Estimating Random Variables from Random Sparse Observations

Andrea Montanari[*]

September 2, 2007

Abstract

Let $X_1, \dots, X_n$ be a collection of iid discrete random variables, and $Y_1, \dots, Y_m$ a set of noisy observations of such variables. Assume each observation Y_a to be a random function of some a random subset of the X_i's, and consider the conditional distribution of X_i given the observations, namely $\mu_i(x_i) \equiv \mathbb{P}\{X_i = x_i|Y\}$ (*a posteriori probability*).

We establish a general decoupling principle among the X_i's, as well as a relation between the distribution of μ_i, and the fixed points of the associated density evolution operator. These results hold asymptotically in the large system limit, provided the average number of variables an observation depends on is bounded. We discuss the relevance of our result to a number of applications, ranging from sparse graph codes, to multi-user detection, to group testing.

1 Introduction

Sparse graph structures have proved useful in a number of information processing tasks, from channel coding [RU07], to source coding [CSV04], to sensing and signal processing [Don06, EJCT06]. Recently similar design ideas have been proposed for code division multiple access (CDMA) communications [MT06, YT06, RS07], and group testing (a classical technique in statistics) [MT07].

The computational problem underlying many of these developments can be described as follows: *infer the values of a large collection of random variables, given a set of constraints, or observations, that induce relations among them.* While such a task is generally computationally hard [BMvT78, Ver89]), sparse graphical structures allow for low-complexity algorithms (for instance iterative message passing algorithms as belief propagation) that were revealed to be very effective in practice. A precise analysis of these algorithms and of their gap to optimal (computationally intractable) inference is however a largely open problem.

In this paper we consider an idealized setting in which we aim at estimating n iid discrete random variables $X = (X_1, \dots, X_n)$ based on noisy observations. We will focus on the large system limit $n \to \infty$, with the number of observations scaling like n. We further restrict our system to be *sparse* in the sense that each observation depends on a bounded (on average) number of variables. A schematic representation is given in Fig. 1.

If $i \in [n]$, and Y denotes collectively the observations, a sufficient statistics for estimating X_i is

$$\mu_i(x_i) = \mathbb{P}\{X_i = x_i|Y\}\, . \tag{1.1}$$

This paper establishes two main results: an asymptotic decoupling among the X_i's, and a characterization of the asymptotic distribution of $\mu_i(\,\cdot\,)$ when Y is drawn according to the source and channel model. In the remainder of the introduction we will discuss a few (hopefully) motivating examples, and we will give an informal summary of our results. Formal definitions, statements and proofs can be found in Sections 2 to 6.

[*]Departments of Electrical Engineering and Statistics, Stanford University, montanari@stanford.edu

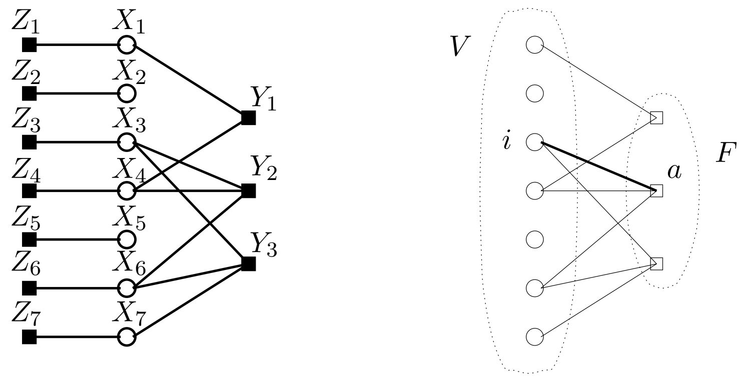

Figure 1: Factor graph representation of a simple sparse observation systems with $n = 7$ hidden variables $\{X_1, \dots, X_7\}$ and $m = 3$ 'multi-variable' observations $\{Y_1, \dots, Y_3\}$. On the right: the bipartite graph G. Highlighted is the edge (i, a).

1.1 Motivating examples

In this Section we present a few examples that fit within the mathematical framework developed in the present paper. The main restrictions imposed by this framework are: (i) The 'hidden variables' X_i's are independent; (ii) The bipartite graph G connecting hidden variables and observations lacks any geometrical structure.

Our results crucially rely on these two features. Some further technical assumptions will be made that partially rule out some of these examples below. However we expect these assumptions to be removable by generalizing the arguments presented in the next sections.

Source coding through sparse graphs. Let $(X_1, \dots, X_n)$ be iid Bernoulli(p). Shannon's theorem implies that such a vector can be stored in nR bits for any $R > h(p)$ (with $h(p)$ the binary entropy function), provided we allow for a vanishingly small failure probability. The authors of Refs. [Mur01, Mur04, CSV04] proposed to implement this compression through a sparse linear transformation. Given a source realization $X = x = (x_1, \dots, x_n)$, the stored vector reads

$$y = \mathbb{H}x \qquad \text{mod } 2\,,$$

with $\mathbb{H}$ a sparse $\{0, 1\}$ valued random matrix of dimensions $m \times n$, and $m = nR$. According to our general model, each of the coordinates of y is a function (mod 2 sum) of a bounded (on average) subset of the source buts $(x_1, \dots, x_n)$.

The i-th information bit can be reconstructed from the stored information by computing the conditional distribution $\mu_i(x_i) = \mathbb{P}\{X_i = x_i|Y\}$. In practice, belief propagation provides a rough estimate of μ_i. Determining the distribution of μ_i (which is the main topic of the present paper) allows to determine the optimal performances (in terms of bit error rate) of such a system.

Low-density generator matrix (LDGM) codes. We take $(X_1, \dots, X_n)$ iid Bernoulli(1/2), encode them in a longer vector $X' = (X'_1, \dots, X'_m)$ via the mapping $x' = \mathbb{H}x \mod 2$, and transmit the encoded bits through a noisy memoryless channel, thus getting output $(Y_1, \dots, Y_m)$ [Lub02]. One can for instance think of a binary symmetric channel BSC(p), whereby $Y_a = X'_a$ with probability $1 - p$ and $Y_a = X'_a \oplus 1$ with probability $1 - p$. Again, decoding can be realized through a belief propagation estimate of the conditional probabilities $\mu_i(x_i) = \mathbb{P}\{X_i = x_i|Y\}$.

If the matrix $\mathbb{H}$ is random and sparse, this problem fits in our framework with the information (uncoded) bits X_i's being hidden variables, while the y_a's correspond to observations.

Low-density parity-check (LDPC) codes. With LDPC codes, one sends through a noisy channel a codeword $X = (X_1, \dots, X_n)$ that is a uniformly random vector in the null space of a random sparse matrix $\mathbb{H}$ [Gal63, RU07]. While in general this does not fit our setting, one can construct an equivalent problem (for analysis purposes) which does, provided the communication channel is binary memoryless symmetric, say BSC(p).

Within the equivalent problem $(X_1, \dots, X_n)$ are iid Bernoulli(1/2) random bits. Given one realization $X = x$ of these bits, one computes its syndrome $y = \mathbb{H}x \mod 2$ and transmits it through a noiseless channel. Further, each of the X_i's is transmitted through the original noisy channel (in our example BSC(p)) yielding output Z_i. If we denote the observations collectively as (Y, Z), it is not hard to show that the conditional probability $\mu_i(x_i) = \mathbb{P}\{X_i = x_i|Y, Z\}$ has in fact the same distribution of the a posteriori probabilities in the original LDPC model.

Characterizing this distribution allows to determine the information capacity of such coding systems, and their performances under MAP decoding [Mon05, KM06, KKM07].

Compressed sensing. In compressed sensing the real vector $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ is measured through a set of linear projections $y_1 = h_1^T x, \dots, y_m = h_m^T x$. In this literature no assumption is made on the distribution of x, which is only constrained to be sparse in a properly chosen basis [Don06, EJCT06]. Further, unlike in our setting, the vector components x_i do not belong to any finite alphabet. However, some applications justify the study of a probabilistic version, whereby the basic variables are quantized. An example is provided by the next item.

Network measurements. The size of flows in the Internet can vary from a few (as in acknowledgment messages) to several million packets (as in content downloads). Keeping track of the sizes of flows passing through a router can be useful for a number of reasons, such as billing, or security, or traffic engineering [EV03].

Flow sizes can be modeled as iid random integers $X = (X_1, \dots, X_n)$. Their common distribution is often assumed to be a heavy-tail one. As a consequence, the largest flow is typically of size n^a for some $a > 0$. It is therefore highly inefficient to keep a separate counter of capacity n^a for each flow. It was proposed in [LMP07] to store instead a shorter vector $Y = \mathbb{H}X$, with $\mathbb{H}$ a properly designed sparse random matrix. The problem of reconstructing the X_i's is, once more, analogous to the above.

Group testing. Group testing was proposed during World War II as a technique for reducing the costs of syphilis tests in the army by simultaneously testing *groups* of soldiers [Dor43]. The variables $X = (X_1, \dots, X_n)$ represent the individuals status (1 = infected, 0 = healthy) and are modeled as iid Bernoulli(p) (for some small p). Test $a \in \{1, \dots, m\}$ involves a subset $\partial a \subseteq [n]$ of the individuals and returns positive value $Y_a = 1$ if $X_i = 1$ for some $i \in \partial a$ and $Y_a = 0$ otherwise[1]. It is interesting to mention that the problem has a connection with random multi-access channels, that was exploited in [BMTW84, Wol85].

One is interested in the conditional probability for the i-th individual to be infected given the observations: $\mu_i(1) = \mathbb{P}\{X_i = 1|Y\}$. Choices of the groups (i.e. of the subsets ∂a) based on random graph structures where recently studied and optimized in [MT07].

Multi-user detection. In a general vector channel, one or more users communicate symbols $X = (X_1, \dots, X_n)$ (we assume, for the sake of simplicity, perfect synchronization). The receiver is given a channel output $Y = (Y_1, \dots, Y_m)$, that is usually modeled as a linear function of the input, plus gaussian noise $Y = \mathbb{H}X + W$, where $W = (W_1, \dots, W_m)$ are normal iid random variables. Examples are CDMA or multiple-input multiple-output channels (with perfect channel state information) [Ver98, TV05].

The analysis simplifies considerably if the X_i's are assumed to be normal as well [TH99, VS99]. However, in many circumstances a binary or quaternary modulation is used, and the normal assumption is therefore unrealistic. The non-rigorous 'replica method' from statistical physics have been used to compute the channel capacity in these cases [Tan02]. A proof of replicas predictions have been obtained in [MT06] in under some condition on the spreading factor. The same techniques were applied in more general settings in [GV05, GW06b, GW06a].

However, specific assumptions on the spreading factor (and noise parameters) were necessary. Such assumptions ensured an appropriate density evolution operator to have unique fixed point. The results

[1]The problem can be enriched by allowing for a small false negative (or false positive) probability.

of the present paper should allow to prove replica results without conditions on the spreading.

As mentioned above we shall make a few technical assumptions on the structure of the sparse observation system. These will concern the distribution of the bipartite graph connecting hidden variables and observations, as well as the dependency of the noisy observations on the X's. While such assumptions rule out some of the example above (for instance, they exclude general irregular LDPC ensembles), we do not think they are crucial for the results to hold.

1.2 An informal overview

We consider two types of observations: single variable observations $Z = (Z_1, \dots, Z_n)$, and multi-variable observations $Y = (Y_1, \dots, Y_m)$. For each $i \in [n]$, Z_i is the result of observing X_i through a memoryless noisy channel. Further for each a, Y_a is an independent noisy function of a subset $\{X_j : j \in \partial a\}$ of the hidden variables. By this we mean that Y_a is conditionally independent from all the other variables, given $\{X_j : j \in \partial a\}$. The subset $\partial a \subseteq [n]$ is itself random with, for each $i \in [n]$, $i \in \partial a$ independently with probability γ/n.

Generalizing the above, we consider the conditional distribution of X_i, given Y *and* Z:

$$\mu_i(x_i) \equiv \mathbb{P}\{X_i = x_i | Y, Z\}\,. \tag{1.2}$$

One may wonder whether additional information can be extracted by considering the correlation among hidden variables. Our first result is that for a generic subset of the variables, these correlation vanish. This is stated informally below:

For any uniformly random set of variable indices $i(1), \dots, i(k) \in [n]$ and any $\xi_1, \dots, \xi_k \in \mathcal{X}$

$$\mathbb{P}\{X_{i(1)} = \xi_1, \dots, X_{i(k)} = \xi_k | Y, Z\} \approx \mathbb{P}\{X_{i(1)} = \xi_1 | Y, Z\} \cdots \mathbb{P}\{X_{i(k)} = \xi_k | Y, Z\}\,. \tag{1.3}$$

This can be regarded as a generalization of the 'decoupling principle' postulated in [GV05]. Here the $\approx$ symbols hides the large system ($n, m \to \infty$) limit, and a 'smoothing procedure' to be discussed below.

Locally, the graph G converges to a random bipartite tree. The locally tree-like structure of G suggests the use of message passing algorithms, in particular belief propagation, for estimating the marginals μ_i. Consider the subgraph including i as well as all the function nodes a such that Y_a depends on i, and the other variables these observations depend on. Refer to the latter as to the 'neighbors of i.' In belief propagation one assumes these to be independent *in absence of i, and of its neighborhood*.

For any j, neighbor of i, let $\mu_{j\to i}$ denote the conditional distribution of X_j in the modified graph where i (and the neighboring observations) have been taken out. Then BP provides a prescription[2] for computing μ_i in terms of the 'messages' $\mu_{j\to i}$, of the form $\mu_i = \mathsf{F}_i^n(\{\mu_{j\to i}\})$. We shall prove that this prescription is asymptotically correct.

Let i be a uniformly random variable node and $i(1), \dots, i(k)$ its neighbors. Then

$$\mu_i \approx \mathsf{F}_i^n(\mu_{i(1)\to i}, \dots, \mu_{i(k)\to i})\,. \tag{1.4}$$

The neighborhood of i converges to a Galton-Watson tree, with Poisson distributed degrees of mean $\gamma\alpha$ (for variable nodes, corresponding to variables X_i's) and γ (for function nodes corresponding to observation Y_a's). Such a tree is generated as follows. Start from a root variable node, generate a Poisson($\gamma\alpha$) number of function node descendants, and for each of them an independent Poisson(γ) number of variable node descendants. This procedure is then repeated recursively.

