The statistical mechanics of turbo codes

Nicolas Sourlas

(アブストラクト訳)

最近，Berrou 等によって提案された“ターボ符号”は不規則系のスピンハミルトニアンによって記述することが出来る．この符号では熱力学極限，すなわち長符号長の極限で，SN 比 v^2/w^2 がある閾値 Θ 以上ならば１ビット当たりの誤り確率がゼロとなることが示される．我々は２つの典型的なターボ符号に関してこの閾値を計算した．その結果，閾値は個々の符号の構成法に依存することが明らかになった．さらにこれを数値実験と比較した．なお，これらは Andrea Montanari との共同研究によって得られた結果である．

The statistical mechanics of turbo codes.

Nicolas Sourlas
Laboratoire de Physique Théorique de l' Ecole Normale Supérieure *
24 rue Lhomond, 75231 Paris CEDEX 05, France.
email: sourlas@lpt.ens.fr

Abstract

The "turbo codes", recently proposed by Berrou et. al. [1] are written as a disordered spin Hamiltonian. It is shown that there is a threshold Θ such that for signal to noise ratios $v^2/w^2 > \Theta$ the error probability per bit vanishes in the thermodynamic limit, i.e. the limit of infinitly long sequences. The value of the threshold has been computed for two particular turbo codes. It is found that it depends on the code. These results are compared with numerical simulations. This work was done in collaboration with Andrea Montanari [8].

The recent invention of "turbo codes" by Berrou and Glavieux [1] is considered a major breakthrough in communications. For the first time one can communicate almost error-free for signal to noise ratios very close to the theoretical bounds of information theory. Turbo codes are fastly becoming the new standard for error correcting codes in digital communications. The invention of turbo codes and their iterative decoding algorithm was empirical. There is no theoretical understanding of why they are so successfull. The decoding algorithm is thought to be an approximate algorithm. We think that turbo codes are interesting, even outside the context of communication theory, because they provide a non trivial example of a disordered system which can be studied numerically with a fast algorithm.
In this paper we will study turbo codes and turbo decoding using the modern tools of statistical mechanics of disordered systems. We have already shown in the past [4] that there is a mathematical equivalence between error correcting codes and theoretical models of spin glasses. In particular the logarithm of the probability for any given signal, conditional on the communication channel output, has the form of a spin glass Hamiltonian.

*UMR 8549, Unité Mixte de Recherche du Centre National de la Recherche Scientifique et de l' Ecole Normale Supérieure.

We will construct the Hamiltonian which corresponds to the turbo codes and study its properties. This will clarify why they are so successfull. In particular we will show that there is a threshold Θ such that for signal to noise ratios $v^2/w^2 > \Theta$ the average error probability per bit $\overline{P_e}$ vanishes in the thermodynamic limit, i.e. the limit of infinitly long sequences. In $\overline{P_e}$ the average is taken over a large class of turbo codes (see later) and over "channel" noise. The value of the threshold has been computed for two particular turbo codes. It was found that it depends on the code. We also compare these results with numerical simulations.

We consider this result as particularly interesting. They were only two known families of codes which can achieve zero error asymptotically: a) orthogonal codes and b) codes based on Derrida's random energy model. Both families become error-free only in the zero rate limit i.e. the limit of infinite redundancy (see later). Turbo codes are the first finite rate codes to be shown to permit error-free communication.

We consider this paper as a new application of statistical mechanics to a problem outside its original domain of applicability. Our results are typical of the statistical mechanics approach: we study only the average performance of turbo codes, not the performance of any particular one. Furthermore there exist (very "few") particular codes (the ones corresponding to permutations close to the identity, see later), performing much worse than the average. For those codes the error probability per bit $P_e \neq 0$, but they are so "few" that this does not prevent the average $\overline{P_e}$ from vanishing.

