Statistical mechanics of lossy compression using multilayer perceptrons

Kazushi Mimura[*]
Faculty of Information Sciences, Hiroshima City University, Hiroshima 731-3194, Japan

Masato Okada
Graduate School of Frontier Sciences, University of Tokyo, Chiba 277-5861, Japan;
Brain Science Institute, RIKEN, Saitama 351-0198, Japan;
and PRESTO, Japan Sciences and Technology Agency, Chiba 277-8561, Japan

(Received 25 August 2005; published 8 August 2006)

Statistical mechanics is applied to lossy compression using multilayer perceptrons for unbiased Boolean messages. We utilize a treelike committee machine (committee tree) and treelike parity machine (parity tree) whose transfer functions are monotonic. For compression using a committee tree, a lower bound of achievable distortion becomes small as the number of hidden units K increases. However, it cannot reach the Shannon bound even where $K \to \infty$. For a compression using a parity tree with $K \geq 2$ hidden units, the rate distortion function, which is known as the theoretical limit for compression, is derived where the code length becomes infinity.

DOI: 10.1103/PhysRevE.74.026108 PACS number(s): 89.70.+c, 02.50.−r, 05.50.+q

I. INTRODUCTION

Cross-disciplinary fields that combine information theory with statistical mechanics have developed rapidly in recent years and achievements in these have become the center of attention. The employment of methods derived from statistical mechanics has resulted in significant progress in providing solutions to several problems in information theory, including problems in error correction [1–4], spreading codes [5,6], and compression codes [7–10]. Above all, data compression plays an important role as one of the base technologies in many aspects of information transmission. Data compression is generally classified into lossless compression and lossy compression [12–14]. Lossless compression is aimed at reducing the size of the message under the constraint of perfect retrieval. In lossy compression, on the other hand, the length of the message can be reduced by allowing a certain amount of distortion. The theoretical framework for the lossy compression scheme is called the rate distortion theory, which consists partly of Shannon's information theory [12,13].

Several lossy compression codes, whose schemes saturate the rate distortion function that represents an optimal performance, were discovered in the case where the code length becomes infinity. For instance, the low density generator matrix (LDGM) code [7,8] and using a nonmonotonic perceptron [9–11] were proposed. In these compression codes, a decoder is first defined to retrieve a reproduced message from a codeword. In the encoding problem, for a given source message, we must find a codeword that minimizes the distortion between the reproduced message and the source message. Therefore, fundamentally, the computational cost of compressing a message is of exponential order of a codeword length. It is important to understand properties of various lossy compression codes saturating the optimal performance for the development of more useful codes.

Since a multilayer network includes a nonmonotonic perceptron as a special case, we employ a treelike committee machine and a parity machine as typical multilayer networks [15–17] to lossy compression and analytically evaluate their performance.

II. LOSSY COMPRESSION

Let us start by defining the concepts of the rate distortion theory [14]. Let y be a discrete random variable with source alphabet $\mathcal{Y}$. We will assume that the alphabet is finite. A source message of M random variables, $\boldsymbol{y}={}^t(y^1,\ldots,y^M)\in\mathcal{Y}^M$, is compressed into a shorter expression, where the operator ^{t} denotes the transpose. Here, the encoder describes the source sequence $\boldsymbol{y}\in\mathcal{Y}^M$ by a codeword $\boldsymbol{s}=\mathcal{F}(\boldsymbol{y})\in\mathcal{S}^N$. The decoder represents $\boldsymbol{y}$ by a reproduced message $\hat{\boldsymbol{y}}=\mathcal{G}(\boldsymbol{s})\in\hat{\mathcal{Y}}^M$, as illustrated in Fig. 1. Note that M represents the length of a source sequence, while N represents the length of a codeword. The code rate is defined by $R=N/M$ in this case. A distortion function is a mapping $d:\mathcal{Y}\times\hat{\mathcal{Y}}\to\mathbb{R}^+$ from the set of source alphabet–reproduction alphabet pair into the set of non-negative real numbers. In most cases, the reproduction alphabet $\hat{\mathcal{Y}}$ is the same as the source alphabet $\mathcal{Y}$. After this, we set $\mathcal{Y}=\hat{\mathcal{Y}}$. An example of common distortion functions is the Hamming distortion given by

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

which results in the probability of error distortion, since $E[d(y,\hat{y})]=P[y\neq\hat{y}]$, where E and P represent the

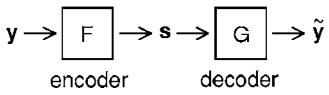

FIG. 1. Rate distortion encoder and decoder.

[*] Electronic address: mimura@cs.hiroshima-cu.ac.jp

1539-3755/2006/74(2)/026108(11)

expectation and the probability of its argument, respectively. The distortion between sequences $\boldsymbol{y},\hat{\boldsymbol{y}} \in \mathcal{Y}^M$ is defined by $d(\boldsymbol{y},\hat{\boldsymbol{y}})=\sum_{\mu=1}^{M} d(y^\mu,\hat{y}^\mu)$. Therefore the distortion associated with the code is defined as $D=E\left[\frac{1}{M}d(\boldsymbol{y},\hat{\boldsymbol{y}})\right]$, where the expectation is with respect to the probability distribution on $\mathcal{Y}$. A rate distortion pair (R,D) is said to be *achievable* if there exists a sequence of rate distortion codes $(\mathcal{F},\mathcal{G})$ with $E\left[\frac{1}{M}d(\boldsymbol{y},\hat{\boldsymbol{y}})\right]\le D$ in the limit $M\to\infty$. We can now define a function to describe the boundary called the *rate distortion function*. The rate distortion function $R(D)$ is the infimum of rates R such that (R,D) is in the rate distortion region of the source for a given distortion D and all rate distortion codes. The infimum of rates R for a given distortion D and given rate distortion codes $(\mathcal{F},\mathcal{G})$ is called the *rate distortion property* of $(\mathcal{F},\mathcal{G})$. We restrict ourselves to a Boolean source $\mathcal{Y}=\{0,1\}$. We assume that the source sequence is not biased to rule out the possibility of compression due to redundancy. The nonbiased Boolean message is one in which each component is generated independently from an identical distribution $P(y^\mu=1)=P(y^\mu=0)=1/2$. For this simple source, the rate distortion function for an unbiased Boolean source with Hamming distortion is given by

$$R(D) = 1 - h_2(D), \tag{2}$$

where $h_2(x)=-x\log_2(x)-(1-x)\log_2(1-x)$ is called the binary entropy function.

III. COMPRESSION USING MULTILAYER PERCEPTRONS

To simplify notations, let us replace all the Boolean representations $\{0,1\}$ with the Ising representation $\{1,-1\}$ throughout the rest of this paper. We set $\mathcal{Y}=\mathcal{S}=\hat{\mathcal{Y}}=\{1,-1\}$ as the binary alphabets. We consider an unbiased source message in which a component is generated independently from an identical distribution:

$$P(y^\mu) = \frac{1}{2}\delta(y^\mu - 1) + \frac{1}{2}\delta(y^\mu + 1), \tag{3}$$

for simplicity. First let us define a decoder. We can construct a nonlinear map $\mathcal{G}:\mathcal{S}^N \to \hat{\mathcal{Y}}^M$ from codeword $\boldsymbol{s}\in\mathcal{S}^N$ to reproduced message $\hat{\boldsymbol{y}}=(\hat{y}^\mu)\in\hat{\mathcal{Y}}^M$. For a given source message $\boldsymbol{y}=(y^\mu)\in\mathcal{Y}^M$, the role of the encoder is to find a codeword $\boldsymbol{s}\in\mathcal{S}^N$ that minimizes the distortion between its reproduced message $\mathcal{G}(\boldsymbol{s})$ and the source message $\boldsymbol{y}$.

