Cutset Width and Spacing for Reduced Cutset Coding of Markov Random Fields

Matthew G. Reyes*
*self-employed
mgreyes@umich.edu

David L. Neuhoff†
†EECS Dept., University of Michigan
neuhoff@umich.edu

Abstract

In this paper we explore tradeoffs, regarding coding performance, between the thickness and spacing of the cutset used in Reduced Cutset Coding (RCC) of a Markov random field image model [10]. Considering MRF models on a square lattice of sites, we show that under a stationarity condition, increasing the thickness of the cutset reduces coding rate for the cutset, increasing the spacing between components of the cutset increases the coding rate of the non-cutset pixels, though the coding rate of the latter is always strictly less than that of the former. We show that the redundancy of RCC can be decomposed into two terms, a correlation redundancy due to coding the components of the cutset independently, and a distribution redundancy due to coding the cutset as a reduced MRF. We provide analysis of these two sources of redundancy. We present results from numerical simulations with a homogeneous Ising model that bear out the analytical results. We also present a consistent estimation algorithm for the moment-matching reduced MRF for the cutset U.

1 Introduction

A Markov random field (MRF) $X = \{X_i : i \in V\}$ is a collection of random variables on an undirected graph graph $G = (V, E)$, where the nodes[1] in V are the random variable indices and the edges in E represent direct dependencies between the random variables [14], and is often proposed as a model for many sources of data, such as images. A family of MRFs on a graph G is defined by a vector statistic t having a component for each edge and each node. An individual MRF within this family is indicated by an exponential parameter vector θ whose components correspond to the components of t. Since there has been relatively little development of algorithms or theory for the compression of MRFs [2, 7–11, 13], we feel that this is an important problem to consider. In this paper we explore design tradeoffs of the lossless Reduced Cutset Coding method introduced in [10].

Reduced Cutset Coding (RCC) is a two-stage algorithm for lossless compression of an MRF defined on an intractable graph, where tractability is with respect to Belief Propagation (BP) [10, 11, 14]. The method consists, first, of suboptimal lossless encoding of a cutset $U \subset V$, chosen such that the subgraphs G_U and G_W induced by U and $W{=}V \setminus U$, respectively, are tractable. The components of X_U are encoded with Arithmetic Coding (AC) using BP to compute a *reduced MRF* coding distribution. A reduced MRF for X_U is an MRF on the subgraph G_U induced by U, with the statistic

An abbreviated version of this paper has been submitted to ISIT 2016.

[1] We use the terms *nodes*, *sites* and *pixels* interchangeably.

t limited to U, and a possibly different exponential parameter vector $\tilde{\theta}_U$. Secondly, conditioned on the encoded cutset X_U, the component subsets of the remaining variables X_W are encoded conditioned on their respective boundaries, again using AC, with BP used to compute the true conditional coding distributions of the variables in X_W with respect to the original MRF.

The rate of this scheme can be expressed as

$$R \;=\; \frac{\mid U \mid}{\mid V \mid} R_U + \frac{\mid W \mid}{\mid V \mid} R_W, \tag{1}$$

where R_U is the rate in bits per pixel for the cutset U, and likewise R_W for the remainder W. Because G_W is tractable for BP, the conditional coding distributions for the components of X_W can be exactly computed. Thus AC will encode each component on average at its conditional entropy plus an overhead of one or two bits [15]. Since we have in mind the components of W having many pixels, the rate R_W is well-approximated by $\frac{1}{|W|}H(X_W|X_U)$, the ideal coding rate for X_W given X_U. Similarly, since U is tractable for BP, the reduced MRF coding distribution can be computed exactly, and R_U is well-approximated by the (normalized) cross entropy $\frac{1}{|U|}H(X_U\|\tilde{X}_U)$ between the marginal distribution for X_U and the reduced MRF distribution for the same variables, which we denote $\tilde{X}_U$, and which equals the entropy $\frac{1}{|U|}H(X_U)$ of X_U plus the divergence $\frac{1}{|U|}D(X_U||\tilde{X}_U)$ between the true and reduced MRF distributions for X_U.

