Conditional Random Fields, Planted Constraint Satisfaction, and Entropy Concentration[*]

Emmanuel Abbe[†] and Andrea Montanari[‡]

Abstract

This paper studies a class of probabilistic models on graphs, where edge variables depend on incident node variables through a fixed probability kernel. The class includes planted constraint satisfaction problems (CSPs), as well as other structures motivated by coding theory and community detection problems. It is shown that under mild assumptions on the kernel and for sparse random graphs, the conditional entropy of the node variables given the edge variables concentrates around a deterministic threshold. This implies in particular the concentration of the number of solutions in a broad class of planted CSPs, the existence of a threshold function for the disassortative stochastic block model, and the proof of a conjecture on parity check codes. It also establishes new connections among coding, clustering and satisfiability.

[*] This paper was presented at RANDOM 2013 [2]

[†] Department of Electrical Engineering and Program in Applied and Computational Mathematics, Princeton University, Email: eabbe@princeton.edu

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

1 Introduction

This paper studies a class of probabilistic models on graphs encompassing models from random combinatorial optimization, coding theory and learning theory. Depending on the context, the class may be described as a family of planted constrained satisfaction problems (CSPs), encoded channels or conditional random fields. We start by providing motivations in CSPs.

CSPs are key components in the theory of computational complexity as well as important mathematical models in various applications of computer science, engineering and physics. In a CSP, a set of variables $x_1, \dots, x_n$ is required to satisfy a collection of constraints involving each a subset of the variables. In many cases of interest, the variables are Boolean and the constraints are all of a common type: e.g., in k-SAT, the constraints require the OR of k Boolean variables or their negations to be TRUE, whereas in k-XORSAT, the XOR of the variables or their negations must equal to zero. Given a set of constraints and a number of variables, the problem is to decide whether there exists a satisfying assignment. In random CSPs, the constraints are drawn at random from a given ensemble, keeping the constraint density[1] constant. In this setting, it is of interest to estimate the *probability* that a random instance is satisfiable. One of the fascinating phenomena occurring for random instances is the phase transition, which makes the task of estimating this probability much easier in the limit of large n. For a large class of CSPs, and as n tends to infinity, the probability of being satisfiable tends to a step function, jumping from 1 to 0 when the constraint density crosses a critical threshold. For random k-XORSAT the existence of such a critical threshold is proved [15, 18, 17, 44]. For random k-SAT, $k \geq 3$, the existence of a n-dependent threshold is proved in [25]. However it remains open to show that this threshold converges when n tends to infinity. Upper and lower bounds are known to match up to a term that is of relative order $k\,2^{-k}$ as k increases [7]. Phase transition phenomena in other types of CSPs are also investigated in [6, 40, 7]

In planted random CSPs, a "planted assignment" is first drawn, and the constraints are then drawn at random so as to keep that planted assignment a satisfying one. Planted ensembles were investigated in [9, 30, 5, 4, 31, 3], and at high density in [8, 14, 21]. In the planted setting, the probability of being SAT is always equal to one by construction, and a more relevant question is to determine the actual *number* of satisfying assignments. One would expect that this problem becomes easier in the limit of large n due to an asymptotic phenomenon. This paper shows that, indeed, a concentration phenomenon occurs: for a large class of planted CSPs (including SAT, NAE-SAT and XOR-SAT) the normalized logarithm of the number or satisfying assignment concentrates (with respect to the graph of the CSP) around a deterministic number. Moreover, this deterministic threshold is n-independent.

It is worth comparing the result obtained in this paper for planted CSPs, with the one obtained in [1] for non planted CSPs. In that case, the number of solution is zero with positive probability and therefore the logarithm of the number of solution does not have a finite expectation. Technically, standard martingale methods do not allow to prove concentration, even around an n-dependent threshold. In [1] an interpolation method [29] is used to prove the existence of the limit of a 'regularized' quantity, namely the logarithm of the number of solutions plus one, divided by the number of variables. A technical consequence of this approach is that the concentration of this quantity around a value that is independent of n can only be proved when the UNSAT probability is known to be $O(1/\log(n)^{1+\varepsilon})$.

[1] The ratio of the expected number of constraints per variables.

This paper shows that –in the planted case– the concentration around an n-independent value holds unconditionally. We use again the interpolation technique [29, 23, 24, 43, 10, 1] but with an interesting twist. While in all the cited references, the entropy (or log-partition function) is shown to be superadditive, in the present setting it turns out to be subadditive.

Let us also mention that a fruitful line of work has addressed the relation between planted random CSPs and their non planted counterparts in the satisfiable phase [3, 33, 47]. These papers show that, when the number of solutions is sufficiently concentrated, planting does not play a critical role in the model. It would be interesting to use these ideas to 'export' the concentration result obtained here to non planted models.

In this paper, we pursue a different type of approach. Motivated by applications[2], in particular in coding theory and community clustering, we consider extensions of the standard planted CSPs to a setting allowing soft probabilistic constraints. Within the setting of soft CSPs, the planted solution is an unknown vector to be reconstructed, and the constraints are regarded as noisy observations of this unknown vector. For instance one can recover the case of planted random k-SAT as follows. Each clause is generated by selecting first k variable indices $i_1, \dots, i_k$ uniformly at random, providing a random hyperedge. Then a clause is drawn uniformly among the ones that are satisfied by the variables $x_{i_1}, \dots, x_{i_k}$ appearing in the planted assignment. The clause can hence be regarded as a noisy observation of $x_{i_1}, \dots, x_{i_k}$. More generally the formula can be seen as a noisy observation of the planted assignment.

Our framework extends the above to include numerous examples from coding theory, learning theory and statistics. Within LDPC or LDGM codes [46], encoding is performed by evaluating the modulo 2 sum of a random subset of information bits and transmitting it through a noisy communication channel. The selection of the information bits is described by a graphs, drawn at random for the code construction, and the transmission of these bits leads to a noisy observation of the graph variables. Similarly, a community clustering block model [27] can be seen as a random graph model, whereby each edge is a noisy observation of the community assignments of the adjacent nodes. Definitions will be made precise in the next section.

The conditional probability of the unknown vector given the noisy observations takes the form of a graphical model, i.e. factorizes according to an hypergraph whose nodes correspond to variables and hyperedges correspond to noisy observations. Such graphical models have been studied by several authors in machine learning [35] under the name of 'conditional random fields', and in [39] in the context of LDPC and LDGM codes. The conditional entropy of the unknown vector given the observations is used here to quantify the residual uncertainty of the vector. This is equivalent to considering the mutual information between the node and edge variables. In such a general setting, we prove that the conditional entropy per variable concentrates around a well defined deterministic limit. This framework allows a unified treatment of a large class of random combinatorial optimization problems, raises new connections among them, and opens up to new models. We obtain in particular a proof of a conjecture posed in [45] on low-density parity-check codes, and the existence of a threshold function for the disassortative stochastic block model [16].

[2] Planted models are also appealing to cryptographic application, as hard instances with known solutions provide good one-way functions [28, 12].

2 The model

Let k and n be two positive integers with $n \geq k$.

- Let $V = [n]$ and $g = (V, E(g))$ be a hypergraph with vertex set V and edge set $E(g) \subseteq E_k(V)$, where $E_k(V)$ denotes the set of all possible $\binom{n}{k}$ hyperedges of order k on the vertex set V. We will often drop the prefix "hyper".

- Let $\mathcal{X}$ and $\mathcal{Y}$ be two finite sets called respectively the input and output alphabets. Let $Q(\cdot|\cdot)$ be a probability transition function (or channel) from $\mathcal{X}^k$ to $\mathcal{Y}$, i.e., for each $u \in \mathcal{X}^k$, $Q(\cdot|u)$ is a probability distribution on $\mathcal{Y}$.

- To each vertex in V, we assign a node-variable in $\mathcal{X}$, and to each edge in $E(g)$, we assign an edge-variable in $\mathcal{Y}$. We define

$$P_{g,Q}(y|x) \equiv \prod_{I \in E(g)} Q(y_I | x[I]), \quad x \in \mathcal{X}^V, y \in \mathcal{Y}^{E(g)}, \tag{1}$$

  where y_I denotes the edge-variable attached to edge I, and $x[I]$ denotes the k node-variables attached to the vertices adjacent to edge I (where the entries of $x[I]$ are placed in the increasing order of I, although the order will not matter for the applications considered in this paper). This defines for a given hypergraph g the probability of the edge-variables given the node-variables.

The above is a type of factor or graphical model, or a planted constraint satisfaction problem with soft probabilistic constraints. For each $x \in \mathcal{X}^V$, $P_{g,Q}(\cdot|x)$ is a product measure on the set of edge-variables. We call $P_{g,Q}$ a **graphical channel with graph g and kernel Q**. We next put the uniform probability distribution on the set of node-variables $\mathcal{X}^V$, and define the a posteriori probability distribution (or reverse channel) by

$$R_{g,Q}(x|y) \equiv \frac{1}{S_{g,Q}(y)} P_{g,Q}(y|x)|\mathcal{X}|^{-n}, \quad x \in \mathcal{X}^V, y \in \mathcal{Y}^{E(g)}, \tag{2}$$

where

$$S_{g,Q}(y) \equiv \sum_{x \in \mathcal{X}^V} P_{g,Q}(y|x)|\mathcal{X}|^{-n} \tag{3}$$

is the marginal distribution of the edge variables.