In such a situation, consider again the BP equation (1.4). Then the function $\mathsf{F}_i^n(\cdots)$ can be approximated by a random function corresponding to a random Galton-Watson neighborhood, do be denoted as F^∞. Further, one can hope that, if the graph G is random, then the $\mu_{i(j)\to i}$ become iid random

[2]The mapping $\mathsf{F}_i^n(\,\cdot\,)$ returns the marginal at i with respect to the subgraph induced by i and its neighbors, when the latter are biased according to $\mu_{j\to i}$. For a more detailed description, we refer to Section 2.2.

variables. Finally (and this is a specific property of Poisson degree distributions) the residual graph with the neighborhood of i taken out, has the same distribution (with slightly modified parameters) as the original one. Therefore, one might imagine that the distribution of μ_i is the same as the one of the $\mu_{i(j)\to i}$'s. Summarizing these observations one is lead to think that the distribution of μ_i must be (asymptotically for large systems) a fixed point of the following distributional equation

$$\nu \stackrel{\mathrm{d}}{=} \mathsf{F}^{\infty}(\nu_1, \dots, \nu_l)\,. \tag{1.5}$$

This is an equation for the distribution of ν (the latter taking values in the set of distributions over the hidden variables X_i) and is read as follows. When $\nu_1, \dots, \nu_l$ are random variables with common distribution ρ, then $\mathsf{F}^{\infty}(\nu_1, \dots, \nu_l)$ has itself distribution ρ (here l and F^{∞} are also random according to the Galton-Watson model for the neighborhood of i). It is nothing but the fixed point equation for density evolution, and can be written more explicitly as

$$\rho(\nu \in A) = \int \mathbb{I}\big(\mathsf{F}^{\infty}(\nu_1, \dots, \nu_l) \in A\big)\, \rho(\mathrm{d}\nu_1) \cdots \rho(\mathrm{d}\nu_l)\,, \tag{1.6}$$

where $\mathbb{I}(\,\cdots)$ is the indicator function. In fact our main result tells that: (i) The distribution of μ_i must be a convex combination of the solutions of the above distributional equation; (ii) If such convex combination is nontrivial (has positive weight on more than one solution) then the correlations among the μ_i's have a peculiar structure.

Assume density evolution to admit the fixed point distributions $\rho_1, \dots, \rho_r$ for some fixed r. Then there exists probabilities $w_1, \dots, w_r$ (which add up to 1) such that, for $i(1) \dots i(k) \in [n]$ uniformly random variable nodes,

$$\mathbb{P}\{\mu_{i(1)} \in A_1, \dots \mu_{i(k)} \in A_k\} \approx \sum_{\alpha=1}^{r} w_\alpha\, \rho_\alpha(\mu \in A_1) \cdots \rho_\alpha(\mu \in A_k)\,. \tag{1.7}$$

In the last statement we have hidden one more technicality: the stated asymptotic behavior might hold only along a subsequence of system sizes. In fact in many cases it can be proved that the above convex combination is trivial, and that no subsequence needs to be taken. Tools for proving this will be developed in a forthcoming publication.

2 Definitions and main results

In this section we provide formal definitions and statements.

2.1 Sparse systems of observations

We consider systems defined on a bipartite graph $G = (V, F, E)$, whereby V and F are vertices corresponding (respectively) to variables and observations ('variable' and 'function nodes'). The edge set is $E \subseteq V \times F$. For greater clarity, we shall use $i, j, k, \dots \in V$ to denote variable nodes and $a, b, c, \dots \in F$ for function nodes. For $i \in V$, we let $\partial i \equiv \{a \in F :\ (i, a) \in E\}$ denote its neighborhood (and define analogously ∂a for $a \in F$). Further, if we let $n \equiv |V|$, and $m \equiv |F|$, we are interested in the limit $n, m \to \infty$ with $\alpha = m/n$ kept fixed (often we will identify $V = [n]$ and $F = [m]$).

A family of iid random variables $\{X_i : i \in V\}$, taking values in a finite alphabet $\mathcal{X}$, is associated with the vertices of V. The common distribution of the X_i will be denoted by $\mathbb{P}\{X_i = x\} = p(x)$. Given $U \subseteq V$, we let $X_U \equiv \{X_i : i \in U\}$ (the analogous convention is adopted for other families of variables). Often we shall write X for X_V.

Random variables $\{Y_a : a \in F\}$ are associated with the function nodes, with Y_a conditionally independent of $Y_{F\backslash a}$, $X_{V\backslash \partial a}$, given $X_{\partial a}$. Their joint distribution is defined by a set of probability kernels $Q^{(k)}$ indexed by $k \in \mathbb{N}$, whereby, for $|\partial a| = k$,

$$\mathbb{P}\{Y_a \in \cdot\, | X_{\partial a} = x_{\partial a}\} = Q^{(k)}(\,\cdot\, | x_{\partial a})\,. \tag{2.1}$$

We shall assume $Q^{(k)}(\,\cdot\,|x_1, \dots, x_k)$ to be invariant under permutation of its arguments $x_1, \dots, x_k$ (an assumption that is implicit in the above equation). Further, whenever clear from the context, we shall drop the superscript (k). Without loss of generality, one can assume Y_a to take values in $\mathbb{R}^b$ for some b which only depends on k.

A second collection of real random variables $\{Z_i :\ i \in V\}$ is associated with the variable nodes, with Z_i conditionally independent of $Z_{V\setminus i}$, $X_{V\setminus i}$ and Y, conditional on X_i. The associated probability kernel will be denoted by R:

$$\mathbb{P}\{Z_i \in \,\cdot\,|X_i = x_i\} = R(\,\cdot\,|x_i)\,. \tag{2.2}$$

Finally, the graph G itself will be random. All the above distributions have to be interpreted as conditional to a given realization of G. We shall follow the convention of using $\mathbb{P}\{\cdots\}$, $\mathbb{E}\{\cdots\}$ etc, for conditional probability, expectation, etc. given G (without writing explicitly the conditioning) and write $\mathbb{P}_G\{\cdots\}$, $\mathbb{E}_G\{\cdots\}$ for probability and expectation with respect to G. The graph distribution is defined as follows. Both node sets V and F are given. Further, for any $(i,a) \in V \times F$, we let $(i,a) \in E$ independently with probability $p_{\rm edge}$. If we let $n \equiv |V|$, and $m \equiv |F|$ (often identifying $V = [n]$ and $F = [m]$), such a random graph ensemble will be denoted as $\mathcal{G}(n, m, p_{\rm edge})$. We are interested in the limit $n, m \to \infty$ with $\alpha = m/n$ kept fixed and $p_{\rm edge} = \gamma/n$.

In particular, we will be concerned with the problem of determining the conditional distribution of X_i given Y and Z cf. Eq. (1.2). Notice that μ_i is a random variable taking values in $\mathsf{M}(\mathcal{X})$ (the set of probability measures over $\mathcal{X}$).

In order to establish our main result we need to 'perturb' the system as follows. Given a perturbation parameter $\theta \in [0,1]$ (that should be thought as 'small'), and a symbol $* \notin \mathcal{X}$, we let

$$Z_i(\theta) = \begin{cases} (Z_i, X_i) & \text{with probability } \theta, \\ (Z_i, *) & \text{with probability } 1-\theta. \end{cases} \tag{2.3}$$

In words, we reveal a random subset of the hidden variables. Obviously $Z(0)$ is equivalent to Z and $Z(1)$ to X. The corresponding probability kernel is defined by (for $A \subseteq \mathbb{R}$ measurable, and $\overline{x} \in \mathcal{X} \cup \{*\}$)

$$R^\theta(\overline{x}, A|x_i) = [(1-\theta)\mathbb{I}(\overline{x} = *) + \theta\, \mathbb{I}(\overline{x} = x_i)]R(A|x_i)\,, \tag{2.4}$$

where $\mathbb{I}(\,\cdots)$ is the indicator function. We will denote by μ_i^θ the analogous of μ_i, cf. Eq. (1.2), with Z being replaced by $Z(\theta)$.

It turns out that introducing such a perturbation is necessary for our result to hold. The reason is that there can be specific choices of the system 'parameters' α, γ, and of the kernels Q and R for which the variables X_i's are strongly correlated. This happens for instance at thresholds noise levels in coding. Introducing a perturbation allows to remove this non-generic behaviors.

We finally need to introduce a technical regularity condition on the laws of Y_a and Z_a (notice that this concerns the *unperturbed model*).

Definition 2.1. *We say that a probability kernel T from $\mathcal{X}$ to a measurable space $\mathcal{S}$ (i.e., a set of probability measures $T(\,\cdot\,|x)$ indexed by $x \in \mathcal{X}$) is* soft *if: (i) $T(\,\cdot\,|x_1)$ is absolutely continuous with respect to $T(\,\cdot\,|x_2)$ for any $x_1, x_2 \in \mathcal{X}$; (ii) We have, for some $M < \infty$, and all $x \in \mathcal{X}$ (the derivative being in the Radon-Nikodyn sense)*

$$\int \frac{\mathrm{d}T(y|x_1)}{\mathrm{d}T(y|x_2)}\, T(\mathrm{d}y|x) \le M\,. \tag{2.5}$$

A system of observations is said to have soft noise *(or* soft noisy observations*), if there exists $M < \infty$ such that the kernels R and and $Q^{(k)}$ for all $k \ge 1$ are M-soft.*

In the case of a finite output alphabet the above definition simplifies considerably: a kernel is soft if all its entries are non-vanishing. Although there exist interesting examples of non-soft kernels (see, for instance, Section 1.1) they can often be treated as limit cases of soft ones.

2.2 Belief propagation and density evolution

Belief propagation (BP) is frequently used in practice to estimate the marginals (1.2). Messages $\nu^{(t)}_{i\to a}$, $\widehat{\nu}^{(t)}_{a\to i}\in \mathsf{M}(\mathcal{X})$ are exchanged at time t along edge $(i,a)\in E$, where $i\in V$, $a\in F$. The update rules follow straightforwardly from the general factor graph formalism [KFL01]

$$\nu^{(t+1)}_{i\to a}(x_i) \;\propto\; p(x_i)R^\theta(z_i|x_i)\prod_{b\in\partial i\setminus a}\widehat{\nu}^{(t)}_{b\to i}(x_i)\,, \tag{2.6}$$

$$\widehat{\nu}^{(t)}_{a\to i}(x_i) \;\propto\; \sum_{x_{\partial a\setminus i}} Q(y_a|x_{\partial a})\prod_{j\in\partial a\setminus i}\nu^{(t)}_{j\to a}(x_j)\,. \tag{2.7}$$

Here and below we denote by $\propto$ equality among measures on the same space 'up to a normalization[3].' The BP estimate for the marginal of variable X_i is (after t iterations)

$$\nu^{(t+1)}_{i}(x_i) \;\propto\; p(x_i)R^\theta(z_i|x_i)\prod_{b\in\partial i}\widehat{\nu}^{(t)}_{b\to i}(x_i)\,. \tag{2.8}$$

Combining Eqs. (2.6) and (2.8), the BP marginal at variable node i can be expressed as a function of variable-to-function node messages at neighboring variable nodes. We shall write

$$\nu^{(t+1)}_{i} \;=\; \mathsf{F}^n_i(\{\nu^{(t)}_{j\to b}:\, j\in\partial b\setminus i;\, b\in\partial i\})\,, \tag{2.9}$$

$$\mathsf{F}^n_i(\cdots)(x_i) \;\propto\; p(x_i)R^\theta(z_i|x_i)\prod_{a\in\partial i}\left\{\sum_{x_{\partial a\setminus i}} Q(y_a|x_{\partial a})\prod_{j\in\partial a\setminus i}\nu^{(t)}_{j\to a}(x_j)\right\}\,. \tag{2.10}$$

Notice that the mapping $\mathsf{F}^n_i(\cdots)$ only depends on the graph G and on the observations $Y, Z(\theta)$, through the subgraph including function nodes adjacent to i and the corresponding variable nodes. Denoting such neighborhood as B, the corresponding observations as Y_{B}, $Z_{\mathsf{B}}(\theta)$, and letting $\nu^{(t)}_{\mathsf{D}} = \{\nu^{(t)}_{j\to b}:\, j\in\partial b\setminus i;\, b\in\partial i\}$, we can rewrite Eq. (2.9) in the form

$$\nu^{(t+1)}_{i} \;=\; \mathsf{F}^n(\nu^{(t)}_{\mathsf{D}};\mathsf{B}, Y_{\mathsf{B}}, Z_{\mathsf{B}}(\theta))\,. \tag{2.11}$$

Here we made explicit all the dependence upon the graph and the observations. If G is drawn randomly from the $\mathcal{G}(n,\alpha n,\gamma/n)$ ensemble, the neighborhood B, as well as the corresponding observations converge in the large system limit, to a well defined limit distribution. Further, the messages $\{\nu^{(t)}_{j\to b}\}$ above become iid and are distributed as $\mu^{(t)}_i$ (this is a consequence of the fact that the edge degrees are asymptotically Poisson). Their common distribution satisfies the density evolution distributional recursion

$$\nu^{(t+1)} \;\stackrel{\rm d}{=}\; \mathsf{F}^\infty(\nu^{(t)}_{\mathsf{D}};\mathsf{B}, Y_{\mathsf{B}}, Z_{\mathsf{B}}(\theta))\,, \tag{2.12}$$

where $\nu^{(t+1)}_{\mathsf{D}} = \{\nu^{(t)}_e:\, e\in\mathsf{D}\}$ are iid copies of $\nu^{(t)}$, and $\mathsf{B}, Y_{\mathsf{B}}, Z_{\mathsf{B}}(\theta)$ are understood to be taken from their asymptotic distribution. We will be particularly concerned with the set of *fixed points* of the above distributional recursion. This is just the set of distributions ρ over $\mathsf{M}(\mathcal{X})$ such that, if $\nu^{(t)}$ has distribution ρ, then $\nu^{(t+1)}$ has distribution ρ as well.

2.3 Main results

For stating our first result, it is convenient to introduce a shorthand notation. For any $U\subseteq V$, we note

$$\widetilde{\mathbb{P}}_U\{x_U\} \equiv \mathbb{P}\{X_U = x_U|Y, Z(\theta)\}\,. \tag{2.13}$$

Notice that, being a function of Y and $Z(\theta)$, $\widetilde{\mathbb{P}}_U\{x_U\}$ is a random variable. The theorem below shows that, if U is a random subset of V of bounded size, then $\widetilde{\mathbb{P}}_U$ factorizes approximately over the nodes

[3]Explicitly, $q_1(x)\propto q_2(x)$ if there exists a constant $C>0$ such that $q_1(x)=q_2(x)$ for all x

$i \in U$. The accuracy of this is measured in terms of *total variation distance*. Recall that, given two distributions q_1 and q_2 on the same finite set $\mathcal{S}$, their total variation distance is

$$||q_1 - q_2||_{\rm TV} = \frac{1}{2}\sum_{x\in\mathcal{S}} |q_1(x) - q_2(x)|\,. \tag{2.14}$$

Theorem 2.2. *Consider an observation system on the graph $G = (V, F, E)$. Let $k \in \mathbb{N}$, and $i(1), \dots, i(k)$ be uniformly random in V. Then, for any $\epsilon > 0$*

$$\int_0^\epsilon \mathbb{E}_{i(1)\cdots i(k)}\mathbb{E}\Big|\Big|\widetilde{\mathbb{P}}_{i(1),\dots,i(k)} - \widetilde{\mathbb{P}}_{i(1)}\cdots\widetilde{\mathbb{P}}_{i(k)}\Big|\Big|_{\rm TV}\,{\rm d}\theta \le (|\mathcal{X}|+1)^k A_{n,k}\sqrt{H(X_1)\epsilon/n} = O(n^{-1/2})\,, \tag{2.15}$$

where $A_{n,k} \le \exp\left(\frac{k^2}{2n}\right)$ for $k < n/2$, and the asymptotic behavior $O(n^{-1/2})$ holds as $n \to \infty$ with if k and $\mathcal{X}$ fixed.

