Critical noise levels for low-density parity check decoding

J. van Mourik and D. Saad
The Neural Computing Research Group, Aston University, Birmingham B4 7ET, United Kingdom

Y. Kabashima
Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 2268502, Japan

(Received 8 March 2002; published 28 August 2002)

We determine the critical noise level for decoding low-density parity check error-correcting codes based on the magnetization enumerator ($\mathcal{M}$), rather than on the weight enumerator ($\mathcal{W}$) employed in the information theory literature. The interpretation of our method is appealingly simple, and the relation between the different decoding schemes such as typical pairs decoding, MAP, and finite temperature decoding (MPM) becomes clear. In addition, our analysis provides an explanation for the difference in performance between MN and Gallager codes. Our results are more optimistic than those derived using the methods of information theory and are in excellent agreement with recent results from another statistical physics approach.

DOI: 10.1103/PhysRevE.66.026705 PACS number(s): 02.50.−r, 89.70.+c, 89.90.+n, 05.50.+q

I. INTRODUCTION

The theory of error-correcting codes is based on the efficient introduction of redundancy to given messages for protecting the information content against corruption. The theoretical foundations of this area were laid by Shannon's seminal work [1] and have been developing ever since. One of the main results obtained in this field is the celebrated *channel coding theorem* stating that there exists a code such that the average message error probability P_E, when maximum likelihood decoding is used, can be made arbitrarily small for sufficiently long messages below the *channel capacity*; and will approach 1 above it. The channel coding theorem is based on unstructured random codes and impractical decoders such as maximum likelihood [2] and typical set decoding [3]. In the case of structured codes, the critical code rate R (message information content/length of the encoded transmission) may lie below the channel capacity, commonly termed *Shannon's bound*, even if optimal (and typically impractical) decoding methods are being used. The proximity of the critical code rate to Shannon's limit provides an indication to the theoretical limitations of a given code. It should be emphasized that the theoretical critical code rate is typically not achievable in practice, as it may require using search methods that scale exponentially with the system size, in the computing time needed.

In 1963 Gallager [4] proposed a coding scheme that involves sparse linear transformations of binary messages that was forgotten soon after, in part due to the success of convolutional codes [2] and the computational limitations of the time. Gallager codes have been recently rediscovered by MacKay and Neal (MN), who independently proposed a closely related code [5]. Variations of this family of codes, known as low-density parity check (LDPC) codes, have displayed performance comparable (and sometimes superior) to other state-of-the-art codes. This family of codes has been thoroughly investigated in the information theory (IT) literature (e.g., [3,5,6]), providing a range of significant theoretical and practical results.

In parallel to studies carried out in the IT community, a different approach has been used to study LDPC codes, using the established methods of statistical physics (SP). This analysis, relying mainly on the replica symmetric analysis of diluted systems [7,8], offers an alternative to information theory methods and has yielded some additional results and insights [9,11,12]. Due to the growing interest in LDPC codes and their successful analysis via the methods of statistical physics, there is growing interest in the relationship between IT and SP methods. As the two communities investigate similar problems, one may expect that standard techniques known in one framework would bring about developments in the other, and vice versa. Here we present a direct SP method to determine the critical noise level of Gallager and MN error-correcting codes, which allows us to focus on the differences between the various decoding criteria and their use for defining the critical noise level for which decoding is theoretically feasible.

The paper is organized as follows: In Sec. II we introduce the general framework, notation and the quantities we focus on, while in Sec. III we will briefly describe the SP calculation. Section IV describes qualitatively the emerging picture of the main quantities calculated for Gallager's code while the corresponding picture for MN codes will be described in Sec. V. Quantitative results for the critical noise level will be presented in Sec. VI followed by conclusions.

II. REGULAR GALLAGER AND MN CODES

In a general scenario, the N-dimensional Boolean message $\vec{s}^o \in \{0,1\}^N$ is encoded to the $M(>N)$ dimensional Boolean vector $\vec{t}^o$, and transmitted via a noisy channel, which is taken here to be a binary symmetric channel (BSC) characterized by an independent flip probability p per bit; other transmission channels may also be examined within a similar framework. At the other end of the channel, the corrupted codeword is decoded utilizing the structured code word redundancy.

The first type of error-correcting code that we focus on here, is Gallager's linear code [4]. Gallager's code is a low-density parity check code defined by the a binary $(M-N)\times M$ matrix $\mathbf{A}=[\mathbf{C}_1|\mathbf{C}_2]$, concatenating two very sparse ma-

trices known to both sender and receiver, with the $(M-N)\times(M-N)$ matrix $\mathbf{C}_2$ being invertible. The matrix $\mathbf{A}$ has K nonzero elements per row and C per column, and the code rate is given by $R=1-C/K=N/M$. Encoding refers to multiplying the original message $\vec{s}^o$ with the $(M\times N)$ matrix $\mathbf{G}^T$ (where $\mathbf{G}=[\mathbb{1}_N|\mathbf{C}_2^{-1}]$), yielding the transmitted vector $\vec{t}^o$. Note that all operations are carried out in (mod 2) arithmetic. Upon sending $\vec{t}^o$ through the BSC with noise level p, the vector $\vec{r}=\vec{t}^o+\vec{n}^o$ is received, where $\vec{n}^o$ is the true noise.

Decoding is carried out by multiplying $\vec{r}$ by $\mathbf{A}$ to produce the syndrome vector $\vec{z}=\mathbf{A}\vec{r}$ $(=\mathbf{A}\vec{n}^o$, since $\mathbf{A}\mathbf{G}^T=\mathbf{0})$. In order to reconstruct the original message $\vec{s}^o$, one has to obtain an estimate $\vec{n}$ for the true noise $\vec{n}^o$. First we select all $\vec{n}$ that satisfy the parity checks $\mathbf{A}\vec{n}=\mathbf{A}\vec{n}^o$,

$$\mathcal{I}_{\mathrm{pc}}(\mathbf{A},\vec{n}^o)\equiv\{\vec{n}|\mathbf{A}\vec{n}=\vec{z}\},$$

and

$$\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{n}^o)\equiv\{\vec{n}\in\mathcal{I}_{\mathrm{pc}}(\mathbf{A},\vec{n}^o)|\vec{n}\neq\vec{n}^o\},\tag{1}$$

the (restricted) parity check set.

