Statistical mechanical evaluation of a spread-spectrum watermarking model with image restoration

Masaki Kawamura*

Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Yoshida 1677-1, Yamaguchi 753-8512, Japan

Kao Hayashi and Tatsuya Uezu

Graduate School of Humanities and Sciences, Nara Women's University, Kitauoyanishi-machi, Nara 630-8506, Japan

Masato Okada

Graduate School of Frontier Sciences, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa 277-8561, Japan

(Received 29 March 2019; published 26 June 2019)

In cases in which an original image is blind, a decoding method where both the image and the messages can be estimated simultaneously is desirable. We propose a spread spectrum watermarking model with image restoration based on Bayes estimation. We therefore need to assume some prior probabilities. The probability for estimating the messages is given by the uniform distribution, and the ones for the image are given by the infinite-range model and two-dimensional (2D) Ising model. Any attacks from unauthorized users can be represented by channel models. We can obtain the estimated messages and image by maximizing the posterior probability. We analyzed the performance of the proposed method by the replica method in the case of the infinite-range model. We first calculated the theoretical values of the bit error rate from obtained saddle-point equations and then verified them by computer simulations. For this purpose, we assumed that the image is binary and is generated from a given prior probability. We also assume that attacks can be represented by the Gaussian channel. The computer simulation retults agreed with the theoretical values. In the case of prior probability given by the 2D Ising model, in which each pixel is statically connected with four-neighbors, we evaluated the decoding performance by computer simulations, since the replica theory could not be applied. Results using the 2D Ising model showed that the proposed method with image restoration is as effective as the infinite-range model for decoding messages. We compared the performances in a case in which the image was blind and one in which it was informed. The difference between these cases was small as long as the embedding and attack rates were small. This demonstrates that the proposed method with simultaneous estimation is effective as a watermarking decoder.

DOI: 10.1103/PhysRevE.99.062132

I. INTRODUCTION

Digital watermarking is attracting attention for its potential application against the misuse of digital content. The basic idea of digital watermarking is that some hidden messages or watermarks such as a copyright or user ID are invisibly embedded in digital cover content. For image watermarking, we need to pay attention to both the hidden messages and the images themselves. Either watermarks are simply embedded by adding them to the cover content [1,2], or the cover content is transformed by discrete cosine transform [3] or wavelet transform [4] and the watermarks are embedded in the transform domain. For the watermarks themselves, random binary bit or Gaussian sequences are usually used for the embedding [1–3]. The messages may be encoded [5]. The spectrum spreading method is an efficient, robust method. In this paper, we consider a decoding algorithm for the spectrum spreading method.

The basic spectrum spreading technique is also used in code division nultiple access (CDMA) [6], where multiple users can transmit their information at the same time and within the same cell. Multiuser interference needs to be considered for the CDMA multiuser demodulator problem. Recently Bayes optimum solutions have been proposed on statistical mechanics [7–10]. In spread spectrum digital watermarking [1–3], watermarks are generated by spreading the messages. Stego images, which are marked images, are generated by embedding these watermarks in the original images. Attacks to or misuses of the stego images can be represented by channel models. We must estimate the hidden messages from tampered images while reducing multiwatermarks interference.

In an informed case—that is, a case in which the original image is known to the decoder—we can determine the difference between the original and the tampered images. Using a framework of the Bayes estimation [7–9], we can estimate these messages from the received messages by maximizing the posterior probability [11]. In contrast, in the blind case—that is, a case in which the original image is unknown—we need to estimate the original images from the tampered images. Watermarks are treated as noises against the image, and therefore, image estimation need to be applied to such a case. Assuming the prior probability of images, we introduce

*kawamura@sci.yamaguchi-u.ac.jp

(a) (b) (c)

FIG. 1. Sample of natural and parity images, (a) original image, (b) parity of uncompressed image, and (c) parity of JPEG image. The parity images are generated from parity bits.

Bayes image estimation [9,12–15] to the blind watermarking model. In order to estimate original images, we must assume the model used to generate the images. Natural images are usually represented as 8 bits per pixel. Using the least significant bit (LSB) or parity of the natural images, binary images can easily be generated. Embedding the watermarks into the binary images is now common [16]. In this paper, we use binary images.

Performance of the blind digital watermarking model has not yet been sufficiently evaluated. We therefore evaluate the average performance of this model. In particular, in the blind case, we propose a method in which both messages and the original image can be estimated at the same time. In order to evaluate the proposed method, we derive saddle-point equations by the replica method and then calculate the theoretical bit error rate. For the theoretical evaluation using the replica method, we assume the infinite-range model as prior probability of images. Moreover, we evaluate the case of the 2D Ising model as a prior probability by computer simulations.

Now we discuss the feasibility of representing original images by the infinite-range model and 2D Ising model. Watermarking methods such as the wet paper code [16] and matrix embedding [17] methods assume that content consists of binary data. Specifically, the original images to be embedded are generated by calculating LSB or parity bits. We refer to a binary image consisting of parity bits as a parity image. Figure 1 shows the parity images generated from a natural image, where Fig. 1(a) is the original natural image and Fig. 1(b) shows the parity image from the uncompressed natural image of Fig. 1(a). The parity image in Fig. 1(c) is generated after JPEG compression of Fig. 1(a). The black and white pixels represent the parity bits 0 and 1, respectively. Figuratively speaking, from these images, we can find that part of the parity images [Figs. 1(b) and 1(c)] can be seen as an image generated from the infinite-range model and other part with some clusters can be seen as one from the 2D Ising model. Since we can evaluate our method in theory, it is reasonable to introduce some image generation models.

The rest of this paper is organized as follows. Section II gives an overview of our watermarking model. We explain that both messages and images can be estimated by maximizing the posterior probability. Section III describes the saddle-point equations derived by the replica method in order to evaluate our method. Section IV shows the results obtained by theory and computer simulations. We conclude the paper in Sec. V.

II. DIGITAL WATERMARKING MODEL

We describe a basic watermarking model in an informed case and an image restoration model before proposing our blind watermarking model.

A. Informed case

When a decoder has been informed of an original image, the informed spread spectrum watermarking model can correspond to the CDMA model. K-bit messages $\boldsymbol{s} = (s_1, s_2, \ldots, s_K)^\top$ are embedded in an original image in layers, where $s_i = \pm 1$. We assume the prior probability of messages is a uniform distribution given by

$$P(\boldsymbol{s}) = \frac{1}{2^K}. \tag{1}$$

Each message s_i is spread by a specific spreading code $\boldsymbol{\xi}_i = (\xi_i^1, \xi_i^2, \ldots, \xi_i^N)^\top$, and watermarks are obtained by summing the K spread messages. The length of the spread codes—that is, the chip rate—is equal to the size of the image, N. Each element of spreading codes ξ_i^μ takes ± 1 with probability

$$P\left(\xi_i^\mu = \pm 1\right) = \frac{1}{2}. \tag{2}$$

Here $(\xi_i^\mu)^2 = 1$. The μth watermark w_μ is represented by

$$w_\mu = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i^\mu s_i \,, \quad \mu = 1, 2, \ldots, N. \tag{3}$$

The stego image or marked image $\boldsymbol{X}$ is created by adding the watermark $\boldsymbol{w}$ to the original image $\boldsymbol{f}$; that is, $X_\mu = f_\mu + w_\mu$. We ignore any embedding errors, because they are almost always small enough to be negligent.

Here assume we have received a tampered stego image that is attacked by an illegal user. We can consider this attack the deterioration process of an image. Attacks can be represented as noise in the communication channel [5,18,19]. We assume

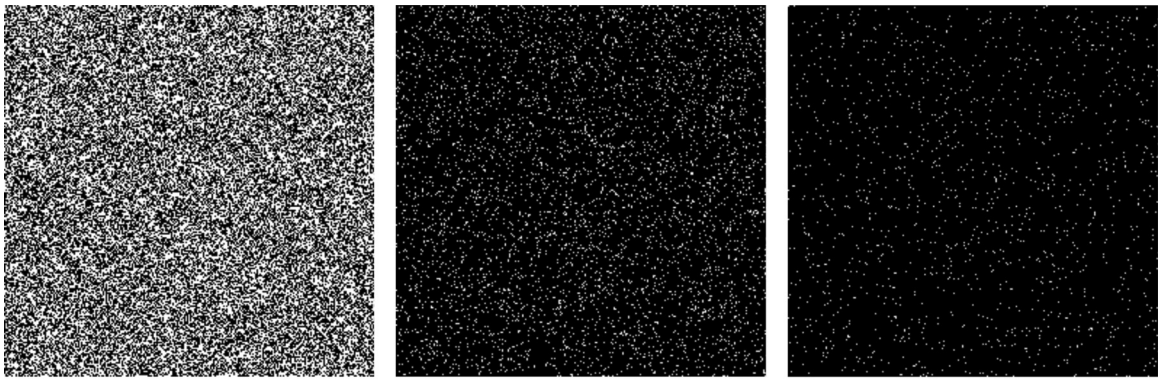

FIG. 2. Images generated by infinite-range model (256 × 256 pixels) with smooth parameters (a) $\alpha_0 = 1.0$, (b) $\alpha_0 = 1.5$, and (c) $\alpha_0 = 2.0$.

the channel is represented by the additive white Gaussian noise channel. Therefore, the conditional probability of the tampered image $\boldsymbol{r}$ given messages $\boldsymbol{s}$ is given by

$$P(\boldsymbol{r}|\boldsymbol{s}) = \prod_{\mu=1}^{N} P(r_\mu|\boldsymbol{s}) \propto \exp\left[-\frac{1}{2\sigma_0^2}\sum_{\mu=1}^{N}(r_\mu - w_\mu)^2\right], \quad (4)$$

where noise obeys the Gaussian distribution $\mathcal{N}(0, \sigma_0^2)$.

