CC

LPTENS 93-32
949402

CERN LIBRARIES, GENEVA

P00019838

**SPIN-GLASSES, ERROR-CORRECTING CODES,
AND FINITE TEMPERATURE DECODING**

Nicolas Sourlas

Laboratoire de Physique Théorique de l' Ecole Normale Supérieure *
24 rue Lhomond, 75231 Paris CEDEX 05, France.
e-mail: sourlas@physique.ens.fr

Abstract

The probability for a sequence of information symbols to have been sent, when the transmission channel's output is known, is simply related to a spin-glass Hamiltonian. The ground state of this Hamiltonian provides the most probable input sequence, given the channel's output. Furthermore, the local magnetization at a certain value of the temperature, is related to the most probable value of the corresponding information symbol, i.e. finite temperature decoding provides the sequence of the most probable symbols, which, in general, is different from the most probable sequence.

LPTENS 93/32

August 93

* Unité Propre du Centre National de la Recherche Scientifique, associée à l' Ecole Normale Supérieure et à l'Université de Paris-Sud.

The mathematical theory of communication[1, 2] is probabilistic in nature. Both the production of information and its transmission are probabilistic events. A source is producing information messages according to a certain probability distribution. Each message consists of a sequence of N bits $\vec{\sigma} = \{\sigma_1, \dots, \sigma_N\}$, $\sigma_i = \pm 1$ and it is assumed that the probability $P_s(\vec{\sigma}) \equiv \exp -H_s(\vec{\sigma})$ of any particular sequence $\vec{\sigma}$ is known. According to Shannon (the "source coding theorem"), it is always possible to encode the source in a way such that all sequences become equally probable ($H_s(\vec{\sigma}) = \text{const.}$, non depending on the σ 's). Source encoding reduces the redundancy in the source messages (not to be confused with "channel encoding", see later). The messages are sent through a transmission channel. In general there is noise during transmission (which may have different origins) which corrupts the transmitted message. If a $\sigma = \pm 1$ is sent through the transmission channel, because of the noise, the output will be a real number u , in general different from σ . Again, the statistical properties of the transmission channel are supposed to be known. Let us call $Q(u|\sigma)$ the probability for the transmission channel output to be between u and $u + du$, when the input was σ , ($Q(u|\sigma)$ is supposed to be known). We assume that the noise is independent for any pair of bits, i.e. $Q(u_1, u_2|\sigma_1, \sigma_2) = Q(u_1|\sigma_1)Q(u_2|\sigma_2)$ ("memoryless channel"). In this framework, communication is a statistical inference problem. Given the transmission channel's output and the statistical properties of the source and of the channel, one has to infer what message was sent as an input. In order to reduce communication errors, one may introduce (deterministic) redundancy into the message ("channel encoding") and use this redundancy to infer the message sent through the channel ("decoding"). The algorithms which transform the source outputs to redundant messages are called error-correcting codes. More precisely, instead of sending the N original bits σ_i , one sends M bits J_k^{in} , $k = 1, \dots, M$, $M > N$, constructed in the following way

$$J_k^{in} = C_{i_1, \dots, i_k}^{(k)}, \quad \sigma_{i_1} \dots \sigma_{i_k}, \quad (1)$$

where the "connectivity" matrix $C_{i_1, \dots, i_k}^{(k)}$ has elements zero or one. For any k , all the $C_{i_1, \dots, i_k}^{(k)}$ except from one are equal to zero, i.e. the J_k^{in} are equal to ± 1 . $C_{i_1, \dots, i_k}^{(k)}$ defines the code, i.e. it tells from which of the σ 's to construct the k th bit of the code. This kind of codes is called parity checking codes.

It has been recently shown[3, 4, 5] that error-correcting codes are mathematically equivalent to theoretical models of disordered spin systems. In particular, minimum error probability decoding (or MED, see later), which is widely used in communications, is equivalent to finding the ground state of a certain spin-glass Hamiltonian. Using the methods of spin-glass theory, it is possible, in certain cases, to compute the error probability as a function of the amount of noise. It has been shown[3, 5], in particular, that the error-correcting code, corresponding to Derrida's random energy model, is ideal, i.e. allows error-free communication, provided the noise is below a certain threshold (i.e. provided one works "below Shannon's channel capacity").

