One step RSB scheme for the rate distortion function

One step RSB scheme for the rate distortion function

Tatsuto Murayama and Masato Okada

RIKEN Brain Science Institute, Wako, Saitama 351-0198, Japan

Abstract
We apply statistical mechanics to an inverse problem of linear mapping to
investigate the physics of the irreversible compression. We use the replica
symmetry breaking (RSB) technique with a toy model to demonstrate the
Shannon result. The rate distortion function, which is widely known as the
theoretical limit of the compression with a fidelity criterion, is derived using
the Parisi one step RSB scheme. The bound cannot be achieved in the sparsely-
connected systems, where suboptimal solutions dominate the capacity.

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

Statistical physics and information science may have been expected to be directed towards
common objectives since Shannon formulated an information theory based on the concept
of entropy. However, envisaging how this actually happened would have been difficult. The
situation has been greatly changed since the field of disordered statistical systems was maturely
established [1]. The areas where these relations are particularly strong are the Shannon theory
[2] and the replica theory on classical spin systems with quenched disorder [3]. Triggered by
the work of Sourlas [4], these links have recently been examined in the area of error corrections
[5, 6], network information theory [7] and turbo decoding [8]. Recent results of these topics
are mostly derived using the replica trick.

However, the research in the cross-disciplinary field so far can be categorized as a so-
called 'zero distortion' decoding scheme in terms of information theory: the system requires
perfect reproduction of the input alphabets [2]. Here, the same spin glass techniques should be
useful to describe the physics of systems with a fidelity criterion; i.e., a certain degree of
information distortion is assumed when reproducing the alphabets. This framework is called
the rate distortion theory [9, 10]. Though processing information requires regarding the
concept of distortions practically, where input alphabets are mostly represented by continuous
variables, statistical physics only employs a few approaches based on highly modified
perceptrons [11].

In this paper, we introduce a simplified model that achieves optimality, only using parity
checks like Gallager's code [12]. We, then, can easily see how information distortion can
be handled by the concepts of statistical physics. More specifically, we study the inverse
problem of a Sourlas-type decoding problem by using the framework of replica symmetry
breaking (RSB) of diluted disordered systems [13]. According to our analysis, this toy model
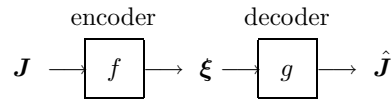
Figure 1. Rate distortion encoder and decoder.

provides an optimal compression scheme for an arbitrary distortion level, though the encoding procedure remains an NP-complete problem without any practical encoders at the moment.

The paper is organized as follows. We first review the concept of the rate distortion theory as well as the main results related to our purpose. We then introduce a toy model. Finally we obtain results consistent with information theory. Detailed derivations will be reported elsewhere.

We start by defining the concepts of the rate distortion theory and stating the simplest version of the main result. Let J be a discrete random variable with alphabet $\mathcal{J}$. Assume that we have a source that produces a sequence $J_1, J_2, \ldots, J_M$, where each symbol is randomly drawn from a distribution. We will assume that the alphabet is finite. Throughout this paper we use vector notation to represent sequences for convenience of explanation: $\boldsymbol{J} = (J_1, J_2, \ldots, J_M)^T \in \mathcal{J}^M$. Here, the encoder describes the source sequence $\boldsymbol{J} \in \mathcal{J}^M$ by a codeword $\boldsymbol{\xi} = f(\boldsymbol{J}) \in \mathcal{X}^N$. The decoder represents $\boldsymbol{J}$ by an estimate $\hat{\boldsymbol{J}} = g(\boldsymbol{\xi}) \in \hat{\mathcal{J}}^M$, as illustrated in figure 1. Note that M represents the length of a source sequence, while N represents the length of a codeword. In this case, the rate is defined by $R = N/M$. Note that the relation $N < M$ always holds when a compression is considered; therefore, $R < 1$ also holds.