We choose a nonlinear map $\mathcal{G}$ utilizing treelike multilayer perceptrons, i.e., a treelike committee machine (committee tree) and a treelike parity machine (parity tree). Figure 2 shows its architecture. The codeword $\boldsymbol{s}$ is divided into N/K-dimensional K disjoint vectors $\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K\in\mathcal{S}^{N/K}$ as $\boldsymbol{s}={}^t(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_K)$. The lth hidden unit receives the vector $\boldsymbol{s}_l$. The outputs of the committee tree and the parity tree are a majority decision and a parity of hidden unit outputs, respectively. The μth bit of the reproduced message $\hat{y}^\mu$ is defined by utilizing the committee tree as

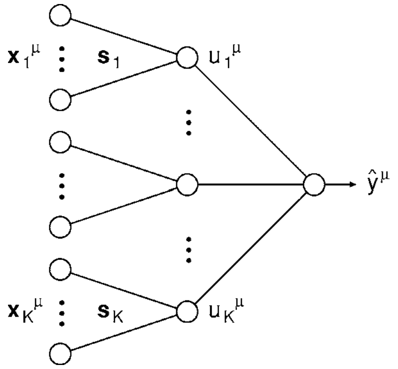

FIG. 2. The architecture of treelike multilayer perceptrons with N input units and K hidden units.

$$\hat{y}^\mu(\boldsymbol{s}) \equiv \mathrm{sgn}\left(\sum_{l=1}^{K} f\left(\sqrt{\frac{K}{N}}\boldsymbol{s}_l\cdot\boldsymbol{x}_l^\mu\right)\right), \tag{4}$$

where $\boldsymbol{x}_l^\mu \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{1})$ are fixed N/K-dimensional vectors and the map $f:\mathbb{R}\to\mathcal{Y}$ is a transfer function. The function $\mathrm{sgn}(x)$ denotes the sign function taking 1 for $x\ge 0$ and -1 for $x<0$. Similarly, the μth bit $\hat{y}^\mu$ of the reproduced message is also defined by utilizing the parity tree as

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

The decoder $\mathcal{G}$ from the codeword $\boldsymbol{s}$ to the reproduced message $\hat{\boldsymbol{y}}=(\hat{y}^\mu)$ is described as

$$\mathcal{G}(\boldsymbol{s}) \equiv \hat{\boldsymbol{y}}(\boldsymbol{s}) = {}^t(\hat{y}^1(\boldsymbol{s}),\ldots,\hat{y}^M(\boldsymbol{s})). \tag{6}$$

In this framework, the encoder $\mathcal{F}$ from the original message $\boldsymbol{y}$ to the codeword $\boldsymbol{s}$ can be written as

$$\mathcal{F}(\boldsymbol{y}) \equiv \underset{\hat{\boldsymbol{s}}}{\mathrm{argmin}}\, d(\boldsymbol{y},\mathcal{G}(\hat{\boldsymbol{s}})), \tag{7}$$

with respect to the case of both the committee tree and the parity tree. Employing the Ising representation, where the length of the codeword is infinite, the average Hamming distortion can be represented as

$$E[d(\boldsymbol{y},\hat{\boldsymbol{y}})] = \sum_{\mu=1}^{M}[1-\Theta(y^\mu\hat{y}^\mu)], \tag{8}$$

where the function $\Theta(x)$ denotes the step function taking 1 for $x\ge 0$ and 0 otherwise. Since we assume the unbiased source message in this paper, we set $f(x)=\mathrm{sgn}(x)$.

This encoding scheme is essentially the same as a learning of the multilayer perceptrons because of a following reason. We first assign the random input vector $\boldsymbol{x}^\mu={}^t(\boldsymbol{x}_1^\mu,\ldots,\boldsymbol{x}_K^\mu)\in\mathbb{R}^N$ to each bit of the original message y^μ. The encoder must find a weight vector $\boldsymbol{s}$ that satisfies input–output relations $\boldsymbol{x}^\mu \mapsto y^\mu$ as much as possible. Then we use this optimal weight vector $\boldsymbol{s}$ as a codeword. Therefore, in a lossless case of $D=0$, an evaluation of the rate distortion property of these codes is entirely identical to the calculation of the storage capacity [18,19].

IV. ANALYTICAL EVALUATION

We analytically evaluate the typical performance, according to Hosaka *et al.* [9], for the proposed compression scheme using the replica method. The minimum permissible average distortion D is calculated, when the code rate R is fixed. For a given original message $\boldsymbol{y}$ and the input vectors $\{\boldsymbol{x}_l^\mu\}$, the number of codewords $\boldsymbol{s}$, which provide a fixed Hamming distortion $MD=d(\boldsymbol{y},\hat{\boldsymbol{y}})$, can be expressed as

$$\mathcal{N}(D,R) = \mathop{\mathrm{Tr}}_{\boldsymbol{s}} \delta(MD; d(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s}))), \tag{9}$$

where $\delta(m;n)$ denotes Kronecker's delta taking 1 if $m=n$ and 0 otherwise. Since the original message $\boldsymbol{y}$ and the input vectors $\{\boldsymbol{x}_l^\mu\}$ are randomly generated predetermined variables, the quenched average of the entropy per bit over these parameters,

$$S(D,R) = \frac{1}{N}\langle \ln \mathcal{N}(D,R)\rangle_{\boldsymbol{y},\boldsymbol{x}}, \tag{10}$$

is naturally introduced for investigating the typical properties, where $\langle\,\rangle_{\boldsymbol{y},\boldsymbol{x}}$ denotes the average over $\boldsymbol{y}$ and $\{\boldsymbol{x}_l^\mu\}$. We calculate the entropy $S(D)$ by the replica method (see Appendix A). The rate-distortion region can be represented by $\{(D,R)\,|\,S(D,R)\geq 0\}$. Therefore a minimum code rate R for a fixed distortion D is given by a solution of $S(D,R)=0$.

Note that a minimum code rate R for $D=0$ coincides with a reciprocal of the critical storage capacity of a multilayer perceptron, i.e., the critical storage capacity $\alpha_c(\equiv M/N)$ can be obtained by $S(0,1/\alpha_c)=0$.

A. Replica symmetric theory of lossy compression using committee tree

1. For general K

In the lossy compression using the committee tree, we obtain average entropy $S_{CT}(D,R)$ as

$$\begin{aligned} S_{CT}(D,R) = \mathop{\mathrm{extr}}_{\beta,q,\hat{q}}\Bigg(& R^{-1}\Bigg\langle \int \left(\prod_{l=1}^{K} Dt_l\right) \\ & \times \ln\{e^{-\beta} + (1-e^{-\beta})\Sigma(\{t_l\};y)\}\Bigg\rangle_y \\ & + \int Du \ln 2\cosh\sqrt{\hat{q}}u - \frac{\hat{q}(1-q)}{2} \\ & + R^{-1}\beta D\Bigg), \end{aligned} \tag{11}$$

where

$$\Sigma(\{t_l\};y) \equiv \mathop{\mathrm{Tr}}_{\{\tau_l=\pm1\}} \Theta\left(-y\sum_{l=1}^{K}\tau_l\right)\prod_{l=1}^{K} H(Q\tau_l t_l), \tag{12}$$

with $Q\equiv\sqrt{q/(1-q)}$ (see Appendix A 1). The operator extr denotes the extremum with respect to the parameters indicated. For any K, we can obtain a minimum code rate R, which gives $S_{CT}(D,R)=0$ for a fixed distortion D.

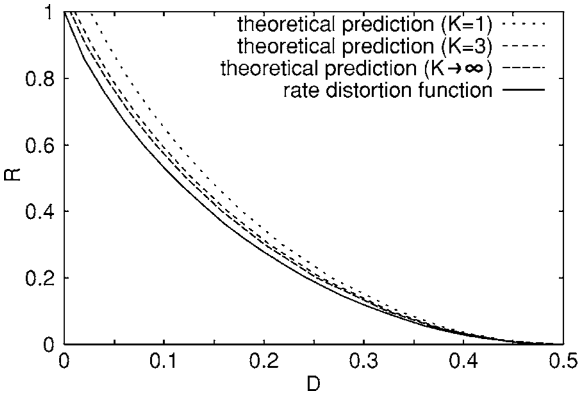

FIG. 3. The rate distortion property of lossy compression using a committee tree. The limit of achievable code rate R expected for $N\to\infty$ plotted versus the distortion D for $K=1$ (dotted line), $K=3$ (short dashed line), and $K\to\infty$ (long dashed line). Solid line denotes rate-distortion function $R(D)$ for binary sequences by Shannon.

