

Typical Performance of Gallager-Type Error-Correcting Codes

Yoshiyuki Kabashima,¹ Tatsuto Murayama,¹ and David Saad²

¹*Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, Yokohama 2268502, Japan*

²*The Neural Computing Research Group, Aston University, Birmingham B4 7ET, United Kingdom*

(Received 26 November 1998)

The performance of Gallager's error-correcting code is investigated via methods of statistical physics. In this approach, the transmitted codeword comprises products of the original message bits selected by two randomly constructed sparse matrices; the number of nonzero row/column elements in these matrices constitutes a family of codes. We show that Shannon's channel capacity is saturated for many of the codes while slightly lower performance is obtained for others which may be of higher practical relevance. Decoding aspects are considered by employing the Thouless-Anderson-Palmer approach which is identical to the commonly used belief-propagation-based decoding.

PACS numbers: 89.90.+n, 02.50.-r, 05.50.+q, 75.10.Hk

The ever increasing information transmission in the modern world is based on communicating messages reliably through noisy transmission channels; these can be telephone lines, magnetic storing media, etc. Error-correcting codes play an important role in correcting errors incurred during transmission; this is carried out by encoding the message prior to transmission and decoding the corrupted received codeword for retrieving the original message. In his groundbreaking papers, Shannon [1] analyzed the capacity of communication channels, setting an upper bound to the achievable noise-correction capability of codes, given their code (or symbol) rate. The latter represents the ratio between the number of bits in the original message and in the transmitted codeword.

Shannon's bound is nonconstructive and does not provide explicit rules for devising practical optimal codes. The quest for more efficient codes, in the hope of saturating the bound set by Shannon, has been going on ever since, providing many useful but suboptimal codes.

One family of codes, presented originally by Gallager [2], attracted significant interest recently as it has been shown to outperform most currently used techniques [3]. In fact, irregular versions of Gallager-type codes have recently been shown to get very close to saturating Shannon's bound in the case of infinitely long messages [4]. Most studies of Gallager-type codes conducted so far have been carried out via numerical simulations. Some analytical results have been obtained via methods of information theory [3], setting bounds on the performance of certain code types, and by combinatorial/statistical methods [4]; no quantitative results have been obtained for their *typical* performance. In this Letter we analyze the typical performance of Gallager-type codes for several parameter choices via methods of statistical mechanics. We then validate the analytical solution by comparing the results to those obtained by the Thouless-Anderson-Palmer (TAP) approach to diluted systems and via numerical methods.

In a general scenario, the N -dimensional Boolean message ξ is encoded to the M -dimensional vector \mathbf{J}^0 which is then transmitted through a noisy channel with flipping

probability p per bit (other noise types may also be considered but will not be examined here). The received message \mathbf{J} is decoded to retrieve the original message.

One can identify several slightly different versions of Gallager-type codes. The one used in this Letter, termed the MN code [3], is based on choosing two randomly selected sparse matrices A and B of dimensionality $M \times N$ and $M \times M$, respectively; these are characterized by K and L nonzero unit elements per row and C and L per column, respectively. The finite, usually small, numbers K , C , and L define a particular code; both matrices are known to both sender and receiver. Encoding is carried out by constructing the modulo 2 inverse of B and the matrix $B^{-1}A$ (modulo 2); the vector $\mathbf{J}^0 = B^{-1}A\xi$ (modulo 2, ξ in a Boolean representation) constitutes the codeword. Decoding is carried out by taking the product of the matrix B and the received message $\mathbf{J} = \mathbf{J}^0 + \zeta$ (modulo 2), corrupted by the Boolean noise vector ζ , which results in $A\xi + B\zeta$. The equation

$$A\xi + B\zeta = AS + B\tau \pmod{2} \quad (1)$$

is solved via the iterative methods of belief propagation (BP) [3] to obtain the most probable Boolean vectors \mathbf{S} and $\boldsymbol{\tau}$; BP methods in the context of error-correcting codes have recently been shown to be identical to a TAP based solution of a similar physical system [5].

The similarity between error-correcting codes of this type and Ising spin systems was first pointed out by Sourlas [6], who formulated the mapping of a simpler code, somewhat similar to the one presented here, onto an Ising spin system Hamiltonian. We recently extended the work of Sourlas, which focused on extensively connected systems, to the finite connectivity case [5].