A *distortion function* is a mapping $d : \mathcal{J} \times \hat{\mathcal{J}} \to \boldsymbol{R}^+$ from the set of source alphabet–reproduction alphabet pairs into the set of non-negative real numbers. Intuitively, the distortion $d(J, \hat{J})$ is a measure of the cost of representing the symbol J by the symbol $\hat{J}$. This definition is quite general. In most cases, however, the reproduction alphabet $\hat{\mathcal{J}}$ is the same as the source alphabet $\mathcal{J}$. Hereafter, we set $\hat{\mathcal{J}} = \mathcal{J}$ and the following distortion measure is adopted as the fidelity criterion; the *Hamming distortion* is given by

$$d(J, \hat{J}) = \begin{cases} 0 & \text{for} \quad J = \hat{J} \\ 1 & \text{for} \quad J \neq \hat{J} \end{cases} \tag{1}$$

which results in a probable error distortion, since the relation $E[d(J, \hat{J})] = P[J \neq \hat{J}]$ holds, where $E[\cdot]$ represents the expectation and $P[\cdot]$ the probability of its argument. The distortion measure is so far defined on a symbol-by-symbol basis. We extend the definition to sequences. The distortion between sequences $\boldsymbol{J}, \hat{\boldsymbol{J}} \in \mathcal{J}^M$ is defined by $d(\boldsymbol{J}, \hat{\boldsymbol{J}}) = (1/M) \sum_{j=1}^{M} d(J_j, \hat{J}_j)$. Therefore, the distortion for a sequence is the average distortion per symbol of the elements of the sequence. The distortion associated with the code is defined as $D = E[d(\boldsymbol{J}, \hat{\boldsymbol{J}})]$, where the expectation is with respect to the probability distribution on $\mathcal{J}$. A rate distortion pair (R, D) should be *achievable* if a sequence of rate distortion codes (f, g) exists with $E[d(\boldsymbol{J}, \hat{\boldsymbol{J}})] \leqslant D$ in the limit $M \to \infty$. Moreover, the closure of the set of achievable rate distortion pairs is called the *rate distortion region* for a source. Finally, we can define a function to describe the boundary; the *rate distortion function* $R(D)$ is the infimum of rates R, so that (R, D) is in the rate distortion region of the source for a given distortion D.

As in [7], we restrict ourselves to a binary source J with a Hamming distortion measure for simplicity. We assume that binary alphabets are drawn randomly, i.e., the source is not biased

to rule out the possibility of compression due to redundancy. We now find the description rate $R(D)$ required to describe the source with an expected proportion of errors less than or equal to D. In this simplified case, according to Shannon, the boundary can be written as follows; the rate distortion function for a binary source with Hamming distortion is given by

$$R(D) = \begin{cases} 1 - h_2(D) & \text{for} \quad 0 \leqslant D \leqslant 1/2 \\ 0 & \text{for} \quad 1/2 < D \end{cases} \tag{2}$$

where $h_2(\cdot)$ represents the binary entropy function.

Next we introduce a simplified model for the lossy compression. We use the inverse problem of Sourlas-type decoding to realize the optimal encoding scheme [4], a variation of which has recently been investigated by information theorists [14]. As in the previous paragraphs, we assume that binary alphabets are drawn randomly from a non-biased source and that the Hamming distortion measure is selected for the fidelity criterion.

We take the Boolean representation of the binary alphabet $\mathcal{J}$, i.e., we set $\mathcal{J} = \{0, 1\}$. We also set $\mathcal{X} = \{0, 1\}$ to represent the codewords throughout the rest of this paper. Let $\boldsymbol{J}$ be an M-bit source sequence, $\boldsymbol{\xi}$ an N-bit codeword and $\hat{\boldsymbol{J}}$ an M-bit reproduction sequence. Here, the encoding problem can be written as follows. Given a distortion D and a randomly-constructed Boolean matrix A of dimensionality $M \times N$, we find the N-bit codeword sequence $\boldsymbol{\xi}$, which satisfies

$$\hat{\boldsymbol{J}} = A\boldsymbol{\xi} \quad (\text{mod } 2) \tag{3}$$

where the fidelity criterion $D = E[d(\boldsymbol{J}, \hat{\boldsymbol{J}})]$ holds, according to every M-bit source sequence $\boldsymbol{J}$. Note that we applied modulo 2 arithmetics for the additive operations in (3). In our framework, decoding will just be a linear mapping $\hat{\boldsymbol{J}} = A\boldsymbol{\xi}$, while encoding remains an NP-complete problem.

Kabashima and Saad recently expanded on the work of Sourlas, which focused on the zero-rate limit, to an arbitrary-rate case [5]. We follow their construction of the matrix A, so we can treat non-trivial situations. Let the Boolean matrix A be characterized by K ones per row and C per column. The finite, and usually small, numbers K and C define a particular code. The rate of our codes can be set to an arbitrary value by selecting the combination of K and C. We also use K and C as control parameters to define the rate $R = K/C$. If the value of K is small, i.e., the relation $K \ll N$ holds, the Boolean matrix A results in a very sparse matrix. By contrast, when we consider densely constructed cases, K must be extensively big and have a value of $O(N)$. We can also assume that K is not $O(1)$ but $K \ll N$ holds.