Let us first briefly remind the connection between error-correction codes and spin-glass models. In the mathematical theory of communication both the production of information and its transmission are considered as probabilistic events. A source is producing information messages according to a certain probability distribution. Messages of length N are sequencies of N symbols or "letters of an alphabet" $a_1, a_2, \cdots, a_N$. We will assume for simplicity a binary alphabet, i.e. $a_i = 0$ or 1 and that all symbols are equally probable. Instead of a_i we can equally well use Ising spins

$$\sigma_i = (-1)^{a_i} = \pm 1 \tag{1}$$

The messages are sent through a noisy transmission channel. If a $\sigma = \pm 1$ is sent through the transmission channel, because of the noise, the output will be a real number σ^{out}, in general different from σ. Again, the statistical properties of the transmission channel are supposed to be known. Let us call $Q(\boldsymbol{\sigma}^{\text{out}}|\boldsymbol{\sigma})d\boldsymbol{\sigma}^{\text{out}}$ the probability for the transmission channel's output to be between $\boldsymbol{\sigma}^{\text{out}}$ and $\boldsymbol{\sigma}^{\text{out}}+d\boldsymbol{\sigma}^{\text{out}}$, when the input was $\boldsymbol{\sigma}$. $Q(\boldsymbol{\sigma}^{\text{out}}|\boldsymbol{\sigma})$ is supposed to be known. We assume that the noise is independent for any pair of bits ("memoryless channel"), i.e.

$$Q(\boldsymbol{\sigma}^{\text{out}}|\boldsymbol{\sigma}) = \prod_i Q(\sigma_i^{\text{out}}|\sigma_i) \tag{2}$$

In the case of a memoryless channel and a gaussian noise:

$$Q_{\text{gauss}}(\sigma^{\text{out}}|\sigma) \equiv \frac{1}{\sqrt{2\pi w^2}} \exp\left\{-\frac{(\sigma^{\text{out}} - \sigma)^2}{2w^2}\right\} \tag{3}$$

Shannon calculated the channels capacity $\mathcal{C}$, i.e. the maximum information per use of the channel that can be transmitted.

$$\mathcal{C}_{\text{gauss}} = \frac{1}{2}\log_2(1+\frac{v^2}{w^2}) \tag{4}$$

where v^2 is the signal power.
Under the above assumptions, 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. 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}, $k = 1, \cdots, M$,$M > N$, constructed in the following way

$$J_k^{\text{in}} \;=\; C^{(k)}_{i_1 \ldots i_{l_k}}\, \sigma_{i_1} \cdots \sigma_{i_{l_k}} \tag{5}$$

where the "connectivity" matrix $C^{(k)}_{i_1 \ldots i_{l_k}}$ has elements zero or one. For any k, all the $C^{(k)}_{i_1 \ldots i_{l_k}}$ except from one are equal to zero, i.e. the J_k^{in} are equal to ± 1. $C^{(k)}_{i_1 \ldots i_{l_k}}$ defines the code, i.e. it tells from which of the σ's to construct the kth bit of the code.
This kind of codes are called parity checking codes because J_k^{in} counts the parity of the minusis among the l_k σ's. The ratio $R = N/M$ which specifies the redudancy of the code, is called the rate of the code.
Knowing the source probability, the noise probability, the code and the channel output, one has to infer the message that was sent. The quality of inference depends on the choice of the code.
According to the famous Shannon's channel encoding theorem, there exist codes which, in the limit of infinitly long messages, allow error-free communication, provided the rate of the code R is less than the channel capacity $\mathcal{C}$. This theorem says that such "ideal" codes exist, but does not say how to construct them.
We have shown that there exists a close mathematical relationship between error-correcting codes and theoretical models of disosdered systems. As we previously said, the output of the channel is a sequence of M real numbers $\boldsymbol{J}^{\text{out}} = \{J_k^{\text{out}},\ k = 1, \cdots, M\}$, which are random variables, obeying the probability distribution $Q(J_k^{\text{out}}|J_k^{\text{in}})$. Once the channel output $\boldsymbol{J}^{\text{out}}$ is known, it is possible to compute the probability $P(\boldsymbol{\tau}|\boldsymbol{J}^{\text{out}})$ for any particular sequence $\boldsymbol{\tau} = \{\tau_i,\ i = 1, \cdots, N\}$ to be the *source* output (i.e. the information message).
More precisely, the equivalence between spin-glass models and error correcting codes is based on the following property.
The probability $P(\boldsymbol{\tau}|\boldsymbol{J}^{\text{out}})$ for any sequence $\boldsymbol{\tau}$ to be the information message, conditional on the channel output $\boldsymbol{J}^{\text{out}}$ is given by

