first time by Morita in [19]. In fact, Morita invented the method to derive the quenched free energy by a partial annealing, fixing part of the Hamilton which is handled in a "quenched way", but where this quenched part can be analyzed much easier than for the full Hamiltonian. This is exactly what we do here by the conditioning. To my knowledge, the method has never been applied to mean-field spin glasses.

From the result we obtain, it easily follows that for small β, the free energy is given by the TAP-variational formula as formulated in [9].

Unfortunately, the argument (for the moment I hope) does not work in the full AT-region.

We actually don't work with a single $\hat{\mathcal{F}} \subset \mathcal{F}$, but instead take the sequence $\{\mathcal{F}_k\}$ of the last lecture. We $\mathbb{E}_k$ for the conditional expectation given $\mathcal{F}_k$. The key result is

Proposition 4.2
If $h > 0$ and β is small enough then

$$\lim_{k\to\infty} \limsup_{N\to\infty} \mathbb{E}\left|\frac{1}{N}\log \mathbb{E}_k(Z_N) - \mathrm{RS}(\beta, h)\right| = 0.$$

Proposition 4.3
Under the same conditions as in Proposition 4.2,

$$\lim_{k\to\infty} \limsup_{N\to\infty} \mathbb{E}\left|\frac{1}{N}\log \mathbb{E}_k\left(Z_N^2\right) - 2\,\mathrm{RS}(\beta, h)\right| \leq 0.$$

Remark 4.4
The requirement on β is rather unsatisfactory. I believe that at least Proposition 4.2 is correct in the full AT-region (2.5).

We give now the proof of Theorem 4.1 based on these propositions.
Proof that the propositions imply Theorem 4.1. We will use that the free energy is self-averaging:

$$\lim_{N\to\infty}\frac{1}{N}\log Z_N=\lim_{N\to\infty}\frac{1}{N}\mathbb{E}\log Z_N, \tag{4.3}$$

assuming the limit on the right hand side exists, which is the result in [16]. This is a simple consequence of the Gaussian isoperimetric inequality, a fact which is well known since long. In fact

$$\left|\frac{1}{N}\log Z_N\left(J\right)-\frac{1}{N}\log Z_N\left(J'\right)\right|\leq\frac{\beta}{\sqrt{2N}}\left\|J-J'\right\|$$

where $\left\|\cdot\right\|$ denotes the Euclidean norm in $\mathbb{R}^{N(N-1)/2}$. Therefore

$$\mathbb{P}\left(\left|\frac{1}{N}\log Z_N-\mathbb{E}\frac{1}{N}\log Z_N\right|\geq t\right)\leq\exp\left[-t^2N/\beta^2\right]$$

This proves the first equality in (4.3).

The upper bound in the second equality of (4.3) follows by Jensen's inequality

$$\limsup_{N\to\infty}\frac{1}{N}\mathbb{E}\log Z_N\leq\limsup_{N\to\infty}\frac{1}{N}\mathbb{E}\log\mathbb{E}_k\left(Z_N\right)$$

for all k. Therefore, by Proposition 4.2,

$$\limsup_{N\to\infty}\frac{1}{N}\mathbb{E}\log Z_N\leq\mathrm{RS}\left(\beta,h\right). \tag{4.4}$$

For the estimate in the other direction, we rely on a second moment argument. For $k,N\in\mathbb{N}$, set $A_{k,N}:=\left\{Z_N\geq\frac{1}{2}\mathbb{E}_k\left(Z_N\right)\right\}$

$$\begin{aligned}\mathbb{E}_k\left(Z_N\right) &= \mathbb{E}_k\left(Z_N;A_{k,N}^c\right)+\mathbb{E}_k\left(Z_N;A_{k,N}\right)\\ &\leq \frac{1}{2}\mathbb{E}_k\left(Z_N\right)+\sqrt{\mathbb{E}_k\left(Z_N^2\right)\mathbb{P}_k\left(A_{k,N}\right)}\end{aligned}$$

and therefore

$$\mathbb{P}_k\left(A_{k,N}\right)\geq\frac{\mathbb{E}_k\left(Z_N\right)^2}{4\mathbb{E}_k\left(Z_N^2\right)}. \tag{4.5}$$