What we want to know is how many messages the decoder can retrieve from the tampered image. We therefore need to estimate messages $\boldsymbol{s}$ and then calculate the bit error rate. In order to estimate the messages, the posterior probability of messages $\boldsymbol{s}$ given the tampered image $\boldsymbol{r}$ should be computed. Since the true parameter σ_0^2 is unknown, we set a parameter as σ^2. From (1) and Bayes theorem, the posterior probability is given by

$$P(\boldsymbol{s}|\boldsymbol{r}) = \frac{P(\boldsymbol{r}|\boldsymbol{s})P(\boldsymbol{s})}{\sum_{\boldsymbol{s}} P(\boldsymbol{r}|\boldsymbol{s})P(\boldsymbol{s})} \quad (5)$$

$$= \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\sum_{\mu=1}^{N}(r_\mu - w_\mu)^2\right], \quad (6)$$

$$Z = \operatorname*{Tr}_{\boldsymbol{s}} \exp\left[-\frac{1}{2\sigma^2}\sum_{\mu=1}^{N}(r_\mu - w_\mu)^2\right], \quad (7)$$

where Z is a normalization factor called a partition function. The watermark w_μ is a function of the messages $\boldsymbol{s}$. $\operatorname*{Tr}_{s}$ stands for the summation over s.

For a maximum *a posteriori* (MAP) estimation, the estimated messages $\widehat{\boldsymbol{s}}$ are given by

$$\widehat{\boldsymbol{s}} = \arg\max_{\boldsymbol{x}} P(\boldsymbol{x}|\boldsymbol{r}), \quad (8)$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)^\top$ are variables that represent messages. For a maximum posterior marginal (MPM) estimation, the estimated messages $\widehat{\boldsymbol{s}}$ are given by

$$\widehat{s}_i = \arg\max_{x_i} \sum_{\boldsymbol{x}\setminus x_i} P(\boldsymbol{x}|\boldsymbol{r}), \quad (9)$$

where summation $\sum_{\boldsymbol{x}\setminus x_i}$ is a summation over $\boldsymbol{x}$ excepting x_i. With that, we can obtain a Bayes optimum estimation.

B. Image restoration model

It is difficult to formulate natural images. In the image restoration method based on Bayes estimation, the original images are assumed to be generated from some probability distribution [12,13,15]. In this paper, we assume that the original images consist of N pixels and that the pixels are binary [12,13,15]. Moreover, we consider the infinite-range model [9] and the 2D Ising model as image generating models. The prior probability of the infinite-range model is given by

$$P(\boldsymbol{f}) \propto \exp\left[\frac{\alpha_0}{N}\sum_{\mu<\nu} f_\mu f_\nu\right], \quad (10)$$

where parameter α_0 represents the smoothness of an image and the summation $\sum_{\mu<\nu}$ runs over all pairs of different indexes μ, ν. Figure 2 shows some images generated by the infinite range model with $\alpha_0 = 1.0, 1.5$, and 2.0. Although sites in the infinite-range model are not intrinsically lined up, we arrange these sites on the two-dimensional lattice like 256×256 pixel images. In the case of (a), the image looks like high-frequency snow noise, while with the larger α_0 in (c), smooth images appear.

For a while, leaving the watermarking scheme aside, we concentrate exclusively on image restoration from a tampered image. In fact, the embedding process of the watermarks can be considered a Gaussian channel. Therefore, we assume the deterioration process from the original image to the tampered image is a Gaussian channel. In this case, the probability of the tampered image $\boldsymbol{r}$ given the original image $\boldsymbol{f}$ is given by

$$P(\boldsymbol{r}|\boldsymbol{f}) = \prod_{\mu=1}^{N} P(r_\mu|\boldsymbol{f}) \propto \exp\left[-\frac{1}{2\sigma_0^2}\sum_{\mu=1}^{N}(r_\mu - f_\mu)^2\right]. \quad (11)$$

For the infinite-range model, from Bayes theorem, the image maximizing the posterior probability,

$$P(\boldsymbol{f}|\boldsymbol{r}) = \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\sum_{\mu=1}^{N}(r_\mu - f_\mu)^2 + \frac{\alpha}{N}\sum_{\mu<\nu} f_\mu f_\nu\right], \quad (12)$$

$$Z = \operatorname*{Tr}_{\boldsymbol{f}} P(\boldsymbol{r}|\boldsymbol{f})P(\boldsymbol{f}), \quad (13)$$

can be chosen as the estimation image. Since the true parameters σ_0^2, α_0 are unknown, parameters σ^2 and α are used.

C. Blind case

When the original image is unknown or blind at the decoder, both the messages and the image should be estimated at the same time. This method requires the posterior probability of messages $\boldsymbol{s}$ and image $\boldsymbol{f}$ given the tampered image $\boldsymbol{r}$. Since the probability of the tampered image $\boldsymbol{r}$ is given by

$$P(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{f}) \propto \exp\left[-\frac{1}{2\sigma_0^2}\sum_{\mu=1}^{N}(r_\mu - w_\mu - f_\mu)^2\right], \quad (14)$$

and the prior probabilities are given by (1) and (10), the posterior probability can be given by

$$P(\boldsymbol{s},\boldsymbol{f}|\boldsymbol{r}) = \frac{P(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{f})P(\boldsymbol{s})P(\boldsymbol{f})}{\sum_{\boldsymbol{s},\boldsymbol{f}} P(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{f})P(\boldsymbol{s})P(\boldsymbol{f})} \quad (15)$$

$$= \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\sum_{\mu=1}^{N}(r_\mu - w_\mu - f_\mu)^2 + \frac{\alpha}{N}\sum_{\mu<\nu} f_\mu f_\nu\right], \quad (16)$$

where

$$Z = \operatorname*{Tr}_{\boldsymbol{s},\boldsymbol{f}} P(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{f})P(\boldsymbol{f}). \quad (17)$$

Constant $P(\boldsymbol{s})$ is reducible. Since the true parameters σ_0^2 and α_0 are unknown, parameters σ^2 and α are used. Now, we rewrite the posterior probability in a different form using the Hamiltonian $H(\boldsymbol{s},\boldsymbol{f})$ as $P(\boldsymbol{s},\boldsymbol{f}|\boldsymbol{r}) = \exp[-H(\boldsymbol{s},\boldsymbol{f})/\sigma^2]/Z$. We can then obtain the Hamiltonian,

$$H(\boldsymbol{s},\boldsymbol{f}) = \frac{1}{2\beta}\sum_{i=1}^{K}\sum_{j=1}^{K} J_{ij}s_i s_j - \frac{1}{\sqrt{\beta}}\sum_{i=1}^{K} h_i s_i - \sum_{\mu=1}^{N} r_\mu f_\mu + \frac{1}{\sqrt{K}}\sum_{\mu=1}^{N}\sum_{i=1}^{K} f_\mu \xi_i^\mu s_i - \frac{\alpha\sigma^2}{N}\sum_{\mu<\nu} f_\mu f_\nu, \quad (18)$$

where β stands for embedding rate $\beta = K/N$ and

$$J_{ij} = \frac{1}{N}\sum_{\mu=1}^{N}\xi_i^\mu \xi_j^\mu, \quad h_i = \frac{1}{\sqrt{N}}\sum_{\mu=1}^{N}\xi_i^\mu r_\mu. \quad (19)$$

From MAP and MPM estimations, the estimated messages $\widehat{\boldsymbol{s}}$ and estimated image $\widehat{\boldsymbol{f}}$ are given by

$$\text{MAP}: (\widehat{\boldsymbol{s}}, \widehat{f}) = \arg\max_{(\boldsymbol{x},\boldsymbol{g})} P(\boldsymbol{x},\boldsymbol{g}|\boldsymbol{r}), \quad (20)$$

$$\text{MPM}: \widehat{S}_i = \arg\max_{X_i}\sum_{\boldsymbol{X}\setminus X_i} P(\boldsymbol{X}|\boldsymbol{r}), \quad (21)$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)^\top$ and $\boldsymbol{g} = (g_1, g_2, \ldots, g_N)^\top$ stand for variables of messages and image, respectively. For MPM estimation, $\boldsymbol{S} = (s_1, \ldots, s_K, f_1, \ldots, f_N)^\top$ stand for the true values of original messages and image. $\widehat{S}_i$ represents each element of the estimated messages $\widehat{\boldsymbol{s}}$ and image $\widehat{\boldsymbol{f}}$, that is, $\widehat{S}_i \in \{\widehat{s}_1, \ldots, \widehat{s}_K, \widehat{f}_1, \ldots, \widehat{f}_N\}$. $\boldsymbol{X} = (x_1, x_2, \ldots, x_K, g_1, g_2, \ldots, g_N)^\top$ represents the corresponding variables of messages and image, and X_i is ith element in $\boldsymbol{X}$ corresponding to $\widehat{S}_i$. The MPM estimation for the CDMA model can be seen in Ref. [9].

