Statistical mechanics of low-density parity-check codes

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **37** (2004) R1–R43 PII: S0305-4470(04)34343-X

TOPICAL REVIEW

Statistical mechanics of low-density parity-check codes

Yoshiyuki Kabashima[1] **and David Saad**[2]

[1] Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 2268502, Japan
[2] Neural Computing Research Group, Aston University, Birmingham B4 7ET, UK

Received 9 October 2003, in final form 23 October 2003
Published 28 January 2004
Online at stacks.iop.org/JPhysA/37/R1 (DOI: 10.1088/0305-4470/37/6/R01)

Abstract

We review recent theoretical progress on the statistical mechanics of error correcting codes, focusing on low-density parity-check (LDPC) codes in general, and on Gallager and MacKay–Neal codes in particular. By exploiting the relation between LDPC codes and Ising spin systems with multi-spin interactions, one can carry out a statistical mechanics based analysis that determines the practical and theoretical limitations of various code constructions, corresponding to dynamical and thermodynamical transitions, respectively, as well as the behaviour of error-exponents averaged over the corresponding code ensemble as a function of channel noise. We also contrast the results obtained using methods of statistical mechanics with those derived in the information theory literature, and show how these methods can be generalized to include other channel types and related communication problems.

PACS numbers: 02.50.−r, 75.10.Hk, 89.70.+c, 89.20.Kk

1. Introduction

1.1. Error correction

Electronic communication plays an important role in modern society and has a profound impact on the way we live. It appears in various forms and in a broad range of applications, from mobile and satellite communication to cable TV and the Internet.

Two features common to most modern digital communication systems are the need for efficient source and channel coding methods. Source coding relates to the compression of redundant information (e.g., pictures, music), even at the expense of fidelity (lossy compression); while channel coding relates to the introduction of some controlled redundancy prior to transmission in order to protect the information against corruption in a noisy transmission medium (e.g., deep space, atmosphere, optical fibres). In this review we mainly focus on error correction (channel coding) although we also mention applications of statistical mechanics analysis to source coding, multi-terminal communication channels, cryptography and other areas of information theory.

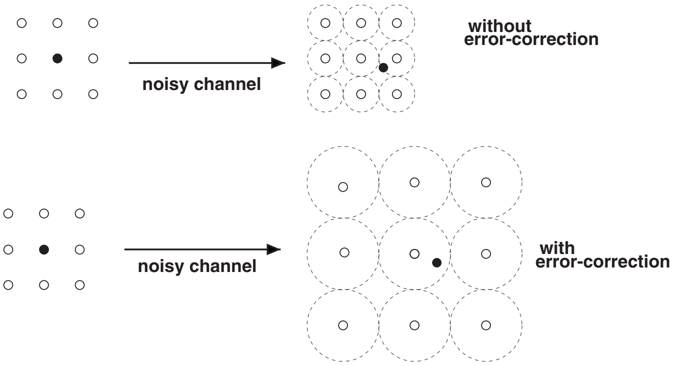


Figure 1. In the top figure we illustrate what happens when a word is transmitted without error correction. White circles represent possible word vectors; the black circle represents the word to be sent. The channel noise corrupts the original word, represented by a drift in the top right picture. The dashed circles indicate decision boundaries in the receiver; in the case depicted, the corruption leads to a transmission error. In the bottom figure we show qualitatively an error-correction mechanism. The redundant information changes the space geometry, increasing the distance between words. The same drift as in the top figure does not result in a transmission error.

In his 1948 papers Shannon [Sha48] proved general results on the limits of compression and error correction by setting up the framework for what is now known as information theory (IT). Shannon's channel coding theorem states that error-free communication is possible if some redundancy is added to the original message in the encoding process. A message encoded at rates R (message information content/code-word length) up to the channel capacity C_{channel} can be decoded with a probability of error that decays exponentially with the message length. Shannon's proof is non-constructive and assumes encoding with unstructured random codes and impractical decoding schemes (requiring a computing effort that grows non-polynomially with the codeword length) [CT91]. Finding practical codes capable of reaching the coding limits established by Shannon has been one of the central issues in coding theory ever since; and only recently, due to some ingenious code designs, are we within reach of closing the remaining gap to the bounds set by Shannon.

Figure 1 illustrates the problem of channel coding. On the top left of figure 1 we represent the space of words (a message is a sequence of words), each circle represents one sequence of binary bits. The word to be sent is represented by a black circle in the left-side figure. Corruption by noise in the channel is represented in the top right figure as a drift in the original word location. The circle around each word represents a decision boundary sphere for the particular word; any signal inside a certain decision region is recognized as representing the word at the centre of the sphere. In the case depicted in figure 1, the drift caused by noise places the received word within the decision boundary of another word vector, causing a transmission error. Error-correction codes are based on mapping the original space of words onto a higher dimensional space in a way that the typical distance between encoded words increases. The collection of all encoded words (codewords) constitutes a codebook. If the original space is transformed, the same drift shown in the top of figure 1 is insufficient to push the received signal outside the decision boundary of the transmitted codeword (bottom figure).

Good codes should be as short as possible, yet should clearly allow for a large number of codewords (for a large set of words), and decision spheres must be as large as possible (for large error-correction capability). The general coding problem consists of optimizing one of these conflicting requirements given the other two.

1.2. Low-density parity-check codes

For a long while, the best practical codes known were variants of Reed–Solomon codes which form the basis for most current technological standards (e.g., in deep-space communications [MS77, VO79]). The situation changed dramatically about a decade ago with the introduction of *Turbo codes* [BGT93]. These codes are composed of two convolutional codes working in parallel and show a practical performance close to Shannon's bound when decoded with iterative methods known as probability propagation [Pea88] or belief propagation; these iterative methods were first studied in the context of coding by Wiberg [Wib96] (excluding Gallager's original formulation [Gal62, Gal63]). The area experienced a second dramatic development when Gallager's low-density parity-check codes (LDPC) were rediscovered by MacKay and Neal in 1995 [MN95, Mac99]; this led to renewed activity in the general area of low-density parity-check codes [RU01a, RSU01, LMSS01] leading to the design of record breaking codes (e.g., [Chu00, Dav99, Dav98]) and greater understanding of their properties.

Gallager codes were first proposed in 1962 [Gal62, Gal63] and then were all but forgotten soon after due to computational limitations of the time and due to the success of convolutional codes. LDPC codes are much easier to understand and analyse than Turbo codes, and arguably represent the future of error correction. Throughout this review we concentrate on LDPC error correcting codes in general and *Gallager* and *MacKay–Neal codes* in particular.

1.3. Information theory and statistical mechanics of coding

The study of error-correcting codes is clearly one of the main topics in information theory. While the main properties of communication channels can be easily obtained from simple entropic considerations [CT91], the construction and analysis of practical codes, particularly LDPC codes of finite connectivity, is rather difficult. In most cases, practical and/or theoretical limitations are derived, in the infinite codeword limit, in the form of bounds as direct average properties are difficult to obtain.

The statistical mechanics of codes represents a completely different approach. By exploiting similarities between error-correcting codes and spin-glass models, as well as methods developed in the study of Ising spin systems, one carries out exact averages over code ensembles, possible messages and noise vectors to calculate the free energy of a given system; studying its properties one obtains exact results for their practical and theoretical limitations.

In section 2 we provide a general description of the communication channels studied and the notation used; in section 3 we briefly review several LDPC code constructions, followed by a more detailed review of recent statistical mechanics based analyses and their relation to analyses carried out in the information theory community (section 4). In section 5 we focus on analytical methods for obtaining the theoretical limitations of codes used in the IT literature and their equivalents in the statistical mechanics based approach; applications of LDPC codes to a range of other problems in information theory and cryptography will be reviewed in section 6 followed by a brief summary.

2. Communication channels

A general communication scenario is described in figure 2(*a*). It is based on encoding a K-dimensional message $\boldsymbol{s}$ to an N-dimensional codeword $\boldsymbol{t}$ which is then transmitted through a noisy communication channel. Codeword corruption during transmission can be described as a probabilistic process defined by the conditional probability $P(\boldsymbol{r}|\boldsymbol{t})$, where $\boldsymbol{t}$ and $\boldsymbol{r}$ represent

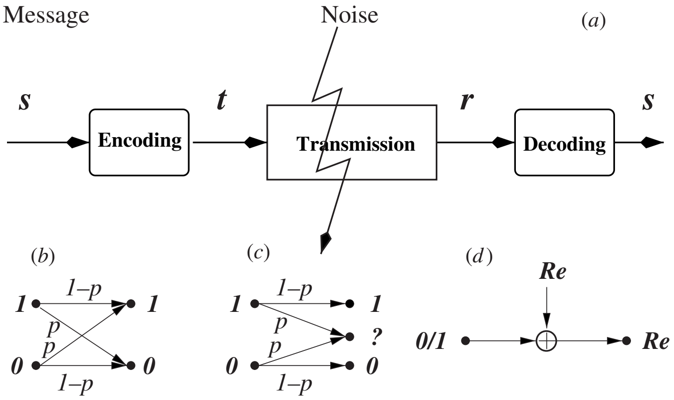


Figure 2. (*a*) Mathematical model for a communication system. (*b*) binary symmetric channel (BSC). (*c*) Binary erasure channel (BEC). (*d*) Real-valued symmetric channels (Gaussian–AWGN, Laplacian etc.).

transmitted and received messages, respectively. We assume no interference effects between codeword components, binary messages/codewords ({0, 1}) and a memoryless channel, so that $P(\boldsymbol{r} \mid \boldsymbol{t}) = \prod_{i=1}^{N} P(r_i \mid t_i)$. The received codeword $\boldsymbol{r}$ is then decoded to retrieve the original message $\boldsymbol{s}$. In this review, we will consider several channel types described schematically in figures 2(*b*)–(*d*), although other channels can also be considered and analysed using similar approaches. The differences between the various channels stem from the corruption probability $P(r_j \mid t_j)$. The binary symmetric channel (BSC), described schematically in figure 2(*b*), is defined by binary input and output alphabets and by the conditional probability

$$P(r \neq t \mid t) = p \qquad P(r = t \mid t) = 1 - p. \tag{1}$$

In the binary erasure channel (BEC) (figure 2(*c*)), binary codeword bits arrive uncorrupted with probability $1 - p$; no information is given in the case of corruption as indicated by the '?'; symbol. The conditional probability of a received bit being identical to the transmitted one is, therefore, $P(r = t \mid t) = 1 - p$. In the case of channels with real-valued noise, described in figure 2(*d*), binary transmitted codeword bits become real received values. Such communication channels are described by some conditional probability $P(r \mid t)$; which, for instance, in the case of a additive-white-Gaussian-noise channel (AWGN), takes the form

$$P(r \mid t) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(r-t)^2}{\sigma^2}\right) \tag{2}$$

where σ^2 represents the variance of the Gaussian noise.

The maximal information per bit that the channel can transport defines the *channel capacity* [CT91] and can be easily derived from entropic considerations; for perfect retrieval, the source vector binary entropy plus that of the noise vector must be smaller than the codebook entropy. Since all codewords may be used with equal probability, the latter (per symbol) equals the (base 2) logarithm of the alphabet size, i.e. 1 in the case of a binary alphabet {0, 1}. The entropy of any binary vector is calculated directly from the probability of having a value of 0/1. For instance, for the binary noise vector (1) the entropy per bit becomes

$$\mathsf{H}_2(p) = -p \log_2(p) - (1-p)\log_2(1-p) \tag{3}$$

and the BSC capacity is given by

$$C_{\mathrm{BSC}} = 1 - \mathsf{H}_2(p). \tag{4}$$

Similarly, for the BEC the channel capacity is

$$C_{\mathrm{BEC}} = 1 - p. \tag{5}$$

Channel capacity expressions for real-valued noisy channels are slightly more complex; for instance, Shannon's bound in the case of AWGN is given by

$$C_{\mathrm{AWGN}} = \tfrac{1}{2}\log_2(1 + \mathrm{SNR}) \tag{6}$$

where SNR is the signal to noise ratio, defined as the ratio of energy per bit of the source (squared amplitude) over the spectral density of the noise (variance). If one constrains the encoded bits to binary values $\{\pm 1\}$ (binary-input additive-white-Gaussian-noise channel—BIAWGNC) the capacity becomes

$$C_{\mathrm{BIAWGNC}} = \int \mathrm{d}r\, P(r \mid 1)\log_2 P(r \mid 1) - \int \mathrm{d}r\, P(r)\log_2 P(r) \tag{7}$$

where $P(r \mid t)$ is as in equation (2).

The analysis presented in this review focuses on the binary symmetric channel but can be easily extended to other channel types [KS99a, VSK99, TS03c, SvMS03, Mon01, FLMRT02] that are arguably of greater practical relevance [VO79, CT91].

3. Low-density parity-check codes

Parity check codes have been used in various error-correction mechanisms almost from the very beginning of the field. One of the most well-known parity check mechanisms is the Hamming code [CT91] and its generalization to the family of linear codes.

Most practical linear codes tend to offer a relatively low error protection for a given transmission ratio, far below the Gilbert–Varshamov distance [Var57, Gil52], bounding all binary linear codes. The performance improves as the number of elements summed in each check grows; however, the decoding process becomes computationally hard and unfeasible for a practical codeword length.

3.1. Gallager's code

LDPC codes were originally introduced by Gallager in 1962 [Gal62]. They rely on a sparse linear transformation of binary messages at the decoding stage, making it computationally feasible; while encoding relies on a dense matrix generated by the inverse of the sparse linear transformation. The significance of Gallager's discovery was not fully appreciated at the time due to the limited computing resources at the time as well as the increasing popularity of convolutional codes that require only a simple system of shift registers to operate effectively.

Gallager's code is defined by a binary matrix $H = [A \mid B]$, concatenating two very sparse matrices known to both sender and receiver, with B (of dimensionality $(N - K) \times (N - K)$) being invertible and A of dimensionality $(N - K) \times K$. The matrix H can be either random or structured, characterized by the number of non-zero elements per row/column. These numbers, which we denote as k and j, respectively, can be constants for all rows/columns (defining a *regular* code) or may vary from row to row (or column to column) giving rise to an *irregular* code.

Irregular codes show superior performance with respect to regular constructions [RU01a, RSU01, KS99b, KS00b, VSK00b] if they are constructed carefully. However, to simplify the presentation, we focus here on regular constructions; the generalization of the methods presented here to irregular constructions is straightforward [VSK02, VSK00b].

Encoding refers to the mapping of a K-dimensional binary vector $\boldsymbol{s} \in \{0, 1\}^K$ (original message) to N-dimensional codewords $\boldsymbol{t} \in \{0, 1\}^N$ $(N > K)$ by the linear product

$$\boldsymbol{t} = G^T \boldsymbol{s} \quad (\mathrm{mod}\, 2) \tag{8}$$

where all operations are performed in the field $\{0, 1\}$ and are indicated by $(\mathrm{mod}\, 2)$. The generator matrix is of the form

$$G = [I \mid B^{-1}A] \quad (\mathrm{mod}\, 2) \tag{9}$$

where I is the $K \times K$ identity matrix. By construction $HG^T = 0\, (\mathrm{mod}\, 2)$ and the first K bits of $\boldsymbol{t}$ correspond to the original message $\boldsymbol{s}$. Note that the generator matrix is dense and each transmitted parity-check carries information about $O(K)$ message bits.

In the case of unbiased messages, with equal bit probability of having the values 1 and 0, the code rate corresponds to the ratio of message to codeword bits $R = K/N$. Counting the number of unit elements in the matrix H one easily establishes the relation $j = (1 - K/N)k$, from which the code rate expression $R = (1 - j/k)$ can be derived. In the case of biased messages, one should replace the number of bits K by the logarithm (base 2) of the corresponding entropy.

To demonstrate the way in which Gallager's code is utilized we consider the BSC, where the encoded vector $\boldsymbol{t}$ is corrupted by a noise vector $\boldsymbol{n} \in \{0, 1\}^N$ with components independently drawn from

$$P(n) = (1 - p)\delta(n) + p\delta(n - 1). \tag{10}$$

The received vector takes the form

$$\boldsymbol{r} = G^T \boldsymbol{s} + \boldsymbol{n} \quad (\mathrm{mod}\, 2). \tag{11}$$

Decoding is carried out by multiplying the received message by the matrix H to produce the *syndrome* vector

$$\boldsymbol{z} = H\boldsymbol{r} = H\boldsymbol{n} \quad (\mathrm{mod}\, 2). \tag{12}$$

Decoding refers to finding an estimate of $\boldsymbol{n}$ knowing $\boldsymbol{z}$ and H; this, of course, enables one to obtain the original message vector $\boldsymbol{s}$ (the first K bits of $\boldsymbol{r} + \boldsymbol{n}\, (\mathrm{mod}\, 2)$). The following estimators may be employed in principle:

- *Maximum a posteriori (MAP)* is based on selecting the noise vector of the lowest weight (smallest number of '1's) that obeys all parity checks (12); this corresponds to mapping the received vector onto the *nearest* codeword. It also implies maximization of the posterior probability $P(\boldsymbol{n}|\boldsymbol{z}, H)$. The noise vector MAP estimator, which is also the maximum likelihood (ML) estimator of the codeword, minimizes the *block error probability* [Iba99] (i.e. of having *any* errors in a decoded message) but is computationally demanding and cannot be used in practice.
- *Marginal posterior maximizer (MPM)* is selecting the most probable *noise-bit* estimator, while marginalizing over all other bits (i.e. summing up over the probabilities of all other variables). This relies on choosing the right prior for the estimated noise vector bits; it has the property of minimizing the *bit error probability* [Iba99] (average error probability per bit) . MPM is in general equally difficult to MAP decoding. However, good approximation methods exist for codes that can be mapped onto sparse graphs, leading to successful decoding in a broad range of noise values.

In practice, decoding is carried out mainly by employing some message passing algorithm such as belief propagation (BP) [Pea88] (also known as probability propagation, Bayesian networks) and its variations.

Irregular Gallager codes decoded using BP offer the best performance to date; these results follow from the work of [RSU01, RSU01, RU01b].

3.2. Sourlas code

In 1989 Sourlas pointed to the relation between simple LDPC codes and spin-glass models [Sou89]. Although the codes presented by Sourlas are of limited practical relevance they made a significant contribution to establishing the links between statistical mechanics and information theory.

The code presented by Sourlas is strongly related to both Gallager and Mackay–Neal (MN) codes. It is based on a regular generator matrix G giving rise to a codeword in the form (11). The decoding problem can be mapped to known physical systems, Sourlas's original paper focuses on the Sherrington–Kirkpatrick [SK75, KS78] and random energy models [Der81, Saa98], where their performance can be analysed.

The results presented are of little practical significance since sparse generator matrices of the form presented (e.g., with two non-vanishing elements per row, $k = 2$) result in a non-vanishing error probability; while using dense generator matrices, which would *potentially* allow for a perfect retrieval of messages, is unfeasible due to decoding difficulties (in fact, decoding codes with $k \geqslant 3$ is already difficult).

3.3. MN code

MacKay and Neal introduced the MN codes in 1995 [MN95, Mac99], a variation on Gallager codes which they discovered independently, giving rise to renewed interest in LDPC codes.

MN codes are defined by two very sparse matrices; the main difference with respect to Gallager codes is that information on both noise and signal is incorporated in the syndrome vector. Both encoding and decoding follow a similar procedure as in (8)–(12) except that the generator and decoding matrices take a different form.

The generator matrix G is an $N \times K$ dense matrix defined by

$$G = B^{-1}A \pmod 2 \tag{13}$$

with B being an $N \times N$ binary invertible sparse matrix and A an $N \times K$ binary sparse matrix. Also MN codes come in both regular and irregular forms; again, for brevity, we concentrate here on regular codes, where the number of unit elements per row/column in A is k and j, respectively, and l in B (for both row/column).

Using communication through a BSC as an example, the transmitted vector $\boldsymbol{t}$ is then corrupted by a binary noise vector $\boldsymbol{n} \in \{0, 1\}^N$ as in (10) and the received vector takes the same form as in (11). Decoding is performed by matrix multiplication of the corrupted codeword by the matrix B, giving rise to the syndrome vector

$$\boldsymbol{z} = B\boldsymbol{r} = A\boldsymbol{s} + B\boldsymbol{r} \pmod 2. \tag{14}$$

Estimating the original message and noise vector from the syndrome $\boldsymbol{z}$ and matrices A and B is carried out in the same way as in Gallager codes.

Specific constructions of MN codes, especially those using Galois fields, rather than the basic binary representation, show very good performance [Dav99, Dav98].

3.4. Designing capacity approaching codes

The main breakthrough in the design of capacity approaching codes came with the work of Richardson and Urbanke [RSU01]. They analysed a BP-based decoding mechanism, by considering a macroscopic representation of the local fields, in the form of probability distributions. The method, termed *density evolution* (DE), is employed for analysing the decoding process and used to derive stability conditions which facilitate the design of capacity approaching codes. In fact, DE is similar to the Bethe approximation [MPV87] used in the study of diluted systems. The relation between BP, density evolution and the Bethe approximation has been pointed out in [KS98, VSK00a, YFW02] (see also section 4.4). Later on, Chung *et al* [CRU01] presented a Gaussian-based approximated DE and applied it to the design of capacity approaching codes.