Using Proposition 4.3, we can for an arbitrary $\varepsilon > 0$ choose k large enough, and given such a k, we find $N_0(\varepsilon, k)$ such that for $N \geq N_0$

$$\mathbb{P}\left(\frac{\mathbb{E}_k(Z_N)^2}{4\mathbb{E}_k(Z_N^2)} \geq \mathrm{e}^{-\varepsilon N}\right) \geq \frac{1}{2},$$

and therefore, by (4.5), and the definition of $A_{k,N}$,

$$\mathbb{P}\left(\mathbb{P}_k\left(\frac{1}{N}\log Z_N \geq \frac{1}{N}\log \mathbb{E}_k(Z_N) - \frac{\log 2}{N}\right) \geq \mathrm{e}^{-\varepsilon N}\right) \geq \frac{1}{2},$$

By (**??**), we find for any $\varepsilon' > 0$, a $c(\varepsilon') > 0$ and a $k_0(\varepsilon') \in \mathbb{N}$ such that for $k \geq k_0(\varepsilon')$, we find $N_0'(\varepsilon', k)$ such that for $N \geq N_0'$, we have

$$\mathbb{P}\left(\frac{1}{N}\log \mathbb{E}_k(Z_N) \geq \mathrm{RS}(\beta, h) - \frac{\varepsilon'}{2}\right) \geq \frac{3}{4},$$

and $N^{-1}\log 2 \leq \varepsilon'/2$. Therefore, for $N \geq \max(N_0', N_0)$

$$\mathbb{P}\left(\mathbb{P}_k\left(\frac{1}{N}\log Z_N \geq \mathrm{RS}(\beta, h) - \varepsilon'\right) \geq \mathrm{e}^{-\varepsilon N}\right) \geq \frac{1}{4},$$

implying by the Markov inequality

$$\mathbb{P}\left(\frac{1}{N}\log Z_N \geq \mathrm{RS}(\beta, h) - \varepsilon'\right) \geq \frac{1}{4}\mathrm{e}^{-\varepsilon N}. \tag{4.6}$$

By Gaussian isoperimetry, we have for any $\eta > 0$ and large enough N

$$\mathbb{P}\left(\left|\frac{1}{N}\log Z_N - \frac{1}{N}\mathbb{E}\log Z_N\right| \leq \eta\right) \geq 1 - \exp\left[-\eta N/\beta^2\right].$$

If we choose $\varepsilon < \eta/\beta^2$, it follows that for N large enough one has

$$\frac{1}{N}\mathbb{E}\log Z_N \geq \mathrm{RS}(\beta, h) - \varepsilon' - \eta$$

and as η and ε' are arbitrary, we get

$$\liminf_{N\to\infty} \frac{1}{N}\mathbb{E}\log Z_N \geq \mathrm{RS}(\beta, h).$$

Together with (4.4), this proves (4.3). ■

I sketch how to prove the above propositions which is quite straightforward, given the iterative construction. The basic idea is to "trust" the physicists that the TAP solutions *are* the means of the σ's, and that under the Gibbs distribution, the σ_i are not too far away from coin tossing with the tilts from the m_i. However, globally, there is an important difference to tilted coin tossing which is also reflected in the fact that the TAP equations need the Onsager correction.

$$Z_N := \sum_{\sigma} 2^{-N} \exp\left[\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij}\sigma_i\sigma_j + h\sum_i \sigma_i\right].$$

Now, we take the TAP solutions:

$$\begin{aligned} h_i &: = h + \beta \textstyle\sum_j \frac{J_{ij}^{\mathrm{sym}}}{\sqrt{N}} m_j - \beta^2 (1-q)\, m_i \\ m_i &= \tanh(h_i) \end{aligned}$$

I am bit sloppy here: In all I am doing below, one has to take the iterations $m^{[k]}$ for finite k, lets first have $N \to \infty$, and $k \to \infty$ afterwards. I am pretending here that one can take $k = \infty$ and do the $N \to \infty$ limit afterwards. So, there are some estimates to do which I put under the carpet.

