Statistical mechanics of lossy compression for nonmonotonic multilayer perceptrons

Florent Cousseau*
Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-5861, Japan

Kazushi Mimura†
Faculty of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan

Toshiaki Omori
Brain Science Institute, RIKEN, Saitama 351-0198, Japan

Masato Okada
Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-5861, Japan
and Brain Science Institute, RIKEN, Saitama 351-0198, Japan
(Received 20 March 2008; published 19 August 2008)

A lossy data compression scheme for uniformly biased Boolean messages is investigated via statistical mechanics techniques. We utilize a treelike committee machine (committee tree) and a treelike parity machine (parity tree) whose transfer functions are nonmonotonic. The scheme performance at the infinite code length limit is analyzed using the replica method. Both committee and parity treelike networks are shown to saturate the Shannon bound. The Almeida-Thouless stability of the replica symmetric solution is analyzed, and the tuning of the nonmonotonic transfer function is also discussed.

DOI: 10.1103/PhysRevE.78.021124 PACS number(s): 02.50.−r, 05.50.+q, 89.70.−a

I. INTRODUCTION

The tools of statistical mechanics have been successfully applied in several problems of information theory in recent years. In particular, in the field of error correcting codes [1–4], spreading codes [5,6], and compression codes [7–11], statistical mechanical techniques have shown great potential. The present paper uses similar techniques to investigate a lossy compression scheme. Lossless compression, which was first pointed out in the pioneering paper of Shannon [12], has been widely studied for many years. After much effort, a set of very good codes have been designed and practical implementations have been proposed [13–15]. Lossy compression, on the other hand, was also first studied in another paper of Shannon [16]. A lot of practical lossy compression schemes have been developed over the years (for example JPEG compression, MPEG compression, etc.) but at the present time none of these schemes saturates the Shannon bound given by the *rate-distortion theorem*. Nevertheless, several theoretical schemes that reach this optimal bound have already been proposed. Recently, Shannon optimal codes based on sparse systems have been discovered [7,17–19] and it is now the general tendency to use such kinds of systems. These codes saturate the Shannon bound asymptotically (i.e., for an infinite codeword length), and in the dense generating matrix limit (but a low-connectivity sparse matrix already gives near-Shannon performance). However, there is still a lot of work to be done for densely connected systems. One such system is given by using a perceptron-based decoder. There have been some recent studies on the encoding problem of such schemes using the belief propagation (BP) algorithm and the results seems promising [20]. The foundations of this encoding method for such lossy compression schemes were originally put forward by Murayama using the Thouless-Anderson-Palmer equations applied to Sourlas-type codes [8]. It is important to study a wide class of decoders to extract a pool of schemes that can give near-Shannon-bound performance prior to fully investigate the encoding problem. The study of such schemes could give interesting clues as to how the lossy compression process works, and it might also help to pinpoint some essential features a scheme should possess in order to achieve Shannon optimal performance.

This paper extends the framework introduced in [9–11] and studies three different decoders based on a nonmonotonic multilayer perceptron. Hosaka *et al.* studied a simple perceptron network featuring a nonmonotonic transfer function in order to have a mirror symmetry property in their model [i.e., $f(u)=f(-u)$]. This was motivated by the belief that the Edwards-Anderson order parameter should be zero to reach the Shannon bound. Consequently, if one codeword $\boldsymbol{s}$ is optimal (note that here optimal denotes a codeword that gives the minimum achievable distortion for the concerned scheme), $-\boldsymbol{s}$ is also optimal. Then, they show that for an infinite length codeword, their scheme effectively saturates the Shannon bound. Next, one interesting feature of the model proposed by Mimura *et al.* [11] is to increase the number of optimal codewords by using a multilayer decoder network. The number of optimal codewords is a function of the number of hidden units K in the decoder network (for example, in their parity-tree model with an even number of hidden units, there are at least 2^{K-1} optimal codewords). Thus, one can expect that finding an optimal codeword becomes more and more easy as the number of hidden units increases. Nevertheless, their model deals only with unbiased messages. The main advantage of the model proposed in this

*florent@mns.k.u-tokyo.ac.jp
†mimura@hiroshima-cu.ac.jp

paper is to combine the benefits of the Hosaka *et al.* model (mirror symmetry and ability to deal with biased messages) with the benefits of Mimura *et al.* model (increasing number of optimal codewords with the number of hidden units). By studying three different schemes we would like to extract some essential characteristics that a good lossy compression framework should possess. Finally, the Almeida-Thouless (AT) stability of the obtained solutions is also studied and presents very good properties with almost no unstable part.

The paper is organized as follows. Section II introduces the framework of lossy compression. Section III exposes our model. Section IV presents the mathematical tools used to evaluate the performance of the present scheme. Section V states some results concerning the validity of the obtained solutions and Sec. VI is devoted to conclusion and discussion.

II. LOSSY COMPRESSION

Let us begin by introducing the framework of lossy data compression [21]. Let $\boldsymbol{y}$ be a discrete random variable defined on a source alphabet $\mathcal{Y}$. An original source message is composed of M random variables, $\boldsymbol{y}=(y^1,\ldots,y^M)\in\mathcal{Y}^M$, and compressed into a shorter expression. The encoder compresses the original message $\boldsymbol{y}$ into a codeword $\boldsymbol{s}$, using the transformation $\boldsymbol{s}=\mathcal{F}(\boldsymbol{y})\in\mathcal{S}^N$, where $N<M$. The decoder maps this codeword $\boldsymbol{s}$ onto the decoded message $\hat{\boldsymbol{y}}$, using the transformation $\hat{\boldsymbol{y}}=\mathcal{G}(\boldsymbol{s})\in\hat{\mathcal{Y}}^M$. The encoding and decoding scheme can be represented as in Fig. 1.

In this case, the code rate is defined by $R=N/M$. A distortion function d is defined as a mapping $d:\mathcal{Y}\times\hat{\mathcal{Y}}\rightarrow\mathbb{R}^+$. With each possible pair of $(y,\hat{y})$, it associates a positive real number. In most cases, the reproduction alphabet $\hat{\mathcal{Y}}$ is the same as the alphabet $\mathcal{Y}$ on which the original message $\boldsymbol{y}$ is defined.

Hereafter, we set $\hat{\mathcal{Y}}=\mathcal{Y}$, and we use the Hamming distortion as the distortion function of the scheme. This distortion function is given by

$$d(y,\hat{y})=\begin{cases}0, & y=\hat{y},\\ 1, & y\neq\hat{y},\end{cases} \quad (1)$$

so that the quantity $d(\boldsymbol{y},\hat{\boldsymbol{y}})=\sum_{\mu=1}^{M}d(y^\mu,\hat{y^\mu})$ measures how far the decoded message $\hat{\boldsymbol{y}}$ is from the original message $\boldsymbol{y}$. In other words, it records the error made on the original message during the encoding and decoding process. The probability of error distortion can be written $E[d(y,\hat{y})]=P[y\neq\hat{y}]$ where E represents the expectation. Therefore, the distortion associated with the code is defined as $D=E[(1/M)d(\boldsymbol{y},\hat{\boldsymbol{y}})]$, where the expectation is taken with respect to the probability distribution $P[\boldsymbol{y},\hat{\boldsymbol{y}}]$. D corresponds to the average error per variable, $\hat{y}^\mu$. Now we define a rate distortion pair (R,D) and we say that this pair is achievable

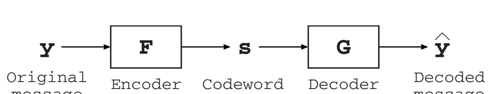

FIG. 1. Rate distortion encoder and decoder.

if there exists a coding-decoding scheme such that, when $M\rightarrow\infty$ and $N\rightarrow\infty$ (note that the rate R is kept finite), we have $E[(1/M)d(\boldsymbol{y},\hat{\boldsymbol{y}})]\leqslant D$. In other words, a rate distortion pair (R,D) is said to be achievable if there exist a pair $(\mathcal{F},\mathcal{G})$ such that $E[(1/M)d(\boldsymbol{y},\hat{\boldsymbol{y}})]\leqslant D$ in the limit $M\rightarrow\infty$ and $N\rightarrow\infty$.

