The Annals of Probability
1987, Vol. 15, No. 3, 1112–1125

A DECOMPOSITION THEOREM FOR BINARY MARKOV RANDOM FIELDS[1]

BY BRUCE HAJEK AND TOBY BERGER

University of Illinois, Urbana-Champaign and Cornell University

Consider a binary Markov random field whose neighbor structure is specified by a countable graph with nodes of uniformly bounded degree. Under a minimal assumption we prove a decomposition theorem to the effect that such a Markov random field can be represented as the nodewise modulo 2 sum of two independent binary random fields, one of which is white binary noise of positive weight. Said decomposition provides the information theorist with an exact expression for the per-site rate-distortion function of the random field over an interval of distortions not exceeding this weight. We mention possible implications for communication theory, probability theory and statistical physics.

1. Introduction. Let $X = (X_i: i \in \mathbb{S})$ denote a random process such that each X_i takes values in $\{0,1\}$ and $\mathbb{S}$ is a countably infinite set. Given d with $0 < d \leq 0.5$, there may exist a representation of X of the form

$$X = Y \oplus U, \tag{1.1}$$

where $Y \oplus U$ represents the componentwise modulo 2 sum of Y and U, where Y and U are independent, where the components of U are mutually independent, and where $P[U_i = 1] = d$ for each i.

Our main result is Theorem 1 of Section 3. It says that, if X is a Markov random field, or equivalently a nearest neighbor Gibbs field, with a uniformly bounded number of neighbors, then under a minimal assumption a representation of the form (1.1) exists for all d sufficiently small. (An enlargement of the probability space may also be needed.) To lead up to this result, we first study similar representations for random vectors in Section 2.

In Section 4 we elucidate the connection between representations of the form (1.1) and the rate-distortion functions of information theory. We show that for a range of sufficiently small distortions the exact expression for the rate-distortion function of a stationary Markov random field depends on the joint statistics solely through the field's entropy density. We also mention possible implications for communication theory, probability theory and statistical physics.

2. Bernoulli extraction from *n*-vectors. Let $\pi_{\mathbf{X}}$ be a probability distribution on $\{0,1\}^n$. We say that $\mathbf{D} = (D_1, \ldots, D_n)$ is *extractable* from $\pi_{\mathbf{X}}$ if the

Received July 1985; revised June 1986.

[1]This work was supported in part by the National Science Foundation under grants ECS-8352030 and ECS-8305681.

AMS 1980 *subject classifications.* Primary 60G60; secondary 94A34, 60K35.

Key words and phrases. Markov random field, Gibbs random field, Ising model, rate-distortion function.

following is true. There are random n-vectors $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{U}$ with binary coordinates such that $\mathbf{X}$ has distribution $\pi_{\mathbf{X}}$,

$$\mathbf{X} = \mathbf{Y} \oplus \mathbf{U},$$

$\mathbf{Y}, U_1, \ldots, U_n$ are mutually independent and

$$P[U_i = 1] = D_i, \qquad 1 \le i \le n.$$

By conditioning on $\mathbf{Y}$, we obtain the following expression for $\pi_{\mathbf{X}}$ in terms of the distribution $\pi_{\mathbf{Y}}$ of $\mathbf{Y}$:

$$\pi_{\mathbf{X}}(\mathbf{x}) = \sum_{\mathbf{y}} T_{\mathbf{D}}(\mathbf{x}, \mathbf{y}) \pi_{\mathbf{Y}}(\mathbf{y}), \tag{2.1}$$

where

$$T_{\mathbf{D}}(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{n} t_{D_i}(x_i, y_i) \tag{2.2}$$

and

$$t_{D_i} = \begin{bmatrix} t_{D_i}(0,0) & t_{D_i}(0,1) \\ t_{D_i}(1,0) & t_{D_i}(1,1) \end{bmatrix} = \begin{bmatrix} 1 - D_i & D_i \\ D_i & 1 - D_i \end{bmatrix}.$$

We can view $T_{\mathbf{D}}$ as a $2^n \times 2^n$ matrix. Equation (2.1) can be rewritten as

$$\pi_{\mathbf{X}} = T_{\mathbf{D}} \pi_{\mathbf{Y}}, \tag{2.3}$$

and (2.2) means that $T_{\mathbf{D}}$ is the Kronecker product of the n 2×2 matrices $t_{D_1}, \ldots, t_{D_n}$.

LEMMA 1. *Suppose that* $0 \le D_i < \frac{1}{2}$ *for* $1 \le i \le n$. *Then* $\mathbf{D}$ *is extractable from* $\pi_{\mathbf{X}}$ *if and only if the vector* $T_{\mathbf{D}}^{-1} \pi_{\mathbf{X}}$ *has nonnegative entries.*

PROOF. The inverse $T_{\mathbf{D}}^{-1}$ exists, for it is the Kronecker product of $t_{D_1}^{-1}, \ldots, t_{D_n}^{-1}$, and the inverse

$$t_{D_j}^{-1} = \frac{1}{(1 - D_j)^2 - D_j^2} \begin{bmatrix} 1 - D_j & -D_j \\ -D_j & 1 - D_j \end{bmatrix} \tag{2.4}$$

exists for each j. If $\mathbf{D}$ is extractable from $\pi_{\mathbf{X}}$, then by (2.3), $T_{\mathbf{D}}^{-1} \pi_{\mathbf{X}}$ is equal to $\pi_{\mathbf{Y}}$, so it has nonnegative entries. On the other hand, if $\mathbf{v}$ defined by $\mathbf{v} = T_{\mathbf{D}}^{-1} \pi_{\mathbf{X}}$ has nonnegative entries, it must be a probability distribution since each $t_{D_i}^{-1}$, and hence $T_{\mathbf{D}}^{-1}$, has columns that sum to 1. Thus, $\pi_{\mathbf{X}} = T_{\mathbf{D}} \mathbf{v}$ for some probability distribution $\mathbf{v}$, which implies that $\mathbf{D}$ is extractable from $\pi_{\mathbf{X}}$. □

LEMMA 2. *Suppose* $\mathbf{X} = (\mathbf{V}_1, \mathbf{V}_2)$ *where* $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *are independent. If* $\mathbf{D}_i$ *is extractable from* $\pi_{\mathbf{V}_i}$ *for* $i = 1, 2$, *then* $(\mathbf{D}_1, \mathbf{D}_2)$ *is extractable from* $\pi_{\mathbf{X}}$.

