Lossless Reduced Cutset Coding of Markov Random Fields

Matthew G. Reyes
EECS Dept., Univ. of Michigan
Ann Arbor, MI 48109, USA
mgreyes@umich.edu

David L. Neuhoff
EECS Dept., Univ. of Michigan
Ann Arbor, MI 48109, USA
neuhoff@umich.edu

Abstract

This paper presents Reduced Cutset Coding, a new Arithmetic Coding (AC) based approach to lossless compression of Markov random fields. In recent work [14], the authors presented an efficient AC based approach to encoding acyclic MRFs and described a Local Conditioning (LC) based approach to encoding cyclic MRFs. In the present work, we introduce an algorithm for AC encoding of a cyclic MRF for which the complexity of the LC method of [14], or the acyclic MRF algorithm of [14] combined with the Junction Tree (JT) algorithm, is too large. For encoding an MRF based on a cyclic graph $G = (V, E)$, a cutset $U \subset V$ is selected such that the subgraph G_U induced by U, and each of the components of $G \setminus U$, are tractable to either LC or JT. Then, the cutset variables X_U are AC encoded with coding distributions based on a reduced MRF defined on G_U, and the remaining components $X_{V \setminus U}$ of X_V are optimally AC encoded conditioned on X_U. The increase in rate over optimal encoding of X_V is the normalized divergence between the marginal distribution of X_U and the reduced MRF on G_U used for the AC encoding. We show this follows a Pythagorean decomposition and, additionally, that the optimal exponential parameter for the reduced MRF on G_U is the one that preserves the moments from the marginal distribution. We also show that the rate of encoding X_U with this moment-matching exponential parameter is equal to the entropy of the reduced MRF with this moment-matching parameter. We illustrate the concepts of our approach by encoding a typical image from an Ising model with a cutset consisting of evenly spaced rows. The performance on this image is similar to that of JBIG.

I. Introduction

Markov random fields (MRFs) are probability distributions on undirected graphs and are members of the broader class of graphical models [10], [17]. MRFs are often proposed as natural models for spatially distributed data such as images [4], [7], [18]. They have been well studied from the points of view of statistical inference, structure learning and parameter estimation [4], [10], [11], [17], [18]. Nevertheless, there has been little development of data compression algorithms for MRFs. Indeed the only work of which the authors are aware are the lossy image compression methods of [9], [12], [13], and the lossless Arithmetic Encoding (AC) based approach of [14].

When applying AC to an image, the principal issues are the ordering of the pixels for encoding and the determining of coding distributions to provide to the arithmetic

This work was supported by NSF Grant CCF 0830438.

1068-0314/10 $26.00 © 2010 IEEE
DOI 10.1109/DCC.2010.41

encoder. In [14], the authors introduced an efficient approach to AC coding for acyclic MRFs, and for cyclic graphs, they introduced an approach based on a new kind of Local Conditioning [6], a variant of the Belief Propagation (BP) algorithm [11]. For cyclic graphs, one could also combine the Junction Tree (JT) algorithm [8] with the method of [14] for encoding acyclic graphs. For either the LC or JT based methods AC encoding methods, the complexity of computing the optimal *coding distributions* is exponential in the graph *treewidth* for the JT algorithm, the size of the largest *relevant cutset* for the LC algorithm. If either the JT or LC algorithms can be used efficiently, we say the graph is *tractable*.

In Section III of this paper, we discuss Reduced Cutset Coding (RCC), a new AC based approach to losslessly compressing an MRF defined on an intractable graph $G = (V, E)$, where the *nodes* (also called *pixels*) in V are the random variable indices and the edges in E represent direct dependencies between the random variables. The method first losslessly encodes a cutset $U \subset V$, chosen so that if we remove the nodes in U and the edges touching them, then the remaining nodes are partitioned into tractable subgraphs. This allows us to efficiently and optimally encode the remaining nodes in $V \setminus U$ conditioned on the random subfield X_U on the cutset. Since computing optimal coding distributions for X_U is rather complex, the approach of this paper is to use coding distributions based on an MRF defined on the subgraph G_U induced by U, which we refer to as a *reduced MRF* on G_U. For this reason, the subgraph G_U should itself be computationally tractable. Since the conditional coding of the remaining components is optimal given the values on the cutset, the encoding rate of this method exceeds that of optimal encoding by the divergence between the marginal distribution of X_U and the *reduced MRF* normalized by the size of V.