We now define two probability distributions on the hypergraph g, which are equivalent for the purpose of this paper:

- A sparse Erdös-Rényi distribution, where each edge is drawn independently with probability $p = \frac{\alpha n}{\binom{n}{k}}$, where $\alpha > 0$ is the edge density.

- A sparse Poisson distribution, where for each $I \in E_k(V)$, a number of edges m_I is drawn independently from a Poisson distribution of parameter $p = \frac{\alpha n}{\binom{n}{k}}$. Note that m_I takes value in $\mathbb{Z}_+$, hence G is now a multi-edge hypergraph. To cope with this more general setting, we allow the edge-variable y_I to take value in $\mathcal{Y}^{m_I}$, i.e., $y_I = (y_I(1), \dots, y_I(m_I))$, and define (with a slight abuse of notation)

$$Q(y_I|x[I]) = \prod_{i=1}^{m_I} Q(y_I(i)|x[I]). \tag{4}$$

This means that for each I, if $m_I \geq 1$, m_I i.i.d. outputs are drawn from the kernel Q, and if $m_I = 0$, no edge is drawn. We denote by $\mathcal{P}_k(\alpha, n)$ this distribution on (multi-edge) hypergraphs.

Since $p = \frac{\alpha n}{\binom{n}{k}}$, the number of edges concentrates around its expectation given by αn and the two models can be shown to be equivalent –for our purposes– in the limit of large n.

3 Main Results

We now define the conditional entropy between the node and edge variables. This is equivalent, up to a constant shift, to the mutual information between the node and edge variables.

Definition 3.1. *Let X be uniformly drawn in $\mathcal{X}^n$, G be a random sparse hypergraph drawn from the $\mathcal{P}_k(\alpha, n)$ ensemble independently of X, and Y be the output of X through the graphical channel $P_{G,Q}$ defined in* (1) *for a kernel Q. We define*

$$H_G(X|Y) \equiv -|\mathcal{X}|^{-n} \sum_{x \in \mathcal{X}^V} \sum_{y \in \mathcal{Y}^{E(G)}} P_{G,Q}(y|x) \log R_{G,Q}(x|y), \tag{5}$$

$$H(X|Y,G) \equiv \mathbb{E}_G H_G(X|Y), \tag{6}$$

where $P_{G,Q}$ and $R_{G,Q}$ are defined in (1) *and* (2) *respectively. Note that $H_G(X|Y)$ is a random variable since G is random, and for a realization $G = g$, $H_g(X|Y)$ is the conditional entropy of X given Y and given $G = g$, which can also be expressed as $H_g(X|Y) = H(X|Y, G = g)$. The notation $H(X|Y,G) = \mathbb{E}_G H_G(X|Y)$ is consistent with the usual notation used for the conditional entropy, where the expectation is over the variables in the conditioning. Note that the mutual information between the node and edge variables is obtained as $I_g(X|Y) = n \log |\mathcal{X}| - H_g(X|Y)$ and the expected value $I(X|Y,G) = n \log |\mathcal{X}| - H(X|Y,G)$. We will work with the conditional entropy is this paper.*

Definition 3.2. *We denote by $M_1(\mathcal{X}^l)$ the set of probability measures on $\mathcal{X}^l$. For a kernel Q from $\mathcal{X}^k$ to $\mathcal{Y}$, we define*

$$\Gamma_l : M_1(\mathcal{X}^l) \to \mathbb{R} \tag{7}$$

$$\nu \mapsto \Gamma_l(\nu) = \frac{1}{|\mathcal{Y}|} \sum_{u^{(1)}, \dots, u^{(l)} \in \mathcal{X}^k} \left[\sum_{y \in \mathcal{Y}} \prod_{r=1}^{l} (1 - Q(y|u^{(r)})) \right] \prod_{i=1}^{k} \nu(u_i^{(1)}, \dots, u_i^{(l)}). \tag{8}$$

Hypothesis H. A kernel Q is said to satisfy hypothesis H if Γ_l is convex for any $l \geq 1$.

Despite the lengthy expression, it is important to note that the definition of Γ_l depends solely on the kernel Q. We will see in Section 4 that a large variety of kernels satisfy this hypothesis, including kernels corresponding to parity-check encoded channels, planted SAT, NAE-SAT, XORSAT, and disassortative stochastic block models.

We first show a sub-additivity property for the expected conditional entropy of graphical channels.

Theorem 3.3. *Let Q be a kernel satisfying hypothesis H. Let $n_1, n_2 \geq k$ and $n = n_1 + n_2$. We denote by G_n a hypergraph drawn from the ensemble $\mathcal{P}_k(\alpha, n)$ and G_{n_i} two hypergraphs drawn independently from the ensembles $\mathcal{P}_k(\alpha, n_i)$, respectively for $i = 1, 2$. Define $f(n) = H(X|Y, G_n)$, and $f(n_i) = H(X|Y, G_{n_i})$, $i = 1, 2$. Then*

$$f(n) \leq f(n_1) + f(n_2). \tag{9}$$

The proof of this theorem is outlined in Section 5.

Corollary 3.4. *Let Q be a kernel satisfying hypothesis H and G_n be a random hypergraph drawn from the ensemble $\mathcal{P}_k(\alpha, n)$. There exists a constant $C_k(\alpha, Q)$ such that*

$$\frac{1}{n} H(X|Y, G_n) \to C_k(\alpha, Q), \quad \textit{as } n \to \infty. \tag{10}$$

The following is obtained using previous corollary and a martingale concentration argument.

Theorem 3.5. *Let Q be a kernel satisfying hypothesis H and G_n be a random hypergraph drawn from the ensemble $\mathcal{P}_k(\alpha, n)$, then, almost surely,*

$$\lim_{n \to \infty} \frac{1}{n} H_{G_n}(X|Y) = C_k(\alpha, Q), \tag{11}$$

with $C_k(\alpha, Q)$ as in Corollary 3.4.

Note that the almost sure convergence is with respect to the graph G_n. The proof is given in Section A.

4 Applications

We next present three applications of the general model described in the previous section. While planted CSPs and parity-check codes are directly derived as particular cases of our model, the stochastic block model is obtained with a limiting argument. Note that the general model described in previous section allows to also generate new hybrid structures. For example, one may consider codes which are not linear but which rely on OR gates as in SAT, community structures whose connectivity rely on collections of k nodes, or network models which have erasures.

4.1 Planted constraint satisfaction problems

Definition 4.1. *A CSP kernel is given by*

$$Q(y|u) = \frac{1}{|A(u)|} \mathbb{1}(y \in A(u)), \quad u \in \mathcal{X}^k, y \in \mathcal{Y}, \tag{12}$$

where $A(u)$ is a subset of $\mathcal{Y}$ containing the "authorized constraints", with the property that $|A(u)|$ is constant (it may depend on k but not on u).

We will next show that a graphical channel with a CSP kernel corresponds to a planted CSP. We derive first a few known examples of CSPs.

- For planted k-SAT, $\mathcal{Y} = \{0,1\}^k$ and $A(u) = \{0,1\}^k \setminus \bar{u}$, where $\bar{u}$ is the vector obtained by flipping each component in u. Using this kernel implies that for any selected edge $I \in E_k(V)$, the edge variable y_I is a vector in $\{0,1\}^k \setminus \bar{x}[I]$ uniformly drawn, representing the negation pattern of the constraint I. Note that using u rather than $\bar{u}$ leads to an equivalent probabilistic model, $\bar{u}$ is simply used here to represent u as a "satisfying assignment". Note that $|A(u)| = 2^k - 1$.
- For planted k-NAE-SAT, $\mathcal{Y} = \{0,1\}^k$ and $A(u) = \{0,1\}^k \setminus \{u, \bar{u}\}$, with $|A(u)| = 2^k - 2$.
- For k-XOR-SAT, $\mathcal{Y} = \{0,1\}$ and $A(u) = \oplus_{i=1}^k u_i$ and $|A(u)| = 1$.

In general, a graphical channel with graph g and kernel Q as in (12) leads to a planted CSP where the constraints are given by $A(x[I]) \ni y_I$ for any $I \in E(g)$. For example, for planted k-SAT, the constraints are equivalent to $\bar{x}[I] \neq y_I$, whereas for planted k-NAE-SAT, the constraints are equivalent to $\bar{x}[I] \notin (y_I, \bar{y}_I)$. If y is drawn from the output marginal distribution $S_{g,Q}$ (cf. (3)), then there exists a satisfying assignment by construction.

Lemma 4.2. *For a graphical channel with graph g and CSP kernel Q as in* (12)*, and for y in the support of $S_{g,Q}$,*

$$H_g(X|Y=y) = \log Z_g(y) \tag{13}$$

where $Z_g(y)$ is the number of satisfying assignments of the planted CSP with graph g and constraints specified by y (where the structure of constraints us specified by Q).

Corollary 4.3. *For a graphical channel with CSP kernel Q as in* (12)*, and for a graph G drawn from the ensemble $\mathcal{P}(\alpha, n)$,*

$$H(X|Y,G) = \mathbb{E}_{G,Y} \log Z_G(Y), \tag{14}$$

where $Z_G(Y)$ is the number of satisfying assignments of the corresponding random planted CSP.