The optimal compression performance that can be obtained in the framework of lossy compression is given by the so-called rate-distortion function $R(D)$ which gives the best achievable code rate R as a function of D [21]. However, despite the fact that the best achievable performance is known, no clues are given about how to construct such an optimal compression scheme. Moreover, finding explicitly the expression of the rate-distortion function is, in general cases, not possible.

Nonetheless, for the special case of uniformly biased Boolean messages in which each component is generated independently by the same probability distribution $P[y=0]=1-P[y=1]=p$, it is possible to calculate analytically the rate distortion function $R(D)$, which becomes

$$R(D)=H_2(p)-H_2(D), \quad (2)$$

where $H_2(x)=-x\log_2 x-(1-x)\log_2(1-x)$. In the following, we restrict ourselves to this particular case (i.e., $P[y=0]=1-P[y=1]=p$ and $\mathcal{Y}=\{0,1\}$).

III. COMPRESSION USING NONMONOTONIC MULTILAYER PERCEPTRONS

In this section we introduce our compression scheme. To make the calculations compatible with the statistical mechanics framework, let us map the Boolean representation {0,1} to the Ising representation $\{-1,1\}$ by means of the mapping $\sigma=(-1)^\rho$, where σ is the Ising variable and ρ is the Boolean one. On top of that, we set $\mathcal{Y}=\mathcal{S}=\hat{\mathcal{Y}}=\{-1,1\}$. Since we consider that all the y^μ are generated independently by an identically biased binary source, we can easily write the corresponding probability distribution

$$P[y^\mu]=p\,\delta(1-y^\mu)+(1-p)\,\delta(1+y^\mu). \quad (3)$$

Next we define the decoder of the compression scheme. We use a nonlinear transformation $\mathcal{G}:\mathcal{S}^N\rightarrow\hat{\mathcal{Y}}^M$ which associates a codeword $\boldsymbol{s}\in\mathcal{S}^N$ with a sequence $\hat{\boldsymbol{y}}\in\hat{\mathcal{Y}}^M$. For a given original message $\boldsymbol{y}$, the encoder is simply defined as follows:

$$\mathcal{F}(\boldsymbol{y})\equiv\arg\min_{\hat{s}}d(\boldsymbol{y},\mathcal{G}(\hat{s})). \quad (4)$$

For the nonlinear transformation $\mathcal{G}$, we utilize nonmonotonic multilayer perceptrons. The codeword $\boldsymbol{s}$ is split into (N/K)-dimensional K disjoint vectors $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K\in\mathcal{S}^{N/K}$ so that $\boldsymbol{s}$ can be written $\boldsymbol{s}=(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K)$. In this paper, we will focus on three different architectures for the nonmonotonic multilayer perceptrons. There are the followings.

(I) A multilayer parity tree with nonmonotonic hidden units. Its output is written

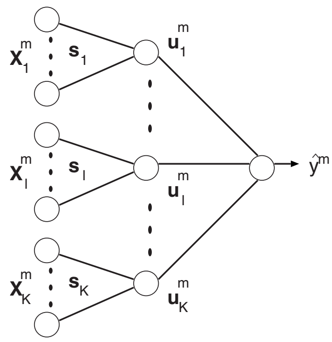

FIG. 2. General architecture of a treelike multilayer perceptron with N input units and K hidden units.

$$\hat{y}^{\mu}(\boldsymbol{s}) \equiv \prod_{l=1}^{K} f_k\left(\sqrt{\frac{K}{N}} \boldsymbol{s}_l \cdot \boldsymbol{x}_l^{\mu} \right). \tag{5}$$

(II) A multilayer committee tree with nonmonotonic hidden units. Its output is written

$$\hat{y}^{\mu}(\boldsymbol{s}) \equiv \mathrm{sgn}\left[\sum_{l=1}^{K} f_k\left(\sqrt{\frac{K}{N}} \boldsymbol{s}_l \cdot \boldsymbol{x}_l^{\mu} \right) \right]. \tag{6}$$

Note that in this case, if the number of hidden units K is even, then there is a possibility to get 0 for the argument of the sign function. We avoid this uncertainty by considering only an odd number of hidden units for the committee tree with nonmonotonic hidden units in the following.

(III) A multilayer committee tree with a nonmonotonic output unit. Its output is written

$$\hat{y}^{\mu}(\boldsymbol{s}) \equiv f_k\left[\sqrt{\frac{1}{K}} \sum_{l=1}^{K} \mathrm{sgn}\left(\sqrt{\frac{K}{N}} \boldsymbol{s}_l \cdot \boldsymbol{x}_l^{\mu} \right) \right]. \tag{7}$$

In each of these structures, f_k is a nonmonotonic function of a real parameter k of the form

$$f_k(x) = \begin{cases} 1 & \text{if } |x| \leqslant k, \\ -1 & \text{if } |x| > k, \end{cases} \tag{8}$$

and the vectors $\boldsymbol{x}_l^{\mu}$ are fixed (N/K)-dimensional independent vectors uniformly distributed on $\{-1,1\}$. "sgn" denotes the sign function, taking the value 1 for $x \geqslant 0$ and -1 for $x<0$. Each of these architectures applies a different transformation to the codeword $\boldsymbol{s}$. The general architecture of these perceptron-based decoders is shown in Fig. 2.

Note that we can also consider a decoder based on a committee tree where both the hidden units and the output unit are nonmonotonic. However, this introduces an extra parameter (we will have one threshold parameter for the hidden units, and one for the output unit) to tune and the performance should not change drastically. For simplicity, we restrict our study to the above three cases only.

Now let us introduce $\mathcal{H}$, an energy function of the system,

$$\mathcal{H}(\boldsymbol{y}, \hat{\boldsymbol{y}}(\boldsymbol{s})) = d(\boldsymbol{y}, \hat{\boldsymbol{y}}(\boldsymbol{s})). \tag{9}$$

This energy function $\mathcal{H}$ is clearly minimized for a codeword $\boldsymbol{s}$ that satisfies Eq. (4). Furthermore, in the Ising representation, the Hamming distance d takes a simple form:

$$d(x,y) = 1 - \Theta(xy), \tag{10}$$

where Θ denotes the unit step function which is 1 for $x \geqslant 0$ and 0 for $x<0$.

The encoding phase can be viewed as a classical perceptron learning problem, where one tries to find the weight vector $\boldsymbol{s}$ that minimizes the energy function $\mathcal{H}$ for the original message $\boldsymbol{y}$ and the random input vector $\boldsymbol{x}$. The vector $\boldsymbol{s}$ which achieves this minimum gives us the codeword to be sent to the decoder. Therefore, in the case of a lossless compression scheme (i.e., $D=0$), evaluating the rate-distortion property of the present scheme is equivalent to finding the number of couplings $\boldsymbol{s}$ that satisfies the input-output relation $\boldsymbol{x}^{\mu} \mapsto y^{\mu}$. In other words, this is equivalent to the calculation of the storage capacity of the network [22,23].

The choice of parity-tree- and committee-tree-based networks is motivated by the thorough literature available on this kind of network. Parity and committee machines have been intensively studied (see [24] for an overview) by the machine-learning community over the years. The techniques used to calculate the storage capacity of such networks give us a starting point for our analytical evaluation of the typical performance of the above schemes.