It follows that the rate of this scheme exceeds the rate of an optimal code, which is

$$\frac{1}{|V|}H(X_U, X_W) = \frac{|U|}{|V|}\frac{1}{|U|}H(X_U) + \frac{|W|}{|V|}\frac{1}{|W|}H(X_W|X_U),$$

by the divergence $\frac{1}{|V|}D(X_U||\tilde{X}_U)$. For a given cutset U, this divergence is minimized by choosing the parameter vector $\tilde{\theta}_U$ to be that which causes the mean of the statistic t_U of the reduced MRF $\tilde{X}_U$ to be the same as the mean of t_U on the marginal X_U of the original MRF X [1, 10, 14]. This is called the *moment-matching parameter* and denoted θ^*_U. In Section 4 we present a consistent algorithm for estimating θ^*_U for a tractable subset U, and as such, for the rest of this paper we let $\tilde{X}_U$ denote this moment-matching reduced MRF. Even when divergence is minimized, one normally expects $\frac{1}{|U|}H(X_U)$ to be larger than $\frac{1}{|W|}H(X_W|X_U)$.

In the present paper we consider an MRF on an $M \times N$ rectangular lattice of sites. The statistic t as well as the parameter θ are both row-invariant, and the image height M is assumed to be very large, so that the sequences of rows of the image are assumed to form a stationary process. The cutset U consists of $k+1$ evenly spaced $n_L \times N$ rectangular regions $L_1, \ldots, L_{k+1}$, referred to as *lines*, so that the k components of G_W are themselves $n_S \times N$ rectangular regions $S_1, \ldots, S_k$, referred to as *strips*. This is an extension of the RCC method of [10], [11], which restricted n_L to be 1.[2] Here, $M = kn_S + (k+1)n_L$, so that lines and strips alternate, beginning with a line and ending with a line. This class of cutsets was chosen to simplify both

[2] Even though now n_L can be larger than one, we continue to use the nomenclature of *lines*.

the algorithm and the analysis. For example, the lines (strips) can be transformed into a simple chain graph by grouping the pixels in each column of a line (strip) into one superpixel. If n_L and n_S are both moderate, for instance at most 10, then BP can be used to perform exact inference efficiently.

An interesting question is how the cutset parameters n_L and n_S affect the individual rates R_U and R_W as well as the weightings of R_U and R_W by the respective sizes of U and W. First, consider R_U. The lines of U are encoded independently with the respective moment-matching reduced MRF coding distributions. From the stationarity assumption, these moment-matching reduced MRFs are the same for each line, and therefore

$$R_U = \frac{1}{|U|} H(X_U || \tilde{X}_U) = \frac{1}{n_L N} H(X_L \| \tilde{X}_L) = \frac{1}{n_L N} H(X_L) + \frac{1}{n_L N} D(X_L \| \tilde{X}_L),$$

where L denotes a block of n_L consecutive rows of the image, X_L is the subset of the MRF on L, and $\tilde{X}_L$ is the same random variables with the moment-matching reduced MRF distribution. Next, by the Markov property and stationarity,

$$R_W = \frac{1}{|W|} H(X_W | X_U) = \frac{1}{n_S N} H(X_S | X_{\partial S}),$$

where S denotes n_S consecutive rows, ∂S denotes the boundary of S, and X_S and $X_{\partial S}$ are the respective subsets of random variables on S and ∂S. Therefore, as a function of line and strip widths, the *per-row rate*[3] is

$$\bar{R}(n_L, n_S) = \frac{(k+1)n_L}{kn_S + (k+1)n_L} \frac{1}{n_L} H(X_L \| \tilde{X}_L) + \frac{kn_S}{kn_S + (k+1)n_L} \frac{1}{n_S} H(X_S | X_{\partial S}).$$

