Finite Temperature Error-Correcting Codes

Pál Ruján
Fachbereich 8 Physik, Carl-von-Ossietzky Universität, Postfach 2503, 2900 Oldenburg, Germany
(Received 16 February 1993)

The correspondence between error-correcting convolution codes and gauge invariant spin-glass models is used to show that the optimal way to recover the original message is by decoding at a finite temperature $T_N(p) > 0$, where p is the strength of the channel noise and $T_N(p)$ the Nishimori temperature. This improves upon the retrieval performance of the $T = 0$ maximal likelihood Viterbi decoding algorithm without increasing its computational complexity. Numerical simulations support the theory.

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

A typical information processing system consists of an information source, a coding device changing the representation of the data according to some optimality criteria, a noisy transmission channel, and a decoding device reconstructing the original data format. This is a simple but very general model valid for many technological applications.

Consider a message $\vec{s} = (s_1, s_2, \ldots, s_N)$ consisting of N bits $s_i = \pm 1$, sampled independently and identically from the source probability distribution $P_S(\vec{s})$. The message is sent to a receiver $\mathcal{R}$ through a noisy, memoryless communication channel $\mathcal{C}$, which can take as input one binary variable per τ time interval. In practice, the input variable is sent as a physical signal $\pm v$. During transmission this signal is corrupted by white noise with zero mean and w^2 variance. If the receiver $\mathcal{R}$ accepts only binary inputs one can simply assume that every single bit passing through the channel is flipped independently with the same probability $p < \frac{1}{2}$ (binary symmetric channel, BSC). If the receiver can deal with the whole perturbed signal with mean vs_i and variance w^2, the channel is said to be Gaussian (GC). The received message will be denoted by $\vec{\sigma}$ ($\sigma_i = \pm 1$, $i = 1, 2, \ldots, N$) and the average error probability per bit by $p_e(\vec{s}, \vec{\sigma})$.

Encoding (decoding) is a procedure introducing (reducing) redundancy so as to minimize the average error. The rate of the code is defined as $R = N/M$, where M is the length of the message after coding. The capacity of the channel is the maximal mutual information obtainable from all possible source distributions $P_S(\vec{s})$,

$$C = \lim_{N\to\infty} \frac{1}{N} \max_{P_S(\vec{s})} I(\vec{s}, \vec{\sigma}) = \lim_{N\to\infty} \frac{1}{N} \max_{P_S(\vec{s})} [H(\vec{s}) + H(\vec{\sigma}) - H(\vec{s}, \vec{\sigma})], \quad (1)$$

where H is the Shannon entropy. C gives the maximal amount of information per bit which can pass through the channel for a given noise type and strength. For the simple channel models considered here

$$C_{\mathrm{BSC}} = 1 - h(p); \qquad C_{\mathrm{GC}} = \frac{1}{2}\log_2\left(1 + \frac{v^2}{w^2}\right), \quad (2)$$

where $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$. The famous channel coding theorem states that in the thermodynamic limit $N \to \infty$ there are codes which will saturate the channel capacity ($R \to C_-$) with a vanishing *average* error $p_e \to 0$ [1]. Unfortunately, Shannon's proof is not constructive, nor does it consider the algorithmic complexity of the encoding and decoding process. All known codes with computationally tractable coding-decoding algorithms do not saturate the channel capacity [2].

Recently, Sourlas [3] suggested a family of codes based on gauge invariant spin-glass models. The original message is stored as the $T = 0$ ground state of a spin-glass gauge-invariant Hamiltonian and only the (binary) coupling constants are transmitted over the noisy channel. The decoded message is defined as the ground state of a similarly structured Hamiltonian but with a set of coupling constants perturbed by the channel noise. It turns out [4] that the widely used convolution codes correspond to one-dimensional spin-glass models with complicated interactions. The maximum likelihood Viterbi decoding algorithm [2] is equivalent to a transfer matrix method for computing the $T = 0$ ground state.