To facilitate the current investigation we first map the problem to that of an Ising model with finite connectivity. We employ the binary representation (± 1) of the dynamical variables \mathbf{S} and $\boldsymbol{\tau}$ and of the vectors \mathbf{J} and \mathbf{J}^0 rather than the Boolean (0, 1) one; the vector \mathbf{J}^0 is generated by taking products of the relevant binary message bits $J_{\langle i_1, i_2, \dots \rangle}^0 = \xi_{i_1} \xi_{i_2} \dots$, where the indices i_1, i_2, \dots correspond to the nonzero elements of $B^{-1}A$, producing a

binary version of \mathbf{J}^0 . As we use statistical mechanics techniques, we consider the message and codeword dimensionality (N and M , respectively) to be infinite, keeping the ratio between them $R = N/M$, which constitutes the code rate, finite. Using the thermodynamic limit is quite natural as Gallager-type codes are usually used for transmitting long (10^4 – 10^5) messages, where finite size corrections are likely to be negligible. To explore the system's capabilities we examine the Hamiltonian

$$\begin{aligned} \mathcal{H} = & \sum_{\langle i_1, \dots, i_K; j_1, \dots, j_L \rangle} \mathcal{D}_{\langle i_1, \dots, i_K; j_1, \dots, j_L \rangle} \\ & \times \delta[-1; \mathbf{J}_{\langle i_1, \dots, i_K; j_1, \dots, j_L \rangle} \cdot S_{i_1} \cdots S_{i_K} \tau_{j_1} \cdots \tau_{j_L}] \\ & - \frac{F_s}{\beta} \sum_{i=1}^N S_i - \frac{F_\tau}{\beta} \sum_{j=1}^M \tau_j. \end{aligned} \quad (2)$$

The tensor product $\mathcal{D}_{\langle i_1, \dots, i_K; j_1, \dots, j_L \rangle} \mathbf{J}_{\langle i_1, \dots, i_K; j_1, \dots, j_L \rangle}$, where $\mathbf{J}_{\langle i_1, \dots, j_L \rangle} = \xi_{i_1} \xi_{i_2} \cdots \xi_{i_K} \zeta_{j_1} \zeta_{j_2} \cdots \zeta_{j_L}$ is the binary equivalent of $A\xi + B\zeta$, treating both signal (S and index i) and noise (τ and index j) simultaneously. Elements of the sparse connectivity tensor $\mathcal{D}_{\langle i_1, \dots, j_L \rangle}$ take the value 1 if the corresponding indices of both signal and noise are chosen (i.e., if all corresponding indices of the matrices A and B are 1) and 0 otherwise; it has C unit elements per i index and L per j index representing the system's degree of connectivity. The δ function provides 1 if the selected sites' product $S_{i_1} \cdots S_{i_K} \tau_{j_1} \cdots \tau_{j_L}$ is in disagreement with the corresponding element $\mathbf{J}_{\langle i_1, \dots, j_L \rangle}$, recording an error, and 0 otherwise. Notice that this term is not frustrated, as there are $M + N$ degrees of freedom and only M constraints from Eq. (1), and can therefore vanish at sufficiently low temperatures. The last two terms on the right represent our prior knowledge in the case of sparse or biased messages F_s and of the noise level F_τ and require assigning certain values to these additive fields. The choice of $\beta \rightarrow \infty$ imposes the restriction of Eq. (1), limiting the solutions to those for which the first term of Eq. (2) vanishes, while the last two terms, scaled with β , survive. Note that the noise dynamical variables τ are irrelevant to measuring the retrieval success $m = \frac{1}{N} \langle \sum_{i=1}^N \xi_i \text{sgn} \langle S_i \rangle_\beta \rangle_\xi$. The latter monitors the normalized mean overlap between the Bayes-optimal retrieved message, shown to correspond to the alignment of $\langle S_i \rangle_\beta$ to the nearest binary value [6], and the original message; the subscript β denotes thermal averaging.