IV. ANALYTICAL EVALUATION

We analyze the performance of these three different compression schemes using the tools of statistical mechanics. We first define the following partition function:

$$Z(\beta, \boldsymbol{y}, \boldsymbol{x}) = \sum_{\boldsymbol{s}} \exp[-\beta \mathcal{H}(\boldsymbol{y}, \hat{\boldsymbol{y}}(\boldsymbol{s}))], \tag{11}$$

where the sum over $\boldsymbol{s}$ represents the sum over all the possible states for the vector $\boldsymbol{s}$. β denotes the inverse temperature parameter. Such a partition function can be identified with the partition function of a spin glass system with dynamical variables $\boldsymbol{s}$ and quenched variables $\boldsymbol{x}$. For a fixed Hamming distortion $MD = E[d(\boldsymbol{y}, \hat{\boldsymbol{y}})]$, the average of this partition function over $\boldsymbol{y}$ and $\boldsymbol{x}$ naturally contains all the interesting typical properties of the scheme, such as the entropy. However, evaluating this average is hard and we need some technique to investigate it. In this paper we use the so-called replica method in order to calculate the average of the partition function. In the case of such a discrete system, the entropy should not be negative, so that the zero-entropy criterion (see [23]) gives us the best achievable code rate limit. The replica method's calculations to obtain the average of the partition function $\langle Z(\beta, \boldsymbol{y}, \boldsymbol{x}) \rangle_{\boldsymbol{y},\boldsymbol{x}}$ are detailed in Appendix A.

A. Replica symmetric solution for the parity tree with nonmonotonic hidden units

In the lossy compression scheme using a parity tree with nonmonotonic hidden units (5), the replica symmetric free energy is given by

$$f(\beta,R,k) = -\frac{1}{\beta}\{p \ln[e^{-\beta} + (1-e^{-\beta})A_k] + (1-p)\ln[e^{-\beta} + (1-e^{-\beta})(1-A_k)] + R \ln 2\}, \quad (12)$$

where

$$A_k = \frac{1}{2} + \frac{1}{2}[1-4H(k)]^K,$$

$$H(k) = \int_k^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt. \quad (13)$$

The internal free energy is

$$u(\beta,k) = p\frac{e^{-\beta}(1-A_k)}{e^{-\beta} + (1-e^{-\beta})A_k} + (1-p)\frac{e^{-\beta}A_k}{e^{-\beta} + (1-e^{-\beta})(1-A_k)}. \quad (14)$$

Minimizing the free energy with respect to A_k, taking the zero-temperature limit $\beta \to \infty$, and using the identity (A6) gives

$$A_k = \frac{p-D}{1-2D}, \quad (15)$$

$$D = \frac{e^{-\beta}}{1+e^{-\beta}}. \quad (16)$$

Finally, using the zero-entropy criterion, one can get

$$R = H_2(p) - H_2(D), \quad (17)$$

which is identical to the rate-distortion function (2).

However, this minimum is reached under the conditions given by Eqs. (15) and (16). Since D is fixed, the condition given by Eq. (16) is easily satisfied by choosing the proper inverse temperature parameter $\beta = \ln[(1-D)/D]$. On the other hand, the condition given by Eq. (15) is satisfied by properly tuning the parameter k of the nonmonotonic function f_k. Let us denote the optimal k that satisfies Eq. (15) by $\hat{k}$. In the case of the parity tree, this optimal $\hat{k}$ is such that the following equation becomes true:

$$H(\hat{k}) = \frac{1}{4}\left(1 - \sqrt[K]{\frac{2p-1}{1-2D}}\right). \quad (18)$$

In this paper we consider that $(p,D) \in \{[0,\frac{1}{2}]\}^2$; therefore, one can easily show that for $p \neq \frac{1}{2}$ then $(2p-1)/(1-2D)$ is negative, which implies that there is no real solution for the above equation if we have an even number of hidden units K (because of the Kth root). However, we can also consider the case where $(p,D) \in \{[\frac{1}{2},1] \times [0,\frac{1}{2}]\}$ without any change (the probabilities of $y=1$ and $y=-1$ are just inverted) and in this case $(2p-1)/(1-2D)$ is positive which implies that there is always a solution for any value of K. The above problem is just a consequence of the definition of p, but is not related to the model. So, in the case of the parity tree, $\hat{k}$ always exists independently of the value of K.

Finally, since we used the replica symmetric (RS) ansatz, we have to verify the Almeida-Thouless stability of the solution to confirm its validity. This is done in the next section.

B. Replica symmetric solution for the committee tree with nonmonotonic hidden units

In the lossy compression scheme using a committee tree with nonmonotonic hidden units (6), the replica symmetric free energy is given by

$$f(\beta,R,k) = -\frac{1}{\beta}\{p \ln[e^{-\beta} + (1-e^{-\beta})B_k] + (1-p)\ln[e^{-\beta} + (1-e^{-\beta})(1-B_k)] + R \ln 2\}, \quad (19)$$

where

$$B_k = \sum_{\tau_l=\pm 1}\left[\Theta\left(\sum_{l=1}^{K}\tau_l\right)\prod_{l=1}^{K}\left(\frac{1+\tau_l(1-4H[k])}{2}\right)\right]. \quad (20)$$

The sum over τ_l represents the sum over each possible state for the dummy variable τ_l ($\tau_l = \pm 1$). The internal free energy is

$$u(\beta,k) = p\frac{e^{-\beta}(1-B_k)}{e^{-\beta} + (1-e^{-\beta})B_k} + (1-p)\frac{e^{-\beta}B_k}{e^{-\beta} + (1-e^{-\beta})(1-B_k)}. \quad (21)$$

As in the parity tree case, after minimizing the free energy with respect to B_k, taking the zero-temperature limit $\beta \to \infty$, and using the identity (A6), we obtain

$$B_k = \frac{p-D}{1-2D}, \quad (22)$$

$$D = \frac{e^{-\beta}}{1+e^{-\beta}}. \quad (23)$$

Finally, using the zero-entropy criterion, one can get

$$R = H_2(p) - H_2(D), \quad (24)$$

which is identical to the rate-distortion function (2). However, here it is not easy to discuss the existence of an optimal $\hat{k}$ that satisfies the condition given by Eq. (22). Such an optimal $\hat{k}$ satisfies the following equation:

$$\sum_{\tau_l=\pm 1}\left[\Theta\left(\sum_{l=1}^{K}\tau_l\right)\prod_{l=1}^{K}\left(\frac{1+\tau_l(1-4H[\hat{k}])}{2}\right)\right] = \frac{p-D}{1-2D}. \quad (25)$$

We will discuss this existence problem a bit more in the next section, when we check the AT stability of the RS solution.

C. Replica symmetric solution for the committee tree with a nonmonotonic output unit

In the lossy compression scheme using a committee tree with a nonmonotonic output unit (7), the replica symmetric free energy is given by

$$f(\beta,R,k) = -\frac{1}{\beta}\{p \ln[e^{-\beta} + (1-e^{-\beta})C_k] + (1-p)\ln[e^{-\beta} + (1-e^{-\beta})(1-C_k)] + R \ln 2\}, \tag{26}$$

where

$$C_k = 2^{-K}\sum_{l=0}^{K}\binom{K}{l}\Theta\left(k^2 - \frac{1}{K}(2l-K)^2\right). \tag{27}$$

The term $\binom{n}{l}$ denotes the binomial coefficient. The internal free energy is

$$u(\beta,k) = p\frac{e^{-\beta}(1-C_k)}{e^{-\beta} + (1-e^{-\beta})C_k} + (1-p)\frac{e^{-\beta}C_k}{e^{-\beta} + (1-e^{-\beta})(1-C_k)}. \tag{28}$$

As in the parity tree case, after minimizing the free energy with respect to C_k, taking the zero-temperature limit $\beta\to\infty$, and using the identity (A6), we obtain

$$C_k = \frac{p-D}{1-2D}, \tag{29}$$

$$D = \frac{e^{-\beta}}{1+e^{-\beta}}. \tag{30}$$

Finally, using the zero-entropy criterion, one can get

$$R = H_2(p) - H_2(D), \tag{31}$$

which is identical to the rate-distortion function (2). However, here also it is not easy to discuss the existence of an optimal $\hat{k}$. Such an optimal $\hat{k}$ satisfies the following equation:

$$2^{-K}\sum_{l=0}^{K}\binom{K}{l}\Theta\left(\hat{k}^2 - \frac{1}{K}(2l-K)^2\right) = \frac{p-D}{1-2D}. \tag{32}$$

This existence problem is discussed later, when checking the AT stability of the RS solution.

