Error Correction on a Tree: An Instanton Approach

V. Chernyak,[1] M. Chertkov,[2] M. G. Stepanov,[2,3,4] and B. Vasic[5]
[1]*Department of Chemistry, Wayne State University, 5101 Cass Avenue, Detroit, Michigan 48202, USA*
[2]*Theoretical Division, LANL, Los Alamos, New Mexico 87545, USA*
[3]*Department of Mathematics, University of Arizona, Tucson, Arizona 85721, USA*
[4]*Institute of Automation and Electrometry, Novosibirsk 630090, Russia*
[5]*Department of Electrical Engineering, University of Arizona, Tucson, Arizona 85721, USA*
(Received 25 March 2004; published 5 November 2004)

We introduce a method that allows analytical or semianalytical estimating of the post-error correction bit error rate (BER) when a forward-error correction is utilized for transmitting information through a noisy channel. The generic method that applies to a variety of error-correction schemes in the regimes where the BER is low is illustrated using the example of a finite-size code approximated by a treelike structure. Exploring the statistical physics formulation of the problem we find that the BER decreases with the signal-to-noise ratio nonuniformly, i.e., crossing over through a sequence of phases. The higher the signal-to-noise ratio the lower the symmetry of the phase dominating BER.

DOI: 10.1103/PhysRevLett.93.198702 PACS numbers: 89.70.+c, 05.20.–y, 89.20.Ff

The last 50 years have witnessed a tremendous increase in the amount of data transmitted through various communication systems. This also leads to tightening of the conditions for error-free data transfer in the presence of noise and other transmission-related impairments. The problem of dealing with errors in information flow has a fundamental importance and has been studied extensively in information and coding theory. In 1948 Shannon [1] proved that applying an error-corrected code can result in an error-free communication in the thermodynamic limit of an infinitely long message, as long as the rate of transmitted information is kept below a certain value known as the channel capacity. Constructing well-performing, capacity-approaching yet practical codes has been a challenge until the discovery that some classes of codes based on random construction [2] can achieve near-optimum performance when used for transmission over white additive Gaussian noise channels [3]. In the past few years several codes have been designed with performances very close to this limit [4]. Generally, these codes are referred to as codes on graphs, and their prime examples are low-density parity-check (LDPC) codes and turbo codes. A linear block code (for which LDPC codes are an example) can be represented as the set of solutions σ to a system of linear equations $\hat{H}\sigma = 0$, where each row of the parity-check matrix $\boldsymbol{H}$ is called a parity-check equation. The codes under consideration are called binary since all the coding operations are in a binary field. The code can be represented by a bipartite (Tanner) graph which consists of variable nodes that represent the code word bits and check nodes that are representatives of its parity-check equations. The number of edges that originate from a node are referred to as its degree. In this Letter we discuss primarily codes with a uniform variable and/or check node degree distribution. Note that relations between the variable and checking nodes on the graph may still be random. The uniform degree codes based on random construction are called regular Gallager codes [2]. These codes show an extraordinarily good performance, however, it has been also shown recently [5] that regularly structured LDPC codes (that have a natural advantage of being memory effective and simpler to build) can be performing comparably well. Another major recent development is associated with reformulating the error-correction problem in terms of statistical mechanics [6], which stimulated a fresh flow of new exciting ideas and analogies. (See, e.g., [7–10].) However, the new approach has mainly focused on comprehensive analysis of the thermodynamic (infinite code length) limit, whereas describing the phenomena related to realistic finite-size codes has attracted much less attention.

Performance of any finite-size error-correcting code is measured in terms of the dependence of the post-error correction bit error rate (BER) on the signal-to-noise ratio (SNR). Error correction aims to decrease BER by adding redundant information (overhead) to the information message. The smaller the post-error correction BER is (for fixed overhead) the better. Any new generation of communication devices creates a new challenge for the error-correction technology as it sets higher standards for the channel capacity, thus lowering the level of BER which can still be tolerated. Straightforward Monte Carlo numerical simulations constitute an efficient method only for the values of a BER $\sim 10^{-7}$ or higher, and it falls short in accessing lower values of BER. Experimental tests are extremely expensive, thus frequently impractical, since they require building a special device prototype for any new suggested coding or decoding strategy. This implies that finding efficient practical ways of extremely low-BER evaluation is under universal demand. Our main objective is constructing a theoretical tool capable of delivering