Both DE and its Gaussian-based approximated version are aimed at designing irregular constructions, we will therefore not review them in detail, but rather point to the similarities between them and the statistical mechanics approach [VSK02].

3.5. Turbo codes

The exciting developments in the area of LDPC codes were preceded by the discovery of another family of capacity approaching codes—the Turbo codes [BGT93]. The introduction of Turbo codes created excitement in the information theory community as they represented a step increase in performance towards saturating Shannon's limit, with respect to previous record holders—Bose–Chaudhuri–Hocquenghem and Reed–Solomon codes [McEon].

Turbo code is a variant of recursive convolutional codes; the latter are based on shift registers (two in most cases, but more in general), used to generate codewords by a recursive convolution of message bits. Various structures can be used in general, although in most cases, the codeword comprises the original message segment and recursively convoluted segments of it. Decoding can be carried out in various ways, in conjunction with the convolution mechanism; for instance, by employing BP techniques for finding the most probable message bits [Fre98, FM98].

In the case of turbo codes two vectors, representing the original message and a permuted version of it, are used as inputs in a recursive convolutional procedure for generating the codeword. The decoding process exploits correlations between bits of the message vector and of the permuted vector, to obtain an estimate of the original message.

An additional advantage of turbo codes is that they can be easily implemented using simple electronic circuits (shift registers); the drawback is that they are difficult to analyse and systematically improve. Turbo codes were also analysed using methods of statistical mechanics [MS00, Mon00]. A brief description of the convolutional mechanics context can be found in [Nis01].

4. Statistical mechanics of coding

The link between error correcting codes and statistical mechanics was first pointed out by Sourlas [Sou89]. He mapped a simple parity check code onto spin-glass models [Sou89], focusing on the SK [SK75] and random energy models [Der81, Saa98] and showing that the latter can be viewed as an ideal code capable of saturating Shannon's bound at vanishing code rates (without taking into account practical decoding considerations).

A few papers relating spin-glass models and coding have been published since then and before the renewed interest in LDPC codes. Among them one should mention several studies of finite temperature decoding [Ruj93, Nis93, Sou94] and the analysis of convolutional codes via transfer-matrix methods and power series expansions [AL95].

The rediscovery of LDPC codes brought with it excitement also to the statistical mechanics community. After extending Sourlas's work to the case of finite code rates [KS99a, VSK99], regular and irregular MN [KMS00b, MKSV00, VSK00b, KMSV00] and Gallager [VSK00a, VSK01, Mon01, KSNS01, vMSK01, vMSK02, NKS01] codes have been studied using statistical mechanics, and a link between the two frameworks has been established

[KS98, VSK02, FLMRT02]. Insight gained from the statistical mechanics analysis also contributed to the design of highly efficient irregular codes [KS99b, KS00b, KS00a, VSK02].

The similarity between Ising spin models and LDPC codes stems from the formulation of the decoding problem. Employing the isomorphism between the additive Boolean group $(\{0, 1\}, \oplus)$ and the multiplicative binary group $(\{+1, -1\}, \times)$, whereby every addition in the Boolean group corresponds to a unique product in the binary group and vice versa, one can map the decoding problem to a Gibbs distribution by constructing an appropriate Hamiltonian.

The decoding problem depends on posteriors such as $P(\boldsymbol{\tau} \mid \boldsymbol{r})$, where $\boldsymbol{r}$ is the observation (received message or syndrome vector), and $\boldsymbol{\tau}$ is a candidate estimate of the unknown original message $\boldsymbol{s}$ (or alternatively a candidate noise vector from which an estimate of the noise can be obtained). Applying Bayes' theorem this posterior takes the form

$$P_{\alpha\gamma}(\boldsymbol{\tau} \mid \boldsymbol{r}) = \frac{1}{Z(\boldsymbol{r})} \exp[\ln P_\gamma(\boldsymbol{r} \mid \boldsymbol{\tau}) + \ln P_\alpha(\boldsymbol{\tau})] \tag{15}$$

where α and γ are hyper-parameters assumed to describe features such as the encoding scheme, source distribution and noise level. This form suggests the following family of Gibbs measures (β being the inverse temperature):

$$P_{\alpha\beta\gamma}(\boldsymbol{\tau} \mid \boldsymbol{r}) = \frac{1}{Z} \exp[-\beta \mathcal{H}_{\alpha\gamma}(\boldsymbol{\tau}; \boldsymbol{r})] \tag{16}$$

$$\mathcal{H}_{\alpha\gamma}(\boldsymbol{\tau}; \boldsymbol{r}) = -\ln P_\gamma(\boldsymbol{r} \mid \boldsymbol{\tau}) - \ln P_\alpha(\boldsymbol{\tau}). \tag{17}$$

The received corrupted codeword depends on the coding mechanism and channel noise, both of which represent the quenched disorder in the system.

The MAP estimator of $\boldsymbol{s}$ is clearly obtained at the ground state of the Hamiltonian, i.e. by the sign of thermal averages $\hat{s}_j^{\mathrm{MAP}} = \mathrm{sgn}(\langle \tau_j \rangle_{\beta\to\infty})$ at zero temperature.

The MPM estimator corresponds to the sign of thermal averages $\hat{s}_j^{\mathrm{MPM}} = \mathrm{sgn}(\langle \tau_j \rangle_{\beta=1})$ at a finite temperature, where true prior probability is assumed [Iba99]. This corresponds to using the Nishimori condition [Nis80, Nis93, Nis01, Ruj93]; and in the notation we use here to a temperature $\beta = 1$.

4.1. Gallager's code

To provide a more detailed description of the analysis we have to focus on a specific code and channel noise. We will explain the analysis for Gallager's code and the BSC; the analyses of the MN code and other channel types follow along the same lines.

A key point is the definition of an appropriate Hamiltonian; this can be done in various ways. We identify two main components in the Hamiltonians that are necessary for the analyses of all LDPC codes: a term that guarantees that all parity checks are satisfied, and a prior term that provides some statistical information on the dynamical variables ($\boldsymbol{\tau}$). In the case of a BSC, the Hamiltonian takes the form

$$\mathcal{H} = \sum_\mu \chi(z_\mu = [H\boldsymbol{\tau}]_\mu) - F \sum_{j=1}^{N} \tau_j. \tag{18}$$

The parity checks $\chi(z_\mu = [H\boldsymbol{\tau}]_\mu) = 0$ if parity check μ is obeyed by the vector $\boldsymbol{\tau}$ and $\chi(\cdot) = \infty$ otherwise; this corresponds to the parity checks (12). The coefficient $F = \frac{1}{2}\ln[(1-p)/p]$, in conjunction with the appropriate choice of temperature $\beta = 1$, corresponds to the correct prior assumption for the noise variables $\boldsymbol{\tau}$.

An explicit expression for $\chi(\cdot)$ in this case takes the form

$$\chi(z_\mu = [H\boldsymbol{\tau}]_\mu) = -\lim_{\gamma\to\infty} \gamma \sum_{\langle i_1\cdots i_k\rangle} \mathcal{D}_{\langle i_1\cdots i_k\rangle} \big(\mathcal{J}_{\langle i_1\cdots i_k\rangle} \tau_{i_1} \cdots \tau_{i_k} - 1\big) \tag{19}$$

where the tensor $\mathcal{J}$ denotes the uncorrupted syndrome (12) in the binary (±1) representation $\mathcal{J}_{\langle i_1,i_2\ldots i_K\rangle} = n_{i_1}n_{i_2}\ldots n_{i_k}$ (ordered indices) and the tensor $\mathcal{D}$ represents the connectivities of the matrix H; it takes the value 1 if the corresponding noise vector indices are chosen (i.e. all corresponding indices of the matrix H are 1) and 0 otherwise. For the time being we assume some fixed value for γ, but later on we will take the limit $\gamma \to \infty$ to obtain the desired properties of $\chi(\cdot)$.

To simplify the analysis and decouple the two quenched variables (true noise vector $\boldsymbol{n}$ and the parity check matrix H) we use the gauge transformation $\tau_i \mapsto \tau_i n_i$ and $\mathcal{J}_{\langle i_1\cdots i_k\rangle} \mapsto \mathcal{J}_{\langle i_1\cdots i_k\rangle} n_{i_1}\cdots n_{i_k} = 1$. This maps any general message to the case $n_i = 1\,\forall\, i$ (ferromagnetic configuration). We rewrite the Hamiltonian in the form:

$$\mathcal{H}_\gamma(\boldsymbol{\tau}) = -\gamma \sum_{\langle i_1\cdots i_k\rangle} \mathcal{D}_{\langle i_1\cdots i_k\rangle}\big(\tau_{i_1}\cdots\tau_{i_k} - 1\big) - F\sum_{i=1}^{N} n_i\tau_i. \tag{20}$$

Once the Hamiltonian has been defined one can calculate the free energy of the system and study emerging solutions for various choices of the parameters k, j and levels of channel noise.

Two main methods can be employed for carrying out the analysis, the replica method for diluted systems [KMS00b, MKSV00, FLMRT02] and the Bethe approximation [VSK99]. In all calculations carried out under the Nishimori condition, the dominant solution is known to be obtained under the replica symmetry (RS) assumption [NS01], providing similar results to those obtained by the Bethe approximation [VSK99].

4.1.1. Replica calculation. Analysing the typical performance of Gallager codes is based on similar studies of diluted systems [WS87a]. The aim is to compute the free energy:

$$\mathcal{F} = -\frac{1}{\beta}\lim_{N\to\infty}\frac{1}{N}\langle\ln\mathcal{Z}\rangle_{\mathcal{D},\boldsymbol{n}} \qquad \text{where} \quad \mathcal{Z} = \mathrm{Tr}_{\boldsymbol{\tau}}\exp(-\beta\mathcal{H}_\gamma(\boldsymbol{\tau};\boldsymbol{n})) \tag{21}$$

from which the typical macroscopic (thermodynamic) behaviour can be obtained using the Hamiltonian (20). Quenched averages are carried out over the connectivity tensor $\mathcal{D}$ and the true noise vector n under the following constraints: the connectivity tensor $\mathcal{D}_{\langle i_1\cdots i_k\rangle} \in \{0, 1\}$ is a random symmetric tensor with the properties:

$$\sum_{\langle i_1\cdots i_k\rangle} \mathcal{D}_{\langle i_1\cdots i_k\rangle} = N - K \qquad \sum_{\langle i_1=l,\ldots,i_k\rangle} \mathcal{D}_{\langle i_1=l,\ldots,i_k\rangle} = j \quad \forall l \tag{22}$$

corresponding to the selection of $N-K$ sets of indices. Noise vector bits n_i take the values $-1/1$ with probabilities $p/1-p$, respectively.

To carry out the calculation one may use the replica approach

$$\mathcal{F} = -\frac{1}{\beta}\lim_{N\to\infty}\frac{1}{N}\frac{\partial}{\partial\,\mathrm{ln}}\bigg|_{\mathrm{n}=0}\ln\langle\mathcal{Z}^{\mathrm{n}}\rangle_{\mathcal{D},n}. \tag{23}$$

Averages over the connectivity tensor $\langle(\cdots)\rangle_{\mathcal{D}}$ and noise vector $\boldsymbol{n}$ take the forms

$$\begin{aligned}\langle(\cdots)\rangle_{\mathcal{D}} &= \frac{1}{\mathcal{N}}\sum_{\{\mathcal{D}\}}\prod_{l=1}^{N}\delta\left(\sum_{\langle i_1=l,i_2,\ldots,i_k\rangle}\mathcal{D}_{\langle i_1=l,\ldots,i_k\rangle} - j\right)(\cdots)\\ &= \frac{1}{\mathcal{N}}\sum_{\{\mathcal{D}\}}\prod_{l=1}^{N}\left[\oint\frac{\mathrm{d}Z_l}{2\pi\mathrm{i}}\frac{1}{Z_l^{j+1}}Z_l^{\sum_{\langle i_1=l,i_2,\ldots,i_k\rangle}\mathcal{D}_{\langle i_1=l,\ldots,i_k\rangle}}\right](\cdots)\end{aligned} \tag{24}$$

and

$$\langle(\cdots)\rangle_n = \sum_{n=-1,+1}[(1-p)\delta(n-1) + p\delta(n+1)](\cdots) \tag{25}$$

respectively. Computing the averages and introducing auxiliary variables (order parameters) through the identity

$$\int \mathrm{d}q_{\alpha_1\cdots\alpha_m}\,\delta\left(q_{\alpha_1\cdots\alpha_m} - \frac{1}{N}\sum_i^N Z_i \tau_i^{\alpha_1}\cdots\tau_i^{\alpha_m}\right) = 1 \tag{26}$$

gives rise to the following expression (details of the calculation can be found in [VSK02, MKSV00]):

$$\begin{aligned}\langle \mathcal{Z}^n\rangle_{\mathcal{D},n} = \frac{1}{\mathcal{N}}\int \left(\frac{\mathrm{d}q_0\,\mathrm{d}\hat{q}_0}{2\pi\mathrm{i}}\right)\left(\prod_{\alpha=1}^{\mathsf{n}}\frac{\mathrm{d}q_\alpha\,\mathrm{d}\hat{q}_\alpha}{2\pi\mathrm{i}}\right)\exp\Bigg[\frac{N^k}{k!}\sum_{m=0}^{\mathsf{n}}\sum_{\langle\alpha_1\cdots\alpha_m\rangle}\mathcal{T}_m q_{\alpha_1\cdots\alpha_m}^k \\ - N\sum_{m=0}^{\mathsf{n}}\sum_{\langle\alpha_1\cdots\alpha_m\rangle} q_{\alpha_1\cdots\alpha_m}\hat{q}_{\alpha_1\cdots\alpha_m}\Bigg]\prod_{i=1}^N \mathrm{Tr}_{\{\tau^\alpha\}}\Bigg[\left\langle\exp\left[F\beta n\sum_{\alpha=1}^{\mathsf{n}}\tau^\alpha\right]\right\rangle_n \\ \times \oint\frac{\mathrm{d}Z}{2\pi\mathrm{i}}\frac{\exp\left[Z\sum_{m=0}^{\mathsf{n}}\sum_{\langle\alpha_1\cdots\alpha_m\rangle}\hat{q}_{\alpha_1\cdots\alpha_m}\tau^{\alpha_1}\cdots\tau^{\alpha_m}\right]}{Z^{j+1}}\Bigg]\end{aligned} \tag{27}$$

where $\mathcal{T}_m = \mathrm{e}^{-\mathsf{n}\beta\gamma}\cosh^{\mathsf{n}}(\beta\gamma)\tanh^m(\beta\gamma)$ and $\mathcal{N}$ is a normalization factor.

4.1.2. Replica symmetric solution. The replica symmetric ansatz consists in assuming the following form for the order parameters:

$$q_{\alpha_1\cdots\alpha_m} = \int \mathrm{d}x\,\pi(x)x^m \qquad \hat{q}_{\alpha_1\cdots\alpha_m} = \int \mathrm{d}\hat{x}\,\hat{\pi}(\hat{x})\hat{x}^m. \tag{28}$$

By performing the limit $\gamma \to \infty$, using (28) in (27), computing the normalization constant $\mathcal{N}$, integrating in the complex variable Z, computing the trace and using the replica identity, $\mathsf{n} \to 0$, one finds

$$\begin{aligned}\mathcal{F} = -\frac{1}{\beta}\,\mathrm{Extr}_{\pi,\hat{\pi}}\Bigg\{\frac{j}{k}\ln 2 + j\int \mathrm{d}x\,\mathrm{d}\hat{x}\,\pi(x)\hat{\pi}(\hat{x})\ln(1+x\hat{x}) \\ - \frac{j}{k}\int\prod_{i=1}^k \mathrm{d}x_i\,\pi(x_i)\ln\left(1+\prod_{i=1}^k x_i\right) \\ - \int\prod_{i=1}^j \mathrm{d}\,\hat{x}_i\,\hat{\pi}(\hat{x}_i)\left\langle\ln\left[\sum_{\sigma=\pm1}\mathrm{e}^{\sigma\beta F n}\prod_{i=1}^j(1+\sigma\hat{x}_i)\right]\right\rangle_n\Bigg\}.\end{aligned} \tag{29}$$

Variation with respect to the parameters yields the saddle-point equations:

$$\begin{aligned}\hat{\pi}(\hat{x}) &= \int\prod_{i=1}^{k-1}\mathrm{d}x_i\,\pi(x_i)\delta\left[\hat{x} - \prod_{i=1}^{k-1}x_i\right] \\ \pi(x) &= \int\prod_{l=1}^{j-1}\mathrm{d}\hat{x}_l\,\hat{\pi}(\hat{x}_l)\left\langle\delta\left[x - \tanh(\beta F n + \sum_{l=1}^{j-1}\mathrm{atanh}\,\hat{x}_l)\right]\right\rangle_n\end{aligned} \tag{30}$$

where $\beta = 1$ and $F = \frac{1}{2}\ln\left(\frac{1-p}{p}\right)$ (Nishimori temperature) for MPM decoding in BSC.

One of the most important macroscopic parameters we would like to find is the typical overlap $\rho = \left\langle\frac{1}{N}\sum_{i=1}^N n_i\hat{n}_i\right\rangle_{\mathcal{D},n}$ between the estimate $\hat{n}_i = \mathrm{sgn}(\langle\tau_i\rangle_\beta)$ and the actual noise n_i;

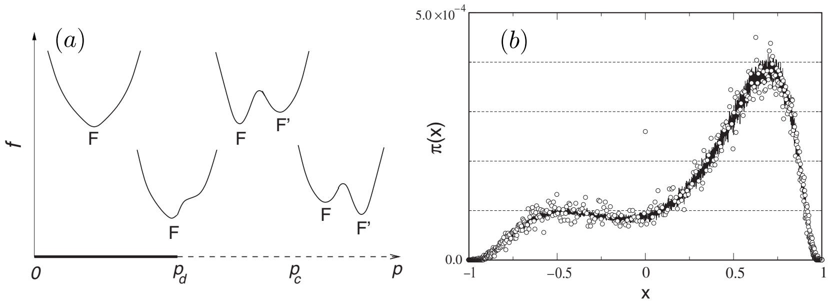


Figure 3. (*a*) Pictorial representation of the RS free energy landscape changing with the noise level p. Up to p_d there is only one stable state F corresponding to the ferromagnetic state with $\rho = 1$. At p_d, a second stable suboptimal ferromagnetic state F' emerges with $\rho < 1$, as the noise level increases, coexistence is attained at p_c. Above p_c, F' becomes the global minimum dominating the system thermodynamics. (*b*) Numerically obtained suboptimal ferromagnetic solution $\pi_{F'}(x)$ for the case $k = 4$, $j = 3$ and $p = 0.2$. Circles correspond to the experimental histogram obtained by decoding with BP in 100 runs for ten different random connectivity matrices.

this can be calculated from

$$\begin{aligned}\rho &= \int \mathrm{d}h\, P(h)\,\mathrm{sgn}(h) \\ P(h) &= \int \prod_{l=1}^{j} \mathrm{d}\hat{x}_l\, \hat{\pi}(\hat{x}_l) \left\langle \delta \left[h - \tanh\left(\beta F n + \sum_{l=1}^{j} \mathrm{atanh}\, \hat{x}_l \right) \right] \right\rangle_n .\end{aligned} \tag{31}$$

4.1.3. Typical performance. To study the various phases of the system one should first solve the saddle point equations (30). In most cases this requires resorting to numerical methods, except for some expected states such as the ferromagnetic and paramagnetic solutions. For instance, the free energy for the ferromagnetic state (F), where

$$\pi_{\mathrm{F}}(x) = \delta[x - 1] \qquad \hat{\pi}_{\mathrm{F}}(\hat{x}) = \delta[\hat{x} - 1] \tag{32}$$

and at Nishimori's temperature, is simply $\mathcal{F}_{\mathrm{F}} = -F(1 - 2p)$, with overlap $\rho = 1$.

The ferromagnetic solution is the only stable solution up to a specific noise level p_d, which identifies the dynamical transition noise level, where meta-stable states first appear. Above p_d, numerical calculations show the emergence of a second stable solution with $\rho < 1$ (suboptimal ferromagnetic); and computationally efficient decoding algorithms cannot identify the dominant solution in feasible time scales. A sketch describing the dependence of the free energy landscape on the noise level is shown in figure 3(*a*) together with a typical numerically obtained suboptimal ferromagnetic solution (figure 3(*b*)) for $k = 4$, $j = 3$ and $p = 0.2$. The ferromagnetic state is *always* a stable solution of (30) and is present for all choices of noise level and construction parameters j and k. It remains dominant up to the thermodynamic transition point p_c, above which the suboptimal ferromagnetic solution becomes the global minimum dominating the system thermodynamics. The identification of both transition points p_d and p_c provides a complete description of the typical performance of infinitely long Gallager codes.