PROOF. Straightforward. □

Let $\mathbf{X}_1 = (X_1, \ldots, X_j)$ and $\mathbf{X}_2 = (X_{j+1}, \ldots, X_n)$, where $j \in \{1, \ldots, n\}$. Define $\mathbf{D}_1 = (D_1, \ldots, D_j)$, $\mathbf{D}_2 = (D_{j+1}, \ldots, D_n)$ and $\mathbf{D} = (\mathbf{D}_1, \mathbf{D}_2)$.

LEMMA 3. *If* $\mathbf{D}_1 = 0$, *then* $\mathbf{D}$ *is extractable from* $\pi_{\mathbf{X}}$ *if and only if, for every* $\mathbf{x}_1 \in \{0,1\}^j$ *satisfying* $\pi_{\mathbf{X}_1}(\mathbf{x}_1) > 0$, $\mathbf{D}_2$ *is extractable from* $\pi_{\mathbf{X}_2|\mathbf{X}_1}(\cdot\,|\mathbf{x}_1)$.

PROOF. Since $T_{\mathbf{D}}^{-1} = I_{2^j} \times T_{\mathbf{D}_2}^{-1}$, where I_m is the m-dimensional identity matrix, we have

$$T_{\mathbf{D}}^{-1}\pi_{\mathbf{X}} = \pi_{\mathbf{X}_1} T_{\mathbf{D}_2}^{-1}\pi_{\mathbf{X}_2|\mathbf{X}_1},$$

which makes the validity of the lemma transparent. □

LEMMA 4. *Suppose* $\mathbf{X} = (\mathbf{V}_1, \mathbf{W}, \mathbf{V}_2)$ *with* $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *conditionally independent given* $\mathbf{W}$. *If both* $(\mathbf{D}_1, \mathbf{0}, \mathbf{0})$ *and* $(\mathbf{0}, \mathbf{0}, \mathbf{D}_2)$ *are extractable from* $\pi_{\mathbf{X}}$, *then* $(\mathbf{D}_1, \mathbf{0}, \mathbf{D}_2)$ *is extractable from* $\pi_{\mathbf{X}}$.

PROOF. Since $(\mathbf{D}_1, \mathbf{0}, \mathbf{0})$ is extractable from $\pi_{\mathbf{X}}$, $(\mathbf{D}_1, \mathbf{0})$ is extractable from $\pi_{\mathbf{V}_1,\mathbf{W}}$. It follows from Lemma 3 that $\mathbf{D}_1$ is extractable from $\pi_{\mathbf{V}_1|\mathbf{W}}(\cdot\,|\mathbf{w})$ for all $\mathbf{w}$ with $\pi_{\mathbf{W}}(\mathbf{w}) > 0$. Similarly, $\mathbf{D}_2$ is extractable from $\pi_{\mathbf{V}_2|\mathbf{W}}(\cdot\,|\mathbf{w})$ for all $\mathbf{w}$ with $\pi_{\mathbf{W}}(\mathbf{w}) > 0$. Since $\mathbf{V}_1$ and $\mathbf{V}_2$ are conditionally independent given $\mathbf{W}$, it follows from Lemma 2 that $(\mathbf{D}_1, \mathbf{D}_2)$ is extractable from $\pi_{\mathbf{V}_1,\mathbf{V}_2|\mathbf{W}}(\cdot,\cdot\,|\mathbf{w})$ for all $\mathbf{w}$ with $\pi_{\mathbf{W}}(\mathbf{w}) > 0$. By applying Lemma 3 again, we conclude that $(\mathbf{D}_1, \mathbf{0}, \mathbf{D}_2)$ is extractable from $\pi_{\mathbf{X}}$. □

LEMMA 5. *Given* $\mathbf{D}'$ *and* $\mathbf{D}''$ *in* $[0, \frac{1}{2})^n$, *let* $\mathbf{D}$ *denote the vector with components*

$$D_i = D_i'(1 - D_i'') + (1 - D_i')D_i'', \qquad 1 \le i \le n.$$

(*This is especially simple if for each* i *either* $D_i' = 0$ *or* $D_i'' = 0$.) *Then* $\mathbf{D}$ *is extractable from* π *if and only if* $\mathbf{D}''$ *is extractable from* π *and* $\mathbf{D}'$ *is extractable from* $T_{\mathbf{D}''}^{-1}\pi$.

PROOF. It is not hard to check that $T_{\mathbf{D}} = T_{\mathbf{D}''}T_{\mathbf{D}'}$, which implies that

$$T_{\mathbf{D}}^{-1}\pi = T_{\mathbf{D}'}^{-1}\left(T_{\mathbf{D}''}^{-1}\pi\right).$$

Obviously the vector on the right-hand side has nonnegative components if and only if the vector on the left-hand side does. Invoking Lemma 1 completes the proof. □

LEMMA 6. *Suppose that* $0 \le \delta_i < \frac{1}{2}$ *for each* i *and that* $(\delta_1, 0, \ldots, 0)$, $(0, \delta_2, 0, \ldots, 0), \ldots, (0, \ldots, 0, \delta_n)$ *are each extractable from* π. *Let* $\mathbf{D} = (D_1, \ldots, D_n)$, *where*

$$D_i = \left(1 - (1 - 2\delta_i)^{t_i}\right)/2$$

for some nonnegative $t_1, \ldots, t_n$ *with* $t_1 + \cdots + t_n \le 1$. *Then* $\mathbf{D}$ *is extractable from* π.

