Diagnosis of Weaknesses in Modern Error Correction Codes: A Physics Approach

M. G. Stepanov,[1,2] V. Chernyak,[3] M. Chertkov,[1] and B. Vasic[4]

[1]*Theory Division and Center for Nonlinear Studies, LANL, Los Alamos, New Mexico 87545, USA*
[2]*Institute of Automation and Electrometry, Novosibirsk 630090, Russia*
[3]*Department of Chemistry, Wayne State University, 5101 Cass Avenue, Detroit, Michigan 48202, USA*
[4]*Department of ECE and Department of Mathematics, University of Arizona, Tucson, Arizona 85721, USA*

(Received 1 June 2005; published 22 November 2005)

One of the main obstacles to the wider use of the modern error-correction codes is that, due to the complex behavior of their decoding algorithms, no systematic method which would allow characterization of the bit-error-rate (BER) is known. This is especially true at the weak noise where many systems operate and where coding performance is difficult to estimate because of the diminishingly small number of errors. We show how the instanton method of physics allows one to solve the problem of BER analysis in the weak noise range by recasting it as a computationally tractable minimization problem.

DOI: 10.1103/PhysRevLett.95.228701 PACS numbers: 89.70.+c, 02.50.−r

Modern technologies, as well as many natural and sociological systems, rely heavily on a wide range of error-correction mechanisms to compensate for their inherent unreliability and to ensure faithful transmission, processing, and storage of information. There has been a great deal of recent research activity in coding theory that has culminated in the recent discovery of coding schemes [1–3] that approach a reliability limit set by classical information theory [4]. The problem considered in this Letter is of a special interest because of a unique feature of the modern coding schemes, which is referred to as an error floor [5,6]. Error floor is a phenomenon characterized by an abrupt degradation of the coding scheme performance, as measured by the bit-error-rate (BER), from the so-called waterfall regime of moderate signal-to-noise ratio (SNR) to the absolutely different error-floor asymptotic achieved at high SNR. To estimate the error-floor asymptotic in the modern high-quality systems is a notoriously difficult task. Typical required BER values are 10^{-12} for an optical communication system, 10^{-15} for hard drive systems in personal computers. However, direct numerical methods, e.g., Monte Carlo simulations, cannot be used to determine the BER below 10^{-9}.

To address this challenge we suggest a physics-inspired approach that ultimately solves the problem of the error-floor analysis. The method is coined the ''instanton'' method, after a theoretical particle in quantum physics that lasts for only an instant, occupying a localized portion of space-time [7]. Statistical physics uses the word instanton to describe a microscopic configuration which, in spite of its rare occurrence, contributes most to the macroscopic behavior of the system [8]. Our instanton is the most probable configuration of the noise to cause a decoding error.

We consider a model of a general communication system with error correction [4]. Data originating from an information source are parsed into fixed length words. Each word is encoded into a longer code word and transmitted through a noisy channel (e.g., a radio or optical link, a magnetic or optical data storage system, etc.). The decoder tries to reconstruct the original code word using the knowledge of the noise statistics and the structure of the code. Error resilience is achieved at the expense of introduced redundancy, and information theory gives conditions for the existence of finite redundancy error-correction codes. However, it does not give a method for realizing decoders of low complexity. There is no better way to reconstruct the code word that was most likely transmitted than to compare the likelihoods of all possible code words. However, this maximal likelihood (ML) algorithm becomes intractable already for code words that are tens of bits long.

A novel exciting era has started in coding theory with the discovery of low-density parity check (LDPC) [1,3,9,10] and turbo [2] codes. These codes are special, not only because they can approach very close to the virtually error-free transmission limit, but mainly because a computationally efficient, so-called iterative, decoding scheme is readily available. When operating at moderate noise values these decoding algorithms show an unprecedented ability to correct errors, a remarkable feature that has attracted a lot of theoretical attention [5,6,11–14]. (Notice also statistical physics-inspired approach [15] that offered an important insight into the extraordinary performance of the iterative decoding [16–18].) It is believed that the error floor is a consequence of iterative decoding, and that the approximate algorithms mentioned above are incapable of matching the performance of ML decoding beyond the error-floor threshold. The importance of error-floor analysis was recognized in the early stages of the turbo codes revolution [19], and it soon became apparent that LDPC codes are also not immune from the error-floor deficiency [6,20]. The main approaches to the error-floor analysis problem proposed to date include: (i) a heuristic approach of the importance sampling type [6], utilizing theoretical considerations developed for a typical randomly constructed LDPC code performing over the very special binary-erasure channel [21], and (ii) deriving lower bounds for BER [22].