Since the first part of Eq. (2) is invariant under the transformations $S_i \rightarrow S_i \xi_i$, $\tau_j \rightarrow \tau_j \zeta_j$, and $\mathbf{J}_{\langle i_1, \dots, j_L \rangle} \rightarrow \mathbf{J}_{\langle i_1, \dots, j_L \rangle} \xi_{i_1} \cdots \xi_{i_K} \zeta_{j_1} \zeta_{j_2} \cdots \zeta_{j_L} = 1$, it would be useful to decouple the correlation between the vectors \mathbf{S} , $\boldsymbol{\tau}$ and $\boldsymbol{\xi}$, $\boldsymbol{\zeta}$. Rewriting Eq. (2) using this gauge, one obtains a similar expression apart from the last terms on the right which become $F_s/\beta \sum_k S_k \xi_k$ and $F_\tau/\beta \sum_k \tau_k \zeta_k$.

The random selection of elements in \mathcal{D} introduces disorder to the system; we calculate the partition function $Z(\mathcal{D}, \mathbf{J}) = \text{Tr}_{\{S, \tau\}} \exp[-\beta \mathcal{H}]$ averaged over the disorder and the statistical properties of the message and noise, using the replica method [5,7]. Taking $\beta \rightarrow \infty$ gives rise to a set of order parameters

$$\begin{aligned} q_{\alpha, \beta, \dots, \gamma} &= \left\langle \frac{1}{N} \sum_{i=1}^N Z_i S_i^\alpha S_i^\beta, \dots, S_i^\gamma \right\rangle_{\beta \rightarrow \infty}, \\ r_{\alpha, \beta, \dots, \gamma} &= \left\langle \frac{1}{M} \sum_{j=1}^M Y_j \tau_j^\alpha \tau_j^\beta, \dots, \tau_j^\gamma \right\rangle_{\beta \rightarrow \infty}, \end{aligned} \quad (3)$$

where α, β, \dots represent replica indices, and the variables Z_i and Y_j come from enforcing the restriction of C and L connections per index, respectively [5]:

$$\begin{aligned} \delta \left(\sum_{\langle i_2, \dots, i_K \rangle} \mathcal{D}_{\langle i_1, i_2, \dots, j_L \rangle} - C \right) &= \oint_0^{2\pi} \frac{dZ}{2\pi} Z \sum_{\langle i_2, \dots, i_K \rangle} \mathcal{D}_{\langle i_1, i_2, \dots, j_L \rangle}^{-(C+1)}, \end{aligned} \quad (4)$$

and similarly for the restriction on the j indices.

To proceed with the calculation one has to make an assumption about the order parameters' symmetry. The assumption made here, and validated later on, is that of replica symmetry in the following representation of the order parameters and the related conjugate variables:

$$\begin{aligned} q_{\alpha, \beta, \dots, \gamma} &= a_q \int dx \pi(x) x^l, & \hat{q}_{\alpha, \beta, \dots, \gamma} &= a_{\hat{q}} \int d\hat{x} \hat{\pi}(\hat{x}) \hat{x}^l, \\ r_{\alpha, \beta, \dots, \gamma} &= a_r \int dy \rho(y) y^l, & \hat{r}_{\alpha, \beta, \dots, \gamma} &= a_{\hat{r}} \int d\hat{y} \hat{\rho}(\hat{y}) \hat{y}^l, \end{aligned} \quad (5)$$

where l is the number of replica indices, a_* are normalization coefficients, and $\pi(x)$, $\hat{\pi}(\hat{x})$, $\rho(y)$ and $\hat{\rho}(\hat{y})$ represent probability distributions. Unspecified integrals are over the range $[-1, +1]$. Extremizing the averaged expression with respect to the probability distributions one obtains the following free energy per spin:

$$\begin{aligned} \frac{1}{N} \langle \ln Z \rangle_{\xi, \zeta, \mathcal{D}} = & \text{Extr}_{\{\pi, \hat{\pi}, \rho, \hat{\rho}\}} \left\{ \frac{C}{K} \int \left[\prod_{k=1}^K dx_k \pi(x_k) \right] \left[\prod_{l=1}^L dy_l \rho(y_l) \right] \ln \left[1 + \prod_{k=1}^K x_k \prod_{l=1}^L y_l \right] \right. \\ & - C \int dx d\hat{x} \pi(x) \hat{\pi}(\hat{x}) \ln(1 + x\hat{x}) - \frac{CL}{K} \int dy d\hat{y} \rho(y) \hat{\rho}(\hat{y}) \ln[1 + y\hat{y}] \\ & + \int \left[\prod_{k=1}^C dx_k \hat{\pi}(\hat{x}_k) \right] \left\langle \ln \left[\prod_{k=1}^C (1 + \hat{x}_k) e^{F_s \xi} + \prod_{k=1}^C (1 - \hat{x}_k) e^{-F_s \xi} \right] \right\rangle_\xi \\ & + \frac{C}{K} \int \left[\prod_{l=1}^L d\hat{y}_l \hat{\rho}(\hat{y}_l) \right] \left\langle \ln \left[\prod_{l=1}^L (1 + \hat{y}_l) e^{F_\tau \zeta} + \prod_{l=1}^L (1 - \hat{y}_l) e^{-F_\tau \zeta} \right] \right\rangle_\zeta - \frac{C}{K} \ln 2 \left. \right\}, \end{aligned} \quad (6)$$

where $\langle \cdot \rangle_\xi$ and $\langle \cdot \rangle_\zeta$ denote averages over the input and noise distributions of the form

$$\langle \cdot \rangle_\xi = \sum_{\xi=\pm 1} \left\{ \frac{1 + \tanh F_s}{2} \delta_{\xi,-1} + \frac{1 - \tanh F_s}{2} \delta_{\xi,1} \right\} (\cdot) \quad (7)$$

and similarly for $\langle \cdot \rangle_\zeta$, where F_s is replaced by F_τ .

The free energy can then be calculated via the saddle point method. Solving the equations obtained by varying Eq. (6) with respect to the probability distributions $\pi(x)$, $\hat{\pi}(\hat{x})$, $\rho(y)$, and $\hat{\rho}(\hat{y})$ is generally difficult. The solutions obtained in the case of unbiased messages (the most interesting case as most messages are compressed prior to transmission) are for the ferromagnetic phase:

$$\begin{aligned} \pi(x) &= \delta(x - 1), & \hat{\pi}(\hat{x}) &= \delta(\hat{x} - 1), \\ \rho(y) &= \delta(y - 1), & \hat{\rho}(\hat{y}) &= \delta(\hat{y} - 1), \end{aligned} \quad (8)$$

and for the paramagnetic phase (there is no spin-glass solution due to lack of frustration):

$$\begin{aligned} \pi(x) &= \delta(x), & \hat{\pi}(\hat{x}) &= \delta(\hat{x}), & \hat{\rho}(\hat{y}) &= \delta(\hat{y}), \\ \rho(y) &= \frac{1 + \tanh F_\tau}{2} \delta(y - \tanh F_\tau) \\ &+ \frac{1 - \tanh F_\tau}{2} \delta(y + \tanh F_\tau). \end{aligned} \quad (9)$$

These solutions clearly obey the saddle point equations. However, it is necessary to validate the stability of the solutions and the replica symmetric ansatz itself. To address these questions we obtained solutions to the system described by the Hamiltonian (2) via the TAP method of finitely connected systems [5]; we solved the saddle point equations derived from Eq. (6) numerically, representing all probability distributions by up to 10^4 bin models and carrying out the integrations via Monte Carlo methods; finally, to show the consistency between theory and practice we carried out large scale simulations which will be presented elsewhere. The results obtained by the various methods are in complete agreement.

The solutions obtained for the various cases fall into two different categories: that of $K = L = 2$ and that of $K \geq 3$ or $L \geq 3$, which we therefore treat separately.

For unbiased messages and either $K \geq 3$ or $L \geq 3$ we obtain the solutions (8) and (9) both by applying the TAP approach and by solving the saddle point equations numerically. The former was carried out at the value of F_τ which corresponds to the true noise and input bias levels (for unbiased messages $F_s = 0$) and thus to Nishimori's condition [8], where no replica symmetry breaking effect is expected. This is equivalent to having the correct prior within the Bayesian framework [9] and enables one to obtain analytic expressions for some observables as long as some gauge requirements are obeyed [8]. Numerical solutions show the emergence of stable dominant delta peaks,

consistent with those of (8) and (9). The question of longitudinal mode stability (corresponding to the replica symmetric solution) was addressed by setting initial conditions for the numerical solutions close to the solutions (8) and (9), showing that they converge back to these solutions which are therefore stable.