The next result establishes that the BP equation (2.9) is approximately satisfied by the actual marginals. For any $i,\, j \in V$, such that $i, j \in \partial b$ for some common function node $b \in F$, let

$$\mu_i^{\theta(j)}(x_i) \equiv \mathbb{P}\Big\{X_i = x_i\Big|Y_a : j \notin \partial a;\; Z_l(\theta) : l \neq j\Big\}\,. \tag{2.16}$$

This is nothing but the conditional distribution of X_i with respect to the graph from which j has been 'taken out.'

Theorem 2.3. *Consider a sparse observation system on a random graph $G = (V, F, E)$ from the $\mathcal{G}_n(\gamma/n, \alpha n)$ ensemble. Assume the noisy observations to be M-soft. Then there exists a constant A depending on $t, \alpha, \gamma, M, |\mathcal{X}|, \epsilon$, such that for any $i \in V$, and any n*

$$\int_0^\epsilon \mathbb{E}_G\mathbb{E}||\mu_i^\theta - \mathsf{F}_i^n(\{\mu_j^{\theta,(i)}\}_{a\in\partial i, j\in\partial a\setminus i})||_{\rm TV}\,{\rm d}\theta \le \frac{\mathsf{A}}{\sqrt{n}}\,. \tag{2.17}$$

Finally, we provide a characterization of the asymptotic distribution of the one variable marginals. Recall that $\mathsf{M}(\mathcal{X})$ denotes the set of probability distributions over $\mathcal{X}$, i.e., the $(|\mathcal{X}|-1)$-dimensional standard simplex. We further let $\mathsf{M}^2(\mathcal{X})$ be the set of probability measures over $\mathsf{M}(\mathcal{X})$ ($\mathsf{M}(\mathcal{X})$ being endowed with the Borel σ-field induced by $\mathbb{R}^{|\mathcal{X}|-1}$). This can be equipped with the smallest σ-field that makes $F_A : \rho \mapsto \rho(A)$ measurable for any Borel subset A of $\mathsf{M}(\mathcal{X})$.

Theorem 2.4. *Consider an observation system on a random graph $G = (V, F, E)$ from the $\mathcal{G}(n, \alpha n, \gamma/n)$ ensemble, and assume the noisy observations to be soft. Let $\varphi : \mathsf{M}(\mathcal{X})^k \to \mathbb{R}$ be a Lipschitz continuous function on $\mathsf{M}(\mathcal{X})^k = \mathsf{M}(\mathcal{X}) \times \cdots \times \mathsf{M}(\mathcal{X})$ (k times).*

Then for almost any $\theta \in [0, \varepsilon]$ there exists an infinite subsequence $R_\theta \subseteq \mathbb{N}$ and a probability distribution S_θ over $\mathsf{M}^2(\mathcal{X})$, supported on the fixed points of the density evolution recursion (2.12), such that the following happens. Given any fixed subset of variable nodes $\{i(1), \dots, i(k)\} \subseteq V$

$$\lim_{n\in R_\theta} \mathbb{E}_G\mathbb{E}\left\{\varphi(\mu_{i(1)}^\theta, \dots, \mu_{i(k)}^\theta)\right\} = \int \left\{\int \varphi(\mu_1, \dots, \mu_k)\,\rho({\rm d}\mu_1)\cdots\rho({\rm d}\mu_k)\right\} S({\rm d}\rho)\,. \tag{2.18}$$

3 Proof of Theorem 2.2 (correlations)

Lemma 3.1. *For any observation system and any $\epsilon > 0$*

$$\frac{1}{n}\sum_{i,j\in V}\int_0^\epsilon I(X_i; X_j|Y, Z(\theta))\,{\rm d}\theta \le 2H(X_1)\,. \tag{3.1}$$

Proof. For $U \subseteq V$, let us denote by $Z^{(U)}(\theta)$ the vector obtained by setting $Z_i^{(U)}(\theta) = Z_i(\theta)$ whenever $i \notin U$, and $Z_i^{(U)}(\theta) = (Z_i, *)$ if $i \in U$. The proof is based on the two identities below

$$\frac{{\rm d}}{{\rm d}\theta}H(X|Y, Z(\theta)) = -\sum_{i\in V} H(X_i|Y, Z^{(i)}(\theta))\,, \tag{3.2}$$

$$\frac{{\rm d}^2}{{\rm d}\theta^2}H(X|Y, Z(\theta)) = \sum_{i\neq j\in V} I(X_i; X_j|Y, Z^{(ij)}(\theta))\,. \tag{3.3}$$

Before proving these identities, let us show that they imply the thesis. By the fundamental theorem of calculus, we have

$$\frac{1}{n}\sum_{i\neq j\in V}\int_0^{\epsilon} I(X_i;X_j|Y,Z^{(ij)}(\theta))\,\mathrm{d}\theta \;=\; \frac{1}{n}\sum_{i\in V}H(X_i|Y,Z^{(i)}(0))-\frac{1}{n}\sum_{i\in V}H(X_i|Y,Z^{(i)}(\epsilon)) \tag{3.4}$$

$$\leq\; \frac{1}{n}\sum_{i\in V}H(X_i|Y,Z^{(i)}(0))\leq H(X_1)\,. \tag{3.5}$$

Further, if $z^{(U)}(\theta)$ is the vector obtained from $z(\theta)$ by replacing $z_i(\theta)$ with $(z_i,*)$ for any $i\in U$, then

$$I(X_i;X_j|Y,Z(\theta)=z(\theta))\leq I(X_i;X_j|Y,Z^{(ij)}(\theta)=z^{(ij)}(\theta))\,. \tag{3.6}$$

In fact the left hand side vanishes whenever $z^{(ij)}(\theta)\neq z(\theta)$. The proof is completed by upper bounding the diagonal terms in the sum (3.1) as $I(X_i;X_i|Y,Z^{(i)}(\theta))=H(X_i|Y,Z^{(i)}(\theta))\leq H(X_1)$.

Let us now consider the identities (3.2) and (3.3). These already appeared in the literature [MMU05, MMRU05, Mac07]. We reproduce the proof here for the sake of self-containedness.

Let us begin with Eq. (3.2). It is convenient to slightly generalize the model by letting the parameter the channel parameter θ be dependent on the variable node. In other words given a vector $\underline{\theta}=(\theta_1,\dots,\theta_n)$, we let, for each $i\in V$, $Z_i(\underline{\theta})=(Z_i,X_i)$ with probability θ_i, and $=(Z_i,*)$ otherwise. Noticing that $H(X|Y,Z(\underline{\theta}))=H(X_i|Y,Z(\underline{\theta}))+H(X|X_i,Y,Z(\underline{\theta}))$ and that the latter term does not depend upon θ_i, we have

$$\frac{\partial}{\partial\theta_i}H(X|Y,Z(\underline{\theta}))=\frac{\partial}{\partial\theta_i}H(X_i|Y,Z(\underline{\theta}))=-H(X_i|Y,Z^{(i)}(\underline{\theta}))\,, \tag{3.7}$$

where the second equality is a consequence of $H(X_i|Y,Z(\underline{\theta}))=(1-\theta_i)H(X_i|Y,Z^{(i)}(\underline{\theta}))$. Equation (3.2) follows by simple calculus taking $\theta_i=\theta_i(\theta)=\theta$ for all $i\in V$.

Equation (3.3) is proved analogously. First, the above calculation implies that the second derivative with respect to θ_i vanishes for any $i\in V$. For $i\neq j$, we use the chain rule to get $H(X|Y,Z(\underline{\theta}))=H(X_i,X_j|Y,Z(\underline{\theta}))+H(X|X_i,X_j,Y,Z^{(ij)}(\underline{\theta}))$, and then write

$$\begin{aligned}H(X_i,X_j|Y,Z(\underline{\theta}))=(1-\theta_i)(1-\theta_j)H(X_i,X_j|Y,Z^{(ij)}(\underline{\theta}))+\theta_i(1-\theta_j)H(X_j|X_i,Y,Z^{(ij)}(\underline{\theta}))+\\+(1-\theta_i)\theta_jH(X_i|X_j,Y,Z^{(ij)}(\underline{\theta}))\,,\end{aligned}$$

whence the mixed derivative with respect to θ_i and θ_j results in $I(X_i;X_j|Y,Z^{(ij)}(\underline{\theta}))$. As above, Eq. (3.3) is recovered by letting $\theta_i=\theta_i(\theta)=\theta$ for any $i\in V$. □

In the next proof we will use a technical device that has been developed within the mathematical theory of spin glasses (see [Tal06], and [GT04, GM07] for applications to sparse models). We start by defining a family of real random variables indexed by a variable node $i\in V$, and by $\xi\in\mathcal{X}$:

$$\mathsf{S}_i(\xi)\equiv\mathbb{I}(X_i=\xi)-\mathbb{P}\{X_i=\xi|Y,Z(\theta)\}\,. \tag{3.8}$$

We will also use $\mathsf{S}(\xi)=(\mathsf{S}_1(\xi)\ \dots,\mathsf{S}_n(\xi))$ to denote the corresponding vector.

Next we let $X^{(1)}=(X^{(1)}_1,\dots,X^{(1)}_n)$ and $X^{(2)}=(X^{(2)}_1,\dots,X^{(2)}_n)$ be two iid assignments of the hidden variables, both distributed according to the conditional law $\mathbb{P}_{X|Y,Z(\theta)}$. If we let $(Y,Z(\theta))$ be distributed according to the original (unconditional) law $\mathbb{P}_{Y,Z(\theta)}$, this defines a larger probability space, generated by $(X^{(1)},X^{(2)},Y,Z)$. Notice that the pair $(X^{(1)},Y,Z)$ and $(X^{(2)},Y,Z)$ is exchangeable, each of the terms being distributed as $(X,Y,Z(\theta))$.

In terms of $X^{(1)}$ and $X^{(2)}$ we can then define $\mathsf{S}^{(1)}(\xi)$ and $\mathsf{S}^{(2)}(\xi)$, and introduce the *overlap*

$$\mathsf{Q}(\xi)\equiv\frac{1}{n}\,\mathsf{S}^{(1)}(\xi)\cdot\mathsf{S}^{(1)}(\xi)=\frac{1}{n}\sum_{i\in V}\mathsf{S}^{(1)}_i(\xi)\,\mathsf{S}^{(2)}_i(\xi)\,. \tag{3.9}$$

Since $|\mathsf{S}_i(\xi)|\leq 1$, we have $|\mathsf{Q}(\xi)|\leq 1$ as well. Our next result shows that the conditional distribution of $\mathsf{Q}(\xi)$ given Y and $Z(\theta)$ is indeed very concentrated, for most valued of θ. The result is expressed in terms of the conditional variance

$$\mathrm{Var}(\mathsf{Q}(\xi)|Y,Z(\theta))\equiv\mathbb{E}\left\{\mathbb{E}[\mathsf{Q}(\xi)^2|Y,Z(\theta)]-\mathbb{E}[\mathsf{Q}(\xi)|Y,Z(\theta)]^2\right\}\,. \tag{3.10}$$

Lemma 3.2. *For any observations system and any $\epsilon > 0$*

$$\int_0^{\epsilon} \mathrm{Var}(\mathsf{Q}(\xi)|Y, Z(\theta))\, \mathrm{d}\theta \leq 4H(X_1)/n\,. \tag{3.11}$$

Proof. In order to lighten the notation, write $\widetilde{\mathbb{E}}\{\cdots\}$ for $\mathbb{E}\{\,\cdot\,|Y=y, Z(\theta)=z(\theta)\}$ (and analogously for $\widetilde{\mathbb{P}}\{\cdots\}$), and drop the argument ξ from $\mathsf{S}_i^{(a)}(\xi)$. Then

$$\begin{aligned}
\mathrm{Var}(\mathsf{Q}(\xi)|Y=y, Z(\theta)=z(\theta)) &= \widetilde{\mathbb{E}}\left\{\left(\frac{1}{n}\sum_{i\in V}\mathsf{S}_i^{(1)}\mathsf{S}_i^{(2)}\right)^2\right\} - \widetilde{\mathbb{E}}\left\{\frac{1}{n}\sum_{i\in V}\mathsf{S}_i^{(1)}\mathsf{S}_i^{(2)}\right\}^2 = \\
&= \frac{1}{n^2}\sum_{i,j\in V}\left\{\widetilde{\mathbb{E}}\left\{\mathsf{S}_i^{(1)}\mathsf{S}_i^{(2)}\mathsf{S}_j^{(1)}\mathsf{S}_j^{(2)}\right\} - \widetilde{\mathbb{E}}\left\{\mathsf{S}_i^{(1)}\mathsf{S}_i^{(2)}\right\}\widetilde{\mathbb{E}}\left\{\mathsf{S}_j^{(1)}\mathsf{S}_j^{(2)}\right\}\right\} = \\
&= \frac{1}{n^2}\sum_{i,j\in V}\left\{\widetilde{\mathbb{E}}\left\{\mathsf{S}_i\mathsf{S}_j\right\}^2 - \widetilde{\mathbb{E}}\left\{\mathsf{S}_i\right\}^2\widetilde{\mathbb{E}}\left\{\mathsf{S}_j\right\}^2\right\}.
\end{aligned}$$

In the last step we used the fact that $\mathsf{S}^{(1)}(\xi)$ and $\mathsf{S}^{(2)}(\xi)$ are conditionally independent given Y and $Z(\theta)$, and used the notation $\mathsf{S}_i(\xi)$ for any of them (recall that $\mathsf{S}^{(1)}(\xi)$ and $\mathsf{S}^{(2)}(\xi)$ are identically distributed). Notice that

$$\widetilde{\mathbb{E}}\{\mathsf{S}_i(\xi)\} = \mathbb{E}\Big\{\mathbb{I}(X_i=\xi) - \mathbb{P}\{X_i=\xi|Y, Z(\theta)\}\Big|Y=y, Z(\theta)=z(\theta)\Big\} = 0\,, \tag{3.12}$$

$$\begin{aligned}
\widetilde{\mathbb{E}}\{\mathsf{S}_i(\xi)\mathsf{S}_j(\xi)\} &= \widetilde{\mathbb{E}}\Big\{\left[\mathbb{I}(X_i=\xi) - \mathbb{P}\{X_i=\xi|Y, Z(\theta)\}\right]\left[\mathbb{I}(X_j=\xi) - \mathbb{P}\{X_j=\xi|Y, Z(\theta)\right]\Big\} = \\
&= \widetilde{\mathbb{P}}\{X_i=\xi,\, X_j=\xi\} - \widetilde{\mathbb{P}}\{X_i=\xi\}\widetilde{\mathbb{P}}\{X_j=\xi\}\,.
\end{aligned} \tag{3.13}$$