2. For large K

We concentrate in the following on the simple case of large K, where the K-multiple integrals can be reduced to a single Gaussian integral. We assume that the number of hidden units K is large but still $K\ll N$. Using the central limit theorem, the averaged entropy is given by

$$\begin{aligned} S_{CT}(D,R) = \mathop{\mathrm{extr}}_{\beta,q,\hat{q}}\Bigg(& R^{-1}\Bigg\langle \int Dt \ln\Bigg\{ e^{-\beta} \\ & + (1-e^{-\beta})H\left(\sqrt{\frac{q_{eff}}{1-q_{eff}}}t\right)\Bigg\}\Bigg\rangle_y \\ & + \int Du \ln 2\cosh\sqrt{\hat{q}}u - \frac{\hat{q}(1-q)}{2} + R^{-1}\beta D\Bigg), \end{aligned} \tag{13}$$

where $q_{eff}\equiv\int Dt[1-2H(Qt)]^2=\frac{2}{\pi}\sin^{-1}q$ and $Q_{eff}\equiv\sqrt{q_{eff}/(1-q_{eff})}$ (see Appendix A 2). Figure 3 shows that the limit of achievable code rate R expected for $N\to\infty$ plotted versus the distortion D for $K=1,3$ and $K\to\infty$. For a fixed code rate R, the achievable distortion decreases as the number of hidden units K increases. However, it does not saturate Shannon's limit even if in the limit $K\to\infty$. For large K, the Edwards-Anderson (EA) order parameter q, which means the average overlap between different codewords, does not converge to zero. Since this means that codewords are correlated, the distribution of codewords is biased in $\mathcal{S}^N$. Note that a nonzero EA order parameter does not mean that the reproduced message has a nonzero average due to the random input vector, which has a zero average.

B. Replica symmetric theory of lossy compression using parity tree

In the lossy compression using the parity tree, on the other hand, we obtain averaged entropy $S_{PT}(D,R)$ as

$$S_{PT}(D,R) = \operatorname*{extr}_{\beta,q,\hat{q}}\Bigg(R^{-1}\Bigg\langle \int \left(\prod_{l=1}^{K} Dt_l\right) \times \ln\{e^{-\beta} + (1-e^{-\beta})\Pi(\{t_l\};y)\}\Bigg\rangle_y + \int Du \ln 2\cosh\sqrt{\hat{q}}u - \frac{\hat{q}(1-q)}{2} + R^{-1}\beta D\Bigg), \tag{14}$$

where

$$\Pi(\{t_l\};y) \equiv \frac{1}{2}\left(1 + y\prod_{l=1}^{K}[1-2H(Qt_l)]\right). \tag{15}$$

For cases utilizing a committee tree and a parity tree, only terms $\Sigma(\{t_l\};y)$ and $\Pi(\{t_l\};y)$ are different. Since both the order parameters q and $\hat{q}$ at the saddle-point of Eq. (14) are less than 1, the average entropy can be expanded with respect to $\Pi_{l=1}^{K}[1-2H(Qt_l)](<1)$. Solutions of the saddle-point equation derived from the expanded form of average entropy are obtained as

$$q = 0,$$

$$\hat{q} = 0, \tag{16}$$

$$D = \frac{e^{-\beta}}{1+e^{-\beta}},$$

in the case $K \geq 2$ (see Appendix A 3). For $K=1$, $q>0$ holds. Note that for $K=1$, a parity tree is equivalent to a committee tree. For $K\geq 2$, the order parameter q becomes zero, namely all codewords are uncorrelated and distributed all around in $\mathcal{S}^N$. Where $K\geq 2$, substituting Eq. (16) into Eq. (14), average entropy is obtained as

$$S_{PT}(D,R) = -R^{-1}\ln 2 + \ln 2 - R^{-1}D\ln D - R^{-1}(1-D)\ln(1-D). \tag{17}$$

A minimum code rate R for a fixed distortion D and $K\geq 2$ is given by $S_{PT}(D,R)=0$. Solving this equation with respect to R, we obtain

$$R = 1 - h_2(D) \equiv R_{RS}(D), \tag{18}$$

which is identical to the rate-distortion function for uniformly unbiased binary sources (2).

However, since the calculation is based on the RS ansatz, we verify the Almeida-Thouless (AT) stability to confirm the validity of this solution. As the RS solution to lossy compression using a parity tree with $K=2$ hidden units can be simply expressed as Eq. (16), the stability condition is analytically obtained as

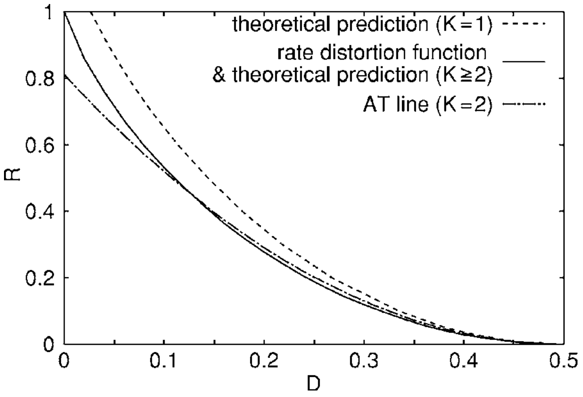

FIG. 4. The rate distortion property of lossy compression using a parity tree. The limits of achievable code rate R expected for $N\to\infty$ is plotted versus the distortion D for $K=1$ (dashed line) and $K\geq 2$ (solid line). The solid line also denotes the rate-distortion function, which is identical to the limit of achievable distortion for $K\geq 2$. The dash-dotted line denotes the AT line for $K=2$. For $K\geq 3$, the RS solution does not exhibit AT instability throughout the achievable region.

$$R > \frac{8}{\pi^2}(1-2D)^2 \equiv R_{AT}(D), \tag{19}$$

where the boundary $R=R_{AT}(D)$ is called the AT line (see Appendix B). For $K\geq 3$, the replica symmetric (RS) solution does not exhibit the AT instability throughout the achievable region of the rate-distortion pair (R,D). Figure 4 shows the limit of achievable distortion D expected for $N\to\infty$ plotted versus code rate R for $K=1$ and $K\geq 2$. In the case $K\geq 2$, the limit of achievable distortion is identical to the rate-distortion function. The dash-dotted line in Fig. 4 denotes the AT line for $K=2$. The region above the AT line denotes that the RS solution is stable. For $K=2$, we found that for the distortion $0.126 \lesssim D \leq 0.5$, $R_{RS}(D)$ can become smaller than $R_{AT}(D)$. Nevertheless, this instability may not be serious in practice, because the region where the RS solution becomes unstable is narrow.

The annealed approximation of the entropy (10) gives a lower bound to the rate distortion property. It coincides with the rate distortion function. According to Opper's discussion [20], the entropy (10) can be represented by the information entropy formally. The annealed information entropy can give an upper bound to the rate distortion property. However, its evaluation is difficult (see Appendix C).