The tilted coin tossing is given by

$$p_i(\sigma_i) = \frac{1}{2} \frac{\exp[h_i\sigma_i]}{\cosh(h_i)}.$$

Then

$$Z_N = \prod_{i=1}^{N} \cosh(h_i)\, Z_N'$$

with

$$Z_N' := \sum_{\sigma} p(\sigma) \exp\left[\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij}\sigma_i\sigma_j - \beta\sum_{i,j} \sigma_i \frac{J_{ij}^{\mathrm{sym}}}{\sqrt{N}} m_j + \beta^2(1-q)\sum \sigma_i m_i\right].$$

(The J_{ij} here are completely i.i.d. Gaussians).

Given the iterative scheme, it is easy to prove that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log\cosh(h_i) = \int \log\cosh(h + \beta\sqrt{q}x)\, \phi(dx),$$

so we obtain the first part of the RS-solution, and it remains to prove that

$$\lim \frac{1}{N} \log Z_N' = \frac{\beta^2(1-q)}{4}.$$

We center the σ_i putting $\hat{\sigma}_i := \sigma_i - m_i$.

$$Z'_N = \sum_\sigma p(\sigma) \exp\left[\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij} \hat{\sigma}_i \hat{\sigma}_j - \frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij} m_i m_j + \beta^2 (1-q) \sum \sigma_i m_i\right]$$

The middle term can be evaluated from the iterative scheme and is easily proved to be

$$\frac{1}{\sqrt{N}} \sum_{i,j} J_{ij} m_i m_j = \sqrt{2} \beta N q (1-q) + o(N).$$

Therefore

$$\begin{aligned} Z'_N &\approx \sum_\sigma p(\sigma) \exp\left[\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij} \hat{\sigma}_i \hat{\sigma}_j - N\beta^2 q (1-q) + \beta^2 (1-q) \sum \sigma_i m_i\right] \\ &= \sum_\sigma p(\sigma) \exp\left[\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij} \hat{\sigma}_i \hat{\sigma}_j + N\beta^2 (1-q) \langle \hat{\sigma}, m \rangle\right] \end{aligned}$$

$$\begin{aligned} \frac{\beta^2}{4} N \|\hat{\sigma}\|^4 &= \frac{\beta^2}{4N} \sum_{i,j} \hat{\sigma}_i^2 \hat{\sigma}_j^2 = \frac{\beta^2}{4N} \sum_{i,j} \left(1 - m_i^2 - 2\hat{\sigma}_i m_i\right) \left(1 - m_j^2 - 2\hat{\sigma}_j m_j\right) \\ &= \frac{N\beta^2 (1-q)^2}{4} - \frac{\beta^2}{4N} \sum_{i,j} \left(1 - m_i^2\right) 2\hat{\sigma}_j m_j \\ &\quad - \frac{\beta^2}{4N} \sum_{i,j} \left(1 - m_j^2\right) 2\hat{\sigma}_i m_i + \frac{\beta^2}{4N} \sum_{i,j} 2\hat{\sigma}_i m_i 2\hat{\sigma}_j m_j + o(N) \\ &= \frac{N\beta^2 (1-q)^2}{4} - N\beta^2 (1-q) \langle \hat{\sigma}, m \rangle + \beta^2 N \left(\frac{1}{N} \sum_i \hat{\sigma}_i m_i\right)^2 + o(N). \end{aligned}$$

Plugging that in, one gets

$$Z'_N = \exp\left[\frac{N\beta^2 (1-q)^2}{4}\right] Z''_N \mathrm{e}^{o(N)}$$

with

$$Z''_N := \sum_\sigma p(\sigma) \exp\left[\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij} \hat{\sigma}_i \hat{\sigma}_j - \frac{\beta^2}{4} N \|\hat{\sigma}\|^4 + \beta^2 N \left(\frac{1}{N} \sum_i \hat{\sigma}_i m_i\right)^2\right],$$

and we have to prove

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}_\infty Z''_N = 0.$$

If we could just regard the J_{ij} as i.i.d. and independent of the m_i, then the first and the second summand inside $\exp[\cdot]$ would cancel, and we would be left with a the last summand which is just a Curie-Weiss type expression. We however have to take the proper conditional distributions of the J_{ij} given in the last lecture. First, we shift J to $J^{[k]}$ whose conditional distribution given $\mathcal{F}_{k-1}$ we know. We are a bit formal pretending that we can take $k = \infty$.