V. ALMEIDA-THOULESS STABILITY OF THE REPLICA SYMMETRIC SOLUTION

In this section we check the AT stability (see [25]) of the RS solution of each scheme. We use the same method as in [11,22]. The main mathematical points of the AT stability study are given in Appendix B.

A. AT stability for the parity tree with nonmonotonic hidden units

In the case of a parity tree with nonmonotonic hidden units, we find

$$P = \frac{8}{\pi}R^{-1}k^2 e^{-k^2}(e^{\beta}-1)^2 \times \left\langle \left(\frac{[1-4H(k)]^{K-1}}{(e^{\beta}+1) + (e^{\beta}-1)y[1-4H(k)]^K} \right)^2 \right\rangle_y,$$

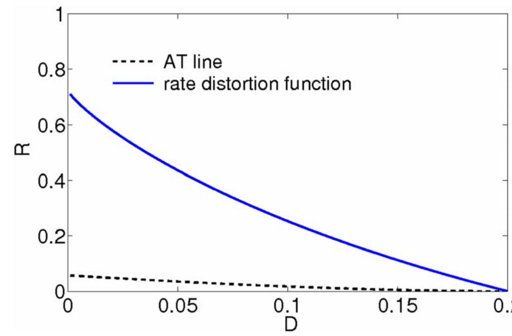

FIG. 3. (Color online) AT line and rate-distortion function for the parity tree with three hidden units. The rate-distortion performance of the parity tree is given by the rate-distortion function. The original message is biased with bias $p=0.2$. The rate-distortion function is always above the AT line and thus the RS solution is always stable.

$$Q = R = P' = Q' = R' = 0. \tag{33}$$

Therefore, using Eq. (B8), the RS stability criterion is given by

$$R > \frac{8}{\pi}K\hat{k}^2 e^{-\hat{k}^2}(e^{\beta}-1)^2 \times \left\langle \left(\frac{[1-4H(\hat{k})]^{K-1}}{(e^{\beta}+1) + (e^{\beta}-1)y[1-4H(\hat{k})]^K} \right)^2 \right\rangle_y, \tag{34}$$

where β is given by (16), and where $\hat{k}$ satisfies Eq. (18). $\langle\cdots\rangle_y$ denotes the expectation with respect to (3).

For $p=\frac{1}{2}$, that is to say for an unbiased message $\boldsymbol{y}$, $\hat{k}$ satisfies the equation $H(\hat{k})=\frac{1}{4}$, which implies $[1-4H(\hat{k})]=0$, and so the AT line is given by the line $R=0$. Consequently, for unbiased messages, the RS solution is always AT stable.

Figure 3 shows the rate-distortion function plotted with the AT stability line for biased messages with $p=0.2$. All the region below the AT line is unstable.

Since no part of the rate-distortion function is under the AT line, the RS solution is always stable. We did the same experiment for higher values of K and never found any unstable part.

The lossy compression scheme using a parity tree with nonmonotonic hidden units presents good properties. It saturates the Shannon bound for any value of $K\geqslant 2$, and the RS solution seems to be always AT stable.

B. AT stability for the committee tree with nonmonotonic hidden units

In the case of a comittee tree with nonmonotonic hidden units, we find

$$P = R^{-1}\left[p\left(\frac{(e^{\beta}-1)(B_k - B_k^*)}{1+(e^{\beta}-1)B_k}\right)^2 + (1-p) \times \left(\frac{(e^{\beta}-1)(B_k^* - B_k)}{1+(e^{\beta}-1)(1-B_k)}\right)^2 \right],$$

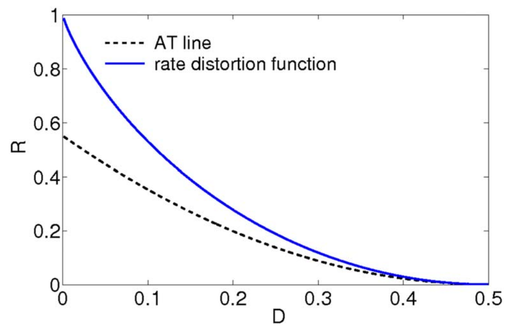

FIG. 4. (Color online) AT line and rate-distortion function for the committee tree with three nonmonotonic hidden units. The rate-distortion performance of the committee tree is given by the rate-distortion function. The original message is unbiased ($p=0.5$). The rate-distortion function is always above the AT line, and thus the RS solution is always stable.

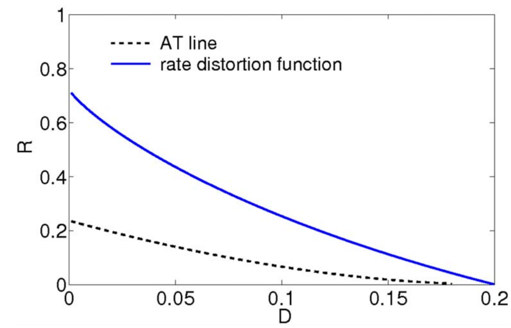

FIG. 5. (Color online) AT line and rate-distortion function for the committee tree with three nonmonotonic hidden units. The rate-distortion performance of the committee tree is given by the rate-distortion function. The original message is biased ($p=0.2$). The rate-distortion function is always above the AT line, and thus the RS solution is always stable.

$$Q = R = P' = Q' = R' = 0, \tag{35}$$

where

$$B_k^* = \sum_{\tau_l=\pm 1}\left[\Theta\left(\sum_{l=1}^{K}\tau_l\right)\left(\frac{1+\tau_1(1-4ke^{-k^2/2}/\sqrt{2\pi}-4H[k])}{2}\right)\right.$$
$$\left.\times\prod_{l=2}^{K}\left(\frac{1+\tau_l(1-4H[k])}{2}\right)\right]. \tag{36}$$

Therefore, using Eq. (B8), the RS stability criterion is given by

$$R > K\left[p\left(\frac{(e^{\beta}-1)(B_{\hat{k}}-B_{\hat{k}}^*)}{1+(e^{\beta}-1)B_{\hat{k}}}\right)^2 + (1-p)\right.$$
$$\left.\times\left(\frac{(e^{\beta}-1)(B_{\hat{k}}^*-B_{\hat{k}})}{1+(e^{\beta}-1)(1-B_{\hat{k}})}\right)^2\right], \tag{37}$$

where β is given by (23), and where $\hat{k}$ satisfies Eq. (25).

However as mentioned in the previous section, it is not clear if there exists a $\hat{k}$ that makes Eq. (25) true. Nevertheless, we did some numerical calculations for $K=3$ and 5 (in this case we consider only odd values of K as mentioned earlier), and always found an optimal k ($\equiv\hat{k}$) in those cases.

We present here the results obtain for $K=3$. Figures 4 and 5 show the rate-distortion function plotted with the AT stability line for unbiased ($p=0.5$) and biased ($p=0.2$) messages. Since no part of the rate-distortion function is under the AT line, the RS solution is always stable. We tried also for higher values of K and no unstable parts were found for the RS solution.