$$\ln P(\boldsymbol{\tau}|\boldsymbol{J}^{\text{out}}) \;=\; \text{const} \;+ \sum_{k=1}^{M} C^{(k)}_{i_1 \ldots i_{l_k}}\, B_k\, \tau_{i_1} \cdots \tau_{i_{l_k}} \;\equiv\; -H(\boldsymbol{\tau}) \tag{6}$$

where

$$B_k \equiv B(J_k^{\rm out}) \equiv \frac{1}{2}\ln\frac{Q(J_k^{\rm out}|1)}{Q(J_k^{\rm out}|-1)} \tag{7}$$

We recognize in this expression the Hamiltonian of a p-spin spin-glass Hamiltonian. The distribution of the couplings is determined by the probability $Q(J^{\rm out}|J^{\rm in})$.
In the case when $Q(J^{\rm out}|J^{\rm in}) = Q(-J^{\rm out}|-J^{\rm in})$ (the case of a "symmetric channel"), $B(J^{\rm out}) = -B_k(-J^{\rm out})$ and one recovers the invariance of the spin-glass Hamiltonian under gauge transformations.
"Minimum error probability decoding" (or MED), which is widely used in communications [2], consists in choosing the most probable sequence $\boldsymbol{\tau}^0$. This is equivalent to finding the ground state of the above spin-glass Hamiltonian.
Instead of considering the most probable instance, one may only be interested in the most probable value $\tau_i^{\rm MAP}$ of the "bit" τ_i (Maximum A posteriori Probability or MAP decoding) [3] which can be expressed in terms of the magnetization at temperature $T = 1/\beta$ equal to one [6]:

$$\tau_i^{\rm MAP} = \text{sign}\ (m_i) \quad ; \quad m_i = \frac{1}{Z}\sum_{\boldsymbol{\tau}} \tau_i\ \exp\{-H(\boldsymbol{\tau})\} \tag{8}$$

where $H(\boldsymbol{\tau})$ is defined by Eq.(6).
It is remarkable that $\beta = 1$ coincides with the Nishimori temperature in spin glasses [7]. MAP decoding is an essential ingredient in turbo decoding (see later).
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

$$P_e = \frac{1}{2}(1 - m^{(d)}) \equiv \frac{1}{2}\left(1 - \frac{1}{N}\sum_{i=1}^{N}\tau_i^{(d)}\right) \tag{9}$$

and $\tau_i^{(d)}$ is the symbol sequence produced by the decoding procedure. One can derive from this a very general lower bound for P_e, using the analog of the low temperature expansion. An obvious bound (for zero temperature decoding) is provided by the probability $P_e^{(1)}$ that only one bit is incorrect, i.e. $\tau_j = -1$ while all other bits are correct, i.e. $\tau_i = 1$ for all $i \neq j$:

$$P_e \geq P_e^{(1)} = \text{Probability of}\ \left\{\sum_{k\in\Omega(j)} B_k < 0\right\} \tag{10}$$

where the $\Omega(j)$ denotes the set of the couplings in which τ_j appears.
A necessary condition for transmitting without errors is that $\sum_{k\in\Omega(j)} B_k > 0$ with probability one. This is only possible if every spin appears in an infinite number of terms in

the Hamiltonian. Let l_k be the number of spins coupled through the coupling B_k.
The total number of spins beeing N, a spin appears on the average in

$$\frac{1}{N}\sum_{k=1}^{M} l_k = \frac{M}{N}\frac{1}{M}\sum_{k=1}^{M} l_k = \frac{\bar{l}}{R} \tag{11}$$