Transitions for Gallager codes with $k = 6$ compared with Shannon's bound (dashed line), the information theory upper bound (full line) and thermodynamic transition points obtained numerically (○) are shown in figure 4(*a*). The thermodynamic transition point obtained, p_c, coincides, within numerical precision, with the information theoretic upper bound [Mac99]. The ferromagnetic and suboptimal ferromagnetic free energies are shown in figure 4(*b*), for $k = 4$ and $R = 1/4$, defining the critical points p_d and p_c.

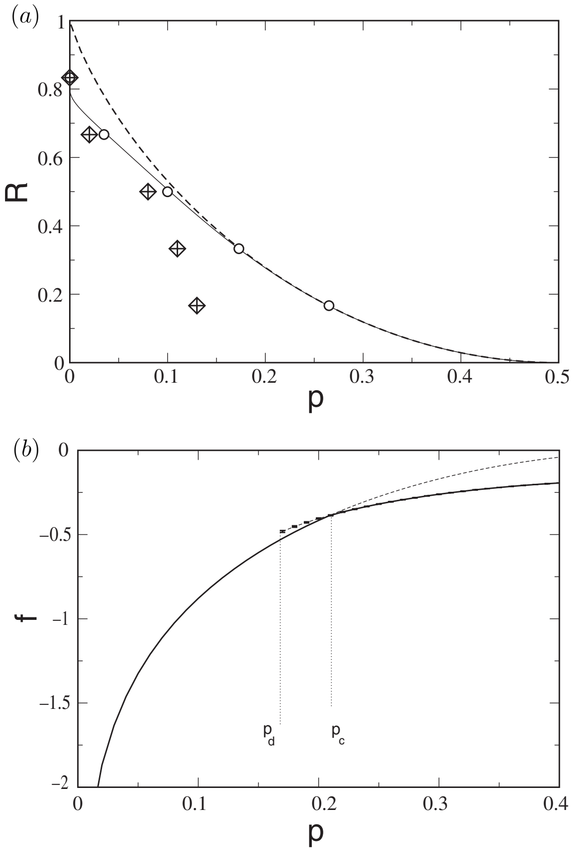


Figure 4. (*a*) Transitions for Gallager codes with $k = 6$ compared with Shannon's bound (dashed line), the information theory upper bound (full line) and thermodynamic transition obtained numerically (○). Transitions obtained by Monte-Carlo integration of the saddle-point equations (◊) and by simulations of BP decoding (+, $M = 5000$ averaged over 20 runs) are also shown. Symbols are chosen larger than the error bars. (*b*) Free energies for $k = 4$, $j = 3$ and $R = 1/4$. The full line corresponds to the free energy of thermodynamic states. Up to p_d only the ferromagnetic state is present. The ferromagnetic state then dominates the thermodynamics up to p_c, where thermodynamic coexistence with suboptimal ferromagnetic states takes place. Dashed lines correspond to RS free energies of non-dominant meta-stable states.

However, the suboptimal ferromagnetic solution has been obtained under the RS ansatz; one can show that above p_d its entropy becomes negative and, therefore, unphysical (at p_c the entropy of the suboptimal ferromagnetic state becomes positive again). This is a clear indication that the replica symmetric solution becomes unstable. A 1-step replica symmetry breaking ansatz has been employed in [FLMRT02] to obtain the solution and complexity of the suboptimal ferromagnetic state and to identify the exact dynamical transition point p_d. The calculation, that considered both BSC and BEC, but focuses on the latter, leads to the same result as that obtained by the RS calculation.

To study the dynamical transition, Franz *et al* [FLMRT02] calculated the number of meta-stable states with a given energy density ϵ, for the BEC, using established methods from the physics of disordered systems [Mon95, FP95]. The number of meta-stable states can be described as

$$\mathcal{N}_{\mathrm{MS}}(\epsilon) \sim \mathrm{e}^{N\Sigma(\epsilon)} \tag{33}$$

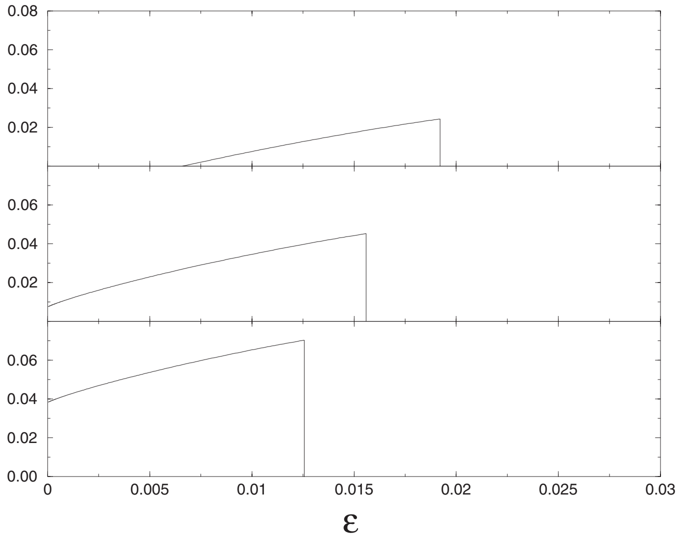


Figure 5. The complexity $\Sigma(\epsilon)$ for (from top to bottom) $p = 0.45$ (below p_c), $p = 0.5$ and $p = 0.55$ (above p_c); calculated for the case of a BEC and a (6, 3) regular code (copied under permission from S Franz, M Leone, A Montanari and F Ricci-Tersenghi 2002. The dynamic phase transition for decoding algorithms *Phys. Rev.* E **66** 046120 [FLMRT02]. Copyright (2002) by the American Physical Society.).

where $\Sigma(\epsilon)$ defines the complexity. Figure 5 shows a plot of the resulting complexity curves for three different values of the erasure probability p in the case of a BEC and a (6, 3) regular code (an 'almost factorized' variational ansatz has been used for calculating the 1-step RSB free energy). The picture that emerges is as follows:

- In the low noise region ($p < p_d$), no meta-stable states exist and local search algorithms are able to recover the erased bits.
- In the intermediate noise region ($p_d < p < p_c$), an exponentially large number of meta-stable states appear with energy densities ϵ in the range $\epsilon_s < \epsilon < \epsilon_d$, defining the static and dynamic energies, with $\epsilon_s > 0$. The best estimated codeword, given the corrupted one, is the original transmitted codeword; however, local algorithms fail to find the best estimate due to a large number of meta-stable solutions.
- Above p_c we have $\epsilon_s = 0$ and a fraction of the meta-stable states consists of valid codewords. Moreover, $\Sigma(0)$ (which gives the number of such codewords) coincides with the complexity of the paramagnetic entropy [FLMRT02].

4.2. MacKay–Neal codes

The analysis of MN codes is quite similar to that of Gallager's codes, the only difference being the consideration of both message and noise vectors in constructing the appropriate Hamiltonian which, after gauging, takes the form

$$\mathcal{H}_\gamma(\boldsymbol{\sigma}, \boldsymbol{\tau}; \boldsymbol{s}, \boldsymbol{n}) = -\gamma \sum_{\langle ir \rangle} \mathcal{D}_{\langle ir \rangle} \big(\sigma_{i_1} \cdots \sigma_{i_k} \tau_{r_1} \cdots \tau_{r_l} - 1\big) - F_s \sum_{i=1}^{k} s_i \sigma_i - F_n \sum_{r=1}^{N} n_r \tau_r \qquad (34)$$

where $\langle \boldsymbol{ir} \rangle$ is shorthand for $\langle i_1 \cdots i_k r_1 \cdots r_l \rangle$; F_s and F_n correspond to the respective Nishimori conditions ($F_s = 0$ in the case of unbiased messages).

A similar analysis to that of Gallager codes results in the following expression for the free energy

$$\mathcal{F} = -\frac{1}{\beta}\,\mathrm{Extr}_{\{\hat{\pi},\pi,\hat{\phi},\phi\}}\left\{\alpha\ln 2 + j\int \mathrm{d}x\,\pi(x)\,\mathrm{d}\hat{x}\,\hat{\pi}(\hat{x})\ln(1+x\hat{x})\right.$$
$$+\alpha l\int \mathrm{d}y\,\phi(y)\,\mathrm{d}\hat{y}\,\hat{\phi}(\hat{y})\ln(1+y\hat{y})$$
$$-\alpha\int\left[\prod_{i=1}^{k}\mathrm{d}x_i\,\pi(x_i)\right]\left[\prod_{r=1}^{l}\mathrm{d}y_r\,\phi(y_r)\right]\ln\left(1+\prod_{i=1}^{k}x_i\prod_{r=1}^{l}y_r\right)$$
$$-\int\left[\prod_{i=1}^{j}\mathrm{d}\hat{x}_i\,\hat{\pi}(\hat{x}_i)\right]\left\langle\ln\left[\sum_{\lambda=\pm1}\mathrm{e}^{\lambda s\beta F_s}\prod_{i=1}^{j}(1+\lambda\hat{x}_i)\right]\right\rangle_s$$
$$\left.-\alpha\int\left[\prod_{r=1}^{l}\mathrm{d}\hat{y}_r\,\hat{\phi}(\hat{y}_r)\right]\left\langle\ln\left[\sum_{\lambda=\pm1}\mathrm{e}^{\lambda n\beta F_n}\prod_{r=1}^{l}(1+\lambda\hat{y}_r)\right]\right\rangle_n\right\}$$

where $\alpha = N/K = j/k$, and $\hat{\pi},\pi,\ \hat{\phi},\phi$ correspond to RS order parameters obtained for both signal and noise vectors, respectively, in the same manner as in section 4.1.2. Full details of the calculation can be found in [VSK02, MKSV00].

The theoretical framework employed for both codes is very similar; however, the solutions obtained analytically and numerically show some interesting differences. In the case of biased messages ($F_s \neq 0$), the results obtained are qualitatively similar to those obtained for Gallager codes, but a different picture emerges when the messages are unbiased, summarized in figure 6 for the cases $k = 1, 2$ and $k \geqslant 3$.

Arguably the most intriguing solution is for the case of $k \geqslant 3$, suggesting that all regular MN codes with $k \geqslant 3$ are theoretically capable of saturating Shannon's limit [KMS00b, MKSV00]. This result has been received with great surprise by the information theory community as it is believed that saturating Shannon's limit is only possible by LDPC codes of infinite connectivity [Mac99, SU03]. One intuitive argument that we can offer [vMSK02] is to do with the randomness of the syndrome vector: any finite connectivity Gallager code takes modulo 2 sums of elements sampled from a biased noise vector and therefore produces a slightly biased syndrome vector; it will only become unbiased once the number of elements sampled diverges. In MN codes, on the other hand, each syndrome bit is obtained from a combination of biased (noise) and unbiased (message) bits, and is therefore truly unbiased even when the number of sampled bits is small.

4.3. Other channels

Extending the analysis above to other channel types is straightforward. The AWGN has been studied in a very similar context in [Ruj93, KS99a, NW99, Mon01, TS03c]. Each real-valued codeword bit can be interpreted as an effective flip rate, leading to a similar Hamiltonian

$$\mathcal{H} = \sum_{\mu=1}^{N-K}\chi(z_\mu = [H\boldsymbol{\tau}]_\mu) - \sum_{i=1}^{N}\log p(\tau_i y_i) \qquad (35)$$

where the last term represents the received real-valued vector $\boldsymbol{y}$ and the effective flip noise vector $\boldsymbol{\tau}$. It is the log-likelihood ratio $h(y_j) \equiv \frac{1}{2}\log(p(y_j)/p(-y_j))$ of the channel noise y_j that serves as the external field acting on site j; the channel characteristics define the field distribution. Analysing the effect of having different communication channels on the code

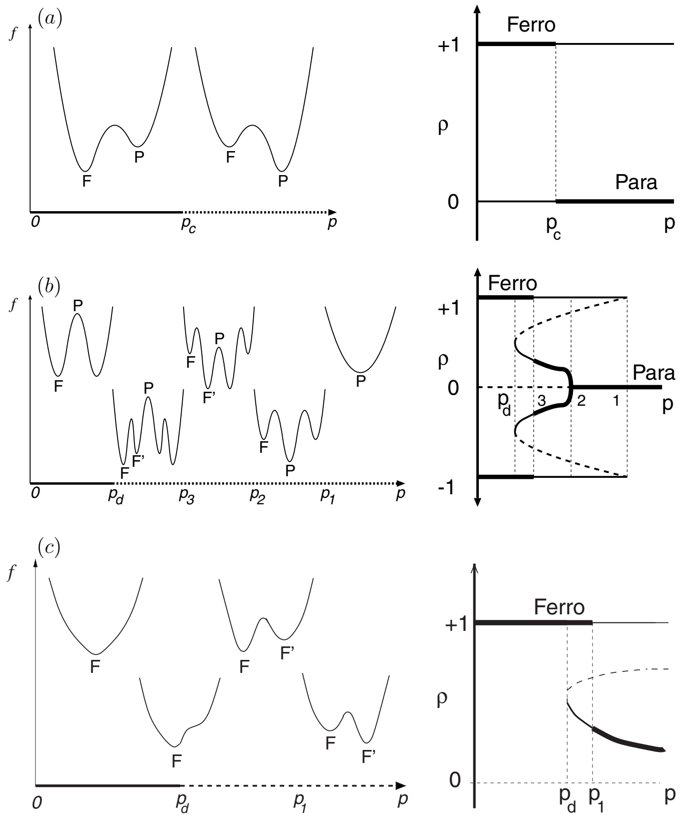


Figure 6. Figures on the left-hand side show schematic representations of free energy landscapes while figures on the right show overlaps ρ as a function of the noise level p; thick and thin lines denote stable solutions of lower and higher free energies, respectively and dashed lines correspond to unstable solutions. (a) $k \geqslant 3$ or $l \geqslant 3, k > 1$: the solid line in the horizontal axis represents the phase where the ferromagnetic solution (F, $\rho = 1$) is thermodynamically dominant. The paramagnetic solution (P, $\rho = 0$) becomes dominant at p_c, that coincides with the channel capacity. (b) $k = 2$ and $l = 2$: the ferromagnetic solution and its mirror image are the only minima of the free energy up to p_d (solid line). Above p_d suboptimal ferromagnetic solutions (F$'$, $\rho < 1$) emerge. The thermodynamic transition occurring at p_3 is below the maximum noise level given by the channel capacity, which implies that these codes do not saturate Shannon's bound even if optimally decoded. (c) $k = 1$: the solid line in the horizontal axis represents the range of noise levels where the ferromagnetic state (F) is the only minimum of the free energy. The suboptimal ferromagnetic state (F$'$) appears in the region represented by the dashed line. The dynamical transition is denoted by p_d, where F$'$ first appears. For higher noise levels, the system becomes bistable and an additional unstable solution of the saddle point equations necessarily appears. The thermodynamical transition occurs at the noise level p_1 (smaller than Shannon's limit) where F$'$ becomes dominant.

properties, therefore reduces to investigating the effect of different field distributions on the physical properties of the system. For instance, for the AWGN, this reduces to (for a detailed description see [TS03c])

$$p_{\mathrm{AWGN}}(h) = \sqrt{\frac{\sigma^2}{2\pi}}\exp(-(h-\sigma^{-2})^2/2\sigma^{-2}). \tag{36}$$

The calculation then follows in a similar way to those described previously and produces qualitatively the same results for all channels studied [TS03c]; the exact numerical details

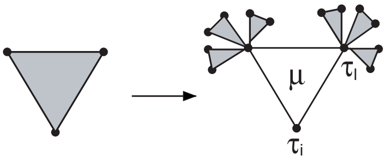


Figure 7. First step in the construction of a Husimi cactus with $k = 3$ and connectivity $j = 4$.

change from channel to channel. Several different channels for both Gallager and MN codes, in a broad parameter range, have been examined in [TS03c]; among the channels studied: the binary-input additive-white-Gaussian-noise channel (BIAWGNC), the binary-input Laplace channel (BILC) and the general binary-input output-symmetric (BIOS) memoryless channel.

4.4. The Bethe approximation

An alternative method for carrying out the analysis is by employing the Bethe approximation [WS87b] (also termed Thouless–Anderson–Palmer (TAP) approach for diluted systems [KS98, VSK99, VSK02] and Husimi cactus [VSK00a]) that is exactly solvable [Guj95, BL82, RK92, Gol91]. It assumes a tree-like graph of connectivity j and a polygon of k vertices with one Ising spin in each vertex. All spins in a polygon interact through a single coupling element $\mathcal{D}_\mu$, where μ represents a shorthand notation for a selection of indices $\langle i_1 \cdots i_k \rangle$; one of the spins is called the base spin (generation 0), as shown in figure 7. In a generic step, the base spins of the $(j-1)(k-1)$ polygons in generation $t-1$ are attached to $k-1$ vertices of a polygon in the next generation t. This process is iterated until a maximum generation $t_{\max}$ is reached, the graph is then completed by attaching j uncorrelated branches of $t_{\max}$ generations at their base spins. In this way each spin inside the graph is connected to j polygons exactly. The local magnetization at the centre m_i can be obtained by fixing boundary (initial) conditions in the 0th generation and iterating the related recursion equations until generation $t_{\max}$ is reached. Carrying out the calculation in the thermodynamic limit corresponds to having $t_{\max} \sim \ln N$ generations and $N \to \infty$.

We adopt here the approach presented in [RK92] for obtaining recursion relations. The probability distribution $P_{\mu i}(\tau_i)$ for the base spin of the polygon μ is connected to $(j-1)(k-1)$ distributions $P_{\nu l}(\tau_l)$, with $\nu \in \mathcal{G}(l)\backslash\mu$ (the set of all polygons linked to l but not μ) of polygons in the previous generation:

$$P_{\mu i}(\tau_i) = \frac{1}{\mathcal{N}} \operatorname{Tr}_{\{\tau_l\}} \exp\left[\beta\gamma\left(\mathcal{J}_\mu \tau_i \prod_{l\in\mathcal{L}(\mu)\backslash i} \tau_l - 1\right) + \beta F \tau_i\right] \prod_{\nu\in\mathcal{G}(l)\backslash\mu} \prod_{l\in\mathcal{L}(\mu)\backslash i} P_{\nu l}(\tau_l) \tag{37}$$

where $\mathcal{L}(\mu)$ denotes the polygon μ of the lattice and the trace is over the spins τ_l such that $l \in \mathcal{L}(\mu)\backslash i$; $\mathcal{J}_\mu$ represents the corresponding syndrome vector.

Calculating the effective field $\hat{x}_{\nu l}$ on a base spin l due to neighbours in polygon ν, taking $\gamma \to \infty$ and $\beta = 1$, one obtains the effective local magnetization due to interactions with the nearest neighbours in one branch $\hat{m}_{\mu l} = \tanh(\hat{x}_{\mu l})$, where

$$\hat{x}_{\mu i} = \operatorname{atanh}\left[\mathcal{J}_\mu \prod_{l\in\mathcal{L}(\mu)\backslash i} \tanh\left(F + \sum_{\nu\in\mathcal{G}(l)\backslash\mu} \hat{x}_{\nu l}\right)\right]. \tag{38}$$

The effective local field on a base spin l of a polygon μ due to $j-1$ branches in the previous generation and due to the external field is

$$x_{\mu l} = F + \sum_{\nu\in\mathcal{G}(l)\backslash\mu} \hat{x}_{\nu l}. \tag{39}$$

The set of equations (38), (39) can be rewritten in terms of $\hat{m}_{\mu l}$ and $m_{\mu l}$ [Mac99, KS98, KF98]

$$m_{\mu i} = \tanh\left(F + \sum_{\nu\in\mathcal{G}(l)\backslash\mu} \operatorname{atanh}(\hat{m}_{\nu i})\right) \qquad \hat{m}_{\mu i} = \mathcal{J}_\mu \prod_{l\in\mathcal{L}(\mu)\backslash i} m_{\mu l} \tag{40}$$

giving rise to a closed set of iterative equations (identical to those of BP) that can also be used for decoding. Iterating the coupled set of equations (40) one converges to a stable minimum and can compute the following approximated free energy:

$$\begin{aligned}\mathcal{F}(\{m_{\mu i}, \hat{m}_{\mu i}\}) = \sum_{\mu=1}^{N-K}\sum_{r\in\mathcal{L}(\mu)} \ln(1 + m_{\mu r}\hat{m}_{\mu r}) - \sum_{\mu=1}^{N-K} \ln\left(1 + \mathcal{J}_\mu \prod_{r\in\mathcal{L}(\mu)} m_{\mu r}\right)\\ - \sum_{l=1}^{N} \ln\left[\mathrm{e}^F \prod_{\mu\in\mathcal{G}(l)} (1 + \hat{m}_{\mu l}) + \mathrm{e}^{-F} \prod_{\mu\in\mathcal{G}(l)} (1 - \hat{m}_{\mu l})\right].\end{aligned} \tag{41}$$

Equations (40) represent the interdependence of microscopic quantities; a macroscopic description can be constructed by retaining only statistical information about the system, namely by describing the evolution of histograms of variables $x_{\mu i}$ and $\hat{x}_{\mu i}$.