More recently, Ruján[6] suggested a very interesting decoding procedure. Instead of the ground state configuration, he proposes to compute the signs of the local magnetizations at the Nishimori temperature[7] (see later). Nishimori[8] has proven that Ruján's procedure produces an error probability per bit smaller or equal to that of MED. In this paper I will consider a family of decoding procedures which have a lower

(or equal) error probability per bit than MED. In doing so I'll derive Ruján's procedure from first principles and generalize it for the cases the original arguments of Ruján and Nishimori do not apply. The results of this paper are quite general and not particular to error correcting codes and may apply to other problems of statistical inference (I am thinking of neural network models in particular).

More precisely, the equivalence between spin-glass models and error correcting codes is the following[4, 5]. The output of the channel is a sequence of M real numbers $J_k^{out} = \{J_k^{out}, k = 1, \dots, M\}$, which are random variables, obeying the probability distribution $Q(J_k^{out}|J_k^{in})$. Once the channel output J_k^{out} is known, it is possible to compute the probability $P(\vec{\tau}|\vec{J}^{out})$ that a particular sequence $\vec{\tau} = \{\tau_i, i = 1, \dots, N\}$ was the source output (i.e. the information message).

The probability $P(\vec{\tau}|\vec{J}^{out})$ for any sequence $\vec{\tau} = \{\tau_i, i = 1, \dots, N\}$ to be the information message, conditional on the channel output $\vec{J}^{out} = \{J_k^{out}, k = 1, \dots, M\}$ is given by

$$\ln P(\vec{\tau}|\vec{J}^{out}) = \text{const} - H_s(\vec{\tau}) + \sum_{k=1}^M C_{i_1, \dots, i_k}^{(k)} B_k \tau_{i_1} \dots \tau_{i_k} \equiv -H_t(\vec{\tau}) \quad (2)$$

where

$$B_k = B_k(J_k^{out}) = \frac{1}{2} \ln \frac{Q(J_k^{out}|1)}{Q(J_k^{out}|-1)} \quad (3)$$

We recognize in this expression the Hamiltonian of a p-spin spin-glass Hamiltonian where the distribution of the couplings is determined by the probability $Q(J_k^{out}|J_k^{in})$.

The proof is the following. The probability $P(\vec{\tau}|\vec{J}^{out})$ for the source output to be $\vec{\tau}$ when the channel output is $\vec{J}^{out}$ is, by Bayes formula,

$$P(\vec{\tau}|\vec{J}^{out}) = \frac{P_s(\vec{\tau})Q(\vec{J}^{out}|\vec{\tau})}{\sum_{\vec{\tau}'} P_s(\vec{\tau}')Q(\vec{J}^{out}|\vec{\tau}')}$$

where $J_k^{in} = C_{i_1, \dots, i_k}^{(k)} \tau_{i_1} \dots \tau_{i_k}$. Because the channel is memoryless and $J_k^{in} = \pm 1$,

$$\ln P(\vec{\tau}|\vec{J}^{out}) = \text{const.} + \sum_{k=1}^M \ln Q(J_k^{out}|J_k^{in}) + \ln P_s(\vec{\tau})$$

$$\ln Q(J_k^{out}|J_k^{in}) = \frac{1}{2} \ln(Q(J_k^{out}|1)Q(J_k^{out}|-1)) + \frac{J_k^{in}}{2} \ln \frac{Q(J_k^{out}|1)}{Q(J_k^{out}|-1)}$$

where const. means independent of J_k^{in} . To complete the proof, one has to insert the value of the J_k^{in} 's.

In the case when $Q(J_k^{out}|J_k^{in}) = Q(-J_k^{out}|J_k^{in})$ (the case of a "symmetric channel"), $B_k(J_k^{out}) = -B_k(-J_k^{out})$ and one recovers the invariance of the spin-glass Hamiltonian under gauge transformations $\tau_i \rightarrow \epsilon_i \tau_i$, $B_k \rightarrow B_k \epsilon_{i_1} \dots \epsilon_{i_k}$, $\epsilon_i = \pm 1$.