terms, where $\bar{l}$ is the average of l_k (the number of spins coupled together) and R is the rate of the code.
So a necessary condition for a finite rate code to achieve zero error probability, is that the average number of spins coupled together diverges in the thermodynamic limit ($N \to \infty$). This condition is realised in Derrida's random energy model [5] which has been shown to be an ideal code [4] (in that case $R = 0$).
We will show in the following that this is also true for the case of recursive turbo codes, while it is not true for non recursive turbo codes.
We first present a review of convolutional codes. Convolutional codes are the building blocks of turbo codes. We shall describe both non recursive and recursive convolutional codes. They correspond to one-dimensional spin models. The information message will be denoted by:

$$\boldsymbol{\tau} \equiv (\tau_1, \ldots, \tau_N) \tag{12}$$

It is convenient to think of the source producing a symbol per unit time, i.e. in τ_i, i denotes the time. For simplicity we consider a code of rate $R = 1/2$. The encoded message has the form:

$$\boldsymbol{J} \equiv (J_1^{(1)}, \ldots, J_N^{(1)}; J_1^{(2)}, \ldots, J_N^{(2)}) \tag{13}$$

Any hardware implementation of a convolutional encoder contains a sequence of r memory registers. We shall call r the range of the code.
Let's denote by $\Sigma_1(t), \ldots, \Sigma_r(t)$ the content of the memory registers at time t. At each time step the content of the memory register is shifted to the right:

$$\Sigma_{j+1}(t+1) = \Sigma_j(t) \quad \text{for} \quad j = 1, \ldots, r-1, \quad \Sigma_0(t) \equiv \Sigma_1(t+1) \tag{14}$$

We define the following sequence of bits which we shall call the register sequence:

$$\boldsymbol{\sigma} \equiv (\sigma_1, \ldots, \sigma_N), \quad \sigma_i \equiv \Sigma_0(i)$$

. For non recursive convolutional codes,

$$\sigma_i(\boldsymbol{\tau}) = \Sigma_0(i) = \tau_i \tag{15}$$

The encoded message $\boldsymbol{J}$ is easily defined in terms of the content of the register sequence:

$$J_i^{(n)} = \prod_{j=0}^{r} (\Sigma_j(i))^{\kappa(j;n)} = \prod_{j=0}^{r} (\sigma_{i-j})^{\kappa(j;n)} \tag{16}$$
$$i = 1, \ldots, N \; ; \; n = 1, 2$$
$$\kappa(j;n) \in \{0, 1\}$$

The numbers $\kappa(j;n)$ define the code.
Recursive convolutional codes are most easily defined by

$$\sigma_i(\boldsymbol{\tau}) = \Sigma_0(i) = \tau_i \prod_{j=1}^{r} \Sigma_j(i)^{\kappa(j;1)} \quad = \quad \tau_i \prod_{j=1}^{r} (\sigma_{i-j})^{\kappa(j;1)} \tag{17}$$

We shall now consider decoding. The probability distribution of the register sequence conditional to some ouput can be written as the Boltzmann weight of a spin model with random couplings. The Hamiltonian is given by:

$$H(\boldsymbol{\sigma}; \boldsymbol{J}^{\text{out}}) \quad = \quad -\sum_{i=1}^{N} \left\{ B(J_i^{(1),\text{out}}) \prod_{j=0}^{r} (\sigma_{i-j})^{\kappa(j;1)} + B(J_i^{(2),\text{out}}) \prod_{j=0}^{r} (\sigma_{i-j})^{\kappa(j;2)} \right\} \tag{18}$$

where $B(\cdot)$ is defined in Eq.(7).