Assuming that the effective fields $x_{\mu i}$ and $\hat{x}_{\mu i}$ are random variables *independently* sampled from the distributions $P(x)$ and $\hat{P}(\hat{x})$, respectively, and that n_i is sampled from $P(n) = (1-p)\delta(n-1) + p\delta(n+1)$, one can then establish the following recursion relation in the space of probability distributions [BL82]:

$$\begin{aligned} P_t(x) &= \int \mathrm{d}n\, P(n) \int \prod_{l=1}^{j-1} \mathrm{d}\hat{x}_l\, \hat{P}_{t-1}(\hat{x}_l)\delta\left[x - Fn - \sum_{l=1}^{j-1}\hat{x}_l\right]\\ \hat{P}_{t-1}(\hat{x}) &= \int \prod_{l=1}^{k-1} \mathrm{d}x_l P_{t-1}(x_l)\delta\left[\hat{x} - \operatorname{atanh}\left(\prod_{l=1}^{k-1}\tanh(x_l)\right)\right]\end{aligned} \tag{42}$$

where $P_t(x)$ is the distribution of effective fields in the tth generation due to the previous generations and external fields; in the thermodynamic limit the distribution far from the boundary is $P_\infty(x)$ (generation $t \to \infty$). The local field distribution at the central site is computed by replacing $j-1$ by j in the first equation of (42):

$$P(h) = \int \mathrm{d}n\, P(n) \int \prod_{l=1}^{C} \mathrm{d}\hat{x}_l\, \hat{P}_\infty(\hat{x}_l)\delta\left[x - Fn - \sum_{l=1}^{C}\hat{x}_l\right]. \tag{43}$$

It is easy to see that $P_\infty(x)$ and $\hat{P}_\infty(\hat{x})$ satisfy equations (30) obtained by the replica symmetric assumption [KMS00b, MKSV00, VSK00b] if the variables describing fields are transformed to those of local magnetizations through $x \mapsto \tanh(\beta x)$. It is therefore not surprising that one obtains identical results to those obtained using the RS analysis and using BP decoding. In fact, the DE method used extensively in the IT community for analysing LDPC codes is similar to the macroscopic iterative equations (42).

4.5. Weight and magnetization enumerators

A different approach to analysing properties of LDPC codes relies on a microscopic calculation where solution vectors are forced to lie on a shell defined by the overlap with the true solution (weight enumerator) or by a certain magnetization value (magnetization enumerator); both can be used to define critical transition points of LDPC codes. We focus here on the magnetization

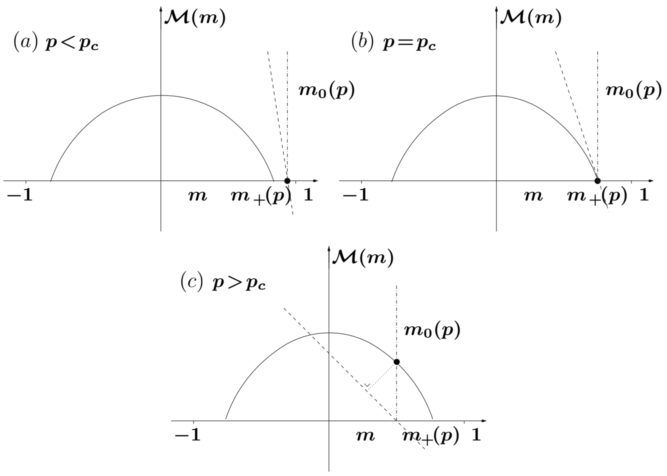


Figure 8. The qualitative picture of $\mathcal{M}(m) \geqslant 0$ (solid curve lines) for different values of p. For MAP, MPM and typical set decoding, only the relative values of $m_+(p)$ and $m_0(p)$ determine the critical noise level. Dashed lines correspond to the energy contribution of $-\beta F$ for Nishimori's condition ($\beta = 1$). The states with the lowest free energy are indicated by a point •. (*a*) Sub-critical 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 (free energy difference corresponds to the distance between the dashed line and the magnetization enumerator curve). (*b*) Critical noise level $p = p_c$, where $m_+(p) = m_0(p)$. The minimal free energy of the sub-optimal solutions coincides with that of the correct solution at Nishimori's condition (all meet at $m_+(p) = m_0(p)$). (*c*) Over-critical noise levels $p > p_c$ where many solutions have a higher magnetization than the true typical one. The minimal free energy of suboptimal solutions is lower than that of the true solution.

enumerator ($\mathcal{M}$); calculations involving the weight enumerator will be mentioned in section 5.2.

The corresponding Hamiltonian is similar to (20) except for the second term that defines the magnetization shell (after gauging)

$$\mathcal{H}_{\gamma,m}(\boldsymbol{\tau}) = -\gamma \sum_{\langle i_1\cdots i_k\rangle} \mathcal{D}_{\langle i_1\cdots i_k\rangle}(\tau_{i_1}\cdots\tau_{i_k} - 1) - \delta\left(\sum_{l=1}^{N} n_l\tau_l - m\right). \qquad (44)$$

Calculating the related entropy as a function of the magnetization m provides an intuitive and transparent explanation of the relation between different decoding schemes such as typical set decoding, MAP and finite temperature decoding (MPM) [vMSK01, vMSK02].

Carrying out the analysis along the same lines as before [Mon01, vMSK01, vMSK02], one obtains expressions for the magnetization enumerator as a function of m, similar to those sketched in figure 8; from these plots one can provide a simple explanation of the relation between various (theoretical) decoding methods, and calculate the thermodynamic transition point p_c. The magnetization enumerator $\mathcal{M}(m)$ (curved solid line) takes positive values only 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 $\boldsymbol{n}$ has the typical magnetization of the noise vector; in the case of a BSC $m(\boldsymbol{n}) = m_0(p) = 1-2p$ (the typical set magnetization is denoted by a dashed–dotted line). States with the lowest free energy are denoted by a point (•).

Selection of the best estimates by the various decoding schemes can be summarized as follows:

- *Maximum likelihood (MAP) decoding*. It selects the solution vector $\boldsymbol{\tau}$ (obeying all parity checks) with the highest magnetization. As the noise level increases, the gap between $m_0(p)$ and $m_+(p)$) closes; the critical noise level p_c is determined by the condition $m_+(p_c) = m_0(p_c)$.
- *Typical set decoding*. It is based on randomly selecting a solution vector $\boldsymbol{\tau}$ with the expected magnetization $m(\boldsymbol{\tau}) = m_0(p)$ [AJK01]; an error is declared when there is no such vector or when there are several solution vectors with magnetization $m(\boldsymbol{\tau}) = m_0(p)$. The critical noise level p_c is determined by the condition $m_+(p_c) = m_0(p_c)$, and is identical to the point obtained by a MAP decoder.
- *Finite temperature (MPM) decoding*. Selection is based on a free energy minimization [KMS00b], where an energy term $-Fm(\boldsymbol{\tau})$ is added to the parity check term (20). Using the thermodynamic relation $\mathcal{F} = \mathcal{U} - \frac{1}{\beta}\mathcal{S}$, β being the inverse temperature (Nishimori's condition corresponds to setting $\beta = 1$), $\mathcal{U}$ the internal energy and $\mathcal{S}$ the entropy; the free energy of suboptimal solutions is given by $\mathcal{F}(m) = -Fm - \frac{1}{\beta}\mathcal{M}(m)$ (for $\mathcal{M}(m) \geqslant 0$), while that of the true solution is given by $-Fm_0(p)$.

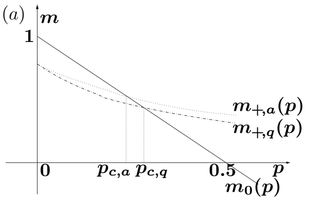
  The selection process is explained graphically in figure 8. The *energy difference* between suboptimal solutions relative to that of the correct solution, is given by the dashed line of slope $-F$ through the point $(m_0(p), 0)$; to calculate the free energy of any suboptimal solution one should also consider its entropy, represented by the magnetization enumerator curve (the true solution is of zero entropy). Therefore, the distance between $\mathcal{M}(m)$ and the dashed line represents the difference between the lowest free energy among suboptimal solutions and that of the true solution. Solutions of magnetization m for which $\mathcal{M}(m)$ lies above/below this line, have a lower/higher free energy, respectively. The critical noise level p_c is defined by the lowest p value for which there are suboptimal solutions with a free energy equal to $-Fm_0(p)$ (i.e. a single contact point between the dashed line and the magnetization enumerator curve). It coincides with the point obtained by MAP [MN00] and typical set decoding [vMSK02].

The critical noise level is defined by following the dependence of m_+ on the noise level and finding the point $m_+(p_c) = m_0(p_c)$ as described in figure 9; results obtained for the critical noise level in the case of Gallager codes of various parameters are also shown (for both quenched and annealed calculations of the free energy related to (44), denoted by a subscript a/q). The annealed approximation gives a much more pessimistic estimate for p_c as it overestimates $\mathcal{M}$ by giving high weight to exponentially rare events. Results obtained by the quenched calculations are similar to those reported in [KSNS01] using another method as explained in section 5.2, but are more optimistic than those reported in the IT literature which rely on bounding techniques.

The analysis has also been carried out for MN codes [vMSK01, vMSK02] and in a range of channel types [SvMS03]. Interestingly, the location of m_+ remains fixed for MN codes with $k \geqslant 3$ and for $k = 2, l \geqslant 3$, leading to a thermodynamical transition point that saturates Shannon's limit in agreement with our previous results [KMS00b, MKSV00].

5. Optimal performance : statistical mechanics versus IT

DE offers a useful framework for evaluating error correction performance achieved by a practical decoding algorithm on the basis of the BP/TAP approach. However, this does not necessarily mean the best performance among all possible decoding schemes. To clarify



(b)

(k,j)	$(6,3)$	$(5,3)$	$(6,4)$	$(4,3)$
Code rate	$1/2$	$2/5$	$1/3$	$1/4$
IT	0.0915	0.129	0.170	0.205
$p_{c,a}$ $(\mathcal{M}_a)$	0.031	0.066	0.162	0.195
$p_{c,q}$ $(\mathcal{M}_q)$	0.0998	0.1365	0.1725	0.2095
Shannon $p_{\rm sh}$	0.109	0.145	0.174	0.214

Figure 9. (*a*) Determining the critical noise levels $p_{c,a/q}$ (quenched and annealed calculations) based on the function $\mathcal{M}_{a/q}$ for Gallager codes. (*b*) Comparison of different critical noise level (p_c) estimates for Gallager codes. Typical set decoding estimates have been obtained via the methods of IT [AJK01], based on the weight enumerator. Shannon's limit denotes the highest theoretically achievable critical noise level $p_{\rm sh}$ for any code [Sha48].

the potential of a code ensemble, it is important to assess the theoretical error correction ability, disregarding computational cost. Several methods have been developed for this purpose in the IT literature. In this section, we introduce two representative schemes, termed *Gallager's methodology* and *typical set analysis*, and relate them to methods known in statistical mechanics (SM). For simplicity, we hereafter focus on (j, k) regular Gallager-type LDPC codes and a BSC of flip probability p; extension to other types of codes such as MN codes and other channels is straightforward.

5.1. Gallager's methodology: error probability for finite code lengths

Shannon's seminal papers indicated that the best code can provide error free communication if code rate R is below Shannon's limit when the code length becomes *infinite*. However, as any code in use has a *finite* code length N, it is practically important and theoretically interesting to assess the probability of error correction failure as a function of the code length.

Gallager's variational method is a systematic scheme for upper bounding the error probability of the best code in a given code ensemble $\mathcal{C}$ by averaging it over the ensemble. In the IT literature, it is usually assumed that decoding is performed directly on *codewords* and, therefore, Gallager's method is conventionally introduced in a manner suitable for this decoding approach. However, this formulation is not convenient here because the decoding problem is provided first with respect to *noise vectors* in Gallager-type codes. We therefore introduce a slightly different representation of Gallager's method, which is applicable to a range of decoding schemes.

5.1.1. Gallager's inequality for the MAP estimator. Suppose that binary vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, which consist of K-bit and N-bit components, respectively, are statistically related via a certain joint distribution $P(\boldsymbol{x}, \boldsymbol{y})$. Let us consider an estimation problem of $\boldsymbol{x}$ given $\boldsymbol{y}$. Following the Bayesian framework, it can be shown that the block error probability, which is the probability that the estimation result $\hat{\boldsymbol{x}}$ given $\boldsymbol{y}$ is not identical to the vector $\boldsymbol{x}$, is minimized by the maximum *a posteriori* probability (MAP) estimator

$$\begin{aligned}\hat{\boldsymbol{x}}_{\rm MAP} &= \underset{\boldsymbol{x}}{\rm argmax}\{P(\boldsymbol{x}|\boldsymbol{y})\} = \underset{\boldsymbol{x}}{\rm argmax}\left\{\frac{P(\boldsymbol{x},\boldsymbol{y})}{\sum_{\boldsymbol{x}'} P(\boldsymbol{x}',\boldsymbol{y})}\right\} \\ &= \underset{\boldsymbol{x}}{\rm argmax}\{P(\boldsymbol{x},\boldsymbol{y})\}.\end{aligned} \tag{45}$$

In order to evaluate the block error probability of this estimator, we introduce an *indicator* *function* $\Delta_{\mathrm{MAP}}(\boldsymbol{x}, \boldsymbol{y})$ which returns 1 if $\hat{\boldsymbol{x}}_{\mathrm{MAP}} \neq \boldsymbol{x}$ and 0, otherwise. Then, the block error probability is computed as

$$P_B = \sum_{\boldsymbol{x},\boldsymbol{y}} P(\boldsymbol{x}, \boldsymbol{y}) \Delta_{\mathrm{MAP}}(\boldsymbol{x}, \boldsymbol{y}). \tag{46}$$

Gallager's methodology relies on upper bounding this probability by utilizing the following inequality for the indicator function

$$\Delta_{\mathrm{MAP}}(\boldsymbol{x}, \boldsymbol{y}) \leqslant \left(\sum_{\boldsymbol{x}' \neq \boldsymbol{x}} \left(\frac{P(\boldsymbol{x}', \boldsymbol{y})}{P(\boldsymbol{x}, \boldsymbol{y})} \right)^{\lambda} \right)^{\rho} \tag{47}$$

which holds for arbitrary $\lambda \geqslant 0$ and $\rho \geqslant 0$. This inequality is proved as follows: if $\hat{\boldsymbol{x}}_{\mathrm{MAP}} = \boldsymbol{x}$, $\Delta_{\mathrm{MAP}}(\boldsymbol{x}, \boldsymbol{y}) = 0$. However, the right-hand side is always non-negative, which means that equation (47) holds. On the other hand, if $\hat{\boldsymbol{x}}_{\mathrm{MAP}} \neq \boldsymbol{x}$, $\Delta_{\mathrm{MAP}}(\boldsymbol{x}, \boldsymbol{y}) = 1$. However, this implies that there exists at least one vector $\boldsymbol{x}'' \neq \boldsymbol{x}$ such that $P(\boldsymbol{x}'', \boldsymbol{y}) \geqslant P(\boldsymbol{x}, \boldsymbol{y})$. This can be generalized as $\Delta_{\mathrm{MAP}}(\boldsymbol{x}, \boldsymbol{y}) = 1 \leqslant (P(\boldsymbol{x}'', \boldsymbol{y})/P(\boldsymbol{x}, \boldsymbol{y}))^{\lambda} \leqslant \sum_{\boldsymbol{x}' \neq \boldsymbol{x}} (P(\boldsymbol{x}', \boldsymbol{y})/P(\boldsymbol{x}, \boldsymbol{y}))^{\lambda}$ for $\forall \rho \geqslant 0$; equation (47) immediately follows because the ratio $P(\boldsymbol{x}', \boldsymbol{y})/P(\boldsymbol{x}, \boldsymbol{y})$ is always non-negative and $\forall \rho \geqslant 0$, $x^{\rho} \geqslant 1$ holds for $\forall x \geqslant 1$.

Inserting equation (47) into equation (46) we obtain *Gallager's inequality*

$$\begin{aligned} P_B &\leqslant \sum_{\boldsymbol{x},\boldsymbol{y}} P(\boldsymbol{x}, \boldsymbol{y}) \left(\sum_{\boldsymbol{x}' \neq \boldsymbol{x}} \left(\frac{P(\boldsymbol{x}', \boldsymbol{y})}{P(\boldsymbol{x}, \boldsymbol{y})} \right)^{\lambda} \right)^{\rho} \\ &= \sum_{\boldsymbol{x},\boldsymbol{y}} P^{1-\lambda\rho}(\boldsymbol{x}, \boldsymbol{y}) \left(\sum_{\boldsymbol{x}' \neq \boldsymbol{x}} P^{\lambda}(\boldsymbol{x}', \boldsymbol{y}) \right)^{\rho} \end{aligned} \tag{48}$$

which provides the tightest inequality by choosing $\lambda = 1/(1+\rho)$ when ρ is fixed. As this inequality holds for $\forall \rho \geqslant 0$ and $\forall \lambda \geqslant 0$, the bound can be optimized by minimization of the right-hand side with respect to $\rho \geqslant 0$ keeping $\lambda = 1/(1+\rho)$.

5.1.2. Application for decoding Gallager-type codes.

Equation (48) can be employed for evaluating the block error probability of the decoding problem of Gallager-type codes. For this, we introduce the joint probability of noise vector $\boldsymbol{n}$ and syndrome vector $\boldsymbol{z}$ given a parity check matrix H; employing the Ising spin representation

$$P(\boldsymbol{n}, \boldsymbol{z}|H) = \prod_{\mu=1}^{N-K} \delta\left(z_{\mu}, \prod_{i \in \mathcal{L}(\mu)} n_i \right) \times \frac{\exp\left(F \sum_{i=1}^{N} n_i\right)}{(2\cosh(F))^N} \tag{49}$$

where $\delta(x, y) = 1$ for $x = y$ and 0 otherwise, $\mathcal{L}(\mu)$ denotes the set of indices $\langle i_1 \cdots i_k \rangle$ for non-zero elements in the μth row of H and $F = \frac{1}{2}\ln[(1-p)/p]$. The first term enforces the parity checks (12) (representing the likelihood term $P(\boldsymbol{z}|\boldsymbol{n}, H)$), while the second represents the appropriate prior term; this is because the noise vector $\boldsymbol{n}$ is generated in the BSC with the prior probability $P(\boldsymbol{n}) = \exp\left(F \sum_{i=1}^{N} n_i\right)\big/(2\cosh(F))^N$.

Using equation (48) in equation (49) leads to an upper bound of the block error probability of the MAP decoding for a given parity check matrix H as

$$P_B(H) \leqslant \sum_{\boldsymbol{n},\boldsymbol{z}} P^{1-\lambda\rho}(\boldsymbol{n}, \boldsymbol{z}|H) \left(\sum_{\boldsymbol{n}' \neq \boldsymbol{n}} P^{\lambda}(\boldsymbol{n}', \boldsymbol{z}|H) \right)^{\rho}$$

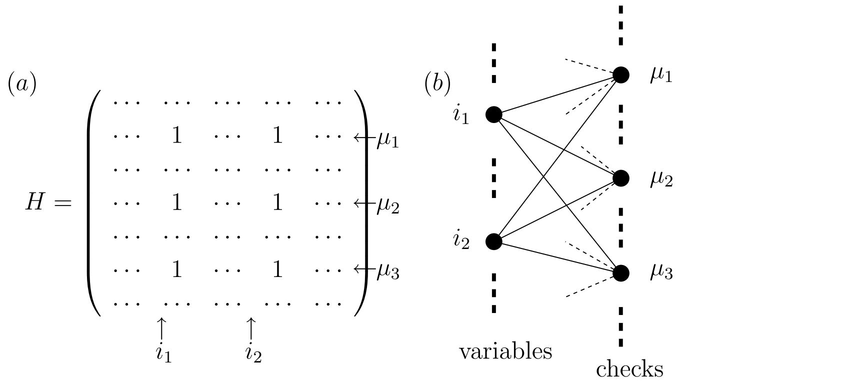


Figure 10. A configuration in a parity check matrix H of $j = 3$ that deteriorates the decoding performance (*a*), represented as a short cycle of a particular type in the graphical expression (*b*). When two variables indexed by i_1 and i_2 share all of the same $j = 3$ checks which are denoted as μ_1, μ_2 and μ_3, simultaneous flips of these two do not break the parity check condition. This makes it difficult to identify correctly the true noise vector $\boldsymbol{n}$. When H is generated uniformly under the (j, k)-constraints, this kind of configuration occurs with a probability of $O(N^{-1})$ in the case of $j = 3$, which yields a polynomially slow decay in equation (51).

$$= \sum_{\boldsymbol{n}} \frac{\exp\big((1-\lambda\rho)F\sum_{i=1}^{N} n_i\big)}{(2\cosh(F))^N} \times \left(\sum_{\boldsymbol{n}'\neq\boldsymbol{n}} \prod_{\mu=1}^{N-K} \delta\left(1, \prod_{i\in\mathcal{L}(\mu)} n_i n_i'\right) \times \exp\left(\lambda F \sum_{i=1}^{N} n_i'\right)\right)^{\rho} \tag{50}$$

where summation over $\boldsymbol{z}$ has already been carried out, resulting in a contribution $\prod_{\mu=1}^{N-K} \delta\big(\prod_{i\in\mathcal{L}(\mu)} n_i, \prod_{i\in\mathcal{L}(\mu)} n_i'\big) = \prod_{\mu=1}^{N-K} \delta\big(1, \prod_{i\in\mathcal{L}(\mu)} n_i n_i'\big)$. For a given code ensemble, the minimum of the block error probability P_B^* is always upperbounded by the average error probability $\langle P_B(H)\rangle_H$, where $\langle(\cdots)\rangle_H$ denotes average over the ensemble of codes (or parity check matrices H) under appropriate constraints. Therefore, we here obtain an upper bound for the block error probability of the best code in the (j, k)-Gallager code ensemble by

$$P_B^* \leqslant \sum_{\boldsymbol{n}} \frac{\exp\big((1-\lambda\rho)F\sum_{i=1}^{N} n_i\big)}{(2\cosh(F))^N} \times \left\langle \left(\sum_{\boldsymbol{n}'\neq\boldsymbol{n}} \prod_{\mu=1}^{N-K} \delta\left(1, \prod_{i\in\mathcal{L}(\mu)} n_i n_i'\right) \times \exp\left(\lambda F \sum_{i=1}^{N} n_i'\right)\right)^{\rho} \right\rangle_H \tag{51}$$

which can be optimized by minimizing the right-hand side with respect to $\rho \geqslant 0$, keeping $\lambda = 1/(1+\rho)$.