Remark 4.4. *This result gives the convergence of the normalized expected logarithm of the number of solutions for any edge density α, in contrast with [1], which obtains the convergence of the same quantity (where the number of solutions is shifted by 1 to avoid taking the logarithm of 0) only for a regime of α small enough, essentially where the probability of being UNSAT decays faster than* $1/\log(n)$*. One should note that an unconditional result of the kind of Corollary 4.3 for the non-planted setting would imply the satisfiability conjecture (the existence of an n-independent threshold for k-SAT).*

Lemma 4.5. *For any $k \geq 1$, and for the CSP kernel corresponding to planted k-SAT, the operator Γ_l is convex for any $l \geq 1$.*

Lemma 4.6. *For any $k \geq 1$, and for the CSP kernel corresponding to planted k-NAE-SAT, the operator Γ_l is convex for any $l \geq 1$.*

Lemma 4.7. *For any k even, and for the CSP kernel corresponding to planted k-XOR-SAT, the operator Γ_l is convex for any $l \geq 1$.*

Using Theorem 3.5 and previous lemmas, the following is obtained.

Corollary 4.8. *Let $Z(F_n)$ denote the number of solutions of a random planted formula $F_n = (G_n, Y)$ with graph G_n and k-SAT, k-NAE-SAT, or k-XOR-SAT (k even) kernel as in definition 4.1. Then $\frac{1}{n}\mathbb{E}_Y \log Z(F_n)$ converges almost surely.*

Note that the almost sure convergence is over G_n, and the limit depends only on k and α.

4.2 Stochastic block model

The problem of community detection is to divide a set of vertices in a network (graph) into groups having a higher connectivity within the groups and lower connectivity across the groups (assortative case), or the other way around (disassortative case). This is a fundamental problem in many modern statistics, machine learning, and data mining problems with a broad range of applications in population genetics, image processing, biology and social science. A large variety of models have been proposed for community detection problems, we refer to [42, 22, 27] for a survey on the subject.

At an algorithmic level, the problem of finding the smallest cut in a graph with two equally sized groups, i.e., the min-bisection problem, is well-known to be NP-hard [19]. Concerning average-case complexity, various random graphs models have been proposed for community clustering. The Erdös-Rényi random graph is typically a very bad model for community structures, since each node is equally likely connected to any other nodes and no communities are typically formed. The stochastic block model is a natural extension of an Erdös-Rényi model with a community structure. Although the model is fairly simple (communities emerge but the average degree is still constant[3]), it is a fascinating model with several fundamental questions still open.

We now describe the stochastic block model (SBM), also called planted bisection model, with two groups and symmetric parameters. Let $V = [n]$ be the vertex set and a, b be two positive real numbers. For a uniformly drawn assignment $X \in \{0,1\}^V$ on the vertices, an edge is drawn between vertex i and j with probability a/n if $X_i = X_j$ and with probability b/n if $X_i \neq X_j$, and each edge is drawn independently. We denote this model by $\mathcal{G}(n, a, b)$. Note that the average degree of an edge is $(a+b)/2$, however, a 0-labelled node is connected in expectation with $a/2$ 0-labeled nodes and with $b/2$ 1-labeled nodes.

This type of model was introduced in [19], in the dense regime. The attention to the sparse regime described above is more recent, with [13] and [32, 16, 41]. In particular, [16] conjectured a phase transition phenomenon, with the detection of clusters[4] being possible if $(a-b)^2 > 2(a+b)$ and impossible otherwise. In [41], a remarkable proof of the impossibility part is obtained, leaving the achievability part open.

There are at least two ways to define the SBM as a graphical channel. The direct way is simply to consider G to be the complete graphs, and for each pair of vertices, to use the kernel $Q(y_{ij}|x_i, x_j) = P(y_{ij}|x_i \oplus x_j)$, where $P(1|0) = a/n$ and $P(1|1) = b/n$. With this approach, however,

[3] Models with corrected degrees have been proposed in [32].

[4] Obtaining a reconstruction positively correlated with the true assignment.

the channel Q depends on the number of vertices n, whereas the edge probability of the graph is constant. We next show how we can "move" the sparsity from the kernel to the graph, obtaining a graphical channel with a sparse graph as in previous section and a fixed (asymmetric) channel. The obtained model will approximate accurately the original model as next shown. Note that this creates a strong connection between coding and clustering, since it expresses the later problem as a particular type of code (a simple degree 2 LDGM code) on a binary input/output channel which is very noisy.

Definition 4.9. *An SBM kernel is given by*

$$Q(z|u_1,u_2) = \begin{cases} a/\gamma & \text{if } u_1 = u_2, \\ b/\gamma & \text{if } u_1 \neq u_2, \end{cases} \tag{15}$$

where $u_1, u_2, z \in \{0,1\}$.

Lemma 4.10. *There exists* $n_0 = n_0(\gamma, a, b)$ *and* $C = C(a,b)$ *such that the following holds true. Let* X *be uniformly drawn on* $\{0,1\}^V$, Y *be the output (the graph) of a sparse stochastic block model of parameters* a, b, *and* Y_γ *be the output of a graphical channel with graph* G_γ *drawn from the ensemble* $\mathcal{P}(\gamma, n)$ *and kernel* (15), *then, for all* $n \geq n_0$

$$\big|H(X|Y) - H(X|G_\gamma, Y_\gamma)\big| \leq \frac{Cn}{\gamma}. \tag{16}$$

Lemma 4.11. *For the SBM kernel given by* (15), $a \leq b$ *(disassortative case) and* γ *large enough, the operator* Γ_l *is convex for any* $l \geq 1$.

Corollary 4.12. *For the disassortative SBM, the limit of* $H(X|Y)/n$ *exists and satisfies*

$$\lim_{n\to\infty} \frac{1}{n} H(X|Y) = \lim_{\gamma\to\infty} \lim_{n\to\infty} \frac{1}{n} H(X|G_\gamma, Y_\gamma)\,. \tag{17}$$

In a work in progress, the assortative case is investigated with a different proof technique. The dependence of the above limit on a, b is also expected to reflect the phase transition of the SBM [16, 41].

4.3 Parity-check encoded channels

Shannon's coding theorem states that for a discrete memoryless channel W from $\mathcal{X}$ to $\mathcal{Y}$, the largest rate at which reliable communication can take place is given by the capacity $I(W) = \max_X I(X;Y)$, where $I(X;Y)$ is the mutual information of the channel W with a random input X. To show that rates up to capacity are achievable, Shannon used random codebooks, relying on a probabilistic argument. Shortly after, Elias [20] showed that random linear codes allow to achieve capacity, reducing the encoding complexity from exponential to quadratic in the code dimension. However, Berlekamp, McEliece, and Van Tilborg showed in [11] that the maximum likelihood decoding of unstructured linear codes is NP-hard.

In order to reduce the complexity of the decoder, Gallager proposed the use sparse linear codes [26], giving birth to the LDPC codes, with sparse parity-check matrices, and LDGM codes, with sparse generator matrices. Various types of LDPC/LDGM codes depend on various types of row and column degree distributions. Perhaps one of the most basic class of such codes is the LDGM code with constant right degree, which corresponds to a generator matrix with column having a fixed number k of ones. This means that each codeword is the XOR of k uniformly selected information bits. In other words, this is a graph based code drawn from an Erdös-Rényi or Poisson ensemble $\mathcal{P}_k(\alpha, n)$. The code can also be seen as a planted k-XOR-SAT formula. The dimension of the code is $m = \alpha n$ and the rate is $r = 1/\alpha$.

Despite the long history of research on the LDPC and LDGM codes, and their success in practical applications of communications, there are still many open questions concerning the behaviour of these codes. In particular, even for the simple code described above, it is still open to show that the mutual information $\frac{1}{n} I(X^n; Y^m)$ concentrates, with the exception of the binary erasure channel for which much more is known [36, 37]. In the case of dense random codes, standard probability arguments show that concentration occurs with a transition at capacity for any discrete memoryless channels. But for sparse codes, the traditional arguments fail. Recently, the following was conjectured in [45] for constant right degree LDGM codes G and binary input symmetric output channels[5],

$$\mathbb{P}_G\{\frac{1}{m} I_G(X;Y) < I(W)\} \to \begin{cases} 0 & \text{if } \alpha < C_k(W) \\ 1 & \text{if } \alpha > C_k(W) \end{cases} \tag{18}$$

where $C_k(W)$ is a constant depending on k and W.

We provide next a concentration result for this model, which implies the above conjecture for even degrees.

Definition 4.13. *An encoded symmetric kernel is given by*

$$Q(z|u) = W(z| \oplus_{i=1}^k u_i), \tag{19}$$

where W is a binary input symmetric output (BISO) channel from $\mathcal{X}$ to $\mathcal{Y}$.

Note that this corresponds to the output of a BISO W when the input to the channel is the XOR of k information bits. This corresponds also to the constant right-degree LDGM codes considered in the conjecture of [45].