We define the decoding at arbitrary temperature $T \equiv 1/\beta$ as follows:

$$\tau_i^{\beta} \quad \equiv \quad \text{sign}(\langle \tau_i(\boldsymbol{\sigma}) \rangle_{\beta}) \tag{19}$$

$$\langle O(\boldsymbol{\sigma}) \rangle_{\beta} \quad \equiv \quad \frac{1}{Z(\boldsymbol{J}^{\text{out}}; \beta)} \sum_{\boldsymbol{\sigma}} O(\boldsymbol{\sigma}) \exp\{-\beta H(\boldsymbol{\sigma}; \boldsymbol{J}^{\text{out}})\} \tag{20}$$

where the expression for $\tau_i(\boldsymbol{\sigma})$ is given by Eq.(17) or by Eq.(15) depending whether the code is recursive or not.
As seen in the introduction there are two widely used decoding strategies:

- Maximum Likelihood decoding which consists in finding the most probable sequence of bits and corresponds to the choice $\beta = \infty$ in Eq.(19): $\tau_i^{ML} \equiv \tau_i^{\beta=\infty}$.
- Maximum A posteriori Probability decoding which consists in finding the most probable sequence of bits and corresponds to the choice $\beta = 1$ in Eq.(19): $\tau_i^{MAP} \equiv \tau_i^{\beta=1}$. This is the strategy which enters in turbo decoding.

Both these strategies can be implemented in a very efficient way using the transfer matrix technique. The corresponding algorithms are known in communication theory as the Viterbi algorithm [2] for the $\beta = \infty$ case and the BCJR algorithm [3] for the $\beta = 1$ case. The complexity of these algorithms grows like $N2^r$.
A turbo code is defined by the choice of a convolutional code and of a permutation of N objects. We use for the permutation the following notation:

$$P : \{1, \ldots, N\} \quad \rightarrow \quad \{1, \ldots, N\}, \quad i \mapsto P(i) \tag{21}$$

The basic idea is to apply the permution P to the source sequence $\boldsymbol{\tau}$ to produce a new sequence $\boldsymbol{\tau}^P$. Both sequences $\boldsymbol{\tau}$ and $\boldsymbol{\tau}^P$ are the inputs to two set of registers, each one implementing a convolutional encoding. In this way the rate of the code is decreased (i.e. greater redundancy).

We illustrate this idea with the example of a rate $1/2$ recursive convolutional code, defined by the constants $\kappa(j;1)$ and $\kappa(j;2)$. The two register sequences are:

$$\sigma_i^{(1)} \equiv \sigma_i(\boldsymbol{\tau}) \quad \Rightarrow \quad \tau_i = \prod_{j=1}^{r} (\sigma_{i-j}^{(1)})^{\kappa(j;1)} \equiv \epsilon_i(\boldsymbol{\sigma}^{(1)}) \tag{22}$$

$$\sigma_i^{(2)} \equiv \sigma_i(\boldsymbol{\tau}^P) \quad \Rightarrow \quad \tau_i^P = \prod_{j=1}^{r} (\sigma_{i-j}^{(2)})^{\kappa(j;1)} \equiv \epsilon_i(\boldsymbol{\sigma}^{(2)}) \tag{23}$$