Therefore

$$\begin{aligned}
\mathrm{Var}(\mathsf{Q}(\xi)|Y=y, Z(\theta)=z(\theta)) &= \frac{1}{n^2}\sum_{i,j\in V}\Big(\widetilde{\mathbb{P}}\{X_i=\xi,\, X_j=\xi\} - \widetilde{\mathbb{P}}\{X_i=\xi\}\widetilde{\mathbb{P}}\{X_j=\xi\}\Big)^2 \leq \\
&\leq \frac{1}{n^2}\sum_{i,j\in V}\sum_{x_1,x_2}\Big(\widetilde{\mathbb{P}}\{X_i=x_1,\, X_j=x_2\} - \widetilde{\mathbb{P}}\{X_i=x_1\}\widetilde{\mathbb{P}}\{X_j=x_2\}\Big)^2 \leq \\
&\leq \frac{2}{n^2}\sum_{i,j\in V} I(X_i;X_j|Y=y, Z(\theta)=z(\theta)))\,.
\end{aligned}$$

In the last step we used the inequality (valid for any two distributions p_1, p_2 over a finite set $\mathcal{S}$)

$$\sum_x \big|p_1(x) - p_2(x)\big|^2 \leq 2D(p_1||p_2)\,, \tag{3.14}$$

and applied it to the joint distribution of X_1 and X_2, and the product of their marginals. The thesis follows by integrating over y and $z(\theta)$ with the measure $\mathbb{P}_{Y,Z(\theta)}$ and using Lemma 3.1. □

Proof (Theorem 2.2). We start by noticing that, since $|\mathsf{Q}(\xi)| \leq 1$, and $\widetilde{\mathbb{E}}\{\mathsf{Q}(\xi)\} = 0$, we have, for any $\xi_1, \ldots, \xi_k \in \mathcal{X}$,

$$\begin{aligned}
|\widetilde{\mathbb{E}}\{\mathsf{Q}(\xi_1)\cdots\mathsf{Q}(\xi_k)\}| &\leq |\widetilde{\mathbb{E}}\{\mathsf{Q}(\xi_1)\mathsf{Q}(\xi_2)\}| \leq \sqrt{\widetilde{\mathbb{E}}\{\mathsf{Q}(\xi_1)^2\}\widetilde{\mathbb{E}}\{\mathsf{Q}(\xi_2)^2\}} \leq \\
&\leq \frac{1}{2}\mathrm{Var}(\mathsf{Q}(\xi_1)|Y=y, Z=z(\theta)) + \frac{1}{2}\mathrm{Var}(\mathsf{Q}(\xi_2)|Y=y, Z=z(\theta))\,,
\end{aligned}$$

(where we assumed, without loss of generality, $k \geq 2$). Integrating with respect to y and $z(\theta)$ with the measure $\mathbb{P}_{Y,Z(\theta)}$, and using Lemma 3.2, we obtain

$$\int_0^{\epsilon} \mathbb{E}\Big|\mathbb{E}\{\mathsf{Q}(\xi_1)\cdots\mathsf{Q}(\xi_k)|Y, Z(\theta)\}\Big|\, \mathrm{d}\theta \leq 4H(X_1)/n\,. \tag{3.15}$$

On the other hand

$$\widetilde{\mathbb{E}}\{\mathsf{Q}(\xi_1)\cdots\mathsf{Q}(\xi_k)\} = \frac{1}{n^k}\sum_{j(1)\dots j(k)\in V}\widetilde{\mathbb{E}}\{\mathsf{S}^{(1)}_{j(1)}(\xi_1)\mathsf{S}^{(2)}_{j(1)}(\xi_2)\cdots\mathsf{S}^{(1)}_{j(k)}(\xi_k)\mathsf{S}^{(2)}_{j(k)}(\xi_k)\} = \tag{3.16}$$

$$= \frac{1}{n^k}\sum_{j(1)\dots j(k)\in V}\widetilde{\mathbb{E}}\{\mathsf{S}_{j(1)}(\xi_1)\cdots\mathsf{S}_{j(k)}(\xi_k)\}^2 \geq \tag{3.17}$$

$$\geq \frac{k!}{n^k}\binom{n}{k}\mathbb{E}_{i(1)\dots i(k)}\widetilde{\mathbb{E}}\{\mathsf{S}_{i(1)}(\xi_1)\cdots\mathsf{S}_{i(k)}(\xi_k)\}^2\,. \tag{3.18}$$

Putting together Eq. (3.15) and (3.18), letting $B_{n,k}\equiv n^k/k!\binom{n}{k}$, and taking expectation with respect to Y and $Z(\theta)$, we get

$$\int_0^\epsilon \mathbb{E}_{i(1)\dots i(k)}\mathbb{E}\Big\{\mathbb{E}\{\mathsf{S}_{i(1)}(\xi_1)\cdots\mathsf{S}_{i(k)}(\xi_k)|Y,Z(\theta)\}^2\Big\}\,\mathrm{d}\theta \leq 4B_{n,k}H(X_1)/n\,, \tag{3.19}$$

which, by Cauchy-Schwarz inequality, implies

$$\int_0^\epsilon \mathbb{E}_{i(1)\dots i(k)}\mathbb{E}\Big\{\big|\mathbb{E}\{\mathsf{S}_{i(1)}(\xi_1)\cdots\mathsf{S}_{i(k)}(\xi_k)|Y,Z(\theta)\}\big|\Big\}\,\mathrm{d}\theta \leq \sqrt{4\epsilon B_{n,k}H(X_1)/n}\,. \tag{3.20}$$

Next notice that

$$\Big|\Big|\widetilde{\mathbb{P}}_{i(1)\dots i(k)}-\widetilde{\mathbb{P}}_{i(1)}\cdots\widetilde{\mathbb{P}}_{i(k)}\Big|\Big|_{\rm TV} = \frac{1}{2}\sum_{\xi_1\dots\xi_k\in\mathcal{X}}\Big|\widetilde{\mathbb{P}}_{i(1)\dots i(k)}\{\xi_1,\dots,\xi_k\}-\widetilde{\mathbb{P}}_{i(1)}\{\xi_1\}\cdots\widetilde{\mathbb{P}}_{i(k)}\{\xi_k\}\Big| =$$

$$=\frac{1}{2}\sum_{\xi_1\dots\xi_k\in\mathcal{X}}\Big|\widetilde{\mathbb{E}}\left\{\mathbb{I}(X_{i(1)}=\xi_1)\cdots\mathbb{I}(X_{i(k)}=\xi_k)-\widetilde{\mathbb{P}}_{i(1)}\{\xi_1\}\cdots\widetilde{\mathbb{P}}_{i(k)}\{\xi_k\}\right\}\Big| =$$

$$=\frac{1}{2}\sum_{\xi_1\dots\xi_k\in\mathcal{X}}\left|\sum_{J\in[k],\,|J|\geq 2}\widetilde{\mathbb{E}}\left\{\prod_{\alpha\in J}\mathsf{S}_{i(\alpha)}(\xi_\alpha)\right\}\prod_{\beta\in[k]\setminus J}\widetilde{\mathbb{P}}_{i(\beta)}\{\xi_\beta\}\right|\,.$$

Using triangular inequality

$$\Big|\Big|\widetilde{\mathbb{P}}_{i(1)\dots i(k)}-\widetilde{\mathbb{P}}_{i(1)}\cdots\widetilde{\mathbb{P}}_{i(k)}\Big|\Big|_{\rm TV} \leq \frac{1}{2}\sum_{J\in[k],\,|J|\geq 2}\;\sum_{\{\xi_\alpha\}_{\alpha\in J}}\left|\widetilde{\mathbb{E}}\left\{\prod_{\alpha\in J}\mathsf{S}_{i(\alpha)}(\xi_\alpha)\right\}\right|\,.$$

Taking expectation with respect to $Y, Z(\theta)$ and to $\{i(1),\dots,i(k)\}$ a uniformly random subset of V, we obtain

$$\mathbb{E}_{i(1)\dots i(k)}\mathbb{E}\Big|\Big|\widetilde{\mathbb{P}}_{i(1)\dots i(k)}-\widetilde{\mathbb{P}}_{i(1)}\cdots\widetilde{\mathbb{P}}_{i(k)}\Big|\Big|_{\rm TV} \leq$$

$$\leq \frac{1}{2}\sum_{l=2}^{k}\binom{k}{l}\sum_{\xi_1\dots\xi_l\in\mathcal{X}}\mathbb{E}_{i(1)\dots i(l)}\mathbb{E}\Big|\mathbb{E}\{\mathsf{S}_{i(1)}(\xi_1)\cdots\mathsf{S}_{i(l)}(\xi_l)\big|Y,Z(\theta)\}\Big|\,.$$

Integrating over θ and using Eq. (3.20), we get

$$\int_0^\epsilon \mathbb{E}_{i(1)\cdots i(k)}\mathbb{E}\Big|\Big|\widetilde{\mathbb{P}}_{i(1),\dots,i(k)}-\widetilde{\mathbb{P}}_{i(1)}\cdots\widetilde{\mathbb{P}}_{i(k)}\Big|\Big|_{\rm TV}\,\mathrm{d}\theta \leq \frac{1}{2}\sum_{l=2}^{k}\binom{k}{l}|\mathcal{X}|^l\sqrt{4\epsilon B_{n,l}H(X_1)/n}\,. \tag{3.21}$$

By using $B_{n,l}\leq B_{n,k}$ the right hand side is bounded as in Eq. (2.15), with $A_{n,k}\equiv\sqrt{B_{n,k}}$. The bound on this coefficient is obtained by a standard manipulation (here we use $-\log(1-x)\leq 2x$ for $x\in[0,1/2]$ and the hypothesis $k\leq n/2$):

$$B_{n,k} = \exp\left\{-\sum_{i=1}^{k-1}\log\left(1-\frac{i}{n}\right)\right\} \leq \exp\left\{\sum_{i=1}^{k-1}\frac{2i}{n}\right\} = \exp\left\{\frac{k(k-1)}{n}\right\}\,, \tag{3.22}$$

hence $A_{n,k} \leq \exp\{k^2/2n\}$ as claimed. □

Obviously, if the graph G is 'sufficiently' random, the expectation over variable nodes $i(1), \dots, i(k)$ can be replaced by the expectation over G.

Corollary 3.3. *Let $G = (V, F, E)$ be a random bipartite graph whose distribution is invariant under permutation of the variable nodes in $V = [n]$. Then, for any observations system on $G = (V, F, E)$, any $k \in \mathbb{N}$ any $\epsilon > 0$, and any (fixed) set of variable nodes $\{i(1), \dots, i(k)$,*

$$\int_0^\epsilon \mathbb{E}_G\mathbb{E}\Big|\Big|\widetilde{\mathbb{P}}_{i(1),\dots,i(k)} - \widetilde{\mathbb{P}}_{i(1)} \cdots \widetilde{\mathbb{P}}_{i(k)}\Big|\Big|_{\rm TV}\, {\rm d}\theta \leq (|\mathcal{X}| + 1)^k A_{n,k}\sqrt{H(X_1)\epsilon/n} = O(n^{-1/2})\,, \tag{3.23}$$

where the constant $A_{k,n}$ is as in Theorem 2.2.

4 Random graph properties

The proofs of Theorems 2.3 and 2.4 rely on some specific properties of the graph ensemble $\mathcal{G}(n, \alpha n, \gamma/n)$.

We begin with some further definitions concerning a generic bipartite graph $G = (V, F, E)$. Given $i, j \in V$, their graph-theoretic distance is defined as the length of the shortest path from i to j on G. We follow the convention of measuring the length of a path on G by the number of function nodes traversed by the path.

Given $i \in V$ and $t \in \mathbb{N}$ we let $\mathsf{B}(i, t)$ be the subset of variable nodes j whose distance from i is at most t. With an abuse of notation, we use the same symbol to denote the subgraph induced by this set of vertices, i.e. the factor graph including those function node a such that $\partial a \subseteq \mathsf{B}(i, t)$ and all the edges incident on them. Further, we denote by $\overline{\mathsf{B}}(i, t)$ the subset of variable nodes j with $d(i, j) \geq t$, as well as the induced subgraph. Finally $\mathsf{D}(i, t)$ is the subset of vertices with $d(i, j) = t$. Equivalently $\mathsf{D}(i, t)$ is the intersection of $\mathsf{B}(i, t)$ and $\overline{\mathsf{B}}(i, t)$.

We will make use of two remarkable properties of the ensemble $\mathcal{G}(n, n\alpha, \gamma/n)$: (i) The convergence of any finite neighborhood in G to an appropriate tree model; (ii) The conditional independence of such a neighborhood from the residual graph, given the neighborhood size.

The limit tree model is defined by the following sampling procedure, yielding a t-generations rooted random tree $\mathsf{T}(t)$. If $t = 0$, $\mathsf{T}(t)$ is the trivial tree consisting of a single variable node. For $t \geq 1$, start from a distinguished root variable node i and connect it to l function nodes, whereby l is a Poisson random variable with mean $\gamma\alpha$. For each such function nodes a, draw an independent Poisson(γ) random variable k_a and connect it to k_a new variable nodes. Finally, for each of the 'first generation' variable node j, sample an independent random tree distributed as $\mathsf{T}(t - 1)$, and attach it by the root to j.

Proposition 4.1 (Convergence to random tree)**.** *Let $\mathsf{B}(i, t)$ be the radius-t neighborhood of any fixed variable node i in a random graph $G \stackrel{\rm d}{=} \mathcal{G}(n, \alpha n, \gamma/n)$, and $\mathsf{T}(t)$ the random tree defined above.*

Given any (labeled) tree T_*, we write $\mathsf{B}(i, t) \simeq \mathsf{T}_*$ if T_* is obtained by the depth-first relabeling of $\mathsf{B}(i, t)$ following a pre-established convention[4]. Then

$$\lim_{n\to\infty} \mathbb{P}\{\mathsf{B}(i, t) \simeq \mathsf{T}_*\} = \mathbb{P}\{\mathsf{T}_t \simeq \mathsf{T}_*\}\,. \tag{4.1}$$

Proposition 4.2 (Bound on the neighborhood size)**.** *Let $\mathsf{B}(i, t)$ be the radius-t neighborhood of any fixed variable node i in a random bipartite graph $G \stackrel{\rm d}{=} \mathcal{G}(n, \alpha n, \gamma/n)$, and denote by $|\mathsf{B}(i, t)|$ its size (number of variable and function nodes). Then, for any $\lambda > 0$ there exists $C(\lambda, t)$ such that, for any n, $M \geq 0$*

$$\mathbb{P}\{|\mathsf{B}(i, t)| \geq M\} \leq C(\lambda, t)\, \lambda^{-M}\,. \tag{4.2}$$

Proof. Let us generalize our definition of neighborhood as follows. If t is integer, we let $\mathsf{B}(i, t + 1/2)$ be the subgraph including $\mathsf{B}(i, t)$ together with all the function nodes that have at least one neighbor in

[4] For instance, one might agree to preserve the original lexicographic order among siblings.