Lemma 4.14. *For an encoded symmetric kernel with k even, the operator Γ_l is convex for any $l \geq 1$.*

Corollary 4.15. *Let X be uniformly drawn in $\{0,1\}^n$, $U = XG$ be the output of a k-degree LDGM code G of dimension αn, and Y be the output of U on a BISO channel W. (Note that we use G for both the graph and the generator matrix; this abuse of notation is however explained by the fact that the generator matrix is indeed the incidence matrix of the graph G.) Then $\frac{1}{n} I_G(X;Y)$ converges almost surely to a constant $C_k(\alpha, W)$.*

Note that for any realization of G, and dropping the subscript G, we have $\frac{1}{m} I(X;Y) = \frac{1}{m} H(Y) - H(W)$, where $H(W)$ denotes the conditional entropy of the channel W. Hence $\frac{1}{m} I(X;Y) < 1 - H(W) \equiv \frac{1}{m} H(Y) < 1$. Since $\frac{1}{m} H(Y)$ converges from previous corollary, and since the limit must be decreasing in α (increasing in r), the conjecture (18) follows.

[5] This means that the channel is a weighted sum of binary symmetric channels

5 Proof outline for Theorem 3.3: Interpolation method for graphical channels

We show in this section the sub-additivity of $H(X|Y,G_n)$, namely

$$H(X|Y,G_n) \leq H(X|Y,G_{n_1}) + H(X|Y,G_{n_2}). \tag{20}$$

Note that if we partition the set of vertices $[n]$ into two disjoint sets of size n_1 and n_2 with $n_1+n_2=n$, and denote by g_1 and g_2 subgraphs of g induced by these subsets (hence obtained by **removing** all the crossing hyperedges), then the following is obtained by basic properties of the entropy

$$H(X|Y,g) \leq H(X|Y,g_1) + H(X|Y,g_2). \tag{21}$$

Hence the above is also true for a random graph G drawn from the ensemble $\mathcal{P}_k(\alpha,n)$. However, the random graph obtained by restricting G_n to a subset of n_i vertices is not equivalent to G_{n_i}, since the edge probability stays at $\frac{\alpha n}{\binom{n}{k}}$ and is not at $\frac{\alpha n_i}{\binom{n_i}{k}}$ as it should be. Consequently, the above does not imply $H(X|Y,G_n) \leq H(X|Y,G_{n_1}) + H(X|Y,G_{n_2})$. To obtain the proper term on the right hand side, one should add the edges lost in the splitting of the vertices (e.g., using a coupling argument), but this gives a lower bound on the right hand side of (21), conflicting with the upper bound. This also shows that it may not be easy to guess the direction of the inequality. We rely on an interpolation method to compare the right quantities.

The interpolation method was first introduced in [29] for the Sherrington-Kirkpatrick model. This is a model for a spin-glass (i.e. a spin model with random couplings) on a complete graph. It was subsequently shown in [23, 24, 43] that the same ideas can be generalized to models on random sparse graphs, and applications in coding theory and random combinatorial optimization were proposed in [38, 34] and [10, 1]. We next develop an interpolation method to estimate the conditional entropy of general graphical channels. Interestingly, we will see that the planting flips the behaviour of the entropy from super to sub-additive.

Definition 5.1. *We define a more general Poisson model for the random graph, where a parameter $\varepsilon_I \geq 0$ is attached to each $I \in E_k(V)$, and the number of edges $m_I(\varepsilon_I)$ is drawn from a Poisson distribution of parameter ε_I. This defines a random hypergraph whose edge probability is not homogenous but depends on the parameters ε_I. Denoting by $\underline{\varepsilon}$ the collection of all $\binom{n}{k}$ parameters ε_I, we denote this ensemble as $\mathcal{P}_k(\underline{\varepsilon},n)$. If for any I, $\varepsilon_I = \frac{\alpha n}{\binom{n}{k}}$, $\mathcal{P}_k(\underline{\varepsilon},n)$ reduces to $\mathcal{P}_k(\alpha,n)$ as previously defined.*

Lemma 5.2. *Let X be uniformly drawn over $\mathcal{X}^n$, $G(\underline{\varepsilon})$ be a random hypergraph drawn from the ensemble $\mathcal{P}_k(\underline{\varepsilon},n)$ independently of X, and let $Y(\underline{\varepsilon})$ be the output of X through $P_{G(\underline{\varepsilon}),Q}$ defined in (1) for the kernel Q. Then*

$$\frac{\partial}{\partial \varepsilon_I} H(X|Y(\underline{\varepsilon}),G(\underline{\varepsilon})) = -I(Y_I;X_I|Y(\underline{\varepsilon}),G(\underline{\varepsilon}))\,, \tag{22}$$

where Y_I and $Y(\underline{\varepsilon})$ are independent conditionally on X (i.e., Y_I is drawn under $Q(\cdot|X[I])$ and $Y(\underline{\varepsilon})$ is drawn independently under $R_{G(\underline{\varepsilon}),Q}(\cdot|X)$).

We define a *path* as a differentiable map $t \mapsto \underline{\varepsilon}(t)$, with $t \in [0,T]$ for some $T \geq 0$. We say that a path is balanced if

$$\sum_{I \in E_k(V)} \frac{\mathrm{d}\varepsilon_I}{\mathrm{d}t}(t) = 0\,. \tag{23}$$

We will write $\dot{\varepsilon}_I(t)$ for the derivative of $\varepsilon_I(t)$ along the path and define $Y(t) = Y(\underline{\varepsilon}(t))$ and $G(t) = G(\underline{\varepsilon}(t))$.

In what follows, we will omit writing the graph $G(t)$ in the conditioning to shorten the expressions, although it is implicitly paired with $Y(t)$ in the conditioning.

Corollary 5.3. *For a balanced path*

$$\frac{\mathrm{d}}{\mathrm{d}t} H(X|Y(t)) = -\sum_{I \in E_k(V)} H(Y_I|Y(t))\, \dot{\varepsilon}_I(t)\,. \tag{24}$$

Given a partition $V = V_1 \sqcup V_2$, we define the associated *canonical path* $\underline{\varepsilon} : t \in [0,1] \to \underline{\varepsilon}(t) \in [0,1]^{E_k(V)}$ as follows. Let $n_i = |V_i|$, $m_i = |E_k(V_i)|$, $i \in \{1,2\}$, and $m = |E_k(V)|$. We define

$$\varepsilon_I(0) \equiv \frac{\alpha n}{m}, \quad \forall I \in E_k(V), \tag{25}$$

$$\varepsilon_I(1) \equiv \begin{cases} \frac{\alpha n_1}{m_1} & \text{if } I \in E_k(V_1) \\ \frac{\alpha n_2}{m_2} & \text{if } I \in E_k(V_2) \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

and

$$\underline{\varepsilon}(t) = (1-t)\underline{\varepsilon}(0) + t\underline{\varepsilon}(1). \tag{27}$$

Note that the canonical path is balanced. Moreover, at time $t = 0$, $\mathcal{P}_k(\underline{\varepsilon}(0), n)$ reduces to the original ensemble $\mathcal{P}_k(\alpha, n)$, and at time $t = 1$, $\mathcal{P}_k(\underline{\varepsilon}(1), n)$ reduces to two independent copies of the original ensemble on the subset of n_1 and n_2 variables: $\mathcal{P}_k(\alpha, n_1) \times \mathcal{P}_k(\alpha, n_2)$.

Applying Lemma 5.3, we obtain the following.

Corollary 5.4. *For the canonical path*

$$\frac{\mathrm{d}}{\mathrm{d}t} H(X|Y(t)) = \alpha n \mathbb{E}_I H(Y_I|Y(t)) - \alpha n_1 \mathbb{E}_{I_1} H(Y_{I_1}|Y(t)) - \alpha n_2 \mathbb{E}_{I_2} H(Y_{I_2}|Y(t))\,, \tag{28}$$

where I is drawn uniformly in $E_k(V)$, and I_i, $i \in \{1,2\}$, are drawn uniformly in $E_k(V_i)$.

We recall that

$$\begin{aligned} H(Y_I|Y(t)) &= -\mathbb{E}_{Y(t),Y_I} \log \sum_x Q(Y_I|x[I]) R_{G(t)}(x|Y(t)) && (29) \\ &= -\mathbb{E}_{Y(t),Y_I} \log \mathbb{E}_{X|Y(t)} Q(Y_I|X[I])\,, && (30) \end{aligned}$$

where $Y(t)$ is the output of $P_{G(t),Q}$ and $\mathbb{E}_{X|Y(t)}$ is the conditional expectation over $R_{G(t),Q}$.

Lemma 5.5.

$$\frac{1}{\alpha|\mathcal{Y}|}\frac{\mathrm{d}}{\mathrm{d}t}H(X|Y(t)) = -\sum_{l=2}^{\infty}\frac{1}{l(l-1)}\mathbb{E}_{X^{(1)},\dots,X^{(l)}}\left[n\Gamma_l(V) - n_1\Gamma_l(V_1) - n_2\Gamma_l(V_2)\right] \tag{31}$$

where

$$\Gamma_l(V) \equiv \mathbb{E}_{I,W_I}\prod_{r=1}^{l}\left(1 - Q(W_I|X^{(r)}[I])\right), \tag{32}$$

I is uniformly drawn in $E_k(V)$, W_I is uniformly drawn in $\mathcal{Y}$, and $X^{(1)},\dots,X^{(l)}$ are drawn under the probability distribution $\sum_y \prod_{i=1}^{l} R_{G(t),Q},(x^{(i)}|y)\sum_u P_{G(t),Q}(y|u)2^{-n}$.