Encoding consists thus of forming a set of coupling constants $\gamma_\alpha = \pm 1$ as

$$\gamma_\alpha = \prod_{i\in\alpha} s_i\,. \quad (3)$$

i refers to $\vec{r}_i$, a lattice vector in $\mathbb{R}^d$, and α is a k-tuplet of indices $(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_k)$. Obviously, γ_α is a binary variable lying on the vertices of a N/R-dimensional hypercube, the coupling space. A vector $\{\gamma\}$ constructed according to (3) is called a codeword. The Hamming distance between two messages $\vec{s}$ and $\vec{\delta}$ is defined as

$$d(\vec{s}, \vec{\delta}) = \frac{1}{2}\sum_{i=1}^{N}(s_i - \delta_i)^2 = N - \sum_{i=1}^{N} s_i\delta_i\,. \quad (4)$$

The same definition applies to the distance between two points in the coupling space, $D(\{\gamma\}, \{K\})$. Hence, the distance between a codeword generated by the message $\vec{s}$ via Eq. (3) and an arbitrary set of couplings $\{K\}$ is

given by

$$D(\{K\},\vec{s}) = \frac{1}{2}\sum_{\alpha}\Big(K_\alpha - v\prod_{i\in\alpha}s_i\Big)^2 = \frac{1}{2}\sum_{\alpha}K_\alpha^2 + \frac{N}{2R}v^2 + E\,, \tag{5}$$

where $v = 1$ for a binary channel and E is the Hamiltonian

$$E(\{K\},\vec{s}) = -\sum_{\alpha}K_\alpha v\prod_{i\in\alpha}s_i\,. \tag{6}$$

Therefore, the energy of a configuration measures its codeword's distance from a set of coupling constants $\{K\}$. The distance between a configuration $\vec{s}$ and its own codeword $\{v\gamma\}$ is zero, so

$$E(\vec{s}) = E_0 = \min_{\vec{\sigma}} E(\{v\gamma\},\vec{\sigma}) = -\frac{N}{R}v^2\,. \tag{7}$$

Again, $v = 1$ for a binary channel. By carefully eliminating spurious symmetries, the ground state E_0 can be made unique. In the absence of noise, Eq. (3) maps a message into a codeword and the minimization of the energy (6) maps back a codeword into a spin configuration. This one-to-one mapping is shown in Fig. 1(a).

Now switch on the noise. At the receiving end of the channel one obtains a set of coefficients K_α, all identically distributed according to

$$P(K_\alpha) = (1-p)\delta(K_\alpha - \gamma_\alpha) + p\delta(K_\alpha + \gamma_\alpha) \tag{8}$$

for a binary channel and

$$P(K_\alpha) = \frac{1}{\sqrt{2\pi}w}\exp\left[\frac{-(K_\alpha - v\gamma_\alpha)^2}{2w^2}\right] \tag{9}$$

for a Gaussian channel. The effect of noise is thus to move at random the original coupling set $\{\gamma\}$ into a new set of variables $\{K\}$. The *maximal likelihood* decoding procedure [2] is then to find the codeword *closest* to $\{K\}$, which via (5) corresponds to minimizing the energy (6). This is illustrated in Fig. 1(b). The main message of this Letter is that the maximal information about the original message can be obtained from a different type of decoding geometry.

The energy functional $E(\{K\},\vec{\sigma})$, (6), defines a generalized spin-glass model invariant under the gauge transformations

$$\sigma_k \to \epsilon_k\sigma_k,\ \forall k, \quad K_\alpha \to K_\alpha\prod_{i\in\alpha}\epsilon_i,\ \forall\alpha\,, \tag{10}$$

where $\{\epsilon_i = \pm 1\}$ is an *arbitrary* configuration of spins. All configurations $\vec{s}\,'$ obtained from the original message by a gauge transformation (10) have the same energy. One can fix the gauge variables $\{\epsilon_i\}$ by transforming an arbitrary message $\vec{s}$ into a ferromagnetically ordered configuration $s_i' = 1$, $\forall i$, implying $\gamma_\alpha' = 1$, $\forall\alpha$. In this ferromagnetic gauge the average error per bit (4) is simply

$$p_e = \frac{1}{2N}\Big(N - \sum_{i=1}^{N}s_i'\sigma_i\Big) = \frac{1-m}{2}\,, \tag{11}$$

where m is the magnetization per spin of the decoded configuration $\vec{\sigma}$. Likewise, *after transmission* the average energy of the original ferromagnetic message will be

$$E_f^{\rm BSC} = -N\frac{(1-2p)}{R}, \quad E_f^{\rm GC} = -N\frac{v^2}{R}\,, \tag{12}$$

implying

$$D_f^{\rm BSC} = 2p\frac{N}{R}, \quad D_f^{\rm GC} = w^2\frac{N}{2R} \tag{13}$$

for the binary and for the Gaussian channel, respectively.