$\mathsf{B}(i,t)$ (as well as the edges to $\mathsf{B}(i,t)$). We also let $\mathsf{D}(i,t+1/2)$ be the set of function nodes that have at least one neighbor in $\mathsf{B}(i,t)$ and at least one outside.

Imagine to explore $\mathsf{B}(i,t)$ in breadth-first fashion. For each t, $|\mathsf{B}(i,t+1/2)|-|\mathsf{B}(i,t)|$ is upper bounded by the sum of $|\mathsf{D}(i,t)|$ iid binomial random variables counting the number of neighbors of each node in $\mathsf{D}(i,t)$, which are not in $\mathsf{B}(i,t)$. For t integer (respectively, half-integer), each such variables is stochastically dominated by a binomial with parameters $n\alpha$ (respectively, n) and γ/n. Therefore $|\mathsf{B}(i,t)|$ is stochastically dominated by $\sum_{s=0}^{2t} Z_n(s)$, where $\{Z_n(t)\}$ is a Galton-Watson process with offspring distribution $\text{Binom}(n,\overline{\gamma}/n)$ and $\overline{\gamma}=\gamma\,\max(1,\alpha)$.

By Markov inequality

$$\mathbb{P}\{|\mathsf{B}(i,t)|\geq M\}\leq g_{2t}^n(\lambda)\,\lambda^{-M}\,,\qquad g_t^n(\lambda)\equiv\mathbb{E}\{\lambda^{\sum_{s=0}^t Z_n(s)}\}.$$

By elementary branching processes theory $g_t^n(\lambda)$ satisfies the recursion $g_{t+1}^n(\lambda)=\lambda\xi_n(g_t^n(\lambda))$, $g_0^n(\lambda)=\lambda$, with $\xi_n(\lambda)=\lambda(1+2\overline{\gamma}(\lambda-1)/n)^n$. The thesis follows by $g_t^n(\lambda)\leq g_t(\lambda)$, where $g_t(\lambda)$ is defined as $g_t^n(\lambda)$ but replacing $\xi_n(\lambda)$ with $\xi(\lambda)=e^{2\gamma(\lambda-1)}\geq\xi_n(\lambda)$. □

Proposition 4.3. *Let $G=(V,F,E)$ be a random bipartite graph from the ensemble $\mathcal{G}(n,m,p)$. Then, conditional on $\mathsf{B}(i,t)=(V(i,t),F(i,t),E(i,t))$, $\overline{\mathsf{B}}(i,t)$ is a random bipartite graph on variable nodes $V\setminus V(i,t-1)$, function nodes $F\setminus F(i,t)$ and same edge probability p.*

Proof. Condition on $\mathsf{B}(i,t)=(V(i,t),F(i,t),E(i,t))$, and let $\mathsf{B}(r,t-1)=(V(i,t-1),F(i,t-1),E(i,t-1))$ (notice that this is uniquely determined from $\mathsf{B}(i,t)$). This is equivalent to conditioning on a given edge realization for any two vertices k, a such that $k\in V(i,t)$ and $a\in F(i,t)$.

On the other hand, $\overline{\mathsf{B}}(i,t)$ is the graph with variable nodes set $\overline{V}\equiv V\setminus V(i,t-1)$, function nodes $\overline{F}\equiv F\setminus F(i,t)$, and edge set $(k,a)\in G$ such that $k\in\overline{V}$, $a\in\overline{F}$. Since this set of vertices couples is disjoint from the one we are conditioning upon, and by independence of edges in G, the claim follows. □

5 Proof of Theorem 2.3 (BP equations)

The proof of Theorem 2.3 hinges on the properties of the random factor graph G discussed in the previous Section as well as on the correlation structure unveiled by Theorem 2.2.

5.1 The effect of changing G

The first need to estimate the effect on changing the graph G on marginals.

Lemma 5.1. *Let X be a random variable taking values in $\mathcal{X}$ and assume $X\to G\to Y_1\to B$ and $X\to G\to Y_1\to B$ to be Markov chains (here G, $Y_{1,2}$ and B are arbitrary random variables, where G stands for* good *and B for* bad*). Then*

$$\mathbb{E}\big|\big|\mathbb{P}\{X\in\cdot\,|Y_1\}-\mathbb{P}\{X\in\cdot\,|Y_2\}\big|\big|_{\text{TV}}\leq 2\,\mathbb{E}\big|\big|\mathbb{P}\{X\in\cdot\,|G\}-\mathbb{P}\{X\in\cdot\,|B\}\big|\big|_{\text{TV}}\,. \tag{5.1}$$

Proof. First consider a single Markov Chain $X\to G\to Y\to B$. Then, by convexity of the total variation distance,

$$\begin{aligned}
\mathbb{E}\big|\big|\mathbb{P}\{X\in\cdot\,|Y\}-\mathbb{P}\{X\in\cdot\,|B\}\big|\big|_{\text{TV}} &= \mathbb{E}\Big|\Big|\mathbb{E}\Big\{\mathbb{P}\{X\in\cdot\,|G,Y\}\Big|Y\Big\}-\mathbb{P}\{X\in\cdot\,|B\}\Big|\Big|_{\text{TV}}\leq & (5.2)\\
&\leq \mathbb{E}\,||\mathbb{P}\{X\in\cdot\,|G,Y\}-\mathbb{P}\{X\in\cdot\,|B\}||_{\text{TV}}= & (5.3)\\
&= \mathbb{E}\,||\mathbb{P}\{X\in\cdot\,|G\}-\mathbb{P}\{X\in\cdot\,|B\}||_{\text{TV}}\,. & (5.4)
\end{aligned}$$

The thesis is proved by applying this bound to both chains $X\to G\to Y_1\to B$ and $X\to G\to Y_1\to B$, and using triangular inequality. □

The next lemma estimates the effect of removing one variable node from the graph. Notice that the graph G is non-random.

Lemma 5.2. *Consider two observation systems associated to graphs $G = (V, F, E)$ and $G' = (V', F', E')$ whereby $V = V' \setminus \{j\}$, $F = F'$ and $E = E' \setminus \{(j, b) : b \in \partial j\}$. Denote the corresponding observations as $(Y, Z(\theta))$ and $(Y', Z'(\theta))$. Then there exist a coupling of the observations such that, for any $i \in V$:*

$$\mathbb{E}||\mathbb{P}\{X_i \in \cdot|Y, Z(\theta)\} - \mathbb{P}\{X_i \in \cdot|Y', Z'(\theta)\}||_{\rm TV} \leq \tag{5.5}$$
$$4\,\mathbb{E}\Big|\Big|\mathbb{P}_{i,\partial^2 j}\{\cdots|Y_{F\setminus\partial j}, Z(\theta)\} - \prod_{l\in\{i,\partial^2 j\}} \mathbb{P}_l\{\,\cdot\,|Y_{F\setminus\partial j}, Z(\theta)\}|\Big|\Big|_{\rm TV},$$

where $\partial^2 j \equiv \{l \in V : d(i, l) = 1\}$ and used the shorthand $\mathbb{P}_U\{\cdots|Y_{F\setminus\partial j}, Z(\theta)\}$ for $\mathbb{P}\{X_U \in \cdots|Y_{F\setminus\partial j}, Z(\theta)\}$.

The coupling consists in sampling $X = \{X_i : i \in V\}$ from its (iid) distribution and then $(Y, Z(\theta))$ and $(Y', Z'(\theta))$ as observations of this configuration X, in such a way that $Z(\theta) = Z'(\theta)$ and $Y_a = Y'_a$ for any $a \in F$ such that $\partial a \in V$.

Proof. Let us describe the coupling more explicitly. First sample $X = \{X_i : i \in V\}$ and $X' = \{X'_i : i \in V'\}$ in such a way that $X_i = X'_i$ for any $i \in V$. Then, for any $i \in V$, sample $Z_i(\theta)$, $Z'_i(\theta)$ conditionally on $X_i = X'_i$ in such a way that $Z_i(\theta) = Z'_i(\theta)$. Sample $Z'_j(\theta)$ conditionally on X'_j. For any $a \in F$ such that $\partial a \in V$, sample Y_a, Y'_a conditionally on $X_{\partial a} = X'_{\partial a}$ in such a way that $Y_a = Y'_a$. Finally for $a \in \partial j$, sample Y_a, Y'_a independently, conditional on $X_{\partial a} \neq X'_{\partial a}$

Notice that the following are Markov Chains

$$X_i \to (X_{\partial^2 j}, Y, Z(\theta)) \;\; \to (Y, Z(\theta)) \to (Y_{F\setminus\partial j}, Z(\theta)), \tag{5.6}$$
$$X_i \to (X_{\partial^2 j}, Y, Z(\theta)) \to (Y', Z'(\theta)) \to (Y_{F\setminus\partial j}, Z(\theta)). \tag{5.7}$$

The only non-trivial step is $(X_{\partial^2 j}, Y, Z(\theta)) \to (Y', Z'(\theta))$. Notice that, once $X_{\partial^2 j}$ is known, $Y_{\partial j}$ is conditionally independent from the other random variables. Therefore we can produce $(Y', Z'(\theta))$ first scratching $Y_{\partial j}$, then sampling X'_j independently, next sampling $Y'_{\partial j}$ and $Z'_j(\theta)$ and finally scratching both X'_j and $X_{\partial^2 j}$.

Applying Lemma 5.1 to the chains above, we get

$$\begin{aligned}
&\mathbb{E}\big|\big|\mathbb{P}\{X_i \in \cdot|Y, Z(\theta)\} - \mathbb{P}\{X_i \in \cdot|Y', Z'(\theta)\}\big|\big|_{\rm TV} \leq \\
&\qquad \leq 2\,\mathbb{E}\big|\big|\mathbb{P}\{X_i \in \cdot|Y_{F\setminus\partial j}, Z(\theta)\} - \mathbb{P}\{X_i \in \cdot|X_{\partial^2 j}, Y, Z(\theta)\}\big|\big|_{\rm TV} = \\
&\qquad = 2\,\mathbb{E}\big|\big|\mathbb{P}\{X_i \in \cdot|Y_{F\setminus\partial j}, Z(\theta)\} - \mathbb{P}\{X_i \in \cdot|X_{\partial^2 j}, Y_{F\setminus\partial j}, Z(\theta)\}\big|\big|_{\rm TV},
\end{aligned}$$

where in the last step, we used the fact that $Y_{\partial j}$ is conditionally independent of X_i, given $X_{\partial^2 j}$. The thesis is proved using the identity (valid for any two random variables U, W)

$$\mathbb{E}||\mathbb{P}\{U \in \cdot\} - \mathbb{P}\{U \in \cdot|W\}||_{\rm TV} = ||\mathbb{P}\{(U, W) \in \cdots\} - \mathbb{P}\{U \in \cdot\}\mathbb{P}\{W \in \cdot\}||_{\rm TV}, \tag{5.8}$$

and the bound (that follows from triangular inequality)

$$\begin{aligned}
&||\mathbb{P}\{(U, W_1 \dots W_k) \in \cdots\} - \mathbb{P}\{U \in \cdots\}\mathbb{P}\{(W_1 \dots W_k) \in \cdots\}||_{\rm TV} \leq \\
&\qquad \leq 2\,||\mathbb{P}\{(U, W_1 \dots W_k) \in \cdots\} - \mathbb{P}\{U \in \cdots\}\mathbb{P}\{W_1 \in \cdot\}\cdots\mathbb{P}\{W_k \in \cdot\}||_{\rm TV}.
\end{aligned}$$

□

An analogous Lemma estimates the effect of removing a function node.

Lemma 5.3. *Consider two observation systems associated to graphs $G = (V, F, E)$ and $G' = (V', F', E')$ whereby $V = V'$, $F = F' \setminus \{a\}$ and $E = E' \setminus \{(j, a) : j \in \partial a\}$. Denote the corresponding observations as $(Y, Z(\theta))$ and $(Y', Z'(\theta))$, with $Z(\theta) = Z'(\theta)$ and $Y = Y' \setminus \{a\}$. Then, for any $i \in V$:*

$$\mathbb{E}||\mathbb{P}\{X_i \in \cdot|Y, Z(\theta)\} - \mathbb{P}\{X_i \in \cdot|Y', Z'(\theta)\}||_{\rm TV} \leq \tag{5.9}$$
$$4\,\mathbb{E}\Big|\Big|\mathbb{P}_{i,\partial a}\{\cdots|Y_{F\setminus\partial a}, Z(\theta)\} - \prod_{l\in\{i,\partial a\}} \mathbb{P}_l\{\,\cdot\,|Y_{F\setminus\partial a}, Z(\theta)\}|\Big|\Big|_{\rm TV}.$$

where we used the shorthand $\mathbb{P}_U\{\cdots|Y_{F\setminus a}, Z(\theta)\}$ for $\mathbb{P}\{X_U \in \cdots|Y_{F\setminus a}, Z(\theta)\}$.

Proof. The proof is completely analogous (and indeed easier) to the one of Lemma 5.2. It is sufficient to consider the Markov chain $X_i \to (X_{\partial a}, Y, Z(\theta)) \to (Y, Z(\theta)) \to (Y_{F\setminus a}, Z(\theta))$, and bound the total variation distance considered here in terms of the first and last term in the chain, where we notice that $(Y_{F\setminus a}, Z(\theta)) = (Y', Z'(\theta))$. We omit details to avoid redundancies. □

Next, we study the effect of removing a variable node from a random bipartite graph.

Lemma 5.4. *Let $G = (V, F, E)$ and $G' = (V', F', E')$ be two random graphs from, respectively, the $\mathcal{G}(n-1, \alpha n, \gamma/n)$ and $\mathcal{G}(n, \alpha n, \gamma/n)$ ensembles. Consider two information systems on such graphs. Let $(Y, Z(\theta))$ and $(Y', Z'(\theta))$ be the corresponding observations, and μ_i^θ, $\mu_i^{\theta\prime}$ the conditional distributions of X_i in the two systems.*

It is then possible to couple G to G' and, for each θ $(Y, Z(\theta))$ to $(Y', Z'(\theta))$ and choose a constant $\mathsf{C} = \mathsf{C}(|\mathcal{X}|, \alpha, \gamma)$ (bounded uniformly for γ and $1/\alpha$ bounded), such that, for any $\epsilon > 0$ and any $i \in V \cap V'$,

$$\int_0^\epsilon \mathbb{E}_G \mathbb{E}||\mu_i^\theta - \mu_i^{\theta\prime}||_{\rm TV} \le \frac{\mathsf{C}}{\sqrt{n}}\,. \tag{5.10}$$

Further, such a coupling can be produced by letting $V' = V \cup \{n\}$, $F' = F$ and $E' = E \cup \{(n, a) : a \in \partial n\}$ where $a \in \partial n$ independently with probability γ/n. Finally $(Y, Z(\theta))$ and $(Y', Z'(\theta))$ are coupled as in Lemma 5.2.