When k is large, this is well approximated by

$$\begin{aligned} \bar{R}(n_L, n_S) &\approx \frac{n_L}{n_L + n_S} \frac{1}{n_L} H(X_L \| \tilde{X}_L) + \frac{n_S}{n_L + n_S} \frac{1}{n_S} H(X_S | X_{\partial S}) \\ &= \frac{n_L}{n_L + n_S} \frac{1}{n_L} \big(H(X_L) + D(X_L \| \tilde{X}_L) \big) + \frac{n_S}{n_L + n_S} \frac{1}{n_S} H(X_S | X_{\partial S}). \end{aligned}$$

Intuitively, as the cutset line width n_L increases, R_U decreases because both $\frac{1}{n_L} H(X_L)$ and the divergence $\frac{1}{n_L} D(X_L || \tilde{X}_L)$ would decrease. However, the fraction of sites $\frac{n_L}{n_L + n_S}$ encoded at the larger R_U rate increases. Hence, there is a potential tradeoff between choosing n_L to be large in order to reduce the cutset rate, and choosing n_L to be small in order to reduce the fraction of sites in the cutset. Similarly, as n_S increases, the fraction of pixels $\frac{n_S}{n_L + n_S}$ encoded at the lower rate increases, but one intuitively expects $\bar{R}_W = \frac{1}{n_S} H(X_S | X_{\partial S})$ to increase. Again, a potential tradeoff.

On the other hand, since the overall rate is $R(n_L, n_S) = \frac{1}{|V|} H(X_V) + \frac{1}{|V|} D(X_U || \tilde{X}_U)$, we see that the divergence term $\frac{1}{|V|} D(X_U || \tilde{X}_U)$ is the redundancy of the code, and

[3] The overall rate is the per-row rate divided by the row width N. From now on, we mainly focus on per-row rate to simplify expressions, and use an overbar to indicate such.

one can therefore focus on what makes it small. Letting $\Delta(n_L, n_S) \overset{\Delta}{=} \frac{1}{|V|} D(X_U || \tilde{X}_U)$ denote the *redundancy of the code*, we will show that the per-row redundancy has the form

$$\begin{aligned} \bar{\Delta}(n_L, n_S) &= \frac{(k+1)n_L}{kn_S + (k+1)n_L} \frac{1}{n_L} D(X_L \| \tilde{X}_L) + \frac{kn_S}{kn_S + (k+1)n_L} \frac{1}{n_S} I(X_{L_i}; X_{L_{i-1}}) \\ &\approx \frac{n_L}{n_L + n_S} \frac{1}{n_L} D(X_L \| \tilde{X}_L) + \frac{n_S}{n_L + n_S} \frac{1}{n_S} I(X_{L_i}; X_{L_{i-1}}) \end{aligned}$$

where $I(X_{L_i}; X_{L_{i-1}})$ is the mutual information between the random variables X_{L_i} on a line and the random variables $X_{L_{i-1}}$ on the previous line. Note that in the above formula for redundancy, which is entirely due to the encoding of the lines, the first term, which we call the *distribution redundancy* is due to use of the reduce MRF coding distribution on each line and the second term, which we call the *correlation redundancy* is due the fact that lines are coded independently. Note also that while the redundancy is entirely due to encoding of the lines, the correlation redundancy depends on the strip width n_S. Moreover, since there is no correlation redundancy in the encoding of the first line, it is appropriate to think of $I(X_{L_i}; X_{L_{i-1}})$ as a penalty per strip. From this viewpoint, one would expect that increasing n_L reduces the divergence per cutset pixel $\frac{1}{n_L} D(X_L || \tilde{X}_L)$, but increases the fraction $\frac{n_L}{n_L + n_S}$ of the image included in the cutset. Hence, it is not clear what is the best value for n_L. Similarly, one would expect that information $I(X_{L_i}; X_{L_{i-1}})$ decreases in n_S, while the fraction of pixels $\frac{n_S}{n_L + n_S}$ increases in n_S. Therefore, it is likewise not clear what n_S should be.