Our approach to the error-floor analysis is different: we suggest an efficient numerical scheme, which is *ab initio* by construction, i.e., the scheme requires no additional assumptions (e.g., no sampling). The numerical scheme is also accurate at producing configurations whose validity, as of actual optimal noise configurations, can be verified theoretically and that provide a tight lower bound for BER. Finally, the instanton scheme is also generic, in that there are no restrictions related to the channel or decoding.

Error-correction scheme.—A message word consisting of K bits is encoded in an N-bit long code word, $N > K$. In the case of binary, linear coding, a convenient representation of the code is given by $M \geq N - K$ constraints, often called parity checks or simply, checks. Formally, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$ with $\sigma_i = \pm 1$, is one of the 2^K code words if and only if $\prod_{i \in \alpha} \sigma_i = 1$ for all checks $\alpha = 1, \ldots, M$, where $i \in \alpha$ if the bit i contributes the check α. The relation between bits and checks (we use $i \in \alpha$ and $\alpha \ni i$ interchangeably) is often described in terms of the $M \times N$ parity-check matrix $\hat{H}$ consisting of ones and zeros: $H_{\alpha i} = 1$ if $i \in \alpha$ and $H_{\alpha i} = 0$ otherwise. A bipartite graph representation of $\hat{H}$, with bits marked as circles, checks marked as squares, and edges corresponding to respective nonzero elements of $\hat{H}$, is usually called the Tanner graph of the code. For an LDPC code, $\hat{H}$ is sparse, i.e., most of the entries are zero. Transmitted through a noisy channel, a code word gets corrupted due to the channel noise, so that the channel output (receiver) is $\boldsymbol{x} \neq \boldsymbol{\sigma}$. Even though information about the original code word is lost at the receiver, one still possesses the full probabilistic information about the channel; i.e., the conditional probability, $P(\boldsymbol{x}|\boldsymbol{\sigma}')$, for a code word $\boldsymbol{\sigma}'$ to be a preimage for the output word $\boldsymbol{x}$, is known. In the case of independent noise samples the full conditional probability can be decomposed into the product, $P(\boldsymbol{x}|\boldsymbol{\sigma}') = \prod_i p(x_i|\sigma_i')$. A convenient characteristic of the channel output at a bit is the so-called log-likelihood, $h_i = \log[p(x_i| + 1)/p(x_i| - 1)]/2s^2$, measured in the units of the SNR squared, s^2. (In the physics formulation [15–18], $\boldsymbol{h}$ is called the magnetic field.) For a common model of the white Gaussian symmetric channel, $p(x|\sigma) = \exp[-s^2(x - \sigma)^2/2]/\sqrt{2\pi/s^2}$ (the example discussed in the Letter) $h_i = x_i$. The decoding goal is to infer the original message from the received output $\boldsymbol{x}$. ML decoding (that generally requires an exponentially large number 2^K of steps) corresponds to finding the maximum argument of $P(\boldsymbol{x}|\boldsymbol{\sigma}')$ with respect to all possible $\boldsymbol{\sigma}'$, i.e., the most probable transmitted code word given $\boldsymbol{x}$. Belief propagation (BP) decoding [1,3,9,18] constitutes a fast (linear in K, N), yet generally approximate alternative to ML. As shown in [1] the set of equations describing BP becomes exactly equivalent to the so-called symbol maximum *a posteriori* (MAP) decoding in the loop-free approximation (a similar construction in physics is known as the Bethe-tree approximation [23]), while in the low-noise limit, $s \to \infty$, ML and MAP become indistinguishable and the BP algorithm reduces to the min-sum algorithm:

$$\eta_{i\alpha}^{(n+1)} = h_i + \sum_{\beta \neq \alpha}^{\beta \ni i} \prod_{j \neq i}^{j \in \beta} \mathrm{sgn}[\eta_{j\beta}^{(n)}] \min_{j \neq i}^{j \in \beta} |\eta_{j\beta}^{(n)}|, \qquad (1)$$

where the message field $\eta_{i\alpha}^{(n)}$ is defined on the edge between bit i and check α at the nth step of the iterative procedure and $\eta_{i\alpha}^{(0)} \equiv 0$. The result of decoding is determined by magnetizations, $m_i^{(n)}$, defined by the right-hand side of Eq. (1) with the restriction $\beta \neq \alpha$ dropped. The BER at a given bit i becomes

$$B_i = \int d\boldsymbol{x} \theta(-m_i\{\boldsymbol{x}\}) P(\boldsymbol{x}|\mathbf{1}), \qquad (2)$$

where $\theta(z) = 1$ if $z > 0$ and $\theta(z) = 0$ otherwise; $\boldsymbol{\sigma} = \mathbf{1}$ is assumed for the input (in a symmetric channel the BER is invariant with respect to the choice of the input).

We aim to find optimal configuration of the output $\boldsymbol{x}$, called instanton, $\boldsymbol{x}_{\mathrm{inst}}$, maximizing $P(\boldsymbol{x}|\mathbf{1})$. The integrand in Eq. (2) decays fast and monotonically with output configurations $\boldsymbol{x}$ moving away from $\mathbf{1}$, guaranteeing that the optimal configuration sits on the error surface, $m_i\{\boldsymbol{x}\} = 0$. When the BER is small the integral in Eq. (2) is approximated as $B_i \sim P(\boldsymbol{x}_{\mathrm{inst}}|\mathbf{1})$. For the Gaussian channel finding the instanton, $\boldsymbol{\varphi}_{\mathrm{inst}} = \mathbf{1} - \boldsymbol{x}_{\mathrm{inst}} \equiv l(\boldsymbol{u})\boldsymbol{u}$, turns into minimizing the length $l(\boldsymbol{u})$ with respect to the unit vector $\boldsymbol{u}$, where $l(\boldsymbol{u})$ measures the distance from the zero-noise point to the point on the error surface corresponding to $\boldsymbol{u}$.

Finding the instanton numerically.—In our numerical scheme, the value of the length $l(\boldsymbol{u})$ for any given unit vector $\boldsymbol{u}$ was found by the bisection method. The minimum of $l(\boldsymbol{u})$ was found by a downhill simplex method also called "amoeba" [24], with accurately tailored annealing. The numerical instanton method was first successfully verified in [25] against analytical loop-free results.

Our demonstrative example is the (155, 64, 20) LDPC code described in [26]. (The parity-check matrix of the code is shown in Fig. S1 of [27].) The code includes 155 bits and 93 checks. Each bit is connected to three checks, any check is connected to 5 bits. The minimal Hamming distance of the code is $l_{\mathrm{ML}}^2 = 20$, i.e., at $s \gg 1$, and if the decoding is ML, $B_i \sim \exp(-20 \times s^2/2)$. (See Fig. S2 of [27] for a Monte Carlo evaluation of the BER vs SNR.) We aim to find and describe the instanton(s) that determine the BER in the error-floor regime (for min-sum decoding): $\sim \exp(-l_{\mathrm{ef}}^2 \times s^2/2)$ with $l_{\mathrm{ef}}^2 < l_{\mathrm{ML}}^2 = 20$. Our numerical, and subsequent theoretical, analyses suggest that the instantons, as well as l_{ef}, do depend on the number of iterations. We do not detail this rich dependence here, focusing primarily on the already nontrivial case of four iterations.