PROOF. By Lemma 1 we need check only that $T_{\mathbf{D}}^{-1}\pi$ has positive entries, i.e., that for each $\mathbf{y} \in \{0,1\}^n$, the sum

$$\sum_{\mathbf{x}} T_D^{-1}(\mathbf{y},\mathbf{x})\pi(\mathbf{x}) \tag{2.5}$$

is nonnegative. Define

$$r_i = D_i/(1 - D_i)$$

and introduce the change of variable $\mathbf{x} = \mathbf{y} \oplus \eta$. The sum in (2.5) may be rewritten as

$$\left\{\sum_{\eta} \pi(\mathbf{y} \oplus \eta) \prod_i (-r_i)^{\eta_i}\right\} \prod_i (1 - D_i).$$

Let

$$s_i = \delta_i/(1 - \delta_i)$$

and $\mathbf{e}_i = (0,\ldots,0,1,0,\ldots,0)$ with ith coordinate equal to 1. Next, introduce the set G_k comprised of real 2^k-tuples of the form

$$\mathbf{a} = \left(a(u_1,\ldots,u_k),\ u_i \in \{0,1\} \text{ for } 1 \le i \le k\right)$$

according to the following prescription in which $a(\mathbf{x}) = a(x_1,\ldots,x_k)$:

$$G_k = \left\{\mathbf{a} \in \mathbb{R}_+^{2^k}\colon\ a(\mathbf{0}) = 1 \text{ and } s_i \le \frac{a(\eta \oplus \mathbf{e}_i)}{a(\eta)} \le \frac{1}{s_i} \text{ for } 1 \le i \le k \text{ and } \eta \in \{0,1\}^{\eta}\right\}.$$

Then define

$$L_k = \min_{\mathbf{a} \in G_k} \sum_{\eta} a(\eta) \prod_{i \le k} (-r_i)^{\eta_i},$$

$$U_k = \max_{\mathbf{a} \in G_k} \sum_{\eta} a(\eta) \prod_{i \le k} (-r_i)^{\eta_i}$$

and

$$\theta_k = L_k/U_k.$$

As a function of η, $\pi(\mathbf{y} \oplus \eta)/\pi(\mathbf{y})$ is in G_n, so it will suffice to show that $L_n \ge 0$, or equivalently, that $\theta_n \ge 0$. We shall accomplish this by showing that $\theta_k \ge 0$ for every $k \le n$.

We can, and do, assume without loss of generality that $s_1 \ge s_2 \ge \cdots \ge s_n$. The relationship between D_i and δ_i can be reexpressed as $r_i = \rho(s_i, t_i)$, where

$$\rho(s,t) = \frac{(1+s)^t - (1-s)^t}{(1+s)^t + (1-s)^t}, \qquad 0 \le s \le 1.$$

Defining $r * r'$ by

$$r * r' = (r + r')/(1 + rr'),$$

we readily verify that

$$\rho(s,t) * \rho(s,t') = \rho(s, t + t'). \tag{2.6a}$$

Also note for future reference that

$$\rho(s,-t) = -\rho(s,t). \tag{2.6b}$$

Since $\rho(s,t) \geq 0$ for $t \geq 0$, Lemma 6 will be established once we prove the following claim.

CLAIM. For $1 \leq k \leq n$,

$$\theta_k \geq \rho(s_k, 1 - (t_1 + \cdots + t_k))/s_k. \tag{2.7}$$

To start the induction proof, we define $s_0 = U_0 = L_0 = \theta_0 = 1$ so that (2.7) is true for $k = 0$. Suppose now that (2.7) is true for some k with $0 \leq k < n$.

For $a \in G_{k+1}$ and $\mathbf{x} \in \{0,1\}^n$, let $a(\mathbf{x})$ denote the component of $\mathbf{a}$ indexed by $(x_1, \ldots, x_{k+1})$. Then

$$\sum_{\eta} a(\eta) \prod_{i \leq k+1} (-r_i)^{\eta_i} = \left[\sum_{\eta:\, \eta_{k+1}=0} a(\eta) \prod_{i \leq k} (-r_i)^{\eta_i} \right], \\ -r_{k+1} a(\mathbf{e}_{k+1}) \left[\sum_{\eta:\, \eta_{k+1}=1} b(\eta) \prod_{i \leq k} (-r_i)^{\eta_i} \right],$$

where

$$b(\eta) = a(\eta)/a(\mathbf{e}_{k+1}).$$

Now $a(\eta_1, \ldots, \eta_k, 0)$ and $b(\eta_1, \ldots, \eta_k, 1)$, as functions of $(\eta_1, \ldots, \eta_k)$, are in G_k, so the two sums in square brackets each must lie in the interval $[L_k, U_k]$. We also know that $a(0,0,\ldots,0,1) \in [s_{k+1}, 1/s_{k+1}]$ and, by the induction hypothesis, that $L_k \geq 0$. It follows that

$$L_{k+1} \geq L_k - U_k r_{k+1}/s_{k+1} \quad \text{and} \quad U_{k+1} \leq U_k - L_k s_{k+1} r_{k+1},$$

which implies that

$$\theta_{k+1} \geq (\theta_k - r_{k+1}/s_{k+1})(1 - s_{k+1} r_{k+1} \theta_k)^{-1}. \tag{2.8}$$

It is not hard to show that $\rho(s,t)/s$ is increasing in s for $t \geq 0$ so that the induction hypothesis (2.7) implies that

$$\theta_k \geq \rho(s_{k+1}, 1 - (t_1 + \cdots + t_k))/s_{k+1}. \tag{2.9}$$

Since the right-hand side of (2.8) is increasing in θ_k, inequality (2.8) remains valid when we replace θ_k by its lower bound from (2.9). Using (2.6) and the fact that $r_{k+1} = \rho(s_{k+1}, t_{k+1})$, we see that this replacement results precisely in the inequality (2.7) with k everywhere replaced by $k+1$. This proves the claim and establishes Lemma 6. □

REMARK 1. Suppose $\delta_1 = \delta_2 = \cdots = \delta_n = \delta > 0$. Let

$$\pi(\mathbf{x}) = \begin{cases} \delta 2^{-(n-1)} & \text{if } x_1 + \cdots + x_n \text{ is even,} \\ (1-\delta) 2^{-(n-1)} & \text{if } x_1 + \cdots + x_n \text{ is odd.} \end{cases}$$

Then the assumptions of Lemma 6 are satisfied and the sum in (2.5) for

$\mathbf{y} = (0, 0, \ldots, 0)$ is nonnegative if and *only if* $t_1 + \cdots + t_n \leq 1$. Thus, in this case, Lemma 6 is tight.