The most interesting quantity to examine is the maximal code rate, for a given corruption process, for which messages can be perfectly retrieved. This is defined in the case of $K, L \geq 3$ by the value of $R = K/C = N/M$ for which the free energy of the ferromagnetic solution becomes smaller than that of the paramagnetic solution, constituting a first order phase transition. A schematic description of the solutions obtained is shown in the inset in Fig. 1a. The paramagnetic solution ($m = 0$) has a lower free energy than the ferromagnetic one (low/high free energies are denoted by thick and thin lines, respectively; there are no axis lines at $m = 0, 1$) for noise levels $p > p_c$ and vice versa for $p \leq p_c$; both solutions are stable. The critical code rate is derived by equating the ferromagnetic and paramagnetic free energies to obtain

$$R_c = 1 - H_2(p) = 1 + [p \log_2 p + (1 - p) \log_2 (1 - p)]. \quad (10)$$

This coincides with *Shannon's capacity*. To validate these results we obtained TAP solutions for the unbiased message case ($K = L = 3, C = 6$), presented in Fig. 1a (plusses), in comparison to Shannon's capacity (solid line).

Analytical solutions for the saddle point equations cannot be obtained for the case of biased patterns and we therefore resort to numerical methods and the TAP approach.

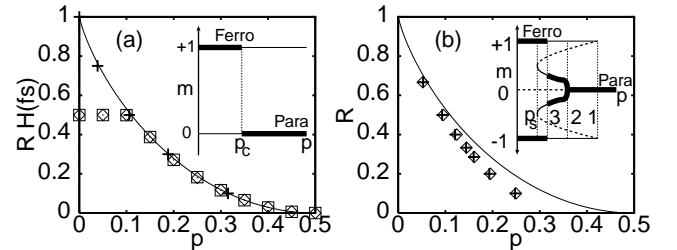


FIG. 1. Critical code rate as a function of the flip rate p , obtained numerically and via the TAP approach ($N = 10^4$), and averaged over ten different initial conditions with error bars much smaller than the symbol size. (a) Numerical solutions for $K = L = 3, C = 6$ and varying input bias f_s (squares) and TAP solutions for both unbiased (plusses) and biased (diamonds) messages; initial conditions were chosen close to the analytical ones. The critical rate is multiplied by the source information content to obtain the maximal information transmission rate. Inset: The ferromagnetic and paramagnetic solutions as functions of p ; thick and thin lines denote stable solutions of lower and higher free energies, respectively. (b) For the unbiased case of $K = L = 2$; initial conditions for the TAP (plusses) and the numerical solutions (diamonds) are of almost zero magnetization. Inset: The ferromagnetic (optimal/suboptimal) and paramagnetic solutions as functions of p ; thick and thin lines are as in (a); dashed lines correspond to unstable solutions.

The maximal information rate (i.e., code rate $\times H_2[f_s = (1 + \tanh F_s)/2]$, the source redundancy) obtained by the TAP method (diamonds) and numerical solutions of the saddle point equations (squares) for each noise level are shown in Fig. 1a. Numerical results have been obtained using 10^3 – 10^4 bin models for each probability distribution and had been run for 10^5 steps per noise level point. The various results are highly consistent and practically saturate Shannon's bound for the same noise level.

The MN code for $K, L \geq 3$ seems to offer optimal performance. However, the main drawback is rooted in the coexistence of the stable $m = 1, 0$ solutions, shown in Fig. 1a (inset), which implies that from some initial conditions the system will converge to the undesired paramagnetic solution. Studying the ferromagnetic solution numerically shows a highly limited basin of attraction, which becomes smaller as K and L increase, while the paramagnetic solution at $m = 0$ always enjoys a wide basin of attraction. As initial conditions for the decoding process are typically of close-to-zero magnetization (almost no prior information about the original message is assumed) it is likely that the decoding process will converge to the paramagnetic solution (as has been observed via computer simulations by us and by others [3]).