The second type of error-correcting code that we focus on here is the MN code [5]. An MN code is a low-density parity check code defined by a binary $M\times(N+M)$ matrix $\mathbf{A}=[\mathbf{C}_s|\mathbf{C}_n]$, concatenating two very sparse matrices known to both sender and receiver, with the $M\times M$ matrix $\mathbf{C}_n$ being invertible. The $M\times N$ matrix $\mathbf{C}_s$ has K nonzero elements per row and C per column, while $\mathbf{C}_n$ has L nonzero elements per row and column. The code rate is given by $R=K/C=N/M$. Encoding refers to multiplying the original message $\vec{s}^o$ by the $(M\times N)$ dense generator matrix $\mathbf{G}=\mathbf{C}_n^{-1}\mathbf{C}_s$, yielding the transmitted vector $\vec{t}^o$. Note that all operations are carried out in (mod 2) arithmetic. Upon sending $\vec{t}^o$ through the BSC with noise level p, the vector $\vec{r}=\vec{t}^o+\vec{n}^o$ is received, where $\vec{n}^o$ is the true noise.

Decoding is carried out by multiplying $\vec{r}$ by $\mathbf{C}_n$ to produce the syndrome vector $\vec{z}=\mathbf{C}_s\vec{s}^o+\mathbf{C}_n\vec{n}^o\equiv\mathbf{A}\vec{c}^o$, where $\vec{c}$ is the concatenated vector $(\vec{s},\vec{n})$. In order to reconstruct the original message $\vec{s}^o$, one has to obtain estimates $\vec{c}$ for the true signal and noise $\vec{c}^o$. First we select all combinations of signal and noise $\vec{c}$ that satisfy the parity checks $\mathbf{A}\vec{c}=\mathbf{A}\vec{c}^o$,

$$\mathcal{I}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)\equiv\{\vec{c}|\mathbf{A}\vec{c}=\vec{z}\},$$

and

$$\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)\equiv\{\vec{c}\in\mathcal{I}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)|\vec{c}\neq\vec{c}^o\},\tag{2}$$

the (restricted) parity check set.

To unify notation for Gallager and MN codes, we will adopt the notation $\vec{c}^o$ for the original noise (and signal) vector, and $\vec{c}$ for the estimate of the noise (and signal) vector. Any general decoding scheme then consists of selecting a vector $\vec{c}^*$ from $\mathcal{I}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)$, on the basis of some noise (and signal) statistics criterion. Upon successful decoding $\vec{c}^o$ will be selected, while a decoding error is declared when a vector $\vec{c}^*\in\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)$ is selected. For each decoding scheme, the average *block error probability* [16]

$$P_e(p_s,p)=\langle\Delta(\text{a vector }\ \vec{c}\in\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)\ \text{ is selected})\rangle_{\mathbf{A},\vec{c}^o}\tag{3}$$

can be defined as a measure of error-correcting ability for a given code ensemble, where $\Delta(\cdot)$ is an indicator function returning 1 if the proposition of the argument is true and 0, otherwise. For BSC, only the number of nonzero components characterizes the statistics of the noise. On the other hand, the signal bits, in general, have an equal probability for being 0 and 1 (i.e., $p_s=\frac{1}{2}$), which implies that they have no useful prior information for the estimation. In the following, we therefore focus on decoding schemes based on the weight of a vector which is the average sum of the noise components $w(\vec{c})\equiv 1/M\Sigma_{j=1}^{M}n_j$. To obtain the error probability, one averages the indicator function over all $\vec{c}^o$ vectors drawn from some distribution and the code ensemble $\mathbf{A}$ as denoted by $\langle\cdot\rangle_{\mathbf{A},\vec{c}^o}$.

Unfortunately, carrying out averages over the indicator function is difficult. Therefore, the error probability (3) is usually upper bounded by averaging over the *number* of vectors $\vec{n}$ obeying a certain condition on the weight $w(\vec{n})$, which characterizes the employed decoding scheme. Alternatively, one can find the average number of vectors with a given weight value w from which one can construct a complete weight distribution of noise vectors $\vec{n}$ in $\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)$. From this distribution one can, in principle, calculate a bound for P_e and derive critical noise values above which successful decoding cannot be carried out.

A natural and direct measure for the average number of states is the entropy of a system under the restrictions described above, which can be calculated via the methods of statistical physics.

It was previously shown (see, e.g., Ref. [9] for technical details) that this problem can be cast into a statistical mechanics formulation, by replacing the field $(\{0,1\},+\mathrm{mod}\ (2))$ by $(\{1,-1\},\times)$, and by adapting the parity checks correspondingly. The statistics of a noise vector $\vec{n}$ is now described by its magnetization $m(\vec{n})\equiv 1/M\Sigma_{j=1}^{M}n_j$, $(m(\vec{n})\in[1,-1])$, which is inversely linked to the vector weight in the $[0,1]$ representation. Similarly, the statistics of a signal vector $\vec{s}$ is now described by its magnetization $m_s(\vec{s})\equiv 1/M\Sigma_{j=1}^{M}s_j$, $(m_s(\vec{s})\in[1,-1])$. With this in mind, we introduce the conditioned magnetization enumerator, for a given code and noise, measuring the noise vector magnetization distribution in $\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{n}^o)$,

$$\mathcal{M}_{\mathbf{A},\vec{n}^o}(m)\equiv\frac{1}{M}\ln[\mathrm{Tr}_{\vec{n}\in\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{n}^o)}\delta(m(\vec{n})-m)].\tag{4}$$