The problem of decoding is to infer, given the channel output, what the original information message was. Let $\vec{\tau}^{(d)}$ be the sequence produced by the decoding procedure. Minimum error probability (or MED) consists in choosing for $\vec{\tau}^{(d)}$ the sequence $\vec{\tau}^0$ which

maximizes the probability $P(\vec{\tau}|\vec{j}^{out})$, i.e. choosing the ground state of the corresponding spin-glass Hamiltonian H_t . When all input sequences are equally probable, H_t is a constant and MED is called maximum likelihood decoding.

When all messages are equally probable and the transmission channel is memoryless and symmetric, the error probability is the same for all input sequences. It is enough to compute it in the case where all input bits are equal to one. In this case, the error probability per bit P_e is $P_e = \frac{1-m^{(d)}}{2}$, where $m^{(d)} = \frac{1}{N} \sum_{i=1}^N \tau_i^{(d)}$.

Let us give a couple of examples of symmetric channels. The first is the case of Gaussian noise (the "Gaussian channel").

$$Q(j^{out}|j^{in}) = c \exp - \frac{(j^{out} - j^{in})^2}{2u^2}, \quad B_k = \frac{j_k^{out}}{u^2} \quad (4)$$

The other example is when the output is again ± 1 (the "binary symmetric channel")

$$Q(j^{out}|j^{in}) = (1-p) \delta_{j^{out}, j^{in}} + p \delta_{j^{out}, -j^{in}} \\ B_k = \frac{\delta_{j_k^{out}, 1}}{2} \ln \frac{1-p}{p} + \frac{\delta_{j_k^{out}, -1}}{2} \ln \frac{p}{1-p} = \frac{j_k^{out}}{2} \ln \frac{1-p}{p} \quad (5)$$

(the last equality holds because in this case $j_k^{out} = \pm 1$).

In the case of spin-glasses, the Nishimori temperature T_N is defined as follows. Let the couplings J of the spin-glass Hamiltonian be independent random variables and their probability distribution be of the form $P(J) = Q(|J|) \exp \gamma J$. Then $T_N = 1/\gamma$. Rujañ proposed for the Gaussian channel or of the binary symmetric channel to use the following decoding procedure $\tau_i \equiv \tau_i^N = \text{sign}(m_i^{(T_N)})$ where $m_i^{(T_N)}$ is the magnetization at site i at T_N . Subsequently Nishimori has proven that $\tau_i^N \geq m_i^{(T_N)}$, where $m_i^{(T)}$ is the magnetization at temperature T and $\bar{\sigma}$ the information sequence. Taking $T = 0$, it follows that, in Rujañ's method, the error probability per bit is lower or equal to the error probability per bit of MED.

How is it possible to find a "better" sequence than the most likely one? Equation (2) contains a lot more information than the one given by the ground state of H_t .

Let us suppose that in addition to the ground state $\vec{\tau}^0$ and its energy E^0 , we can compute the L first excited states $\vec{\tau}^\alpha$, $\alpha = 1, \dots, L$ and their energies E^α . How robust is our estimate τ_i^0 of the information bit σ_i ? This depends on how different the energies E^α are. If the first excited states have much higher energy, all other sequences are much less probable than the ground state and our estimate is reliable. But even if there exist a large number of states almost degenerate to the ground state, if $\tau_i^\alpha = \tau_i^0$ for all those states, it is clear that τ_i^0 is still a very good estimate. In this way, it is possible to know which bits of the ground state sequence are very reliable and which are less. We will suggest, later in this paper, an algorithm to generate, in certain cases, the first excited states.