D_f has a simple geometrical meaning [see Fig. 1(b)]: it is the average distance between the original codeword (message) and the perturbed coupling vector $\{K\}$. This distance is gauge invariant, self-averaging, and depends only on the noise strength and the coding rate. This suggests that instead of choosing the codeword closest to $\{K\}$ ($E = E_0$) one should search for codewords whose energy is between E_f and $E_f + dE$. All configurations $\vec{\sigma}$ whose energy matches that of the original message, (12) are *equally probable* candidates for representing the original message. This type of decoding geometry corresponds to the principle of maximal entropy [5].

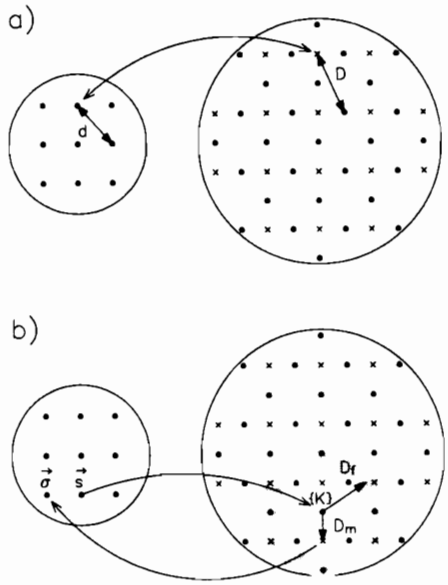

FIG. 1. (a) The message (left) and the coupling space (right). There is a one-to-one map between messages and codewords (crosses) but not every coupling vector (dots) is a codeword. d measures the distance in message, D in the coupling space. (b) After transmission the message $\vec{s}$ "lands" at a coupling vector $\{K\}$. The usual decoding strategy is to look for the closest codeword. D_m corresponds to the minimal energy; the ground state is $\vec{\sigma}$. The maximal entropy decoding suggested in this Letter considers all codewords contained in the spherical shell of D_f radius and dE width (shaded area).

In information theory the decoding procedure must deliver a unique message. This requirement is not always feasible. For the BSC, for example, the $T = 0$ ground state of (6) is macroscopically degenerate due to frustration effects. Hence, one has a large number of possible ground states, none being "better" than the others. In such cases one might resort to the Bayes criterion, which suggests a majority decision for each component σ_i (or any particular "word" formed by σ_i's). Likewise, in order to implement the maximal entropy decoding strategy, one must construct for a given $\{K\}$ *all* configurations of spins with a *fixed* energy E_f and then apply a Bayesian decision. This implies the use of the microcanonical ensemble. In the canonical ensemble formalism the corresponding procedure is to determine the *temperature* at which the *average energy* equals E_f. Fortunately, the solution to this problem is already known.

Several years ago, Nishimori [6, 7] made the remarkable observation that spin-glass models which are invariant under the gauge transformations (10) can be solved analytically at special temperatures. More precisely, all gauge-invariant physical quantities can be expressed as averages over the quenched coupling distribution (8,9) at the following temperatures:

$$\beta_N = \frac{1}{2} \ln \frac{1-p}{p} \tag{14}$$

for the binary distribution (BSC) and

$$\beta_N = \frac{v}{w^2} \tag{15}$$

for the Gaussian distribution (GC), where $\beta = \frac{1}{kT}$ is the inverse temperature. At the Nishimori temperature the internal energy (which is gauge invariant) is thus exactly equal [6] to E_f, (12). Therefore, the loss of information due to channel noise can be simulated by heating up the spin-glass system to the Nishimori temperature.

This observation suggests the following strategy for maximal entropy decoding: (a) Compute the local spin averages

$$m_i(\{K\}, \beta_N) = \langle \sigma_i \rangle = \frac{\sum_{\{\sigma_k\}} \sigma_i \exp(-\beta_N E)}{\sum_{\{\sigma_k\}} \exp(-\beta_N E)} \tag{16}$$

at the Nishimori temperature β_N with E given by (6). (b) Apply the Bayes criterion

$$\sigma_i^{\text{decoded}} = \operatorname{sgn}(m_i) . \tag{17}$$

As an example, consider the following $R = \frac{1}{2}$ code:

$$\gamma_3(k+1) = s_k s_{k+1} s_{k+2}, \quad \gamma_2(k+1) = s_k s_{k+2} , \tag{18}$$

where the variables s_k, $k = 1, 2, \ldots, N$, form a one dimensional chain with free boundary conditions. The passage of the couplings through the channel is simulated by flipping independently the variables $\gamma_{2,3}(k)$ with probability p (BSC). This results on a set of couplings $K_{2,3}(k)$. The energy functional is now

$$E = -\sum_{k=2}^{N-1} E_k ;$$
$$E_k = K_3(k) s_{k-1} s_k s_{k+1} + K_2(k) s_{k-1} s_{k+1} . \tag{19}$$