The lossy compression scheme using a committee tree with nonmonotonic hidden units also presents good properties. If an optimal $\hat{k}$ exists (which seems to be always true), it saturates the Shannon bound, and the RS solution seems to be always AT stable.

C. AT stability for the committee tree with a nonmonotonic output unit

In the case of a committee tree with a nonmonotonic output unit, we find

$$P' = \frac{R^{-1}}{\pi^2}(1-e^{\beta})^2 2^{-2(K-2)}\left\langle\left(\frac{\left(\begin{array}{c}K-2\\ \left\lceil\frac{K-y\sqrt{K-2k}}{2}-1\right\rceil\end{array}\right)-\left(\begin{array}{c}K-2\\ \left\lfloor\frac{K+y\sqrt{K-2k}}{2}\right\rfloor\end{array}\right)}{e^{-\beta/2(y-1)}-y(1-e^{\beta})C_k}\right)^2\right\rangle_y,$$

$$P = Q = R = Q' = R' = 0, \tag{38}$$

where $\lceil x\rceil$ denotes the ceiling function ($\lceil x\rceil=\min\{n\in\mathbb{Z}|n\leqslant x\}$) and $\lfloor x\rfloor$ denotes the floor function ($\lfloor x\rfloor=\max\{n\in\mathbb{Z}|n\leqslant x\}$). Therefore, using Eq. (B8), the RS stability criterion is given by

$$R > \frac{K(K-1)}{\pi^2}(1-e^{\beta})^2 2^{-2(K-2)} \left\langle \left(\frac{\binom{K-2}{\left\lceil \frac{K-y\sqrt{K-2}\hat{k}}{2} - 1 \right\rceil} - \binom{K-2}{\left\lfloor \frac{K+y\sqrt{K-2}\hat{k}}{2} \right\rfloor}}{e^{-\beta/2(y-1)} - y(1-e^{\beta})C_{\hat{k}}} \right)^2 \right\rangle_y, \tag{39}$$

where β is given by (30), and where $\hat{k}$ satisfies Eq. (32). $\langle \cdots \rangle_y$ denotes the same expectation as in the parity tree case.

However as mentioned in the previous section, here also it is not clear if there exists $\hat{k}$ such that Eq. (32) is satisfied. On top of that, the function C_k, which depends on k, is not continuous but discrete. C_k is a step function of k. Therefore, we might have no k satisfying Eq. (32). On the other hand, since C_k is a step function of k, if we find a k that satisfies Eq. (32), then it implies that $\hat{k}$ is not given by a unique solution but by an optimal interval where all the elements in this interval satisfy (32). We did some numerical experiments for an unbiased message ($p=0.5$). For the special case of $K=2$, Eq. (32) is clearly satisfied for any $k \in]0,\sqrt{2}[$ so that in this case $\hat{k}$ is given by any element of the interval $]0,\sqrt{2}[$. But for $K>2$ (we checked up to $K=100$), we did not found any optimal k. We did the same thing for a biased message ($p=0.2$) with a fixed distortion $D=0.1$, and for any $K \leqslant 100$ no optimal k exists. This implies that, in the general case, the committee tree with a nonmonotonic output unit does not saturate the Shannon bound. However, if the number of hidden units K becomes very large, we can apply the central limit theorem to replace the scalar product $\boldsymbol{s}_l \cdot \boldsymbol{x}_l$ by a Gaussian variable. Under these conditions, the expression of C_k becomes very simple,

$$C_k = 1 - 2H(k). \tag{40}$$

In this case, C_k is no longer a step function, but a continuous function of k, and it is easy to see that there is always a k that satisfies the equation

$$1 - 2H(k) = \frac{p-D}{1-2D}. \tag{41}$$

Let us denote it by $\hat{k}_{\text{inf}}$. So in the large-K limit $\hat{k}=\hat{k}_{\text{inf}}$ exists and is unique. The compression scheme with a committee tree using a nonmonotonic output unit saturates the Shanon bound in this limit. It is, however, hard to check the AT stability for an infinite number of hidden units K (the binomial coefficient follows a factorial growth), but we claim that the solution obtained by the RS ansatz is always AT stable (except for some very narrow region where $D \approx p$). We show in Figs. 6 and 7 the rate distortion function plotted with the AT line for $K=50$ hidden units for unbiased ($p=0.5$) and biased ($p=0.2$) messages.

To sum up this section, the lossy compression scheme using a committee tree with a nonmonotonic output unit presents a quite complex structure which does not saturate the Shannon bound in most cases. However, it does saturate it when the number of hidden units becomes infinite. Concerning the AT stability, the committee tree with a nonmonotonic output unit does not seem to exhibit any critical instability for the RS solution.

VI. CONCLUSION AND DISCUSSION

We investigated a lossy compression scheme for uniformly biased Boolean messages using nonmonotonic parity

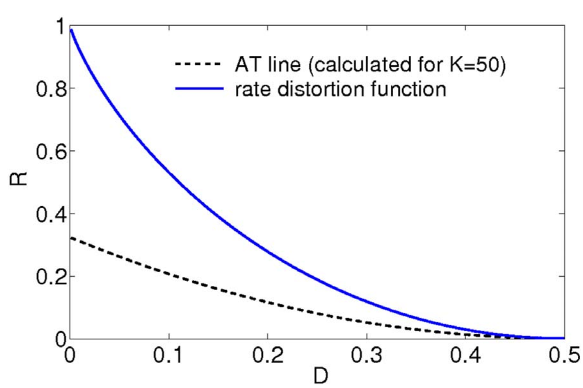

FIG. 6. (Color online) AT line and rate-distortion function for the committee tree with a nonmonotonic output unit. The rate-distortion performance of the committee tree is given by the rate-distortion function. The original message is unbiased ($p=0.5$). The AT line is calculated using $K=50$. The rate-distortion function is always above the AT line, and thus the RS solution is always stable.

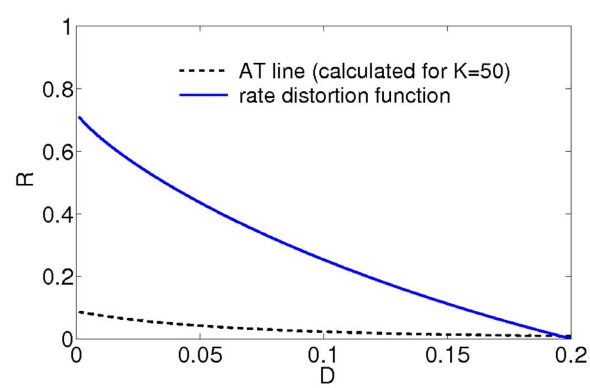

FIG. 7. (Color online) AT line and rate-distortion function for the committee tree with a nonmonotonic output unit. The rate-distortion performance of the committee tree is given by the rate-distortion function. The original message is biased ($p=0.2$). The AT line is calculated using $K=50$. The rate distortion function is always above the AT line except for a very narrow region where $D \approx p$.

tree and nonmonotonic committee tree multilayer perceptrons. All the schemes were shown to saturate the Shannon bound under some specific conditions. The replica symmetric solution is always stable, which tends to confirm the validity of the replica symmetric ansatz.

The Edwards-Anderson order parameter q was always found to be 0, meaning that codewords are uncorrelated in the codeword space. As already mentioned in [9,11], one may conjecture that this is a necessary condition for a lossy compression scheme to achieve the Shannon limit. The mirror symmetry seems then to be an essential feature to saturate the Shannon bound. The committee tree with nonmonotonic hidden units corresponds to the same committee tree model as in the Mimura *et al.* paper [11], with the exception of the hidden layer transfer function, which is given by the nonmonotonic transfer function f_k in this paper. By enforcing mirror symmetry in the hidden layer, we were able to get Shannon optimal performance for an infinite length codeword independently of the number of hidden units, whereas this was not possible using the Mimura *et al.* model, even for an infinite number of hidden units. In the same way, keeping the monotonic sgn function as the transfer function of the hidden layer and transforming only the output unit into a nonmonotonic one by the use of f_k (i.e., the committee tree with a non monotonic output unit), we were able to reach the Shannon limit with an infinite number of hidden units. Once again, enforcing mirror symmetry enabled us to get Shannon optimal performance.