III. THEORETICAL EVALUATION

A. Bit error rate

The accuracy for estimated messages can be measured by bit error rate (BER), as

$$\text{BER}_m = \frac{1 - d_m}{2}, \quad (22)$$

where d_m represents the overlap between the original message s_i and the estimated message $\widehat{s}_i$ and is defined as

$$d_m = \frac{1}{K}\sum_{i=1}^{K} s_i \widehat{s}_i. \quad (23)$$

Image quality is usually measured by peak signal-to-noise ratio (PSNR). However, since we deal with binary images, the image quality can also be measured by BER as

$$\text{BER}_R = \frac{1 - d_R}{2}, \quad (24)$$

where the overlap, d_R, between the original image f_μ and the estimated image $\widehat{f}_\mu$ is defined as

$$d_R = \frac{1}{N}\sum_{\mu=1}^{N} f_\mu \widehat{f}_\mu. \quad (25)$$

Because mean squared error $\text{MSE} = 4\,\text{BER}_R$, PSNR can be calculated from BER_R.

Using the BER, we evaluate the performance of our proposed method, which estimates both messages and image at the same time. We want to know the average performance rather than specific messages and image. Therefore, we average the BER over all possible messages $\boldsymbol{s}$, images $\boldsymbol{f}$, and spread codes ξ_i^μ. We assume random diffusion by spread codes and a large system limit. Under this assumptions, we can derive saddle-point equations of overlaps m and R from the posterior probability and can then theoretically evaluate the performance.

B. Replica method

In order to determine the average performance, Helmholtz free energy F is averaged over messages, the pixel value of images, and spread codes. That is, $[F] = -T[\log Z]$, where $[\cdot]$ denotes a configurational average defined by

$$[x] = \int\prod_{\mu=1}^{N} dr_\mu \operatorname*{Tr}_{\boldsymbol{s},\boldsymbol{f}} \langle P(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{f})P(\boldsymbol{s})P(\boldsymbol{f})x\rangle_\xi, \quad (26)$$

and $\langle\cdot\rangle_\xi$ denotes an average over ξ_i^μ. By using the replica method, we can obtain this averaged free energy $[F]$ from the relation

$$[\log Z] = \lim_{n\to 0}\frac{[Z^n] - 1}{n}. \quad (27)$$

In other words, $[\log Z]$ can be calculated from n replicas of the original system using the configurational average of the product of the partition functions, Z^n. We therefore start to

calculate from

$$[Z^n] = \int \prod_{\mu=1}^{N} dr_\mu \langle \operatorname*{Tr}_{\boldsymbol{s},\boldsymbol{f}} P(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{f}) P(\boldsymbol{s}) P(\boldsymbol{f}) Z^n \rangle_\xi \tag{28}$$

$$= \int \prod_{\mu=1}^{N} dr_\mu \operatorname*{Tr}_{\boldsymbol{s},\boldsymbol{x}^a} \operatorname*{Tr}_{\boldsymbol{f},\boldsymbol{g}^a} \left\langle \left(2\pi\sigma_0^2\right)^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma_0^2} \sum_{\mu=1}^{N} \left(r_\mu - \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i^\mu s_i - f_\mu \right)^2 \right.\right.$$
$$\left.\left. -\frac{1}{2\sigma^2} \sum_{a=1}^{n} \sum_{\mu=1}^{N} \left(r_\mu - \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i^\mu x_i^a - g_\mu^a \right)^2 + \frac{\alpha_0}{N} \sum_{\mu<\nu} f_\mu f_\nu + \frac{\alpha}{N} \sum_{a=1}^{n} \sum_{\mu<\nu} g_\mu^a g_\nu^a \right] \right\rangle_\xi , \tag{29}$$

where a is the replica index.

According to the replica analysis of the CDMA model [7,9], we first need to carry out the terms of the messages. Let us average over the spread codes ξ_i^μ. By introducing the following notations to (29):

$$v_0^\mu = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i^\mu s_i \ , \quad v_a^\mu = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i^\mu x_i^a , \tag{30}$$

we obtain

$$[Z^n] = \int dv_0^\mu \prod_a dv_a^\mu e^{N(g_1+g_2)}, \tag{31}$$

$$e^{Ng_1} = \operatorname*{Tr}_{\boldsymbol{s},\boldsymbol{x}} \prod_\mu \left\langle \delta\left(v_0^\mu - \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i^\mu s_i \right) \prod_a \delta\left(v_a^\mu - \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \xi_i^\mu x_i^a \right) \right\rangle_\xi , \tag{32}$$

$$e^{Ng_2} = \operatorname*{Tr}_{\boldsymbol{f},\boldsymbol{g}} \prod_\mu \int \frac{dr_\mu}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2\sigma_0^2} \left(r_\mu - v_0^\mu - f_\mu \right)^2 - \frac{1}{2\sigma^2} \sum_a \left(r_\mu - v_a^\mu - g_\mu^a \right)^2 + \frac{\alpha_0}{N} \sum_{\mu<\nu} f_\mu f_\nu + \frac{\alpha}{N} \sum_{a=1}^{n} \sum_{\mu<\nu} g_\mu^a g_\nu^a \right]. \tag{33}$$

In the term e^{Ng_1}, using the integral representation of delta function $\delta(\cdot)$, we can carry out the average over the spread codes ξ_i^μ and then introduce order parameters to the terms of the messages s_i, x_i^a, given by

$$q_{ab} = \frac{1}{K} \sum_{i=1}^{K} x_i^a x_i^b , \quad m_a = \frac{1}{K} \sum_{i=1}^{K} s_i x_i^a . \tag{34}$$

The term e^{Ng_1} can be represented as

$$e^{Ng_1} = \operatorname*{Tr}_{\boldsymbol{s},\boldsymbol{x}} \left\{ \prod_{a<b} \int dq_{ab} \delta\left(K q_{ab} - \sum_{i=1}^{K} x_i^a x_i^b \right) \prod_a \int dm_a \delta\left(K m_a - \sum_{i=1}^{K} s_i x_i^a \right) \right\}$$
$$\times \prod_\mu \int \frac{d\widehat{v}_0^\mu}{2\pi} \prod_a \frac{d\widehat{v}_a^\mu}{2\pi} \exp\left[i\widehat{v}_0^\mu v_0^\mu + i \sum_a \widehat{v}_a^\mu v_a^\mu - \frac{1}{2} \left(\widehat{v}_0^\mu \right)^2 - \frac{1}{2} \sum_a \left(\widehat{v}_a^\mu \right)^2 - \sum_{a<b} q_{ab} \widehat{v}_a^\mu \widehat{v}_b^\mu - \sum_a m_a \widehat{v}_0^\mu \widehat{v}_a^\mu \right] \tag{35}$$

$$= \int \prod_{a<b} \frac{i dq_{ab} d\widehat{q}_{ab}}{2\pi} \prod_a \frac{i dm_a d\widehat{m}_a}{2\pi} \exp\left(-K \sum_{a<b} \widehat{q}_{ab} q_{ab} - K \sum_a \widehat{m}_a m_a \right)$$
$$\times \prod_\mu \int \frac{d\widehat{v}_0^\mu}{2\pi} \prod_a \frac{d\widehat{v}_a^\mu}{2\pi} \exp\left[i\widehat{v}_0^\mu v_0^\mu + i \sum_a \widehat{v}_a^\mu v_a^\mu - \frac{1}{2} \sum_a \left(\widehat{v}_a^\mu \right)^2 - \sum_{a<b} q_{ab} \widehat{v}_a^\mu \widehat{v}_b^\mu - \sum_a m_a \widehat{v}_0^\mu \widehat{v}_a^\mu - \frac{1}{2} \left(\widehat{v}_0^\mu \right)^2 \right]$$
$$\times \operatorname*{Tr}_{\boldsymbol{s},\boldsymbol{x}} \prod_{k=1}^{K} \exp\left(\sum_{a<b} \widehat{q}_{ab} x_k^a x_k^b + \sum_a \widehat{m}_a s_k x_k^a \right). \tag{36}$$

From integrating $[Z^n]$ into the terms of r_μ, v_0^μ, and $\widehat{v}_0^\mu$, we can obtain the term (A5). (See Appendix A for more details on this derivation.) Now, we assume symmetry between replicas for the order parameters of messages; that is, $q_{ab} = q$, $\widehat{q}_{ab} = \widehat{q}$,