In Section III-C, we show that the divergence between the marginal distribution of X_U and a reduced MRF used as coding distribution for X_U follows a Pythagorean decomposition. This allows us to argue that the *moment-matching* reduced MRF on G_U is optimal in terms of minimizing the rate increase introduced by RCC. The optimal rate of encoding X_U is the marginal entropy of X_U and it is known that this is upper bounded by the entropy of the moment-matching reduced MRF on G_U [5], [15]. We show in Section III-D that the rate obtained by encoding X_U with coding distribution based on the moment-matching reduced MRF is exactly the entropy of the moment-matching reduced MRF. The Pythagorean decomposition and said entropy result are reduced MRF analogues of standard results from information geometry [1], [17].

In the final section of this paper we present experimental results from encoding an Ising model on an $N \times N$ grid graph with uniform coupling parameter and no external field. We find experimentally that the encoding rate is indeed close to the entropy-rate of the Ising MRF, implying that the divergence between the marginal distribution of X_U and the reduced MRF is small. We also compare RCC with JBIG [16] on a typical Ising image and find that the encoding rates are very similar for the given example.

The next section provides the necessary background on MRFs, BP and AC.

II. Background

A. Markov Random Fields

We consider a Markov random field on a graph $G = (V, E)$, where V is a set of N *nodes* (also called *pixels*), and E is a set of undirected *edges*, each connecting a pair of elements of V. A *path* in a graph is a sequence of nodes, each successive pair joined by an edge in E. A graph is said to be *connected* if every pair of nodes $i, j \in V$ can be joined by some path, and *disconnected* otherwise. For any $U \subset V$, its *boundary* ∂U is the set of nodes not in U connected by an edge to a member of U. As a shorthand, ∂i denotes $\partial\{i\}$, $i \in V$. For a subset U, we let $E_U \subset E$ be the subset of edges both of whose endpoints are contained in U. Then, the graph $G_U = (U, E_U)$ is the *subgraph induced by* U. The graph $G \setminus U$ is obtained by removing U and all edges incident to it from G. If $G \setminus U$ is disconnected, each maximal connected subset $C_i \subset V$ of nodes is called a *component*, and $G \setminus U$ is simply the collection of the (disjoint) subgraphs $\{G_{C_i}\}$ induced by the respective components. A subset $U \subset V$ is said to *separate* two other subsets $C_1, C_2 \subset V$ if C_1 and C_2 are contained in distinct components of $G \setminus U$.

A random variable X_i, taking values in a common alphabet $\mathcal{X}$, is associated with each node i. A family of MRFs is specified by a vector statistic $t = (t_i, i \in V; t_{i,j}, \{i, j\} \in E)$ defined on the values at individual nodes and at the endpoints of the edges.[1] That is, for a given image $\mathbf{x} = \{x_i : i \in V\}$, the function $t_{ij} : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ determines the contribution of the pair (x_i, x_j) to the probability of $\mathbf{x}$, and similarly for $t_i : \mathcal{X} \longrightarrow \mathbb{R}$. Specifically, an MRF $\mathbf{X}$ based on t is generated by introducing an exponential parameter (vector) $\theta = (\theta_i, i \in V; \theta_{ij}, \{i, j\} \in E)$ and the specification that the probability distribution $p_G(\theta)$ for $\mathbf{X}$ has the exponential Gibbs form

$$p_G(\mathbf{x}; \theta) \quad = \quad \frac{1}{Z(\theta)} \exp\{\langle \theta, t(\mathbf{x}) \rangle\}, \tag{1}$$

where $\langle\ ,\ \rangle$ denotes inner product, $Z(\theta)$ is the partition function, and the subscript G on p indicates the graph on which the MRF is defined. For each node i and neighbor $j \in \partial i$, θ_i and θ_{ij} scale the sensitivity of the distribution $p_G(\theta)$ to t_i and t_{ij}, respectively. The set $\Theta(G) = \{\theta \in \mathbb{R}_+^{|V|+|E|}\}$ is the set of *admissable exponential parameters* for MRFs on G, while $\mathcal{F}(G) = \{p_G(\theta) \mid \theta \in \Theta(G)\}$ is the family of all MRFs on G based on t. The set $\Theta = \Theta(G)$ is a coordinate system for MRFs in $\mathcal{F} = \mathcal{F}(G)$. For a subset of nodes $U \subset V$ we let t_U and θ_U be the components of θ and t, respectively, corresponding to nodes and edges in U. We will sometimes partition an exponential parameter θ in the form $\theta = (\theta_U, \setminus\theta_U)$, where $\setminus\theta_U$ is the complement of θ_U in θ.