quantitative estimates for these low probability events analytically. The approach we propose to adopt and develop for achieving this goal is known under the names of saddle point, optimal fluctuation, or instanton calculus. (Instanton calculus, introduced initially in the context of disordered systems [11] and aimed at estimating a low probability event, is common in modern theoretical physics.) Therefore, we start with a general but brief introduction to the subject. We describe the basic principles of coding for an LDPC code, introduce the optimal maximal *a posteriori* (MAP) decoding strategy along with generally suboptimal yet very efficient belief propagation (BP) decoding, and finally define the post-error correction BER that characterizes the code performance. Next we argue, following [3,4,9,10,12], that a finite-size treelike structure offers a good approximation for a uniform degree LDPC code if the length of the shortest loop on the corresponding Tanner graph is long enough. We further focus on the BER computation for the central site on the tree, presenting it as an integral over noise configuration (fields) on the tree. Instantons—special configurations of the field giving the major contribution into the integral/BER—are first found numerically through complete variational procedure. We show that all the relevant instantons correspond to different symmetries visualized in terms of the partially colored Tanner graph. Finally, we describe a sequence of phase transitions, between phases and/or instantons of different symmetries, thus fulfilling the task of describing BER dependence on SNR in the low-BER domain.

Error correction consists of: (i) coding the original message (word) represented as a set of L binary ± 1 symbols into a longer word consisting of N binary signals; (ii) transmitting the N-bit long code word through a noisy channel; (iii) decoding the corrupted message detected at the output. The Tanner graph consists of N variable nodes (marked by Latin indices) that correspond to the bits of the transmitted message and $M = N - L > 0$ checking nodes (marked by Greek indices) that represent the parity checks; and the connections occur between those bits j and parity checks α so that the bit j participates in the parity-check α, i.e., $j \in \alpha$. (In this representation all the parity checks should be linearly independent.) More formally, $\sigma = (\sigma_1, \cdots, \sigma_N)$ with $\sigma_i = \pm 1$ represents one of 2^L code words if and only if $\prod_{j\in\alpha}\sigma_j = 1$ for all the checking nodes, $\alpha = 1, \cdots, M$. The code redundancy is described by the overhead $M/L = R^{-1} - 1$, with $R = L/N < 1$ being the code rate. Transmitted through a noisy channel a code word gets corrupted due to the channel noise, so that at the channel output one detects, $\mathbf{x} \neq \sigma$, where in the simplest model case of the additive white Gaussian channel considered here $\mathbf{x} = \sigma + \varphi$, $\langle\varphi\rangle = 0$, and $\langle\varphi_i\varphi_j\rangle = \delta_{ij}/s^2$, where s measures the SNR.

The goal of decoding is inferring the best approximation for the original message from a corrupted word. Optimal decoding, also known under the name of MAP symbol decoding, can be represented in terms of the generating function of an effective "spin" model [6,7] $\exp[-F(h)] = \sum_\sigma \prod_{\alpha=1}^{M} \delta(\prod_{j\in\alpha}\sigma_j, 1) \exp(\sum_{k=1}^{N} h_k\sigma_k)$, where the "external magnetic field" $\boldsymbol{h}$ is related to the channel noise φ, $\boldsymbol{h} = s^2(1 + \varphi)$, $\delta(x, 1)$ is the Kronecker δ symbol, and the "magnetization", defined as $\psi_j(\boldsymbol{h}) \equiv \langle\sigma_j\rangle = -\partial F(h)/\partial h_j$, is interpreted as the result of decoding, or more accurately $\mathrm{sgn}[\psi_j]$ gives the decoded value for the bit j. The code performance can be characterized via the density of errors at the given site j known as the post-error correction BER that can be also described as the probability of a spin flip

$$B_j = \int_{-1}^{0} d\xi \int d\boldsymbol{h}\,\delta(\psi_j\{\boldsymbol{h}\} - \xi) \prod_{j=1}^{N} f(h_j), \qquad (1)$$

where $f(x) \equiv \exp[-(x - s^2)^2/(2s^2)]/\sqrt{2\pi s^2}$ and $\sigma = \mathbf{1}$ is assumed for the code word input.

