Thouless-Anderson-Palmer approach for lossy compression

Tatsuto Murayama

RIKEN Brain Science Institute, Hirosawa 2-1, Wako, Saitama 351-0198, Japan

We study an ill-posed linear inverse problem, where a binary sequence will be reproduced using a sparse matrix. According to the previous study, this model can theoretically provide an optimal compression scheme for an arbitrary distortion level, though the encoding procedure remains an NP-complete problem. In this paper, we focus on the consistency condition for a dynamics model of Markov-type to derive an iterative algorithm, following the steps of Thouless-Anderson-Palmer approach. Numerical results show that the algorithm can empirically saturate the theoretical limit for the sparse construction of our codes, which is also very close to the rate-distortion function.

Lossy compression is quite important in our modern life. One first encodes information into an appropriate form, which will be decoded to reproduce similar sequence. The theoretical framework for this kind of compression scheme with a fidelity criterion is called rate-distortion theory, which consists an important part of the Shannon's information theory [1].

We start by defining the concepts of the rate-distortion theory and stating the simplest version of the main result [2]. Let J be a discrete random variable with alphabet $\mathcal{J}$. Assume that we have a source that produces a sequence $J_1, J_2, \ldots, J_M$, where each symbol is randomly drawn from a distribution. We will assume that the alphabet is finite. Throughout this paper, we use vector notation to represent sequences for convenience of explanation: $\boldsymbol{J} = (J_1, J_2, \ldots, J_M)^T \in \mathcal{J}^M$. Here, the encoder describes the source sequence $\boldsymbol{J} \in \mathcal{J}^M$ by a codeword $\boldsymbol{\xi} = f(\boldsymbol{J}) \in \mathcal{X}^N$. The decoder represents $\boldsymbol{J}$ by an estimate $\hat{\boldsymbol{J}} = g(\boldsymbol{\xi}) \in \hat{\mathcal{J}}^M$. Note that M represents the length of a source sequence, while N represents the length of a codeword. In this case, the rate is defined by $R = N/M$. Note that the relation $N < M$ always holds when a compression is considered; therefore, $R < 1$ also holds.

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

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

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

In this paper, we restrict ourselves to a binary source $\boldsymbol{J}$ with a Hamming distortion measure for simplicity. We assume that binary alphabets are drawn randomly, i.e., the source is not biased to rule out the possibility of compression due to redundancy. We now find the description rate $R(D)$ required to describe the source with an expected proportion of errors less than or equal to D. In this simplified case, according to Shannon, the boundary can be written as follows; the rate-distortion function for a binary source with Hamming distortion is given by

$$R(D) = \begin{cases} 1 - h_2(D), & 0 \leq D \leq \dfrac{1}{2} \\ 0, & \dfrac{1}{2} < D, \end{cases} \tag{2}$$

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

Next we introduce a toy model for the lossy compression. We use the inverse problem of Sourlas-type decoding to realize the optimal encoding scheme, a variation of which has recently been investigated by some information theorists [3]. As in the previous paragraphs, we assume that binary alphabets are drawn randomly from a nonbiased source and that the Hamming distortion measure is selected for the fidelity criterion. Theoretically speaking, it has been reported that the typical distortion could be well captured by the Parisi one-step RSB scheme, giving the physical interpretation of $R(D)$ [4]. In this paper, we will discuss an actual encoding technique for this optimal family of codes.

First, we take the Boolean representation of the binary alphabet $\mathcal{J}$, i.e., we set $\mathcal{J}=\{0,1\}$. We also set $\mathcal{X}=\{0,1\}$ to represent the codewords. Let $\boldsymbol{J}$ be an M-bit source sequence, $\boldsymbol{\xi}$ an N-bit codeword, and $\hat{\boldsymbol{J}}$ an M-bit reproduction sequence. Here, the encoding problem can be written as follows. Given a distortion D and a randomly constructed Boolean matrix A of dimensionality $M \times N$, we find the N-bit codeword sequence $\boldsymbol{\xi}$, which satisfies $\hat{\boldsymbol{J}}=A\boldsymbol{\xi}$ (mod2), where the fidelity criterion $D=E[d(\boldsymbol{J},\hat{\boldsymbol{J}})]$ holds, according to every M-bit source sequence $\boldsymbol{J}$. Note that we applied modulo 2 arithmetics for the additive operations. In our framework, decoding will just be a linear mapping $\hat{\boldsymbol{J}}=A\boldsymbol{\xi}$, while encoding remains an NP-complete problem.