Given an exponential coordinate vector θ, we let $\mu = \mu(\theta)$ denote the expected value of the statistic t under the MRF induced by θ, and we refer to μ as the *moment* parameter. The set of moment parameters arising from exponential parameters in $\Theta(G)$ is $\mathcal{M}(G)$, which is referred to as the set of *achievable moment parameters* for MRFs on G based on t. If the components of t are affinely independent, t is called *minimal*, and the mapping between Θ and $\mathcal{M} = \mathcal{M}(G)$ is one-to-one. Therefore, $\mathcal{M}$ provides a second coordinate system for MRFs in $\mathcal{F}$ [1]. We can then index an MRF p by either the exponential parameter θ or the corresponding moment parameter μ, which can similarly be expressed in partitioned form as $\mu = (\mu_U, \setminus\mu_U)$. Moreover, we can index an MRF p in mixed notation, for example, $p \sim (\mu_U; \setminus\theta_U)$.

For a subset of nodes $U \subset V$, the marginal distribution of X_U for the original MRF on G under exponential parameter $\theta \in \Theta(G)$ is denoted $p_G^U(x_U; \theta)$. We can also consider MRF distributions for X_U. We say that an MRF on the subgraph G_U based on t_U is a *reduced MRF*. Specified in exponential coordinates $\theta_U \in \Theta(G_U)$, the reduced MRF

[1] Properly, this is a *pairwise* MRF. Generalizations to other MRFs are straightforward.

distribution is denoted $p_{G_U}(\theta_U)$ and has the form given in (1). If t is minimal for the family of MRFs on G, the subvector t_U is minimal for the family of MRFs $\mathcal{F}(G_U)$ on the induced subgraph G_U, so $\Theta(G_U)$ and $\mathcal{M}(G_U)$ are dual coordinate systems for MRFs on G_U. For $\mu_U \in \mathcal{M}(G_U)$, the reduced MRF on G_U is denoted $p_{G_U}(\mu_U)$. The entropy of an MRF $p_G(\theta)$ will be denoted by $H_G(X;\theta)$, or $H_G(\theta)$ for short, and the marginal entropy of a subfield X_U, $U \subset V$, by $H_G^U(X;\theta)$ or $H_G^U(\theta)$.

We can express the probability distribution of an MRF in product form by introducing, for each node i and each edge $\{i,j\}$, the *self-* and *edge-potentials* $\Phi_i \overset{\Delta}{=} \exp\{\theta_i t_i\}$ and $\Psi_{i,j} \overset{\Delta}{=} \exp\{\theta_{ij} t_{ij}\}$, respectively. Suppressing now the dependence on θ, we write

$$p_G(\mathbf{x}) \;=\; \frac{1}{Z} \prod_{\{i,j\} \in E} \Psi_{i,j}(x_i, x_j) \prod_{i \in V} \Phi_i(x_i). \tag{2}$$

It is straightforward to show using (1) or (2) that $\mathbf{X}$ is *Markov* with respect to G in the sense that for any two subsets $C_1, C_2 \subset V$ of nodes separated by a third subset U, then random subfields X_{C_1} and X_{C_2} are conditionally independent of each other given the values on X_U [10]. Depending on whether the underlying graph for an MRF is cyclic or acyclic, we say that we have a *cyclic* or *acyclic* MRF, respectively.

B. Arithmetic Encoding

To apply Arithmetic Encoding (AC) to an MRF, we first order or *scan* the nodes, i.e., arrange them into a one-dimensional sequence, $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. Then for $i = 1, \ldots, N$, the ith node value x_i is fed to the arithmetic *encoder* along with a *coding distribution* f_i, which is a function $f_i : \mathcal{X} \to [0,1]$, $\sum_{x \in \mathcal{X}} f_i(x) = 1$. Ordinarily, f_i will also depend on some or all of the previous pixel values $x_1^{i-1} = (x_1, \ldots, x_{i-1})$, but this is not reflected in the notation. The encoder outputs a sequence of bits, referred to as a *codeword* for the image $\mathbf{x}$, whose length is denoted by $l(\mathbf{x})$. The decoder uses a prefix of the codeword to decode x_1^{i-1}, and uses subsequent bits and f_i (which it can compute since $x_1, \ldots, x_{i-1}$ are known) in the decoding of x_i. In a slight abuse of notation we refer to the distribution $f(\mathbf{x}) \overset{\Delta}{=} \prod_{i=1}^N f_i(x_i)$ as *the coding distribution*.