Note that each bulk spin is "anchored" in place by five different couplings. Flipping one spin requires a $\sim p^3$ order process at low p. A cluster of two neighboring spins has, however, a local field of six coupling constants. If three of those are flipped, the cluster is free to choose its orientation at random (frustration effect). Hence, a macroscopic number of spins will have a vanishing local magnetization at $T = 0$. A more detailed low temperature analysis will be published elsewhere.

In order to compute effectively the local magnetizations I use a simple variant of the transfer formalism [8]. First, assume that the spin variables are summed successively from left to right. In step n one has to perform the sum

$$\sum_{s_n} \Psi_n^{>}(s_n, s_{n+1}) \mathrm{e}^{-\beta E_{n+1}(s_n, s_{n+1}, s_{n+2})} = \lambda_{n+1}^{>} \Psi_{n+1}^{>}(s_{n+1}, s_{n+2}) , \tag{20}$$

where $-\ln \lambda_{n+1}^{>}$ contributes to the free energy. A possible parametrization of the vector Ψ is

$$\Psi_n^{>}(s_n, s_{n+1}) = \exp[\, h_1^{>}(n) s_n + h_2^{>}(n) s_{n+1} + h_{12}^{>}(n) s_n s_{n+1}] \tag{21}$$

with $\Psi_1^{>} = 1$. Similar expressions are used for summing up successively the spins when starting from the right end. The parameters $h_{1,2,12}^{>,<}(n)$ (21) defining the right $\Psi_n^{<}$ and left vectors $\Psi_n^{>}$ are stored during the left and right iterations corresponding to (20). The local magnetizations $\{m_i\}$ (16) (or other local correlations) can be now computed by summing up the chain as far as possible from both the left and the right end. This leads to the expression

$$m_i = \frac{\sum_{s_{i-1}, s_i, s_{i+1}} \Psi_{i-1}^{>} s_i \mathrm{e}^{-\beta E_i} \Psi_{i+1}^{<}}{\sum_{s_{i-1}, s_i, s_{i+1}} \Psi_{i-1}^{>} \mathrm{e}^{-\beta E_i} \Psi_{i+1}^{<}} . \tag{22}$$

Finally, the decoded spin is assigned the sign of the local magnetization.

Table I contains numerical values of the average overlap at different temperatures and noise strengths. The numerical values have been obtained by averaging 10 different transmissions of a message consisting of 10^5 bits. The finite temperature decoding procedure delivers systematically better results (large overlap) than the $T = 0$ maximal likelihood method. The most remarkable improvement is observed at low noise levels. This can be understood as follows: while at $T = 0$ a finite concentra-

TABLE I. The average overlap $o = \sum_{i=1}^{N} s_i^{\text{input}} s_i^{\text{output}}$, $N = 10^5$, and error p_e as a function of noise and inverse temperature ($\beta_N = \frac{1}{2} \ln \frac{1-p}{p}$). The average is taken over ten independent transmissions. Each single run provides the same temperature dependence. $\epsilon - n$ denotes $\times 10^{-n}$.