The similarity between codes of this type and Ising spin systems was first pointed out by Sourlas, who formulated the mapping of a code onto an Ising spin system Hamiltonian in the context of error correction [4]. To facilitate the current investigation, we first map the problem to that of an Ising model with finite connectivity following the Sourlas method. We use the Ising representation $\{1, -1\}$ of the alphabet $\mathcal{J}$ and $\mathcal{X}$ rather than the Boolean one $\{0, 1\}$; the elements of the source $\boldsymbol{J}$ and the codeword sequences $\boldsymbol{\xi}$ are rewritten in Ising values, and the reproduction sequence $\hat{\boldsymbol{J}}$ is generated by taking products of the relevant binary codeword sequence elements in the Ising representation $\hat{J}_\mu = \prod_{i \in \mathcal{S}(\mu)} S_i$. Here, we denote the set of codeword indices i that participate in the message index μ by $\mathcal{S}(\mu) = \{i | a_{\mu i} = 1\}$ with $A = (a_{\mu i})$. Therefore, chosen i correspond to the ones per row, producing a Ising version of $\hat{\boldsymbol{J}}$. Note that the additive operation in the Boolean representation is translated into multiplication in the Ising one. Hereafter, we set $J_j, \hat{J}_j, \xi_i = \pm 1$ while we do not change the notation for simplicity. As we use statistical-mechanics techniques, we consider the source

and codeword-sequence dimensionality (M and N, respectively) to be infinite, keeping the rate $R = N/M$ finite. To explore the system's capabilities, we examine the Hamiltonian

$$H(\boldsymbol{S}|\boldsymbol{J}) = \sum_{\mu=1}^{M} G[\boldsymbol{S}|J_\mu] \tag{4}$$

with

$$G[\boldsymbol{S}|J_\mu] = -J_\mu \prod_{i \in \mathcal{S}(\mu)} S_i \tag{5}$$

where we have introduced the dynamical variable S_i to find the optimal value of ξ_i, and $G[\boldsymbol{S}|J_\mu]$ denotes the local connectivity of a random hypergraph neighbouring the message bit J_μ [15]. In addition, we now introduce the connectivity tensor $\mathcal{G}$ satisfying the relation

$$\sum_{\mu=1}^{M} \prod_{i \in \mathcal{S}(\mu)} S_i = \sum_{\langle i_1, \dots, i_K \rangle} \mathcal{G}_{\langle i_1, \dots, i_K \rangle} S_{i_1}, \dots, S_{i_K} \tag{6}$$

for any configuration of $\boldsymbol{S}$. Elements of the connectivity tensor $\mathcal{G}_{\langle i_1, \dots, i_K \rangle}$ take the value one if the corresponding indices of codeword bits are chosen (i.e., if all corresponding indices of the matrix A are one) and zero otherwise; C ones per i index represent the system's degree of connectivity.

For calculating the partition function $Z(\boldsymbol{J}) = \mathrm{Tr}_{\{S\}} \exp[-\beta \mathcal{H}(\boldsymbol{S}|\boldsymbol{J})]$, we apply the replica method following the calculation of Kabashima and Saad [5]. To calculate replica free energy, we have to calculate the annealed average of the nth power of the partition function by preparing n replicas. Here we introduce the inverse temperature β, which can be interpreted as a measure of the system's sensitivity to distortions. Although larger values of β seem to be preferable to realize smaller reproduction errors, taking the limit $\beta \to \infty$ fails to provide the optimal solution. This is a direct consequence of the system's irreversibility. As we see in the following calculation, the optimal value of β is naturally determined when the consistency of the replica symmetry breaking scheme is considered [3, 13]. We use integral representations of the Dirac δ function to enforce the restriction, C bonds per index, on $\mathcal{G}$ [16]:

$$\delta \left(\sum_{\langle i_2, i_3, \dots, i_K \rangle} \mathcal{G}_{\langle i, i_2, \dots, i_K \rangle} - C \right) = \oint_0^{2\pi} \frac{\mathrm{d}Z}{2\pi} Z^{-(C+1)} Z^{\sum_{\langle i_2, i_3, \dots, i_K \rangle} \mathcal{G}_{\langle i, i_2, \dots, i_K \rangle}} \tag{7}$$