To obtain the *magnetization enumerator* $\mathcal{M}(m)$,

$$\mathcal{M}(m)=\langle\mathcal{M}_{\mathbf{A},\vec{c}^o}(m)\rangle_{\mathbf{A},\vec{c}^o},\tag{5}$$

which is the entropy of the noise vectors in $\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{n}^0)$ with a given m, one carries out uniform explicit averages over all codes $\mathbf{A}$ with given parameters K,C (and L), and the weighted average over all possible noise vectors generated by the BSC, (and all possible signal vectors), i.e.,

$$P(\vec{n}^o)=\prod_j^M [(1-p)\,\delta(n_j^o-1)+p\,\delta(n_j^o+1)], \tag{6}$$

$$P(\vec{s}^o)=\prod_j^N [(1-p_s)\,\delta(s_j^o-1)+p_s\,\delta(s_j^o+1)], \tag{7}$$

with here $p_s=\frac{1}{2}$. It is important to note that, in calculating the entropy, the average quantity of interest is the magnetization enumerator rather than the actual number of states. For physicists, this is the natural way to carry out the averages for three main reasons. (a) The entropy obtained in this way is believed to be *self-averaging*, i.e., its average value (over the disorder) coincides with its *typical* value. (b) This quantity is *extensive* and grows linearly with the system size. (c) This averaging distinguishes between *annealed* variables that are averaged or summed for a given set of *quenched* variables that are averaged over later on. In this particular case, summation over all $\vec{c}$ vectors is carried for a *fixed* choice of code $\mathbf{A}$ and vector $\vec{c}^o$; averages over these variables are carried out at the next level.

One should point out that in somewhat similar calculations, we showed that this method of carrying out the averages provides more accurate results in comparison to averaging over both sets of variables simultaneously [14].

A positive magnetization enumerator, $\mathcal{M}(m)>0$ indicates that there is an exponential number of solutions (in M) with magnetization m, for typically chosen $\mathbf{A}$ and $\vec{c}^o$, while $\mathcal{M}(m)\to 0$ indicates that this number vanishes as $M\to\infty$ (note that negative entropy is unphysical in discrete systems).

Another important indicator for successful decoding is the overlap ω between the selected estimate $\vec{n}^*$, and the true noise $\vec{n}^o$: $\omega(\vec{n},\vec{n}^o)\equiv 1/M\Sigma_{j=1}^M n_j n_j^o$, $(\omega(\vec{n},\vec{n}^o)\in[-1,1])$, with $\omega=1$ for successful (perfect) decoding. However, this quantity cannot be used for decoding as $\vec{n}^o$ is unknown to the receiver. The (code and noise dependent) noise overlap enumerator is now defined as

$$\mathcal{W}_{\mathbf{A},\vec{c}^o}(\omega)\equiv\frac{1}{M}\ln[\mathrm{Tr}_{\vec{c}\in\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)}\delta(\omega(\vec{n},\vec{n}^o)-\omega)], \tag{8}$$

and the average quantity being

$$\mathcal{W}(\omega)=\langle\mathcal{W}_{\mathbf{A},\vec{c}^o}(\omega)\rangle_{\mathbf{A},\vec{c}^o}. \tag{9}$$

This measure is directly linked to the *weight enumerator* [3], although according to our notation, averages are carried out distinguishing between annealed and quenched variables unlike the common definition in the IT literature. However, as we will show below, the two types of averages provide identical results *in this particular case*.

Similarly, for MN codes one defines the signal magnetization and weight enumerators as

$$\mathcal{M}_s(m_s)\equiv\frac{1}{N}\langle\ln[\mathrm{Tr}_{\vec{c}\in\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)}\delta(m(\vec{s})-m_s)]\rangle_{\mathbf{A},\vec{c}^o}, \tag{10}$$

$$\mathcal{W}_s(\omega_s)\equiv\frac{1}{N}\langle\ln[\mathrm{Tr}_{\vec{c}\in\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^o)}\delta(\omega(\vec{s},\vec{s}^o)-\omega_s)]\rangle_{\mathbf{A},\vec{c}^o}. \tag{11}$$

In what follows, we perform all calculations as if both m and ω (and m_s and ω_s for MN codes), are constrained to particular values. As we will show, omitting a constraint in the final expressions can then easily be done by assigning the zero value to the corresponding Lagrange multiplier.

III. THE STATISTICAL PHYSICS APPROACH

Quantities of the type $\mathcal{Q}(c)=\langle\mathcal{Q}_y(c)\rangle_y$, with $\mathcal{Q}_y(c)=1/M\ln[\mathcal{Z}_y(c)]$ and $\mathcal{Z}_y(c)\equiv\mathrm{Tr}_x\delta(c(x,y)-Mc)$, are very common in the SP of disordered systems; the macroscopic order parameter $c(x,y)$ is fixed to a specific value and may depend both on the disorder y and on the microscopic variables x. Although we will not prove this here, such a quantity is generally believed to be *self-averaging* in the large system limit, i.e., obeying a probability distribution $P(\mathcal{Q}_y(c))=\delta(\mathcal{Q}_y(c)-\mathcal{Q}(c))$. The direct calculation of $\mathcal{Q}(c)$ is known as a *quenched* average over the disorder, but is typically hard to carry out and requires using the replica method [8]. The replica method makes use of the identity $\langle\ln\mathcal{Z}\rangle=\langle\lim_{\mathrm{n}\to 0}[\mathcal{Z}^{\mathrm{n}}-1]/\mathrm{n}\rangle$, by calculating averages over a product of partition function replicas. Employing assumptions about replica symmetries and analytically continuing the variable n to zero, one obtains solutions that enable one to determine the state of the system.

To simplify the calculation, one often employs the so-called *annealed* approximation, which consists of performing an average over $\mathcal{Q}_y(c)$ first, followed by the logarithm operation. This avoids the replica method and provides (through the convexity of the logarithm function) an upper bound to the quenched quantity,

$$\begin{aligned} Q_a(c)&\equiv\frac{1}{M}\ln[\langle\mathcal{Z}_y(c)\rangle_y]\\ &\geqslant Q_q(c)\\ &\equiv\frac{1}{M}\langle\ln[\mathcal{Z}_y(c)]\rangle_y\\ &=\lim_{\mathrm{n}\to 0}\frac{\langle\mathcal{Z}_y^{\mathrm{n}}(c)\rangle_y-1}{\mathrm{n}M}. \end{aligned} \tag{12}$$

The technical details of the calculation are similar to those in Ref. [9]. It turns out that it is useful to perform the gauge transformation $c_j\to c_j c_j^o$, such that the averages over the code $\mathbf{A}$ and noise/signal $\vec{c}^o$ can be separated, $\mathcal{W}_{\mathbf{A},\vec{c}^o}$ becomes independent of $\vec{c}^o$, leading to an equality between the quenched and annealed results, $\mathcal{W}(m)=\mathcal{M}_a(m)|_{p=0}=\mathcal{M}_q(m)|_{p=0}$. For any finite noise value p one should multiply $\exp[\mathcal{W}(\omega)]$ by the probability that a state obeys all parity checks $\exp[-\mathcal{K}(\omega,p)]$ given an overlap ω and a noise level p [3]. In calculating $\mathcal{W}(\omega)$ and $\mathcal{M}_{a/q}(m)$, the δ functions fixing m and ω, are enforced by introducing Lagrange multipliers $\hat{m}$ and $\hat{\omega}$.