		$10\beta_N$	$2.0\beta_N$	$1.5\beta_N$	β_N	$0.75\beta_N$	$0.5\beta_N$
$p = 0.025$	β	18.32	3.66	2.75	1.83	1.37	0.92
	o	99945 ± 5.68	99963 ± 2.38	99963 ± 2.38	99963 ± 2.38	99963 ± 2.38	99972 ± 2.98
	p_e	0.275e-3	0.185e-3	0.185e-3	0.185e-3	0.185e-3	0.140e-3
$p = 0.05$	β	14.722	2.94	2.21	1.47	1.10	0.74
	o	98844.2 ± 9.57	98824.6 ± 9.42	98824.6 ± 9.42	988834.2 ± 9.68	98880.6 ± 8.98	98786 ± 38.1
	p_e	0.577e-2	0.588e-2	0.588e-2	0.583e-2	0.56e-2	0.61e-2
$p = 0.075$	β	12.56	2.51	1.88	1.26	0.94	0.63
	o	94429.4 ± 33.2	94437.4 ± 32.0	94477.55 ± 30.9	94519.4 ± 32.3	94481.8 ± 32.3	93589.2 ± 28.3
	p_e	0.0279	0.0278	0.0276	0.0274	0.0276	0.0321
$p = 0.100$	β	10.99	2.20	1.65	1.10	0.82	0.55
	o	87512.8 ± 48.5	87517.6 ± 48.58	87567.4 ± 50.3	87590.8 ± 48.3	87309.6 ± 40.52	85213.2 ± 42.38
	p_e	0.0624	0.0624	0.0622	0.0620	0.0635	0.0739
$p = 0.125$	β	9.73	1.95	1.46	0.97	0.73	0.49
	o	78609.8 ± 89.3	78732.4 ± 93.18	78852 ± 96.86	78875.4 ± 109.3	78329.8 ± 92.43	73752.4 ± 94.71
	p_e	0.107	0.106	0.106	0.1056	0.108	0.131
$p = 0.150$	β	8.67	1.73	1.30	0.87	0.65	0.43
	o	66208.6 ± 112.5	66364.2 ± 109.8	66676.4 ± 105.6	66951 ± 93.7	66178.2 ± 78.9	60739.4 ± 68.45
	p_e	0.169	0.168	0.167	0.165	0.169	0.196

tion of spins is free to flip due to frustration effects, the finite temperature entropy provides an effective stabilizing field. The results obtained at low p suggest that the optimal decoding temperature should be higher than T_N. However, for values of p below 0.05 ($p^3 < 1.25 \times 10^{-4}$) the chain is too short for generating a statistically reliable number of events. The results obtained for $p > 0.05$ clearly show that the best overlaps are obtained at the Nishimori point and that the error increases steeply at higher temperatures.

It is worthwhile remarking that around $p = 0.13$ the code (18) loses its error correcting ability. The convolution codes correct errors by spreading the local spin information over the range of the couplings. As p increase, so does the typical correlation length associated with the disorder. When the mean cluster size of the flipped couplings becomes of the same order of magnitude as the maximal interaction range, the coding fails. This suggests the presence of a geometric phase transition in the coding ability of convolution codes.

The convolution code (18) does not have a particularly good performance, especially not in the BSC setup. It was used here only as an example substantiating the claim that the transmission error is systematically reduced by decoding at finite temperature. Since the Viterbi decoding algorithm is itself equivalent to a one-dimensional transfer matrix method, this performance improvement comes at no additional computational cost.

Sourlas [4] has developed coding schemes based on large non-Abelian alphabets and suggested simulated annealing as a possible decoding procedure. The results presented in this Letter indicate that a simple Monte Carlo simulation at the Nishimori temperature combined with a Boltzmann factor weighted Bayes majority rule might function rather well as decoding algorithm. The above theory applies also to many signal processing applications, like image reconstruction with the random Markov fields method.

Further work is needed in order to analyze and design codes which are well suited for error correction at relatively high noise levels. It seems also possible that biological systems use extensively this type of thermalized information extraction.

I am indebted to Robert Németh for a long collaboration on the mysteries surrounding the Nishimori solution. Moshe Schwartz has helped me with valuable comments and his encouragement. My stay in Tel Aviv was made possible by a Grant from the German–Israeli Foundation for Scientific Research and Development.

[1] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948); **27**, 623 (1948).
[2] R. J. Eliece, *The Theory of Information and Coding*, Encyclopedia of Mathematics and its Applications (Addison-Wesley, Reading, MA, 1977).
[3] N. Sourlas, Nature (London) **339**, 693 (1989).
[4] N. Sourlas, in *Statistical Mechanics of Neural Networks*, edited by L. Garrido, Lecture Notes in Physics Vol. 368 (Springer-Verlag, Berlin, 1991), pp. 317–330.
[5] R. D. Levine, J. Phys. A **13**, 91 (1980).
[6] H. Nishimori, J. Phys. C **13**, 4071 (1980).
[7] H. Nishimori, Progr. Theor. Phys. **66**, 1169 (1981).
[8] P. Rujan, Physica (Amsterdam) **91A**, 549 (1978).