The instanton with the minimal length of $l_{\mathrm{a}}^2 = 46^2/210 \approx 10.076$ is shown in the upper part of Fig. 1(a), see also Fig. S3 of [27]. Everywhere away from the 12-bit pattern the noise is numerical zero. The resulting nonzero noise values are proportional to integers (within numerical precision). If decoding starts from the instanton configuration of the noise, magnetization is ex-

actly zero at the bit number 77. This minimal length instanton controls BER at $s \to \infty$, however, for any large but finite s one should also account for many other "close" instantons with $l(\boldsymbol{u}) \approx l_{\mathsf{a}}$, thus approximating $B_i \sim \sum_{\text{inst}} P(\boldsymbol{x}_{\text{inst}}|1)$. The two instanton configurations shown in Fig. 1(b) and 1(c) represent two local minima $l_{\mathsf{b}}^2 = 806/79 \approx 10.203$ and $l_{\mathsf{c}}^2 = 44^2/188 \approx 10.298$, respectively, which are the closest to the minimal one. (See also Figs. S4–S5 of [27].) These instantons were found as a result of multiple attempts at amoeba minimization.

Interpretation of the instantons found.—The remarkable integer/rational structure of the instantons found numerically by amoeba admits a theoretical explanation. Our algebraic construction generalizes the computational tree approach of Wiberg [12]. The computational tree is built by unwrapping the Tanner graph of a given code into a tree from a bit for which we would like to determine the probability of error. (The erroneous bit is shaded in Fig. 1.) The number of generations in the tree is equal to the number of BP iterations (for more details see [11]). As observed in [12], the result of decoding at the shaded bit of the original code is exactly equal to the decoding result in the tree center. Any check node processes messages coming from the tree periphery in the following way: (i) the message with the smallest absolute value (we assume no degeneracy in the beginning) is passed, (ii) the source bit of the smallest message is colored, and (iii) the sign of the product of inputs is assigned to the outcome. At any bit that lies on the colored leaves-to-center path the incoming messages are summed up. The initial messages at any bit of the tree are magnetic fields and, therefore, the result obtained in the tree center is a linear combination of the magnetic fields with integer coefficients. The integer n_i corresponding to bit i of the original graph is the sum of the aforementioned signs over all colored replicas of i on the computational tree. Therefore, the condition at the tree center becomes $\sum_i n_i h_i = 0$. Returning to the original graph and maximizing the integrand of Eq. (2) with the condition enforced, we arrive at the following expressions for the instanton configuration and the effective weight:

$$\varphi_j = n_j \Big(\sum_i n_i\Big) \Big/ \Big(\sum_i n_i^2\Big), \quad l^2 = \Big(\sum_i n_i\Big)^2 \Big/ \Big(\sum_i n_i^2\Big), \quad (3)$$

where the equation applies to the Gaussian channel, however its generalizations to any other channel is straightforward. One can check directly [e.g., looking at Fig. 1(a)], that Eqs. (3) are satisfied for the minimum weight instanton. In this case we find that the signature of any colored message before and after processing through a check remains intact, and thus the resulting n_i for any colored bit is just a total number of the bit's replicas. The structure of this instanton is exactly equivalent to one of the code words on the computational tree, called a pseudocode word as generically it does not correspond to a code word on the original graph [12]. Equation (3) also suggests another possibility that goes beyond the standard pseudocode word construction [12]. In the case shown in Fig. 1(c) the colored part of the tree does correspond to a pseudocode word (by structure), however, the pale part of the computational tree cannot be neglected as the noise values at these nodes are nonzero. This peculiarity is due to the fact that some of the checks shown in the upper part of Fig. 1(c) are connected to more than two colored bits. One finds that the signature of the message propagating from bit 75 to bit 127 alternates because of the pale 0 lying on a leaf, $8/47 = |h_0| > |h_{75}| = 3/47$, $h_0 = -8/47 < 0$. This modifies n_{75} making it equal to 4, as one of the 6 replicas of the bit 75 contributes to the total count -1 instead of $+1$. Moreover,