While all codes with $K, L \geq 3$ saturate Shannon's bound and are characterized by a first order, paramagnetic to ferromagnetic, phase transition, codes with $K = L = 2$ show lower performance and different physical characteristics. The analytical solutions (8) and (9) are unstable at some flip rate levels and one resorts to solving the saddle point equations numerically and to TAP based solutions. The picture that emerges is sketched in the inset in Fig. 1b: The paramagnetic solution dominates the high flip rate regime (appearing as a dominant delta peak in the numerical solutions) up to the point p_1 (denoted as 1 in the inset) in which a stable ferromagnetic solution of higher free energy appears (thin lines at $m = \pm 1$). At a lower flip rate value p_2 the paramagnetic solution becomes unstable (dashed line) and is replaced by two stable suboptimal ferromagnetic (broken symmetry) solutions which appear as a couple of peaks in the various probability distributions; typically, these have a lower free energy than the ferromagnetic solution until p_3 , after which the ferromagnetic solution becomes dominant (at some code rate values it is dominant directly following the disappearance of the paramagnetic solution). Still, only once the suboptimal ferromagnetic solutions disappear, at the spinodal point p_s , a unique ferromagnetic solution emerges as a single delta peak in the numerical results (plus a mirror solution). The point in which the suboptimal ferromagnetic solutions disappear constitutes the maximal practical flip rate for the current code rate

and was defined numerically (diamonds) and via TAP solutions (plusses) as shown in Fig. 1b.

Notice that initial conditions for both TAP and the numerical solutions were chosen almost randomly, with a very slight bias of $\mathcal{O}(10^{-12})$, in the initial magnetization. The TAP dynamical equations are identical to those used for practical BP decoding [5] and therefore provide equivalent results to computer simulations with the same parametrization, supporting the analytical results. The excellent convergence results obtained point out the existence of a unique pair of global solutions to which the system converges (below p_s) from practically all initial conditions. This observation and the practical implications of using the $K = L = 2$ code have not been obtained by information theory methods (e.g., [3]); these prove the existence of very good codes for $C, L \geq 3$ and examine decoding properties only via numerical simulations.

In summary, we discovered that for certain Gallager-type codes with $K \geq 3$ or $L \geq 3$, one potentially obtains optimal performance, saturating Shannon's bound. This comes at the expense of a decreasing basin of attraction making the decoding process increasingly impractical. A more practical code, $K = L = 2$, shows close to optimal performance with a very large basin of attraction. Studying the typical performance of Gallager-type codes, which complements the methods used in the information theory literature, is the first step towards understanding their exceptional performance and in the search for a principled method for designing optimal Gallager-type codes. Important aspects that are yet to be investigated include other noise types, irregular constructions, and the significance of finite size effects.

Support by the JSPS (RFTF), The Royal Society, and EPSRC (GR/L52093) is acknowledged.

-
- [1] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).
 - [2] R. G. Gallager, IRE Trans. Inf. Theory **IT-8**, 21 (1962).
 - [3] D. J. C. MacKay, IEEE Trans. Inf. Theory **45**, 399 (1999).
 - [4] T. Richardson, A. Shokrollahi, and R. Urbanke, <http://cm.bell-labs.com/cm/ms/who/tjr/pub.html> (1999); I. Kanter and D. Saad, Phys. Rev. Lett. **83**, 2660 (1999).
 - [5] Y. Kabashima and D. Saad, Europhys. Lett. **44**, 668 (1998); **45**, 97 (1999); R. Vicente, D. Saad, and Y. Kabashima, Phys. Rev. E **60**, 5352 (1999).
 - [6] N. Sourlas, Nature (London) **339**, 693 (1989).
 - [7] K. Y. M. Wong and D. Sherrington, J. Phys. A **20**, L793 (1987); C. De Dominicis and P. Mottishaw, J. Phys. A **20**, L1267 (1987).
 - [8] H. Nishimori, Prog. Theor. Phys. **66**, 1169 (1981).
 - [9] N. Sourlas, Europhys. Lett. **25**, 159 (1994).