MAP decoding is optimal, however inefficient, since it requires an exponentially large number (2^L) of steps. BP decoding [3,12] constitutes a fast (linear in N), yet generally approximate alternative, corresponding to replacing the generating function in MAP by solving the following set of nonlinear equations (hereafter referred to as the BP equations) $\eta_{j\alpha} = h_j + \sum_{\beta\neq\alpha}^{j\in\beta} \tanh^{-1} \times [\prod_{i\neq j}^{i\in\beta} \tanh(\eta_{i\beta})]$, and $\eta_j = h_j + \sum_{\beta}^{j\in\beta} \tanh^{-1} \times [\sum_{i\neq j}^{i\in\beta} \tanh(\eta_{i\beta})]$, where $\tanh^{-1}(\psi_j) \equiv \eta_j$. Iterative solutions of the BP equations truncated at a finite step is known as the message passing (MP) algorithm. As shown in [12] the set of BP equations becomes exactly equivalent to MAP in the loop-free approximation. Using physics jargon, it is equivalent to the Bethe-lattice approximation [13]. This basic approximation involves generating a tree with the number of generations, counted from the central variable node to be equal to the shortest loop length on a realistic graph. Note that for Gallager codes the typical length of the shortest loop is estimated as $\sim \ln N$ [4]. Although the method of BER computation proposed in this Letter is generally applicable for any kind of codes, we will focus solely on the regular codes for which each variable node participates in the $m \geq 2$ checking node, and each checking node constraint includes $l \geq 3$ variable nodes, with $l > m$.

The set of the δ-functional BP constraints, leads to essential complications in the generic case resulting in a nontrivial statistical mechanical model. However, in the treelike case (no loops) the constraints become fairly easy to handle. Indeed, in this case each variable site can be described by one "inbound" field $\eta_{j\alpha}$ with the checking site α belonging to the only path from the given variable site to the tree center, and the other $m - 1$, by the "outbound" field $\eta_{j\beta}$ with $\beta \in j$ and $\beta \neq \alpha$. It is a remarkable feature of the tree structure that the integrand in Eq. (1) can be expressed solely in terms of the inbound fields on the tree, and only the outbound field is defined

exactly in the center of the tree. Therefore, the only nontrivial integrations go over the inbound fields, hereafter denoted by simply η_j, and Eq. (1) is simplified to $B_0 = \int_{-1}^{0} d\zeta P_0(\zeta) \sim P_0(0)$ and $P_0(\zeta) = \int(\prod_j d\eta_j)\exp(-\mathcal{Q})$, where the effective action is

$$\mathcal{Q} \equiv \frac{1}{2s^2}\left[\zeta - \sum_{\beta\in 0}\tanh^{-1}\left(\prod_{k\neq 0}^{k\in\beta}\eta_k\right) - s\right]^2 + \frac{1}{2s^2}\sum_{j\neq 0}\left[\eta_j - \sum_{\beta\in j}^{\beta>j}\tanh^{-1}\left(\prod_{k\neq j}^{k\in\beta}\eta_k\right) - s^2\right]^2. \quad (2)$$

$j = 0$ marks the tree center, $\beta > j$ denotes that the check node, and β is positioned above the variable node j in the tree hierarchy.

Integrations over noise fields η_j will be performed in the saddle point instanton fashion that corresponds to the assumption that the major contribution to the integral originates from the special (instanton) configurations related to the minimum of the effective action $\mathcal{Q}$: $\delta\mathcal{Q}/\delta\eta = 0$. Alternatively, one can solve the BP equations on the tree using the MP algorithm (i.e., making some fixed number of iterations), by substituting it into the resulting expression for the magnetization/BER, and maximizing it with respect to the noise field. The two variational schemes should be equivalent in the limit of the infinite number of iterations in the MP case (we found the convergence with the number of iterations to be relatively fast and monotonic in the loop-free case). The result of the MP variational procedure for $m = 2$, $l = 3$, and $n = 3$ (the number of generations on the tree), for ten iterations is shown in Fig. 1. Full variation over all noise fields on the tree (thus containing no symmetry assumption) shows a rich bifurcation picture corresponding to a symmetry breakdown. At small values of SNR the optimal solution is of maximal symmetry with all noise fields

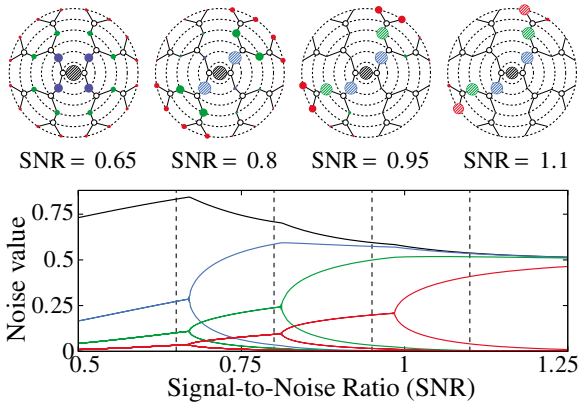

FIG. 1 (color). $m = 2, l = 3$, and $n = 3$. Instantons and bifurcation picture for a complete optimization procedure (no symmetry was *a priori* assumed). The first line area of a circle surrounding any variable node is proportional to the value of the noise on the node. Different colors correspond to different generations on the tree.