The results of this paper are to show the following results, most of which have been conjectured above. Under the stationarity assumption, the coding rate $R^S_{n_S}$ of a strip increases with n_S, the coding rate $R^L_{n_L}$ of a line decreases with n_L when the moment-matching reduced MRF is used to encode the lines, and $R^S_{n_S} < R^L_{n_L}$ for all choices of n_S and n_L. We also present a consistent estimation algorithm for the moment-matching parameter θ^*_U. We show that the divergence $D(X_U || \tilde{X}_U)$, equivalently the redundancy, can be decomposed into a correlation redundancy due to encoding the lines independently and a distribution redundancy due to approximating the lines as reduced MRFs, and present analysis of these two sources of redundancy. Numerical simulations with an Ising model illustrate the propositions.

In the rest of this paper, Section 2 provides background on MRFs and lossless coding and Section 3 provides an overview Reduced Cutset Coding in the current setting. Section 4 presents an estimation algorithm for θ^*_U, Section 5 establishes the anticipated tradeoffs between cutset thickness and spacing, and finally, Section 6 discusses numerical simulations with an Ising model.

2 Background

We introduce notation for lossless coding of MRFs.

2.1 Graphs and Markov Random Fields

A *path* in a graph $G = (V, E)$ is a sequence of nodes, each successive pair of nodes being 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. The subgraph $G_U = (U, E_U)$ *induced by* U is the graph consisting of nodes and edges contained in U. Likewise, the subgraph $G_{V \setminus U}$ is obtained by removing U and all edges incident to it from G. If $G_{V \setminus U}$ is disconnected, each maximal connected subset of $G_{V \setminus U}$ is called a *component*, and $G_{V \setminus U}$ is simply the collection of the (disjoint) subgraphs induced by the respective components. A subset $U \subset V$ is called a *cutset* if $G_{V \setminus U}$ consists of more than one component.

A family of MRFs is specified by an alphabet $\mathcal{X}$ and a vector statistic $t = (t_i, i \in V; t_{i,j}, \{i, j\} \in E)$ defined on the site values at individual nodes and the endpoints of edges.[4] 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 $t_i : \mathcal{X} \longrightarrow \mathbb{R}$. We say that X is an MRF based on t. The entire family of MRFs based on t is generated by introducing an exponential parameter vector $\theta = (\theta_i, i \in V; \theta_{ij}, \{i, j\} \in E)$ where for each node i, and neighbor $j \in \partial i$, θ_i and θ_{ij} scale the sensitivity of the distribution $p(G; \mathbf{x}; \theta)$ to the functions t_i and t_{ij}, respectively. Specifically, for an MRF X on G based on t with exponential parameter θ, configuration $\mathbf{x}$ has probability $p(G; \mathbf{x}; \theta)$ given by

$$p(G; \mathbf{x}; \theta) \quad = \quad \exp\{\langle \theta, t(\mathbf{x}) \rangle - \Phi(\theta)\}, \tag{2}$$

where $\langle \ , \ \rangle$ denotes inner product, $\Phi(\theta)$ is the *log-partition function*, and the arguments of $p(\cdot; \cdot; \cdot)$ indicate, respectively, the graph on which the MRF is defined, the configuration in question, and the exponential parameter on the graph. For a given 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* of the MRF. The MRF distribution over all configurations is denoted $p(G; X; \theta)$, and the entropy of an MRF is denoted $H(G; X; \theta)$.