REMARK 2. Lemma 6 might lead one to conjecture that the set of vectors extractable from a given π is convex. The conjecture is true if $n = 2$. In the remainder of this remark, we present a counterexample to the conjecture for $n = 3$.

Given ε with $0 < \varepsilon < 0.5$, we define a probability vector π^ε on $\{0,1\}^3$ to be the probability distribution of (X_1, X_2, X_3), where $P[X_2 = 0] = P[X_2 = 1] = 0.5$, X_1 and X_3 are conditionally independent given X_2 and $P[X_1 \neq X_2 | X_2] = P[X_3 \neq X_2 | X_2] = \varepsilon$.

Note that $P[X_2 = 0 | X_1 = 1,\ X_3 = 1] = \delta$, where $\delta = \varepsilon^2/[\varepsilon^2 + (1-\varepsilon)^2]$. Let $\mathbf{D}^a = (\varepsilon, 0, \varepsilon)$ and $\mathbf{D}^b = (0, \delta, 0)$. It is easy to check that $\mathbf{D}^a$ and $\mathbf{D}^b$ each are extractable from π^ε. We will investigate conditions under which $(\mathbf{D}^a + \mathbf{D}^b)/2$ is extractable from π^ε. Notice that $T^{-1}_{\mathbf{D}^a/2}\pi^\varepsilon = \pi^\alpha$, where α is defined by

$$\alpha(1 - \varepsilon/2) + (\varepsilon/2)(1 - \alpha) = \varepsilon,$$

and π^α is defined in the same way as π^ε with ε replaced by α. Thus, by Lemma 3, $\mathbf{D}$ is extractable from π^ε if and only if $\mathbf{D}^b/2$ is extractable from π^α, or equivalently, if and only if

$$\delta/2 \leq \alpha^2\left[\alpha^2 + (1-\alpha)^2\right]^{-1}. \tag{2.10}$$

The left-hand side of (2.10) is equal to $\varepsilon^2/2 + o(\varepsilon^2)$, while the right-hand side is equal to $\varepsilon^2/4 + o(\varepsilon^2)$. Thus, the inequality is violated for all sufficiently small ε, which implies that for such values of ε the set of vectors extractable from π^ε is *not* convex.

3. Bernoulli extraction from Markov random fields. Let $\mathbb{S}$ denote a countably infinite set, and let $\Omega = \{0,1\}^{\mathbb{S}}$. For i in $\mathbb{S}$, let X_i denote the ith coordinate function on Ω,

$$X_i(\omega) = \omega_i \quad \text{for } \omega \in \Omega,$$

and let Σ denote the smallest σ-algebra of subsets of Ω with respect to which X_i is measurable for each i. Let $\mathbf{D} = (D_i\colon i \in \mathbb{S})$, where $0 \leq D_i \leq \frac{1}{2}$ for each i, and let $\pi_{\mathbf{X}}$ be a probability measure on Σ. We say that $\mathbf{D}$ is extractable from $\pi_{\mathbf{X}}$ if the following is true. There is a probability measure $\pi_{\mathbf{Y}}$ on (Ω, Σ) such that, if $\mathbf{Y} = (Y_i\colon i \in \mathbb{S})$ has distribution $\pi_{\mathbf{Y}}$, if $\mathbf{W} = (W_i\colon i \in \mathbb{S})$ is such that $\mathbf{Y}$ and all the W_j's are mutually independent, if $P[W_i = 1] = 1 - P[W_i = 0] = D_i$, and if $X_i = Y_i \oplus W_i$, then $(X_i\colon i \in \mathbb{S})$ has distribution $\pi_{\mathbf{X}}$.

Assume that for each site i there is a finite set $N(i) \subset \mathbb{S}$ of neighbors of site i, that $i \notin N(i)$, and that $j \in N(i)$ if $i \in N(j)$ for all sites i and j. Define the boundary ∂A of a set of sites A by

$$\partial A = \{i\colon i \text{ is not in } A \text{ and } i \text{ is a neighbor of a site in } A\}.$$

Throughout the rest of this section we assume that π is a probability distribution on (Ω, Σ) satisfying the following assumptions:

A.1 $\pi(\{x_i = \eta_i\colon i \in A\}) > 0$ for any finite subset A of $\mathbb{S}$ and any η in Ω.

A.2 If A and B are finite subsets of Ω such that $A \cap B = \varnothing$ and $\partial A \subset B$, and if $\eta \in \Omega$ and $\mathbf{v} \in \{0,1\}^A$, then

$$\pi(\{X_i = v_i\colon i \in A\} | \{X_i = \eta_i\colon i \in B\}) = \pi_{A,\eta}(\mathbf{v}),$$

where we define $\pi_{A,\eta}$ by

$$\pi_{A,\eta}(\mathbf{v}) = \pi(\{X_i = v_i\colon i \in A\} | \{X_i = \eta_i\colon i \in \partial A\}).$$

A.3 There is an integer M such that any site has at most $M - 1$ neighbors.