5.1.3. Rigorous bound. It has been shown, using the methods of IT, that the right-hand side of equation (51) can be decomposed into two parts as

$$O(N^{-\gamma}) + O(\exp[-NE]) \tag{52}$$

for naively (and completely randomly) constructed (j, k)-Gallager code ensembles, where γ is a certain power determined by parameters j, k and N is assumed large [Gal63, MB01]. This implies that the bound vanishes to 0 as $N \to \infty$ if the exponent E, which depends on the adjustable parameters $\rho, \lambda \geqslant 0$, can be maximized to a positive value. The rate of convergence is quite slow due to a polynomially small fraction of poor codes in the ensemble, which have short cycles of particular kinds in the parity check matrices (figure 10) [vMK03]. Therefore, the behaviour of the average bound (51), (52) can be improved by *expurgating* such codes from the ensemble. In [MB01], it is shown that the expurgated ensemble exhibits an exponential behaviour, characterized by the second term of equation (52).

For expurgated ensembles, one can evaluate a rigorous lower bound of the exponent E as a function of ρ and λ, with an extra constraint, by employing Jensen's inequality $\langle X^\rho \rangle \leqslant \langle X \rangle^\rho$, which is valid for a non-negative random number X and $0 \leqslant \rho \leqslant 1$. This yields

$$E_a(\rho, \lambda; R, p) = \operatorname*{Extr}_{|x|<1, |\hat{x}|<1} \left\{ \rho \left[-\frac{j}{k} \ln \left(\frac{1+x^k}{2} \right) + j \ln \left(\frac{1+\hat{x}x}{2} \right) \right. \right.$$
$$\left. - \ln \left[\left(\sum_{n'=\pm 1} \mathrm{e}^{\lambda F n n'} \left(\frac{1+\hat{x}n'}{2} \right)^j \right) \right]_{\lambda\rho} \right]$$
$$\left. - \ln 2\cosh(1-\lambda\rho)F + \ln 2\cosh F \right\} \tag{53}$$

where $[\cdots]_{\lambda\rho} = \sum_{n=\pm 1} (\cdots)\, \mathrm{e}^{(1-\lambda\rho)Fn}/(2\cosh(1-\lambda\rho)F)$ and $\operatorname{Extr}(\cdots)$ denotes extremization over the variables $|x| < 1$ and $|\hat{x}| < 1$. This procedure is analogous to the *annealed approximation* of SM, similar to the approach taken in [SST92].

For $j, k \to \infty$, while keeping $R = 1 - j/k = K/N$ finite, the maximization of equation (53) with respect to $0 \leqslant \rho \leqslant 1$ keeping $\lambda = 1/(1+\rho)$ reproduces the *random coding exponent*

$$E_{\mathrm{RC}}(R, p) = \begin{cases} (1-R)\ln 2 - \ln(\sqrt{p} + \sqrt{1-p})^2, & 0 \leqslant p \leqslant p_b \\ p_c \ln \frac{p_c}{p} + (1-p_c) \ln \frac{1-p_c}{1-p}, & p_b < p \leqslant p_c \\ 0, & p_c < p \end{cases} \tag{54}$$

which is known in IT literature [Gal68], where the BSC flip rate $p = (1 - \tanh(F))/2$, p_c is a critical noise rate that satisfies Shannon's limit $R = 1 - \mathsf{H}_2(p_c)$ and $p_b = p_c^2 \big/ \big(p_c^2 + (1-p_c)^2\big)$, is often termed *Bhattachalya's limit*. For relatively high rates R, it is known that this exponent represents the exact decay rate of the best possible codes, which implies that there is no room for improving the bound (54) in the case of $j, k \to \infty$ (but obviously not for finite j, k values, where no exact expression exists in the IT literature).

5.1.4. Improving the bound by the replica method. However, the exact result for infinite j, k does not necessarily mean that the exponent of (54) provides the tightest bound for *finite* j, k as well. Actually, direct evaluation of equation (51) using the replica method yields another exponent [KSNS01]

$$E_q(\rho, \lambda; R, p) = \operatorname*{Extr^*}_{\pi(\cdot), \hat{\pi}(\cdot)} \left\{ -\frac{j}{k} \ln \left\langle \left(\frac{1 + \prod_{i=1}^k x_i}{2} \right)^\rho \right\rangle_\pi + j \ln \left\langle \left(\frac{1+\hat{x}x}{2} \right)^\rho \right\rangle_{\pi, \hat{\pi}} \right.$$
$$- \ln \left\langle \left[\left(\sum_{n'=\pm 1} \mathrm{e}^{\lambda F n n'} \prod_{\mu=1}^{j} \left(\frac{1+\hat{x}_\mu n'}{2} \right) \right)^\rho \right]_{\lambda\rho} \right\rangle_{\hat{\pi}}$$
$$\left. - \ln 2\cosh(1-\lambda\rho)F + \ln 2\cosh F \right\} \tag{55}$$

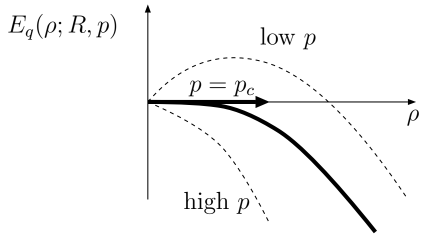


Figure 11. Schematic profiles of $E_q(\rho;\, R,\, p)$.

under the RS ansatz, where $\langle \cdots \rangle_\pi$ denotes an average with respect to dummy variables $x_i \in [-1, 1]$ $(i = 1, 2, \ldots, k)$ over an identical variational distribution $\pi(x)$, and similarly for $\hat{x}_\mu \in [-1, 1]$ $(\mu = 1, 2, \ldots, j)$ and $\langle \cdots \rangle_{\hat{\pi}}$. The functional extremization $\text{Extr}^*{}_{\pi(\cdot),\hat{\pi}(\cdot)}\{\cdots\}$ excludes the ferromagnetic solution of $\pi_{\text{F}}(x) = \delta(x-1)$ and $\hat{\pi}_{\text{F}}(\hat{x}) = \delta(\hat{x}-1)$.

For finite j, k, $E_q(\rho, \lambda;\, R,\, p)$ is maximized by $\lambda = 1/(1+\rho)$ for any given $\rho \geqslant 0$, whereas $E_a(\rho, \lambda;\, R,\, p)$ is not. For the partially maximized exponent $E_q(\rho;\, R,\, p) \equiv E_q(\rho, 1/(1+\rho);\, R,\, p)$, the following properties generally hold (figure 11):

$$\lim_{\rho\to 0} E_q(\rho;\, R,\, p) = 0 \tag{56}$$

$$\frac{\partial^2}{\partial\rho^2} E_q(\rho;\, R,\, p) < 0. \tag{57}$$

This implies that for a given R, the *noise threshold* p_c below which $\max_{\rho\geqslant 0}\{E_q(\rho;\, R,\, p)\}$ becomes positive, indicating that the average error bound vanishes for $N \to \infty$, is determined by a condition

$$\lim_{\rho\to 0} \frac{\partial}{\partial\rho} E_q(\rho;\, R,\, p_c) = 0. \tag{58}$$

Inserted into equation (55), this reduces to the phase boundary condition

$$\mathcal{F}_{\text{NF}} - \mathcal{F}_{\text{F}} = 0 \tag{59}$$

where $\mathcal{F}_{\text{F}} = -F\tanh(F)$ and $\mathcal{F}_{\text{NF}}$ are the free energies of the ferromagnetic and non-ferromagnetic solutions, respectively, calculated from the quenched variational free energy (29) for $\beta = 1$; the latter validates the RS ansatz, used here, as no replica symmetry breaking effect is expected for the Nishimori condition [NS01]. This also implies that the noise threshold of MAP decoding, which corresponds to the zero temperature state in statistical mechanics, is identical to that of the MPM decoding, the performance of which is optimized at Nishimori's temperature, in agreement with results obtained in the IT literature [MN00].

As the exponent $E_q(\rho, \lambda;\, R,\, p)$ is directly evaluated from equation (51) without employing additional inequalities, the optimized bound obtained should be tighter and provide more optimistic lower bounds for noise threshold p_c than that from $E_a(\rho, \lambda;\, R,\, p)$. Clearly one of the main drawbacks of the replica method is the lack of mathematical rigour; recent research [Gue03, Tal03] proved the exactness of results obtained using the replica methods in extensively connected systems. One can hope that similar proofs for diluted systems will follow, making these results much stronger. In any case the difference between the two exponents becomes smaller as $j, k \to \infty$ given a code rate R (table 1).

5.1.5. Reliability exponent. The exponent that represents the fastest decay rate of decoding error probability achievable by the best codes in the ensemble is termed the *reliability exponent* (RE) [Gal68]. The random coding exponent (54) coincides with the RE for relatively high

Table 1. Comparison between different evaluation schemes of the noise threshold p_c for MAP decoding. ANNEAL1 indicates the lower bound of p_c obtained by maximizing $E_a(\rho, \lambda; R, p)$ with respect to ρ keeping $\lambda = 1/(1+\rho)$. Lower bounds for ANNNEAL2 are evaluated by maximizing the same exponent with respect to $\rho \geqslant 0$ and $\lambda \geqslant 0$ without imposing additional conditions; it provides a tighter bound since the optimization with respect to λ, for a fixed ρ, is not commutable with the average over a code ensemble. QUENCH denotes the estimates of p_c obtained from $E_q(\rho, \lambda; R, p)$, evaluated directly from equation (48) using the replica method without employing any extra inequalities; it therefore provides the most optimistic estimate. SHANNON offers critical noise rates p_{sh} at Shannon's limit for given code rates R. The difference in the estimates between the three evaluation schemes becomes smaller as j and k increase, keeping the code rate finite for $j \geqslant 3$. On the other hand, ANNEAL2 and QUENCH generally provide the same estimates for $j = 2$ since p_c for this particular parameter choice is determined by the local instability of the ferromagnetic solution for which the two methods coincidently provide an identical condition, whereas a discontinuous phase transition between the ferro- and paramagnetic solutions determines p_c for $j \geqslant 3$.

R	(j, k)	ANNEAL1	ANNEAL2	QUENCH	SHANNON
1/2	(3, 6)	0.0678	0.0915	0.0998	0.109
2/5	(3, 5)	0.115	0.129	0.136	0.145
1/3	(4, 6)	0.1705	0.1709	0.173	0.174
1/3	(2, 3)	0	0.0670	0.0670	0.174
1/2	(2, 4)	0	0.0286	0.0286	0.109

code rates R. However, for a low code rate region, there still exists a narrow gap between the current tightest lower and upper bounds of the RE, and the exact expression is yet to be determined [MB01, KSNS01, Bar03].

Exact evaluation of RE by improving lower or upper bounds of the error probability, the preferred approach in the IT community, may be difficult since using inequalities has the potential to provide loose bounds. In fact, starting from inequality (47), one cannot improve the bound further, since inequality (47) itself does not provide a tight bound for the low R region [Gal68, KSNS01]. Instead, evaluation based on an *equality* with respect to the error indicator

$$\Delta_{\text{MAP}}(\boldsymbol{n}|H) = \lim_{\beta_\pm \to +\infty, \lambda_\pm \to \pm 1} Z_+^{\lambda_+}(\beta_+|\boldsymbol{n}, H) Z_-^{\lambda_-}(\beta_-|\boldsymbol{n}, H) \tag{60}$$

might provide the exact expression of RE, where $\boldsymbol{n}$ and H are the true noise and parity check matrix, respectively, and

$$\begin{aligned} Z_+(\beta|\boldsymbol{n}, H) &\equiv \sum_{\boldsymbol{n}' \neq \boldsymbol{n}} \prod_{\mu=1}^{N-K} \delta\left(1, \prod_{i \in \mathcal{L}(\mu)} n_i n_i'\right) \times \exp\left(\lambda \beta F \sum_{i=1}^N n_i'\right) \\ Z_-(\beta|\boldsymbol{n}, H) &\equiv \sum_{\boldsymbol{n}'} \prod_{\mu=1}^{N-K} \delta\left(1, \prod_{i \in \mathcal{L}(\mu)} n_i n_i'\right) \times \exp\left(\lambda \beta F \sum_{i=1}^N n_i'\right) \end{aligned} \tag{61}$$

are the two partition sums.

Equation (60) provides an expression for the block error probability

$$P_B(H) = \lim_{\beta_\pm \to +\infty, \lambda_\pm \to \pm 1} \sum_{\boldsymbol{n}} P(\boldsymbol{n}) Z_+^{\lambda_+}(\beta_+|\boldsymbol{n}, H) Z_-^{\lambda_-}(\beta_-|\boldsymbol{n}, H) \tag{62}$$

for a given parity check matrix H. Note that the ability to separate suboptimal solutions from the ferromagnetic solution relies on the gap in the magnetization enumerator that exists for all $p < p_c$ (see figure 8). Furthermore, employing an equality $P_B^* = \min_H\{P_B(H)\} = \lim_{r \to -\infty}(\langle P_B^r(H)\rangle_H)^{1/r}$, a direct expression of RE for a given code ensemble is obtained as

$$E_{\mathrm{RE}}(R,p) = -\frac{1}{N}\ln P_B^* = -\lim_{r\to-\infty,\beta_\pm\to+\infty,\lambda_\pm\to\pm1}\left\{\frac{1}{rN}\right.$$
$$\left.\times\ln\left\langle\left[\sum_{\boldsymbol{n}} P(\boldsymbol{n})Z_+^{\lambda_+}(\beta_+|\boldsymbol{n},H)Z_-^{\lambda_-}(\beta_-|\boldsymbol{n},H)\right]^r\right\rangle_H\right\} \tag{63}$$

which can be evaluated by the replica method, considering $\lambda_\pm$ and r as replica powers.

A recent study in this direction revealed that an expression

$$E_{\mathrm{RE}}(R,p) = \begin{cases}\max_{0<r\leqslant1}\left\{\frac{(1-R)\ln 2}{r} - \frac{1}{r}\ln(1+2^r p^{r/2}(1-p)^{r/2})\right\} & 0\leqslant p\leqslant p_a\\ E_{\mathrm{RC}}(R,p) & p_a < p \leqslant 1\end{cases} \tag{64}$$

is derived for LDPC code ensembles in the limit $j,k\to\infty$, where $E_{\mathrm{RC}}(R,p)$ is the random coding (RC) exponent (54) and p_a a critical noise rate for which $\frac{(1-R)\ln 2}{r} - \frac{1}{r}\ln(1+2^r p^{r/2}(1-p)^{r/2})$ is maximized at $r=1$ [SvMSK03]. It is worthwhile mentioning that this is identical to the existing *lower bound* of the RE evaluated for the ensemble of all possible codes (in expurgated ensembles) [Gal68]. It is well known that LDPC code ensembles for $j,k\to\infty$ have very similar properties to those of the ensemble of all possible codes [MB01, Mac99]; therefore, this result suggests that the existing tightest lower bound of the RE represents the exact expression of the fastest error exponent achievable by the best possible codes, as is widely believed, while a rigorous proof is still sought after [Bar03].

5.2. *Typical set analysis: simpler method for assessing critical noise levels*

Although Gallager's variational method is powerful enough to tightly bound the block error probability of MAP decoding for a wide class of code ensembles, it generally requires rather complicated computation even just for evaluating the noise threshold. In addition, it is quite technical and provides few insights for intuitive understanding of the various types of decoding errors.

Typical set (pairs) analysis is an alternative approach to lower bound the noise threshold for a given code ensemble focusing on *typical set (pairs) decoding*, which is slightly weaker than the MAP decoding scheme (e.g., in rare cases, the true noise may have a higher magnetization than that of the typical set; in such a case the two decoding schemes will differ). Error evaluation in this scheme is relatively easy to understand because occurrences of decoding failure are directly studied using the law of large numbers and the *weight enumerator*; the latter is a standard quantity in the IT literature characterizing the distribution of distances between codewords. This method was pioneered by Shannon for the ensemble of all codes more than 50 years ago [Sha49]; but was not applied to other ensembles until recently. Only since MacKay successfully employed it for analysis of certain LDPC code ensembles, it is now becoming more popular in the IT community [Mac99, AJK01].

5.2.1. Typical sequences and classification of errors. In order to introduce the typical set decoding approach, let us first provide the definition of a noise vector being *typical*. Due to the law of large numbers, a noise vector $\boldsymbol{n}'$ generated by a BSC satisfies a condition

$$\left|\frac{1}{N}\sum_{i=1}^{N} n_i' - p\right| \leqslant \epsilon_N \tag{65}$$

with a high probability for large N and a positive number $\epsilon_N \sim O(N^{-\gamma})(0<\gamma<1/2)$, where $0 < p = (1-\tanh F)/2 < 1/2$ is the flip rate characterizing the BSC. We define as typical any noise vector $\boldsymbol{n}'$ for which this condition holds. We also term the set of all typical vectors the *typical set*.

In typical set decoding one selects a vector that belongs to the typical set and satisfies the parity check equation (12), as a valid noise vector estimate (see also section 4.5). Two types of decoding errors can occur in this decoding scheme: type I error occurs when the true noise vector $\boldsymbol{n}$ is atypical. Type II error occurs when $\boldsymbol{n}$ is typical, and there are multiple typical vectors that satisfy the parity check equation. By a straightforward extension of the law of large numbers, it can be shown that the occurrence probability of type I errors, $P_{\rm I}$, vanishes in the limit $N \to \infty$ [AJK01]. Therefore, the noise threshold p_c is determined only by the condition that probability of type II errors $P_{\rm II}$ vanishes. Since $P_{\rm II}$ depends on each realization of the parity check matrix H, we define p_c for a given code ensemble $\mathcal{C}$ as the highest flip rate below which the average type II error probability $\langle P_{\rm II}(H)\rangle_H$ vanishes in the limit $N \to \infty$.

5.2.2. Lower bound of noise thresholds and weight enumerator. In order to evaluate $\langle P_{\rm II}(H)\rangle_H$, it is convenient to introduce an indicator function $\Delta_{\rm II}(\boldsymbol{n}|H)$ that returns 1, if type II error occurs, and 0 otherwise, for a true noise vector $\boldsymbol{n}$ and parity check matrix H. Then, the type II error probability for a given H is calculated as

$$P_{\rm II}(H) = \sum_{\boldsymbol{n}} P(\boldsymbol{n})\Delta_{\rm II}(\boldsymbol{n}|H) \tag{66}$$

and $\langle P_{\rm II}(H)\rangle_H$ is obtained by averaging this over the code ensemble.