The conditional probability of a configuration $\mathbf{x}_W$ on subset $W \subset V$ given the values $\mathbf{x}_U$ on another subset $U \subset V$ is denoted $p(G; \mathbf{x}_W | \mathbf{x}_U; \theta)$. It is straightforward to check that $p(G; \mathbf{x}_W | \mathbf{x}_{\partial W}; \theta) = p(G; \mathbf{x}_W | \mathbf{x}_{V \setminus W}; \theta)$ for all W, $\mathbf{x}_W$, and $\mathbf{x}_{\partial W}$. This is the *Markov Property.* The conditional distributions of random subfield X_W given a specific configuration $\mathbf{x}_{\partial W}$, or on the random subfield $X_{\partial W}$, are denoted $p(G; X_W | \mathbf{x}_{\partial W}; \theta)$ and $p(G; X_W | X_{\partial W}; \theta)$, respectively. Likewise, $H(G; X_W | \mathbf{x}_{\partial W}; \theta)$ and $H(G; X_W | X_{\partial W}; \theta)$ are the respective conditional entropies of X_W given a specific configuration $\mathbf{x}_{\partial W}$ or the random subfield $X_{\partial W}$.

For subset U, the marginal probability distribution on X_U is denoted $p(G; X_U; \theta)$, where $p(G; \mathbf{x}_U; \theta)$ denotes the marginal probability of configuration $\mathbf{x}_U$. The *reduced MRF* distribution for X_U on G_U based on statistic t_U with exponential parameter $\tilde{\theta}_U$ is denoted $p(G_U; X_U; \tilde{\theta}_U)$ and has the same form as in (2), where $\Phi_U(\tilde{\theta}_U)$ denotes

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

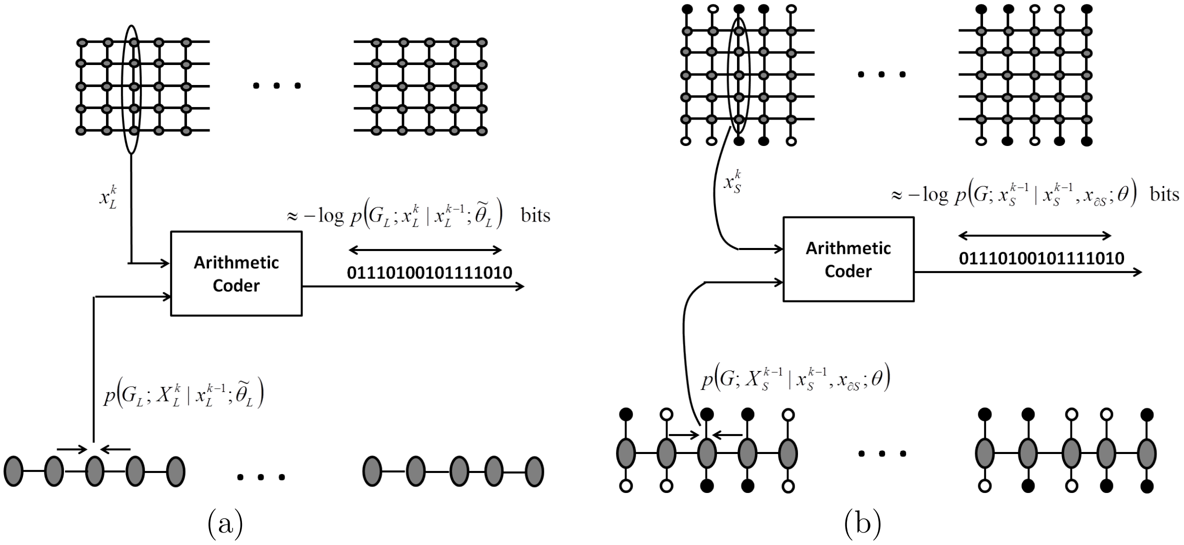

x_L^k
$\approx -\log p\left(G_L; x_L^k \mid x_L^{k-1}; \tilde{\theta}_L\right)$ bits
Arithmetic Coder
01110100101111010
$p\left(G_L; X_L^k \mid x_L^{k-1}; \tilde{\theta}_L\right)$
(a)
x_S^k
$\approx -\log p\left(G; x_S^{k-1} \mid x_S^{k-1}, x_{\partial S}; \theta\right)$ bits
Arithmetic Coder
01110100101111010
$p\left(G; X_S^{k-1} \mid x_S^{k-1}, x_{\partial S}; \theta\right)$
(b)