The average value m of a variable $\tau = \pm 1$ is simply related to its probability distribution $P(\tau = 1) = (1 + m)/2$, so

$$P(\tau_i = 1) = \frac{1 + m_i}{2}, \quad m_i = \frac{\sum_\alpha \tau_i \exp -E^\alpha}{\sum_\alpha \exp -E^\alpha} \quad (6)$$

When $L = 2^N$ is the total number of configurations of the system, the sum in eq. (6) is a thermal average, and

$$P(\tau_i = 1) = \frac{1 + m_i}{2} = \frac{1 + \langle \tau_i \rangle}{2}, \quad m_i = \frac{1}{Z} \sum_{\{\tau^j\}} \tau_i \exp -H(\vec{\tau}) \quad (7)$$

where $\langle \dots \rangle$ means thermal average at temperature one. The most probable value of τ_i is $\tau_i^{\dagger} = \text{sign}(m_i)$, and its reliability is better when $|m_i|$ is large. So $\tau_i^{\dagger}$ is the best estimate of the information bit σ_i , if one disregards correlations with the other bits. We can then consider the sequence $\vec{\tau}^{\dagger}$ as an estimate of the information sequence. We call this procedure $T = 1$ decoding. We may say that $\vec{\tau}^0$ is the most probable sequence, while $\vec{\tau}^{\dagger}$ is the sequence of the most probable bits. In other words, MED has the maximum probability to provide the correct sequence, but when the sequence it provides is incorrect, it is probable that it has many wrong bits, while $T = 1$ decoding has a lower probability to provide the correct sequence, but at the average it makes fewer errors per bit. How "worse" is $\vec{\tau}^{\dagger}$ as a sequence, compared to $\vec{\tau}^0$? It is easy to compute from equation (2) their relative probability.

One can ask, more generally, what is the probability $P(\sigma_1 = \epsilon_1, \dots, \sigma_n = \epsilon_n)$ for n specified σ 's to be equal to n given ϵ 's. The answer is a straightforward generalization of the previous formula $P(\sigma_i = \epsilon_i, \dots, \sigma_n = \epsilon_n) = 2^{-n} < (\tau_1 + \epsilon_1) \dots (\tau_n + \epsilon_n) >$. This means that it is possible to use $T = 1$ decoding to look also at correlations between bits, provided one computes more complicated correlation functions (rather than single site magnetization).

What is the relation of $T = 1$ decoding and Rujañ's proposal of $T = T_N$ decoding? Considering the case of uncorrelated information sequences (i.e. H_t absent from eq. (2)) and looking at the couplings of the Gaussian channel or of the binary symmetric channel, we see that, in both cases, Nishimori's temperature $T_N = 1$. In these cases (and more generally when the Hamiltonian is gauge invariant and the coupling distribution is of the Nishimori form) $T = T_N$ and $T = 1$ decoding are the same. So we have rededuced Rujañ's method and in doing so, we have found what its meaning is in terms of probability theory. However, our derivation is very general and is valid also when the previous conditions are not fulfilled, for example when the transmission channel is not symmetric or when the information sequence is correlated. In both these cases there is no gauge invariance of the Hamiltonian H_t .

In this paper, we have limited ourselves to the case of parity checking codes (i.e. to the case where the σ 's belong to the group Z_2) and memoryless channels. It is clear, however, that the considerations on $T = 1$ decoding are also valid in the more general case of non linear codes[5], where the σ 's are elements of a finite (non abelian) group G , when, in general, there is no gauge invariance of H_t , and in the case of channels with short range memory.

Is it possible to implement the previous ideas in numerical algorithms?

I would like to point out that if one has a maximum likelihood decoding algorithm, i.e. an algorithm for finding the ground state of H_t , then one can find the first excited states of H_t using the same algorithm. The idea is inspired from the study of spin-glasses[10, 11] and is the following. One first determines the ground state $\vec{\tau}^0$ of H_t and its energy E^0 and then considers the following Hamiltonian $H'(\vec{\tau}) = H_t(\vec{\tau}) + \epsilon \sum_{i=1}^N \tau_i \tau_i^0$

where ϵ is a positive constant. The ground state $\vec{\tau}^0$ plays the role of an external magnetic field whose magnitude is determined by ϵ . ϵ plays the role of a repulsion to the ground state. In general the same algorithm can be used to find the ground state $\vec{\tau}'$ of H' . Let us call $E' = H_i(\vec{\tau}')$ and $q' = \sum_i \tau'_i \tau_i^0 \equiv N - 2h'$. h' is called the Hamming distance of the two configurations. If $E^0 + \epsilon N \leq E' + \epsilon q'$ then $\vec{\tau}' = \vec{\tau}^0$, otherwise $\vec{\tau}'$ is a new configuration. Starting from a not too small ϵ from which $\vec{\tau}' \neq \vec{\tau}^0$ we repeat the procedure with a new ϵ , one half the previous one, until we reach some ϵ^0 for which we find again $\vec{\tau}^0$. We can repeat the procedure with other values of $\epsilon > \epsilon^0$ and produce in this way as many as desirable low energy configurations.