$$\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij} \hat{\sigma}_i \hat{\sigma}_j = \frac{\beta}{\sqrt{2N}} \sum_t \sum_{i,j} \rho_{ij}^{[t]} \hat{\sigma}_i \hat{\sigma}_j + \frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij}^{[\infty]} \hat{\sigma}_i \hat{\sigma}_j$$

Then

$$\mathbb{E}_\infty \exp\left[\frac{\beta}{\sqrt{2N}} \sum_{i,j} J_{ij}^{[\infty]} \hat{\sigma}_i \hat{\sigma}_j\right] = \exp\left[\frac{\beta^2}{4N} \mathbb{E}_\infty \left(\sum_{i,j} J_{ij}^{[\infty]} \hat{\sigma}_i \hat{\sigma}_j\right)^2\right]$$

which, given (3.13) can be computed as

$$\frac{\beta^2}{4} N \left[\|\hat{\sigma}\|^2 - \sum_t \left\langle \hat{\sigma}, \phi^{[t]} \right\rangle^2\right]^2 \leq N \|\hat{\sigma}\|^4, \qquad (4.7)$$

and so

$$\mathbb{E}_\infty Z_N'' \leq \sum_\sigma p(\sigma) \exp\left[\frac{\beta}{\sqrt{2N}} \sum_t \sum_{i,j} \rho_{ij}^{[t]} \hat{\sigma}_i \hat{\sigma}_j + \beta^2 N \langle \hat{\sigma}, m \rangle^2 + o(N)\right],$$

and a lower bound, taking the lhs of (4.7) into account. Taking the explicit expression of the $\rho^{[t]}$ into account, one has

$$\frac{1}{\sqrt{2N}} \sum_{i,j} \rho_{ij}^{[t]} \hat{\sigma}_i \hat{\sigma}_j = N \left\langle \hat{\sigma}, \zeta^{[t]} \right\rangle \left\langle \hat{\sigma}, \phi^{[t]} \right\rangle + O\left(\sqrt{N}\right).$$

The crucial point is that all terms which are linear in $\langle \hat{\sigma}, x \rangle$ with $x \in \mathbb{R}^N$ measurable with respect to TAP, have canceled, and what remains are quadratic terms. These can be handled with a Curie-Weiss type estimate, provided that β is small enough giving

$$\mathbb{E}_\infty Z_N' \leq \exp[o(N)]$$

The lower bound is easy it works in the full AT-region, but for the upper bound, I have not (yet) been able to prove that this holds for the full high-temperature region.

The conditional second moment can be handled essentially in the same way.

Remark 4.5

a) The most challenging problem is if this approach is restricted to high temperature. For BP equations on locally tree like graphs, there is a (non-rigorous) method to count the relevant solutions of the BP-equations, but it seems also to be restricted to 1RSB cases ([20] Ch. 19). It would be interesting to try to apply that to TAP equations.

   The SK-model is certainly the wrong one to try something like that as it is (supposed to be) full RSB as soon as the AT-line is crossed.

   In the recent paper [18], Mézard derives (heuristically) for the Hopfield model TAP equations also in the so-called "retrieval phase" which is not high-temperature, and uses iterations similar to the ones used here, and numerical experiments.

a) A somewhat puzzling point is that the above approach is non-trivial at $h = 0$, where the TAP-solutions are of course $m_i = 0$, and quenched=annealed is easy. If one take a small $h > 0$, and computes the ξ, η, then they don't go to 0 for $h \to 0$. So, the $\mathcal{F}_k$ stay non-trivial. One may regard that as a sign that the method is not the "right one", or also the contrary: For $h = 0$, one has $N^{-1} \log \mathbb{E} Z_N \to \beta^2/4$ for all β, but certainly $N^{-1} \log \mathbb{E}_\infty Z_N \to \beta^2/4$ can be true at most up to AT-line, i.e. $\beta = 1$.

c) It is tempting to conjecture that the random Curie-Weiss terms give a description for the finite N correction in the high-temperature region, i.e. that the Gibbs distribution is in leading order coin tossing with tilts to the m_i with corrections which are described by the series of random Curie-Weiss terms.

d) For the perceptron, the formal derivation of the Gardner formula works as well, except that we have not yet been able to handle the Curie-Weiss type correction when replacing J with $J^{[k]}$. I hope that the proof will be more transparent than Talagrand's. The latter is based on approximations of $u = -\infty 1_{x<0}$ by smooth functions, and careful estimates of the errors.