Next, one can easily derive a lower bound for the number of optimal codewords (here "optimal" means a codeword that gives the minimum achievable distortion of the corresponding scheme) for each of the three schemes. In the case of the parity tree and committee tree with nonmonotonic hidden units, there are at least 2^K optimal codewords in the codeword space. Indeed, if $\boldsymbol{s}$ denotes an optimal codeword, we can replace any of its components $\boldsymbol{s}_l$ by $-\boldsymbol{s}_l$ without altering the output of the hidden layer, and thus leave the output of the network unchanged. In the case of the committee tree with a non-monotonic output unit, because of the more complex structure of the hidden layer, we can guarantee only the existence of two optimal codewords, which are given by $\boldsymbol{s}$ and $-\boldsymbol{s}$.

However, a formal encoder for those schemes would require a computational cost which grows exponentially with the original message length to perform its task. We need more efficient algorithms to reduce the encoding time. A preliminary study made by Hosaka *et al.* [20] uses the BP algorithm for this task. This could be a good solution to achieve the encoding phase in a polynomial time. Another possibility is to use the survey propagation algorithm approach which was developed for satisfiability problems [26]. Furthermore, as mentioned above, the parity tree and committee tree with nonmonotonic hidden units have a number of optimal codewords that grows exponentially with the number of hidden units. This could make the search for one optimal codeword easier to achieve using some proper heuristics. This issue will be studied in a future presentation.

ACKNOWLEDGMENTS

This work was partially supported by a Grant-in-Aid for Encouragement of Young Scientists (B) (Grant No. 18700230), Grants-in-Aid for Scientific Research on Priority Areas (Grants No. 18079003 and No. 18020007), a Grant-in-Aid for Scientific Research (C) (Grant No. 16500093), and a JSPS Grant-in-Aid (Grant No. 06J06774) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. T.O. acknowledges support from the Japan Society for the Promotion of Science.

APPENDIX A: ANALYTICAL EVALUATION USING THE REPLICA METHOD

The free energy can be evaluated by the replica method (the parameter k is fixed here),

$$f(\beta,R) = -\frac{1}{\beta N}\lim_{n\to 0}\frac{\langle Z(\beta,\boldsymbol{y},\boldsymbol{x})^n\rangle_{\boldsymbol{y},\boldsymbol{x}} - 1}{n}, \tag{A1}$$

where $Z(\beta,\boldsymbol{y},\boldsymbol{x})^n$ denotes the n-times replicated partition function,

$$Z(\beta,\boldsymbol{y},\boldsymbol{x})^n = \sum_{\boldsymbol{s}^1,\ldots,\boldsymbol{s}^n}\prod_{a=1}^{n}\exp[-\beta\mathcal{H}(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s}^a))]. \tag{A2}$$

The vector $\boldsymbol{s}^a$ is given by $\boldsymbol{s}^a=(\boldsymbol{s}_1^a,\ldots,\boldsymbol{s}_K^a)$ and the superscript a denotes the replica index.

By using the zero-entropy criterion [23], we have

$$0=\beta[U-F],$$

$$u=f, \tag{A3}$$

where U denotes the internal energy and F the free energy. u and f denote the same quantity per bit ($u=U/N$, $f=F/N$). In the zero-entropy limit, only one state of the dynamical variable $\boldsymbol{s}$ achieves a distortion per bit inferior or equal to D. The free energy per bit,

$$f(\beta,R) = -\frac{1}{\beta N}\ln\langle Z(\beta,\boldsymbol{y},\boldsymbol{x})\rangle_{\boldsymbol{y},\boldsymbol{x}}, \tag{A4}$$

is equal to the internal energy per bit,

$$u(\beta)=\frac{\partial\beta f}{\partial\beta}. \tag{A5}$$

This result $f(\beta,R)=u(\beta)$ gives us an explicit relation between the code rate R and the inverse temperature β.

Since this temperature was artificially introduced by means of the parameter β, we should get rid of it by taking the zero-temperature limit ($\beta\to+\infty$) where the dynamical variable freezes. At this limit, one can retrieve the codeword that minimizes the free energy and gives the best achievable code rate. However, since a distortion per bit D is tolerated, at the zero-temperature limit the internal energy per bit should be equal to this distortion. This motivates the introduction of the following identity:

$$\lim_{\beta\to+\infty} u(\beta) = D. \tag{A6}$$

Finally, at this zero-temperature limit, one can get an explicit relation which binds the best achievable code rate R with the distortion D:

$$f(D,R) = D. \tag{A7}$$

We now proceed to the calculation of the replicated partition function (A2). Inserting the identity

$$1 = \prod_{a<b}^{n}\prod_{l=1}^{K}\int_{-\infty}^{+\infty} dq_l^{ab}\,\delta\!\left(\boldsymbol{s}_l^a\cdot\boldsymbol{s}_l^b - \frac{N}{K}q_l^{ab}\right) = \left(\frac{1}{2\pi i}\right)^{n(n-1)K/2}\int\left(\prod_{a<b}\prod_{l} dq_l^{ab}d\hat{q}_l^{ab}\right)\times\exp\left[\sum_{a<b}\sum_{l}\hat{q}_l^{ab}\left(\boldsymbol{s}_l^a\cdot\boldsymbol{s}_l^b - \frac{N}{K}q_l^{ab}\right)\right] \tag{A8}$$

into (A2) enables us to separate the relevant order parameter, and to calculate the average moment $\langle Z(\beta,\boldsymbol{y},\boldsymbol{x})^n\rangle_{\boldsymbol{y},\boldsymbol{x}}$ for natural numbers n as

$$\langle Z(\beta,\boldsymbol{y},\boldsymbol{x})^n\rangle_{\boldsymbol{y},\boldsymbol{x}} \simeq \int\left(\prod_{a<b}\prod_{l} dq_l^{ab}d\hat{q}_l^{ab}\right)\times\exp N\Bigg[R^{-1}\ln\Bigg\langle\int\left(\prod_l d\boldsymbol{u}_l d\boldsymbol{v}_l\right)\prod_l\left(e^{-(1/2)^t\boldsymbol{v}_l\mathcal{Q}_l\boldsymbol{v}_l + i\boldsymbol{v}_l\cdot\boldsymbol{u}_l}\right) \times\prod_a\left[e^{-\beta}+(1-e^{-\beta})\Theta(y,\{u_l^a\})\right]\Bigg\rangle_y + \frac{1}{K}\ln\sum_{\{s_l^a\}}\exp\left(\sum_{a<b}\sum_l\hat{q}_l^{ab}s_l^a s_l^b\right) - \frac{1}{K}\sum_{a<b}\sum_l q_l^{ab}\hat{q}_l^{ab}\Bigg], \tag{A9}$$

where $\mathcal{Q}_l$ is an $n\times n$ matrix having elements $\{q_l^{ab}\}$ and where $\langle\cdots\rangle_y$ denotes the expectation with respect to (3). The function $\Theta(y,\{u_l^a\})$ depends on the decoder and will be discussed in the following sections. We analyze the scheme in the thermodynamic limit $N,M\to+\infty$, while the code rate R is kept finite. In this limit, (A9) can be evaluated using the saddle point method with respect to q_l^{ab} and $\hat{q}_l^{ab}$. To continue the calculation, we have to make some assumptions about the structure of these order parameters. We use here the so-called replica symmetric ansatz

$$q_l^{ab} = (1-q)\delta_{ab} + q,$$

$$\hat{q}_l^{ab} = (1-\hat{q})\delta_{ab} + \hat{q}, \tag{A10}$$

where δ_{ab} denotes the Kronecker delta. This ansatz means that all the hidden units are equivalent after averaging over the disorder.