V. DISTRIBUTION OF CODEWORDS

It has already been shown that both compression using a sparse matrix and compression using a nonmonotonic perceptron also achieve optimal performance known as Shannon's limit [7,9]. All these schemes and compression using a parity tree with $K\geq 2$ hidden units becomes the common EA order parameter $q=0$. In compression using a nonmonotonic perceptron, the μth bit of the reproduced message is defined as $\hat{y}^{\mu}(\boldsymbol{s}) \equiv \hat{f}(N^{-1/2}\boldsymbol{s}\cdot\boldsymbol{x}^{\mu})$, where $\hat{f}$ is the transfer function with mirror symmetry, i.e., $\hat{f}(-x)=\hat{f}(x)$ [9]. Due to the mirror

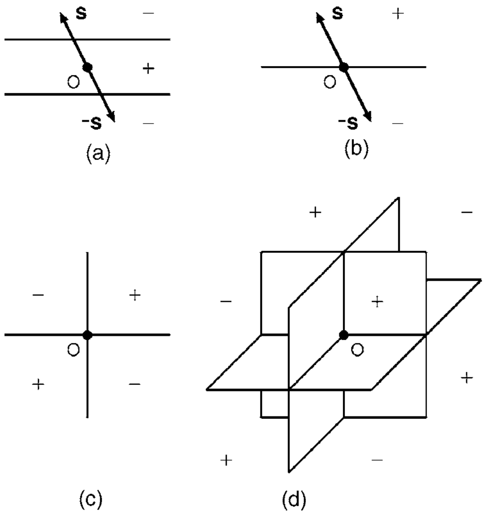

FIG. 5. Relationship between codeword and bit of reproduced message in lossy compression using parity tree with K hidden units. Symbol + denotes bit of the reproduced message is 1 and − denotes that it is −1. Set $\mathcal{S}^N$ is divided by K hyperplanes, whose normal vectors are orthogonal each other. For $K \geq 2$, vectors with same distortion as codeword $\boldsymbol{s}$ are distributed throughout $\mathcal{S}^N$. (a) A nonmonotonic perceptron, $q=0$, (b) a $K=1$ parity tree, $q>0$, (c) a $K=2$ parity tree, $q=0$, and (d) a $K=3$ parity tree, $q=0$.

symmetry of $\hat{f}$, both $\boldsymbol{s}$ and $-\boldsymbol{s}$ provide identical output for any $\boldsymbol{x}^\mu$. Hence the EA order parameter is likely to become zero. The transfer function $\hat{f}$ with parameter κ is defined as taking 1 for $|x| \leq \kappa$ and −1 otherwise. Figure 5 shows the relationship between a codeword and a bit of the reproduced message. Figure 5(a) is the case of compression using a nonmonotonic perceptron.

In compression using a parity tree, on the other hand, the μth bit of the reproduced message is

$$\hat{y}^\mu(-\boldsymbol{s}) = \prod_{l=1}^{K} \operatorname{sgn}\left(\sqrt{\frac{K}{N}}\boldsymbol{x}_l^\mu \cdot (-\boldsymbol{s}_l)\right) = (-1)^K \hat{y}^\mu(\boldsymbol{s}). \quad (20)$$

For $K=1$, i.e., a parity tree is identical to a monotonic perceptron, $\hat{y}^\mu(-\boldsymbol{s}) = -\hat{y}^\mu(\boldsymbol{s})$ holds. Here, the EA order parameter becomes $q>0$. Therefore the distribution of codewords is biased in $\mathcal{S}^N$. Compression using a parity tree with $K=1$ hidden unit cannot achieve Shannon's limit. Figure 5(b) shows the case of compression using a monotonic perceptron, i.e., a committee tree and a $K=1$ parity tree. However, for an even number of hidden units K, a parity tree also has the same effect as mirror symmetry.

We will next discuss the case of $K \geq 2$. Let $\mathcal{V}(\boldsymbol{s}) \subset \mathcal{S}^N$ be a set of vectors that reversed the signs of an arbitrary even number of blocks of a codeword $\boldsymbol{s} = {}^t(\boldsymbol{s}_1, \dots, \boldsymbol{s}_K)$, e.g., ${}^t(-\boldsymbol{s}_1, -\boldsymbol{s}_2, \boldsymbol{s}_3, \dots, \boldsymbol{s}_K) \in \mathcal{V}(\boldsymbol{s})$. The cardinality of the set $\mathcal{V}(\boldsymbol{s})$ is

$$\|\mathcal{V}(\boldsymbol{s})\| = \sum_{n=0}^{\lfloor K/2 \rfloor} {}_K C_{2n} = 2^{K-1}, \quad (21)$$

where $\lfloor x \rfloor$ is the largest integer $\leq x$. According to Eq. (5), all $\hat{\boldsymbol{s}} \in \mathcal{V}(\boldsymbol{s})$ provide identical output for any $\boldsymbol{x}_l^\mu$. The summation of all $\hat{\boldsymbol{s}} \in \mathcal{V}(\boldsymbol{s})$ becomes

$$\sum_{\hat{\boldsymbol{s}} \in \mathcal{V}(\boldsymbol{s})} \hat{\boldsymbol{s}} = {}^t(\dots, 2^{K-2}\boldsymbol{s}_l + 2^{K-2}(-\boldsymbol{s}_l), \dots) = \boldsymbol{0}. \quad (22)$$

This means that 2^{K-1} vectors with the same distortion as codeword $\boldsymbol{s}$ are distributed throughout $\mathcal{S}^N$. For instance, Fig. 5(c) shows the distribution of codewords obtained by compression using a $K=2$ parity tree. The set $\mathcal{S}^N$ is divided by two $N-1$-dimensional hyperplanes whose normal vectors are orthogonal to each other. For the μth bit of the reproduced message, the normal vectors of hyperplanes are ${}^t(\boldsymbol{x}_1^\mu, \boldsymbol{0})$ and ${}^t(\boldsymbol{0}, \boldsymbol{x}_2^\mu) \in \mathbb{R}^N$. Figure 5(d) shows the case of compression using a $K=3$ parity tree. Here, although the same effect as mirror symmetry cannot be seen, nevertheless, EA order parameter q becomes zero for the reason mentioned above. This situation is the same for $K \geq 4$.

With respect to the LDGM code [7], Murayama succeeded in developing a practical encoder using the Thouless-Anderson-Palmer (TAP) approach which introduced inertia term heuristically [8]. The TAP approach is called belief propagation (BP) in the field of information theory. Hosaka *et al.* applied this inertia term introduced BP to compression using a nonmonotonic perceptron [11]. In compression using a parity tree with K hidden units, the number of codewords which give a minimum distortion is 2^{K-1}. Therefore it may become easy to find codewords as the number of hidden units K becomes large. But, in a practical encoding problem, it may not be easy to use a large K since $K \ll N$ is needed.

VI. CONCLUSION

We investigated a lossy compression scheme for unbiased Boolean messages employing a committee tree and a parity tree, whose transfer functions were monotonic. The lower bound for achievable distortion in using a committee tree became small when the number of hidden units K was large. It did not reach Shannon's limit, even in the case where $K \to \infty$. However, lossy compression using a parity tree with $K \geq 2$ hidden units could achieve Shannon's limit where the code length became infinity. We assumed the RS ansatz in our analysis using the replica method. In using a parity tree with $K \geq 2$, the RS solution was unstable in the narrow region. For $K \geq 3$, the RS solution did not exhibit the AT instability throughout the achievable region.

There is generally more than one code with the same distortion as a codeword. The EA order parameter, which means an average overlap between different codewords, need to be zero to reach Shannon's limit like several known schemes which saturate this limit. Therefore it may be a necessary condition that the EA order parameter becomes zero to reach Shannon's limit.

Since the encoding with our method needs exponential time, we need to employ various efficient polynominal-time approximation encoding algorithms. It is under way to investigate the influence of the number of hidden units on the accuracy of approximation encoding algorithms. In future work, we intend to evaluate the upper bound to the rate distortion property without replica.