We let $q_{i|*}(x_i | x_1^{i-1})$ denote the conditional probability, under distribution q, that the ith scanned node assumes value x_i when the preceding $i-1$ symbols have the value x_1^{i-1}. If $f_i = q_{i|*}$ for all i, the encoding is said to be *optimal for* q. If the true distribution is p and the encoding is optimal for q, it is well-known [5] that

$$\mathbb{E}\, l(\mathbf{X}) \;\approx\; H(\mathbf{X}) + D(p||q) \;,$$

where $\mathbb{E}\, l(\mathbf{X})$ denotes expected value of codeword length and $D(p||q)$ is the KL-divergence between distributions p and q. This implies that optimal lossless compression of an MRF corresponds to exact inference in the MRF.

C. Belief Propagation

When the MRF is defined on an acyclic graph G, *Belief Propagation* (BP) [11] can be used to compute the optimal coding distributions. We first form a tree rooted at the first node to be scanned, and choose a scan order such that for each $i = 2, \ldots, |V|$, the parent $\pi(i)$ of the i-th node has already been scanned. Then, each leaf node j sends to its parent $\pi(j)$ the message component

$$m_{j\to\delta j}(x_{\delta j}) \;=\; \sum_{x_j} \Phi_j(x_j)\Psi_{j,\delta j}(x_j, x_{\delta j}) \tag{3}$$

for each $x_{\delta j} \in \mathcal{X}$. Each non-leaf node j except for the root node passes to its parent $\pi(j)$ the message $m_{j\to\pi(j)}$ with components determined by

$$m_{j\to\pi(j)}(x_{\pi(j)}) = \sum_{x_j} \Psi_{j,\pi(j)}(x_j, x_{\pi(j)})\Phi_j(x_j) \prod_{k\in\partial\sigma(j)} m_{k\to j}(x_j) \;,$$

with the rule that node j does not form the message to send to $\pi(j)$ until receiving incoming messages from all of its children.

Once the root node receives messages from all of its children, it can compute its coding distribution as $p_{1|*}(x_1) \propto \prod_{k\in\sigma(1)} m_{k\to 1}(x_1)$. Then, moving through the scan the optimal coding distribution for the i-th node in the scan is computed as [14]

$$p_{i|*}(x_i \mid x_{\pi(i)}) \propto \Psi_{\pi(i),i}(x_{\pi(i)}, x_i) \prod_{j\in\sigma(i)} m_{j\to i}(x_i). \tag{4}$$

The potential $\Psi_{\pi(i),i}(x_{\pi(i)}, x_1)$ in (4) can be viewed as a message from $\pi(i)$ to i that is conditioned on the observed value of $x_{\pi(i)}$

For a cyclic graph, one can perform optimal AC encoding by modifying the Junction Tree (JT) algorithm [8]. In the JT algorithm, nodes are grouped into supernodes to form an acyclic graph of supernodes, called a *junction tree*. Thus to encode a cyclic MRF, we could form the junction tree, then follow the procedure described above for acyclic graphs. The difference is that we will encode the variables in groups, as supernodes, and that a given supernode may have a non-empty intersection with its parent supernode. One can also AC encode a cyclic MRF using Local Conditioning (LC) [6], [14], another variant of belief propagation that permits exact inference in cyclic graphs. In LC, certain nodes are split into multiple copies to remove all cycles, and each message is now a matrix where each column is a message conditioned on a given configuration of the relevant set for that edge. Again, the same two-stage process as described above can be used, with some modifications being required to account for the multiple copies of some nodes.

III. Reduced Cutset Coding

We assume a Markov random field $\mathbf{X}$ with distribution p_G, indexed by either a fixed but arbitrary exponential vector $\theta \in \Theta$ or the corresponding moment vector $\mu = \mu(\theta)$. As discussed earlier, with RCC we choose a cutset U, encode X_U, and then conditionally encode $X_{V\setminus U}$ given X_U. The resulting encoding rate R, defined as the expected codeword length normalized by the number of nodes in V, decomposes into

$$R = \frac{|U|}{|V|} R_U + \frac{|V \setminus U|}{|V|} R_{V\setminus U} \;,$$

where R_U and $R_{V\setminus U}$ are the rates of encoding the subfields X_U and $X_{V\setminus U}$, respectively.