1. Replica symmetric evaluation for the parity tree with nonmonotonic hidden units

In the case of a parity tree with nonmonotonic hidden units, the function $\Theta(y,\{u_l^a\})$ in (A9) is given by

$$\Theta(y,\{u_l^a\}) = \theta\left(y\prod_{l=1}^{K}\operatorname{sgn}[k^2-(u_l^a)^2]\right). \tag{A11}$$

We can then obtain the expression of the free energy as

$$f(\beta,R,k,q,\hat{q}) = -\frac{1}{\beta}\operatorname*{extr}_{q,\hat{q}}\Bigg[\Bigg\langle\int_{-\infty}^{+\infty}\left(\prod_{l=1}^{K}Dt_l\right)\times\ln[e^{-\beta}+(1-e^{-\beta})\Pi_k(\{t_l\},y)]\Bigg\rangle_y + R\int_{-\infty}^{+\infty}Du\,\ln[2\cosh(\sqrt{\hat{q}}u)] - R\frac{\hat{q}(1-q)}{2}\Bigg], \tag{A12}$$

where extr denotes extremization and where

$$Dx = \frac{e^{-x^2/2}dx}{\sqrt{2\pi}},$$

$$\Pi_k(\{t_l\},y) = \frac{1}{2} + \frac{y}{2}\prod_{l=1}^{K}\left[1 - 2H\left(\frac{k+\sqrt{q}t_l}{\sqrt{1-q}}\right) - 2H\left(\frac{k-\sqrt{q}t_l}{\sqrt{1-q}}\right)\right]. \tag{A13}$$

Taking the derivative of (A12) with respect to q and $\hat{q}$ gives the saddle point equations for the order parameters:

$$\hat{q} = 2R^{-1}\left\langle\int_{-\infty}^{+\infty}\left(\prod_{l=1}^{K}Dt_l\right)\times\frac{-(1-e^{-\beta})\Pi_k'(\{t_l\},y)}{e^{-\beta}+(1-e^{-\beta})\Pi_k(\{t_l\},y)}\right\rangle_y,$$

$$q = \int_{-\infty}^{+\infty}Du\,\tanh^2(\sqrt{\hat{q}}u), \tag{A14}$$

where $\Pi_k'(\{t_l\},y) = \partial\Pi_k(\{t_l\},y)/\partial q$.

We solved this saddle point equation numerically and find that the solution is given for $q=\hat{q}=0$. According to [9,11] this result was expected and implies that all the codewords are uncorrelated and distributed all around $\mathcal{S}^N$. Substituting q

$=\hat{q}=0$ into (A12), one can finally find the free energy given by (12).

2. Replica symmetric evaluation for the committee tree with nonmonotonic hidden units

In the case of a committee tree with nonmonotonic hidden units, the function $\Theta(y,\{u_l^a\})$ in (A9) is given by

$$\Theta(y,\{u_l^a\}) = \theta\left(y\sum_{l=1}^{K} \mathrm{sgn}[k^2-(u_l^a)^2]\right). \tag{A15}$$

We can then obtain the expression of the free energy as

$$f(\beta,R,k,q,\hat{q}) = -\frac{1}{\beta}\underset{q,\hat{q}}{\mathrm{extr}}\left[\left\langle \int_{-\infty}^{+\infty}\left(\prod_{l=1}^{K} Dt_l\right) \times \ln[e^{-\beta}+(1-e^{-\beta})\Sigma_k(\{t_l\},y)]\right\rangle_y + R\int_{-\infty}^{+\infty} Du \ln[2\cosh(\sqrt{\hat{q}}u)] - R\frac{\hat{q}(1-q)}{2}\right], \tag{A16}$$

where

$$\Sigma_k(\{t_l\},y) = \sum_{\tau_l=\pm 1}\left\{\theta\left[y\sum_{l=1}^{K}\tau_l\right]\times\prod_{l=1}^{K}\left[\frac{1+\tau_l}{2} - \tau_l H\left(\frac{k+\sqrt{q}t_l}{\sqrt{1-q}}\right) - \tau_l H\left(\frac{k-\sqrt{q}t_l}{\sqrt{1-q}}\right)\right]\right\}. \tag{A17}$$

Taking the derivative of (A16) with respect to q and $\hat{q}$ gives the saddle point equations for the order parameters,

$$\hat{q} = 2R^{-1}\left\langle \int_{-\infty}^{+\infty}\left(\prod_{l=1}^{K} Dt_l\right) \times \frac{-(1-e^{-\beta})\Sigma_k'(\{t_l\},y)}{e^{-\beta}+(1-e^{-\beta})\Sigma_k(\{t_l\},y)}\right\rangle_y,$$

$$q = \int_{-\infty}^{+\infty} Du \tanh^2(\sqrt{\hat{q}}u), \tag{A18}$$

where $\Sigma_k'(\{t_l\},y)=\partial\Sigma_k(\{t_l\},y)/\partial q$.

We solved this saddle point equation numerically and here also we find that the solution is given for $q=\hat{q}=0$. Substituting $q=\hat{q}=0$ into (A16), one can finally find the free energy given by (19).

3. Replica symmetric evaluation for the committee tree with a nonmonotonic output unit

In the case of a committee tree with a non-monotonic output unit, the function $\Theta(y,\{u_l^a\})$ in (A16) is given by

$$\Theta(y,\{u_l^a\}) = \theta\left\{y\left[k^2 - \frac{1}{K}\left(\sum_{l=1}^{K}\mathrm{sgn}[u_l^a]\right)^2\right]\right\}. \tag{A19}$$

We can then obtain the expression of the free energy as

$$f(\beta,R,k,q,\hat{q}) = -\frac{1}{\beta}\underset{q,\hat{q}}{\mathrm{extr}}\left[\left\langle \int_{-\infty}^{+\infty}\left(\prod_{l=1}^{K} Dt_l\right) \times \ln[e^{-\beta}+(1-e^{-\beta})F_{\Sigma,k}(\{t_l\},y)]\right\rangle_y + R\int_{-\infty}^{+\infty} Du \ln[2\cosh(\sqrt{\hat{q}}u)] - R\frac{\hat{q}(1-q)}{2}\right], \tag{A20}$$

where

$$F_{\Sigma,k}(\{t_l\},y) = \sum_{\tau_l=\pm 1}\left\{\theta\left[yk^2 - \frac{y}{K}\left(\sum_{l}\tau_l\right)^2\right] \times \prod_{l=1}^{K} H\left(-t_l\tau_l\sqrt{\frac{q}{1-q}}\right)\right\}. \tag{A21}$$

Taking the derivative of (A20) with respect to q and $\hat{q}$ gives the saddle point equations for the order parameters,

$$\hat{q} = 2R^{-1}\left\langle \int_{-\infty}^{+\infty}\left(\prod_{l=1}^{K} Dt_l\right) \times \frac{-(1-e^{-\beta})F_{\Sigma,k}'(\{t_l\},y)}{e^{-\beta}+(1-e^{-\beta})F_{\Sigma,k}(\{t_l\},y)}\right\rangle_y,$$

$$q = \int_{-\infty}^{+\infty} Du \tanh^2(\sqrt{\hat{q}}u), \tag{A22}$$

where $F_{\Sigma,k}'(\{t_l\},y)=\partial F_{\Sigma,k}(\{t_l\},y)/\partial q$.