A.4 There is a strictly positive constant D_0 such that

$$\pi_{\{i\},\eta}(v) \geq D_0 \quad \text{for any } \eta \in \Omega,\ v \in \{0,1\}.$$

REMARKS ABOUT ASSUMPTIONS.

1. Assumptions A.1 and A.2 together are often taken as the defining properties of a Markov random field. So defined, the set of Markov random fields is well known to be equal to the set of Gibbs states with nearest neighbor potentials; see Dobrushin (1968), Lanford and Ruelle (1969) and Preston (1974).

2. Given the other assumptions, A.4 is easily seen to be equivalent to the assumption that the vector indexed by $\mathbb{S}$ with ith coordinate D_0 and all other coordinates equal to zero is extractable from π.

3. If A.1 were dropped and if A.2 were weakened to cover only η and B such that $\pi(\{X_i = \eta_i\colon i \in B\}) > 0$, then A.2–A.4 would imply A.1 anyway.

For any $d \in [0, \frac{1}{2}]$, define

$$d^* = \left[1 - (1 - 2d)^{1/M}\right]/2,$$

and for $k \geq 0$ define

$$g(k) = \left[1 - (1 - 2D_0)^{1/M^k}\right]/2.$$

Note that

$$g(0) = D_0 \quad \text{and} \quad g(k+1) = g(k)^* \quad \text{for } k \geq 0.$$

For a vector $\mathbf{D} = (D_1, \ldots, D_n)$, we define $\mathbf{D}^* = (D_1^*, \ldots, D_n^*)$.

Given a finite set F and a probability distribution π on $\{0,1\}^F$, we let $R_F(\pi)$ denote the set of vectors $\mathbf{D} = (D_j\colon j \in F)$ such that $\mathbf{D}$ is extractable from π.

LEMMA 7. *Given a finite subset F of $\mathbb{S}$, define D_j for j in F by*

$$D_j = g(k),$$

where k is the number of neighbors that j has in F. Let $\mathbf{D} = (D_j\colon j \in F)$. Then

$$\mathbf{D} \in R_F(\pi_{F,\eta}) \quad \textit{for all } \eta.$$

PROOF. We will use induction on the number of sites in F. The theorem is clearly true if F contains only one site. Suppose the theorem is true for sets containing n sites, and let F be a set containing $n + 1$ sites. Choose a site i in F and write $F = A \cup B \cup \{i\}$, where B is the set of sites in F which are neighbors of site i. For j in $A \cup B$, let $D_j = g(k')$, where k' is the number of neighbors of j which are in $A \cup B$. We shall abuse notation by writing $(\mathbf{D}_A, \mathbf{D}_B)$ instead of $(D_j\colon\ j \in A \cup B)$, with the understanding that $\mathbf{D}_A$ represents the vector elements indexed by A and $\mathbf{D}_B$ represents the vector elements indexed by B.

The induction hypothesis applied to the n-element set $A \cup B$ yields

$$(\mathbf{D}_A, \mathbf{D}_B) \in R_{A\cup B}(\pi_{A\cup B,\eta}) \quad \text{for all } \eta,$$

which directly implies that

$$(\mathbf{D}_A, \mathbf{D}_B, 0) \in R_F(\pi_{F,\eta}) \quad \text{for all } \eta. \tag{3.1}$$

We also have

$$(\mathbf{0}, \mathbf{0}, D_0) \in R_F(\pi_{F,\eta}) \quad \text{for all } \eta. \tag{3.2}$$

For all η, the Markov property assures us that $\pi_{F,\eta}(v)$ assigns probability such that, conditional on the binary random variables $\{v_j,\ j \in B\}$, v_i is independent of $\{v_k,\ k \in A\}$. Consider first the case in which $B = \varnothing$. Then v_i is independent of $\{v_k,\ k \in A\}$, so it follows from Lemma 2, (3.1) and (3.2) that $(\mathbf{D}_A, D_0) \in R_F(\pi_{F,\eta})$ for all η. Thus, the theorem is true when B is empty. Henceforth, we will assume that B is not empty.

Equation (3.1) implies that $(\mathbf{D}_A, \mathbf{0}, 0)$ is in $R_F(\pi_{F,\eta})$, which, by (3.2) and Lemma 4, implies that

$$(\mathbf{D}_A, \mathbf{0}, D_0) \in R_F(\pi_{F,\eta}). \tag{3.3}$$

Next, note that Lemma 5 implies that (3.1) and (3.3), respectively, are equivalent to

$$(\mathbf{0}, \mathbf{D}_B, 0) \in R_F\left(T^{-1}_{(\mathbf{D}_A,\mathbf{0},0)}\pi_{F,\eta}\right) \quad \text{for all } \eta \tag{3.4}$$

and

$$(\mathbf{0}, \mathbf{0}, D_0) \in R_F\left(T^{-1}_{(\mathbf{D}_A,\mathbf{0},0)}\pi_{F,\eta}\right) \quad \text{for all } \eta. \tag{3.5}$$

Hence both $(\mathbf{D}_B, 0)$ and $(\mathbf{0}, D_0)$ are extractable from the conditional distribution for $\{v_k,\ k \in B \cup \{i\}\}$ derived from $T^{-1}_{(\mathbf{D}_A,\mathbf{0},0)}\pi_{F,\eta}$ by conditioning on any fixed value of $\{v_k,\ k \in A\}$. Since $B \cup \{i\}$ contains at most M sites, it follows from Lemma 6 that $(\mathbf{D}_B^*, D_0^*)$ also is extractable from each of these conditional distributions. This permits us to deduce from Lemma 3 that

$$(\mathbf{0}, \mathbf{D}_B^*, D_0^*) \in R_F\left(T^{-1}_{(\mathbf{D}_A,\mathbf{0},0)}\pi_{F,\eta}\right) \quad \text{for all } \eta. \tag{3.6}$$

A final application of Lemma 5 shows that (3.6) is equivalent to

$$(\mathbf{D}_A, \mathbf{D}_B^*, D_0^*) \in R_F(\pi_{F,\eta}) \quad \text{for all } \eta. \tag{3.7}$$

The jth coordinate on the left-hand side of (3.7) is at least as large as $g(k)$, where k is the number of neighbors of j in F. (Observe that $k = k' + 1$ for

$j \in B$, and recall that $B \neq \varnothing$, so $k \geq 1$ for site i.) Thus, the theorem is true for F, which completes our proof by induction. □

THEOREM 1. *Let* $\mathbf{D}$ *denote the vector indexed by* $\mathbb{S}$ *with each component equal to* δ, *where*

$$\delta = g(M-1) = \left[1 - (1-2D_0)^{M^{-(M-1)}}\right]/2. \tag{3.8}$$

Then $\mathbf{D}$ *is extractable from* π.