$m_a = m$, and $\widehat{m}_a = \widehat{m}$. Under this assumption, we obtain

$$[Z^n] = \int \frac{id\widehat{q}dq}{2\pi}\frac{id\widehat{m}dm}{2\pi} e^{N(G_1+G_2+G_3)}, \tag{37}$$

$$G_1 = -\frac{1}{2}n(n-1)\beta\widehat{q}q - n\beta\widehat{m}m, \tag{38}$$

$$G_2 = -\frac{n\beta\widehat{q}}{2} + n\beta\int D_z \log 2\cosh(z\sqrt{\widehat{q}} + \widehat{m}), \tag{39}$$

$$e^{NG_3} = \operatorname*{Tr}_{f,g}\prod_\mu \int \frac{dv_0^\mu d\widehat{v}_0^\mu}{2\pi}\prod_a \frac{dv_a^\mu d\widehat{v}_a^\mu}{2\pi}\frac{dr_\mu}{\sqrt{2\pi}\sigma_0}\exp\left[i\widehat{v}_0^\mu v_0^\mu + i\sum_a \widehat{v}_a^\mu v_a^\mu - \frac{1}{2}\sum_a\left(\widehat{v}_a^\mu\right)^2 - \frac{1}{2}\left(\widehat{v}_0^\mu\right)^2 - q\sum_{a<b}\widehat{v}_a\widehat{v}_b\right.$$
$$\left. - m\sum_a \widehat{v}_0\widehat{v}_a - \frac{1}{2\sigma_0^2}\left(r_\mu - v_0^\mu - f_\mu\right)^2 - \frac{1}{2\sigma^2}\sum_a\left(r_\mu - v_a^\mu - g_\mu^a\right)^2 + \frac{\alpha_0}{N}\sum_{\mu<\nu} f_\mu f_\nu + \frac{\alpha}{N}\sum_{a=1}^n\sum_{\mu<\nu} g_\mu^a g_\nu^a\right], \tag{40}$$

where $D_z = dz/\sqrt{2\pi}e^{-z^2/2}$. By integrating over v_a^μ, $\widehat{v}_a^\mu$, the term e^{NG_3} is given by

$$e^{NG_3} = \operatorname*{Tr}_{f,g}\prod_\mu \sigma^n(\sigma^2+1-q)^{-\frac{n}{2}}\left[1 + \frac{n\left(2m - q - \sigma_0^2 - 1\right)}{2(\sigma^2+1-q)} + \frac{n\left(\sigma_0^2+1\right)}{2\sigma^2} + \Upsilon\left(\sum_a g_\mu^a\right)^2\right]$$
$$\times \exp\left[\Phi + \Psi\sum_{a<b} g_\mu^a g_\mu^b + \Omega f_\mu\sum_a g_\mu^a + \frac{\alpha_0}{2N}\left(\sum_{\mu=1}^N f_\mu\right)^2 + \frac{\alpha}{2N}\sum_a\left(\sum_{\mu=1}^N g_\mu^a\right)^2\right]. \tag{41}$$

(See Appendix B for more details on this derivation.) This term represents contribution from the image.

Next, for term e^{NG_3}, we introduce various order parameters of the images, given by

$$r_0 = \frac{1}{N}\sum_{\mu=1}^N f_\mu, \quad r_a = \frac{1}{N}\sum_{\mu=1}^N g_\mu^a, \tag{42}$$

$$R_a = \frac{1}{N}\sum_{\mu=1}^N f_\mu g_\mu^a, \quad Q_{ab} = \frac{1}{N}\sum_{\mu=1}^N g_\mu^a g_\mu^b. \tag{43}$$

Using these order parameters, we can rewrite it as

$$e^{NG_3} = \int dr_0\prod_a dr_a\prod_a dR_a\prod_{a<b} dQ_{ab} e^{N(G_4+G_5+G_6+G_7)}, \tag{44}$$

where

$$e^{NG_4} = \operatorname*{TrTr}_{f\ g}\exp\left[-\widehat{r}_0\left(Nr_0 - \sum_{\mu=1}^N f_\mu\right) - \sum_a \widehat{r}_a\left(Nr_a - \sum_{\mu=1}^N g_\mu^a\right)\right.$$
$$\left. - \sum_a \widehat{R}_a\left(NR_a - \sum_{\mu=1}^N f_\mu g_\mu^a\right) - \sum_{a<b}\widehat{Q}_{ab}\left(NQ_{ab} - \sum_{\mu=1}^N g_\mu^a g_\mu^b\right)\right], \tag{45}$$

$$e^{NG_5} = [\sigma^n(\sigma^2+1-q)^{-\frac{n}{2}}]^N, \tag{46}$$

$$e^{NG_6} = \exp N\left[\frac{n\left(2m - q - \sigma_0^2 - 1\right)}{2(\sigma^2+1-q)} + \frac{n\left(\sigma_0^2+1\right)}{2\sigma^2} + \Upsilon\left(2\sum_{a<b} Q_{ab} + n\right)\right], \tag{47}$$

$$e^{NG_7} = \exp N\left[\Phi + \Psi\sum_{a<b} Q_{ab} + \Omega\sum_a R_a + \frac{\alpha_0}{2}r_0^2 + \frac{\alpha}{2}\sum_a r_a^2\right]. \tag{48}$$

The variables Υ, Φ, Ψ, and Ω are given by (B5)–(B8). We assume the replica symmetry for these order parameters; that is, $r_a = r$, $\widehat{r}_a = \widehat{r}$, $R_a = R$, $\widehat{R}_a = \widehat{R}$, $Q_{ab} = Q$, and $\widehat{Q}_{ab} = \widehat{Q}$. They lead to

$$e^{NG_3} = \int \frac{id\widehat{r}_0 dr_0}{2\pi}\int\frac{id\widehat{r}dr}{2\pi}\int\frac{id\widehat{R}dR}{2\pi}\int\frac{id\widehat{Q}dQ}{2\pi} e^{N(G_4+G_5+G_6+G_7)}, \tag{49}$$

where

$$G_4 = -r_0\widehat{r}_0 - nr\widehat{r} - nR\widehat{R} - \frac{n(n-1)}{2}Q\widehat{Q} - \frac{n}{2}\widehat{Q} + \log\left[2\cosh(\widehat{r}_0) + n\operatorname*{Tr}_f \exp(\widehat{r}_0 f)\int D_s \log 2\cosh\left(s\sqrt{\widehat{Q}} + \widehat{r} + \widehat{R}f\right)\right], \tag{50}$$

$$G_5 = n\log\sigma - \frac{n}{2}\frac{\sigma_0^2+1}{\sigma^2} - \frac{n}{2}\log(\sigma^2+1-q), \tag{51}$$

$$G_6 = \frac{n\left(2m-q-\sigma_0^2-1\right)}{2(\sigma^2+1-q)} + \frac{n\left(\sigma_0^2+1\right)}{2\sigma^2} + n\left[-\frac{2m-q-\sigma_0^2-1}{2(\sigma^2+1-q)^2} + \frac{\left(\sigma_0^2+1\right)\left(1-\frac{nm}{\sigma^2}\right)}{\sigma^2(\sigma^2+1-q)} + \frac{\sigma_0^2+1}{2\sigma^4}\right](1-Q+nQ), \tag{52}$$

$$G_7 = \Phi + \frac{n(n-1)}{2}\Psi Q + n\Omega R + \frac{\alpha_0}{2}r_0^2 + \frac{n\alpha}{2}r^2. \tag{53}$$

Thus, we obtain

$$[Z^n] = \int \frac{id\widehat{q}dq}{2\pi}\frac{id\widehat{m}dm}{2\pi}\frac{id\widehat{r}_0 dr_0}{2\pi}\frac{id\widehat{r}dr}{2\pi}\frac{id\widehat{R}dR}{2\pi}\frac{id\widehat{Q}dQ}{2\pi}e^{N(G_1+G_2+G_4+G_5+G_6+G_7)}. \tag{54}$$

In the large-system limit $N\to\infty$, the integral can be evaluated by the saddle-point method. From (27), the free energy F is given in the limit $n\to 0$ as

$$F = \frac{1}{2}\beta\widehat{q}q - \beta\widehat{m}m - \frac{\beta\widehat{q}}{2} + \beta\int D_z \log 2\cosh(z\sqrt{\widehat{q}}+\widehat{m}) + \log\sigma - \frac{1}{2}\log(\sigma^2+1-q) + \frac{2m-q-\sigma_0^2-1}{2(\sigma^2+1-q)} + \frac{\alpha}{2}r^2 - r\widehat{r} - R\widehat{R} - \frac{1}{2}(1-Q)\widehat{Q} - (1-Q)\frac{2m-q-\sigma_0^2-1}{2(\sigma^2+1-q)^2} - \frac{1-R}{\sigma^2+1-q} + \frac{\operatorname*{Tr}_f \exp(\widehat{r}_0 f)\int D_s \log 2\cosh(s\sqrt{\widehat{Q}}+\widehat{r}+\widehat{R}f)}{2\cosh(\widehat{r}_0)}. \tag{55}$$

Since there are n-independent constant terms in F, we define them as

$$F_0 = -r_0\widehat{r}_0 + \frac{\alpha_0}{2}r_0^2 + \log 2\cosh(\widehat{r}_0). \tag{56}$$

Extremization of the free energy yields the saddle-point equations as

$$m = \int D_z \tanh(z\sqrt{\widehat{q}}+\widehat{m}), \tag{57}$$

$$\widehat{m} = \frac{1}{\beta(\sigma^2+1-q)} - \frac{1-Q}{\beta(\sigma^2+1-q)^2}, \tag{58}$$

$$q = \int D_z \tanh^2(z\sqrt{\widehat{q}}+\widehat{m}), \tag{59}$$

$$\widehat{q} = \frac{q-2m+\sigma_0^2+2(1-R)+Q}{\beta(\sigma^2+1-q)^2} - 2(1-Q)\frac{q-2m+\sigma_0^2+1}{\beta(\sigma^2+1-q)^3}, \tag{60}$$

$$r = \frac{1}{2\cosh(\widehat{r}_0)}\operatorname*{Tr}_f e^{\widehat{r}_0 f}\int D_z \tanh(\alpha r + z\sqrt{\widehat{Q}} + \widehat{R}f), \tag{61}$$

$$R = \frac{1}{2\cosh(\widehat{r}_0)}\operatorname*{Tr}_f f e^{\widehat{r}_0 f}\int D_z \tanh(\alpha r + z\sqrt{\widehat{Q}} + \widehat{R}f), \tag{62}$$

$$Q = \frac{1}{2\cosh(\widehat{r}_0)}\operatorname*{Tr}_f e^{\widehat{r}_0 f}\int D_z \tanh^2(\alpha r + z\sqrt{\widehat{Q}} + \widehat{R}f), \tag{63}$$

$$\widehat{R} = \frac{1}{\sigma^2+1-q}, \quad \widehat{Q} = \frac{q-2m+\sigma_0^2+1}{(\sigma^2+1-q)^2}, \tag{64}$$

$$r_0 = \tanh(\widehat{r}_0), \quad \widehat{r}_0 = \alpha_0 r_0. \tag{65}$$

In these equations, we can find two sets of equations for both the CDMA model [7–9] and the image restoration model [9]. These two equations depend on each other.