ACKNOWLEDGMENTS

This work was partially supported by a Grant-in-Aid for Scientific Research on Priority Areas (Grant No. 14084212), and for Scientific Research (C) (Grant No. 16500093), and for Encouragement of Young Scientists (B) (Grant No. 15700141) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

APPENDIX A: ANALYTICAL EVALUATION USING THE REPLICA METHOD

The entropy $S(D,R)$ can be evaluated by the replica method:

$$S(D,R) = \lim_{n\to 0} \frac{1}{nN} \ln\langle \mathcal{N}^n(D,R)\rangle_{\boldsymbol{y},\boldsymbol{x}}. \tag{A1}$$

A moment $\mathcal{N}^n(D,R)$, which is the number of codewords with respect to an n-replicated system, can be represented as

$$\mathcal{N}^n(D,R) = \mathop{\mathrm{Tr}}_{\boldsymbol{s}^1,\ldots,\boldsymbol{s}^n} \prod_{a=1}^{n} \delta(MD; d(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s}^a))), \tag{A2}$$

where $\boldsymbol{s}^a = {}^t(\boldsymbol{s}_1^a,\ldots,\boldsymbol{s}_K^a)$ and the superscript a denotes a replica index. Inserting an identity

$$\begin{aligned} 1 &= \prod_{a<b}\prod_{l=1}^{K} \int_{-\infty}^{\infty} dq_l^{ab}\, \delta\!\left(\boldsymbol{s}_l^a\cdot\boldsymbol{s}_l^b - \frac{N}{K}q_l^{ab}\right) \\ &= \left(\frac{1}{2\pi i}\right)^{n(n-1)K/2} \int \left(\prod_{a<b}\prod_{l} dq_l^{ab} d\hat{q}_l^{ab}\right) \\ &\quad \times \exp\left[\sum_{a<b}\sum_{l} \hat{q}_l^{ab}\left(\boldsymbol{s}_l^a\cdot\boldsymbol{s}_l^b - \frac{N}{K}q_l^{ab}\right)\right], \end{aligned} \tag{A3}$$

into this expression to separate the relevant order parameter. Utilizing the Fourier expression of Kronecker's delta,

$$\delta(MD; d(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s}^a))) = \int_{i(c-\pi)}^{i(c+\pi)} \frac{d\beta^a}{2\pi i} e^{\beta^a(D - d(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s}^a)))}, \quad \forall\ c \in \mathbb{R}, \tag{A4}$$

we can calculate the average moment $\langle\mathcal{N}^n(D,R)\rangle_{\boldsymbol{y},\boldsymbol{x}}$ for natural numbers n as

$$\begin{aligned} \langle\mathcal{N}^n(D,R)\rangle_{\boldsymbol{y},\boldsymbol{x}} \simeq{}& \int\left(\prod_a d\beta^a\right)\int\left(\prod_{a<b}\prod_l dq_l^{ab}d\hat{q}_l^{ab}\right) \\ &\times\exp N\Bigg[R^{-1}\ln\left\langle\int\left(\prod_l d\boldsymbol{u}_l d\boldsymbol{v}_l\right)\prod_l e^{-(1/2)^t\boldsymbol{v}_l Q_l \boldsymbol{v}_l + i\boldsymbol{v}_l\cdot\boldsymbol{u}_l}\prod_a\{e^{-\beta^a} + (1-e^{-\beta^a})\Theta(y,\{u_l^a\})\}\right\rangle_y \\ &+ \frac{1}{K}\ln\mathop{\mathrm{Tr}}_{\{s_l^a\}}\exp\left(\sum_{a<b}\sum_l \hat{q}_l^{ab}s_l^a s_l^b\right) - \frac{1}{K}\sum_{a<b}\sum_l q_l^{ab}\hat{q}_l^{ab} + R^{-1}D\sum_a \beta^a\Bigg], \end{aligned} \tag{A5}$$

where Q_l is an $n\times n$ matrix having matrix elements $\{q_l^{ab}\}$ and

$$\langle h(y)\rangle_y = \sum_{y\in\{-1,1\}} \left[\tfrac{1}{2}\delta(y-1) + \tfrac{1}{2}\delta(y+1)\right]h(y).$$

Function $\Theta(y,\{u_l^a\})$ included in the right hand side of Eq. (A5) depends on the decoder (details are discussed in the following sections). We analyze a system in the thermodynamic limit $N,M\to\infty$, while code rate R is kept finite. This integral (A5) will be dominated by the saddle point of the extensive exponent and can be evaluated via a saddle point problem with respect to β^a, q_l^{ab}, and $\hat{q}_l^{ab}$. Here, we assume the replica symmetric (RS) ansatz

$$\beta_a = \beta,$$

$$q_l^{ab} = (1-q)\,\delta_{ab} + q, \tag{A6}$$

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

where $\delta_{k,k'}$ is Kronecker's delta taking 1 if $k=k'$ and 0 otherwise. This ansatz means that all the hidden units are equivalent after averaging over the disorder.

1. Lossy compression using committee tree for general *K*

In the lossy compression using the committee tree, the $\Theta(y,\{u_l^a\})$ included in Eq. (A5) is obtained as

$$\Theta(y,\{u_l^a\}) = \Theta\left(y\sum_{l=1}^{K}\mathrm{sgn}(u_l^a)\right). \tag{A7}$$

Therefore we obtain average entropy $S_{CT}(D,R)$ as

$$S_{CT}(D,R)=\underset{\beta,q,\hat{q}}{\mathrm{extr}}\left(R^{-1}\left\langle\int\left(\prod_{l=1}^{K}Dt_l\right)\times\ln\{e^{-\beta}+(1-e^{-\beta})\Sigma(\{t_l\};y)\}\right\rangle_y+\int Du\ln 2\cosh\sqrt{\hat{q}}u-\frac{\hat{q}(1-q)}{2}+R^{-1}\beta D\right),\tag{A8}$$

where

$$\Sigma(\{t_l\};y)\equiv\underset{\{\tau_l=\pm1\}}{\mathrm{Tr}}\,\Theta\left(-y\sum_{l=1}^{K}\tau_l\right)\prod_{l=1}^{K}H(Q\tau_l t_l),\tag{A9}$$

with $Q\equiv\sqrt{q/(1-q)}$. Utilizing the Fourier expression of the step function $\Theta(x)=\int_0^\infty d\lambda\int_{-i\infty}^{i\infty}\frac{d\hat{\lambda}}{2\pi i}e^{\hat{\lambda}(\lambda-x)}$, the saddle-point equations $\frac{\partial S}{\partial\beta}=\frac{\partial S}{\partial q}=\frac{\partial S}{\partial\hat{q}}=0$ become

$$q=\int Du\tanh^2\sqrt{\hat{q}}u,\tag{A10}$$

$$\hat{q}=2R^{-1}\left\langle\int\left(\prod_{l=1}^{K}Dt_l\right)\frac{-(1-e^{-\beta})\Sigma'(\{t_l\};y)}{e^{-\beta}+(1-e^{-\beta})\Sigma(\{t_l\};y)}\right\rangle_y,\tag{A11}$$

$$D=\left\langle\int\left(\prod_{l=1}^{K}Dt_l\right)\frac{e^{-\beta}-e^{-\beta}\Sigma(\{t_l\};y)}{e^{-\beta}+(1-e^{-\beta})\Sigma(\{t_l\};y)}\right\rangle_y,\tag{A12}$$

where $\Sigma'(\{t_l\};y)\equiv\partial\Sigma(\{t_l\};y)/\partial q$. Substituting the solutions to the saddle-point equations into Eq. (A8), the average entropy $S_{CT}(D,R)$ is obtained. Thus, for any K, we can obtain a minimum code rate R, which gives $S_{CT}(D,R)=0$ for a fixed distortion D.

2. Lossy compression using committee tree for large *K*

We concentrate in the following on the simple case of large K, where the K-multiple integrals can be reduced to a single Gaussian integral. We assume that the number of hidden units K is large but still $K\ll N$. Here, the term $\Sigma(\{t_l\};y)$ included in Eq. (A8) does not depend on all the individual integration variables t_l but only on the combination $\Sigma_{l=1}^{K}[2H(Qt_l)-1]$. With the central limit theorem, the term is given by

$$\Sigma(\{t_l\};y)=\int_0^\infty d\lambda\int_{-\infty}^{\infty}\frac{d\hat{\lambda}}{2\pi}\exp\left\{i\hat{\lambda}\lambda+i\hat{\lambda}y\frac{1}{\sqrt{K}}\sum_l[2H(Qt_l)-1]-\hat{\lambda}^2\left(1-\frac{1}{K}\sum_l[2H(Qt_l)-1]^2\right)\right\}.\tag{A13}$$

Therefore we obtain the averaged entropy as

$$S_{CT}(D,R)=\underset{\beta,q,\hat{q}}{\mathrm{extr}}\left(R^{-1}\left\langle\int Dt\ln\left\{e^{-\beta}+(1-e^{-\beta})H\left(\sqrt{\frac{q_{eff}}{1-q_{eff}}}t\right)\right\}\right\rangle_y+\int Du\ln 2\cosh\sqrt{\hat{q}}u-\frac{\hat{q}(1-q)}{2}+R^{-1}\beta D\right),\tag{A14}$$

where $q_{eff}\equiv\int Dt[1-2H(Qt)]^2=\frac{2}{\pi}\sin^{-1}q$ and the saddle-point equations are

$$q=\int Du\tanh^2\sqrt{\hat{q}}u,\tag{A15}$$

$$\hat{q}=2R^{-1}\left\langle\int Dt\frac{-(1-e^{-\beta})H'(Q_{eff}t)}{e^{-\beta}+(1-e^{-\beta})H(Q_{eff}t)}\right\rangle_y,\tag{A16}$$

$$D=\left\langle\int Dt\frac{e^{-\beta}-e^{-\beta}H(Q_{eff}t)}{e^{-\beta}+(1-e^{-\beta})H(Q_{eff}t)}\right\rangle_y,\tag{A17}$$

with $Q_{eff}\equiv\sqrt{q_{eff}/(1-q_{eff})}$ and $H'(Q_{eff}t)\equiv\partial H(Q_{eff}t)/\partial q$.