This means that $X^{(1)},\dots,X^{(l)}$ are drawn i.i.d. from the channel $R_{G(t)}$ given a hidden output Y, these are the 'replica' variables, which are exchangeable but not i.i.d.. Note that denoting by ν the empirical distribution of $X^{(1)},\dots,X^{(l)}$, the above definition of $\Gamma_l(V)$ coincides with that of $\Gamma_l(\nu)$, hence the abuse of notation with definition (8). Hypothesis H ensures that Γ_l is convex for any distribution on $\mathcal{X}^l$, hence in particular for the empirical distribution of the replicas. Therefore, previous lemma implies Lemma 3.3 and Theorem 3.4 follows by the sub-additivity property.

References

[1] E. Abbe and A. Montanari, *On the concentration of the number of solutions of random satisfiability formulas*, Random Structures and Algorithm DOI 10.1002/rsa.20501, 2013, Available at arXiv:1006.3786v1 [cs.DM], 2010.

[2] E. Abbe and A. Montanari, *Conditional random fields, planted constraint satisfaction and entropy concentration*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (Prasad Raghavendra, Sofya Raskhodnikova, Klaus Jansen, and JosD.P. Rolim, eds.), Lecture Notes in Computer Science, vol. 8096, Springer Berlin Heidelberg, 2013, pp. 332–346.

[3] D. Achlioptas and A. Coja-Oghlan, *Algorithmic barriers from phase transitions*, Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science (Washington, DC, USA), FOCS '08, IEEE Computer Society, 2008, pp. 793–802.

[4] D. Achlioptas, H. Jia, and C. Moore, *Hiding satisfying assignments: two are better than one*, In Proceedings of AAAI, 2004, pp. 131–136.

[5] D. Achlioptas, H. Kautz, and C. Gomes, *Generating satisfiable problem instances*, In Proceedings of AAAI, 2000, pp. 256–261.

[6] D. Achlioptas, J. Han Kim, M. Krivelevich, and P. Tetali, *Two-coloring random hypergraphs*, Random Structures and Algorithms **20** (2002), no. 2, 249–259.

[7] D. Achlioptas, A. Naor, and Y. Peres, *Rigorous Location of Phase Transitions in Hard Optimization Problems*, Nature **435** (2005), 759–764.

[8] F. Altarelli, R. Monasson, and F. Zamponi, *Can rare SAT formulas be easily recognized? On the efficiency of message passing algorithms for K-SAT at large clause-to-variable ratios*, Computing Research Repository **abs/cs/060** (2006).

[9] W. Barthel, A. K. Hartmann, M. Leone, F. Ricci-Tersenghi, M. Weigt, and R. Zecchina, *Hiding solutions in random satisfiability problems: A statistical mechanics approach*, Phys. Rev. Lett. **88** (2002), 188701.

[10] M. Bayati, D. Gamarnik, and P. Tetali, *Combinatorial approach to the interpolation method and scaling limits in sparse random graphs*, 42nd Annual ACM Symposium on Theory of Computing (Cambridge, MA), June 2010, pp. 105–114.

[11] E. Berlekamp, R.J. McEliece, and H. C A Van Tilborg, *On the inherent intractability of certain coding problems (corresp.)*, Information Theory, IEEE Transactions on **24** (1978), no. 3, 384–386.

[12] A. Bogdanov and Y. Qiao, *On the security of Goldreich's one-way function*, computational complexity **21** (2012), no. 1, 83–127 (English).

[13] A. Coja-Oghlan, *Graph partitioning via adaptive spectral techniques*, Comb. Probab. Comput. **19** (2010), no. 2, 227–284.

[14] A. Coja-Oghlan, M. Krivelevich, and D. Vilenchik, *Why almost all satisfiable k-cnf formulas are easy*, Proceedings of the 13th International Conference on Analysis of Algorithms, 2007, pp. 89–102.

[15] H. Daudé and V. Ravelomanana, *Random 2-XORSAT at the satisfiability threshold*, Proceedings of the 8th Latin American conference on Theoretical informatics (Berlin, Heidelberg), LATIN'08, Springer-Verlag, 2008, pp. 12–23.

[16] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E 84, 066106 (2011).

[17] M. Dietzfelbinger, A. Goerdt, M. Mitzenmacher, A. Montanari, R. Pagh, and M. Rink, *Tight thresholds for cuckoo hashing via XORSAT*, Proceedings of the 37th international colloquium conference on Automata, languages and programming (Berlin, Heidelberg), ICALP'10, Springer-Verlag, 2010, pp. 213–225.

[18] O. Dubois and J. Mandler, *The 3-XORSAT threshold*, Proceedings of the 43rd Symposium on Foundations of Computer Science (Washington, DC, USA), FOCS '02, IEEE Computer Society, 2002, pp. 769–778.

[19] M.E Dyer and A.M Frieze, *The solution of some random NP-hard problems in polynomial expected time*, Journal of Algorithms **10** (1989), no. 4, 451 – 489.

[20] P. Elias, *Coding for noisy channels*, IRE Convention Record **4** (1955), 37–46.

[21] U. Feige, E. Mossel, and D. Vilenchik, *Complete convergence of message passing algorithms for some satisfiability problems*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, Springer, 2006, pp. 339–350.

[22] S. Fortunato, *Community detection in graphs*, Physics Reports **486 (3-5)** (2010), 75–174.

[23] S. Franz and M. Leone, *Replica bounds for optimization problems and diluted spin systems*, J. Stat. Phys. **111** (2003), 535.

[24] S. Franz, M. Leone, and F.L. Toninelli, *Replica bounds for diluted non-Poissonian spin systems*, J. Phys. A **36** (2003), 10967.

[25] E. Friedgut, *Sharp thresholds of graph properties, and the k-sat problem*, J. Amer. Math. Soc. **12** (1999), 1017–1054, appendix by J. Bourgain.

[26] R. G. Gallager, *Low-density parity-check codes*, MIT Press, Cambridge, Massachussetts, 1963.

[27] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, *A survey of statistical network models*, Foundations and Trends in Machine Learning **2** (2010), no. 2, 129–233.

[28] O. Goldreich, *Studies in complexity and cryptography*, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 76–87.

[29] F. Guerra and F. L. Toninelli, *The thermodynamic limit in mean field spin glasses*, Commun. Math. Phys. **230** (2002), 71–79.

[30] H. Haanpää, M. Järvisalo, P. Kaski, and I. Niemelä, *Hard satisfiable clause sets for benchmarking equivalence reasoning techniques*, Journal on Satisfiability, Boolean Modeling and Computation 2 (2005), 27–46.

[31] H. Jia, C. Moore, and D. Strain, *Generating hard satisfiable formulas by hiding solutions deceptively*, In AAAI, AAAI Press, 2005, pp. 384–389.

[32] B. Karrer and M. E. J. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E **83** (2011), 016107.

[33] F. Krzakala and L. Zdeborová, *Hiding quiet solutions in random constraint satisfaction problems*, Phys. Rev. Lett. **102** (2009), 238701.

[34] S. Kudekar and N. Macris, *Sharp bounds for optimal decoding of Low-Density Parity-Check codes*, IEEE Trans. on Inform. Theory **55** (2009), 4635–4650.

[35] J. Lafferty, *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, Morgan Kaufmann, 2001, pp. 282–289.

[36] M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. A. Spielman, *Efficient erasure correcting codes*, IEEE Trans. on Inform. Theory **47** (2001), no. 2, 569–584.

[37] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. A. Spielman, and V. Stemann, *Practical loss-resilient codes*, 29th annual ACM Symposium on Theory of Computing, 1997, pp. 150–159.

[38] A. Montanari, *Tight bounds for LDPC and LDGM codes under MAP decoding*, IEEE Trans. on Inform. Theory **51** (2005), 3221–3246.

[39] A. Montanari, *Estimating random variables from random sparse observations*, European Transactions on Telecommunications **19** (2008), no. 4, 385–403.

[40] A. Montanari, R. Restrepo, and P. Tetali, *Reconstruction and Clustering in Random Constraint Satisfaction Problems*, CoRR abs/0904.2751, 2009.

[41] E. Mossel, J. Neeman, and A. Sly, *Stochastic Block Models and Reconstruction*, (2012), arXiv:1202.1499 [math.PR], 2012.

[42] M. E. J. Newman, *Communities, modules and large-scale structure in networks*, Nature Physics **8** (2011), no. 1, 25–31.

[43] D. Panchenko and M. Talagrand, *Bounds for diluted mean-field spin glass models*, Prob. Theor. Rel. Fields **130** (2004), 319–336.

[44] B. Pittel and G. B. Sorkin, *The Satisfiability Threshold for k-XORSAT*, arXiv:1212.1905 (2012).

[45] K. Raj Kumar, P. Pakzad, A.H. Salavati, and A. Shokrollahi, *Phase transitions for mutual information*, Turbo Codes and Iterative Information Processing (ISTC), 2010 6th International Symposium on, 2010, pp. 137–141.

[46] T. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, Cambridge, 2008.

[47] L. Zdeborová and F. Krzakala, *Quiet planting in the locked constraint satisfaction problems*, SIAM Journal on Discrete Mathematics **25** (2011), no. 2, 750–770.

In the following sections, we omit writing the graph in the conditioning to shorten the expressions.