IV. COMPUTER SIMULATIONS

A. Overlaps and BER

Let us derive the overlaps d_m and d_R. The overlaps are averaged over all realization of the spreading codes and noises [8,9]. Therefore, the overlaps are given by

$$d_m = \lim_{n\to 0}\lim_{K\to\infty}\left[\frac{1}{K}\sum_{i=1}^{K} s_i \mathrm{sgn}(\langle \widehat{s}_i\rangle_\sigma)\right], \tag{66}$$

$$d_R = \lim_{n\to 0}\lim_{N\to\infty}\left[\frac{1}{N}\sum_{\mu=1}^{N} f_\mu \mathrm{sgn}(\langle \widehat{f}_\mu\rangle_\sigma)\right], \tag{67}$$

where $\langle\cdot\rangle_\sigma$ denotes the average over the posterior distribution and $[\cdot]$ denotes the average over the spreading codes, noises, messages and images [8]. We have

$$d_m = \int_{-\infty}^{\infty} Dz\ \mathrm{sgn}(z\sqrt{\widehat{q}}+\widehat{m}) \tag{68}$$

$$= \mathrm{erf}\left(\frac{\widehat{m}}{\sqrt{2\widehat{q}}}\right), \tag{69}$$

$$d_R = \frac{1}{2\cosh(\widehat{r}_0)}\mathop{\mathrm{Tr}}_f f e^{\widehat{r}_0 f}\int_{-\infty}^{\infty} D_z \mathrm{sgn}(\alpha r + z\sqrt{\widehat{Q}}+\widehat{R}f) \tag{70}$$

$$= \frac{1}{2\cosh(\widehat{r}_0)}\left[e^{\widehat{r}_0}\mathrm{erf}\left(\frac{\alpha r+\widehat{R}}{\sqrt{2\widehat{Q}}}\right) - e^{-\widehat{r}_0}\mathrm{erf}\left(\frac{\alpha r-\widehat{R}}{\sqrt{2\widehat{Q}}}\right)\right], \tag{71}$$

where $\mathrm{erf}(x)$ is the error function defined by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt. \tag{72}$$

From the overlaps, the BERs can be given by

$$\mathrm{BER}_m = \frac{1}{2}\left[1-\mathrm{erf}\left(\frac{\widehat{m}}{\sqrt{2\widehat{q}}}\right)\right], \tag{73}$$

$$\mathrm{BER}_R = \frac{1}{2}\left\{1-\frac{1}{2\cosh(\widehat{r}_0)}\left[e^{\widehat{r}_0}\mathrm{erf}\left(\frac{\alpha r+\widehat{R}}{\sqrt{2\widehat{Q}}}\right) - e^{-\widehat{r}_0}\mathrm{erf}\left(\frac{\alpha r-\widehat{R}}{\sqrt{2\widehat{Q}}}\right)\right]\right\}. \tag{74}$$

B. Verification of saddle-point equations

We verify the obtained saddle-point equations by computer simulations. First, we consider the infinite-range model for the image restoration model. Figure 2 shows the sample images generated that satisfy the prior probability (10), where $\alpha_0 = 1.0, 1.5$, and 2.0. In the case of (b) $\alpha_0 = 1.5$, the average value of the pixels is $r_0 = 0.859$ from (65). The size of sample images in Fig. 2 is 256×256 pixels. Since the length of the spread codes is $N = 1024$, we use smaller original images of 32×32 pixels for the computer simulations. The message lengths are $K = 128$, 512, and 1024. Figure 3 shows the BER as a function of the channel noise. The parameters in the decoder, α and σ^2, are given by true values $\alpha = \alpha_0$ and $\sigma^2 = \sigma_0^2$. The abscissa axis represents E_b/N_0 given by

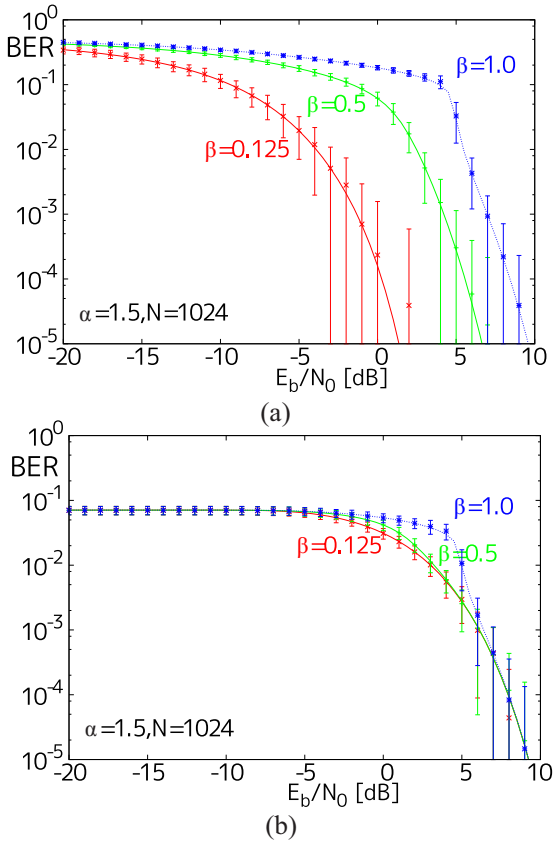

FIG. 3. (a) Bit error rate BER_m for messages and (b) BER_R for image. The smooth parameter is $\alpha = 1.5$. The embedding rates are $\beta = 0.125$, 0.5, and 1.0.

$$\frac{E_b}{N_0} = 10\log_{10}\left(\frac{1}{2\sigma^2}\right)(\mathrm{dB}), \tag{75}$$

where σ^2 is the variance of the Gaussian channel. The axis of ordinate represents the BERs for both messages BER_m and images BER_R. BER_m is averaged over 200 trials in the computer simulations. The average BER_m is shown with error bars. BER_R is calculated on whole image and is shown with points. The initial values of the estimated messages and estimated image are set by the true values, and then we obtain one of the best solutions. The theoretical values obtained by the saddle-point equations are plotted by a solid line for embedding rate $\beta = K/N = 0.125$ ($K = 128$), a dashed line for $\beta = 0.5$ ($K = 512$), and a double-dashed line for $\beta = 1.0$ ($K = 1024$). The computer simulations results agreed with those derived theoretically. In Fig. 3, the BER_m for the messages worsened according to the embedding rate β, while the BER_R for the images were slightly influenced by β under the fixed smooth parameter α.

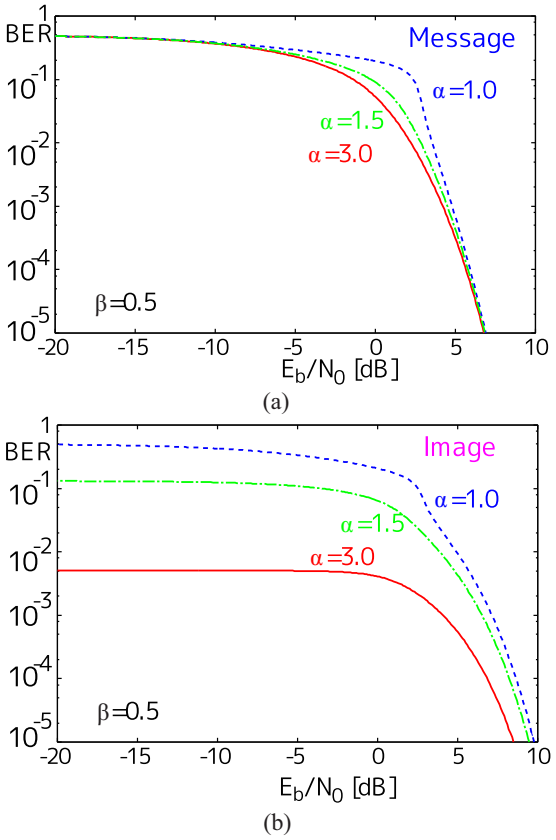

FIG. 4. (a) Bit error rate BER_m for messages and (b) BER_R for image. The smooth parameters are $\alpha = 1.0, 1.5, 3.0$. The embedding rate is $\beta = 0.5$.

Next, we evaluate the bit error rate for the smooth parameters $\alpha = 1.0, 1.5$, and 3.0 under the fixed embedding rate β. Figure 4 shows the BERs for the embedding rate $\beta = 0.5$. Because of the fixed embedding rate, the BER_m for the messages were slightly influenced by the parameter α, while the BER_R for the images became better according to α. In other words, smoother images can be easily restored.

C. Advantages of image restoration

The key concept underlying the proposed method is that it can estimate both the messages and image at the same time in the decoder. Here, we compare the performance of the blind decoder with that of the informed decoder. Cases in which the original image is known or informed to the decoder correspond to the CDMA model, and only messages are estimated.

Figure 5 shows the bit error rate BER_m for messages in the blind and informed decoders. The embedding rates are $\beta = 0.125, 0.5$, and 1.0. The BERs in the blind decoder are larger than those in the informed decoder because images are also estimated. However, in cases in which the embedding rate β is small enough, or in which there is not much noise in the communication channel, there is not much difference between the blind and informed decoders, i.e., blind decoder can successfully carry out image estimation.