A. Encoding the Cutset

The cutset U should be chosen so the induced subgraph G_U is tractable to either JC or LC, in which case we can efficiently compute the coding distribution $f = p_{G_U}(\theta^c_U)$, where θ_U^c is an exponential parameter chosen for the encoding of X_U. In this case

$$R_U = \frac{1}{|U|}\left[H_G^U(X_U) + D(p_G^U(\theta)||p_{G_U}(\theta^c))\right]. \tag{5}$$

B. Conditional Component Coding

The cutset U should be chosen so that the disjoint subgraphs $G_{C_1}, G_{C_2}, \ldots, G_{C_K}$, induced by the components $C_1, \ldots, C_K$ of $G \setminus U$, are tractable to either JC or LC. We first note that for any subset $C \subset V$ of nodes, the conditional probability $p_G(x_C|x'_{\partial C}; \theta)$ of realization x_C given the configuration $x'_{\partial C}$ on its boundary can be expressed as

$$p_G(x_C|x_{\partial C}; \theta) \quad \propto \quad \prod_{\{i,j\}\in E_C} \Psi_{i,j}(x_i, x_j) \prod_{i\in C} \Phi_i(x_i) \prod_{k\in\partial i\setminus C} \Psi_{i,k}(x_i, x'_k).$$

In other words, it is a reduced MRF distribution on G_C with the only modifications to the original potentials being updates to the self-potentials on those nodes on the surface of C, which take into account the boundary values $x_{\partial C}$. Therefore, if each of the disjoint subgraphs of $G \setminus U$ is tractable, then LC or JT can be used to optimally encode $X_{V\setminus U}$ conditioned on X_U. Since the components are conditionally independent given the values on their boundaries, the rate of encoding the remainder $X_{V\setminus U}$ is

$$R_{V\setminus U} \quad = \quad \frac{1}{|V \setminus U|} H_G^{V\setminus U}(X_{V\setminus U}|X_U) \quad = \quad \frac{1}{|V \setminus U|}\sum_{i=1}^{K} H_G^{C_i}(X_{C_i}|X_{\partial C_i}).$$

C. Pythagorean Decomposition

Given an MRF on G specified by $\theta \in \Theta(G)$ and corresponding $\mu \in \mathcal{M}(G)$, consider the exponential parameter $\bar{\theta} = (\theta_U, 0)$, which induces an MRF $p_G(\bar{\theta})$. Setting to zero those exponential parameters corresponding to nodes and edges not contained in U has the effect of isolating each node $V \setminus U$ from all others. We can analyze the divergence in (5) by considering the submanifold

$$\mathcal{F}' \quad = \quad \{p' \in \mathcal{F}(G) : \backslash\theta'_U = 0\}$$

of MRFs on the graph consisting of G_U and $|V \setminus U|$ isolated nodes[2] and the submanifold

$$\mathcal{F}'' \quad = \quad \{p'' \in \mathcal{F}(G) : \mu''_U = \mu_U\}$$

of all MRFs whose moment coordinates μ''_U for edges and nodes in U are equal to the corresponding coordinates μ_U from the original MRF $p \sim \mu$. It is known that $\mathcal{F}'$ and $\mathcal{F}''$ intersect uniquely at the MRF with exponential parameter $\theta^* = (\theta^*_U, 0)$ that corresponds to mixed coordinates $(\mu_U; 0)$, and that the MRF $p_G(\theta^*)$ is referred to as the *m-projection* of $p_G(\theta)$ onto the $\mathcal{F}'$ [1]. Given a subset U and a distinct exponential parameter $\theta' \in \mathcal{F}'$, the divergence between $p_G(\theta)$ and $p_G(\theta')$ can be decomposed as [1]

$$D(p_G(\theta)||p_G(\theta')) = D(p_G(\theta)||p_G(\theta^*)) + D(p_G(\theta^*)||p_G(\theta')). \tag{6}$$

This is a well-known Pythagorean relation of information geometry. One may conclude from it that of all MRFs $p_G(\theta')$ in the submanifold $\mathcal{F}'$, the one with minimum reverse

[2] Some of the remaining coordinates of θ for edges in U can be set to zero, thus creating a proper subgraph of G_U.

divergence with the original MRF $p_G(\theta)$ is the moment-matching MRF $p_G(\theta^*)$.

We can use the above Pythagorean formula to simplify the divergence in (5) by first noting that MRFs in $\mathcal{F}'$ have a particularly simple form. Specifically, if $\theta' = (\theta'_U, 0)$ is the exponential parameter for an MRF in the submanifold $\mathcal{F}'$, then $p_G(\theta')$ is simply the product of the reduced MRF $p_{G_U}(\theta'_U)$ and the product of the independent, uniform distributions for the isolated nodes. As a result, the divergence between two MRFs $p_G(\theta')$ and $p_G(\theta'')$ in $\mathcal{F}'$ can be decomposed into the divergence between the reduced MRFs $p_{G_U}(\theta'_U)$ and $p_{G_U}(\theta''_U)$, plus the divergence between the two independent parts. But the independent parts are identical so the divergence is simply the divergence between the two reduced MRFs, as summarized in the following.