A Proofs of Theorems 3.4 and 3.5

We now prove the lemmas used in Section 5 to prove Theorem 3.4. We then prove Theorem 3.5.

Proof of Lemma 5.2. Note that for a random variable Z_ε which is Poisson distributed of parameter ε, and a function f,

$$\frac{\partial}{\partial \varepsilon}\mathbb{E}f(Z_\varepsilon) = \mathbb{E}f(Z_\varepsilon + 1) - \mathbb{E}f(Z_\varepsilon). \tag{33}$$

Therefore,

$$\frac{\partial}{\partial \varepsilon_I} H(X|Y(\underline{\varepsilon})) = H(X|Y(\underline{\varepsilon}), Y_I) - H(X|Y(\underline{\varepsilon})), \tag{34}$$

where Y_I is an extra output drawn independently from $Y(\underline{\varepsilon})$ but conditionally on the same X. We also recall the definition of the mutual information, $I(A;B) = H(A) - H(A|B) = H(B) - H(B|A)$. We then have

$$\begin{aligned}
\frac{\partial}{\partial \varepsilon_I} H(X|Y(\underline{\varepsilon})) &= -I(Y_I; X|Y(\underline{\varepsilon})) && (35)\\
&= -H(Y_I|Y(\underline{\varepsilon})) + H(Y_I|X, Y(\underline{\varepsilon})) && (36)\\
&= -H(Y_I|Y(\underline{\varepsilon})) + H(Y_I|X_I, Y(\underline{\varepsilon})) && (37)\\
&= -I(Y_I; X_I|Y(\underline{\varepsilon})), && (38)
\end{aligned}$$

where we used the fact that Y_I depends only on the components of X indexed by I. □

Proof of Corollary 5.3. Note that from previous proof, and using the fact that $Y_I - X - Y(\underline{\varepsilon})$ form a Markov chain,

$$\begin{aligned}
\frac{\partial}{\partial \varepsilon_I} H(X|Y(\underline{\varepsilon})) &= -I(Y_I; X_I|Y(\underline{\varepsilon})) && (39)\\
&= -H(Y_I|Y(\underline{\varepsilon})) + H(Y_I|X_I) && (40)\\
&= -H(Y_I|Y(\underline{\varepsilon})) + H(Q) && (41)
\end{aligned}$$

where

$$H(Q) \equiv -2^{-k} \sum_{u \in \mathcal{X}^k, z \in \mathcal{Y}} Q(z|u) \log Q(z|u)$$

is a constant depending on Q only. Therefore, if the path is balanced, the chain rule yields the result. □

Proof of Lemma 5.5. By definition

$$H(Y_I|Y(t)) = -\mathbb{E}_{Y(t),Y_I} \log \mathbb{E}_{X|Y(t)} Q(Y_I|X[I]), \tag{42}$$

and expanding the logarithm in its power series,

$$\log \mathbb{E}_{X|Y(t)} Q(Y_I|X[I]) = -\sum_{l=1}^{\infty} \frac{1}{l} (\mathbb{E}_{X|Y(t)}(1 - Q(Y_I|X[I])))^l. \tag{43}$$

We now introduce the 'replicas' $X^{(1)}, \dots, X^{(l)}$, which are i.i.d. under $Q_{X|Y(t)}$, i.e., we have the Markov relation $X - Y(t) - (X^{(1)}, \dots, X^{(l)})$. Denoting by $\widetilde{Q} = 1 - Q$, we obtain

$$\log \mathbb{E}_{X|Y(t)} Q(Y_I|X[I]) = -\sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}_{X^{(1)},\dots,X^{(l)}|Y(t)} \prod_{r=1}^{l} \widetilde{Q}(Y_I|X^{(r)}[I])). \tag{44}$$

As opposed to the non-planted case, where supper-additivity is achieved by showing that terms weighted by $1/l$ are convex in the empirical distribution of the replicas, convexity does not hold with the above expression and the following is needed. Collecting terms we have

$$H(Y_I|Y(t)) = \mathbb{E}_X \mathbb{E}_{Y(t)|X} \mathbb{E}_{Y_I|X} \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}_{X^{(1)},\dots,X^{(l)}|Y(t)} \prod_{r=1}^{l} \widetilde{Q}(Y_I|X^{(r)}[I])) \tag{45}$$

$$= \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}_{X,X^{(1)}\dots,X^{(l)}} \mathbb{E}_{Y_I|X} \prod_{r=1}^{l} \widetilde{Q}(Y_I|X^{(r)}[I])). \tag{46}$$

We next switch measure in the expectation $\mathbb{E}_{Y_I|X}$, defining W_I to be uniformly distributed over $\mathcal{Y}$, and writing

$$H(Y_I|Y(t)) = |\mathcal{Y}| \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}_{X,X^{(1)}\dots,X^{(l)}} \mathbb{E}_{W_I} \prod_{r=1}^{l} \widetilde{Q}(W_I|X^{(r)}[I]))Q(W_I|X[I])). \tag{47}$$

Renaming X by $X^{(0)}$, and using the fact that $X^{(0)}, X^{(1)} \dots, X^{(l)}$ are exchangeable, we can write

$$H(Y_I|Y(t)) = |\mathcal{Y}| \sum_{l=1}^{\infty} \frac{1}{l} \mathbb{E}_{X^{(0)},X^{(1)}\dots,X^{(l)}} \left(\mathbb{E}_{W_I} \prod_{r=1}^{l} \widetilde{Q}(W_I|X^{(r)}[I])) - \mathbb{E}_{W_I} \prod_{r=0}^{l} \widetilde{Q}(W_I|X^{(r)}[I])) \right) \tag{48}$$

$$= |\mathcal{Y}| \mathbb{E}_{X^{(1)}} \mathbb{E}_{W_I} \widetilde{Q}(W_I|X^{(1)}[I])) - |\mathcal{Y}| \sum_{l=2}^{\infty} \frac{1}{l(l-1)} \mathbb{E}_{X^{(1)}\dots,X^{(l)}} \mathbb{E}_{W_I} \prod_{r=1}^{l} \widetilde{Q}(W_I|X^{(r)}[I])). \tag{49}$$

Recall that from Corollary 5.4,

$$\frac{\mathrm{d}}{\mathrm{d}t} H(X|Y(t)) = \alpha n \mathbb{E}_I H(Y_I|Y(t)) - \alpha n_1 \mathbb{E}_{I_1} H(Y_{I_1}|Y(t)) - \alpha n_2 \mathbb{E}_{I_2} H(Y_{I_2}|Y(t)) \,.$$

Hence, carrying out the above expansions for each term, and since the term $\mathbb{E}_{X^{(1)}} \mathbb{E}_{W_I} \widetilde{Q}(W_I|X^{(1)}[I]))$ cancels out, the lemma follows. □

Proof of Theorem 3.5. Since we know that $H_G(X|Y)/n$ converges in expectation, it is sufficient to show that it concentrates around its expectation. Indeed we claim that there exists $B > 0$ such that

$$\mathbb{P}\{|H_G(X|Y) - H(X|Y)| \geq n\Delta\} \leq 2\, e^{-nB\Delta^2} \,, \tag{50}$$

whence our thesis follows from Borel-Cantelli.

The proof of (50) is a direct application of Azuma-Hoeffding inequality, and we limit ourselves to sketching its main steps. We condition on the number $m = O(n)$ of hyperedges in G and regard $H_G(X|Y)$ as a function of the choice of the m hyperedges. We claim that $|H_G(X|Y) - H_{G'}(X|Y)| \leq 2C$, for some constant C, if G and G' differ only in one of their hyperedges, whence Eq. (50) follows by Azuma-Hoeffding.

In order to prove the last claim, let $G' = G + a$ denote the graph G to which hyperedge $a = (i_1, \dots, i_k)$ has been added. Then, writing explicitly the component of Y corresponding to hyperedge a by Y_a, we need to prove that $|H_{G+a}(X|Y, Y_a) - H_G(X|Y)| \leq C$. We have, dropping the subscripts and superscript for the sake of simplicity,

$$\begin{aligned} 0 \leq H(X|Y) - H(X|Y, Y_a) &= H(X|Y) - H(X, Y_a|Y) + H(Y_a|Y) && (51)\\ &= H(Y_a|Y) - H(Y_a|X, Y) && (52)\\ &\leq \log_2 |\mathcal{Y}|\,, && (53) \end{aligned}$$

where the last inequality follows from the fact that Y_a takes value in the finite set $\mathcal{Y}$. □

B Proofs of Lemmas 4.2, 4.5, 4.6 and 4.7

Proof of Lemma 4.2. We have

$$\begin{aligned} P_{g,Q}(y|x) &= \prod_{I \in E(g)} Q(y_I|x[I]) && (54)\\ &= \prod_{I \in E(g)} \frac{1}{|A|} \mathbb{1}(y_I \in A(x[I])) && (55)\\ &= \begin{cases} \frac{1}{|A|^{|E(g)|}} & \text{if } x \sim y, \\ 0 & \text{otherwise,} \end{cases} && (56) \end{aligned}$$

where $x \sim y$ means that x is a satisfying assignment for y (and using the constrained defined by the kernel Q). Hence for a given x, $P_{g,Q}(\cdot|x)$ is uniform on the set of all y's verifying x, which has cardinality $|A|^{|E(g)|}$. Since X is uniform, for a given y, $R_{g,Q}(\cdot|y)$ is a uniform measure on a set of cardinality

$$\sum_{x \in \mathcal{X}^n} \prod_{I \in E(g)} \mathbb{1}(y_I \in A(x[I])) = |\{x \in \mathcal{X}^n : y_I \in A(x[I]), \forall I \in E(g)\}| = Z_g(y) \qquad (57)$$