giving rise to a set of order parameters

$$q_{\alpha, \beta, \dots, \gamma} = \frac{1}{N} \sum_{i=1}^{N} Z_i S_i^\alpha S_i^\beta \cdots S_i^\gamma \tag{8}$$

where $\alpha, \beta, \dots, \gamma$ represent replica indices, and the average over $\boldsymbol{J}$ is taken with respect to the probability distribution

$$P[J_\mu] = \tfrac{1}{2}\delta(J_\mu - 1) + \tfrac{1}{2}\delta(J_\mu + 1) \tag{9}$$

as we consider the non-biased source sequences for simplicity. Assuming the replica symmetry, we use a different representation for the order parameters and the related conjugate variables [16]:

$$q_{\alpha, \beta, \dots, \gamma} = q \int \pi(x) \, \mathrm{d}x \, \tanh^l(\beta x) \tag{10}$$

$$\hat{q}_{\alpha, \beta, \dots, \gamma} = \hat{q} \int \hat{\pi}(\hat{x}) \, \mathrm{d}\hat{x} \, \tanh^l(\beta \hat{x}) \tag{11}$$

where $q = [(K-1)!NC]^{1/K}$ and $\hat{q} = [(K-1)!]^{-1/K}[NC]^{(K-1)/K}$ are normalization constants, and $\pi(x)$ and $\hat{\pi}(\hat{x})$ represent probability distributions related to the integration variables. Here l denotes the number of related replica indices. Throughout this paper, integrals with unspecified limits denote integrals over the range of $(-\infty, +\infty)$. We then obtain an expression for the free energy per source bit expressed in terms of the probability distributions $\pi(x)$ and $\hat{\pi}(\hat{x})$:

$$
\begin{aligned}
-\beta f &= \frac{1}{M} \langle\langle \ln Z(\boldsymbol{J}) \rangle\rangle \\
&= \ln \cosh \beta + \int \left[\prod_{l=1}^{K} \pi(x_l) \, \mathrm{d}x_l \right] \left\langle \ln \left(1 + \tanh \beta J \prod_{l=1}^{K} \tanh \beta x_l \right) \right\rangle_J \\
&\quad - K \int \pi(x) \, \mathrm{d}x \int \hat{\pi}(\hat{x}) \, \mathrm{d}\hat{x} \ln(1 + \tanh \beta x \tanh \beta \hat{x}) \\
&\quad + \frac{C}{K} \int \left[\prod_{l=1}^{C} \pi(\hat{x}_l) \, \mathrm{d}\hat{x}_l \right] \ln \left[\mathrm{Tr}_S \prod_{l=1}^{C} (1 + S \tanh \beta \hat{x}_l) \right]
\end{aligned}
\tag{12}
$$

where $\langle\langle \cdots \rangle\rangle$ denotes the average over quenched randomness of $\boldsymbol{J}$, and also $\mathcal{G}$. The saddle-point equations with respect to probability distributions provide a set of relations between $\pi(x)$ and $\hat{\pi}(\hat{x})$:

$$
\pi(x) = \int \left[\prod_{l=1}^{C} \pi(\hat{x}_l) \, \mathrm{d}\hat{x}_l \right] \delta \left(x - \sum_{l=1}^{C-1} \hat{x}_l \right)
\tag{13}
$$

$$
\hat{\pi}(\hat{x}) = \int \left[\prod_{l=1}^{K} \pi(x_l) \, \mathrm{d}x_l \right] \left\langle \delta \left[\hat{x} - \frac{1}{\beta} \tanh^{-1} \left(\tanh \beta J \prod_{l=1}^{K-1} \tanh \beta x_l \right) \right] \right\rangle_J .
\tag{14}
$$

By using the result obtained for the free energy, we can easily perform further straightforward calculations to find all the other observable thermodynamical quantities, including internal energy:

$$
\begin{aligned}
e &= \frac{1}{M} \langle\langle \mathrm{Tr}_S \, H(\boldsymbol{S}|\boldsymbol{J}) \, \mathrm{e}^{-\beta H(\boldsymbol{S}|\boldsymbol{J})} \rangle\rangle \\
&= -\frac{1}{M} \frac{\partial}{\partial \beta} \langle\langle \ln Z(\boldsymbol{J}) \rangle\rangle
\end{aligned}
\tag{15}
$$

which records reproduction errors. Therefore, in terms of the considered replica symmetric ansatz, a complete solution of the problem seems to be easily obtainable; unfortunately, it is not.