PROOF. Let $A_1, A_2, \ldots$ be finite subsets of $\mathbb{S}$ such that $A_1 \subset A_2 \subset \cdots$ and $\mathbb{S} = \bigcup_n A_n$. Let $\mathbf{D}(n)$ denote the vector with ith coordinate δ for i in A_n and zero for i in $\mathbb{S} - A_n$. By Lemma 7, $\mathbf{D}(n)$ is extractable from π. Let $\Sigma \otimes \Sigma$ denote the product σ-algebra of subsets of $\Omega \times \Omega$ and let $(\mathbf{Y}, \mathbf{W})$ denote the coordinate functions on $\Omega \times \Omega$. Since $\mathbf{D}(n)$ is extractable from π, there exists a probability measure ν_n on $(\Omega \times \Omega, \Sigma \otimes \Sigma)$ such that, under measure ν_n, $\mathbf{W}$ is a process of independent Bernoulli variables independent of $\mathbf{Y}$, with $\nu_n(\{W_i = 1\}) = D_i(n)$, and $\mathbf{Y} \oplus \mathbf{W}$ has distribution π.

Ω and $\Omega \times \Omega$ with the usual product topologies are compact metrizable spaces and $\Sigma \otimes \Sigma$ is generated by the open subsets of $\Omega \times \Omega$. Thus, by passing to a subsequence if necessary, we can assume that ν_n converges weakly to a distribution ν as n tends to infinity. $\mathbf{Y}$ and $\mathbf{W}$ are continuous mappings from $\Omega \times \Omega$ to Ω so that, under measure ν, $\mathbf{Y}$ and $\mathbf{W}$ are independent and $\mathbf{Y} \oplus \mathbf{W}$ has distribution π. Finally, under measure ν, $\mathbf{W}$ is a process of independent Bernoulli variables such that $\nu(\{W_i = 1\}) = \delta$. Thus $\mathbf{D}$ is extractable from π. □

4. Applications to rate-distortion theory. The concept of extractability is important in information theory for the computation of rate-distortion functions; see Gallager (1968) or Berger (1971). This is elucidated by Theorem 2, a result familiar to many information theorists, but nonetheless proved here for completeness. We then couple Theorem 2 with our preceding results about extraction in order to obtain explicit lower bounds for the so-called critical distortion d_c for the one-dimensional and two-dimensional Ising models.

Let $I(X; Y)$ denote the Shannon mutual information between the pair of random binary n-vectors X and Y; see Gallager (1968). The rate-distortion function for a probability distribution $\pi_{\mathbf{X}}$ on $\{0,1\}^n$ is the function $R_X\colon [0, \frac{1}{2}]^n \to \mathbb{R}^+$ defined by

$$R_X(\mathbf{D}) = \inf I(X; Y),$$

where the infimum is over all pairs of random vectors (X, Y) such that

$$P[X_i \neq Y_i] \leq D_i, \qquad 1 \leq i \leq n, \tag{4.1}$$

and

$$X \text{ has distribution } \pi_{\mathbf{X}}. \tag{4.2}$$

An important related function, the per-site rate-distortion function $\overline{R}_X\colon [0, \frac{1}{2}] \to \mathbb{R}^+$, is defined by

$$\overline{R}_X(d) = \inf n^{-1} I(X; Y),$$

where the infimum is over all pairs of random vectors (X, Y) such that

$$n^{-1} \sum_{i=1}^{n} P[X_i \neq Y_i] \leq d$$

and X has distribution $\pi_{\mathbf{X}}$.

Under broad conditions the limit as n tends to infinity of $\overline{R}_X(d)$ exists and equals the minimum number of encoding bits per site required for X to be recoverable from these bits with an average per-site distortion of d. Rigorous statements of this fact are called source coding theorems; again, see Gallager (1968) or Berger (1971).

THEOREM 2. (a) *If* $\mathbf{D}$ *is extractable from* $\pi_{\mathbf{X}}$, *then*

$$R_X(\mathbf{D}) = H(X) - \sum_{i=1}^{n} h(D_i),$$

where $H(X)$ *denotes the Shannon entropy of the random vector* X, *and*

$$h(x) = -x \log x - (1 - x)\log(1 - x).$$

(b) *If* $(d, d, \ldots, d)$ *is extractable from* $\pi_{\mathbf{X}}$, *then*

$$\overline{R}_X(d) = H(X)/n - h(d).$$

PROOF. If (X, Y) satisfies (4.1) and (4.2), then

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(X) - \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}, Y). \end{aligned}$$

Now

$$H(X_i|X_1, \ldots, X_{i-1}, Y) \leq H(X_i|Y_i) \leq h(D_i),$$

so

$$I(X; Y) \geq H(X) - \sum_{i=1}^{n} h(D_i). \tag{4.3}$$

On the other hand, if $X = Y \oplus U$, where Y and U are independent, $U_1, \ldots, U_n$ are independent and $P[U_i] = D_i$, then

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(X) - H(U) = H(X) - \sum_{i=1}^{n} h(D_i). \end{aligned}$$

This proves part (a) of the theorem.