3. Lossy compression using parity tree for general *K*

In the lossy compression using the parity tree, on the other hand, the $\Theta(y,\{u_l^a\})$ included in Eq. (A5) is obtained as

$$\Theta(y,\{u_l^a\})=\Theta\left(y\prod_l\mathrm{sgn}(u_l^a)\right).\tag{A18}$$

Hence we obtain averaged entropy $S_{PT}(D,R)$ as

$$S_{PT}(D,R)=\underset{\beta,q,\hat{q}}{\mathrm{extr}}\left(R^{-1}\left\langle\int\left(\prod_{l=1}^{K}Dt_l\right)\times\ln\{e^{-\beta}+(1-e^{-\beta})\Pi(\{t_l\};y)\}\right\rangle_y+\int Du\ln 2\cosh\sqrt{\hat{q}}u-\frac{\hat{q}(1-q)}{2}+R^{-1}\beta D\right),\tag{A19}$$

where

$$\Pi(\{t_l\};y)\equiv\frac{1}{2}\left(1+y\prod_{l=1}^{K}[1-2H(Qt_l)]\right).\tag{A20}$$

For cases utilizing a committee tree and a parity tree, only terms $\Sigma(\{t_l\};y)$ and $\Pi(\{t_l\};y)$ are different. Since both the order parameters q and $\hat{q}$ at the saddle point of Eq. (A19) are less than 1, the average entropy $S_{PT}(D,R)$ can be expanded with respect to $\Pi_{l=1}^{K}[1-2H(Qt_l)](<1)$ as

$$S_{PT}(D,R)=\underset{\beta,q,\hat{q}}{\mathrm{extr}}\left(R^{-1}\left\{\ln\frac{1+e^{-\beta}}{2}-\sum_{m=1}^{\infty}\frac{1}{2m}\left(\frac{1-e^{-\beta}}{1+e^{-\beta}}\right)^{2m}\times\left[\int Dt[1-2H(Qt)]^{2m}\right]^{K}\right\}+\int Du\ln 2\cosh\sqrt{\hat{q}}u-\frac{\hat{q}(1-q)}{2}+R^{-1}\beta D\right). \tag{A21}$$

We obtain saddle-point equations using this expanded form of the averaged entropy:

$$q=\int Du\tanh^{2}\sqrt{\hat{q}}u, \tag{A22}$$

$$\hat{q}=2R^{-1}K\sum_{m=1}^{\infty}\left(\frac{1-e^{-\beta}}{1+e^{-\beta}}\right)^{2m}\left[\int Dt(1-2H(Qt))^{2m}\right]^{K-1}\times\int Dt(1-2H(Qt))^{2m-1}\frac{te^{-(Qt)^{2}/2}}{\sqrt{2\pi q}(1-q)^{3/2}}, \tag{A23}$$

$$D=\frac{e^{-\beta}}{1+e^{-\beta}}+\sum_{m=1}^{\infty}\frac{2e^{-\beta}}{(1+e^{-\beta})^{2}}\left(\frac{1-e^{-\beta}}{1+e^{-\beta}}\right)^{2m-1}\times\left[\int Dt(1-2H(Qt))^{2m}\right]^{K}. \tag{A24}$$

For $K\geq 2$, because of the existence of term $[\int Dt(1-2H(Qt))^{2m}]^{K-1}$ in Eq. (A23), solutions to the saddle-point equations can become $q=\hat{q}=0$. We can find no other solutions except for $q=\hat{q}=0$ by solving Eqs. (A22)–(A24) numerically for $K\geq 2$. Substituting this into Eq. (A24), we obtain $D=e^{-\beta}/(1+e^{-\beta})$.

APPENDIX B: ALMEIDA-THOULESS INSTABILITY OF REPLICA SYMMETRIC SOLUTION

1. General case

The Hessian computed at the replica symmetric saddle-point characterizes fluctuations in the order parameters β^{a}, 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 $\mathcal{M}(\{\beta^{a}\},\{q_{l}^{ab}\},\{\hat{q}_{l}^{ab}\})$ be the exponent of the integrand of the integral (A5). Equation (A5) can be represented as

$$\langle\mathcal{N}^{n}(D,R)\rangle_{\mathbf{y},\mathbf{x}}=\int\left(\prod_{a}d\beta^{a}\right)\int\left(\prod_{a<b}\prod_{l}dq_{l}^{ab}d\hat{q}_{l}^{ab}\right)\times\exp(N\mathcal{M}(\{\beta^{a}\},\{q_{l}^{ab}\},\{\hat{q}_{l}^{ab}\})). \tag{B1}$$

We expand $\mathcal{M}$ around β, q, and $\hat{q}$ in $\delta\beta^{a}$, δq_{l}^{ab}, and $\delta\hat{q}_{l}^{ab}$ and then find up to second order

$$\mathcal{M}(\{\beta+\delta\beta^{a}\},\{q+q_{l}^{ab}\},\{\hat{q}+\delta\hat{q}_{l}^{ab}\})=\mathcal{M}(\{\beta\},\{q\},\{\hat{q}\})+\frac{1}{2}{}^{t}\boldsymbol{\nu}G\boldsymbol{\nu}+\mathcal{O}(\|\boldsymbol{\nu}\|^{3}), \tag{B2}$$

where

$$\boldsymbol{\nu}={}^{t}(\{\delta\beta^{a}\},\{\delta q_{1}^{ab}\},\{\delta\hat{q}_{1}^{ab}\},\ldots,\{\delta q_{K}^{ab}\},\{\delta\hat{q}_{K}^{ab}\}) \tag{B3}$$

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

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

where $n\times n$ matrix S, $n\times n(n-1)$ matrix T, and $n(n-1)\times n(n-1)$ matrices U,V are

$$S=(\{S^{a,b}\}),$$

$$T=(\{T^{a,bc}\},\{\hat{T}^{a,bc}\}),$$

$$U=\begin{pmatrix}\{U^{ab,cd}\} & \{\tilde{U}^{ab,cd}\}\\ \{\tilde{U}^{ab,cd}\} & \{\hat{U}^{ab,cd}\}\end{pmatrix},$$

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

with

$$S^{a,b}=\partial^{2}\mathcal{M}/\partial\beta^{a}\,\partial\beta^{b},$$

$$T^{a,bc}=\partial^{2}\mathcal{M}/\partial\beta^{a}\,\partial q_{l}^{bc},$$

$$\hat{T}^{a,bc}=\partial^{2}\mathcal{M}/\partial\beta^{a}\,\partial\hat{q}_{l}^{bc},$$

$$U^{ab,cd}=\partial^{2}\mathcal{M}/\partial q_{l}^{ab}\,\partial q_{l}^{cd},$$

$$\hat{U}^{ab,cd}=\partial^{2}\mathcal{M}/\partial\hat{q}_{l}^{ab}\,\partial\hat{q}_{l}^{cd},$$

$$\tilde{U}^{ab,cd}=\partial^{2}\mathcal{M}/\partial q_{l}^{ab}\,\partial\hat{q}_{l}^{cd},$$