Lemma 3.1: If $\theta', \theta'' \in \Theta(G)$ index distinct MRFs $p(\theta'), p(\theta'') \in \mathcal{F}'_U$, then

$$D(p_G(\theta')||p_G(\theta'')) = D(p_{G_U}(\theta'_U)||p_{G_U}(\theta''_U)). \tag{7}$$

Now, due to the simple form of MRFs in $\mathcal{F}'$, we can decompose the divergence between the original exponential parameter θ and an exponential parameter θ' in $\mathcal{F}'$.

Lemma 3.2: Let $\theta \in \Theta(G)$ be given, and let θ' index an MRF $p_G(\theta') \in \mathcal{F}'$. Then,

$$D(p_G(\theta)||p_G(\theta')) = D(p_G^U(\theta)||p_{G_U}(\theta'_U)) + |V\backslash U| \log |\mathcal{X}| - H_G^{V-U|U}(\theta), \tag{8}$$

where $H_G^{V-U|U}(\theta)$ is the conditional entropy of $X_{V\backslash U}$ given X_U in the original MRF.

From the above two lemmas and (6) we derive the following decomposition for the divergence between the marginal distribution $p_G^U(\theta)$ and the reduced MRF $p_{G_U}(\theta'_U)$.

Theorem 3.3: Let $\theta \in \Theta(G)$ and $\mu \in \mathcal{M}(G)$ be associated exponential and moment parameters for MRFs on G, and let $\theta'_U \in \Theta(G_U)$ be an exponential parameter for reduced MRFs on G_U. Then

$$D(p_G^U(\theta)||p_{G_U}(\theta'_U)) = D(p_G^U(\theta)||p_{G_U}(\mu_U)) + D(p_{G_U}(\mu_U)||p_{G_U}(\theta'_U)). \tag{9}$$

This implies that the coding parameter θ_U^c that minimizes the divergence with the marginal distribution, and hence minimizes the increase in rate due to using a reduced MRF coding distribution, is the moment-matching exponential parameter θ_U^*, or in moment coordinates, the subvector μ_U of the original moment parameter $\mu \sim \theta$.

The following theorem summarizes the rates achieved through the reduced cutset coding method introduced.

Theorem 3.4: Let $G = (V, E)$ be an undirected graph on which an MRF $p_G(\theta)$ is defined and let μ be the moment coordinates corresponding to exponential parameter θ. If a subset $U \subset V$ is encoded using coding distribution $p_{G_U}(\theta_U^c)$, then

- $R_U = \frac{1}{|U|}\left[H_G^U(\theta) + D(p_G^U(\theta)||p_{G_U}(\mu_U)) + D(p_{G_U}(\mu_U)||p_{G_U}(\theta_U^c))\right],$
- $R_{V\backslash U} = \frac{1}{|V\backslash U|}\left[H_G^{V\backslash U}(X_{V\backslash U}|X_U)\right].$

D. Notes

In this section we shed further light on the relationship between the optimal rate obtainable by any method for encoding the cutset X_U and the optimal rate obtainable through Reduced Cutset Coding. It is known that $H_G^U(\theta) \leq H_{G_U}(\theta_U^*)$ from the maximum entropy property of MRFs [5], [15]. We showed further in [14] that $H_{G_U}(\theta_U^*) \leq H_{G_U}(\theta_U)$. We are now in a position to quantify the gap of the first inequality. In the last subsection we discussed how, given an exponential parameter θ and a cutset U, the MRF in the submanifold $\mathcal{F}'_U$ with minimum reverse divergence to $p_G(\theta)$ is induced by the moment-matching exponential parameter θ^*. It is well-known [1], [17] and straightforward to show that

$$H_G(\theta^*) = H_G(\theta) + D(p_G(\theta)||p_G(\theta^*)). \tag{10}$$

Again because of their simple form, we can decompose the entropy of MRFs in the submanifold $\mathcal{F}'_U$ in the following way.