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

The similarity between codes of this type and Ising spin systems was first pointed out by Sourlas, who formulated the mapping of a code onto an Ising spin system Hamiltonian in the context of error-correcting codes [6]. To facilitate the current investigation, we first map the problem to that of an Ising model with finite connectivity following Sourlas' method. We use the Ising representation $\{1,-1\}$ of the alphabets $\mathcal{J}$ and $\mathcal{X}$ rather than the Boolean one $\{0,1\}$; the elements of the source $\boldsymbol{J}$ and the codeword sequences $\boldsymbol{\xi}$ are rewritten in Ising values, and the reproduction sequence $\hat{\boldsymbol{J}}$ is generated by taking products of the relevant binary codeword sequence elements in the Ising representation $\hat{J}_\mu = \Pi_{i \in \mathcal{L}(\mu)}\xi_i$. Here, we denote the set of codeword indices i that participate in the source index μ by $\mathcal{L}(\mu)=\{i|a_{\mu i}=1\}$ with $A=(a_{\mu i})$. Therefore, chosen i's correspond to the ones per row, producing an Ising version of $\hat{\boldsymbol{J}}$. Note that the additive operation in the Boolean representation is translated into multiplication in the Ising one. Hereafter, we set $J_\mu,\hat{J}_\mu,\xi_i=\pm 1$ while we do not change the notations for simplicity. Furthermore, as we use statistical-mechanics techniques, we consider the source and codeword-sequence dimensionality (M and N, respectively) to be infinite, keeping the rate $R=N/M$ finite.

To explore the system's capabilities, we examine the Hamiltonian $H(\boldsymbol{S}|\boldsymbol{J})=\Sigma_{\mu=1}^{M}G[\boldsymbol{S}|J_\mu]$, with $G[\boldsymbol{S}|J_\mu]=-J_\mu\Pi_{i \in \mathcal{L}(\mu)}S_i$, where we have introduced the dynamical variable S_i to find the optimal value of ξ_i, and $G[\boldsymbol{S}|J_\mu]$ denotes the local connectivity of a random hypergraph neighboring the source bit J_μ. It is convenient to represent the posterior probability of codeword $\boldsymbol{S}$, given a source $\boldsymbol{J}$ in the form $\mathcal{P}(\boldsymbol{S}|\boldsymbol{J})=\exp[-\beta H(\boldsymbol{S}|\boldsymbol{J})]/Z(\boldsymbol{J})$ with the inverse temperature β, where $Z(\boldsymbol{J})=\mathrm{Tr}_{S}\exp[-\beta H(\boldsymbol{S}|\boldsymbol{J})]$ is the partition function.

We obtain an expression for the free energy per source bit expressed in terms of the probability distributions $\pi(x)$ and $\hat{\pi}(\hat{x})$, providing a set of the saddle point relations:

$$\pi(x)=\int \left[\prod_{l=1}^{C} \hat{\pi}(\hat{x}_l)d\hat{x}_l\right]\delta\left(x-\sum_{l=1}^{C-1}\hat{x}_l\right), \qquad (3)$$

$$\hat{\pi}(\hat{x})=\int \left[\prod_{l=1}^{K}\pi(x_l)dx_l\right]\left\langle\delta\left[\hat{x}-\frac{1}{\beta}\tanh^{-1}\right.\right.$$
$$\left.\left.\times\left(\tanh(\beta J)\prod_{l=1}^{K-1}\tanh(\beta x_l)\right)\right]\right\rangle_J. \qquad (4)$$

By using the result obtained for the free energy, we can easily perform further straightforward calculations to find all the other observable thermodynamic quantities, including internal energy which records reproduction errors. This set of Eqs. (3) and (4) may be solved numerically for general β, K, and C. The spin glass solution can be calculated for both the replica symmetric and the one-step RSB ansatz. The former reduces to the paramagnetic solution ($f_{\mathrm{RS}}=-1$), which is unphysical for $R<1$, while the latter yields continuous distributions $\pi(x)$ and $\hat{\pi}(\hat{x})$ at the freezing point β_g, which can be obtained from the root of the equation enforcing the non-negative replica symmetric entropy. The random-energy-model limit $K,C\rightarrow\infty$, and simple algebra gives the relation between the rate $R=N/M$ and the distortion D in the form $R=1-h_2(D)$, which coincides with the rate-distortion function in the Shannon's theorem [4].