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

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

$$\tilde{V}^{ab,cd}=\partial^{2}\mathcal{M}/\partial q_{l}^{ab}\,\partial\hat{q}_{l'}^{cd}\quad(l\neq l'). \tag{B6}$$

For $(\beta,q,\hat{q})$ to be a local maximum of $\mathcal{M}$, it is necessary for the Hessian G to be negative definite, i.e., all of its eigenvalues must be negative. Matrices U and V contain the quadratic fluctuations of the order parameters in the same and different hidden units, respectively. Because of the block form of G, the eigenproblem splits into an uncoupled diagonalization of the two matrices: $U-V$ and

$$\hat{G}=\begin{pmatrix} S & T \\ K^tT & U+(K-1)V \end{pmatrix}. \tag{B7}$$

The eigenvectors of $U-V$ correspond to fluctuations in directions that break the permutation symmetry (PS). The eigenvectors of $\hat{G}$ represent fluctuations that do not break this symmetry. The most unstable mode corresponds to an eigenvector of $\hat{G}$ that breaks the replica symmetry (RS). We can write the eigenvalue equation as

$$\hat{G}\boldsymbol{\mu}=\lambda\boldsymbol{\mu}, \tag{B8}$$

with

$$\boldsymbol{\mu}={}^t(\{\epsilon^a\},\{\eta^{ab}\},\{\hat{\eta}^{ab}\}). \tag{B9}$$

There are three types of eigenvectors, i.e., $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\mu}_3$ [21]. The first $\boldsymbol{\mu}_1$ has the form

$$\epsilon^a=\epsilon, \quad \eta^{ab}=\eta, \quad \hat{\eta}^{ab}=\hat{\eta}. \tag{B10}$$

Using the orthogonality of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, the second type of eigenvector $\boldsymbol{\mu}_2$ has the form

$$\epsilon^a=\begin{cases}(1-n)\epsilon, & (a=\theta),\\ \epsilon, & (\text{otherwise}),\end{cases}$$

$$\eta^{ab}=\begin{cases}\frac{1}{2}(2-n)\eta, & (a=\theta \text{ or } b=\theta),\\ \eta, & (\text{otherwise}),\end{cases}$$

$$\hat{\eta}^{ab}=\begin{cases}\frac{1}{2}(2-n)\hat{\eta}, & (a=\theta \text{ or } b=\theta),\\ \hat{\eta}, & (\text{otherwise}),\end{cases} \tag{B11}$$

for a specific replica θ. In the limit $n\to 0$ this eigenvector $\boldsymbol{\mu}_2$ converges to $\boldsymbol{\mu}_1$ therefore the eigenvalue of the eigenvector $\boldsymbol{\mu}_2$ becomes degenerate with $\boldsymbol{\mu}_1$'s.

Similarly, using the orthogonality of $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$, the third type of eigenvector $\boldsymbol{\mu}_3$ has the form

$$\epsilon^a=0,$$

$$\eta^{ab}=\begin{cases}\frac{1}{2}(2-n)(3-n)\eta & (a=\theta,b=\nu),\\ \frac{1}{2}(3-n)\eta, & (a=\theta \text{ or } a=\nu \text{ or } b=\theta \text{ or } b=\nu),\\ \eta & (\text{otherwise}),\end{cases}$$

$$\hat{\eta}^{ab}=\begin{cases}\frac{1}{2}(2-n)(3-n)\hat{\eta} & (a=\theta,b=\nu),\\ \frac{1}{2}(3-n)\hat{\eta}, & (a=\theta \text{ or } a=\nu \text{ or } b=\theta \text{ or } b=\nu),\\ \hat{\eta} & (\text{otherwise}),\end{cases} \tag{B12}$$

for two specific replicas θ and μ. In the limit $n\to 0$, perturbations keep symmetry of the eigenvectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ across the replicas. Therefore $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are irrelevant to replica symmetry breaking (RSB) but only determine the stability within the RS ansatz. Hence the third eigenvector $\boldsymbol{\mu}_3$, which is called the replicon mode, causes RSB. The eigenvalue equation $\hat{G}\boldsymbol{\mu}_3=\lambda_3\boldsymbol{\mu}_3$ with respect to Eq. (B12) splits into $T\boldsymbol{\mu}_3=\mathbf{0}$ and $[U+(K-1)V]\boldsymbol{\mu}_3'=\lambda_3\boldsymbol{\mu}_3'$, where $\boldsymbol{\mu}_3={}^t(\mathbf{0},\boldsymbol{\mu}_3')$. Therefore the eigenproblem of $\hat{G}$ is equivalent to that of $U+(K-1)V$.

Let us calculate the elements of U and V. The second derivative $\mathcal{M}$ by q_l^{ab} related to the $U^{ab,cd}$,$V^{ab,cd}$ is

$$\frac{\partial^2\mathcal{M}}{\partial q_l^{ab}\,\partial q_{l'}^{cd}}=R^{-1}\langle v_l^a v_l^b v_{l'}^c, v_{l'}^d\rangle_{u,v}-R^{-1}\langle v_l^a v_l^b\rangle_{u,v}\langle v_{l'}^c, v_{l'}^d\rangle_{u,v}, \tag{B13}$$

where

$$\langle g(\{v_l^a\})\rangle_{u,v}=\frac{\left\langle\int\left(\prod_l d\boldsymbol{u}_l d\boldsymbol{v}_l e^{-(1/2)^t\boldsymbol{v}Q_l\boldsymbol{v}_l+i\boldsymbol{v}_l\cdot\boldsymbol{u}_l}\right)g(\{v_l^a\})\prod_a\{e^{-\beta^a}+(1-e^{-\beta^a})\Theta(y,\{u_l^a\})\}\right\rangle_y}{\left\langle\int\left(\prod_l d\boldsymbol{u}_l d\boldsymbol{v}_l e^{-(1/2)^t\boldsymbol{v}Q_l\boldsymbol{v}_l+i\boldsymbol{v}_l\cdot\boldsymbol{u}_l}\right)\prod_a\{e^{-\beta^a}+(1-e^{-\beta^a})\Theta(y,\{u_l^a\})\}\right\rangle_y}, \tag{B14}$$

for any function $g(\{v_l^a\})$. The second derivative $\mathcal{M}$ by $\hat{q}_l^{ab}$ related to the $\hat{U}^{ab,cd}$,$\hat{V}^{ab,cd}$ is

$$\frac{\partial^2 \mathcal{M}}{\partial \hat{q}_l^{ab}\,\partial \hat{q}_{l'}^{cd}} = \begin{cases} K^{-1}\langle s^a s^b s^c s^d\rangle_s \quad - K^{-1}\langle s^a s^b\rangle_s \langle s^c s^d\rangle_s & (l=l'), \\ 0 & (l\neq l'), \end{cases} \tag{B15}$$

where

$$\langle g(\{s^a\})\rangle_s = \frac{\displaystyle\int Dz \operatorname*{Tr}_{\{s^a\}} g(\{s^a\}) \exp\left(\sqrt{\hat{q}} z \sum_a s^a\right)}{\displaystyle\int Dz \operatorname*{Tr}_{\{s^a\}} \exp\left(\sqrt{\hat{q}} z \sum_a s^a\right)} \tag{B16}$$

for any function $g(\{s^a\})$. The second derivative $\mathcal{M}$ by $q_l^{ab},\hat{q}_l^{ab}$ related to the $\tilde{U}^{ab,cd},\tilde{V}^{ab,cd}$ is

$$\frac{\partial^2 \mathcal{M}}{\partial q_l^{ab}\,\partial \hat{q}_{l'}^{cd}} = \begin{cases} K^{-1} & (l=l',\ a=c,\ b=d), \\ 0 & (\text{otherwise}). \end{cases} \tag{B17}$$

Using Gardner's method [18], we find that the RS stability criterion is

$$K\gamma < 1, \tag{B18}$$

where

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

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

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

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

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

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

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

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

$$R' \equiv V^{ab,cd} \quad (a \neq c, b \neq d). \tag{B19}$$

The line $K\gamma=1$ is called the AT line. Setting $K=0$, on the other hand, the matrix $U+(K-1)V$ is equal to $U-V$. When $K=0$, inequality $K\gamma=0<1$ of Eq. (B18) always holds. Therefore permutation symmetry breaking (PSB) does not occur in this system.