Lemma 3.5: Let $\theta' \in \Theta(G)$ index an MRF $p(\theta') \in \mathcal{F}'_U$. Then,

$$H_G(\theta) = H_{G_U}(\theta_U) + |V \setminus U| \log |\mathcal{X}|.$$

For MRFs in $\mathcal{F}'_U$ the random subfields X_U and $X_{V\setminus U}$ are independent, which implies that $H_G(X;\theta) = H_G(X_U : \theta) + H_G(X_{V\setminus U};\theta)$. Using this, the above lemma, and equation (10) leads to the following theorem.

Theorem 3.6: Let $G = (V, E)$ be an undirected graph on which MRF X is defined. For an arbitrary subset $U \subset V$,

$$H_{G_U}(\theta^*_U) = H^U_G(\theta) + D(p^U_G(\theta)||p_{G_U}(\theta^*_U)). \tag{11}$$

Therefore, the rate of encoding the cutset X_U with reduced MRF exponential parameter θ^*_U, the rate of the encoding is the same as if the reduced MRF induced by coding parameter θ^*_U was encoded optimally.

IV. Example: Ising Model on $N \times N$ Grid

In this section we consider RCC in a specific example. The node set V is an $N \times N$ square array of nodes, and the edge set E contains all horizontally and vertically adjacent nodes. As illustrated in Fig. 1(a), the cutset U consists of evenly spaced rows of the graph, and the components of $G \setminus U$ are $(M-1) \times N$ rectangular strips, where M is the spacing between successive rows of the cutset. It is known that the complexity of both JT and LC on an $M \times N$ strip is exponential in M [11], [14]. Thus, the line spacing M should be chosen to have a moderate value, say, 10 or less.

The MRF for this example is an Ising model with uniform interactions in the absence of an external field [3]. More precisely, for each node i the random variable X_i takes values in $\{-1, 1\}$ and $t_i \equiv 0$, and for each edge we assign statistic $t(x_i, x_j) = x_i x_j$ and exponential coordinate $\theta_{ij} = \theta = 0.5$. Using a standard Gibbs Sampling technique [7], we generates the sample $\mathbf{x}_{GS}$ with $N = 421$, shown in Figure 1(b). Instead of using a distinct optimal coding coordinate θ^*_{ij} for each edge of G_U, we approximate this by using a single optimal coding parameter θ^* to parameterize each edge. We searched for θ^* by simply encoding the lines of U with several values of θ^c and chose the one that yields the smallest value of $l(\mathbf{X}_{GS})$, namely, $\theta^* = 1.21$.

Plots of the coding rates, in bits per pixel, for the cutset of lines, the strip, and the total image are shown in Figure 1(c). By the properties of the Gibbs Sampler [7], the image $\mathbf{x}_{GS}$ is *typical* for the Ising model. Hence the rates R_U and $R_{V\setminus U}$ of encoding the lines and strips of $\mathbf{x}_{GS}$ should be very close to the rates specified in Theorem 3.4. We see that the rate R_U for the lines is essentially constant. This is because the same coding distribution is used for each line and because the lines are roughly stationary. In addition, as the line spacing increases, the strip coding rate $R_{V\setminus U}$ increases because the ineriors of the strips have decreasing dependence on X_U. Note further that RCC achieves nearly

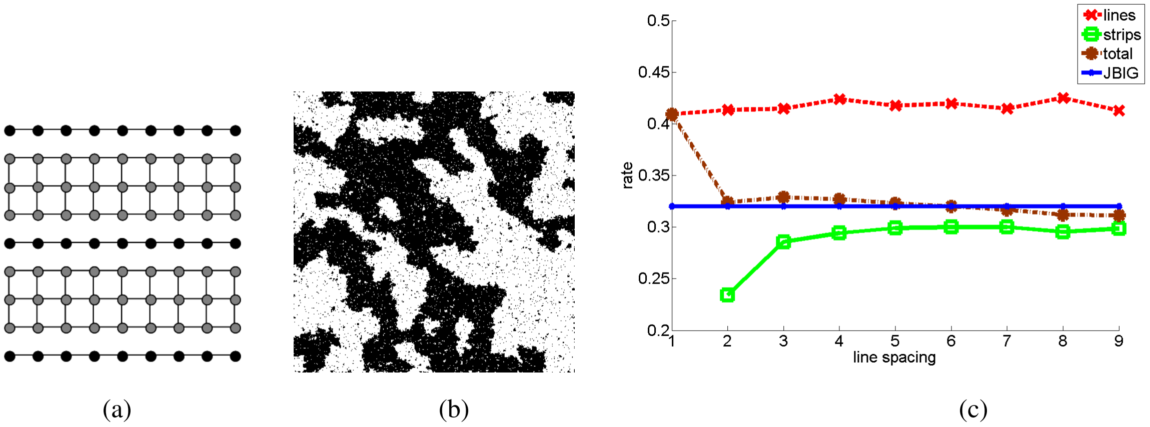

Figure 1. (a) Cutset of disjoint lines (black pixels) and disjoint strip components (gray pixels); (b) Typical image $\mathbf{x}_{GS}$; (c) Coding rates: R_U (crosses), $R_{V\setminus U}$ (squares), total rate R (circles), JBIG (solid line).

its least rate with line spacing two, in which case the cutset consists of every other row, which leads to a very simple encoding of the single-row strips.