Carrying out the averages explicitly, one then employs the saddle point method to extremize the averaged quantity with respect to the parameters introduced while carrying out the calculation. These lead, in both quenched and annealed calculations, to a set of saddle point equations that are solved either analytically or numerically to obtain the final expression for the averaged quantity (entropy).

The final expressions for the annealed entropy per noise degree of freedom for Gallager codes, under both overlap (ω) and magnetization (m) constraints, are of the form

$$\mathcal{Q}_a = -\frac{C}{K}\{(\ln(2)+(K-1)\ln[1+c_1^K]\} + \ln\langle \mathrm{Tr}_{n=\pm 1}\exp[n(\hat{\omega}+\hat{m}n^o)](1+nc_1^{K-1})^C\rangle_{n^o} - (\hat{\omega}\omega+\hat{m}m), \tag{13}$$

where the average cavity magnetization c_1 has to be obtained from the saddle point equation $\partial\mathcal{Q}_a/\partial c_1=0$. Similarly, the final expression in the quenched calculation, employing the simplest replica symmetry assumption [8], is of the form

$$\mathcal{Q}_q = -C\int dx d\hat{x}\,\pi(x)\hat{\pi}(\hat{x})\ln[1+x\hat{x}] + \frac{C}{K}\int\left\{\prod_{k=1}^{K} dx_k\pi(x_k)\right\}\ln\left[\frac{1}{2}\left(1+\prod_{k=1}^{K}x_k\right)\right] + \int\left\{\prod_{c=1}^{C} d\hat{x}_c\hat{\pi}(\hat{x}_c)\right\}\left\langle \ln\left[\mathrm{Tr}_{n=\pm 1}\exp(n(\hat{\omega}+\hat{m}n^o))\prod_{c=1}^{C}(1+n\hat{x}_c)\right]\right\rangle_{n^o} - (\hat{\omega}\omega+\hat{m}m). \tag{14}$$

The probability distributions $\pi(x)$ and $\hat{\pi}(\hat{x})$ emerge from the calculation; the former represents a probability distribution with respect to the noise vector local magnetization [15], while the latter relates to a field of conjugate variables that emerge from the introduction of δ functions while carrying out the averages (for details see Ref. [9]). Their explicit forms are obtained from the functional saddle point equations $\delta\mathcal{Q}_q/\delta\pi(x)$, $\delta\mathcal{Q}_q/\delta\hat{\pi}(\hat{x})=0$, and all integrals are from -1 to 1.

The final expressions for the annealed entropy per noise degree of freedom for MN codes, under both signal and noise overlap (ω,ω_s) and magnetization (m,m_s) constraints, are of the form

$$\mathcal{Q}_a = -\{\ln(2)+(K+L-1)\ln[1+c_1^K d_1^L]\} - R(\hat{m}_s m_s+\hat{\omega}_s\omega_s)-(\hat{m}m+\hat{\omega}\omega) + R\ln\langle \mathrm{Tr}_{s=\pm 1}\exp[s(\hat{\omega}_s+\hat{m}_s s^o)](1+s\hat{c}_1)^C\rangle_{s^o} + \ln\langle \mathrm{Tr}_{n=\pm 1}\exp[n(\hat{\omega}+\hat{m}n^o)](1+n\hat{d}_1)^L\rangle_{n^o}, \tag{15}$$

where c_1, d_1 have to be obtained from the saddle point equations $\partial\mathcal{Q}_a/\partial c_1, \partial\mathcal{Q}_a/\partial d_1=0$. Similarly, the final expression in the quenched calculation, employing the simplest replica symmetry assumption [8], is of the form

$$\mathcal{Q}_q = \int\prod_{k=1}^{K} dx_k\;\pi(x_k)\prod_{l=1}^{L} dy_l\rho(y_l)\ln\left[\frac{1}{2}\left(1+\prod_{k=1}^{K}x_k\prod_{l=1}^{L}y_l\right)\right] - R(\hat{m}_s m_s+\hat{\omega}_s\omega_s)-(\hat{m}m+\hat{\omega}\omega) - K\int dx d\hat{x}\,\pi(x)\hat{\pi}(\hat{x})\ln[1+x\hat{x}] + R\int\prod_{c=1}^{C} d\hat{x}_c\hat{\pi}(\hat{x}_c)\left\langle \ln\left[\mathrm{Tr}_{s=\pm 1}\exp(s(\hat{\omega}_s+\hat{m}_s s^o))\prod_{c=1}^{C}(1+s\hat{x}_c)\right]\right\rangle_{s^o} - L\int dy d\hat{y}\,\rho(y)\hat{\rho}(\hat{y})\ln[1+y\hat{y}] + \int\prod_{l=1}^{L} d\hat{y}_l\hat{\rho}(\hat{y}_l)\left\langle \ln\left[\mathrm{Tr}_{n=\pm 1}\exp(n(\hat{\omega}+\hat{m}n^o))\prod_{l=1}^{L}(1+n\hat{y}_l)\right]\right\rangle_{n^o}. \tag{16}$$

The probability distributions $\pi(x),\rho(y)$ and $\hat{\pi}(\hat{x}),\hat{\rho}(\hat{y})$ emerge from the calculation; the former represent probability distributions with respect to the signal/noise vector local magnetizations [15], while the latter relate to fields of conjugate variables that emerge from the introduction of δ functions while carrying out the averages (for details see Ref. [9]). Their explicit forms are obtained from the functional saddle point equations $\delta\mathcal{Q}_q/\delta\pi(x)$, $\delta\mathcal{Q}_q/\delta\hat{\pi}(\hat{x})$, $\delta\mathcal{Q}_q/\delta\rho(y)$, $\delta\mathcal{Q}_q/\delta\hat{\rho}(\hat{y})=0$, and all integrals are from -1 to 1.