We now take the Thouless-Anderson-Palmer approach to build a dynamics model using the Markov process assumption on prior beliefs, showing that it is possible to obtain a closed set of equations for practical encoding. At this point, we assume a mean field behavior for the dependence of the dynamical variables $\boldsymbol{S}$ on a certain realization of the source sequence $\boldsymbol{J}$, i.e., the dependence is factorizable and might be replaced by a product of mean fields. Furthermore, we treat the Boltzmann weights for specific codeword bit S_i are factorizable with respect to the source bit J_μ [7]. On the other hand, from a physicist's point of view, it is natural to introduce the Markov process assumption on the priors to find a solution in the spin glass state. This term can be considered as the prior knowledge at a certain time $t+1$, given the previous one on the variables at t. Hereafter, we introduce a parameter $t=1,2,\ldots$ to represent time evolution. In this scenario, we can derive a set of consistency equations

$$\mathcal{W}_t(J_\mu|S_i,\{J_{\mu'\neq\mu}\})=\sum_{S_{i'\neq i}}e^{-\beta G_t(\boldsymbol{S}|J_\mu)}\prod_{i'\neq i}\mathcal{P}_t(S_{i'}|\{J_{\mu'\neq\mu}\}), \qquad (5)$$

$$\mathcal{P}_{t+1}(S_i|\{J_{\mu'\neq\mu}\})$$
$$=a_{\mu i}\mathcal{Q}_{t+1}(S_i)\prod_{\mu'\in\mathcal{M}(i)\backslash\mu}\mathcal{W}_t(J_{\mu'}|S_i,\{J_{\mu''\neq\mu'}\}), \qquad (6)$$

with

$$Q_{t+1}(S) = \langle \exp(\alpha S + \tanh^{-1}\langle\gamma\rangle SS_i)\rangle_{\mathcal{P}_t(S_i|\{J_{\mu'\neq\mu}\})}, \quad (7)$$

where $a_{\mu i}$ is a normalization factor. In Eq. (7), we have two parameters to determine; α denotes the ferromagnetic bias and γ introduces the autocorrelation for sequences.

Here, we introduce another set $\mathcal{M}(i)$ such that it defines the set of source indices linked to the codeword index i. Equation (5) evaluates the average influence of the newly added parity bit J_μ to S_i, when $\{S_{i'}|i' \in \mathcal{L}(\mu)\setminus i\}$ obeys a posterior distribution, which should be determined by the rest of the data set $\{J_{\mu'}|\mu' \in \mathcal{M}(i)\setminus\mu\}$. This calculation corresponds to the cavity method in the conventional framework, representing the effective Boltzmann weight $\mathcal{W}_t(J_\mu|S_i(t),\{J_{\mu'\neq\mu}\})$ produced by J_μ, in which the self-induced contributions are eliminated by assuming the tree description for loopy interactions [8]. On the other hand, Eq. (6) indicates the stack of the cavity fields determining the posterior distribution $\mathcal{P}_t(S_i|\{J_{\mu'\neq\mu}\})$.

In this formula, the approximated marginal posterior will be

$$\mathcal{P}_{t+1}(S_i|J) = a_i Q_{t+1}(S_i) \prod_{\mu\in\mathcal{M}(i)} \mathcal{W}_t(J_\mu|S_i,\{J_{\mu'\neq\mu}\}), \quad (8)$$

taking the full set of the cavity fields, determined self-consistently by Eqs. (5) and (6), into account, where a_i is a normalization factor again.

Next, we present a more convenient form of the above equations. The conditional probability $\mathcal{P}_t(J_\mu|S_i,\{J_{\mu'\neq\mu}\})$ is a normalized effective Boltzmann weight:

$$\mathcal{P}_t(J_\mu|S_i,\{J_{\mu'\neq\mu}\}) = b_{\mu i}\mathcal{W}_t(J_\mu|S_i,\{J_{\mu'\neq\mu}\})$$

$$= b_{\mu i}\sum_{S_{i'\neq i}} e^{-\beta G_t(S|J_\mu)} \prod_{i'\neq i} \mathcal{P}_t(S_i|\{J_{\mu'\neq\mu}\}), \quad (9)$$

where $b_{\mu i}$ is a normalization constant. This relation is obtained by taking the connection μ out of the system, and taking into consideration the dependance of the variables S on all other connections.