By the definition of $\overline{R}_X$, we see that

$$\overline{R}_X(d) = \min_{\mathbf{D}} n^{-1} R_X(\mathbf{D}), \tag{4.4}$$

where the minimum is over vectors $\mathbf{D}$ in $[0, \frac{1}{2}]^n$ satisfying $D_1 + D_2 + \cdots + D_n \leq nd$. For each such $\mathbf{D}$ we deduce from (4.3), the concavity of $h(\cdot)$ and the fact that

$h(\cdot)$ is monotonic increasing on $[0, \frac{1}{2}]$ that

$$n^{-1}R_X(\mathbf{D}) \geq n^{-1}H(X) - n^{-1}\sum_{i=1}^{n} h_i(D_i) \geq n^{-1}H(X) - h(d). \tag{4.5}$$

On the other hand part (a) of the theorem implies that

$$n^{-1}R_X((d,\ldots,d)) = n^{-1}H(X) - h(d). \tag{4.6}$$

Combining (4.4)–(4.6) proves part (b). □

The specific entropy of a stationary random field X with site space $\mathbb{Z}^k$ is defined as the limit as $n \to \infty$ of $H(X)/n$ as the n sites spread out appropriately to cover the entirety of $\mathbb{Z}^k$. Follmer (1973) has shown that its limit $H_0(X)$ exists for any stationary random field on $\mathbb{Z}^k$. By combining Theorems 1 and 2 and taking limits, we obtain

$$\overline{R}_X(d) = H_0(X) - h(d), \qquad 0 \leq d \leq \delta \tag{4.7}$$

for any stationary binary random field X with site space $\mathbb{Z}^k$ that satisfies assumptions A.1–A.4, where δ is given by (3.8). The critical distortion is defined as

$$d_c = \sup\{d\colon \overline{R}_X(d) = H_0(X) - h(d)\}.$$

Ising's celebrated model of a ferromagnet over $\mathbb{Z}^2$ with no external magnetic field is a Markov random field with

$$\pi_{\{i\},\eta}(v) = \frac{\exp\left(\gamma(-1)^v \Sigma_{j\in N(i)}(-1)^{\eta_j}\right)}{\exp\left(\gamma\Sigma_{j\in N(i)}(-1)^{\eta_j}\right) + \exp\left(-\gamma\Sigma_{j\in N(i)}(-1)^{\eta_j}\right)}, \tag{4.8}$$

where γ is a positive constant. For this example, $D_0 = e^{-4\gamma}/(e^{-4\gamma} + e^{4\gamma})$. Onsager (1944) derived the formula for $H_0(x)$ for this model. Hence, (4.7) yields the exact formula for the per-site rate-distortion function of the homogeneous two-dimensional Ising model for $0 \leq d \leq \delta \leq d_c$; the exact value of d_c remains unknown. Since each site has 4 neighbors in the two-dimensional Ising model, M of Section 3 is 5 so (3.8) gives us $d_c \geq g(4)$. However, the special structure of $\mathbb{Z}^2$ permits us to cover any finite set of sites sequentially such that the following is true: At most two of the neighbors of any site are chosen before the site is chosen, and at most two of them are chosen after the site is chosen. It follows from detailed inspection of the anatomy of Lemma 7 that $d_c \geq g(3)$ even when $M = 5$ is replaced by $M = 3$ in the definition of $g(k)$. Therefore,

$$d_c \geq \left[1 - (1 - 2D_0)^{1/27}\right]\Big/2 = \left[1 - \tanh(4\gamma)^{1/27}\right]\Big/2 \tag{4.9}$$

for the two-dimensional Ising model. For all other interaction models over $\mathbb{Z}^k$, $k \geq 2$, only approximations to $H_0(X)$ are known, so we are unable to evaluate (4.7) analytically.

Gray (1971) was the first to prove that (4.7) always holds for some $d_c > 0$ in the special case of a finite-alphabet homogeneous Markov chain on $\mathbb{Z}$ possessing

a strictly positive one-step transition matrix. Furthermore, in the special case of a binary symmetric Markov chain with $m = \min(p, 1-p)$, where $p = P[X_{i+1} = X_i]$, Gray (1970) derived the exact formula

$$d_c = \left[1 - \left(1 - (m/1-m)^2\right)^{1/2}\right]\Big/2. \tag{4.10}$$

In addition, Gray (1970, 1971) and Avram and Berger (1985) have calculated and/or bounded d_c for several other classes of random sequences on $\mathbb{Z}$.

The binary symmetric Markov source is, in essence, the random field that statistical physicists refer to as the one-dimensional Ising model with no external magnetic field. It is instructive to compare the exact value of d_c for the one-dimensional Ising model with the lower bound provided by the theory developed in Section 3. Since M is 3 for the one-dimensional Ising model, Theorem 1 gives $d_c \geq g(2)$. However, by covering the sites sequentially according to the usual one-dimensional indexing, we see from the proof of Lemma 7 that we can reduce the effective value of M from 3 to 2 without disturbing the validity of $d_c \geq g(2)$. Therefore,

$$\begin{aligned} d_c &\geq \frac{1}{2}\left(1 - (1 - 2D_0)^{1/2^2}\right) \\ &= \frac{1}{2}\left(1 - \left(1 - 2\frac{m^2}{m^2 + (1-m)^2}\right)^{1/4}\right) \\ &= \frac{1}{2}\left(1 - \left(\frac{(1-m)^2 - m^2}{(1-m)^2 + m^2}\right)^{1/4}\right). \end{aligned} \tag{4.11}$$

It is straightforward to show that the bound given by (4.11) is indeed strictly less than the true d_c of (4.10) in all but the trivial cases $m = \frac{1}{2}$, in which they both equal $\frac{1}{2}$, and $m = 0$, in which they both equal 0.