Enforcing a δ function corresponds to taking $\hat{\omega},\hat{m},\hat{\omega}_s,\hat{m}_s$ such that $\partial\mathcal{Q}_{a/q}/\partial\hat{\omega}$, $\partial\mathcal{Q}_{a/q}/\partial\hat{m}$, $\partial\mathcal{Q}_{a/q}/\partial\hat{\omega}_s$, $\partial\mathcal{Q}_{a/q}/\partial\hat{m}_s=0$, while not enforcing it corresponds to putting

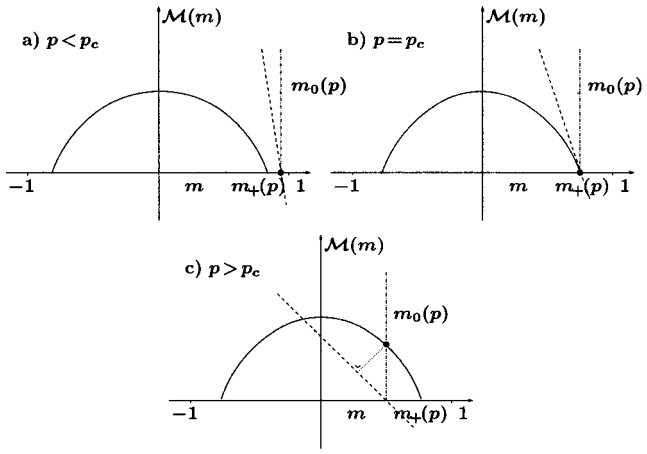


FIG. 1. The qualitative picture of $\mathcal{M}(m)\geqslant 0$ (solid lines) for different values of p. For MAP, MPM, and typical set decoding, only the relative values of $m_+(p)$ and $m_0(p)$ determine the critical noise level. Dashed lines correspond to the energy contribution of $-\beta F$ at Nishimori's condition ($\beta=1$). The states with the lowest free energy are indicated by a point ●. (a) Subcritical noise levels $p<p_c$, where $m_+(p)<m_0(p)$, there are no solutions with higher magnetization than $m_0(p)$, and the correct solution has the lowest free energy. (b) Critical noise level $p=p_c$, where $m_+(p)=m_0(p)$. The minimum of the free energy of the suboptimal solutions is equal to that of the correct solution at Nishimori's condition. (c) Overcritical noise levels $p>p_c$ where many solutions have a higher magnetization than the true typical one. The minimum of the free energy of the suboptimal solutions is lower than that of the correct solution.

$\hat{\omega},\hat{m},\hat{\omega}_s,\hat{m}_s$ equal to 0. Since ω,m,ω_s,m_s, follow from $\partial\mathcal{Q}_{a/q}/\partial\hat{\omega},\partial\mathcal{Q}_{a/q}/\partial\hat{m},\partial\mathcal{Q}_{a/q}/\partial\hat{\omega}_s,\partial\mathcal{Q}_{a/q}/\partial\hat{m}_s=0$, all the relevant quantities can be recovered with appropriate choices of $\hat{\omega},\hat{m},\hat{\omega}_s,\hat{m}_s$.

IV. QUALITATIVE PICTURE

We now discuss the qualitative behavior of $\mathcal{M}(m)$, and the interpretation of the various decoding schemes. To obtain separate results for $\mathcal{M}(m)$ and $\mathcal{W}(m)$ we calculate the results of Eqs. (13) and (14) [and Eqs. (15) and (16)], corresponding to the annealed and quenched cases, respectively, setting $\hat{\omega}=0$ to obtain $\mathcal{M}(m)$ and $\hat{m}=0$ to obtain $\mathcal{W}(\omega)$ (that becomes $\mathcal{M}(m)|_{p=0}$ after gauging). In Fig. 1, we have qualitatively plotted the resulting function $\mathcal{M}(m)$ for relevant values of p. $\mathcal{M}(m)$ (solid line) only takes positive values in the interval $[m_-(p),m_+(p)]$; for even K, $\mathcal{M}(m)$ is an even function of m and $m_-(p)=-m_+(p)$. The maximum value of $\mathcal{M}(m)$ is always $(1-R)\ln(2)$ for Gallager codes, and $R\ln(2)$ for MN codes. The true noise $\vec{n}^o$ has (with probability 1) the typical magnetization of the BSC: $m(\vec{n}^o)=m_0(p)=1-2p$ (dashed-dotted line).

The various decoding schemes can be summarized as follows.

(1) *Maximum likelihood (MAP) decoding* minimizes the block error probability [16] and consists of selecting the $\vec{n}$ from $\mathcal{T}_{\rm pc}(\mathbf{A},\vec{n}^0)$ with the highest magnetization. Since the probability of error below $m_+(p)$ vanishes, $P(\exists\vec{n}\in\mathcal{T}^{\rm r}_{\rm pc}:m(\vec{n})>m_+(p))=0$, and since $P(m(\vec{n}^o)=m_0(p))=1$, the critical noise level p_c is determined by the condition $m_+(p_c)=m_0(p_c)$. The selection process is explained in Figs. 1(a–c).

(2) *Typical pairs decoding* is based on randomly selecting a $\vec{n}$ from $\mathcal{T}_{\rm pc}$ with $m(\vec{n})=m_0(p)$ [3]; an error is declared when $\vec{n}^0$ is not the only element of $\mathcal{T}_{\rm pc}$. For the same reason as above, the critical noise level p_c is determined by the condition $m_+(p_c)=m_0(p_c)$.