Unfortunately, it is difficult to directly express $\Delta_{\rm II}(\boldsymbol{n}|H)$ in a rigorously treatable form. However, one can easily produce an upper bound

$$\Delta_{\rm II}(\boldsymbol{n}|H) \leqslant \mathcal{V}_{\rm II}(\boldsymbol{n}|H) \times \delta\left(\sum_{i=1}^{N} n_i - N\tanh F\right) \tag{67}$$

in the Ising spin representation, where

$$\begin{aligned}\mathcal{V}_{\rm II}(\boldsymbol{n}|H) &\equiv \sum_{\boldsymbol{n}'\neq\boldsymbol{n}} \prod_{\mu=1}^{N-K} \delta\left(1, \prod_{i\in\mathcal{L}(\mu)} n_i n_i'\right) \delta\left(\sum_{i=1}^{N} n_i' - N\tanh F\right) \\ &= \sum_{\boldsymbol{x}\neq\boldsymbol{1}} \prod_{\mu=1}^{N-K} \delta\left(1, \prod_{i\in\mathcal{L}(\mu)} x_i\right) \delta\left(\sum_{i=1}^{N} n_i x_i - N\tanh F\right).\end{aligned} \tag{68}$$

Since $\Delta_{\rm II}(\boldsymbol{n}|H) = 1$ when errors do occur, it is always upper bounded by the number of solution vectors of the parity check equation (excluding the true noise $\boldsymbol{n}$) that belong to the typical set, $\mathcal{V}_{\rm II}(\boldsymbol{n}|H)$. In the last expression (68), we rewrote the summation over the dummy variable $\boldsymbol{n}'$ using a new variable $\boldsymbol{x} = (x_i) \equiv (n_i'n_i)$; the N-dimensional vector $\mathbf{1}$, with all elements being 1, represents the true noise vector $\boldsymbol{n}$ in the new expression.

Inserting equations (67) and (68) into equation (66), and taking an average over the expurgated (j, k)-Gallager code ensemble (i.e. with no atypically poor codes) in conjunction with the identity $1 = \int \mathrm{d}w\, \delta\big(\sum_{i=1}^{N} x_i - Nw\big)$, an upper bound of the average type II error probability is obtained as

$$\langle P_{\rm II}(H)\rangle_H \leqslant \int \mathrm{d}w \exp[N(-\mathcal{K}(w, p) + \mathcal{R}(w))] \tag{69}$$

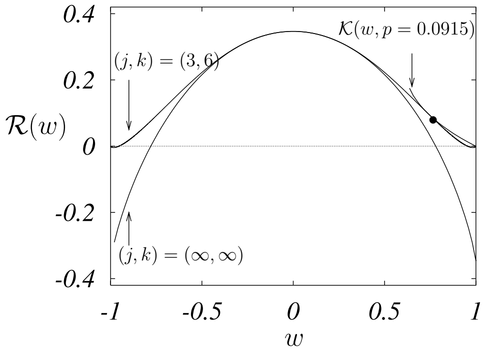


Figure 12. The weight enumerator $\mathcal{R}(w)$ for $(j,k)=(3,6)$ and in the limit of $j,k\to\infty$ keeping the code rate $R=1-j/k=1/2$. For $p=0.0915$, the function $\mathcal{K}(w,p)$ has a contact with the weight enumerator of $(j,k)=(3,6)$ at $w^*\simeq 0.735$, which implies $p_c\geqslant 0.0915$ holds for the $(3,6)$-Gallager code ensemble. $\mathcal{K}(w,p)$ is generally defined only for $1-4p<w\leqslant 1$ and becomes lower as p increases. Therefore, roughly speaking, the lower bound of p_c becomes higher as a code ensemble has a narrower weight enumerator. For a fixed code rate R, the code ensemble of $j,k\to\infty$ has the narrowest possible profile of $\mathcal{R}(w)$, which provides the exact estimate of the noise threshold $p_c=p_{\rm sh}$ where $p_{\rm sh}$ is Shannon's limit that satisfies $R=1-\mathsf{H}_2(p_{\rm sh})$.

where $\mathcal{K}(w,p)$ is derived independently of the code ensemble as $\exp[-N\mathcal{K}(w,p)]\sim\sum_{\boldsymbol{n}}P(\boldsymbol{n})\delta\big(\sum_{i=1}^N n_ix_i-N\tanh F\big)\delta\big(\sum_{i=1}^N n_i-N\tanh F\big)$ imposing a constraint $(1/N)\sum_{i=1}^N x_i=w$; the *weight enumerator*

$$\mathcal{R}(w)=\frac{1}{N}\ln\left\langle\left[\sum_{\boldsymbol{x}\neq\boldsymbol{1}}\prod_{\mu=1}^{N-K}\delta\left(1,\prod_{i\in\mathcal{L}(\mu)}x_i\right)\delta\left(\sum_{i=1}^N x_i-Nw\right)\right]\right\rangle_H \qquad (70)$$

characterizes the code ensemble. Equation (69) implies that $\langle P_{\rm II}(H)\rangle_H$ vanishes in the limit $N\to\infty$ as long as $\max_w\{-\mathcal{K}(w,p)+\mathcal{R}(w)\}<0$, which yields a lower bound for p_c.

The meaning of the exponent in the right-hand side of equation (69) is intuitively understandable by considering the mechanism that gives rise to a decoding failure. Firstly, $\exp[-N\mathcal{K}(w,p)]$ represents the probability that a 'gauged noise vector' $\boldsymbol{n}+\boldsymbol{x}$ (mod 2) is typical, as well as the true noise vector $\boldsymbol{n}$, under the condition that the number of non-zero elements of $\boldsymbol{x}$, $\sum_{i=1}^N x_i$, is constrained to $N(1-w)/2$ (also termed *weight* in this Boolean representation). In practice, a codeword vector $\boldsymbol{t}=G^T\boldsymbol{s}\,(\mathrm{mod}\,2)$, alternatively characterized by the equation $H\boldsymbol{t}=H(G^T\boldsymbol{s})=0\,(\mathrm{mod}\,2)$, plays the role of $\boldsymbol{x}$; a type II error occurs if both $\boldsymbol{n}$ and the gauged vector $\boldsymbol{n}+\boldsymbol{x}$ (mod 2) become typical because there are at least two typical noise vectors satisfying the parity check equation. However, this just provides an error probability caused by a single codeword $\boldsymbol{x}$. Therefore, secondly, we have to evaluate the number of codewords that have a weight w, which is provided by the weight enumerator $\mathcal{R}(w)$. Multiplying this number of codewords to $\exp[-N\mathcal{K}(w,p)]$ and taking a summation over the possible weight w, we finally obtain equation (69).

In the bound (69), all relevant properties of the code ensemble are represented by the weight enumerator $\mathcal{R}(w)$. This function is maximized to $R\ln 2$ at $w=0$, in general, and has a mirror symmetry $\mathcal{R}(-w)=\mathcal{R}(w)$, in particular, for even k. Pictorially, the lower bound of p_c can be obtained through the value for which $\mathcal{K}(w,p)$ makes contact with $\mathcal{R}(w)$ (somewhat similar to the magnetization enumerator of figure 8) at a certain point w^*, marked by (●) in figure 12. This can be analytically performed in the case of $j,k\to\infty$ as $\mathcal{R}(w)$ can be expressed analytically, providing Shannon's limit $p_{\rm sh}$ as a lower bound for p_c. However,

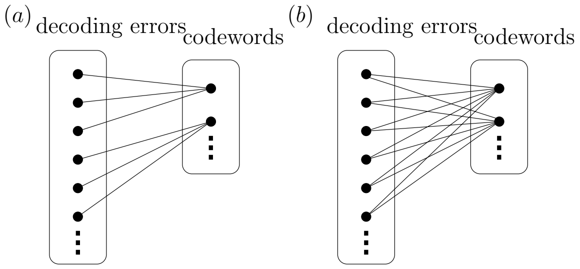


Figure 13. A possible shortcoming of typical set analysis. (*a*) If each decoding error in noise estimation were associated with a single codeword, a simple product $\exp[-N\mathcal{K}(w^*, p)] \times \exp[N\mathcal{R}(w^*)]$ would have correctly evaluated $\langle P_{\mathrm{II}}(H)\rangle_H$. (*b*) However, when a single decoding error is associated with multiple codewords, the product overestimates $\langle P_{\mathrm{II}}(H)\rangle_H$.

p_{sh} also serves as the upper bound of p_c for any code ensembles, this means that $p_c = p_{\mathrm{sh}}$ indicating that the Gallager code saturates Shannon's limit when $j, k \to \infty$.

Thus, typical set analysis can exactly evaluate p_c of the Gallager code ensembles in the limit $j, k \to \infty$. Unfortunately, this may not necessarily be the case for finite j, k. It can be shown that the lower bounds of p_c offered by the typical set analysis are the same as those obtained by Gallager's methodology for MAP decoding [AJK01], which in itself provides more pessimistic evaluations than the replica method as shown in table 1. The gap between SM and typical set analysis results may be attributed to the different decoding schemes used. However, one can show that the replica method yields more optimistic lower bounds for p_c also when typical set decoding is used, which implies that evaluation of the noise threshold utilizing typical set analysis is rigorous but not tight enough for finite j, k.

5.2.3. Improving the bound by the replica method. A possible shortcoming of typical set analysis relates to the upper bounding of the average type II error probability by a product of the error probability caused by a single codeword ($\exp[-N\mathcal{K}(w, p)]$) and the number of codewords ($\exp[N\mathcal{R}(w)]$), focusing on the most relevant weight $w = w^*$. This bound would have been tight if each codeword brought about estimation errors *exclusively* (i.e. each noise vector estimation error is generated by a different codeword). However, since each noise vector estimation error may be associated with multiple codewords belonging to the same codebook, the simple product $\exp[-N\mathcal{K}(w^*, p)] \times \exp[N\mathcal{R}(w^*)]$ may overestimate the correct type II error probability (figure 13). Therefore, it is necessary to take correlations between multiple codewords associated with a single error into account in order to improve the evaluation of p_c.

An analysis based on an *equality* with respect to the error indicator

$$\Delta_{\mathrm{II}}(\boldsymbol{n}|H) = \lim_{\rho\to+0} \mathcal{V}_{\mathrm{II}}^{\rho}(\boldsymbol{n}|H) \left(\sum_{i=1}^{N} n_i - N \tanh F\right) \tag{71}$$

might naturally introduce such correlations as

$$\mathcal{V}_{\mathrm{II}}^{\rho}(\boldsymbol{n}|H) = \left(\sum_{x\neq\mathbf{1}} \prod_{\mu=1}^{N-K} \delta\left(1, \prod_{i\in\mathcal{L}(\mu)} x_i\right) \delta\left(\sum_{i=1}^{N} n_i x_i - N \tanh F\right)\right)^{\rho}$$

creating certain interactions among 'codeword vectors' $\boldsymbol{x}$. Substituting equation (71) into equation (66) and taking an average over the code ensemble provides an equality

$$\langle P_{\mathrm{II}}(H)\rangle_H = \lim_{\rho\to+0} \exp[-N E_{\mathrm{II}}(\rho; R, p)] \tag{72}$$

where

$$E_{\rm II}(\rho;\, R,\, p) = -\frac{1}{N}\ln\left\langle\left[\sum_{n} P(\boldsymbol{n})\mathcal{V}_{\rm II}^{\rho}(\boldsymbol{n}|H)\delta\left(\sum_{i=1}^{N} n_i - N\tanh F\right)\right]\right\rangle_H \tag{73}$$

which can be evaluated by the replica method. Equation (72) indicates that p_c can be assessed from the limit where $\lim_{\rho\to+0} E_{\rm II}(\rho;\, R,\, p)$ becomes positive.

A recent study showed that noise thresholds obtained by the SM typical set decoding scheme are identical to those assessed by the replica approach to MAP decoding [KNvM02]. This indicates that differences of error correction abilities between the typical set and MAP decoding schemes are relatively small and vanish in the limit of long message lengths.

6. Applications of LDPC codes

So far we have focused on LDPC as error-correcting codes. However, coding techniques are required for various purposes in digital communication. In this section, we mention how LDPC codes can be utilized for various purposes, other than simple error correction.

6.1. Lossless data compression

Data compression, or source coding, is a scheme to reduce the message size (data) by modifying the information representation. This is usually carried out prior to transmission in order to optimize communication efficiency by minimizing the data to be sent. The possibility of data compression was first pointed by Shannon in his celebrated *source coding theorem* [Sha48]. He showed that for an information source represented by a distribution $P(\boldsymbol{s})$ of N-dimensional Boolean vectors $\boldsymbol{s}$, one can employ another representation of $K(\leqslant N)$ dimensions without any distortion, if the code rate $R = K/N$ satisfies $R \geqslant \mathsf{H}_2(\mathcal{S})$ in the limit $K, N \to \infty$, where $\mathsf{H}_2(\mathcal{S}) \equiv -(1/N)\sum_{\boldsymbol{s}} P(\boldsymbol{s})\log_2 P(\boldsymbol{s})$ denotes the binary entropy per bit of the source ($\mathcal{S}$) distribution $P(\boldsymbol{s})$. On the other hand, it can also be shown that such reduction is impossible when $R < \mathsf{H}_2(\mathcal{S})$. Therefore, $\mathsf{H}_2(\mathcal{S})$ represents the optimal compression rate, or *compression limit*.

Unfortunately, the source coding theorem is non-constructive and suggests few clues for designing good practical compression methods. However, after much effort, a practical code that asymptotically saturates the optimal limit was finally discovered more than a decade later [Jel68]. Therefore, the compression scheme based on LDPC codes presented below may not compete with existing good practical codes such as the arithmetic codes [Jel68] and Lempel–Ziv (LZ) compression [ZL77]. Nevertheless, this still serves as a useful prototype for constructing a more advanced compression scheme used in network communication [SW73, Mur02], described in the following section.

In order to compress an N-dimensional Boolean source vector $\boldsymbol{s}$ to a $K(< N)$-dimensional codeword $\boldsymbol{z}$ on the basis of an LDPC scheme, let us introduce a $K \times N$ sparse Boolean matrix H with j and k non-zero elements per column and row, respectively. Using this matrix, one can compress $\boldsymbol{s}$ to a shorter vector $\boldsymbol{z}$ by

$$\boldsymbol{z} = H\boldsymbol{s} \pmod 2. \tag{74}$$

On the other hand, decoding $\boldsymbol{z}$ to retrieve the original representation $\boldsymbol{s}$ is performed with knowledge of the source distribution $P(\boldsymbol{s})$ utilizing the posterior distribution

$$P(\boldsymbol{\sigma}|\boldsymbol{z}) = \frac{P(\boldsymbol{\sigma})\delta(\boldsymbol{z} = H\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta(\boldsymbol{z} = H\boldsymbol{\sigma})} \tag{75}$$

which can be practically carried out employing the BP/TAP algorithm.

Similarly to the case of error correction, the performance of this scheme can be evaluated utilizing the replica method [Mur02]. In the Ising spin representation, the free energy per element can be evaluated from

$$\mathcal{F} = \operatorname*{Extr}_{\pi(\cdot),\hat{\pi}(\cdot)} \Bigg\{ -\frac{j}{k}\left\langle \ln\left(\frac{1+\prod_{i=1}^{k} x_i}{2}\right)\right\rangle_{\pi} + j\left\langle \ln\left(\frac{1+\hat{x}x}{2}\right)\right\rangle_{\hat{\pi},\pi} - \frac{1}{N}\left\langle \ln\left[\sum_{\boldsymbol{\sigma}}\left(\prod_{i=1}^{N}\prod_{\mu=1}^{j}\left(\frac{1+\hat{x}_{\mu i}\tau_i}{2}\right)\right)P(\boldsymbol{\sigma}\otimes\boldsymbol{s})\right]\right\rangle_{\hat{\pi},\boldsymbol{s}} \Bigg\} \tag{76}$$

under the RS ansatz, where $\boldsymbol{\sigma}\otimes\boldsymbol{s} = (\sigma_i s_i)$ $(i = 1, 2, \ldots, N)$ stands for source vectors gauged by the true source vector $\boldsymbol{s}$ in the Ising spin expression and $P(\boldsymbol{\sigma}\otimes\boldsymbol{s})$ represents the source distribution in this expression. $\langle\cdots\rangle_{\boldsymbol{s}}$ denotes an average over the source distribution.

For $j \geqslant 3$, the ferromagnetic solution $\pi_{\mathrm{F}}(x) = \delta(x-1)$ and $\hat{\pi}_{\mathrm{F}}(\hat{x}) = \delta(\hat{x}-1)$, which represents decoding success, always extremizes the free energy (76) to

$$\mathcal{F}_{\mathrm{F}} = -\frac{1}{N}\sum_{\boldsymbol{s}} P(\boldsymbol{s})\ln P(\boldsymbol{s}) = \mathsf{H}_2(\mathcal{S})\ln 2. \tag{77}$$

In addition to this, another solution, which stands for decoding failure, appears when R is below a certain critical rate R_d, which is determined by j and k. For finite j, this solution is obtained only numerically. However, this solution can be analytically expressed as $\pi_{\mathrm{NF}}(x) = \delta(x)$ and $\hat{\pi}_{\mathrm{NF}}(\hat{x}) = \delta(\hat{x})$ in the case of $j, k \to \infty$ under the fixed code rate. Inserting this solution into equation (76) provides the free energy

$$\mathcal{F}_{\mathrm{NF}} = \frac{j}{k}\ln 2 = R\ln 2. \tag{78}$$

This, in conjunction with equation (77), means that the decoding success solution is thermodynamically dominant and, therefore, the original expression $\boldsymbol{s}$ is potentially decodable from the compressed vector $\boldsymbol{z}$ for $R \geqslant \mathsf{H}_2(\mathcal{S})$ and an arbitrary source distribution $P(\boldsymbol{s})$. This implies that the current scheme achieves Shannon's compression limit for $j, k \to \infty$.

However, this does not imply that $\boldsymbol{z}$ can be decoded in practical time scales. The BP/TAP algorithm is likely to be trapped in suboptimal solutions for $R < R_d$; the compression limit for practical decoding is therefore provided by R_d which is always higher than a critical rate R_c, determined by the thermodynamic transition between the decoding success and failure solutions. Roughly speaking, as j grows at a fixed rate $R = j/k$, R_c decreases, while R_d increases. In particular, in the case of $j \to \infty$, the potential and practical limits R_c and R_d converge to $\mathsf{H}_2(\mathcal{S})$ and 1, respectively, which means that the current scheme is impractical in this limit although the theoretical performance can saturate Shannon's limit.

On the other hand, other existing schemes such as the LZ codes are executable on practical time scales and asymptotically achieve the compression limit even if details of the source distribution are unknown [ZL77]. Therefore, the LDPC-based compression scheme may not be competitive when used for the purpose of simple noiseless data compression.

6.2. *Lossless compression of distributed sources*

Although the practical significance of the LDPC-based scheme seems weaker than that of existing state-of-the-art methods for simple lossless compression, it may not be the case for more advanced problems. This is because optimal strategies sometimes cannot be employed when conditions change. A data compression problem of distributed sources, first addressed by Slepian and Wolf for data transmission in a network [SW73], offers one such example.

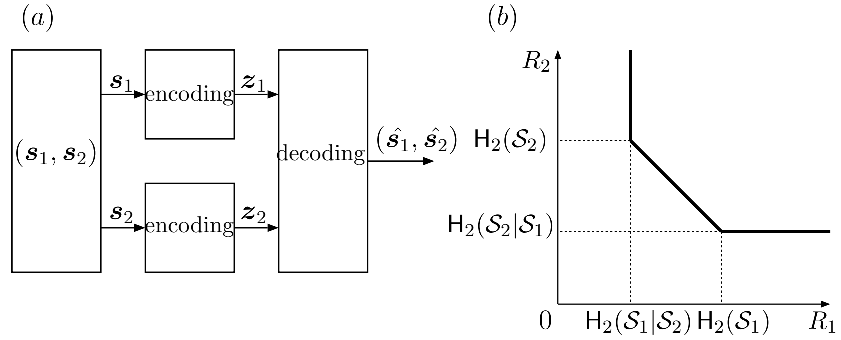

(a)
(b)
(s_1, s_2)
s_1
s_2
encoding
encoding
z_1
z_2
decoding
$(\hat{s}_1, \hat{s}_2)$
R_2
$\mathsf{H}_2(\mathcal{S}_2)$
$\mathsf{H}_2(\mathcal{S}_2|\mathcal{S}_1)$
0
$\mathsf{H}_2(\mathcal{S}_1|\mathcal{S}_2)$
$\mathsf{H}_2(\mathcal{S}_1)$
R_1

Figure 14. (*a*) The Slepian–Wolf system. Encoding is carried out independently at distributed sites, whereas decoding is simultaneously performed by a single user. (*b*) The achievable limit of the Slepian–Wolf system.

Let us assume that two correlated source vectors s_1 and s_2 of N dimensions are generated from a joint source distribution $P(s_1, s_2)$. In a general scenario of the Slepian–Wolf problem, s_1 and s_2 (from sources $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively) are *independently* compressed to K_1- and K_2-dimensional vectors z_1 and z_2, respectively. On the other hand, a *single* decoder *simultaneously* retrieves the original expressions s_1 and s_2 from the codewords z_1 and z_2 utilizing the knowledge of $P(s_1, s_2)$ at the decoding stage (figure 14(*a*)). For instance, this kind of problem arises when two satellites covering overlapping regions transmit digital images to a single base station on earth.