This set of equations (13) and (14) may be solved numerically for general β, K, and C. However, there exists an analytical solution of these equations. We first consider this case. Two dominant solutions emerge that correspond to the paramagnetic and the spin glass phases. The paramagnetic solution, which is also valid for general β, K and C, is in the form of $\pi(x) = \delta(x)$ and $\hat{\pi} = \delta(\hat{x})$; it has the lowest possible free energy per bit $f_{\text{PARA}} = -1$, although its entropy $s_{\text{PARA}} = (R-1) \ln 2$ is positive only for $R \geqslant 1$. It means that the true solution must be somewhere beyond the replica symmetric ansatz. As a first step, which is called one step replica symmetry breaking (RSB), n replicas are usually divided into n/m groups, each containing m replicas. Pathological aspects due to the replica symmetry may be avoided making use of the newly-defined freedom m. Actually, this one step RSB scheme is considered to provide the exact solutions when the random energy model limit is considered [17], while our analysis is not restricted to this case so far.

The spin glass solution can be calculated for both the replica symmetric and the one step RSB ansatz. The former reduces to the paramagnetic solution ($f_{RS} = f_{PARA}$), which is unphysical for $R < 1$, while the latter yields $\pi_{1RSB}(x) = \delta(x)$, $\hat{\pi}_{1RSB}(\hat{x}) = \delta(\hat{x})$ with $m = \beta_g(R)/\beta$ and β_g obtained from the root of the equation enforcing the non-negative replica symmetric entropy

$$s_{RS} = \ln\cosh\beta_g - \beta_g\tanh\beta_g + R\ln 2 = 0 \qquad (16)$$

with a free energy

$$f_{1RSB} = -\frac{1}{\beta_g}\ln\cosh\beta_g - \frac{R}{\beta_g}\ln 2. \qquad (17)$$

The simple expression (17) is derived analytically without using any approximations. However, the stability of the solution must be taken into account when considering the validity.

Since the target bit of the estimation in this model is $J_{\langle i_1,\dots,i_K\rangle}$ and its estimator the product $S_{i_1}\cdots S_{i_K}$, a performance measure for the information corruption could be the per-bond energy e. According to the one step RSB framework, the lowest free energy can be calculated from the probability distributions $\pi_{1RSB}(x)$ and $\hat{\pi}_{1RSB}(\hat{x})$ satisfying the saddle-point equations (13) and (14) at the characteristic inverse temperature β_g, when the replica symmetric entropy s_{RS} disappears. Therefore, f_{1RSB} equals e_{1RSB}. Let the Hamming distortion be our fidelity criterion. The distortion D associated with this code is given by the fraction of free energies that arise in the spin glass phase:

$$D = \frac{f_{1RSB} - f_{RS}}{2|f_{RS}|} = \frac{1 - \tanh\beta_g}{2}. \qquad (18)$$

Here, we substitute the spin glass solutions into the expression, making use of the fact that the replica symmetric entropy s_{RS} disappears at a consistent β_g, which is determined by (16). Using (16) and (18), simple algebra gives the relation between the rate $R = N/M$ and the distortion D in the form

$$R = 1 - h_2(D) \qquad (19)$$

which coincides with the rate distortion function in the Shannon theorem. We do *not* define any nonlinear mappings in the decoding stage but we implicitly do in the encoding stage. This situation is due to the duality between channel coding and lossy compression; the channel capacity can be achieved using linear encoders. Furthermore, we do *not* observe any first-order jumps between analytical solutions. Recently, we have seen that many approaches to the family of codes, characterized by the linear encoding operations, result in a quite different picture; the optimal boundary is constructed in the random energy model limit and is well captured by the concept of a first-order jump. Our analysis of this model, viewed as a kind of inverse problem, provides an exception.