The identity

$$e^{-\beta G_t(S|J_\mu)} = \frac{1}{2}\cosh(\beta J_\mu)\left(1 + \tanh(\beta J_\mu)\prod_{i\in\mathcal{L}(\mu)} S_i\right) \quad (10)$$

and simple algebra with respect to the newly defined variables $m_{ij}(t)$, $\hat{m}_{ij}(t) \in [-1,+1]$ satisfying the relations

$$\mathcal{P}_t(S_i|\{J_{\mu'\neq\mu}\}) = \frac{1+m_{\mu i}(t)S_i}{2}, \quad (11)$$

$$\mathcal{P}_t(J_\mu|S_i,\{J_{\mu'\neq\mu}\}) = \frac{1+\hat{m}_{\mu i}(t)S_i}{2} \quad (12)$$

give the set of consistency equations in the form

$$\hat{m}_{\mu i}(t+1) = \tanh(\beta J_\mu)\prod_{i'\in\mathcal{L}(\mu)\setminus i} m_{\mu i'}(t), \quad (13)$$

$$m_{\mu i}(t+1) = \tanh\left(\sum_{\mu'\in\mathcal{M}(i)\setminus\mu} \tanh^{-1}[\hat{m}_{\mu'i}(t)] + \alpha \right.$$
$$\left. + \tanh^{-1}[\gamma m_i(t)]\right), \quad (14)$$

with the pseudoposterior expression

$$m_i(t) = \tanh\left(\sum_{\mu\in\mathcal{M}(i)} \tanh^{-1}[\hat{m}_{\mu i}(t)] + \alpha \right.$$
$$\left. + \tanh^{-1}[\gamma m_i(t)]\right). \quad (15)$$

The set of Eqs. (13) and (14) give an iterative algorithm for code generation. In our dynamics model, the choice of parameter $\gamma=0$ results in naive Thoulers-Anderson-Palmer equation without the reaction term [9]. Therefore, in this case, the dynamics can be strictly captured using the method
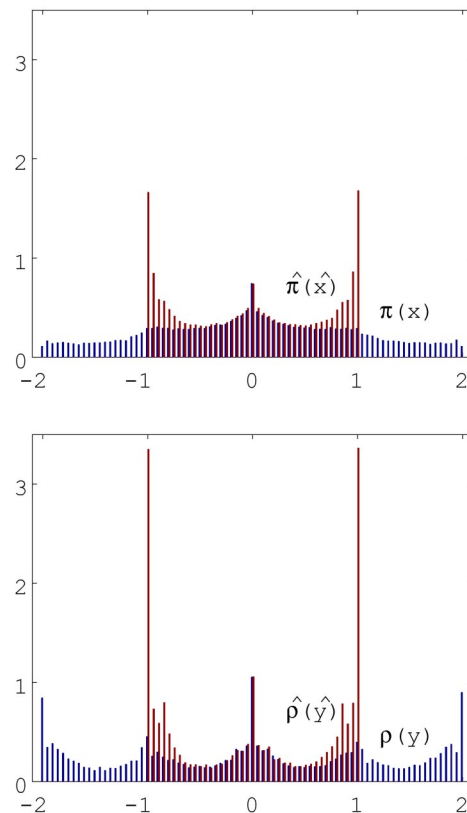

FIG. 1. Snapshots of probability distributions for $K=2$, $L=3$, and $\beta_g=2.35$. (Top) Stable solution of Eqs. (3) and (4) is calculated by Monte Carlo methods. We use 10^5 bin models to approximate the probability distributions $\pi(x)$ and $\hat{\pi}(\hat{x})$, starting from various initial conditions. (Bottom) Fixed-point condition for the density evolution of Eqs. (13) and (14) is represented in terms of the probability distributions in the same bin model. We use the relation $m_{\mu i}=\tanh\beta y$ and $\hat{m}_{\mu i}=\tanh\beta\hat{y}$, where the variables are assumed to be generated from common densities ρ and $\hat{\rho}$, respectively.

of "density evolution" proposed by Richardson and Urbanke in the context of determining the capacity of low-density parity-check codes under message-passing decoding [10]. Let $\rho(\ )$ denote the common density of $(1/\beta)\tanh^{-1} m_{\mu i}$ and $\hat{\rho}(\ )$ the density of $(1/\beta)\tanh^{-1}\hat{m}_{\mu i}$; it is easy to see that the set of probability distributions should satisfy the saddle point equations (3) and (4). It is quite interesting to find the consistency between the information theoretic density evolution technique and the replica theory for disordered statistical systems. Therefore the similarity between π and ρ (or $\hat{\pi}$ and $\hat{\rho}$) can be considered as an important measure for good encoding, giving the design principle for dynamics model (Fig. 1). Finally, Eq. (15) provides the Bayes optimal encoding $\hat{m}_i(t)=\mathrm{sgn}(m_i(t))$ in the Ising representation.