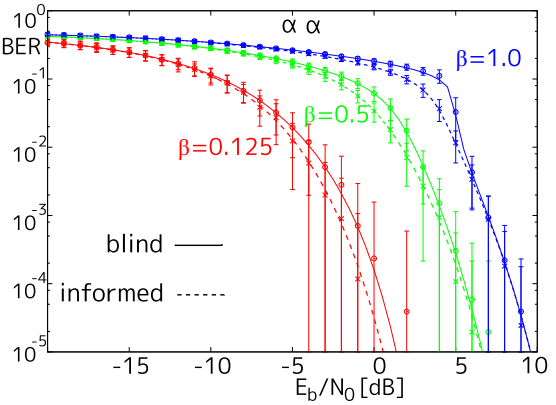

FIG. 5. BER_m in the cases of blind and informed images.

D. Two-dimensional Ising model

In addition to the infinite-range model, we also consider the 2D Ising model for image restoration, in which each pixel is statically connected with four-neighbors. This model is natural for the image restoration. In this model, there are some clusters in generated images because the pixels interact with their nearest neighbors. These cluster patterns can be seen in the parity of JPEG images. In this section, we treat the 2D Ising model as an image generating model; that is, the prior probability is given by

$$P(\boldsymbol{f}) \propto \exp\left[\alpha_0 \sum_{\langle \mu,\nu\rangle} f_\mu f_\nu\right], \tag{76}$$

where $\langle \mu, \nu\rangle$ denotes pairs of nearest neighbor sites. Figure 6 shows the generated images for parameters $\alpha_0 = 0.4$, 1.5, and 10 in the 2D Ising model. In this manner, once the generating models have been changed, the generated images are much different. Since it is difficult to construct a generating model of natural images, it is necessary to consider various generating models in which as many characteristics of natural images are applied as possible.

The posterior probability of the original image $\boldsymbol{f}$ given the tampered image $\boldsymbol{r}$ is given by

$$P(\boldsymbol{f}|\boldsymbol{r}) = \frac{1}{Z}\exp\left[-\frac{1}{2\sigma^2}\sum_{\mu=1}^{N}(r_\mu - f_\mu)^2 + \alpha \sum_{\langle \mu,\nu\rangle} f_\mu f_\nu\right]. \tag{77}$$

Although the replica method can be applied to a certain 2D Ising model with diluted random connections by using a mean field approximation [20], the exact treatment of the 2D Ising model is technically difficult and the replica method does not yield the accurate assessment. We therefore evaluate its performance by computer simulations. Since we can see the continuous structure in Fig. 6, the image size in the 2D Ising model is 256×256 pixels unlike ones of the infinite-range model. The images are divided into 256 blocks, whose size is 256 pixels per a block. So, the spread code length is $N = 256$. Figure 7 shows the bit error rates BER_m and BER_R for the 2D Ising model. The parameters α and σ^2 are set to the true value $\alpha = \alpha_0$ and $\sigma^2 = \sigma_0^2$. The BERs are averaged over all blocks.

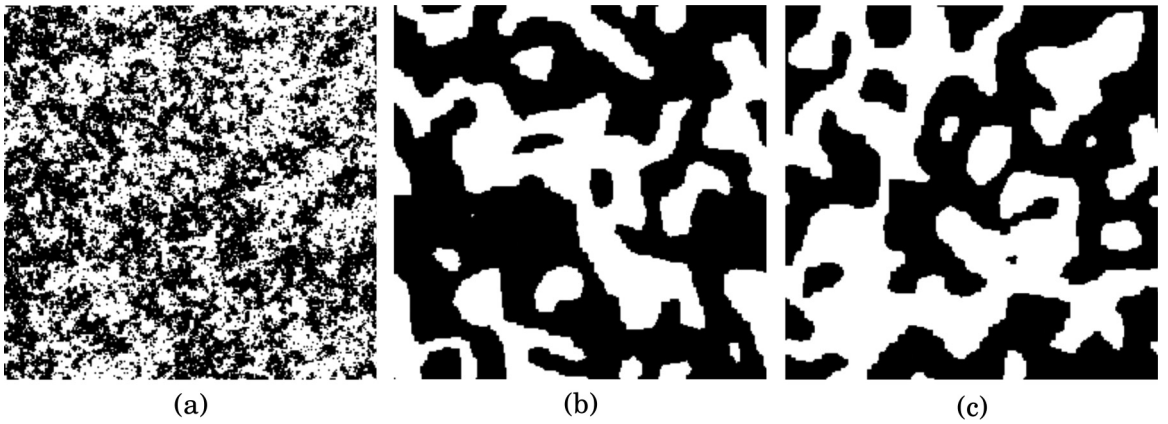

FIG. 6. Images generated by 2D-Ising model (256 × 256 pixels) with the smooth parameters (a) $\alpha_0 = 0.4$, (b) $\alpha_0 = 1.5$, and (c) $\alpha_0 = 10.0$.

BER_R for the images are slightly influenced by the embedding rate β under the fixed parameter α.

Next, we evaluate the performance under the fixed embedding rate $\beta = 0.25$. Figure 8 shows BERs for the smooth parameters $\alpha_0 = 0.4$, 1.5, and 10.0. BER_m for messages are slightly influenced by α. BER_R for images in cases of $\alpha_0 > 1$ are smaller than those of $\alpha = 0.4$. For large $\alpha_0 = 10$, phase transition may occur.

Figure 9 shows the bit error rate BER_m for messages in both blind and informed decoders. The embedding rates were $\beta = 0.125$, 0.5, and 1.0. Curved lines denote the theoretical values for the informed decoder. The blind decoder had just as good a performance as the informed decoder. That is, the blind decoder could successfully restore the image and estimate the messages.

V. CONCLUSION

We proposed an estimation method that can estimate messages and an image at the same time when using a blind

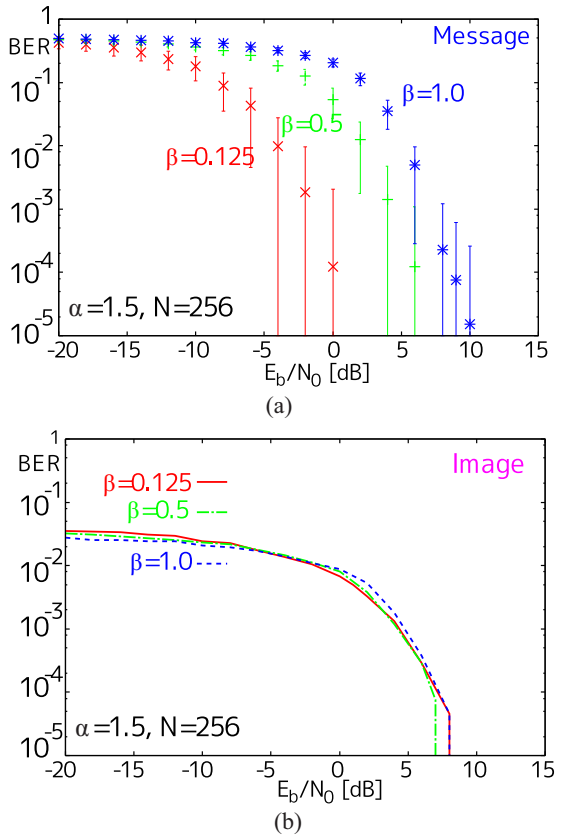

FIG. 7. (a) Bit error rates BER_m for messages and (b) BER_R for image in 2D Ising model. The smooth parameter is $\alpha = \alpha_0 = 1.5$. The embedding rates are $\beta = 0.125$, 0.5, 1.0.

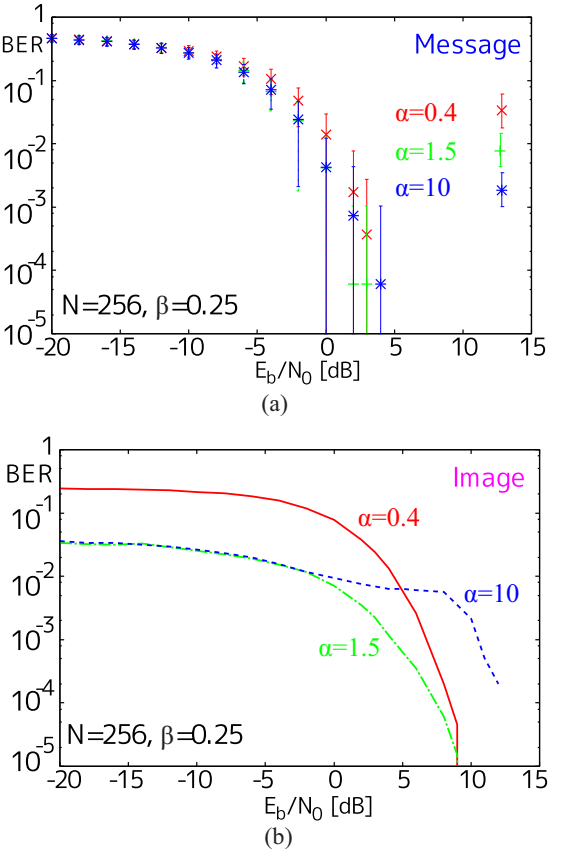

FIG. 8. (a) Bit error rate BER_m for messages and (b) BER_R for image in 2D Ising model. The smooth parameters are $\alpha_0 = 0.4$, 1.5, and 10.0. The embedding rate is $\beta = 0.25$.