For any line spacing, the normalized encoding rate of the strips, $R_{V\setminus U}$, is a lower bound to the entropy-rate of the Ising MRF, and the total encoding rate is an upper bound. From Figure 1(c), we see that these are close and become closer as line spacing increases, which indicates that the coding is very nearly optimal. It also provides a good estimate of the entropy-rate, which is useful because the integral formula for entropy-rate given in [2] is very difficult to compute. Such estimated upper and lower bounds are given in Table I for different values of θ.

For comparison, we encoded $\mathbf{x}_{GS}$ using JBIG, a state-of-the-art bilevel image compression method [16], and found that JBIG performs as well as RCC. This is quite surprising in that RCC is essentially optimal for the MRF, whereas JBIG was not at all designed for MRFs. However, as seen in Figure 1, $\mathbf{x}_{GS}$ is a relatively simple image with large homogenous regions so the that perhaps it is not so surprising that JBIG works well.

We also applied RCC and JBIG to images generated by values of θ ranging from 0.1 to 1, and found similar results. For example, we found line spacing two was essentially as good as any line spacing.

References

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Oxford University Press, 2000.

[2] D. Anastassiou and D.J. Sakrison, "Some Results Regarding the Entropy Rate of Random Fields," *IEEE Trans. Inform. Thy.*, vol. IT-28, pp. 340–343, Mar. 1982.

θ	0.1	0 .2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$H(\theta)$ upper bound	0.988	0.949	0.910	0.687	0.311	0.182	0.124	0.103	0.092	0.086
$H(\theta)$ lower bound	0.988	0.947	0.86	0.676	0.298	0.174	0.119	0.099	0.089	0.083

Table I
Estimated upper and lower bounds to entropy rate.

[3] R.J. Baxter, *Exactly Solved Models in Statistical Mechanics*, New York: Academic, 1982.

[4] J. Besag, "On the Statistical Analysis of Dirty Pictures," *J. of Roy. Stat. Soc. B*, vol. 48, no. 3, pp. 256–302, 1986.

[5] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 2005.

[6] F.J. Diez, "Local conditioning in Bayesian networks," *Artificial Intelligence*, 87, pp. 1–20, 1996.

[7] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Trans. PAMI*, vol. 6, pp. 721–741, Nov. 1984.

[8] F. Jensen, *An Introduction to Bayesian Networks*, UCL Press, London,1996.

[9] I. Kontoyiannis, "Pattern Matching and Lossy Data Compression on Random Fields," *IEEE Trans. Inform. Thy.*, vol. 49, pp. 1047–1051 April 2003.

[10] S. Lauritzen, *Graphical Models*, Oxford University Press, 1996.

[11] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Francisco, CA, 1988.

[12] M. G. Reyes, X. Zhao, D. L. Neuhoff, T. N. Pappas, "Lossy Compression of Bilevel Images Based on Markov Random Fields," *Proc. ICIP*, San Antonio, pp. II-373–376, Sept. 2007.

[13] M. G. Reyes, X. Zhao, D. L. Neuhoff, T. N. Pappas, "Structure-preserving properties of bilevel image compression," HVEI XII, San Jose, CA, pp. 680617–680617, Jan. 2008.

[14] M. G. Reyes and D. L. Neuhoff, "Arithmetic Encoding of Markov Random Fields," *Proc. ISIT*, Seoul, Korea, pp. 532–536, July 2009.

[15] M. G. Reyes and D. L. Neuhoff, "Entropy Bounds for a Markov Random Subfield," *Proc. ISIT*, Seoul, Korea, pp. 309–313, July 2009.

[16] K. Sayood, *Introduction to Data Compression*, Morgan Kaufmann, San Francisco, 2006.

[17] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families and variational inference*, Berkeley Tech. Report 649, Sept. 2003.

[18] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carko Methods: A Mathematical Introduction*. Berlin: Springer-Verlag, 1995.