It is clear that a region specified by $R_1 = K_1/N \geqslant \mathsf{H}_2(\mathcal{S}_1)$ and $R_2 = K_2/N \geqslant \mathsf{H}_2(\mathcal{S}_2)$ is achievable without any distortion by optimal compression codes for a single source, dealing with s_1 and s_2 as vectors that independently follow marginal distributions $P(s_1) = \sum_{s_2} P(s_1, s_2)$ and $P(s_2) = \sum_{s_1} P(s_1, s_2)$, respectively. However, Slepian and Wolf showed that the achievable region can be further extended potentially as

$$R_1 \geqslant \mathsf{H}_2(\mathcal{S}_1|\mathcal{S}_2) \qquad R_2 \geqslant \mathsf{H}_2(\mathcal{S}_2|\mathcal{S}_1) \qquad R_1 + R_2 \geqslant \mathsf{H}_2(\mathcal{S}_1, \mathcal{S}_2) \tag{79}$$

(figure 14(*b*)) if the knowledge of the joint distribution $P(s_1, s_2)$ is fully utilized, where $\mathsf{H}_2(\mathcal{S}_1, \mathcal{S}_2) = -(1/N)\sum_{s_1,s_2} P(s_1, s_2)\log_2 P(s_1, s_2)$, $\mathsf{H}_2(\mathcal{S}_1|\mathcal{S}_2) = \mathsf{H}_2(\mathcal{S}_1, \mathcal{S}_2) - \mathsf{H}_2(\mathcal{S}_2)$ and similarly for $\mathsf{H}_2(\mathcal{S}_2|\mathcal{S}_1)$. Unfortunately, it is difficult to achieve this limit by the optimal codes for a single source since incorporating the correlation between s_1 and s_2 with such schemes is generally non-trivial.

On the other hand, the LDPC-based compression scheme is easily extended for the distributed source by using the LDPC matrices H_1 and H_2, of dimensionalities $K_1 \times N$ and $K_2 \times N$, respectively, such that

$$z_1 = H_1 s_1 \pmod 2 \qquad z_2 = H_2 s_2 \pmod 2. \tag{80}$$

In this scheme, one can directly incorporate the source distribution $P(s_1, s_2)$ in the decoding stage through the Bayes formula

$$P(\sigma_1, \sigma_2|z_1, z_2) = \frac{P(\sigma_1, \sigma_2)\delta(z_1 = H_1\sigma_1)\delta(z_2 = H_2\sigma_2)}{\sum_{\sigma_1,\sigma_2} P(\sigma_1, \sigma_2)\delta(z_1 = H_1\sigma_1)\delta(z_2 = H_2\sigma_2)}. \tag{81}$$

Murayama showed that this scheme achieves the Slepian–Wolf limit (79) when the numbers of non-zero elements per column/row in H_1 and H_2 become infinite [Mur02]. Furthermore, he illustrated that utilizing LDPC matrices of finite non-zero elements per column/row, practical decoding by BP/TAP becomes possible beyond the single source coding limit $R_1 \geqslant \mathsf{H}_2(\mathcal{S}_1)$

and $R_2 \geqslant \mathsf{H}_2(\mathcal{S}_2)$ for certain distributed sources; this implies that LDPC-based compression schemes may be a promising direction for distributed data compression problems of this type.

6.3. Lossy data compression

The source coding theorem indicates that it is impossible to reduce the size of data below the compression limit without allowing for any distortion. However, if a certain level of distortion is allowed, one can further reduce the data size. Compression of this type is termed *lossy compression*. JPEG and MPEG, which are examples of current standard schemes in use for compressing data of images and movies, fall into this category.

In general, as the allowed distortion increases, the achievable data size decreases; namely, there is a trade-off between the optimal compression rate and the distortion, which is provided by the *rate-distortion theorem* presented by Shannon more than a decade after the source coding theorem [Sha59].

Unlike lossless compression, no practical algorithm capable of saturating the optimal performance predicted by the rate-distortion theory is known for lossy compression, even for simple information sources. Therefore, the quest for better lossy compression codes remains one of the important research areas in IT [YZB97].

Let us here focus on a simple lossy data compression problem of an unbiased Boolean source of N-dimensional vector $\boldsymbol{s}$, the distribution of which is assumed uniform $P(\boldsymbol{s}) = 1/2^N$. The *distortion function* $d(\boldsymbol{s}, \tilde{\boldsymbol{s}})$ is used to evaluate the distortion, where $\tilde{\boldsymbol{s}}$ is an N-dimensional representative Boolean vector used to approximate $\boldsymbol{s}$ with a reduced information content. Here, we employ the Hamming distance

$$d(\boldsymbol{s}, \tilde{\boldsymbol{s}}) = \sum_{i=1}^{N} (1 - \delta_{s_i, \tilde{s}_i}) \tag{82}$$

where $\delta_{x,y} = 1$ if $x = y$ and 0, otherwise.

In the current case, the *lossless* compression limit is given by the binary entropy per bit of the source distribution $R \geqslant \mathsf{H}_2(\mathcal{S}) = -(1/N) \sum_{\boldsymbol{s}} 2^{-N} \log_2 2^{-N} = 1$, which implies that it is impossible to reduce the size of the data any more without allowing some level of distortion. However, when a distortion up to ND measured by the Hamming distance is allowed, it can be shown analytically that one can compress $\boldsymbol{s}$ into a $K = NR$-dimensional Boolean vector $\boldsymbol{z}$ if $R \geqslant R(D)$, where

$$R(D) = 1 - \mathsf{H}_2(D) \tag{83}$$

is termed the *rate-distortion function* of the current unbiased Boolean source [CT91]; such analytical expressions of the rate-distortion functions are not known for most other sources.

In order to devise a lossy compression scheme, it is necessary to appropriately design a map from the compressed expression $\boldsymbol{z}$ to the representative vector $\tilde{\boldsymbol{s}}$. One possible construction of this map is to employ an $N \times K$ LDPC matrix H such that

$$\tilde{\boldsymbol{s}} = \tilde{\boldsymbol{s}}(\boldsymbol{z}) = H\boldsymbol{z} \pmod 2. \tag{84}$$

Then, given an N-dimensional source vector $\boldsymbol{s}$, encoding is carried out by selecting such a vector $\boldsymbol{z}$ that satisfies the distortion constraint $d(\boldsymbol{s}, \tilde{\boldsymbol{s}}(\boldsymbol{z})) \leqslant ND$ as the compressed representation of $\boldsymbol{s}$. On the other hand, one can easily decode $\boldsymbol{z}$ to approximate the original vector $\boldsymbol{s}$ employing equation (84). It can be shown that this scheme saturates the rate-distortion function (83) when the numbers of non-zero elements per column/row of H become infinite [MO03, MY02].

One shortcoming of this LDPC-based scheme in the current suggestion is the computational difficulty at the encoding stage. Since finding $\boldsymbol{z}$ for a given $\boldsymbol{s}$, where both are discrete variables, is a non-trivial search problem that becomes practically difficult as the message length N increases. A naive use of the BP/TAP approach does not serve as a satisfiable approximation algorithm in this case since encoding requires selection of a single vector $\boldsymbol{z}$, whereas the BP/TAP method generally calculates variable averages over the posterior distribution in which clues for selecting a single vector are erased. However, this difficulty may be resolved by certain advanced methods [MPZ02] although further investigation is necessary.

Another drawback of the current method is the difficulty in directly extending the scheme to biased sources. It can be shown that for a uniformly biased Boolean source characterized by $P(\boldsymbol{s}) = \prod_{i=1}^{N} p^{s_i}(1-p)^{1-s_i}$, where $0 \leqslant p \leqslant 1$, the rate-distortion function (83) is modified to

$$R(D) = \begin{cases} \mathsf{H}_2(p) - \mathsf{H}_2(D) & \text{for} \quad 0 < D < p \\ 0 & \text{for} \quad p \leqslant D \leqslant 1 \end{cases} \tag{85}$$

which indicates that the data size can be reduced further than equation (83) for biased sources because the original message distribution in itself includes some redundancy. This limit can be achieved by appropriately constructing biased representative vectors that approximate the biased vectors with the required distortion using as little information as possible. However, as addition modulo 2 generally reduces the statistical bias of each bit, construction of such representative vectors by a linear map (84) is difficult; this prevents the current method from saturating the rate-distortion function of biased source (85). In a recent study [HKN02], this difficulty has been resolved by replacing the linear map (84) with a non-linear map constructed by perceptrons which are characterized by non-monotonic transfer functions of a specific type [vMWB00].

6.4. Error correction in a broadcast channel

As most existing codes are constructed for simple point-to-point communication, they do not necessarily offer the optimal performance for multi-terminal communication such as the Internet, mobile phones and satellite communication. Designing codes that utilize characteristic features of these communication channels may enhance their performance; this would greatly benefit overloaded communication channels that suffer from an ever increasing information flow.

The *broadcast channel*, which models television and radio broadcasting, is one of the most fundamental examples of multi-terminal communication [CT91]. We here show how LDPC codes can be utilized for improving the communication performance in a broadcasting set-up.

In a general scenario, a single sender (station) broadcasts a codeword composed of different messages to multiple receivers. For simplicity, we focus on the case of two receivers; a single codeword $\boldsymbol{t}$ of N bits, comprising two messages $\boldsymbol{s}_1$ (NR_1 bits) and $\boldsymbol{s}_2$ (NR_2 bits), is transmitted to two receivers. As each channel is noisy, receivers 1 and 2 obtain two corrupted codewords $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$, respectively, which is modelled by a conditional probability $P(\boldsymbol{r}_1, \boldsymbol{r}_2|\boldsymbol{t})$. The received codewords are decoded by respective receivers to retrieve only the message addressed to each of them.

Combining codes is a known empirical strategy for designing high performance communication schemes for broadcast channels on the basis of multiple linear error-correcting codes of relatively short message lengths [MS77, vG83, vG84]. Inspired by this, the

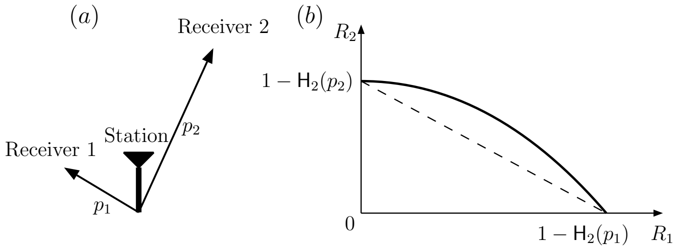


Figure 15. (*a*) A broadcast channel of a single station and two receivers. (*b*) A schematic profile of Cover's limit (thick full curve). The dashed line indicates the time-sharing limit achievable by concatenating two independent codes.

performance of a linearly combined coding scheme was recently examined for LDPC codes [NKMZS03]. The code is specified by a parity check matrix of upper triangular form

$$H = \begin{pmatrix} H_1 & H_2 \\ 0 & H_3 \end{pmatrix} \tag{86}$$

where the sizes of sub-matrices H_1, H_2, H_3 are $[(1-\alpha)N - R_1 N] \times (1-\alpha)N$, $[(1-\alpha)N - R_1 N] \times \alpha N$ and $[\alpha N - R_2 N] \times \alpha N$, respectively.

Based on this matrix, the generator matrix G^T is constructed as

$$G^T = \begin{pmatrix} G_1^T & G_2^T \\ 0 & G_3^T \end{pmatrix} \tag{87}$$

where $G_i^t (i = 1, 3)$ are systematically designed so as to satisfy $H_i G_i^T = 0 \,(\mathrm{mod}\, 2)$ and $G_2^T = -H_1^T \left[H_1 H_1^T\right]^{-1} \left[H_2 G_3^T\right]$. In this scheme, two messages are encoded into a single codeword using G^T as $\boldsymbol{t} = G^T (\boldsymbol{s}_1 \boldsymbol{s}_2)^T \,(\mathrm{mod}\, 2)$. On the other hand, two corrupted codewords $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$ are independently decoded by each receiver solving the parity check equations $\boldsymbol{z}_i = H\boldsymbol{r}_i = H\boldsymbol{n}_i \,(\mathrm{mod}\, 2)$ $(i = 1, 2)$.

Analogous to the case of single channels, error free communication becomes theoretically possible if the corresponding code rate vector (R_1, R_2) is placed within a certain convex region, which is termed the *capacity region*, when the code length grows infinite. In particular, the capacity region can be analytically expressed as

$$R_2 < 1 - \mathsf{H}_2(\delta * p_2) \qquad R_1 < \mathsf{H}_2(\delta * p_1) - \mathsf{H}_2(p_2) \tag{88}$$

where the noise models for receivers 1 and 2 are assumed as BSC specified by flip rates p_1 and $p_2 \,(< p_1)$, respectively. Here, we introduce the notation $\delta * p = \delta(1-p) + (1-\delta)p$. Equation (88) is often termed *Cover's capacity*, depicted by a solid curve in figure 15. Unfortunately, the derivation of Cover's capacity is non-constructive and offers few clues for designing efficient practical codes. Furthermore, even achieving the *time-sharing capacity* (a dotted straight line in figure 15), which is theoretically achievable by simple concatenation of two independent codewords, separately optimized for each channel, is in practice never trivial, as there are no known codes that saturate Shannon's bound even for a single channel.

A statistical mechanics based analysis for the broadcast channel of this type reveals that the suggested linearly combined LDPC coding scheme provides an improved performance over the simple concatenation method, in both potential and practical limits, when the number of non-zero elements per column/row in the parity check matrix is finite [NKMZS03]. Unfortunately, it was also shown that the optimal performance achievable by this scheme cannot go beyond the

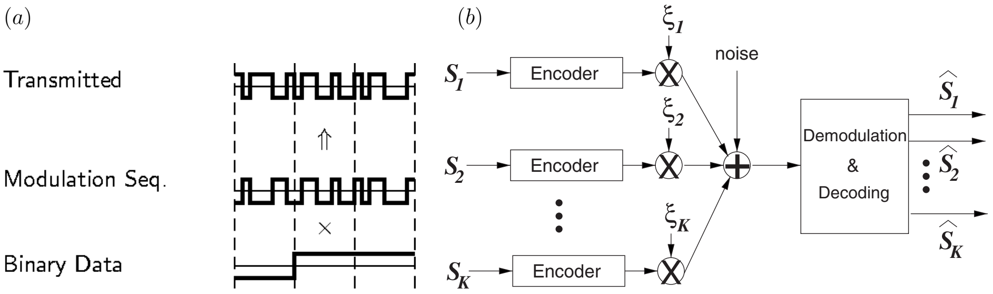


Figure 16. (*a*) Modulation in conventional CDMA, where random modulation sequences are used to generate the transmitted signal from the original message. (*b*) LDPC coding of the source sequences s_i prior to modulation by random modulation sequences ξ_i. Demodulation and decoding provide the estimates $\hat{s}_i$.

time-sharing capacity even theoretically. This analysis implies that different coding schemes such as non-linear codes should be examined for achieving Cover's limit.

6.5. LDPC for CDMA

Multiple access communication is at the opposite end to broadcasting, where multiple sources transmit simultaneously to a single receiver; the task of the receiver is to separate the combined (possibly corrupted) signal and retrieve the original sources. Several methods can be used for separating the sources; two obvious solutions are for the different sources to transmit at different times or using different frequencies [Ver98]. A different, arguably more efficient, approach is based on code division multiple access (CDMA), where messages are encoded prior to transmission.

Conventional modulation techniques are based on modulating each signal by a random modulation vector shown schematically in figure 16(*a*). Demodulation is then carried out by multiplying the received signal by the modulation sequence for each source and estimating the original message. A statistical mechanics based analysis of conventional CDMA modulation was recently introduced by Tanaka [Tan02].

The idea of combining LDPC codes with CDMA systems was originally introduced in [dBD03, dBD02, ADU03]. The idea is to encode the messages by different LDPC codes prior to the modulation stage as described schematically in figure 16(*b*). Results obtained by computer simulations, and after carefully designing LDPC codes by DE, show excellent performance [ADU03]. However, these studies are limited to cases where the number of users is $O(1)$ (one exception is in [RCGV02], where the number of users is expected to be large; however, it relies on the *assumption* of near-capacity-approaching LDPC codes being available).

A recent study [TS03b, TS03a] offers a statistical mechanics based analysis of the joint detection/decoding for a LDPC-coded CDMA system in the large-system limit. The analysis provides both practical and theoretical limitations of the suggested method obtained from the statistical mechanics based analysis, in the form of dynamical and thermodynamical transition points, respectively. The results reported indicate that while the theoretical limits of the new methods are excellent, the practical performance is limited by a relatively low dynamical transition point [TS03b, TS03a]. However, the analysis was carried out for regular LDPC codes; it is highly likely that practical performance can be pushed close to the theoretical limits by clever irregular code designs.

$$(a)\quad \boldsymbol{s} \to \boldsymbol{r}\text{: Easy} \qquad (b)\quad \boldsymbol{r} \to \boldsymbol{s}\text{:}\begin{cases}\text{Easy with a secret key}\\ \text{Hard without a secret key}\end{cases}$$

Figure 17. Required properties of a public key cryptosystem. (*a*) A plain text $\boldsymbol{s}$ is encrypted into a cipher text $\boldsymbol{r}$ using the public key with a low computational cost. (*b*) Decryption of the cipher text $\boldsymbol{r}$ is computationally hard without utilizing the secret key, while it can be easily carried out if the secret key is available.

6.6. Public key cryptography

Public-key cryptography plays an important role in many aspects of modern information transmission, for instance, in the areas of electronic commerce and Internet-based communication. It makes it possible for the service provider to distribute a public key which may be used to encrypt messages in a manner that can only be decrypted by the service provider [DH76] (figure 17). The on-going quest for safer and more efficient cryptosystems produced many useful methods over the years such as the Rivest–Shamir–Adleman (RSA) [RSA78], ElGammal [ElG85] and McEliece cryptosystems [McE78] to name but a few. We here show that another example of such a system, which is somewhat similar to the one presented by McEliece, can be devised on the basis of significantly different behaviour for LDPC codes of the MN- and Sourlas-types [KMS00a, SKM01].

In the suggested cryptosystem, a plaintext represented by a K-dimensional Boolean vector $\boldsymbol{s}$ is encrypted to the N-dimensional Boolean ciphertext $\boldsymbol{r}$ utilizing a predetermined Boolean matrix G^T of dimensionality $N \times K$, and a corrupting N-dimensional vector $\boldsymbol{n}$, the elements of which become 1 with probability p and 0, otherwise, in the following manner

$$\boldsymbol{r} = G^T \boldsymbol{t} + \boldsymbol{n} \pmod 2. \tag{89}$$

The matrix G^T and the flip probability p constitute the *public key*. The corrupting vector $\boldsymbol{n}$ is generated in the transmitting terminal.

The matrix G^T, which is at the centre of the encryption/decryption process, is constructed by randomly choosing a $K \times K$ dense invertible matrix D and two randomly selected LDPC matrices A (of dimensionality $N \times K$) and B (of dimensionality $N \times N$ and invertible), via $G^T = B^{-1}AD \pmod 2$. Similarly for the MN codes, the matrices A and B are characterized by j and l non-zero elements per column and k and l non-zero elements per row, respectively, in the simplest case, whereas irregular construction using varying k, j and l for each column/row may also be considered. The parameters j, k and l define a particular cryptosystem while the matrices A, B and D constitute the *private* key.

The authorized user may decrypt the ciphertext $\boldsymbol{r}$ in a similar manner to the MN codes. Namely, a parity check equation of the form

$$\boldsymbol{z} = B\boldsymbol{r} = A(D\boldsymbol{s}) + B\boldsymbol{n} \pmod 2 \tag{90}$$

which is offered by multiplying the ciphertext $\boldsymbol{r}$ (89) by the private key B, is first solved for $\tilde{s} = D\boldsymbol{s}$ using the BP/TAP algorithm. Due to properties of the MN codes, this is easy if p is set below the dynamical transition point p_d that is determined by the set of (j, k, l). After that, the plain text is finally retrieved as $\boldsymbol{s} = D^{-1}\tilde{s}$.

On the other hand, an unauthorized user must extract $\boldsymbol{s}$ from equation (89) knowing only the ciphertext $\boldsymbol{r}$ and the public key (G^T, p). The first straightforward attempt to enumerate all possible $\boldsymbol{s}$ is clearly doomed, unless p is vanishingly small, enough to corrupt just a few bits. Decomposing G^T into a combination of sparse and dense matrices is known to belong to a class of NP-complete problems [GJ79].

Another approach is to approximately decrypt $\boldsymbol{r}$ using the BP/TAP scheme, which yields an effectively identical decoding problem to that of the Sourlas-type codes, with the generator matrix G^T being dense. However, due to properties of the Sourlas codes, finding solutions to equation (89) is strongly dependent on initial conditions. In particular, when G^T is dense, which is the case in the current problem, for all initial conditions other than the plaintext itself, the BP/TAP algorithm fails to converge to the plaintext solution [KMS00a, Mac99, KS87]. Obtaining the correct solution for equation (89) without knowledge of the private key will therefore become unfeasible, which implies that decryption by unauthorized users is practically impossible. Several attacks by unauthorized parties who have acquired partial knowledge of private key components and/or of the plaintext have been recently studied, showing that the cryptosystem is fairly secure [SSK03].