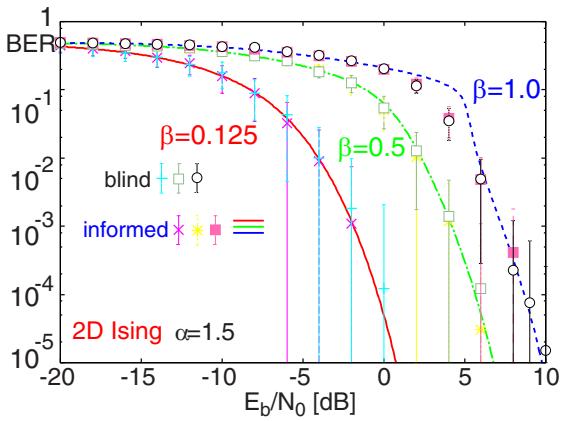

FIG. 9. BER_m in blind and informed cases. The smooth parameter is $\alpha = 1.5$. The embedding rates are $\beta = 0.125, 0.5, 1.0$.

decoder. When this method is used with Bayes estimation, prior probabilities for both the messages and the images are required. In this paper, we assumed that the prior probability for messages had a uniform distribution and that those for images were the infinite-range model and 2D Ising model.

For the infinite-range model, we derived the saddle-point equations by the replica method in order to evaluate the average performance. Since there are two terms—the messages term and the image term—we implemented a two-step approach: First, we introduced order parameters for the messages and assumed replica symmetry for them, and second, we introduced order parameters for the image and assumed replica symmetry for them. The obtained saddle-point equations consist of two indivisible parts: the equations of the CDMA model and those of the image restoration model. We verified the saddle-point equations by computer simulations. The theoretical results agreed with those of the simulations.

Next, we evaluated the performance of the 2D Ising model by computer simulations. When the smooth parameter α_0 was fixed, there was little change in the BER for images, and the BER for messages depended on the embedding rate β. In contrast, when the embedding rate β was fixed, there was little change in the BER for messages, and the BER for images depended on the smooth parameter. However, there was a lower bound in the 2D Ising model.

We also evaluated the performance differences between blind and informed decoders. Results showed that the difference was very small when the embedding or attack rates were small, since the image restoration could still be carried out well. This demonstrates the effectiveness of the proposed method.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grants No. 21700255, No. JP16K00156, and No. JP16K05474. The computer simulations were carried out on PC clusters at Yamaguchi University and on multicore processors at Nara Women's University.

APPENDIX A: INTEGRAL OF $[Z^n]$ WITH RESPECT TO $r_\mu, v_0^\mu, \hat{v}_0^\mu$

From (36), we obtain

$$[Z^n] = \int \prod_{a<b} \frac{id q_{ab} d\hat{q}_{ab}}{2\pi} \prod_a \frac{id m_a d\hat{m}_a}{2\pi} e^{N(G_1+G_2+G_3)}, \tag{A1}$$

where

$$e^{G_1} = \exp\left(-\beta \sum_{a<b} \hat{q}_{ab} q_{ab} - \beta \sum_a \hat{m}_a m_a\right), \tag{A2}$$

$$e^{NG_2} = \mathrm{Tr}_{s,x} \prod_{k=1}^{K} \exp\left(\sum_{a<b} \hat{q}_{ab} x_k^a x_k^b + \sum_a \hat{m}_a s_k x_k^a\right), \tag{A3}$$

$$\begin{aligned} e^{NG_3} = \mathrm{Tr}_{f,g} \prod_\mu \int \frac{d v_0^\mu d\hat{v}_0^\mu}{2\pi} \prod_a \frac{d v_a^\mu d\hat{v}_a^\mu}{2\pi} \frac{dr_\mu}{\sqrt{2\pi}\sigma_0} \exp\Bigg[& i\hat{v}_0^\mu v_0^\mu + i \sum_a \hat{v}_a^\mu v_a^\mu - \frac{1}{2} \sum_a \left(\hat{v}_a^\mu\right)^2 - \frac{1}{2}\left(\hat{v}_0^\mu\right)^2 \\ & - \sum_{a<b} q_{ab} \hat{v}_a \hat{v}_b - \sum_a m_a \hat{v}_0 \hat{v}_a - \frac{1}{2\sigma_0^2}\left(r_\mu - v_0^\mu - f_\mu\right)^2 - \frac{1}{2\sigma^2} \sum_a \left(r_\mu - v_a^\mu - g_\mu^a\right)^2 \\ & + \frac{\alpha_0}{N} \sum_{\mu<\nu} f_\mu f_\nu + \frac{\alpha}{N} \sum_{a=1}^{n} \sum_{\mu<\nu} g_\mu^a g_\nu^a \Bigg], \end{aligned} \tag{A4}$$

Now, we integrate e^{NG_3} by $r_\mu, v_0^\mu, \widehat{v}_0^\mu$:

$$
\begin{aligned}
e^{NG_3} = \operatorname*{Tr}_{f,g} \prod_\mu \sqrt{\frac{\sigma^2}{\sigma^2+n(\sigma_0^2+1)}} \int \prod_a \frac{dv_a^\mu d\widehat{v}_a^\mu}{2\pi} \int D_{t_\mu} \exp\Bigg(& -\frac{1}{2\sigma^2}\sum_a \left(v_a^\mu\right)^2 + i\sum_a \widehat{v}_a^\mu v_a^\mu - \frac{1}{2}\sum_a \left(\widehat{v}_a^\mu\right)^2 \\
& + \left\{ t_\mu \sqrt{\frac{\sigma_0^2+1}{\sigma^2[\sigma^2+n(\sigma_0^2+1)]}} + \frac{1}{\sigma^2+n(\sigma_0^2+1)} \left(-i\sum_a m_a \widehat{v}_a^\mu + f_\mu + \frac{\sigma_0^2+1}{\sigma^2}\sum_a g_\mu^a \right) \right\} \sum_a v_a^\mu \\
& - \frac{1}{\sigma^2}\sum_a g_\mu^a v_a^\mu + \frac{n}{2[\sigma^2+n(\sigma_0^2+1)]} \left(\sum_a m_a \widehat{v}_a^\mu \right)^2 + \frac{in}{\sigma^2+n(\sigma_0^2+1)} f_\mu \sum_a m_a \widehat{v}_a^\mu \\
& + \frac{1}{\sigma^2+n(\sigma_0^2+1)} \left(-i\sum_a m_a \widehat{v}_a^\mu + f_\mu \right) \sum_a g_\mu^a + \frac{\sigma_0^2+1}{2\sigma^2\left[\sigma^2+n(\sigma_0^2+1)\right]} \left(\sum_a g_\mu^a \right)^2 \\
& - \sum_{a<b} q_{ab} \widehat{v}_a^\mu \widehat{v}_b^\mu + \frac{\alpha_0}{N}\sum_{\mu<\nu} f_\mu f_\nu + \frac{\alpha}{N}\sum_{a=1}^{n}\sum_{\mu<\nu} g_\mu^a g_\nu^a \Bigg).
\end{aligned} \tag{A5}
$$

Under the assumption of the replica symmetry, we obtain

$$
\begin{aligned}
e^{NG_3} = \operatorname*{Tr}_{f,g} \prod_\mu \sqrt{\frac{\sigma^2}{\sigma^2+n(\sigma_0^2+1)}} \int \prod_a \frac{dv_a^\mu d\widehat{v}_a^\mu}{2\pi} \int D_{t_\mu} \exp\Bigg(& -\frac{1}{2\sigma^2}\sum_a \left(v_a^\mu\right)^2 + i\sum_a \widehat{v}_a^\mu v_a^\mu - \frac{1}{2}\sum_a \left(\widehat{v}_a^\mu\right)^2 \\
& + \left\{ t_\mu \sqrt{\frac{\sigma_0^2+1}{\sigma^2\left[\sigma^2+n(\sigma_0^2+1)\right]}} + \frac{1}{\sigma^2+n(\sigma_0^2+1)} \left(-im\sum_a \widehat{v}_a^\mu + f_\mu + \frac{\sigma_0^2+1}{\sigma^2}\sum_a g_\mu^a \right) \right\} \sum_a v_a^\mu \\
& - \frac{1}{\sigma^2}\sum_a g_\mu^a v_a^\mu + \frac{nm^2}{2\left[\sigma^2+n(\sigma_0^2+1)\right]} \left(\sum_a \widehat{v}_a^\mu \right)^2 + \frac{inm}{\sigma^2+n(\sigma_0^2+1)} f_\mu \sum_a \widehat{v}_a^\mu \\
& + \frac{1}{\sigma^2+n(\sigma_0^2+1)} \left(-im\sum_a \widehat{v}_a^\mu + f_\mu \right) \sum_a g_\mu^a + \frac{\sigma_0^2+1}{2\sigma^2\left[\sigma^2+n(\sigma_0^2+1)\right]} \left(\sum_a g_\mu^a \right)^2 \\
& - \frac{q}{2}\left[\left(\sum_a \widehat{v}_a^\mu \right)^2 - \sum_a \left(\widehat{v}_a^\mu\right)^2 \right] + \frac{\alpha_0}{2N} \left(\sum_{\mu=1}^{N} f_\mu \right)^2 - \frac{\alpha_0}{2} + \frac{\alpha}{2N}\sum_{a=1}^{n} \left(\sum_{\mu=1}^{N} g_\mu^a \right)^2 - \frac{\alpha}{2} \Bigg).
\end{aligned} \tag{A6}
$$