We will now investigate the possibility of other solutions satisfying (13) and (14) in the case of finite K and C. Since the saddle-point equations appear difficult for analytical arguments, we resort to numerical evaluations representing the probability distributions $\pi_{1RSB}(x)$ and $\hat{\pi}_{1RSB}(\hat{x})$ by up to 10^5 bin models and carrying out the integrations by using Monte Carlo methods. Note that the characteristic inverse temperature β_g is also evaluated numerically by using (12) and (15). We firstly calculate the entropy numerically, following the basic relation $f = e - \beta^{-1}s$. Then we choose the proper value of β which provides $s = 0$. We set $K = 2$ and selected various values of C to demonstrate the performance of stable solutions. The numerical results obtained by the one step RSB scenario show suboptimal properties (figure 2). This strongly implies that the analytical solution is not the only stable solution. Furthermore, there have been recent works on the one step RSB solution of the model considered in this paper. The stability of the solution is well examined for some value of K and C [18].
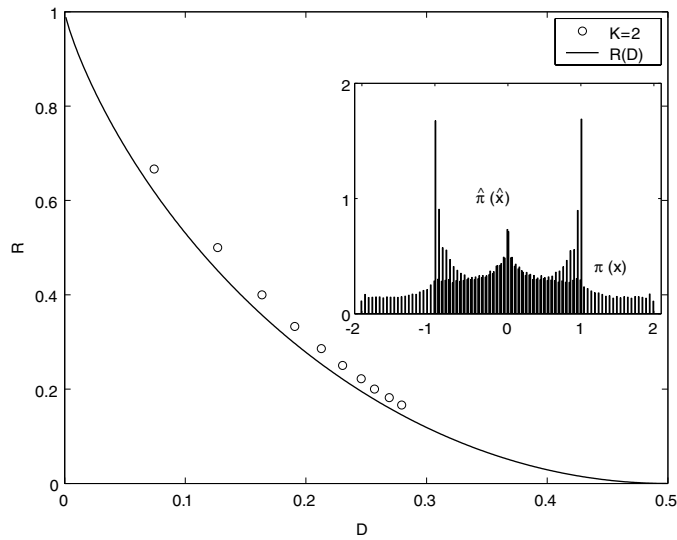
Figure 2. Numerically-constructed stable solutions: stable solutions of (13) and (14) for the finite values of K and L are calculated using Monte Carlo methods. We use 10^5 bin models to approximate the probability distributions $\pi_{1RSB}(x)$ and $\hat{\pi}_{1RSB}(\hat{x})$, starting from various initial conditions. The distributions converge to the continuous ones, giving suboptimal performance. ($\circ$) $K = 2$ and $L = 3, 4, \ldots, 12$; solid line indicates the rate distortion function $R(D)$. Inset: snapshots of the distributions, where $L = 3$ and $\beta_g = 2.35$.

In this paper two points should be noted. Firstly, we find consistency between the Shannon rate distortion theory and the Parisi one step RSB scheme. Secondly, we confirm that the analytical solution, which is consistent with the Shannon result, cannot be stable in the sparsely-connected systems. In the case of sparse models, one might find a polynominal-time algorithm which calculates the suboptimal solutions, providing a practical method of lossy compression. We are currently working on the verification.

Acknowledgments

References

[1] Nishimori H 2001 *Statistical Physics of Spin Glasses and Information Processing* (Oxford: Oxford University Press)
[2] Cover T M and Thomas J A 1991 *Elements of Information Theory* (New York: Wiley)
[3] Dotsenko V 2001 *Introduction to the Replica Theory of Disordered Statistical Systems* (Cambridge: Cambridge University Press)
[4] Sourlas N 1989 *Nature* **339** 693
[5] Kabashima Y and Saad D 1999 *Europhys. Lett.* **45** 97
[6] Kabashima Y, Murayama T and Saad D 2000 *Phys. Rev. Lett.* **84** 1355
[7] Murayama T 2002 *J. Phys. A: Math. Gen.* **35** L95
[8] Montanari A and Sourlas N 2000 *Eur. Phys. J.* B **18** 107
[9] Shannon C E 1959 *IRE National Convention Record,* Part 4 p 142

[10] Berger T 1971 *Rate Distortion Theory: A Mathematical Basis for Data Compression* (Englewood Cliffs, NJ: Prentice-Hall)
[11] Hosaka T, Kabashima Y and Nishimori H 2002 *Phys. Rev.* E **66** 066126
[12] Gallager R G 1962 *IRE Trans. Inf. Theory* **8** 21
[13] Mezard M, Parisi G and Virasoro M 1987 *Spin-Glass Theory and Beyond* (Singapore: World Scientific)
[14] Matsunaga Y and Yamamoto H 2002 *Proc. 2002 IEEE Int. Symp. on Information Theory* p 461
[15] Franz S, Leone M, Ricci-Tersenghi F and Zecchina R 2001 *Phys. Rev. Lett.* **87** 127209
[16] Wong K Y M and Sherrington D 1987 *J. Phys. A: Math. Gen.* **20** L793
[17] Derrida B 1981 *Phys. Rev.* B **24** 2613
[18] Montanari A and Ricci-Tersenghi F 2003 *Preprint* cond-mat/0301591