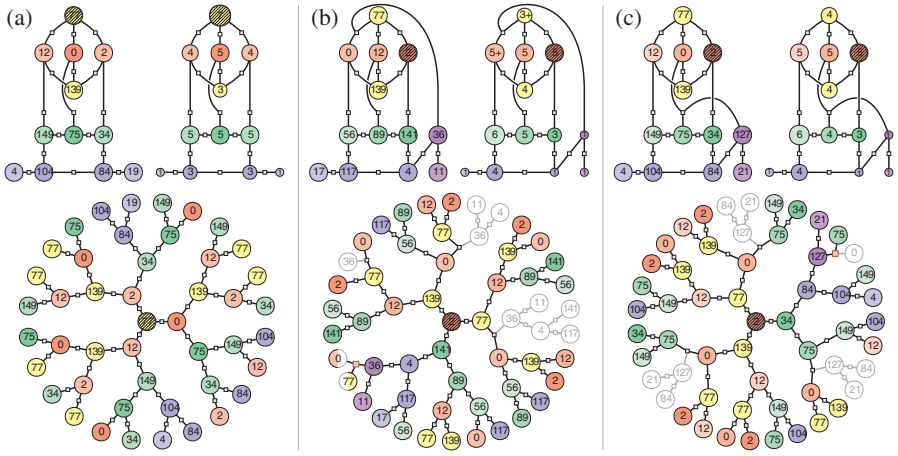

FIG. 1 (color online). Parts of the full Tanner graph with nonzero noise for the instantons, corresponding to (a) simple, (b) degenerate, and (c) sign-alternating pseudocode words, are shown in three panels each consisting of three diagrams. In the top left and bottom diagrams bits are numbered according to the (155, 64, 20)-code definition, where the numbering is fixed up to the full symmetry transformation of the code. The numbers appearing on the top right graphs (and areas of the corresponding circles) are (a) $(1 - x_i) \times 210/46$, (b) $(1 - x_i) \times 79/18$, (c) $(1 - x_i) \times 188/44$. For the computational tree (bottom panel) the bits drawn in color participate in the pseudocode word and the shaded bit marks the error position. The marked checks and squares correspond to the points of (b) degeneracy and of (c) sign alternation.

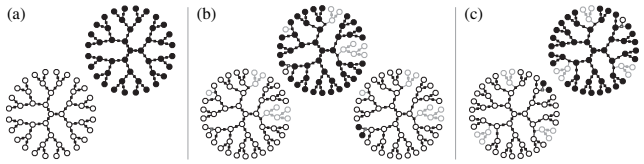

FIG. 2. Interpretation of the instanton as a median within a set of pseudocode words. Three panels show the set of pseudocode words for the three instantons described in Fig. 1. Bits on the computational tree painted in white and black correspond to $+1/-1$. Other notations and marks are in accordance with the captions of Fig. 1.

looking at Fig. 1(b) one finds that the instanton can be even more elaborate as the number of replicas for some bits becomes fractional. (''+'' sign on Fig. 1(b) corresponds to $+7/18$.) This is actually the degenerate case with a colored structure bifurcating at a check (connected to the bits 0, 77, and 36), so that the messages entering the check from two distinct periphery have different signatures but are exactly equal to each other by the absolute value, $h_0 = -h_{77} = 18/79$. Equation (3) does not work for this case, but the following generalization corrects the problem: one needs to introduce an additional condition accounting for the degeneracy. In our example this extra condition can be simply stated as $h_0 = -h_{77}$. (See Fig. S6 of [27].)

Instantons also allow for a complementary interpretation. A decoding error occurs when the magnetization in the computational tree center, which can be considered as a sum over all pseudocode words weighted by, $\exp(s^2 \sum_i h_i p_i)$, turns to zero (with p_i being the number of bit i replicas with the -1 sign in the pseudocode word). In the case of high SNR the sum is dominated by the pseudocode words of maximal weight. Therefore, any instanton, as a configuration of magnetic fields, should be equidistant from some set of $k \geq 2$ pseudocode words: $\sum_i h_i p_i^{(1)} = \cdots = \sum_i h_i p_i^{(k)}$, where at least one of them has a $+1$ value and at least one has a -1 value in the tree center to achieve zero magnetization. And indeed the set of relevant pseudocode words shown in Fig. 2 is a pair in the cases (a),(c) and a triple in the case (b). (See [27] for details.)