Therefore $H_g(X|Y = y) = \log Z_g(y)$. □

Proof of Lemma 4.5. For planted k-SAT, $\mathcal{Y} = \mathcal{X}^k = \{0, 1\}^k$,

$$Q(z|u) = \frac{1}{2^k - 1} \mathbb{1}(z \neq \bar{u}) \qquad (58)$$

and

$$\Gamma_l(\nu) = \frac{1}{2^k}\left(\frac{1}{2^k-1}\right)^l \sum_{u^{(1)},\dots,u^{(l)}\in\mathcal{X}^k} \left[\sum_{z\in\mathcal{Y}}\prod_{r=1}^{l} \mathbb{1}(\bar{z} = u^{(r)})\right] \prod_{i=1}^{k} \nu(u_i^{(1)},\dots,u_i^{(l)}) \tag{59}$$

$$= \frac{1}{2^k}\left(\frac{1}{2^k-1}\right)^l \sum_{u\in\mathcal{X}^k} \prod_{i=1}^{k} \nu(u_i,\dots,u_i) \tag{60}$$

$$= \frac{1}{2^k}\left(\frac{1}{2^k-1}\right)^l \sum_{u_1,\dots,u_k\in\mathcal{X}} \prod_{i=1}^{k} \nu(u_i,\dots,u_i) \tag{61}$$

$$= \frac{1}{2^k}\left(\frac{1}{2^k-1}\right)^l \left(\sum_{u_1\in\mathcal{X}} \nu(u_1,\dots,u_1)\right)^k, \tag{62}$$

which is convex in ν for any $k, l \geq 1$. □

Proof of Lemma 4.6. For planted k-NAE-SAT, $\mathcal{Y} = \mathcal{X}^k = \{0,1\}^k$,

$$Q(z|u) = \frac{1}{2^k-2}\mathbb{1}(z \notin (u,\bar{u})) \tag{63}$$

and

$$\Gamma_l(\nu) = \frac{1}{2^k}\left(\frac{1}{2^k-2}\right)^l \sum_{u^{(1)},\dots,u^{(l)}\in\mathcal{X}^k} \left[\sum_{z\in\mathcal{Y}}\prod_{r=1}^{l} \mathbb{1}(u^{(r)} \in (z,\bar{z}))\right] \prod_{i=1}^{k} \nu(u_i^{(1)},\dots,u_i^{(l)}) \tag{64}$$

$$= \frac{1}{2^k}\left(\frac{1}{2^k-2}\right)^l \sum_{b_1,\dots,b_l\in\mathcal{X}} \sum_{u\in\mathcal{X}^k} \prod_{i=1}^{k} \nu(u_i \oplus b_1,\dots,u_i \oplus b_l) \tag{65}$$

$$= \frac{1}{2^k}\left(\frac{1}{2^k-2}\right)^l \sum_{b_1,\dots,b_l\in\mathcal{X}} \left(\sum_{u_1\in\mathcal{X}} \nu(u_1 \oplus b_1,\dots,u_1 \oplus b_l)\right)^k, \tag{66}$$

which is convex in ν for any $k, l \geq 1$. □

Proof of Lemma 4.7. This is a special case of Lemma C.2 for $s = 1$, $d = -1$. □

C Proofs of Lemmas 4.10 and 4.11

Proof of Lemma 4.10. We introduce a new collection of random variables $\{Z_{ij}\}_{(ij)\in E_2(V)}$, taking values in $\{0,1,*\}$, and indexed by the $\binom{n}{2}$ edges of the complete graph over vertex set V. These are conditionally independent given X with distribution given as follows:

$$Z_{ij}\big|_{X_i=X_j} = \begin{cases} 1 & \text{with probability } a/n, \\ 0 & \text{with probability } (2\gamma - a)/n, \\ * & \text{with probability } 1 - 2\gamma/n, \end{cases} \tag{67}$$

$$Z_{ij}\big|_{X_i\neq X_j} = \begin{cases} 1 & \text{with probability } b/n, \\ 0 & \text{with probability } (2\gamma - b)/n, \\ * & \text{with probability } 1 - 2\gamma/n, \end{cases} \tag{68}$$

The following claim is proved below.

Lemma C.1. *There exists a constant $C = C(\gamma) < \infty$ such that, uniformly in n*

$$\big|H(X|Y_\gamma) - H(X|Z)\big| \leq C(\gamma)\,. \tag{69}$$

It is therefore sufficient to bound the difference $|H(X|Y) - H(X|Z)|$. Notice that the variables X, Y, Z can be constructed on the same probability space in such a way that $X - Z - Y$ form a Markov chain. Namely, it is sufficient to let the Y_{ij} be conditionally independent, and independent from X given Z, with

$$Y_{ij} = \begin{cases} 1 & \text{if } Z_{ij} = 1, \\ 0 & \text{if } Z_{ij} \in \{0, *\}. \end{cases} \tag{70}$$

We therefore have that $H(X|Z) \leq H(X|Y)$ and we are left with the task of upper bounding $H(X|Y) - H(X|Z)$.

We have, by the Markov property and the chain rule of conditional entropy

$$\begin{aligned} H(X|Z) &= H(X|Y,Z) && (71)\\ &= H(X,Z|Y) - H(Z|Y) && (72)\\ &= H(X|Y) + H(Z|X,Y) - H(Z|Y)\,, && (73) \end{aligned}$$

and therefore

$$H(X|Y) - H(X|Z) = H(Z|Y) - H(Z|X,Y)\,. \tag{74}$$

Note that, by subadditivity of the entropy, and since conditioning reduces entropy, we have

$$H(Z|Y) \leq \sum_{(i,j)\in E_2(V)} H(Z_{ij}|Y) \leq \sum_{(i,j)\in E_2(V)} H(Z_{ij}|Y_{ij})\,. \tag{75}$$

Further, by conditional independence of the $\{Z_{ij}\}$ given X, Y, we have

$$H(Z|X,Y) = \sum_{(i,j)\in E_2(V)} H(Z_{ij}|X,Y) = \sum_{(i,j)\in E_2(V)} H(Z_{ij}|X_i,X_j,Y_{ij})\,. \tag{76}$$

We therefore conclude that

$$H(X|Y) - H(X|Z) \leq \sum_{(i,j)\in E_2(V)} \big\{H(Z_{ij}|Y_{ij}) - H(Z_{ij}|X_i,X_j,Y_{ij})\big\}\,. \tag{77}$$

A simple calculation yields $H(Z_{ij}|Y_{ij}) = f_n((a+b)/2)$ and $H(Z_{ij}|X_i, X_j, Y_{ij}) = f_n(a) + f_n(b))/2$, where

$$f_n(c) \equiv -\frac{2\gamma - c}{n}\log\Big(\frac{2\gamma-c}{n}\Big) - \Big(1-\frac{2\gamma}{n}\Big)\log\Big(1-\frac{2\gamma}{n}\Big) + \Big(1-\frac{c}{n}\Big)\log\Big(1-\frac{c}{n}\Big)\,, \tag{78}$$

Subtracting an affine term, we can write

$$f_n(c) = f_{n,0} + f_{n,1}c + g_n(c)\,, \tag{79}$$

$$g_n(c) = -\frac{2\gamma}{n}\Big[\Big(1-\frac{c}{2\gamma}\Big)\log\Big(1-\frac{c}{2\gamma}\Big) + \frac{c}{2\gamma}\Big] + \Big[\Big(1-\frac{c}{n}\Big)\log\Big(1-\frac{c}{n}\Big) + \frac{c}{n}\Big]\,. \tag{80}$$

Note that, for $x \in [-1/2, 41/2]$, we have $0 \le (1-x)\log(1-x) + x \le x^2$. Hence, for $\gamma \ge c$, $n \ge 2c$,

$$0 \ge g_n(c) \ge \frac{c^2}{n}\Big[\frac{1}{n} - \frac{1}{2\gamma}\Big] \ge -\frac{c^2}{2n\gamma}\,, \tag{81}$$

and therefore

$$\begin{aligned} H(X|Y) - H(X|Z) &\le \binom{n}{2}\Big[f_n\Big(\frac{a+b}{2}\Big) - \frac{1}{2}f_n(a) - \frac{1}{2}f_n(b)\Big] \\ &= \binom{n}{2}\Big[g_n\Big(\frac{a+b}{2}\Big) - \frac{1}{2}g_n(a) - \frac{1}{2}g_n(b)\Big] \\ &\le \frac{n^2}{2}\,\frac{1}{2n\gamma}\,\frac{(a-b)^2}{4} = \frac{n(a-b)^2}{16\gamma}\,. \end{aligned}$$

This finishes the proof. □

Proof of Lemma C.1. We write $Y_\gamma = \{Y_{\gamma,ij}\}_{(i,j)\in E_2(V)}$ where, for each $(i,j) \in E_2(V)$, $Y_{\gamma,ij}$ is a vector containing Poisson$(n\gamma/\binom{n}{2})$ entries, each being an independent output of the channel Q in Eq. (15) on input (X_i, X_j). Analogously, we can interpret Z_{ij} as a vector of length Bernoulli$(2\gamma/n)$, with the length 0 corresponding to the value $*$. When the length of the vector is equal to one, its entry is distributed as the output of the same channel Q.