A strengthened version of Lemma 6 can be proved which yields the correct value of d_c for the one-dimensional Ising model. This strengthened lemma concerns for which values of D_1 and D it is true that, if both $(D_0, 0)$ and $(0, D_1)$ can be extracted from the joint distribution of any pair of binary random variables, then (D_1, D) also can be extracted therefrom. It can be shown that $d_c \geq D$ for any value of D that meets this condition for some D_1, and that for the one-dimensional Ising model the value of d_c given by (4.10) is indeed such a D. Similarly, it can be shown that d_c for the two-dimensional Ising model is greater than or equal to the largest value of D for which there exist D_1 and D_2 such that, if $(D_0, 0, 0)$ and $(0, D_1, D_2)$ both are extractable from the joint distribution of any three binary random variables, then (D_1, D_2, D) is extractable therefrom. We conjecture that the maximum such D actually equals d_c. It does not appear to be a simple matter to calculate said maximum D analytically, but it should be feasible to evaluate it numerically.

The exact formula for $\overline{R}_X(d)$ for the case in which X is a stationary Gaussian random field over either $\mathbb{Z}^k$ or $\mathbb{R}^k$ and distortion is measured by quadratic error

stems back to work by Kolmogorov (1956) and was first displayed explicitly for $k \geq 2$ by Hayes, Habibi and Wintz (1970).

5. Concluding remarks. From the viewpoint of communication theory, our representation theorem provides a sufficient condition for a discrete-parameter binary random field to be viewed as the output of a communication channel that adds binary, white, possibly nonstationary noise modulo 2 at each space-time index. It clearly is possible to generalize to certain nonbinary alphabets which support appropriate definitions of additive noise. The generalization to Hamming noise on channels from GF(q) to GF(q) is particularly straightforward.

We close with some speculations about the possible significance of our results in probability theory and in statistical physics. Suppose one were able to determine the maximum possible weights D_i for the U_i in a decomposition $\{X_i\} = \{Y_i \oplus U_i\}$ of the sort we have been considering. Then the associated random field $\{Y_i\}$ would exhibit in relatively transparent form the long-range dependency structure of the field $\{X_i\}$, since the field $\{U_i\}$ possesses no memory. Hence, for models of interest in statistical physics or in other areas of application, the indecomposable process $\{Y_i\}$ associated with the maximal-weight $\{U_i\}$ in our decomposition should be an object worthy of study by probabilists concerned with the regularity of memory in random fields. Such decompositions also may prove helpful in establishing the effective equivalence of certain models from the viewpoint of their ability to support long-range order. That is, by decomposing a random field that has been proposed as a mathematical model of a phenomenon of interest in statistical physics into a portion that is innately memoryless and, independent of that, a portion that is responsible for the model's long-range memory, we should be better able to assess the model's suitability for the task at hand.

Motivated by the suggestion of Berger and Bonomi (1984) that such problems merit study, Bassalygo and Dobrushin (1987) have independently pursued the problem addressed here. Using cluster expansion techniques they have proved the existence of a decomposition of the desired sort for quite general additive noises of sufficiently low weight, provided the temperature parameter in their Gibbs interaction potential over $\mathbb{Z}^k$ is large and the number of terms in the potential to which any site contributes is uniformly bounded. In recent work Newman (1987) has uncovered interesting links between our decomposition theorem and Lee–Yang theory in statistical mechanics and has obtained improved estimates of d_c in certain instances.

REFERENCES

AVRAM, F. and BERGER, T. (1985). On critical distortion for Markov sources. *IEEE Trans. Inform. Theory* **IT-31** 688–690.

BASSALYGO, L. A. and DOBRUSHIN, R. L. (1987). ε-entropy of the Gibbs field. *Problemy Peredači Informacii*. To appear.

BERGER, T. (1971). *Rate Distortion Theory*. Prentice-Hall, Englewood Cliffs, N.J.

BERGER, T. and BONOMI, F. (1984). Coding for the Ising source and Ising channel. In *Proc. Sixth Internat. Symp. Inform. Theory, Tashkent, Uzbekistan, USSR, September 18–22, 1984*. Abstracts of papers, Part II, 271–272.

DOBRUSHIN, R. L. (1968). The description of a random field by means of conditional probabilities and conditions for its regularity. *Theory Probab. Appl.* **13** 197–224.

FOLLMER, H. (1973). On entropy and information gain in random fields. *Z. Wahrsch. verw. Gebiete* **26** 207–217.

GALLAGER, R. G. (1968). *Information Theory and Reliable Communication*. Wiley, New York.

GRAY, R. M. (1970). Information rates of autoregressive processes. *IEEE Trans. Inform. Theory* **IT-16** 412–421.

GRAY, R. M. (1971). Rate distortion functions for finite-state finite-alphabet Markov sources. *IEEE Trans. Inform. Theory* **IT-17** 127–134.

HAYES, J. F., HABIBI, A. and WINTZ, P. (1970). Rate distortion function for a Gaussian source model of images. *IEEE Trans. Inform. Theory* **IT-16** 507–509.

KOLMOGOROV, A. N. (1956). On the Shannon theory of information transmission in the case of continuous signals. *IRE Trans. Inform. Theory* **IT-2** 102–108.

LANFORD, D. E. and RUELLE, D. (1969). Observables of infinity and states with short range correlations in statistical mechanics. *Comm. Math. Phys.* **13** 194–215.

NEWMAN, C. (1987). Decomposition of binary random fields and zeros of partition functions. *Ann. Probab.* **15** 1126–1130.

ONSAGER, L. (1944). Crystal statistics: A two-dimensional model with order–disorder transitions. *Phys. Rev.* **65** 117–147.

PRESTON, C. J. (1974). *Gibbs States on Countable Sets*. Cambridge Tracts, Cambridge Univ. Press, Cambridge.

SCHOOL OF ELECTRICAL ENGINEERING
308 PHILLIPS HALL
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING AND
COORDINATED SCIENCE LABORATORY
UNIVERSITY OF ILLINOIS, URBANA-CHAMPAIGN
1101 WEST SPRINGFIELD AVENUE
URBANA, ILLINOIS 61801