APPENDIX B: INTEGRAL OF e^{NG_3} WITH RESPECT TO $v_a^\mu, \widehat{v}_a^\mu$

Under the assumption of the replica symmetry, we integrate by v_a^μ, and eliminate the terms at the limit $n \to 0$. We obtain

$$
\begin{aligned}
e^{NG_3} = \operatorname*{Tr}_{f,g} \prod_\mu \sqrt{\frac{\sigma^2}{\sigma^2+n(\sigma_0^2+1)}} \int \prod_a \frac{dv_a^\mu d\widehat{v}_a^\mu}{2\pi} \int D_{t_\mu} \exp\Bigg(& -\frac{1}{2\sigma^2}\sum_a \left(v_a^\mu\right)^2 + i\sum_a \widehat{v}_a^\mu v_a^\mu \\
& + \left\{ t_\mu \sqrt{\frac{\sigma_0^2+1}{\sigma^2\left[\sigma^2+n(\sigma_0^2+1)\right]}} + \frac{1}{\sigma^2+n(\sigma_0^2+1)} \left(-im\sum_a \widehat{v}_a^\mu + f_\mu + \frac{\sigma_0^2+1}{\sigma^2}\sum_a g_\mu^a \right) \right\} \sum_a v_a^\mu \\
& - \frac{1}{\sigma^2}\sum_a g_\mu^a v_a^\mu - \frac{1}{2}(1-q)\sum_a \left(\widehat{v}_a^\mu\right)^2 + \left\{ \frac{nm^2}{2\left[\sigma^2+n(\sigma_0^2+1)\right]} - \frac{q}{2} \right\} \left(\sum_a \widehat{v}_a^\mu \right)^2 \\
& + \frac{im}{\sigma^2+n(\sigma_0^2+1)} \left(nf_\mu - \sum_a g_\mu^a \right) \sum_a \widehat{v}_a^\mu + \frac{1}{\sigma^2+n(\sigma_0^2+1)} f_\mu \sum_a g_\mu^a \\
& + \frac{\sigma_0^2+1}{2\sigma^2\left[\sigma^2+n(\sigma_0^2+1)\right]} \left(\sum_a g_\mu^a \right)^2 + \frac{\alpha_0}{2N} \left(\sum_{\mu=1}^{N} f_\mu \right)^2 + \frac{\alpha}{2N}\sum_{a=1}^{n} \left(\sum_{\mu=1}^{N} g_\mu^a \right)^2 \Bigg),
\end{aligned} \tag{B1}
$$

$$= \operatorname*{Tr}_{f,g} \prod_{\mu} \sigma^n \int \prod_a \frac{d\widehat{v}_a^{\mu}}{\sqrt{2\pi}} \int D_{t_\mu} \exp\left[-\frac{1}{2}(\sigma^2+1-q)\sum_a \left(\widehat{v}_a^{\mu}\right)^2 + \left(m-\frac{q}{2}\right)\left(\sum_a \widehat{v}_a^{\mu}\right)^2 \right.$$
$$+\frac{inm}{\sigma^2} f_\mu \sum_a \widehat{v}_a^{\mu} - i\sum_a g_\mu^a \widehat{v}_a^{\mu} + i\left(1-\frac{nm}{\sigma^2}\right)\left(f_\mu + \frac{\sigma_0^2+1}{\sigma^2}\sum_b g_\mu^b + t_\mu\sqrt{\sigma_0^2+1}\right)\sum_a \widehat{v}_a^{\mu}$$
$$+\frac{n}{2\sigma^2}\left(f_\mu + \frac{\sigma_0^2+1}{\sigma^2}\sum_a g_\mu^a\right)^2 - \frac{\sigma_0^2+1}{2\sigma^4}\left(\sum_a g_\mu^a\right)^2 + \frac{nt_\mu\sqrt{\sigma_0^2+1}}{\sigma^2}\left(f_\mu + \frac{\sigma_0^2+1}{\sigma^2}\sum_a g_\mu^a\right)$$
$$\left. -\frac{t_\mu\sqrt{\sigma_0^2+1}}{\sigma^2}\sum_a g_\mu^a + \frac{nt_\mu^2(\sigma_0^2+1)}{2\sigma^2} + \frac{\alpha_0}{2N}\left(\sum_{\mu=1}^N f_\mu\right)^2 + \frac{\alpha}{2N}\sum_{a=1}^n\left(\sum_{\mu=1}^N g_\mu^a\right)^2\right]. \tag{B2}$$

Using Hubbard-Stratonovich transformation,

$$\exp\left[\left(m-\frac{q}{2}\right)\left(\sum_a \widehat{v}_a^{\mu}\right)^2\right] = \int D_{z_\mu} \exp\left[z_\mu\sqrt{2m-q}\sum_a \widehat{v}_a^{\mu}\right], \tag{B3}$$

we integrate by $\widehat{v}_a^{\mu}, t_\mu, z_\mu$:

$$e^{NG_3} = \operatorname*{Tr}_{f,g}\prod_\mu \sigma^n(\sigma^2+1-q)^{-\frac{n}{2}}\left[1+\frac{n(2m-q-\sigma_0^2-1)}{2(\sigma^2+1-q)} + \frac{n(\sigma_0^2+1)}{2\sigma^2} + \Upsilon\left(\sum_a g_\mu^a\right)^2\right]$$
$$\times \exp\left[\Phi + \Psi\sum_{a<b} g_\mu^a g_\mu^b + \Omega f_\mu\sum_a g_\mu^a + \frac{\alpha_0}{2N}\left(\sum_{\mu=1}^N f_\mu\right)^2 + \frac{\alpha}{2N}\sum_a\left(\sum_{\mu=1}^N g_\mu^a\right)^2\right], \tag{B4}$$

where

$$\Upsilon = \frac{1}{2}\left[-\frac{2m-q-\sigma_0^2-1}{(\sigma^2+1-q)^2} + \frac{2(\sigma_0^2+1)\left(1-\frac{nm}{\sigma^2}\right)}{\sigma^2(\sigma^2+1-q)} + \frac{\sigma_0^2+1}{\sigma^4}\right], \tag{B5}$$

$$\Phi = n\left[\frac{\sigma_0^2+1}{\sigma^2(\sigma^2+1-q)}\left(1-\frac{nm}{\sigma^2}\right) - \frac{\sigma_0^2+1}{2\sigma^4} - \frac{1}{\sigma^2+1-q}\right], \tag{B6}$$

$$\Psi = \frac{\sigma_0^2+1}{\sigma^2(\sigma^2+1-q)}\left(1-\frac{nm}{\sigma^2}\right)\left[2-\frac{n(\sigma_0^2+1)}{\sigma^2}\left(1-\frac{nm}{\sigma^2}\right)\right] - \frac{\sigma_0^2+1}{\sigma^4}, \tag{B7}$$

$$\Omega = \frac{1}{\sigma^2+1-q}\left[1-n(\sigma_0^2+1)\left(1-\frac{nm}{\sigma^2}\right)^2\right] + \frac{n(\sigma_0^2+1)}{\sigma^4}. \tag{B8}$$

[1] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, IEEE Int. Conf. Image Process. **3**, 243 (1996).

[2] I. J. Cox, M. Miller, J. A. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, 2nd ed. (Morgan Kaufmann Publishers Inc., San Francisco, CA, 2008).

[3] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, IEEE Trans. Image Process. **6**, 1673 (1997).

[4] J. Ohnishi and K. Matsui, in *Proceedings of the Third IEEE International Conference on Multimedia Computing and Systems (ICMCS'96)* (IEEE, 1996), pp. 514–521.

[5] I. J. Cox, M. L. Miller, and A. L. McKellips, Proc. IEEE **87**, 1127 (1999).

[6] S. Verdú, Algorithmica **1**, 303 (1989).

[7] T. Tanaka, Europhys. Lett. **54**, 540 (2001).

[8] T. Tanaka, IEEE Trans. IT **48**, 2888 (2002).

[9] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing* (Oxford University Press, Oxford, 2001).

[10] M. Yoshida, T. Uezu, T. Tanaka, and M. Okada, J. Phys. Soc. Jpn. **76**, 054003 (2007).

[11] K. Senda and M. Kawamura, *Information Theoretic Security*, Lecture Notes in Computer Science Vol. 5973 (Springer, Berlin, Heidelberg, 2010), pp. 231–247.

[12] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 (1984).

[13] H. Nishimori and K. Y. Michael Wong, Phys. Rev. E **60**, 132 (1999).

[14] J.-i. Inoue and D. M. Carlucci, Phys. Rev. E **64**, 036121 (2001).

[15] K. Tanaka, J. Phys. A: Math. Gen. **35**, R81 (2002).

[16] J. Fridrich, M. Goljan, P. Lisoněk, and D. Soukal, IEEE Trans. Sign. Process. **53**, 3923 (2005).
[17] J. Fridrich, IEEE Trans. Inf. Forens. Secur. **1**, 390 (2006).
[18] J. R. Hernández and F. P.-González, Proc. IEEE **87**, 1142 (1999).
[19] J. Su, F. Hartung, and B. Girod, in *Security and Watermarking of Multimedia Contents*, Proc. SPIE Vol. 3657 (SPIE, 1999), pp. 159–170.
[20] S. F. Edwards and P. W. Anderson, J. Phys. F **5**, 965 (1975).