Let $\ell(Y_{\gamma,ij})$ and $\ell(Z_{ij})$ denote the length of vectors $Y_{\gamma,ij}$ and Z_{ij}. It follows from standard estimates on Poisson random variables that Y and Z can be coupled in such a way that $\mathbb{E}\{|\ell(Y_{\gamma,ij}) - \ell(Z_{ij})|\} \le C/n^2$ with $C = C(\gamma)$ and further, whenever $\ell(Z_{ij}) = 1$ and $\ell(Y_{\gamma,ij}) \ge 1$, the first entry of the vector $Y_{\gamma,ij}$ is equal to the only entry in Z_{ij}.

Finally notice that

$$\begin{aligned} &H(X|Y_\gamma, \ell(Y_{\gamma,ij}) = \ell_0 + 1) - H(X|Y_\gamma, \ell(Y_{\gamma,ij}) = \ell_0) && (82)\\ &\qquad = H(Y'_{ij}|X, \ell(Y_{\gamma,ij}) = \ell_0) - H(Y'_{ij}|Y_\gamma, \ell(Y_{\gamma,ij}) = \ell_0)\,, && (83) \end{aligned}$$

with Y'_{ij} distributed as an independent output of the channel Q on input X_i, X_j, It follows that $|H(X|Y_\gamma, \ell(Y_{\gamma,ij}) = \ell_0 + 1) - H(X|Y_\gamma, \ell(Y_{\gamma,ij}) = \ell_0)| \le 1$ and therefore

$$|H(X|Y_\gamma) - H(X|Z)| \le \sum_{(i,j)\in E_2(V)} \mathbb{E}\{|\ell(Y_{\gamma,ij}) - \ell(Z_{ij})|\} \le n^2\,\frac{C}{n^2} \le C\,. \tag{84}$$

□

Proof of Lemma 4.11. This is a special case of Lemma C.2 below, with $s = (a+b)/\gamma$ and $d = (a-b)/\gamma$. If γ is large enough, then $s \leq 1$ and if $a \leq b$, then $d \geq 0$ and all the coefficients in (92) are positive. □

Lemma C.2. *If $\mathcal{X} = \mathcal{Y} = \{0,1\}$, $Q(y|x_1,\ldots,x_k) = W(y|\oplus_{i=1}^k x_i)$ and W is an arbitrary binary input/output channel, then*

$$\Gamma_l(\nu) = \frac{1}{2}\sum_{w\in\mathbb{F}_2^l} d^{|w|}\left[s^{l-|w|} + (-1)^{|w|}(2-s)^{l-|w|}\right]\mathcal{F}(\nu)^k(w) \tag{85}$$

where $s = W(1|0)+W(1|1)$, $d = W(1|0)-W(1|1)$, $|w| = \sum_{i=1}^l w_i$ and $\mathcal{F}(\nu)(w) = \sum_{x\in\mathbb{F}_2^l}(-1)^{x\cdot w}\nu(x)$ is the Fourier-Walsh transform of ν (where $x \cdot w$ denotes the dot product of x and w).

Note that $\mathcal{F}(\nu)(w)$ is linear in ν, hence

- For $s = 1$, i.e., for symmetric channels,

$$\Gamma_l(\nu) = \sum_{w\in\mathbb{F}_2^k:\, |w| \text{ even}} d^{|w|}\mathcal{F}(\nu)^k(w) \tag{86}$$

  and Γ_l is convex when k is even.

- If $s \geq 1$, $d \geq 0$ or $s \leq 1$, $d \leq 0$, then Γ_l is convex when k is even.

Proof of Lemma C.2. We have

$$\Gamma_l(\nu) = \frac{1}{2}\sum_{u^{(1)},\ldots,u^{(l)}\in\mathbb{F}_2^k}\left[\sum_{y\in\mathbb{F}_2}\prod_{r=1}^{l}(1-P(y|u^{(r)}))\right]\prod_{i=1}^{k}\nu(u_i^{(1)},\ldots,u_i^{(l)}) \tag{87}$$

and using the fact that $P(y|u^{(r)}) = W(y|\oplus_{i=1}^k u_i^{(r)})$,

$$\Gamma_l(\nu) = \frac{1}{2}\sum_{v^{(1)},\ldots,v^{(l)}\in\mathbb{F}_2}\left[\sum_{y\in\mathbb{F}_2}\prod_{r=1}^{l}(1-W(y|v^{(r)}))\right]\nu^{\star k}(v^{(1)},\ldots,v^{(l)}) \tag{88}$$

$$= \frac{1}{2}\sum_{v\in\mathbb{F}_2^l}\gamma(v)\nu^{\star k}(v) \tag{89}$$

where

$$\gamma(v) \equiv \sum_{y\in\mathbb{F}_2}\prod_{r=1}^{l}(1-W(y|v^{(r)})) = (1-a)^{l-|v|}(1-b)^{|v|} + a^{l-|v|}b^{|v|} \tag{90}$$

and $a = W(1|0)$, $b = W(1|1)$. Note that

$$a^{l-|v|}b^{|v|} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad (a+b)^{l-|w|}(a-b)^{|w|}, \tag{91}$$

hence

$$\mathcal{F}(\gamma)(w) = (2-(a+b))^{l-|w|}(a-b)^{|w|}(-1)^{|w|} + (a+b)^{l-|w|}(a-b)^{|w|}\,. \tag{92}$$

□

Proof of (91). To show that

$$\mathbb{F}_2^l \ni v \mapsto \rho^{|v|} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad \mathbb{F}_2^l \ni w \mapsto (1+\rho)^{l-|w|}(1-\rho)^{|w|} \tag{93}$$

note that the identity is true when $l = 1$ and assume it to be true for l. Then for $l+1$

$$\begin{aligned} \sum_{v \in \mathbb{F}_2^{l+1}} \rho^{|v|}(-1)^{|vw|} &= \sum_{v \in \mathbb{F}_2^l} \rho^{|v|}(-1)^{|vw_1^l|} + \rho^{|v|+1}(-1)^{|vw_1^l|}(-1)^{w_{l+1}} && (94) \\ &= \sum_{v \in \mathbb{F}_2^l} \rho^{|v|}(-1)^{|vw_1^l|}(1+\rho(-1)^{w_{l+1}}) && (95) \\ &= (1+\rho)^{l-|w_1^l|}(1-\rho)^{|w_1^l|}(1+\rho(-1)^{w_{l+1}})\,. && (96) \end{aligned}$$

□

D Proofs of Lemma 4.14

Proof. We represent the channel W as a $2 \times |\mathcal{Y}|$ stochastic matrix. By definition of BISO channels, this matrix can be decomposed into pairs of columns which are symmetric as

$$\begin{pmatrix} c & d \\ d & c \end{pmatrix} \tag{97}$$

with $c, d \geq 0$, or into single columns which have constant values. Let us assume that W contains m such matrices and s such constant columns. We have

$$\begin{aligned} \Gamma_l(\nu) &= \frac{1}{|\mathcal{Y}|} \sum_{u^{(1)},\ldots,u^{(l)} \in \mathbb{F}_2^k} \left[\sum_{y \in \mathcal{Y}} \prod_{r=1}^{l} (1 - P(y|u^{(r)})) \right] \prod_{i=1}^{k} \nu(u_i^{(1)}, \ldots, u_i^{(l)}) && (98) \\ &= \frac{1}{|\mathcal{Y}|} \sum_{v^{(1)},\ldots,v^{(l)} \in \mathbb{F}_2} \left[\sum_{y \in \mathcal{Y}} \prod_{r=1}^{l} (1 - W(y|v^{(r)})) \right] \nu^{\star k}(v^{(1)}, \ldots, v^{(l)}) && (99) \\ &= \frac{1}{|\mathcal{Y}|} \sum_{v \in \mathbb{F}_2^l} g(v)\nu^{\star k}(v) && (100) \\ &= \frac{1}{|\mathcal{Y}|} \sum_{w \in \mathbb{F}_2^l} \mathcal{F}(g)(w)\mathcal{F}(\nu)^k(w) && (101) \end{aligned}$$

where

$$g(v) = \sum_{i=1}^{m} \left(C_i^{l-|v|} D_i^{|v|} + D_i^{l-|v|} C_i^{|v|} \right) + \sum_{i=1}^{s} E_i^l, \tag{102}$$

for some positive constants C_i, D_i, $i \in [m]$, E_i, $i \in [s]$. Moreover, using (91),

$$\begin{aligned} C^{l-|v|}D^{|v|} + D^{l-|v|}C^{|v|} \quad \overset{\mathcal{F}}{\longleftrightarrow} \quad & (C+D)^{l-|w|}(C-D)^{|w|} + (C+D)^{l-|w|}(D-C)^{|w|} && (103) \\ &= (C+D)^{l-|w|}(C-D)^{|w|}(1+(-1)^{|w|}), && (104) \end{aligned}$$

and $\mathcal{F}(g)(w)$ has only positive coefficients since only the terms with $|w|$ even survive. Hence Γ_l is convex when k is even. □