We solved this saddle point equation numerically and here also we find that the solution is given for $q=\hat{q}=0$. Substituting $q=\hat{q}=0$ into (A20), one can finally find the free energy given by (26).

APPENDIX B: ALMEIDA-THOULESS STABILITY CRITERION

The Hessian computed at the RS saddle point characterizes fluctuations in the order parameters q_l^{ab} and $\hat{q}_l^{ab}$ around the RS saddle point. Instability of the RS solution is signaled by a change of sign of at least one of the eigenvalues of the Hessian. Let $\mathcal{M}(\{q_l^{ab}\},\{\hat{q}_l^{ab}\})$ be the exponent of the integrand of integral (A9). Equation (A9) can be represented as

$$\langle Z(\beta,\boldsymbol{y},\boldsymbol{x})^n\rangle_{\boldsymbol{y},\boldsymbol{x}} = \int\left(\prod_{a<b}\prod_{l} dq_l^{ab}d\hat{q}_l^{ab}\right) \times \exp[N\mathcal{M}(\{q_l^{ab}\},\{\hat{q}_l^{ab}\})]. \tag{B1}$$

We expand $\mathcal{M}$ around q and $\hat{q}$ in δq_l^{ab} and $\delta\hat{q}_l^{ab}$ and then find up to the second order

$$\mathcal{M}(\{q+\delta q_l^{ab}\},\{\hat{q}+\delta\hat{q}_l^{ab}\}) = \mathcal{M}(\{q\},\{\hat{q}\}) + \frac{1}{2}{}^t\boldsymbol{v}\boldsymbol{G}\boldsymbol{v} + O(\|\boldsymbol{v}\|^3), \tag{B2}$$

where

$$\boldsymbol{v} = {}^t(\{\delta q_1^{ab}\},\{\delta\hat{q}_1^{ab}\},\ldots,\{\delta q_K^{ab}\},\{\delta\hat{q}_K^{ab}\}) \tag{B3}$$

is the perturbation to the RS saddle point. The Hessian $\boldsymbol{G}$ is the following $[Kn(n-1)]\times[Kn(n-1)]$ matrix:

$$G = \begin{pmatrix} U & V & \cdots & V \\ V & U & \cdots & V \\ \vdots & \vdots & \ddots & \vdots \\ V & V & \cdots & U \end{pmatrix}, \tag{B4}$$

where $n(n-1)\times n(n-1)$ matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are

$$\boldsymbol{U} = \begin{pmatrix} \{\boldsymbol{U}^{ab,cd}\} & \{\tilde{\boldsymbol{U}}^{ab,cd}\} \\ \{\tilde{\boldsymbol{U}}^{ab,cd}\} & \{\tilde{\boldsymbol{U}}^{ab,cd}\} \end{pmatrix}, \tag{B5}$$

$$\boldsymbol{V} = \begin{pmatrix} \{\boldsymbol{V}^{ab,cd}\} & \{\tilde{\boldsymbol{V}}^{ab,cd}\} \\ \{\tilde{\boldsymbol{V}}^{ab,cd}\} & \{\tilde{\boldsymbol{V}}^{ab,cd}\} \end{pmatrix}, \tag{B6}$$

with

$$\boldsymbol{U}^{ab,cd} = \partial^2 \mathcal{M}/\partial q_l^{ab}\partial q_l^{cd},$$

$$\tilde{\boldsymbol{U}}^{ab,cd} = \partial^2 \mathcal{M}/\partial \hat{q}_l^{ab}\partial \hat{q}_l^{cd},$$

$$\tilde{\boldsymbol{U}}^{ab,cd} = \partial^2 \mathcal{M}/\partial q_l^{ab}\partial \hat{q}_l^{cd},$$

$$\boldsymbol{V}^{ab,cd} = \partial^2 \mathcal{M}/\partial q_l^{ab}\partial q_{l'}^{cd} \quad (l \neq l'),$$

$$\tilde{\boldsymbol{V}}^{ab,cd} = \partial^2 \mathcal{M}/\partial \hat{q}_l^{ab}\partial \hat{q}_{l'}^{cd} \quad (l \neq l'),$$

$$\tilde{\boldsymbol{V}}^{ab,cd} = \partial^2 \mathcal{M}/\partial q_l^{ab}\partial \hat{q}_{l'}^{cd} \quad (l \neq l'). \tag{B7}$$

For $q,\hat{q}$ to be a local maximum of $\mathcal{M}$, it is necessary for the Hessian $\boldsymbol{G}$ to be negative definite (i.e., all of its eigenvalues must be negative).

To check these eigenvalues, we use the same method as in [11]. We do not give the mathematical details here. Finally, using Gardner's method [22], we can derive the criterion for the RS solution to be stable as

$$K\gamma < 1, \tag{B8}$$

where

$$\gamma \equiv \gamma_0 + (K-1)\gamma_1,$$

$$\gamma_0 \equiv P - 2Q + R,$$

$$\gamma_1 \equiv P' - 2Q' + R',$$

$$P \equiv \boldsymbol{U}^{ab,ab},$$

$$Q \equiv \boldsymbol{U}^{ab,ac} \quad (b \neq c),$$

$$R \equiv \boldsymbol{U}^{ab,cd} \quad (a \neq c, b \neq d),$$

$$P' \equiv \boldsymbol{V}^{ab,ab},$$

$$Q' \equiv \boldsymbol{V}^{ab,ac} \quad (b \neq c),$$

$$R' \equiv \boldsymbol{V}^{ab,cd} \quad (a \neq c, b \neq d). \tag{B9}$$

The line $K\gamma=1$ defines the so-called AT line.

[1] N. Sourlas, Nature (London) **339**, 693 (1989).
[2] Y. Kabashima, T. Murayama, and D. Saad, Phys. Rev. Lett. **84**, 1355 (2000).
[3] H. Nishimori and K. Y. Michael Wong, Phys. Rev. E **60**, 132 (1999).
[4] A. Montanari and N. Sourlas, Eur. Phys. J. B **18**, 107 (2000).
[5] T. Tanaka, Europhys. Lett. **54**, 540 (2001).
[6] T. Tanaka and M. Okada, IEEE Trans. Inf. Theory **51**, 700 (2005).
[7] T. Murayama and M. Okada, J. Phys. A **36**, 11123 (2003).
[8] T. Murayama, Phys. Rev. E **69**, 035105(R) (2004).
[9] T. Hosaka, Y. Kabashima, and H. Nishimori, Phys. Rev. E **66**, 066126 (2002).
[10] T. Hosaka and Y. Kabashima, J. Phys. Soc. Jpn. **74**, 488 (2005).
[11] K. Mimura and M. Okada, Phys. Rev. E **74**, 026108 (2006).
[12] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).
[13] R. Gallager, IEEE Trans. Inf. Theory **IT-8**, 21 (1968).
[14] D. J. C. MacKay and R. M. Neal, Electron. Lett. **33**, 457 (1997).
[15] T. J. Richardson and R. L. Urbanke, IEEE Trans. Inf. Theory **47**, 638 (2001).
[16] C. E. Shannon, IRE Natl. Conv. Rec. **4**, 142 (1959).
[17] Y. Matsunaga and H. Yamamoto, IEEE Trans. Inf. Theory **49**, 2225 (2003).
[18] E. Martinian and M. J. Wainwright, in Proceedings of the Data Compression Conference, 2006, edited by IEEE Computer Society (unpublished).
[19] M. J. Wainwright, IEEE Signal Process. Mag. **24**, 47 (2007).
[20] T. Hosaka and Y. Kabashima, Physica A **365**, 113 (2006).
[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[22] E. Gardner, J. Phys. A **21**, 257 (1988).
[23] W. Krauth and M. Mézard, J. Phys. (Paris) **50**, 3057 (1989).
[24] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, U.K., 2001).
[25] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).
[26] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).