(3) *Finite temperature (MPM) decoding.* An energy $-Fm(\vec{n})$ (with $F=\frac{1}{2}\ln[1-p/p]$) according to Nishimori's condition (corresponding to the selection of an accurate prior within the Bayesian framework) is attributed to each $\vec{n}\in\mathcal{T}_{\rm pc}$, and a solution is chosen from those with the magnetization that minimizes the free energy [9]. This procedure is known to minimize the *bit error probability* [16]. Using the thermodynamic relation $\mathcal{F}=\mathcal{U}-1/\beta\mathcal{S}$, β being the inverse temperature (Nishimori's condition corresponds to setting $\beta=1$), the free energy of the suboptimal solutions is given by $\mathcal{F}(m)=-Fm-(1/\beta)\mathcal{M}(m)$ [for $\mathcal{M}(m)\geqslant 0$], while that of the correct solution is given by $-Fm_0(p)$ (its entropy being 0). The selection process is explained graphically in Figs. 1(a)–1(c). The free energy differences between suboptimal solutions relative to that of the correct solution in the current plots are given by the orthogonal distance between $\mathcal{M}(m)$ and the line with slope $-\beta F$ through the point $(m_0(p),0)$. Solutions with a magnetization m for which $\mathcal{M}(m)$ lies above this line, have a lower free energy, while those for which $\mathcal{M}(m)$ lies below, have a higher free energy. Since negative entropy values are unphysical in discrete systems, only suboptimal solutions with $\mathcal{M}(m)\geqslant 0$ are considered. The lowest p value for which there are suboptimal solutions with a free energy equal to $-Fm_0(p)$ is the critical noise level p_c for MPM decoding. In fact, using the convexity of $\mathcal{M}(m)$ and Nishimori's condition, one can show that the slope $\partial\mathcal{M}(m)/\partial m>-\beta F$ for any value $m<m_o(p)$ and any p, and equals $-\beta F$ only at $m=m_o(p)$; therefore, the critical noise level for MPM decoding $p=p_c$ is identical to that of MAP, in agreement with results obtained in the information theory community [17].

The statistical physics interpretation of finite temperature decoding corresponds to making the specific choice for the Lagrange multiplier $\hat{m}=\beta F$ and considering the free energy instead of the entropy. In earlier work on MPM decoding in the SP framework [9], negative entropy values were treated by adopting different replica symmetry assumptions, which effectively result in changing the inverse temperature, i.e., the Lagrange multiplier $\hat{m}$. This effectively sets $m=m_+(p)$, i.e., to the highest value with non-negative entropy. The suboptimal states with the lowest free energy are then those with $m=m_+(p)$.

The central point in all decoding schemes is to select the correct solution only on the basis of its magnetization. As long as there are no suboptimal solutions with the same magnetization, this is, in principle, possible. As shown here, all

three decoding schemes discussed above manage to do so. To find whether at a given p there exists a gap between the magnetization of the correct solution and that of the nearest suboptimal solution, just requires plotting $\mathcal{M}(m)(>0)$ and $m_0(p)$, thus allowing a graphical determination of p_c. Since MPM decoding is done at Nishimori's temperature, the simplest replica symmetry assumption is sufficient to describe the thermodynamically dominant state [8]. At p_c the states with $m_+(p_c)=m_0(p_c)$ are thermodynamically dominant, and the p_c values that we obtain under this assumption are exact.

V. MN CODES—AN ALTERNATIVE VIEW

For MN codes there is a way to obtain the *exact* expression for $\mathcal{M}$, in the case of unbiased messages, by employing a single highly plausible assumption. We first note that every parity check bit $z_{\langle\rangle}=s^o_{i_1}\cdots s^o_{i_K}n^o_{j_1}\cdots n^o_{j_L}$ is made up of a combination of K unbiased (i.e., $p_s=\frac{1}{2}$) signal bits, and L biased (i.e., $p\neq\frac{1}{2}$) noise bits. As a result, every syndrome element $z_{\langle\rangle}$ is unbiased independently of the noise bit statistics. It is therefore plausible to assume that the noise bit statistics (i.e., p) have no influence on the distribution of the parity check bits $z_{\langle\rangle}$, and therefore on $\mathcal{M}$ (which only depends on the true noise through the $z_{\langle\rangle}$). If this assumption is satisfied, one can invoke Nishimori's condition to obtain an exact expression for $\mathcal{M}$.

Independently of the assumption, Nishimori's condition gives the following identity for the thermodynamically dominant state:

$$\left.\frac{\partial\mathcal{M}(m)}{\partial m}\right|_{m=m_o(p)} = -F(p) = -\frac{1}{2}\ln\left(\frac{1-p}{p}\right) = -\frac{1}{2}\ln\left(\frac{1+m_o}{1-m_o}\right). \tag{17}$$

Since states characterized by any magnetization value $m<m_0(p_t)$ will become dominant for an appropriately chosen value of p, and since we assume that $\mathcal{M}$ is independent of p, the identity

$$\frac{\partial\mathcal{M}(m)}{\partial m} = -\frac{1}{2}\ln\left(\frac{1+m}{1-m}\right) \tag{18}$$

must hold for any value of m. Furthermore, the maximum of $\mathcal{M}(m)$ is reached at $m=0$ with $\mathcal{M}(0)=R\ln(2)$, and we have that

$$\mathcal{M}(m) = \mathcal{M}(0) - \frac{1}{2}\int_0^m du\,\ln\left(\frac{1+u}{1-u}\right) = \ln(2)\left[R-1+H_2\left(\frac{1+m}{2}\right)\right], \tag{19}$$

where $H_2(p)$ is the binary entropy per bit for vectors with bias p. Hence, under this assumption, we do not only obtain the exact expression for $\mathcal{M}(m)$, but we see that the critical noise level p_c is given by $R=1-H_2(p_c)$, saturating Shannon's bound for this type of codes.