Before closing this section, it may be worthwhile to briefly compare the current LDPC-based method to the leading public key cryptosystem of RSA [RSA78]. RSA decryption takes $O(K^3)$ operations while the current method naively requires $O(K^2)$ operations, which can be further reduced to $O(K \log K)$ by constructing a dense matrix D as a product of random permutation and triangular matrices. From this aspect, the LDPC-based scheme may be superior to the RSA cryptosystem. Encryption cost is $O(K^2)$, which is similar to that of RSA, whereas inverting the matrices B and D is carried out only once and is of $O(K^3)$. A major drawback of the current method is the size of the public key. Since G^T is a dense matrix, the size of the public key is of $O(N \times K)$, while that for RSA is only $O(K)$. However, as the transmission of the public key is carried out only once, this may not be of great significance.

7. Summary

In summary, we have surveyed recent progress in statistical mechanics research on low-density parity-check codes. Identifying the similarity between codes defined by a sparse matrix and Ising spin systems of multi-spin interaction makes it possible to analyse and develop a family of high-performance error correcting codes. This relies on employing methods from statistical mechanics in general and the theory of spin glasses in particular. The efficacy of this approach is not limited to basic error correction, similar approaches have also been successfully applied to several other coding schemes such as data compression, multi-terminal data transmission, cryptography, etc.

Research activities in these directions revealed great similarity and some differences, in both the problems studied and methods used, between information sciences and physics, which makes it much easier than ever before to apply methods of one discipline to problems in another. We hope that the current review will contribute to promoting such cross-disciplinary studies.

Acknowledgments

Support from Grant-in-Aid, MEXT, Japan, no 14084206 (YK), The Royal Society (DS) and the European Community's Human Potential Programme under contract HPRN-CT-2002-00319, STIPCO (DS) is acknowledged. Contributions, discussions and comments from Renato Vicente, Jort van Mourik, Tatsuto Murayama, Kazutaka Nakamura, Toshiyuki Tanaka and Tomohiko Uematsu are greatly appreciated.

References

[ADU03] Amraoui A, Dusad S and Urbanke R 2003 Achieving general points in the 2-user Gaussian MAC without time-sharing or rate-splitting by means of iterative coding *Proc. 2002 IEEE Int. Symp. on Information Theory (Lausanne, Switzerland)* p 334

[AJK01] Aji S, Jin H, Khandekar A, MacKay D and McEliece R 2001 BSC thresholds for code ensembles based on 'typical pairs' decoding *Codes, Systems and Graphical Models* ed B Marcus and J Rosenthal (New York: Springer) pp 195–210

[AL95] Amic C D E and Luck J 1995 Zero-temperature error-correcting code for a binary symmetric channel *J. Phys. A: Math. Gen.* **28** 135–47

[Bar03] Barg A 2003 More on the reliability function of the BSC *Proc. 2003 IEEE Int. Symp. on Information Theory (Yokohama, Japan)* p 115

[BGT93] Berrou C, Glavieux A and Thitimajshima P 1993 Near Shannon limit error-correcting coding and decoding: turbo codes *Proc. IEEE Int. Conf. Commun. (ICC) (Geneva, Switzerland)* pp 1064–70

[BL82] Bowman D and Levin K 1982 Spin-glass in the Bethe approximation: insights and problems *Phys. Rev.* B **25** 3438–41

[Chu00] Chung S-Y 2000 On the construction of some capacity-approaching coding schemes *PhD Thesis* Massachusetts Institute of Technology

[CRU01] Chung S, Richardson T and Urbanke R 2001 Analysis of sum–product decoding of low-density parity-check codes using a Gaussian approximation *IEEE Trans. Inform. Theory* **47** 657–70

[CT91] Cover T and Thomas J 1991 *Elements of Information Theory* (New York: Wiley)

[Dav98] Davey M 1998 Record-breaking correction using low-density parity-check codes *Hamilton Prize essay* Gonville and Caius College, Cambridge

[Dav99] Davey M 1999 Error-correction using low-density parity-check codes *PhD Thesis* University of Cambridge

[dBD02] de Baynast A and Declercq D 2002 Random code division multiple access (RCDMA) with Gallager codes *Tech. Report* unpublished

[dBD03] de Baynast A and Declercq D 2003 Gallager codes for multiple user applications *Proc. 2002 IEEE Int. Symp. on Information Theory (Lausanne, Switzerland)* p 335

[Der81] Derrida B 1981 Random energy model: an exactly solvable model of disordered systems *Phys. Rev.* B **24** 238–51

[DH76] Diffie W and Hellman M E 1976 New directions in cryptography *IEEE Trans. Inform. Theory* **22** 644–54

[ElG85] ElGammal T 1985 A public key cryptosystem and a signature scheme based on discrete logarithms *IEEE Trans. Inform. Theory* **31** 469–72

[FLMRT02] Franz S, Leone M, Montanari A and Ricci-Tersenghi F 2002 Dynamic phase transition for decoding algorithms *Phys. Rev.* E **66** 046120

[FM98] Frey B and MacKay D 1998 A revolution: belief propagation in graphs with cycles *Advances in Neural Information Processing Systems* vol 10 ed M Jordan, M Kearns and S Solla (Cambridge, MA: The MIT Press) pp 479–85

[FP95] Franz S and Parisi G 1995 Recipes for metastable states in spin glasses *J. Physique* I **5** 1401

[Fre98] Frey B 1998 *Graphical Models for Machine Learning and Digital Communication* (Cambridge, MA: MIT Press)

[Gal62] Gallager R 1962 Low density parity check codes *IRE Trans. Inform. Theory* **8** 21–8

[Gal63] Gallager R 1963 *Low-Density Parity-Check codes (Research Monograph Series* no 21*)* (Cambridge, MA: MIT Press) p 21

[Gal68] Gallager R 1968 *Information Theory and Reliable Communication* (New York: Wiley)

[Gil52] Gilbert E N 1952 A comparison of signaling alphabets *Bell Syst. Tech. J.* **31** 504–22

[GJ79] Garey M R and Johnson D S 1979 *Computers and Intractability* vol 251 (San Francisco, CA: Freeman)

[Gol91] Goldschmidt Y 1991 Spin glass on the finite-connectivity lattice: the replica solution without replicas *Phys. Rev.* B **43** 8148–52

[Gue03] Guerra F 2003 Replica broken bounds in the mean field spin glass model *Commun. Math. Phys.* **233** 1–12

[Guj95] Gujrati P 1995 Bethe or Bethe-like lattice calculations are more reliable than conventional mean-field calculations *Phys. Rev. Lett.* **74** 809–12

[HKN02] Hosaka T, Kabashima Y and Nishimori H 2002 Statistical mechanics of lossy data compression using a nonmonotonic perceptron *Phys. Rev.* E **66** 066126

[Iba99] Iba Y 1999 The Nishimori line and Bayesian statistics *J. Phys. A: Math. Gen.* **32** 3875–88

[Jel68] Jelinek F 1968 *Probabilistic Information Theory* (New York: McGraw-Hill)

[KF98] Kschischang F and Frey B 1998 Iterative decoding of compound codes by probability propagation in graphical models *IEEE J. Sel. Areas Commun.* **2** 153–9

[KMS00a] Kabashima Y, Murayama T and Saad D 2000 Cryptographical properties of Ising spin systems *Phys. Rev. Lett.* **84** 2030–3

[KMS00b] Kabashima Y, Murayama T and Saad D 2000 Typical performance of Gallager-type error-correcting codes *Phys. Rev. Lett.* **84** 1355–8

[KMSV00] Kabashima Y, Murayama T, Saad D and Vicente R 2000 Regular and irregular Gallager-type error-correcting codes *Advances in Neural Information Processing Systems* vol 12 ed S Solla, T Leen and K Müler (Cambridge, MA: MIT Press) pp 272–8

[KNvM02] Kabashima Y, Nakamura K and van Mourik J 2002 Statistical mechanics of typical set decoding *Phys. Rev.* E **66** 036125

[KS78] Kirkpatrick S and Sherrington D 1978 Infinite-ranged models of spin-glasses *Phys. Rev.* B **17** 4384–403

[KS87] Kanter I and Sompolinsky H 1987 Mean-field theory of spin-glasses with finite coordination number *Phys. Rev. Lett.* **58** 164–7

[KS98] Kabashima Y and Saad D 1998 Belief propagation vs. TAP for decoding corrupted messages *Europhys. Lett.* **44** 668–74

[KS99a] Kabashima Y and Saad D 1999 Statistical physics of error-correcting codes *Europhys. Lett.* **45** 97–103

[KS99b] Kanter I and Saad D 1999 Error-correcting codes that nearly saturate Shannon's bound *Phys. Rev. Lett.* **83** 2660–3

[KS00a] Kanter I and Saad D 2000 Cascading parity-check error-correcting codes *Phys. Rev.* E **61** 2137–40

[KS00b] Kanter I and Saad D 2000 Finite-size effects and error-free communication in Gaussian channels *J. Phys. A: Math. Gen.* **33** 1675–81

[KSNS01] Kabashima Y, Sazuka N, Nakamura K and Saad D 2001 Tighter decoding reliability bound for Gallager's error-correcting codes *Phys. Rev.* E **64** 046113

[LMSS01] Luby M, Mitzenmacher M, Shokrollahi A and Spielman D 2001 Improved low-density parity-check codes using irregular graphs and belief propagation *IEEE Trans. Inform. Theory* **47** 585–98

[Mac99] MacKay D 1999 Good error-correcting codes based on very sparse matrices *IEEE Trans. Inform. Theory* **45** 399–431

[MB01] Miller G and Burshtein D 2001 Bounds on the maximum-likelihood decoding error probability of low-density parity-check codes *IEEE Trans. Inform. Theory* **47** 2696–710

[McE78] McEliece R 1978 A public-key cryptosystem based on algebraic coding theory *Tech. Report DSN Progress Report* 42-44 JPL-Caltech, CA

[McEon] McEliece R 2002 *Theory of Information & Coding* 2nd edn (Cambridge, MA: Cambridge University Press)

[MKSV00] Murayama T, Kabashima Y, Saad D and Vicente R 2000 Statistical physics of regular low-density parity-check error-correcting codes *Phys. Rev.* E **62** 1577–91

[MN95] MacKay D and Neal R 1995 Good codes based on very sparse matrices *Lecture Notes in Computer Science* vol 1025 (Berlin: Springer) pp 100–11

[MN00] MacKay D and Nakamura K 2000 Thresholds of low-density parity check codes (discussion document) available online at http://www. inference. phy. cam. ac. uk/mackay/Discussion. html

[MO03] Murayama T and Okada M 2003 One step RSB scheme for the rate distortion function *J. Phys. A: Math. Gen.* **36** 11123–30

[Mon95] Monasson R 1995 The structural glass transition and the entropy of the metastable states *Phys. Rev. Lett.* **75** 2847–50

[Mon00] Montanari A 2000 Turbo codes: the phase transition *Eur. Phys. J.* B **18** 121–36

[Mon01] Montanari A 2001 The glassy phase of Gallager codes *Eur. Phys. J.* B **23** 121–36

[MPV87] Mézard M, Parisi G and Virasoro M 1987 *Spin Glass Theory and Beyond* (Singapore: World Scientific)

[MPZ02] Mézard M, Parisi G and Zecchina R 2002 Analytic and algorithmic solution of random satisfiability problems *Science* **297** 812–5

[MS77] MacWilliams F and Sloane N 1977 *The Theory of Error-Correcting Codes* (Amsterdam: North-Holland)

[MS00] Montanari A and Sourlas N 2000 The statistical mechanics of turbo codes *Eur. Phys. J.* B **18** 107–19

[Mur02] Murayama T 2002 Statistical mechanics of the data compression theorem *J. Phys. A: Math. Gen.* **35** L95–100

[MY02] Matsunaga Y and Yamamoto Y 2002 A coding theorem for lossy data compression by LDPC codes *Proc. 2002 IEEE Int. Symp. on Information Theory (Lausanne, Switzerland)* p 461

[Nis80] Nishimori H 1980 Exact results and critical properties of the Ising model with competing interactions *J. Phys. C: Solid State Phys.* **13** 4071–6

[Nis93] Nishimori H 1993 Optimal decoding for error-correcting codes *J. Phys. Soc. Japan* **62** 2973–5

[Nis01] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing* (Oxford: Oxford University Press)

[NKMZS03] Nakamura K, Kabashima Y, Morelos-Zaragoza R and Saad D 2003 Statistical mechanics of broadcast channels using low density parity check codes *Phys. Rev.* E **67** 036703

[NKS01] Nakamura K, Kabashima Y and Saad D 2001 Statistical mechanics of low-density parity check error-correcting codes over Galois fields *Eurphys. Lett.* **56** 610–6

[NS01] Nishimori H and Sherrington D 2001 Absence of replica symmetry breaking in a region of the phase diagram of the Ising spin glass *Disordered and Complex Systems* ed P Sollich, A Coolen, L Hughston and R Streater (New York: AIP) p 67–72

[NW99] Nishimori H and Wong M 1999 Statistical mechanics of image restoration and error-correcting codes *Phys. Rev.* E **60** 132–44

[Pea88] Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems* (San Francisco, CA: Morgan Kaufmann)

[RCGV02] Roumy A, Caire G, Guemghar S and Verdú S 2002 Maximizing the spectral efficiency of LDPC-encoded CDMA *Tech. Report* unpublished

[RK92] Rieger H and Kirkpatrick T 1992 Disordered p-spin interaction models on Husimi trees *Phys. Rev.* B **45** 9772–7

[RSA78] Rivest R I, Shamir A and Adleman L 1978 A method for obtaining digital signatures and public key cryptosystems *Commun. ACM* **21** 120–6

[RSU01] Richardson T, Shokrollahi A and Urbanke R 2001 Design of capacity-approaching irregular low-density parity-check codes *IEEE Trans. Inform. Theory* **47** 619–37

[RU01a] Richardson T and Urbanke R 2001 The capacity of low-density parity check codes under message-passing decoding *IEEE Trans. Inform. Theory* **47** 599–618

[RU01b] Richardson T and Urbanke R 2001 Efficient decoding of low-density parity-check codes *IEEE Trans. Inform. Theory* **47** 638–56

[Ruj93] Ruján P 1993 Finite temperature error-correcting codes *Phys. Rev. Lett.* **70** 2968–71

[Saa98] Saakian D 1998 Diluted generalized random energy model *JETP Lett.* **67** 440–4

[Sha48] Shannon C 1948 Mathematical theory of communication: I *Bell. Syst. Tech. J.* **27** 379–423
Shannon C 1948 Mathematical theory of communication: II *Bell. Syst. Tech. J.* **27** 623–56

[Sha49] Shannon C E 1949 *The Mathematical Theory of Information* (Urbana, IL: University of Illinois Press)

[Sha59] Shannon C E 1959 Coding theorems for a discrete source with a fidelity criterion *IRE National Convention Record* vol 4 pp 142–63

[SK75] Sherrington D and Kirkpatrick S 1975 Solvable model of a spin-glass *Phys. Rev. Lett.* **35** 1792–6

[SKM01] Saad D, Kabashima Y and Murayama T 2001 Public key cryptography and error correcting codes as Ising models *Disordered and Complex Systems* ed P Sollich, A Coolen, L Hughston and R Streater (New York: AIP) pp 89–94

[Sou89] Sourlas N 1989 Spin-glass models as error-correcting codes *Nature* **339** 693–5

[Sou94] Sourlas N 1994 Spin-glasses, error-correcting codes and finite-temperature decoding *Europhys. Lett.* **25** 159–64

[SSK03] Skantzos N, Saad D and Kabashima Y 2003 Analysis of common attacks in public-key cryptosystems based on low-density parity-check codes *Phys. Rev.* E **68** at press

[SST92] Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev.* A **45** 6056–91

[SU03] Sason I and Urbanke R 2003 Parity-check density versus performance of binary linear block codes over memoryless symmertic channels *IEEE Trans. Inform. Theory* **49** 1611–35

[SvMS03] Skantzos N, van Mourik J and Saad D 2003 Magnetization enumerator of real-valued symmetric channels in gallager error-correcting codes *Phys. Rev.* E **67** 037101

[SvMSK03] Skantzos N, van Mourik J, Saad D and Kabashima Y 2003 Average and reliability error exponents in low-density parity-check codes *J. Phys. A: Math. Gen.* **36** 11131–41

[SW73] Slepian D and Wolf J K 1973 Noiseless coding of correlated information sources *IEEE Trans. Inform. Theory* **19** 471–80

[Tal03] Talagrand M 2003 The generalised Parisi formula *Tech. Report* unpublished

[Tan02] Tanaka T 2002 A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors *IEEE Trans. Inform. Theory* **11** 2888–910

[TS03a] Tanaka T and Saad D 2003 A statistical-mechanical analysis of coded CDMA with regular LDPC codes *Proc. 2003 IEEE Int. Symp. on Information Theory (Yokohama, Japan)* p 444

[TS03b] Tanaka T and Saad D 2003 Statistical-mechanical analysis of LDPC-coded CDMA *Tech. Report* unpublished

[TS03c] Tanaka T and Saad D 2003 Typical performance of low-density parity-check codes over general symmetric channels *J. Phys. A: Math. Gen.* **36** 11143–57

[Var57] Varshamov R R 1957 Estimate of the number of signals in error-correcting codes *Dokl. Akad. Nauk. SSSR 117* **5** 739–41

[Ver98] Verdú S 1998 *Multiuser Detection* (Cambridge: Cambridge University Press)

[vG83] van Gils W 1983 Two topics on linear unequal error protection codes: bounds on their length and cyclic code classes *IEEE Trans. Inform. Theory* **29** 866–76

[vG84] van Gils W 1984 Linear unequal error protection codes from shorter codes (corresp.) *IEEE Trans. Inform. Theory* **30** 544–46

[vMK03] van Mourik J and Kabashima Y 2003 The polynomial error probability for LDPC codes *Preprint* cond-mat/0310177

[vMSK01] van Mourik J, Saad D and Kabashima Y 2001 Weight vs. magnetization enumerator for Gallager codes *Cryptography and Coding: 8th IMA Int. Conf. (Berlin: Germany)* ed B Honary (Berlin: Springer) pp 148–57

[vMSK02] van Mourik J, Saad D and Kabashima Y 2002 Critical noise levels for LDPC decoding *Phys. Rev.* E **66** 026705

[vMWB00] van Mourik J, Wong K Y M and Bollé D 2000 From shrinking to percolation in an optimization model *J. Phys. A: Math. Gen.* **33** L53

[VO79] Viterbi A and Omura J 1979 *Principles of Digital Communication and Coding* (Singapore: McGraw-Hill)

[VSK99] Vicente R, Saad D and Kabashima Y 1999 Finite-connectivity systems as error-correcting codes *Phys. Rev.* E **60** 5352–66

[VSK00a] Vicente R, Saad D and Kabashima Y 2000 Error-correcting code on a cactus: a solvable model *Europhys. Lett.* **51** 698–704

[VSK00b] Vicente R, Saad D and Kabashima Y 2000 Statistical mechanics of irregular low-density parity-check codes *J. Phys. A: Math. Gen.* **33** 6527–42

[VSK01] Vicente R, Saad D and Kabashima Y 2001 Error-correcting codes on a Bethe-like lattice *Advances in Neural Information Processing Systems* vol 13 ed T Leen, T Dietterich and V Tresp (Cambridge, MA: The MIT Press) pp 322–8

[VSK02] Vicente R, Saad D and Kabashima Y 2002 Low density parity check codes: a statistical physics perspective *Advances in Imaging and Electron Physics* vol 125 ed P Hawkes (New York: Academic) p 232–353

[Wib96] Wiberg N 1996 Codes and decoding on general graphs *PhD Thesis* Department of Electrical Engineering, Linköping University

[WS87a] Wong K and Sherrington D 1987 Graph bipartitioning and spin glasses on a random network of fixed finite valence *J. Phys. A: Math. Gen.* **20** L793–9

[WS87b] Wong K and Sherrington D 1987 Graph bipartitioning and the Bethe spin-glass *J. Phys. A: Math. Gen.* **20** L785–91

[YFW02] Yedidia J S, Freeman W T and Weiss Y 2002 Constructing free energy approximations and generalised belief propagation algorithms *Tech. Report* TR2002-35 Mitsubishi Electric Research Laboratories

[YZB97] Yang E, Zhang Z and Berger T 1997 Fixed-slope universal lossy data compression *IEEE Trans. Inform. Theory* **43** 1465–76

[ZL77] Ziv J and Lempel A 1977 A universal algorithm for sequential data compression *IEEE Trans. Inform. Theory* **23** 337–43