that belong to a given generation (counted from the tree center) being identical. With a SNR increase the symmetry of the optimal configuration degrades discreetly through n steps. The symmetry of the k-th order instanton can be described by a set of variable nodes (shown as striped on Fig. 1) that extend from the center (which is always striped) towards the k-th generation according to the following rule: All checking nodes connected to a marked variable node of the previous generation are marked, while for any marked checking node exactly one variable node of the next generation is marked. The rule is generic, i.e., it applies for any values of m and l.

Taking the symmetry assumption for granted, one can substantially simplify and improve the process of finding the set of instanton solutions and getting a better estimate for the BER. Thus the independent fields that correspond to an instanton with the symmetry broken up to the k-th order can be conveniently represented in terms of the two-index quantities $\eta_j^{(p)}$ using the following agreement. The variable $\eta_j^{(p)}$, where $p = 0, \cdots, k$ and $j = 0, \cdots n - 1 - p$, represents the field on a nonmarked node located in generation j (counting from the leaves), so that the first marked node on the only path to the center lies in generation p (counted from the tree center). The variable $\eta_{n-p}^{(p)}$ with $p = 1, \cdots, k$ represents the field on a marked node that is located in the generation p (counting from the center). Replacing the full set of the η fields on the graph by the described above restricted symmetry set $\{\eta_j^{(p)}\}$, substituting it into the effective action $\mathcal{Q}$ described by Eq. (2), and minimizing the resulted k-th order effective equation with respect to the k-th order restricted set of η fields one arrives at a system of equations for the k-th order instanton that are bulky and are not presented here. The set of equations for the k-th instanton can be formulated in terms of a $k + 1$-dimensional minimization problem. We have found, however, that the system can be approximately reduced to a one-dimensional chain minimization problem if either of the following conditions holds: (i) $l \gg 1$; (ii) $n, n - p \gg 1$ and $s > s_c$, where s_c is defined as s, which formally solves the system, $\eta = g(\eta)$ and $1 = g'(\eta)$ where $g(\eta) = s^2 + (m-1)\tanh^{-1} \times [\tanh(\eta^{l-1})]$; (iii) $s \gg s_c$. Note, that in the thermodynamic limit action of the high-symmetry instanton, which is finite at $s < s_c$, then becomes infinite at $s > s_c$, with s_c being finite for $m > 2$. In all three cases the instantons have the following structure: The unmarked variables $\eta_l^{(p)}$ with $p > 0$ grow while approaching the center according to the equation $\eta_j^{(p)} = g(\eta_{j-1}^{(p)})$, whereas for the marked variables $\eta_{n-p}^{(p)} \approx 0$. Therefore, the only dynamical field to be optimized is the unmarked portion of the $p = 0$ branch. Note that although the approximation is justified only in either of the three aforementioned limits, it actually works quantitatively well even for the moderate values of the key parameters l, m, n and s, as follows from comparing the numerical solutions of the

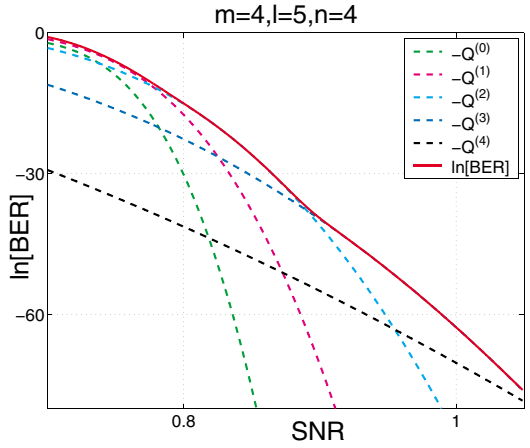

FIG. 2 (color). $m = 4$, $l = 5$, and $n = 4$. Comparative plots of BER (full sum, but the phase volume factors were not counted: $c_p \approx 1$), and individual instanton contributions, calculated within the single-chain approximation, vs SNR.

full (i.e., making no *a priori* symmetry assumptions), $k+1$-dimensional and approximate one-dimensional minimization problems.