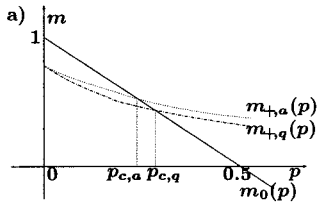
Unfortunately, the assumption cannot be verified easily without the replica method. To verify whether indeed $\partial\mathcal{M}(m)/\partial p=0$, we have to take the derivative of expression (16) (setting $\hat{\omega}=\hat{\omega}_s=\hat{m}_s=0$) with respect to p. It turns out that $\mathcal{M}$ is only independent of p, when $\rho(\hat{y})$ is an even function of $\hat{y}$, which, in turn, requires that $\rho(y)$ and $\pi(x)$ are even functions of their arguments. Numerical analysis shows that this is the case for any $K\geqslant 3$ or $K=2,\ L\geqslant 3$, while not so for $K=1$ or $K=L=2$. This result is consistent with those reported in Ref. [9], i.e., typical MN codes with $K\geqslant 3$ or $K=2,\ L\geqslant 3$ do saturate Shannon's bound, while those with $K=1$ and $K=L=2$ do not.

Intuitively this result can be understood in the following way. There are M parity check bits and only $N(<M)$ signal bits, such that parity check bits, although individually unbiased, are not uncorrelated. These correlations do seem to have an effect on $\mathcal{M}(m)$ for $K=1$ and $K=L=2$, while for $K\geqslant 3$ and $K=2,\ L\geqslant 3$ the signal bits seem to be "scrambled" enough in the parity checks for the correlations to be insignificant. Note that this argument does not hold for Gallager codes and MN codes with biased messages, where the parity check bits exclusively comprise biased bits, and are therefore biased themselves. They only become unbiased as $K\to\infty$ for Gallager codes (for which it was already reported in the literature [5] that such codes can saturate Shannon's bound), and for $K\to\infty$ or $L\to\infty$ for MN codes.

In fact, numerical analysis reveals that for $K\geqslant 3$ and for $K=2,\ L\geqslant 3$ we have that $\rho(\hat{y})=\delta(\hat{y})$, $\rho(y)=\delta(y)$, $\pi(x)=\delta(x)$ at least up to $m_+(p)=m_0(p_t)$ which is independent of p. This allows us to calculate $\mathcal{M}$ analytically from expression (16), and we recover, as expected, the exact expression (19).

For $K=1$ or $K=L=2$, as in the case of Gallager codes, one can only obtain $m_+(p)$ numerically. The results of this procedure are presented in the following section. Furthermore, for $K=1$ and for $K=L=2$, we find that spontaneously $m_s\neq 0$ for some values of $p<p_c$, when no restriction is enforced (i.e., for $\hat{m}_s=0$). This implies that one may improve the decoding performance by imposing the condition of unbiased signal (similar to the conditions for typical set decoding), i.e., by adjusting the Lagrange multiplier $\hat{m}_s$ such that $m_s=0$. Unfortunately, this only happens for values of p for which there is an exponential number of suboptimal solutions $\vec{c}\in\mathcal{I}^{\mathrm{r}}_{\mathrm{pc}}(\mathbf{A},\vec{c}^{\,o})$ with the same weight as $\vec{c}^{\,o}$, and imposing this constraint on the signal estimator only reduces this number, leaving it, nevertheless, exponential.

It was shown [10] that MN codes, in principle, contain sufficient information to saturate Shannon's bound for unbiased messages. For codes with $K=1$, or $K=L=2$, some of this information is wasted in a region where errorless decoding is impossible anyway, such that Shannon's bound is not saturated. For codes with $K\geqslant 3$, or $K=2,\ L\geqslant 3$, our analysis indicates that all information is used optimally, and that Shannon's bound can be theoretically saturated. Our argu-



b)

(K,C)	$(6,3)$	$(5,3)$	$(6,4)$	$(4,3)$
Code rate	1/2	2/5	1/3	1/4
IT ($\mathcal{W}_a$)	0.0915	0.129	0.170	0.205
SP	0.0990	0.136	0.173	0.209
$p_{c,a}$ ($\mathcal{M}_a$)	0.031	0.066	0.162	0.195
$p_{c,q}$ ($\mathcal{M}_q$)	0.0998	0.1365	0.1725	0.2095
Shannon p_t	0.109	0.145	0.174	0.214

FIG. 2. (a) Determining the critical noise levels $p_{c,a/q}$ based on the function $\mathcal{M}_{a/q}$ for Gallager codes and for MN codes with $K=1$ or $K=L=2$, a qualitative picture. (b) Comparison of different critical noise level (p_c) estimates for Gallager codes. Typical set decoding estimates have been obtained using the methods of IT [13], based on having a unique solution to $\mathcal{W}(m)=\mathcal{K}(m,p_c)$, as well as using the methods of SP [18]. The numerical precision is up to the last digit for the current method. Shannon's limit denotes the highest theoretically achievable critical noise level p_t for any code [1].

ment also explains the relative importance of the parameters K and L for the behavior of the code in comparison with C.

VI. CRITICAL NOISE LEVEL—RESULTS

Some general comments can be made about the critical MAP (or typical set) values obtained via the annealed and quenched calculations. Since $\mathcal{M}_q(m)\leqslant\mathcal{M}_a(m)$ [for given values of K, C (L), and p], we can derive the general inequality $p_{c,q}\geqslant p_{c,a}$. For all K, C (L) values that we have numerically analyzed, for both annealed and quenched cases, $m_+(p)$ is a nonincreasing function of p, and p_c is unique. The estimates of the critical noise levels $p_{c,a/q}$, based on $\mathcal{M}_{a/q}$, are obtained by numerically calculating $m_{c,a/q}(p)$, and by determining their intersection with $m_0(p)$. This is explained graphically in Fig. 2(a).

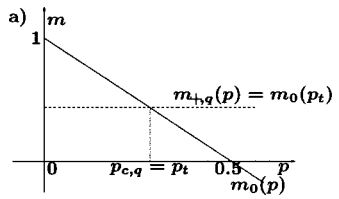
As the results for MPM decoding have already been presented elsewhere [11], we will now concentrate on the critical results p_c obtained for a typical set and MAP decoding for Gallager codes; these are presented in Fig. 2(b), showing the values of $p_{c,a/q}$ for various choices of K and C compared with those reported in the literature.