Very widely used are the so called "convolutional" codes. Their connectivity matrix C (see eq. (1)) corresponds to one-dimensional spin system with finite range interaction. The zero temperature transfer matrix algorithm is used for MED. (In communication theory, it is called dynamic programming or Viterbi decoding algorithm). The computational complexity of this algorithm scales as $N2^\nu$ where N is the length of the sequence and ν the range of the interaction. For finite temperature decoding of convolutional codes, Ruján[6] proposed a variation of the transfer matrix algorithm[9]. This algorithm has essentially the same computational complexity as the zero temperature transfer matrix algorithm but requires storing $2N(2^\nu - 1)$ coupling constants.

We have presented two extreme decoding strategies, most likely sequence or most likely symbol decoding. In the case when a lot of states are quasi-degenerate to the ground state and the most likely sequence has much higher energy, it makes sense to choose, among the low energy configurations, the sequence with most probable symbols (I also added below the probabilities of certain correlations), i.e. the sequence $\vec{\tau}$ which maximizes $\ln P(\vec{\tau}|\vec{J}^{out}) + \rho \ln P(\vec{\tau} = \vec{\tau}^0) + \rho' \sum_i \ln P(\tau_i \tau_{i+1} = s_i^0)$, where $\tau_i^0 = \text{sign} < \tau_i >, s_i^0 = \text{sign} < \tau_i \tau_{i+1} >$, and averages are at $T = 1$. ρ and ρ' determine the relative weight of the different probabilities one has chosen. This means to take the ground state of

$$H'' = H_i(\vec{\tau}) - \frac{\rho}{2} \sum_i \tau_i \ln \frac{1+m_i}{1-m_i} - \frac{\rho'}{2} \sum_i \tau_i \tau_{i+1} \ln \frac{1+s_i}{1-s_i}$$

Finally, I would like to point out that all the arguments in this paper are only based on probability theory and therefore are not particular to error-correcting codes. It is well known for example in the theory of learning from examples in neural network problems, that finite temperature learning is more efficient than zero temperature learning. As "learning" is another example of statistical inference, it would be interesting to know if probabilistic arguments similar to those presented here, apply to this case also (see in this context the discussion in ref. [12]).

I would like to thank P. Ruján for explaining me his work and H. Nishimori for sending me his paper prior to publication.

References

- 1) Shannon, C. E., A Mathematical Theory of Communication, *Bell Syst. Tech. J.* 27, 379 and 623 (1948)
- 2) Shannon, C. E. and Weaver W., A Mathematical Theory of Communication, (Univ. of Illinois Press, 1963)
- 3) Sourlas, N., *Nature* 339, 693 (1989)
- 4) Sourlas, N., in *Statistical Mechanics of Neural Networks*, Lecture Notes in Physics 368, ed. L. Garrido, Springer Verlag (1990)
- 5) Sourlas, N., Ecole Normale Supérieure preprint (April 1993)
- 6) Ruján, P., *Phys. Rev. Lett.* 70, 2968 (1993)
- 7) Nishimori, H., *Progr. Theor. Phys.* 66, 1169 (1981)
- 8) Nishimori, H., *J. Phys. Soc. Jpn.*, to be published
- 9) Ruján, P., *Physica* 91A, 549 (1978)
- 10) Caracciolo, S., Parisi, G., Patarnello, S. and Sourlas, N., *Europhys. Lett.*, 11, 783 (1990)
- 11) Caracciolo, S., Parisi, G., Patarnello, S. and Sourlas, N., *J. Physique, France*, 51 (1990) 1877
- 12) Oppen, M. and Hausser D., in COLT 91, p. 75