Proof. Take $V = [n-1]$, $V' = [n]$, $F = F' = [n\alpha]$ and sample the edges by letting, for any $i \in [n-1]$, $(i, a) \in E$ if and only if $(i, a) \in E'$. Therefore $E = E' \setminus \{(n, a) :\ a \in \partial n\}$ (here ∂n is the neighborhood of variable node n with respect to the edge set E'). Coupling $(Y, Z(\theta))$ and $(Y', Z'(\theta))$ as in Lemma 5.2, and using the bound proved there, we get

$$\int_0^\epsilon \mathbb{E}_G \mathbb{E}||\mu_i^\theta - \mu_i^{\theta\prime}||_{\rm TV} \le 4\int_0^\epsilon \mathbb{E}_G \mathbb{E}\Big|\Big|\mathbb{P}_{i,\partial^2 n}\{\,\cdots|Y_{F\setminus\partial n}, Z(\theta)\} - \prod_{l\in\{i,\partial^2 n\}} \mathbb{P}_l\{\,\cdot\,|Y_{F\setminus\partial n}, Z(\theta)\}\Big|\Big|_{\rm TV} {\rm d}\theta\,,\,. \tag{5.11}$$

In order to estimate the total variation distance on the right hand side, we shall condition on $|\partial n|$ and $|\partial^2 n|$. Once this is done, the conditional probability $\mathbb{P}_{i,\partial^2 n}\{\,\cdots|Y_{F\setminus\partial n}, Z(\theta)\}$ is distributed as the conditional probability of $|\partial^2 n|+$ variables, in a system $\widehat{G}(|\partial n|)$ with $n-1$ variable nodes and $n\alpha - |\partial n|$ function nodes. Let us denote by $(\widehat{Y}, \widehat{Z}(\theta))$ the corresponding observations (and by $\widehat{\mathbb{P}}$, $\widehat{\mathbb{E}}$ probability and expectations). Then the right hand side in Eq. (5.11) is equal to

$$\begin{aligned}
4\int_0^\epsilon & \mathbb{E}_{|\partial n|,|\partial^2 n|}\mathbb{E}_G\Big\{\mathbb{E}\Big|\Big|\mathbb{P}_{i,\partial^2 n}\{\,\cdots|Y_{F\setminus\partial n}, Z(\theta)\} - \prod_{l\in\{i,\partial^2 n\}} \mathbb{P}_l\{\,\cdot\,|Y_{F\setminus\partial n}, Z(\theta)\}\Big|\Big|_{\rm TV}\Big|\,|\partial n|, |\partial^2 n|\Big\}\,{\rm d}\theta = \\
&= 4\int_0^\epsilon \mathbb{E}_{|\partial n|,|\partial^2 n|}\mathbb{E}_{\widehat{G}(|\partial n|)}\widehat{\mathbb{E}}\Big|\Big|\widehat{\mathbb{P}}_{1\dots|\partial^2 n|+1}\{\,\cdots|\widehat{Y}, \widehat{Z}(\theta)\} - \prod_{l=}^{|\partial^2 n|} \widehat{\mathbb{P}}_l\{\,\cdot\,|\widehat{Y}, \widehat{Z}(\theta)\}\Big|\Big|_{\rm TV} {\rm d}\theta \le \\
&\le 4\mathbb{E}_{|\partial n|,|\partial^2 n|}\Big\{(|\mathcal{X}|+1)^{|\partial^2 n+1|}\sqrt{H(X_1)\epsilon/(n-1)}\Big\} + 4\epsilon\mathbb{P}\{|\partial^2 n| + 1 \ge \sqrt{n}/10\}\,,\,.
\end{aligned}$$

In the last step we applied Corollary 3.3 and distinguished the cases $|\partial^2 n| + 1 \ge \sqrt{n}/10$ (then bounding the total variation distance by 1) and $|\partial^2 n| + 1 < \sqrt{n}/10$ (then bounding $A_{n-1,|\partial^2 n|+1}$ by $\sqrt{2}$ thanks to the estimate in Theorem 2.2). The thesis follows using Proposition 4.2 to bound both terms above (notice in fact that $|\partial^2 n| \le |\mathsf{B}(i, 1)|$). □

Again, an analogous estimate holds for the effect of removing one function node. The proof is omitted as it is almost identical to the previous one.

Lemma 5.5. *Let $G = (V, F, E)$ and $G' = (V', F', E')$ be two random graphs from, respectively, the $\mathcal{G}(n, \alpha n - 1, \gamma/n)$ and $\mathcal{G}(n, \alpha n, \gamma/n)$ ensembles. Consider two information systems on such graphs. Let $(Y, Z(\theta))$ and $(Y', Z'(\theta))$ be the corresponding observations, and μ_i^θ, $\mu_i^{\theta\prime}$ the conditional distributions of X_i in the two systems.*

It is then possible to couple G to G' and, for each θ $(Y, Z(\theta))$ to $(Y', Z'(\theta))$ and choose a constant $\mathsf{C} = \mathsf{C}(|\mathcal{X}|, \alpha, \gamma)$ (bounded uniformly for γ and $1/\alpha$ bounded), such that, for any $\epsilon > 0$ and any $i \in V \cap V'$,

$$\int_0^\epsilon \mathbb{E}_G \mathbb{E}||\mu_i^\theta - \mu_i^{\theta\prime}||_{\rm TV} \le \frac{\mathsf{C}}{\sqrt{n}}\,. \tag{5.12}$$

Further, such a coupling can be produced by letting $V' = V$, $F' = F \setminus \{a\}$, for a fixed function node a, and $E' = E \cup \{(j,a) : j \in \partial a\}$ where $j \in \partial a$ independently with probability γ/n. Finally $(Y, Z(\theta))$ and $(Y', Z'(\theta))$ are coupled as in Lemma 5.3.

5.2 BP equations

We begin by proving a useful technical Lemma.

Lemma 5.6. *Let p_1, p_2 be probability distribution over a finite set $\mathcal{S}$, and $q : \widehat{\mathcal{S}} \times \mathcal{S} \to \mathbb{R}_+$ be a non-negative function. Define, for $a = 1, 2$ the probability distributions*

$$\widehat{p}_a(x) \equiv \frac{\sum_{y\in\mathcal{S}} q(x,y)\, p_a(y)}{\sum_{x'\in\widehat{\mathcal{S}}, y'\in\mathcal{S}} q(x',y')\, p_a(y')}\,. \tag{5.13}$$

Then

$$||\widehat{p}_1 - \widehat{p}_2||_{\rm TV} \le 2\left(\frac{\max_{y\in\mathcal{S}}\sum_x q(x,y)}{\min_{y\in\mathcal{S}}\sum_x q(x,y)}\right)\, ||p_1 - p_2||_{\rm TV}\,. \tag{5.14}$$

Proof. Using the inequality $|(a_1/b_1)-(a_2/b_2)| \le |a_1-a_2|/b_1 + (a_2/b_2)|b_1-b_2|/b_1$ (valid for $a_1, a_2, b_1, b_2 \ge 0$), we get

$$|\widehat{p}_1(x) - \widehat{p}_2(x)| \le \frac{\sum_y q(x,y)|p_1(y)-p_2(y)|}{\sum_{x',y'} q(x',y')p_1(y')} + \frac{\sum_y q(x,y)p_2(y)}{\sum_{x',y'} q(x',y')p_2(y')}\, \frac{\left|\sum_{x',y'} q(x',y')(p_1(y')-p_2(y'))\right|}{\sum_{x',y'} q(x',y')p_1(y')}\,.$$

Summing over x we get

$$||\widehat{p}_1 - \widehat{p}_2||_{\rm TV} \le \frac{\sum_y (\sum_x q(x,y))|p_1(y)-p_2(y)|}{\sum_{y'}(\sum_{x'} q(x',y'))p_1(y')}\,,$$

whence the thesis follows. □

Given a graph G, $i \in V$, $t \ge 1$, we let $\mathsf{B} \equiv \mathsf{B}(i,t)$, $\overline{\mathsf{B}} \equiv \overline{\mathsf{B}}(i,t)$ and $\mathsf{D} \equiv \mathsf{D}(i,t)$, Further, we introduce the shorthands

$$W_{\mathsf{B}} \equiv \{Y_a : \partial a \subseteq \mathsf{B},\, \partial a \not\subseteq \mathsf{D}\} \cup \{Z_i : i \in \mathsf{B}\setminus\mathsf{D}\}\,, \tag{5.15}$$

$$W_{\overline{\mathsf{B}}} \equiv \{Y_a : \partial a \subseteq \overline{\mathsf{B}}\} \cup \{Z_i : i \in \overline{\mathsf{B}}\}\,. \tag{5.16}$$

Notice that W_{B}, $W_{\overline{\mathsf{B}}}$ form a partition of the variables in $Y, Z(\theta)$. Further W_{B}, $W_{\overline{\mathsf{B}}}$ are conditionally independent given X_{D}. As a consequence, we have the following simple bound.

Lemma 5.7. *For any two non-negative functions f and g, we have*

$$\mathbb{E}\{f(W_{\mathsf{B}})\, g(W_{\overline{\mathsf{B}}})\} \le \max_{x_{\mathsf{B}}} \mathbb{E}\{f(W_{\mathsf{B}})|X_{\mathsf{D}} = x_{\mathsf{D}}\}\, \mathbb{E}\{g(W_{\overline{\mathsf{B}}})\}\,. \tag{5.17}$$

Proof. Using the conditional independence property we have

$$\mathbb{E}\Big\{f(W_{\mathsf{B}})\, g(W_{\overline{\mathsf{B}}})\Big\} = \mathbb{E}\{\mathbb{E}[f(W_{\mathsf{B}})|X_{\mathsf{D}}]\, \mathbb{E}[g(W_{\overline{\mathsf{B}}})|X_{\mathsf{D}}]\Big\} \le \max_{x_{\mathsf{B}}} \mathbb{E}\{f(W_{\mathsf{B}})|X_{\mathsf{D}} = x_{\mathsf{D}}\}\, \mathbb{E}\Big\{\mathbb{E}[g(W_{\overline{\mathsf{B}}})|X_{\mathsf{D}}]\Big\}\,,$$

which proves our claim. □

It is easy to see that the conditional distribution of (X_i, X_{D}) takes the form (with an abuse of notation we write $\mathbb{P}\{X_U|\cdots\}$ instead of $\mathbb{P}\{X_U = x_U|\cdots\}$)

$$\begin{aligned}\mathbb{P}\{X_i, X_{\mathsf{D}}|Y, Z(\theta)\} &= \frac{\mathbb{P}\{X_i, W_{\mathsf{B}}|X_{\mathsf{D}}, W_{\overline{\mathsf{B}}}\}\,\mathbb{P}\{X_{\mathsf{D}}|W_{\overline{\mathsf{B}}}\}}{\sum_{X_i', X_{\mathsf{D}}'}\mathbb{P}\{X_i', W_{\mathsf{B}}|X_{\mathsf{D}}', W_{\overline{\mathsf{B}}}\}\,\mathbb{P}\{X_{\mathsf{D}}'|W_{\overline{\mathsf{B}}}\}} = &(5.18)\\ &= \frac{\mathbb{P}\{X_i, W_{\mathsf{B}}|X_{\mathsf{D}}\}\,\mathbb{P}\{X_{\mathsf{D}}|W_{\overline{\mathsf{B}}}\}}{\sum_{X_i', X_{\mathsf{D}}'}\mathbb{P}\{X_i', W_{\mathsf{B}}|X_{\mathsf{D}}'\}\,\mathbb{P}\{X_{\mathsf{D}}'|W_{\overline{\mathsf{B}}}\}}\,. &(5.19)\end{aligned}$$

If B is a small neighborhood of i, the most intricate component in the above formulae is the probability $\mathbb{P}\{X_{\mathsf{D}}|W_{\overline{\mathsf{B}}}\}$. It would be nice if we could replace this term by the product of the marginal probabilities of X_j, for $j \in \mathsf{D}$. We thus define

$$\mathbb{Q}\{X_i, X_{\mathsf{D}}||Y, Z(\theta)\} = \frac{\mathbb{P}\{X_i, W_{\mathsf{B}}|X_{\mathsf{D}}\}\prod_{j\in\mathsf{D}}\mathbb{P}\{X_j|W_{\overline{\mathsf{B}}}\}}{\sum_{X_i', X_{\mathsf{D}}'}\mathbb{P}\{X_i', W_{\mathsf{B}}|X_{\mathsf{D}}'\}\prod_{j\in\mathsf{D}}\mathbb{P}\{X_j'|W_{\overline{\mathsf{B}}}\}}\,. \tag{5.20}$$

Notice that this is a probability kernel, but not a conditional probability (to stress this point we used the double separator $||$).

Finally, we recall the definition of local marginal $\mu_i^\theta(x_i)$ and introduce, by analogy, the approximation $\mu_i^{\theta,t}(\,\cdot\,)$

$$\mu_i^\theta(x_i) = \sum_{X_{\mathsf{D}}}\mathbb{P}\{X_i = x_i, X_{\mathsf{D}}|Y, Z(\theta)\}\,, \qquad \mu_i^{\theta,t}(x_i) \equiv \sum_{X_{\mathsf{D}}}\mathbb{Q}\{X_i = x_i, X_{\mathsf{D}}|Y, Z(\theta)\}\,. \tag{5.21}$$

It is easy to see that, for $t = 1$, $\mu_i^{\theta,t}$ is nothing but the result of applying belief propagation to the marginals of the neighbors of i with respect to the reduced graph that does not include i. Formally, in the notation of Theorem 2.3:

$$\mu_i^{\theta,1}(x_i) = \mathsf{F}_i^n(\{\mu_{j\to a}^\theta\}_{a\in\partial i, j\in\partial a\setminus i})(x_i)\,. \tag{5.22}$$

The result below shows that indeed the boundary condition on X_{D} can be chosen as factorized, thus providing a more general version of Theorem 2.3.