where $\boldsymbol{\tau}^P$ is the permuted message ($\tau_i^P \equiv \tau_{P(i)}$).
The relation between the two register sequences is rather involved and nonlocal for a general choice of the permutation. Moreover $\sigma_i^{(1)}$ can be expressed only in terms of a large number of $\sigma_j^{(2)}$'s.
It turns out that it is convenient to write the corresponding Hamiltonian as a function of both register sequences. This introduces new degrees of freedom and the Hamiltonian is a function of $2N$ instead of N spin. The unwanted degrees of freedom are eliminated by imposing the constraint $\tau_i^P = \tau_{P(i)}$. This constraint can be written in terms of the σ's using Eqs.(22) and (23). In this way the probability distribution is a local function of the spin variables $\sigma^{(1)}$ and $\sigma^{(2)}$.
Solving the constraint produces an infinite connectivity Hamiltonian for recursive turbo codes. For non recursive codes, the connectivity remains finite. This finite versus infinite connectivity is the essential difference between non recursive and recursive turbo codes and explains why recursive turbo codes are so better and why they can achieve zero error probability for low enough noise.
We now discuss decoding. There is no exact decoding algorithm for turbo codes. Berrou et al. have proposed a very ingenious algorithm, called turbo decoding, which is thought to be approximate. Turbo decoding is an iterative procedure. At each step of the iteration, one considers one of the two chains and proceeds to MAP decoding. The information so obtained is injected to the next step by adding appropriate external fields to the Hamiltonian of the other chain. The algorithm terminates if a fixed point is reached.
Turbo decoding can be seeing as the union of two one dimensional subsystem. Each subsystem acts on the other one through a magnetic field (in the non recursive case) or through an additional coupling (in the recursive case).
We would like to compute the error probability per bit. As explained above, in the case of a symmetric transmission channel, it is enough to compute the magnetization in the case of all inputs $\tau_i = 1$.
The similarity of the Hamiltonian of turbo codes with the Hamiltonians of disordered spin systems is obvious. The disorder in the case of turbo codes has two origins. One is due to the (random) permutation which defines the particular code. The other is more conventional and is related to the randomness of the couplings which is due to the transmission noise. As usual in disordered systems, we can only compute the average over disorder and for that we have to introduce replicas. We briefly report the main results of this approach [8] for the gaussian channel described by Eq.(3).
For recursive turbo codes there exists a low noise phase $w^2 < w_c^2$ where the error probability vanishes in the thermodynamic limit (i.e. for infinitely long sequences). In this

phase the model is completely ordered. A local stability analysis yields the critical value w_{loc}^2 such that for $w^2 > w_{loc}^2$ the no-error phase is destroyed by small fluctuations. Clearly $w_{loc}^2 \geq w_c^2$. We computed w_{loc}^2 for the two cases of turbo codes.

It is well known that the replica method is not mathematically rigorous. So it is natural to question the validity of our results. For this purpose we have carried out numerical simulations. We used the Berrou et al. turbo decoding algorithm and averaged over 200 to 500 realizations of the disorder.
The first conclusion is that recursive turbo codes are much better codes that non recursive ones. Furthemore our results for recursive turbo codes are compatible with the existence of a threshold w_c^2 such that for $w^2 < w_c^2$ the error probability per bit is zero, while no such threshold seems to exist for non recursive codes. This is in agreement with replica theory. Zero error probability can only be achieved in the $N \to \infty$ limit. Our simulations are for $N = 10^5$. It would be interesting to perform a detailed study of finite size corrections, i.e. of the N dependance of the error probability per bit.
Another important issue is the breaking of replica symmetry. Since turbo-decoding is thought to be an approximate algorithm, it may be not the best tool to look for replica symmetry breaking. We have started an analytical investigation of replica symmetry breaking.

References

[1] C.Berrou, A.Glavieux, and P.Thitimajshima. Proc.1993 Int.Conf.Comm. 1064-1070

[2] A.J.Viterbi. IEEE Trans.Com.Technology **COM-19**(1971) 751-771

[3] L.Bahl, J.Cocke, F.Jelinek, and J.Raviv. IEEE Trans.Inf.Theory **IT-20**(1974) 248-287

[4] N.Sourlas. Nature **339**(1989) 693-694
N.Sourlas, in *Statistical Mechanics of Neural Networks*, Lecture Notes in Physics 368, ed. L. Garrido, Springer Verlag (1990)
N.Sourlas, Ecole Normale Supérieure preprint (April 1993)
N.Sourlas, in *From Statistical Physics to Statistical Inference and Back,* ed. P. Grassberger and J.-P. Nadal, Kluwer Academic (1994), page 195.

[5] B.Derrida. Phys.Rev. **B 24**(1981) 2613-2626

[6] P.Ruján. Phys.Rev.Lett. **70**(1993) 2968-2971
N.Sourlas. Europhys.Lett. **25**(1994) 159-164
H.Nishimori. J. Phys. Soc. Jpn. **62**(1993) 2973

[7] H.Nishimori. J.Phys. **C 13**(1980) 4071-4076

[8] A.Montanari and N.Sourlas, to appear.
A.Montanari, in preparation.