Within the instanton approximation the BER is estimated as $B_0 \sim \sum_{k=0}^{n} N_k \exp[-\mathcal{Q}^{(k)}]c_k$, with $\mathcal{Q}^{(k)}$ being the action of the k-th order instanton. The combinatorial factor N_k, with $N_0 = 1$ and $N_k = m(m-1)^{k-1}(l-1)^k$ accounts for the symmetry-induced instantons' degeneracy. The phase space volume c_k occupied by a given instanton accounts for Gaussian fluctuations in the instanton neighborhood. It is possible to show that both $\ln[c_k]$ and $\ln[N_k]$ are subdominant to $\mathcal{Q}^{(k)}$ in any of the three asymptotic limits (i–iii), where low-BER is dominated by a single instanton contribution that determines the relevant SNR phase. Calculations of c_k are to be detailed in a forthcoming more technical publication. At the lowest SNR the major contribution to the BER originates from the most symmetric instanton. With the SNR increasing, the system is coming through a series of phase transitions from $\mathcal{Q}^{(0)}$ to $\mathcal{Q}^{(1)}$, $\mathcal{Q}^{(2)}$, etc., to $\mathcal{Q}^{(n)}$, that take place at $s_1 < s_2 < \cdots < s_n$, respectively. Note, that at $n \to \infty$ an infinite sequence of s_k, with $k < n$, converges to s_c from below. In the case of a finite tree shown in Fig. 2 the transitions are not that sharp, yet are still recognizable.

Emergence of the sequence of phases reported above can also be understood intuitively: If the noise is large, correlations between the noise values on different nodes are weak, thus no symmetry breaking (marked) structure on the Tanner graph is possible and therefore the most symmetric noise configuration is optimal. The correlation length growth due to the SNR increase leads to developing a preferred and/or marked structure that breaks the full symmetry. The structure grows from the tree center toward the leaves, simply because the tree center is chosen for the local measurement of the BER. In the extreme case of large SNR the symmetry breakdown is obviously associated with the structure of the code word closest to the original one, thus making the logarithm of the BER to be proportional to the Hamming distance between the two special code words, and also rationalizes why (for any instanton solution), the marked structure locally resembles the structure of the next-to-original code word. Note also that the emergence of a finite correlation length (on the graph) growing with the SNR increase, suggests that the tree approximation works well for a finite LDPC code as long as the correlation length is short compared to the length of the shortest loop on the LDPC graph. Thus, the no loops/tree approximation is perfectly justified for at least some number of low SNR phases. For the higher SNR phases the approximation may still be reasonable, however, resolving this challenging question requires going beyond the tree approximation. We conclude with noting that emergence of the sequence of transitions suggests a substantial flattening of the BER dependence on the SNR at moderate values of the latter. This observation may have an interesting relation to the error floor phenomenon reported for the frame (code word) error rate [14].

We are thankful to I. Gabitov for many fruitful discussions and support. We also acknowledge the very useful comments of D. Sherrington and A. Montanari that stimulated the development of the project in its early stages.

[1] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[2] R. G. Gallager, *Low Density Parity Check Codes* (MIT Press, Cambridge, MA, 1963).
[3] D. J. C. MacKay, IEEE Trans. Inf. Theory **45**, 399 (1999).
[4] T. J. Richardson and R. L. Urbanke, IEEE Trans. Inf. Theory **47**, 599 (2001).
[5] B. Vasic and O. Milenkovic, IEEE Trans. Inf. Theory **50**, 1156 (2004).
[6] N. Sourlas, Nature (London) **339**, 693 (1989).
[7] P. Ruján, Phys. Rev. Lett. **70**, 2968 (1993).
[8] A. Montanari, Eur. Phys. J. A **23**, 121 (2001).
[9] J. S. Yedidia, W. T. Freeman, and Y. Weiss, www.merl.com/papers/TR2001-16/.
[10] R. Vicente, D. Saad, and Y. Kabashima, Europhys. Lett. **51**, 698 (2000).
[11] I. M. Lifshitz, Usp. Fiz. Nauk **83**, 617 (1964).
[12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference* (Morgan Kaufmann, San Francisco, 1988).
[13] H. A. Bethe, Proc. R. Soc. London A, **150**, 552 (1935).
[14] C. Di, D. Proietti, I. E. Telatar, T. J. Richardson, and R. L. Urbanke, IEEE Trans. Inf. Theory **48**, 1570 (2002).