Theorem 5.8 (BP equations, more general version)**.** *Consider an observations system on a random bipartite graph $G = (V, F, E)$ from the $\mathcal{G}(n, \alpha n, \gamma/n)$ ensemble, and assume the noisy observations to be M-soft. Then there exists a constant A depending on $t, \alpha, \gamma, M, |\mathcal{X}|, \epsilon$, such that for any $i \in V$, and any n*

$$\int_0^\epsilon \mathbb{E}_G\mathbb{E}||\mu_i^\theta - \mu_i^{\theta,t}||_{\rm TV}\,{\rm d}\theta \le \frac{\mathsf{A}}{\sqrt{n}}\,. \tag{5.23}$$

Proof. Notice that the definitions of μ_i^θ, $\mu_i^{\theta,t}$ have the same form as $\widehat{p}_1$, $\widehat{p}_2$ in Lemma 5.6, whereby x corresponds to X_i and y to X_{D}. We have therefore

$$||\mu_i^\theta - \mu_i^{\theta,t}||_{\rm TV} \le 2\left(\frac{\max_{x_{\mathsf{D}}}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{D}} = x_{\mathsf{D}}\}}{\min_{x_{\mathsf{D}}}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{D}} = x_{\mathsf{D}}\}}\right)\left|\left|\mathbb{P}\{X_{\mathsf{D}} = \,\cdot\,|W_{\overline{\mathsf{B}}}\} - \prod_{j\in\mathsf{D}}\mathbb{P}\{X_j = \,\cdot\,|W_{\overline{\mathsf{B}}}\}\right|\right|_{\rm TV}. \tag{5.24}$$

Given observations $Z(\theta)$ and $U \subseteq V$, let us denote as $\mathfrak{C}(Z(\theta), U)$ the values of x_U such that, for any $i \in U$ with $Z_i(\theta) = (Z_i, x_i^0)$ with $x_i^0 \neq *$, one has $x_i = x_i^0$ (i.e. the set of assignments x_U that are compatible with direct observations). Notice that the factor in parentheses can be upper bounded as

$$\begin{aligned}\frac{\max_{x_{\mathsf{D}}}\sum_{x_{\mathsf{B}\setminus\mathsf{D}}}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}}{\min_{x_{\mathsf{D}}}\sum_{x_{\mathsf{B}\setminus\mathsf{D}}}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}} &\le \frac{\max_{x_{\mathsf{D}}}\max_{x_{\mathsf{B}\setminus\mathsf{D}}\in\mathfrak{C}(Z(\theta),\mathsf{B}\setminus\mathsf{D})}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}}{\min_{x_{\mathsf{D}}}\min_{x_{\mathsf{B}\setminus\mathsf{D}}\in\mathfrak{C}(Z(\theta),\mathsf{B}\setminus\mathsf{D})}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}} = &(5.25)\\ &\le \frac{\max_{x_{\mathsf{B}}\in\mathfrak{C}(Z(\theta),\mathsf{B}\setminus\mathsf{D})}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}}{\min_{x_{\mathsf{B}}\in\mathfrak{C}(Z(\theta),\mathsf{B}\setminus\mathsf{D})}\mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}}\,. &(5.26)\end{aligned}$$

Using Lemma 5.7 to take expectation with respect to the observations $(Y, Z(\theta))$, we get

$$\mathbb{E}||\mu_i^\theta - \mu_i^{\theta,t}||_{\rm TV} \le C(\mathsf{B})\, \mathbb{E}\left|\left|\mathbb{P}\{X_{\mathsf{D}} = \,\cdot\,|W_{\overline{\mathsf{B}}}\} - \prod_{j\in\mathsf{D}} \mathbb{P}\{X_j = \,\cdot\,|W_{\overline{\mathsf{B}}}\}\right|\right|_{\rm TV}, \tag{5.27}$$

where (with the shorthand $\mathbb{P}\{\,\cdot\,|x_U\}$ for $\mathbb{P}\{\,\cdot\,|X_U = x_U\}$, and omitting the arguments from $\mathfrak{C}$, since they are clear from the context)

$$\begin{aligned}
C(\mathsf{B}) &= 2\max_{x'_{\mathsf{D}}} \mathbb{E}\left\{\left(\frac{\max_{x_{\mathsf{D}}} \mathbb{P}\{W_{\mathsf{B}}|x_{\mathsf{D}}\}}{\min_{x_{\mathsf{D}}} \mathbb{P}\{W_{\mathsf{B}}|x_{\mathsf{D}}\}}\right)\middle| X_{\mathsf{D}} = x'_{\mathsf{D}}\right\} \le \\
&\le 2\max_{x'_{\mathsf{B}}} \mathbb{E}\left\{\left(\frac{\max_{x_{\mathsf{B}}\in\mathfrak{C}} \mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}}{\min_{x_{\mathsf{B}}\in\mathfrak{C}} \mathbb{P}\{W_{\mathsf{B}}|X_{\mathsf{B}} = x_{\mathsf{B}}\}}\right)\middle| X_{\mathsf{B}} = x'_{\mathsf{B}}\right\} \le \\
&\le 2\max_{x'_{\mathsf{B}}} \mathbb{E}\left\{\prod_{a\in\mathsf{B}} \frac{\max_{x_{\partial a}} \mathbb{P}\{Y_a|x_{\partial a}\}}{\min_{x_{\partial a}} \mathbb{P}\{Y_a|x_{\partial a}\}} \prod_{i\in\mathsf{B}\setminus\mathsf{D}} \frac{\max_{x_i\in\mathfrak{C}} \mathbb{P}\{Z_i(\theta)|x_i\}}{\min_{x_i\in\mathfrak{C}} \mathbb{P}\{Z_i(\theta)|x_i\}}\middle| X_{\mathsf{B}} = x'_{\mathsf{B}}\right\} \le \\
&\le \prod_{a\in\mathsf{B}} \max_{x'_{\partial a}} \mathbb{E}\left\{\frac{\max_{x_{\partial a}} \mathbb{P}\{Y_a|x_{\partial a}\}}{\min_{x_{\partial a}} \mathbb{P}\{Y_a|x_{\partial a}\}}\middle| X_{\partial a} = x'_{\partial a}\right\} \prod_{i\in\mathsf{B}\setminus\mathsf{D}} \max_{x'_i} \mathbb{E}\left\{\frac{\max_{x_i} \mathbb{P}\{Z_i|x_i\}}{\min_{x_i} \mathbb{P}\{Z_i|x_i\}}\middle| X_i = x'_i\right\} \le M^{|\mathsf{B}|}\,.
\end{aligned}$$

In the last step we used the hypothesis of soft noise, and before we changed $Z_i(\theta)$ in Z_i because the difference is irrelevant under the restriction $x \in \mathfrak{C}$, and subsequently removed this restriction.

We now the expectation of Eq. (5.27) over the random graph G, conditional on B

$$\mathbb{E}_G\left\{\mathbb{E}||\mu_i^\theta - \mu_i^{\theta,t}||_{\rm TV}\middle|\mathsf{B}\right\} \le M^{|\mathsf{B}|}\mathbb{E}_G\left\{\mathbb{E}\left|\left|\mathbb{P}\{X_{\mathsf{D}} = \,\cdot\,|W_{\overline{\mathsf{B}}}\} - \prod_{j\in\mathsf{D}} \mathbb{P}\{X_j = \,\cdot\,|W_{\overline{\mathsf{B}}}\}\right|\right|_{\rm TV}\middle|\mathsf{B}\right\}, \tag{5.28}$$

Notice that the the conditional expectation is equivalent to an expectation over a random graph on variable nodes $(V \setminus V(\mathsf{B})) \cup \mathsf{D}$, and function nodes $F \setminus F(\mathsf{B})$ (where $V(\mathsf{B})$ and $F(\mathsf{B})$ denotes the variable and function node sets of B). The distribution of this 'residual graph' is the same as for the original ensemble: for any $j \in (V \setminus V(\mathsf{B})) \cup \mathsf{D}$ and any $b \in F \setminus F(\mathsf{B})$, the edge (j, b) is included independently with probability γ/n. We can therefore apply Corollary 3.3

$$\int_0^\epsilon \mathbb{E}_G\left\{\mathbb{E}||\mu_i^\theta - \mu_i^{\theta,t}||_{\rm TV}\middle|\mathsf{B}\right\}\, {\rm d}\theta \le M^{|\mathsf{B}|}(|\mathcal{X}|+1)^{|\mathsf{B}|}A_{n-|\mathsf{B}|,|\mathsf{B}|}\sqrt{H(X_1)\epsilon/(n-|\mathsf{B}|)}\,. \tag{5.29}$$

We can now take expectation over $\mathsf{B} = \mathsf{B}(i,t)$, and invert expectation and integral over θ, since the integrand is non-negative and bounded. We single out the case $|\mathsf{B}| > \sqrt{n}/10$ and upper bound the total variation distance by 1 in this case. In the case $|\mathsf{B}| \le \sqrt{n}/10$ we upper bound $A_{|\mathsf{B}|,n-|\mathsf{B}|}$ by $\sqrt{2}$ and lower bound $n - |\mathsf{B}|$ by $n/2$, thus yielding, for $\widetilde{M} \equiv M(1+\mathcal{X})$:

$$\int_0^\epsilon \mathbb{E}_G\mathbb{E}||\mu_i^\theta - \mu_i^{\theta,t}||_{\rm TV}\, {\rm d}\theta \quad\le\quad \sqrt{4H(X_1)\epsilon/n}\,\mathbb{E}\{\widetilde{M}^{|\mathsf{B}|}\} + \mathbb{P}\{|\mathsf{B}| > \sqrt{n}/10\}\,. \tag{5.30}$$

The thesis follows by applying Proposition 4.2 to both terms. □

6 Proof of Theorem 2.4 (density evolution)

Given a probability distribution S over $\mathsf{M}^2(\mathcal{X})$, we define the probability distribution $\mathsf{P}_{S,k}$ over $\mathsf{M}(\mathcal{X}) \times \cdots \times \mathsf{M}(\mathcal{X})$ (k times) by

$$\mathsf{P}_{S,k}\{(\mu_1,\dots,\mu_k) \in A\} = \int \rho^k(A)\, S({\rm d}\rho)\,. \tag{6.1}$$

where ρ^k is the law of k iid random μ's with common distribution ρ. We shall denote by $\mathsf{E}_{S,k}$ expectation with respect to the same measure.

Lemma 6.1. *Let $G = (V, F, E)$ be any bipartite graph whose distribution is invariant under permutation of the variable nodes in $V = [n]$, and $\mu_i(\,\cdot\,) \equiv \mathbb{P}\{X_i = \cdot|Y, Z\}$ the marginals of an observations system on G. Then, for any diverging sequence $R_0 \subseteq \mathbb{N}$ there exists a subsequence R and a distribution S on $\mathsf{M}^2(\mathcal{X})$ such that, for any subset $\{i(1), \dots, i(k)\} \subseteq [n]$ of distinct variable nodes, and any bounded Lipschitz function $\varphi : \mathsf{M}(\mathcal{X})^k \to \mathbb{R}$:*

$$\lim_{n\in R} \mathbb{E}_G\mathbb{E}\{\varphi(\mu_{i(1)}, \dots, \mu_{i(k)})\} = \mathsf{E}_{S,k}\{\varphi(\mu_1, \dots, \mu_k)\}\,. \tag{6.2}$$

Proof. We shall assume, without loss of generality, that $R_0 = \mathbb{N}$. Notice that $(\mu_1, \dots, \mu_n)$ is a family of exchangeable random variables. By tightness, for each $i = 1, 2, \dots$, there exists a subsequence R_i such that $(\mu_1, \dots, \mu_i)$ converges in distribution, and $R_{i+1} \subseteq R_i$. Construct the subsequence R whose j-th element is the j-th element of R_j. Then for any k, $(\mu_{i(1)}, \dots, \mu_{i(k)})$ converges in distribution along R to an exchangeable set $(\mu_1^{(k)}, \dots, \mu_k^{(k)})$. Further the projection of the law of $(\mu_1^{(k)}, \dots, \mu_k^{(k)})$ on the first $k-1$ variables is the law of $(\mu_1^{(k-1)}, \dots, \mu_{k-1}^{(k-1)})$. Therefore, this defines an exchangeable distribution over the infinite collection of random variables $\{\mu_i : i = 1, 2, \dots\}$. By de Finetti, Hewitt-Savage Theorem [dF69, HS55] there exists S such that, for any k, the joint distribution of $(\mu_1, \dots, \mu_k)$ is $\mathsf{P}_{S,k}$. In particular

$$\lim_{n\in R} \mathbb{E}_G\mathbb{E}\{\varphi(\mu_{i(1)}, \dots, \mu_{i(k)})\} = \mathbb{E}\{\varphi(\mu_1, \dots, \mu_k)\} = \mathsf{E}_{S,k}\{\varphi(\mu_1, \dots, \mu_k)\}\,.$$

□

Proof. [Main Theorem] By Lemma 6.1, Eq. (2.18) holds for some probability distribution S_θ on $\mathsf{M}^2(\mathcal{X})$. It remains to prove that S_θ is supported over the fixed points of the density evolution equation (2.12).