From Fig. 2(b) it is clear that the annealed approximation gives a much more pessimistic estimate for p_c. This is due to the fact that it overestimates $\mathcal{M}$ in the following way: $\mathcal{M}_a(m)$ describes the combined entropy of $\vec{n}$ and $\vec{n}^o$ as if $\vec{n}^o$ were thermal variables as well. Therefore, exponentially rare events for $\vec{n}^o$ [i.e., $m(\vec{n}^o)\neq m_0(p)$] still may carry positive entropy due to the addition of a positive entropy term from $\vec{n}$. In a separate study [18] these effects have been taken care of by the introduction of an extra exponent; this is not necessary in the current formalism as the quenched calculation automatically suppresses such contributions. The similarity between the results reported here and those obtained in Ref. [14] is not surprising as the equations obtained in quenched calculations are similar to those obtained by averaging the upper bound to the reliability exponent using a method presented originally by Gallager [4]. Numerical differences between the two sets of results are probably due to the higher numerical precision here.

We have also obtained the critical noise levels for some parameter choices in MN codes. We only present the quenched (exact) values, and compare them only with the highest theoretically achievable critical noise level p_t for any code [1], as we are not aware of values obtained with other methods in the literature. Note that although still strictly below p_t, the critical noise levels p_c for $K=L=2$ with increasing values of C rapidly approach p_t to within the current numerical precision (see Fig. 3).

VII. CONCLUSIONS

In this paper we have shown how both weight and magnetization enumerators can be calculated using the methods of statistical physics in the case of regular LDPC codes. We study the role played by the *magnetization enumerator* $\mathcal{M}(m)$ in determining the achievable critical noise level for various decoding schemes. The formalism based on the magnetization enumerator $\mathcal{M}$ offers a intuitively simple alternative to the weight enumerator formalism used in conjunction with typical pairs decoding in the IT literature [3,18]. The SP based analysis employs the replica method given the very



b)

(K,C,L)	$(1,3,2)$	$(2,6,2)$	$(2,3,2)$	$(3,9,3)$
Code rate	1/3	1/3	2/3	1/3
$p_{c,q}$ ($\mathcal{M}_q$)	0.15	$\lesssim$0.174	0.06	0.174
Shannon p_t	0.174	0.174	0.0615	0.174

FIG. 3. (a) Determining the critical noise levels $p_{c,q}$ based on the function $\mathcal{M}_q$ for MN codes with $K\geqslant 3$ or $K=2$, $L\geqslant 3$, a qualitative picture. (b) Comparison of different critical noise level ($p_{c,q}$) estimates for MN codes. The numerical precision is up to the last digit for the current method. Shannon's limit denotes the highest theoretically achievable critical noise level p_t for any code [1].

low critical values obtained by the annealed approximation calculation. Furthermore, the powerfull gauge theory as proposed by Nishimori [8], proves that the replica symmetric assumption is correct (at least at the critical noise level), and thus that the critical noise levels as obtained by our method are *exact*. Although we have concentrated here on the critical noise level for the BSC, other channel types as well as other quantities of interest can be treated using a similar formalism. The predictions for the critical noise level are more optimistic than those reported in the IT literature, and are up to numerical precision in agreement with those reported in Ref. [18]. We have also shown that the critical noise levels for typical pairs, MAP and MPM decoding must coincide, and we have provided an intuitive explanation to the difference between MAP and MPM decoding. Finally, an extension of this analysis to MN codes reveals the mechanism that allows them to saturate Shannon's limit for finite $K \geqslant 3$ and for $K=2$, $L \geqslant 3$ values (if impractical algorithms such as maximum likelihood are used). This result, which is consistent with previous SP based analyses [9], is considered as surprising in the IT community.

We believe that SP based analysis will provide more insight into the performance and characteristics of random LDPC codes, complementing the analysis provided by IT methods.

ACKNOWLEDGMENTS

Support by Grant-in-Aid Nos. 13680400 and 13780208 (Y.K.), The Royal Society, and Grant No. EPSRC-GR/N00562 (D.S./J.v.M.) is acknowledged.

[1] C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).
[2] A.J Viterbi and J.K. Omura, *Principles of Digital Communication and Coding* (McGraw-Hill, Singapore, 1979).
[3] S. Aji, H. Jin, A. Khandekar, D.J.C. MacKay, and R.J. McEliece, in *Codes, Systems and Graphical Models*, edited by B. Marcus and J. Rosenthal (Springer, New York, 2001), p. 195.
[4] R.G. Gallager, IRE Trans. Inf. Theory **8**, 21 (1962).
[5] D.J.C. MacKay, IEEE Trans. Inf. Theory **45**, 399 (1999).
[6] T. Richardson, A. Shokrollahi, and R. Urbanke, IEEE Trans. Inf. Theory **47**, 619 (2001).
[7] M. Mezard, G. Parisi, and M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[8] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing* (Oxford University Press, Oxford, UK, 2001).
[9] Y. Kabashima, T. Murayama, and D. Saad, Phys. Rev. Lett. **84**, 1355 (2000); Y. Kabashima, T. Murayama, D. Saad, and R. Vicente, Phys. Rev. E **62**, 1577 (2000).
[10] R. Vicente, D. Saad, and Y. Kabashima, in Advances in Imaging and Electron Physics, edited by P. Hawkes (Academic, New York, in press).
[11] R. Vicente, D. Saad, and Y. Kabashima, Europhys. Lett. **51**, 698 (2000).
[12] A. Montanari, Eur. Phys. J. B **23**, 121 (2001).
[13] R.G. Gallager, *Information Theory and Reliable Communication* (Wiley, New York, 1968).
[14] Y. Kabashima, N. Sazuka, K. Nakamura, and D. Saad, Phys. Rev. E **64**, 046113 (2001).
[15] M. Opper and D. Saad, *Advanced Mean Field Methods—Theory and Practice* (MIT Press, Cambridge, MA, 2001).
[16] Y. Iba, J. Phys. A **32**, 3875 (1999).
[17] D.J.C. MacKay, URL http://wol.ra.phy.cam.ac.uk/mackay/CodesTheory.html.
[18] Y. Kabashima, K. Nakamura, and J. van Mourik, e-print cond-mat/0106323, Phys. Rev. E (to be published).