To conclude, we demonstrated that the instanton approach is a very powerful and practical instrument for quantitative analysis of the error floor. The success makes us confident that this novel method will be indispensable for design of new error-correcting schemes.

The authors acknowledge very useful and inspiring discussions with participants of workshop on ''Applications of Statistical Physics to Coding Theory'' sponsored by Los Alamos National Laboratory, that took place in Santa Fe, NM, USA. This work was supported in part by the DOE under the LDRD program at Los Alamos National Laboratory, by the NSF under Grant No. CCR-0208597 and No. ITR-0325979, by the INSIC, and by a start-up grant from Wayne State University.

[1] R. G. Gallager, *Low Density Parity Check Codes* (MIT, Cambridge, MA, 1963).
[2] C. Berrou, A. Glavieux, and P. Thitimajshima, in *Proceedings of the IEEE International Conference on Communications, Geneva, Switzerland, 23-26 May 1993* (IEEE, New York, 1993), Vol. 2, p. 1064.
[3] D. J. C. MacKay, IEEE Trans. Inf. Theory **45**, 399 (1999).
[4] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[5] D. J. C. MacKay and M. S. Postol, Electronic Notes in Theoretical Computer Science **74** (2003).
[6] T. Richardson, *Error Floors of LDPC Codes*, in *Proceedings of the 41st Annual Allerton Conference on Communication, Control and Computing, Monticello, Illinois, October, 2003.*
[7] A. A. Belavin, A. M. Polyakov, A. S. Schwartz, and Y. S. Tyupkin, Phys. Lett. **59B**, 85 (1975).
[8] I. M. Lifshitz, Usp. Fiz. Nauk **83**, 617 (1964) [Sov. Phys. Usp. **7**, 549 (1965)].
[9] T. Richardson and R. Urbanke, IEEE Communications Magazine **41**, 126 (2003).
[10] B. Vasic and O. Milenkovic, IEEE Trans. Inf. Theory **50**, 1156 (2004).
[11] N. Wiberg, H.-A. Loeliger, and R. Kotter, Eur. Trans. Telecommun. Relat. Technol. **6**, 513 (1995).
[12] N. Wiberg, Ph.D. thesis, Linköping University, 1996.
[13] G. D. Forney Jr., R. Koetter, F. R. Kschischang, and A. Reznik, in *Codes, Systems, and Graphical Models* (Springer, New York, 2001), p. 101.
[14] Y. Weiss and W. T. Freeman, IEEE Trans. Inf. Theory **47**, 736 (2001).
[15] N. Sourlas, Nature (London) **339**, 693 (1989).
[16] A. Montanari, Eur. Phys. J. B **23**, 121 (2001).
[17] J. S. Yedidia, W. T. Freeman, and Y. Weiss, http://www.merl.com/papers/TR2002-35/.
[18] M. Mezard, Science **301**, 1685 (2003).
[19] S. Benedetto and G. Montorsi, IEEE Trans. Inf. Theory **42**, 409 (1996).
[20] Y. Mao and A. H. Banihashemi, in *Proceedings of the IEEE International Conference on Communications* (IEEE, New York, 2001), Vol. 1, p. 41.
[21] C. Di, D. Proietti, I. E. Telatar, T. J. Richardson, and R. L. Urbanke, IEEE Trans. Inf. Theory **48**, 1570 (2002).
[22] P. O. Vontobel and R. Koetter, in *Proceedings of the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004* (IEEE, New York, 2004), p. 70.
[23] H. A. Bethe, Proc. R. Soc. A **150**, 552 (1935).
[24] W. H. Press *et al. Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press, Cambridge, England, 1988).
[25] V. Chernyak, M. Chertkov, M. G. Stepanov, and B. Vasic, Phys. Rev. Lett. **93**, 198702 (2004).
[26] R. M. Tanner, D. Srdhara, and T. Fuja, in *Proceedings of the 6th International Symposium on Communication Theory and Applications, Ambleside, England, July 2001*.
[27] Figures S1–S6 and supplementary materials are presented in the extended version of this manuscript available at http://arxiv.org/abs/cond-mat/0506037.