Practical encoding scheme for this compression model will be as follows. Given the source sequence J we first translate the Boolean alphabets into that of Ising ones. Then, for a certain set of control parameters α, β and γ, Eqs. (13) and (14) are recursively calculated until they converge to a certain fixed point. Finally, according to Eq. (15), we calculate the codeword sequence ξ from the Boolean translation of $\hat{m}=\mathrm{sgn}(m)$. Notice that the decoding process will be just a linear mapping. The most interesting quantity to examine is clearly the minimum typical distortion for a given compression rate. Empirical results are shown in Fig. 2, together with the theoretical evaluation for the code constructions using the replica method. We use the optimal inverse temperature β_g for code generation, where per-bit entropy vanishes at the freezing point. Recent works in the information science reveal that designing the codes which approach to the Shannon's limit $R(D)$ is quite difficult in the practical sense; we do not have good coding methods of low complexity, especially of $O(N)$. Our code construction, however, takes only $O(N)$, and the performance is surprisingly good. We believe that the physicist approach can play an important role in the lossy compression schemes, as we have already seen in the context of the error-correcting codes.

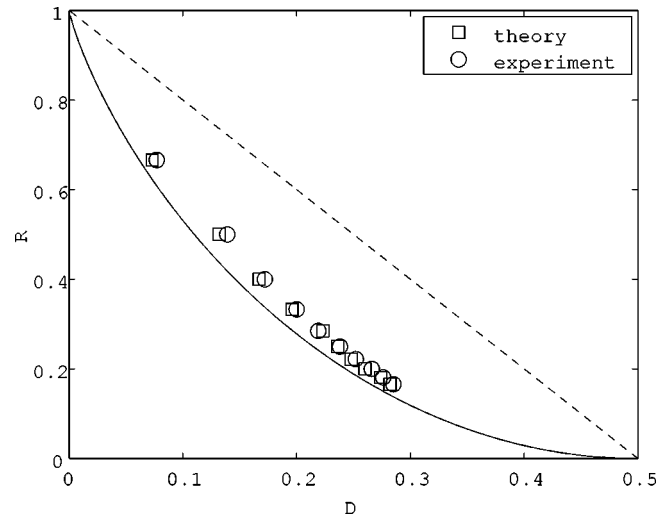Future directions of the current research include utilizing

FIG. 2. Empirical performance: Numerical experiments show that the algorithm with optimal $\alpha=0$ and $\beta=\beta_g$ can achieve the bound for sparse construction of the codes, where $K=2$ and $L=3,4,\ldots,12$. We choose $\gamma=0.01$ for $C=3$ and $\gamma=0.1$ for the rest. Solid line denotes the rate-distortion function $R(D)$ for binary sequences by Shannon, while dashed line can be easily achieved by universal lossless coding techniques. ($\bigcirc$) Numerical results with the system size $N=20\,000$, averaged over ten trials for each evaluation. ($\square$) Theoretical bound for the sparse construction obtained by the 10^5 bin model for Monte Carlo sampling in the replica framework.

more refined approximation techniques to find better coding schemes for lossy compression, as well as the evaluation of the trade-off relation between performance and computational costs. These tasks are interesting and challenging.

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

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

[3] Y. Matsunaga and H. Yamamoto, IEEE Trans. Inf. Theory **49**, 2225 (2003).

[4] T. Murayama and M. Okada, in *Advances in Neural Information Processing Systems 15*, edited by S. Becker, S. Thrun, and K. Obermayer (MIT Press, Cambridge, MA, 2003).

[5] Y. Kabashima and D. Saad, Europhys. Lett. **45**, 97 (1999).

[6] N. Sourlas, Nature (London) **339**, 693 (1989).

[7] Y. Kabashima and D. Saad, Europhys. Lett. **44**, 668 (1998).

[8] M. Mezard, G. Parisi, and M. Virasoro, *Spin-Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[9] M. Opper and D. Saad, *Advanced Mean Field Methods: Theory and Practice* (MIT Press, Cambridge, MA, 2001).

[10] T.J. Richardson and R.L. Urbanke, IEEE Trans. Inf. Theory **47**, 599 (2001).