2. For lossy compression using a parity tree with $K=2$ hidden units

Let us consider the RS stability of lossy compression using a parity tree with $K=2$ hidden units. Here, $\Theta(y,\{u_l^a\})$ is given by $\Theta(y,\{u_l^a\})=\Theta(y\Pi_l \mathrm{sgn}(u_l^a))$ therefore solutions to the saddle-point equations are

$$q = \hat{q} = 0, \quad D = \frac{e^{-\beta}}{1+e^{-\beta}}. \tag{B20}$$

Substituting Eq. (B20) into Eqs. (B13) and (B15), we obtain

$$P' = R^{-1}\frac{4}{\pi^2}(1-2D)^2,$$

$$P = Q = R = Q' = R' = 0. \tag{B21}$$

Therefore, using Eq. (B18), the RS stability can be obtained as

$$R > \frac{8}{\pi^2}(1-2D)^2 \equiv R_{AT}(D). \tag{B22}$$

This proves Eq. (19).

3. For lossy compression using a parity tree with $K\geq 3$ hidden units

Next, let us consider the RS stability of lossy compression using a parity tree with $K\geq 3$ hidden units. Here, the solutions to the saddle-point equations are $q=\hat{q}=0$, $D=e^{-\beta}/(1+e^{-\beta})$ as well as for $K=2$. Substituting Eq. (B20) into Eqs. (B13) and (B15), we obtain

$$P = Q = R = P' = Q' = R' = 0. \tag{B23}$$

Since the inequality $K\gamma=0<1$ of Eq. (B18) always holds, the RS solution does not exhibit the AT instability throughout the achievable region for $K\geq 3$.

APPENDIX C: A LOWER BOUND TO THE RATE DISTORTION PROPERTY OF LOSSY COMPRESSION USING A PARITY TREE

In order to derive a lower bound to the rate distortion property, an upper bound to the entropy is necessary. Using Jensen's inequality, an upper bound to the entropy $S^{upper}(D,R)$ is given by

$$\begin{aligned} S(D,R) &= \frac{1}{N}\langle \ln \mathcal{N}(D,R)\rangle_{y,x} \leq \frac{1}{N}\ln\langle \mathcal{N}(D,R)\rangle_{y,x} \\ &\equiv S^{upper}(D,R). \end{aligned} \tag{C1}$$

After a simple calculation, we obtain the upper bound to the entropy of lossy compression using a parity tree $S_{PT}^{upper}(D,R)$ as

$$\begin{aligned} S_{PT}^{upper}(D,R) &= \ln 2 + \operatorname*{extr}_{\beta}\left(R^{-1}\ln\frac{1+e^{-\beta}}{2} + \beta R^{-1} D\right) \\ &= -R^{-1}\ln 2 + \ln 2 - R^{-1}D\ln D \\ &\quad - R^{-1}(1-D)\ln(1-D). \end{aligned} \tag{C2}$$

Note that this annealed entropy $S_{PT}^{anneal}(D,R)$ is not depend on the number of hidden units K. Solving $S_{PT}^{anneal}(D,R)=0$ with respect to R, we obtain

$$R = 1 - h_2(D). \tag{C3}$$

This shows that the rate distortion function for uniformly unbiased binary sources (2) can be also derived as a lower bound to the rate distortion property of compression using a parity tree.

We next discuss a upper bound to the rate distortion property. In order to derive a upper bound to the rate distortion property, we need an lower bound to the entropy. Using Jensen's inequality, an upper bound to the entropy $S^{upper}(D,R)$ is represented by

$$\begin{aligned} S(D,R) &= \frac{1}{N}\langle \ln \mathcal{N}(D,R)\rangle_{\boldsymbol{y},\boldsymbol{x}} = \frac{1}{N}\left\langle -\ln\frac{1}{\mathcal{N}(D,R)}\right\rangle_{\boldsymbol{y},\boldsymbol{x}} \\ &\geq -\frac{1}{N}\ln\left\langle \frac{1}{\mathcal{N}(D,R)}\right\rangle_{\boldsymbol{y},\boldsymbol{x}} \equiv S^{lower}(D,R). \end{aligned} \tag{C4}$$

This inequality can be also obtained by an annealed information entropy as follows. According to Opper's discussion [20], we first define a function that characterizes a version space as follows:

$$\rho(\boldsymbol{s}) \equiv \frac{\delta(MD;d(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s})))}{\operatorname*{Tr}_{\boldsymbol{s}}\,\delta(MD;d(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s})))}. \tag{C5}$$

Since this function $\rho(\boldsymbol{s})$ is non-negative and normalized to $\mathrm{Tr}_{\boldsymbol{s}}\rho(\boldsymbol{s})=1$, it defines a probability with respect to $\boldsymbol{s}$. Therefore we obtain the information entropy per bit $\mathcal{H}(D,R)$ as

$$\begin{aligned} \mathcal{H}(D,R) &\equiv \frac{1}{N}\left\langle \operatorname*{Tr}_{\boldsymbol{s}}\rho(\boldsymbol{s})\ln\frac{1}{\rho(\boldsymbol{s})}\right\rangle_{\boldsymbol{y},\boldsymbol{x}} = \frac{1}{N}\left\langle \ln\frac{1}{\rho(\boldsymbol{s})}\right\rangle_{\boldsymbol{s},\boldsymbol{y},\boldsymbol{x}} \\ &\geq -\frac{1}{N}\ln\langle\rho(\boldsymbol{s})\rangle_{\boldsymbol{s},\boldsymbol{y},\boldsymbol{x}} \\ &= -\frac{1}{N}\ln\left\langle \operatorname*{Tr}_{\boldsymbol{s}}\rho(\boldsymbol{s})^2\right\rangle_{\boldsymbol{y},\boldsymbol{x}} = -\frac{1}{N}\ln\left\langle\frac{1}{\mathcal{N}(D,R)}\right\rangle_{\boldsymbol{y},\boldsymbol{x}}, \end{aligned} \tag{C6}$$

where $\langle g(\boldsymbol{s})\rangle_{\boldsymbol{s}}=\mathrm{Tr}_{\boldsymbol{s}}\rho(\boldsymbol{s})g(\boldsymbol{s})$. Using the identity

$$\rho(\boldsymbol{s})\ln\frac{1}{\rho(\boldsymbol{s})} = \begin{cases} 0, & \text{if } \delta(MD;d(\boldsymbol{y},\hat{\boldsymbol{y}}(\boldsymbol{s})))=0, \\ \mathcal{N}(D,R)^{-1}\ln\mathcal{N}(D,R), & \text{otherwise}, \end{cases} \tag{C7}$$

we can easily confirm $\mathcal{H}(D,R)=S(D,R)$.

However, it is difficult to evaluate the lower bound $S^{lower}(D,R)$ directly because $\langle\mathcal{N}(D,R)^{-1}\rangle_{\boldsymbol{y},\boldsymbol{x}}\geq\langle\mathcal{N}(D,R)\rangle_{\boldsymbol{y},\boldsymbol{x}}^{-1}$. This difficulty is caused by a limitation of the version space due to the distortion. This limitation complicates the probability $\rho(\boldsymbol{s})$.

[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] T. Hosaka and Y. Kabashima (unpublished).

[12] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[13] C. E. Shannon, IRE Natl. Conv. Rec. **4**, 142 (1959).

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[15] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).

[16] E. Barkai and I. Kanter, Europhys. Lett. **14**, 107 (1991).

[17] E. Barkai, D. Hansel, and H. Sompolinsky, Phys. Rev. A **45**, 4146 (1992).

[18] E. Gardner, J. Phys. A **21**, 257 (1988).

[19] W. Krauth and M. Mézard, J. Phys. (France) **50**, 3057 (1989).

[20] M. Opper, Phys. Rev. E **51**, 3613 (1995).

[21] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).