Let $\varphi : \mathsf{M}(\mathcal{X}) \to \mathbb{R}$ be a test function that we can assume, without loss of generality bounded by 1, and with Lipschitz constant 1. Further, let $\mathsf{D}(i) \equiv \mathsf{D}(i, 1)$ and $\mu_{\mathsf{D}(i)}^{\theta(i)} \equiv \{\mu_j^{\theta(i)};\ j \in \mathsf{D}(i)\}$. By Theorem 2.3, together with the Lipschitz property and boundedness, we have

$$\int_0^\epsilon \mathbb{E}_G\mathbb{E}\,\left\{\Big[\varphi\big(\mu_i^\theta\big) - \varphi\big(\mathsf{F}_i^n\big(\mu_{\mathsf{D}(i)}^{\theta,(i)}\big)\big)\Big]^2\right\}\,\mathrm{d}\theta \le \frac{\mathsf{A}'}{\sqrt{n}}\,. \tag{6.3}$$

Fix now two variable nodes, say $i = 1$ and $i = 2$. Using Cauchy-Schwarz, this implies

$$\int_0^\epsilon \Big|\mathbb{E}_G\mathbb{E}\,\left\{\Big[\varphi\big(\mu_1^\theta\big) - \varphi\big(\mathsf{F}_1^n\big(\mu_{\mathsf{D}(1)}^{\theta,(1)}\big)\big)\Big]\Big[\varphi\big(\mu_2^\theta\big) - \varphi\big(\mathsf{F}_2^n\big(\mu_{\mathsf{D}(2)}^{\theta,(2)}\big)\big)\Big]\right\}\Big|\,\mathrm{d}\theta \le \frac{\mathsf{A}'}{\sqrt{n}}\,.$$

Applying dominated convergence theorem, it follows that, for almost all $\theta \in [0, \epsilon]$,

$$\lim_{n\to\infty} \mathbb{E}_G\mathbb{E}\,\left\{\Big[\varphi\big(\mu_1^\theta\big) - \varphi\big(\mathsf{F}_1^n\big(\mu_{\mathsf{D}(1)}^{\theta,(1)}\big)\big)\Big]\Big[\varphi\big(\mu_2^\theta\big) - \varphi\big(\mathsf{F}_2^n\big(\mu_{\mathsf{D}(2)}^{\theta,(2)}\big)\big)\Big]\right\} = 0\,. \tag{6.4}$$

By Lemma 6.1, we can find a sequence R_θ, and a distribution S_θ over $\mathsf{M}(\mathcal{X})^2$, such that Eq. (6.2) holds. We claim that along such a sequence

$$\lim_{n\in R} \mathbb{E}_G\mathbb{E}\,\{\varphi\big(\mu_1^\theta\big)\varphi\big(\mu_1^\theta\big)\} = \mathsf{E}_{S_\theta,2}\{\varphi(\mu_1)\varphi(\mu_2)\}\,, \tag{6.5}$$

$$\lim_{n\in R} \mathbb{E}_G\mathbb{E}\,\{\varphi\big(\mu_1^\theta\big)\varphi\big(\mathsf{F}_2^n\big(\mu_{\mathsf{D}(2)}^{\theta,(2)}\big)\big)\} = \mathbb{E}\,\mathsf{E}_{S_\theta,k+1}\{\varphi\big(\mu_1\big)\varphi\big(\mathsf{F}^\infty\big(\mu_2, \cdots, \mu_{k+1}\big)\big)\}\,, \tag{6.6}$$

$$\lim_{n\in R} \mathbb{E}_G\mathbb{E}\,\left\{\varphi\big(\mathsf{F}_1^n\big(\mu_{\mathsf{D}(1)}^{\theta,(1)}\big)\big)\varphi\big(\mathsf{F}_2^n\big(\mu_{\mathsf{D}(2)}^{\theta,(2)}\big)\big)\right\} = \mathbb{E}\,\mathsf{E}_{S_\theta,k_1+k_2}\{\varphi\big(\mathsf{F}_1^\infty\big(\mu_1, \cdots, \mu_{k_1}\big)\big)\varphi\big(\mathsf{F}_2^\infty\big(\mu_{k_1+1}, \cdots, \mu_{k_1+k_2}\big)\big)\}\,. \tag{6.7}$$

Here the expectations on the right hand sides are with respect to marginals $\mu_1, \mu_2, \dots$ distributed according to $\mathsf{P}_{S_\theta,\cdot}$ (this expectation is denoted as $\mathsf{E}_{S_\theta,\cdot}$) as well as with respect to independent random mappings $\mathsf{F}^\infty : \mathsf{M}(\mathcal{X})^* \to \mathsf{M}(\mathcal{X})$ defined as in Section 2.2, cf. Eq. (2.12) (this includes expectation with respect to k, k_1, k_2 and is denoted as $\mathbb{E}$).

Before proving the above limits, let us show that they implies the thesis. Substituting Eqs. (6.5) to (6.7) in Eq. (6.4) and re-ordering the terms we get

$$\int \Delta(\rho)^2 \; S_\theta(\mathrm{d}\rho) = 0\,, \tag{6.8}$$

$$\Delta(\rho) \equiv \int \varphi(\mu)\,\rho(\mathrm{d}\mu) - \mathbb{E}\int \varphi(\mathsf{F}^\infty(\mu_1,\cdots,\mu_k))\;\rho(\mathrm{d}\mu_1)\cdots\rho(\mathrm{d}\mu_k)\,. \tag{6.9}$$

Therefore $\Delta(\rho) = 0$ S_θ-almost surely, which is what we needed to show in order to prove Theorem 2.4.

Let us now prove the limits above. Equation (6.5) follows is an immediate consequence of Lemma 6.1. Next consider Eq. (6.6), and condition the expectation on the left-hand side upon $\mathsf{B}(i=2,t=1)=\mathsf{B}$, as well as upon W_{B}, cf. Eq. (5.15). First notice that, by Lemma 5.4

$$\mathbb{E}_G\mathbb{E}\left\{||\mu_1^\theta - \mu_1^{\theta,(2)}||_{\rm TV}|\mathsf{B},W_{\mathsf{B}}\right\} \le \frac{\mathsf{C}}{\sqrt{n-|\mathsf{B}|}}\,. \tag{6.10}$$

and condition the expectation on the left-hand side upon $\mathsf{B}(i,1)=\mathsf{B}$, as well as upon W_{B}, cf. Eq. (5.15). As a consequence, by Lipschitz property and boundedness of φ

$$\left|\mathbb{E}_G\mathbb{E}\left\{\varphi(\mu_1^\theta)\varphi(\mathsf{F}_2^n(\mu^{\theta,(2)}_{\mathsf{D}(2)})|\mathsf{B},W_{\mathsf{B}}\right\} - \mathbb{E}_G\mathbb{E}\left\{\varphi(\mu_1^{\theta,(2)})\varphi(\mathsf{F}_2^n(\mu^{\theta,(2)}_{\mathsf{D}(2)})|\mathsf{B},W_{\mathsf{B}}\right\}\right| \le \frac{\mathsf{C}}{\sqrt{n-|\mathsf{B}|}}\,. \tag{6.11}$$

In the second term the $\mu_j^{\theta,(2)}$ are independent of the conditioning, of the function F_2^n (which is deterministic once B, W_{B} are given). Therefore, by Lemma 6.1 (here we are taking the limit on the joint distribution of the $\mu_j^{\theta(2)}$, but not on F_2^n; to emphasize this point we note the latter as $\mathsf{F}_2^{n_*}$)

$$\lim_{n\in R_\theta}\mathbb{E}_G\mathbb{E}\left\{\varphi(\mu_1^{\theta,(2)})\varphi(\mathsf{F}_2^{n_*}(\mu^{\theta,(2)}_{\mathsf{D}(2)})|\mathsf{B},W_{\mathsf{B}}\right\} = \mathsf{E}_{S,k}\{\varphi(\mu_1)\varphi(\mathsf{F}_2^{n_*}(\mu_2,\dots,\mu_{1+|\mathsf{D}(2)|}))\}\,. \tag{6.12}$$

(Notice that the graph whose expectation is considered on the left hand side is from the ensemble $\mathcal{G}(n-|V(\mathsf{B})|,\alpha n-|F(\mathsf{B})|,\gamma/n)$. The limit measure S_θ could *a priori* be different from the one for the ensemble $\mathcal{G}(n,\alpha n,\gamma/n)$. However, Lemmas 5.4, 5.5 imply that this cannot be the case.)

By using dominated convergence and Eq. (6.11) we get

$$\begin{aligned}\lim_{n\in R_\theta}\mathbb{E}_G\mathbb{E}\left\{\varphi(\mu_1^\theta)\varphi(\mathsf{F}_2^{n_*}(\mu^{\theta,(2)}_{\mathsf{D}(2)})\right\} &= \mathbb{E}_{\mathsf{B},W_{\mathsf{B}}}\left\{\lim_{n\in R_\theta}\mathbb{E}_G\mathbb{E}\left\{\varphi(\mu_1^\theta)\varphi(\mathsf{F}_2^{n_*}(\mu^{\theta,(2)}_{\mathsf{D}(2)})|\mathsf{B},W_{\mathsf{B}}\right\}\right\} = \\ &= \mathbb{E}_{\mathsf{B},W_{\mathsf{B}}}\mathsf{E}_{S,k}\left\{\varphi(\mu_1)\varphi(\mathsf{F}_2^{n_*}(\mu_2,\dots,\mu_{1+|\mathsf{D}(2)|}))\right\}\,.\end{aligned}$$

Finally we can take the limit $n_*\to\infty$ as well. By local convergence of the graph to the tree model, we have uniform convergence of $\mathsf{F}_2^{n_*}$ to F^∞ and thus Eq. (6.6).

The proof of Eq. (6.7) is completely analogous to the latter and is omitted to avoid redundancies. □

References

[BMTW84] Toby Berger, Nader Mehravari, Don Towsley, and Jack Wolf. Random multiple-access communication and group testing. *IEEE Trans. Communications*, 32:769–779, 1984.

[BMvT78] Elwyn Berlekamp, Robert J. McEliecee, and Henk C.A. van Tilborg. On the inherent intractability of certain coding problems. *IEEE Trans. Inform. Theory*, IT-29:384–386, 1978.

[CSV04] Giuseppe Caire, Shlomo Shamai, and Sergio Verdú. Noiseless data compression with low density parity check codes. In P. Gupta and G. Kramer, editors, *Dimacs Series in Mathematics and Theoretical Computer Science*, pages 224–235. AMS, 2004.

[dF69] Bruno de Finetti. Sulla prosequibilitá di processi aleatori scambiabili. *Rend. Ist. Mat. Trieste*, 1:53–67, 1969.

[Don06] David L. Donoho. Compressed Sensing. *IEEE Trans. on Inform. Theory*, 52:1289– 1306, 2006.

[Dor43] Robert Dorfman. The detection of defective members of large populations. *Ann. Math. Statist.*, 14:436–440, 1943.

[EJCT06] Justin K. Romberg Emmanuel J. Candes and Terence Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Inform. Theory*, 52:489– 509, 2006.

[EV03] Cristian Estan and George Varghese. New Directions in Traffic Measurement and Accounting: Focusing on the Elephants, Ignoring the Mice. *ACM Trans. on Comp. Syst.*, 21:270–313, 2003.

[Gal63] Robert G. Gallager. *Low-Density Parity-Check Codes*. MIT Press, Cambridge, Massachussetts, 1963.

[GM07] Antoine Gerschenfeld and Andrea Montanari. Reconstruction for models on random graphs. In *48nd Annual Symposium on Foundations of Computer Science*, Providence, October 2007.

[GT04] Francesco Guerra and Fabio L. Toninelli. The high temperature region of the Viana-Bray diluted spin glass model. *J. Stat. Phys.*, 115:531–555, 2004.

[GV05] Dongning Guo and Sergio Verdu. Randomly spread cdma: Asymptotics via statistical physics. *IEEE Trans. on Inform. Theory*, 51:1982–2010, 2005.

[GW06a] Dongning Guo and Chih-Chun Wang. Asymptotic Mean-Square Optimality of Belief Propagation for Sparse Linear Systems. In *IEEE Information Theory Workshop*, Chengdu, China, October 2006.

[GW06b] Dongning Guo and Chih-Chun Wang. Belief Propagation Is Asymptotically Equivalent to MAP Estimation for Sparse Linear Systems. In *44th Annual Allerton Conference on Communication, Control and Computing*, Monticello, IL, October 2006.

[HS55] Edwin Hewitt and Leonard J. Savage. Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, 80:470–501, 1955.

[KFL01] Frank R. Kschischang, Brendan J. Frey, and H.-Andrea Loeliger. Factor Graphs and the Sum-Product Algorithm. *IEEE Trans. on Inform. Theory*, 47(2):498–519, 2001.

[KKM07] Satish Babu Korada, Shrinivas Kudekar, and Nicolas Macris. Exact solution for the conditional entropy of Poissonian LDPC codes over the Binary Erasure Channel. In *Proc. of the IEEE Int. Symposium on Inform. Theory*, Nice, July 2007.

[KM06] Shrinivas Kudekar and Nicolas Macris. Sharp Bounds for MAP Decoding of General Irregular LDPC. In *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seattle, July 2006.

[LMP07] Yi Lu, Andrea Montanari, and Balaji Prabhakar. Detailed network measurements using sparse graph counters: The theory. In *46th Allerton Conf. on Communication, Control, and Computing*, Monticello, ILL, September 2007.

[Lub02] Michael Luby. LT codes. In *43rd Annual Symposium on Foundations of Computer Science*, Vancouver, November 2002.

[Mac07] Nicolas Macris. Griffith-Kelly-Sherman Correlation Inequalities: A Useful Tool in the Theory of Error Correcting Codes. *IEEE Trans. Inform. Theory*, 53:664–683, 2007.

[MMRU05] Cyril Méasson, Andrea Montanari, Tom Richardson, and Rüdiger Urbanke. The Generalized Area Theorem and Some of its Consequences. submitted to IEEE IT, 2005.

[MMU05] Cyril Méasson, Andrea Montanari, and Rüdiger Urbanke. Maxwell Construction: The Hidden Bridge between Iterative and Maximum a Posteriori Decoding. IEEE Trans. Inform. Theory, to appear, 2005.

[Mon05] Andrea Montanari. Tight bounds for LDPC and LDGM codes under MAP decoding. *IEEE Trans. Inform. Theory*, 51:3221–3246, 2005.

[MT06] Andrea Montanari and David Tse. Analysis of Belief Propagation for Non-Linear Problems: The Example of CDMA (or: How to Prove Tanaka's Formula). In *IEEE Information Theory Workshop*, Punta del Este, Uruguay, March 2006.

[MT07] Marc Mézard and Cristina Toninelli. Group testing with random pools: optimal two-stage algorithms. 2007. `arXiv:0706.3104`.

[Mur01] Tatsuto Murayama. Statistical mechanics of linear compression codes in network communication. `arXiv:cond-mat/0106209`, 2001.

[Mur04] Tatsuto Murayama. Near rate-distortion bound performance of sparse matrix codes. In *Proc. of the IEEE Int. Symposium on Inform. Theory*, page 299, Chicago, July 2004.

[RS07] Jack Raymond and David Saad. Sparsely-Spread CDMA-a Statistical Mechanics Based Analysis. 2007. `arXiv: 0704.0098`.

[RU07] Tom Richardson and Rüdiger Urbanke. *Modern Coding Theory*. Cambridge University Press, 2007. In preparation, available online at `http://lthcwww.epfl.ch/mct/index.php`.

[Tal06] Michel Talagrand. The Parisi formula. *Ann. of Math.*, 163:221–263, 2006.

[Tan02] Toshyuki Tanaka. A statistical-mechanics approach to large-system analysis of cdma multiuser detectors. *IEEE Trans. on Inform. Theory*, 48:2888–2910, 2002.

[TH99] David N.C. Tse and Stephen V. Hanly. Linear multiuser receivers: effective interference, effective bandwidth and user capacity. *IEEE Trans. on Inform. Theory*, 45:641–657, 1999.

[TV05] David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, Cambridge, UK, 2005.

[Ver89] Sergio Verdú. Computational complexity of optimum multiuser detection. *Algorithmica*, 4:303–312, 1989.

[Ver98] Sergio Verdu. *Multiuser Detection*. Cambridge University Press, Cambridge, UK, 1998.

[VS99] Sergio Verdu and Shlomo Shamai. Spectral efficiency of CDMA with random spreading. *IEEE Trans. on Inform. Theory*, 45:622–640, 1999.

[Wol85] Jack K. Wolf. Born again group testing: Multiaccess communications. *IEEE Trans. Inform. Theory*, 31:185–191, 1985.

[YT06] Mika Yoshida and Toshiyuki Tanaka. Analysis of Sparsely-Spread CDMA via Statistical Mechanics. In *Proc. of the IEEE Int. Symposium on Inform. Theory*, Seattle, July 2006.