References

[1] Auffinger, A., and Jagannath, A.: *Thouless-Anderson-Palmer equations for conditional Gibbs measures in the generic p-spin glass model.* Preprint arXiv:1612.06359

[2] Barra, A., Genovese, G., and Guerra F. *Equilibrium statistical mechanics of bipartite spin systems.* J. Phys **A 44**, 245002 (2011).

[3] Bayati, M. Lelarge, M., and Montanari, A.: *Universality in polytope phase transitions and message passing algorithms.* Ann. Appl. Prob. **25**, 753-822 (2015).

[4] Berthier, R., Montanari, A., and Nguyen, P.-M.: *State evolution for approximate message passing with non-separable functions.* arXiv 1708.03950v1.

[5] Bolthausen, E.: *An iterative construction of solutions of the TAP equations for the Sherrington-Kirkpatrick model.* Comm. Math. Phys. **325**, 333-366 (2014).

[6] Bolthausen, E.: *A Morita type proof of the replica symmetric formula for SK.* Preprint

[7] Bolthausen, E., and Nakajima, S.: *On the Gardner formula and the TAP equations for the perceptron.* In preparation.

[8] Candès, E.: *Mathematics of sparsity (and a few other things).* Proceedings of the International Congress of Mathematicians, Seoul, South Korea, 2014.

[9] Chen, W.-K., and Panchenko, D.: *On the TAP free energy in the mixed p-spin models.* Preprint arXiv:1709.03468

[10] Coja-Oghlan, A, and Perkins, W.: *Belief propagation on replica symmetric random factor graph models.* Proc. 20th RANDOM **27**, 1-15 (2016).

[11] Coja-Oghlan, A, and Perkins, W.: *Bethe states of random factor graphs.* Preprint arXiv: 1709.03827v1.

[12] Donoho, D.: *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension.* Discrete Comput. Enom. **35**, 617-652 (2006).

[13] Donoho, D. and Tanner, J.: *Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing.* Phil. Trans. R. Soc. A **367,** 4273-4293 (2009).

[14] Donoho, D., Maleki, A., and Montanari, M.: *Message Passing Algorithms for Compressed Sensing.* PNAS **106**, 18914-18919 (2009).

[15] Gardner, E., and Derrida, B.: *Optimal storage properties of neural network models.* J. Phys. A: Math. Gen. **21**, 271-284 (1988).

[16] Guerra, F. and Toninelli, F. L.: *The thermodynamic limit in mean field spin glass models.* Comm. Math. Phys. **230**, 71-79 (2002).

[17] Ledoux, Michel: *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs **89**, AMS.

[18] Mézard, M.: *The space of interactions in neural networks: Gardner's computation with the cavity method.* J. Phys. A: Math. Gen. **22**, 2181-2190 (1989)

[19] Morita,T.: *Statistical mechanics of quenched solid solutions with application to magnetically dilute alloys.* J. Math. Phys. **5**, 1401-1405 (1966).

[20] Mézard, M., and Montanari, M.: *Information, physics, and computation.* Oxford University Press 2009.

[21] Mézard, M., Parisi, G., and Virasoro, M.A.: *Spin glass theory and beyond.* World Scientific LN in Physics, Vol 9. World Scientific 1987.

[22] Mézard, M.: *Mean-field message-passing equations in the Hopfield model and its generalizations.* Phys. Rev. E **95,** 22117-22132 (2017).

[23] Panchenko, D.: *The Sherrington-Kirkpatrick model.* Springer, New York, 2013.

[24] Plefka, T.: *Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model.* J. Phys. A: Math. Gen. **15**, 1971–1978 (1982).

[25] Talagrand, M.: *Mean Field Models for Spin Glasses.* Vol I and Vol II. Springer 2011.

[26] Sherrington, D., and Kirkpatrick, S.: *Solvable model of a spin-glass.* Phys. Rev. Lett. **35**, 1792–1795 (1975).

[27] Thouless, D.J., Anderson, P.W., and Palmer, R.G.: *Solution of "solvable model in spin glasses"*. Philosophical Magazin **35,** 593-601 (1977).

[28] Zdeborova, L., and Krzakala, F.: *Statistical physics of inference: Thresholds and algorithms.* Advances in Physics **65** (2016).