Figure 1: (a) AC encoding of a line X_L with reduced MRF $P(G_L; X_L; \tilde{\theta}_L)$ coding distribution using Belief Propagation, and (b) AC encoding of a strip X_S conditioned on its boundary $X_{\partial S}$ with conditional distribution $p(G; X_S|X_{\partial S}; \theta)$ using Belief Propagation.

the log-partition function for the reduced MRF. Similarly, $p(G_U; \mathbf{x}_U; \tilde{\theta}_U)$ denotes the probability of configurations $\mathbf{x}_U$ under the reduced MRF distribution. The statistic t_U is inherited from the original statistic t. The marginal entropy of X_U is denoted $H(G; X_U; \theta)$ while the entropy of a reduced MRF $p(G_U; X_U; \tilde{\theta}_U)$ is denoted $H(G_U; X_U; \tilde{\theta}_U)$.

2.2 Belief Propagation and Lossless Coding

In general, ones uses Belief Propagation (BP) [14] to compute $p(G; \mathbf{x}_U; \theta)$ for a configuration $\mathbf{x}_U$. Since the inner product $\langle t_U(\mathbf{x}_U), \theta_U \rangle$ can be computed directly, BP is used to compute the log-partition function $\Phi(\theta)$, and more generally, to marginalize over $X_{V \setminus U}$. If G has no cycles, then $p(G; \mathbf{x}_U; \theta)$ can be computed with complexity linear in the number of nodes in V. If G has cycles, one can compute $p(G; \mathbf{x}_U; \theta)$ by grouping subsets of V into supernodes such that the new graph is acyclic [14]. In this case, complexity is exponential in the size of the largest supernode. A graph is said to be *tractable* if either G has no cycles or if G can be clustered into an acyclic graph where the size of the largest supernode is moderate. Similarly, a subset U is said to be tractable if G_U is tractable, in which case $p(G_U; \mathbf{x}_U; \tilde{\theta}_U)$ can be computed for the reduced MRF on G_U. Also, for tractable subset W, $p(G; \mathbf{x}_W | \mathbf{x}_{\partial W}; \theta)$ can be computed for configurations $\mathbf{x}_W$ and $\mathbf{x}_{\partial W}$.

For the purposes of this paper it suffices to say that lossless compression with an *optimal encoder* involves computation of a *coding distribution.* For a tractable subset U, if configuration $\mathbf{x}_U$ is losslessly compressed with reduced MRF coding distribution $p(G_U; X_U; \tilde{\theta}_U)$, then the average number of bits produced is the *cross entropy* $H(G; X_U; \theta || G_U; X_U; \tilde{\theta}_U)$ between the marginal distribution $p(G; X_U; \theta)$ and the reduced MRF coding distribution $p(G_U; X_U; \tilde{\theta}_U)$ for X_U, defined as

$$H(G; X_U; \theta || G_U; X_U; \tilde{\theta}_U) \;=\; H(G; X_U; \theta) + D(p(G; X_U; \theta) || p(G_U; X_U; \tilde{\theta}_U))$$

where $D(p(G; X_U; \theta)||p(G_U; X_U; \tilde{\theta}_U))$ is the *divergence* from $p(G; X_U; \theta)$ to $p(G_U; X_U; \tilde{\theta}_U)$ and is the *redundancy* in the code [4].

We showed in [10] that the above divergence is minimized at θ^*_U, the exponential parameter on G_U such that the corresponding moment μ^*_U is equal to the moment subvector μ_U under the original MRF $p(G; X; \theta)$. The distribution of the reduced MRF $p(G_U; X_U; \theta^*_U)$ is called the *moment-matching* reduced MRF distribution for X_U, denoted $\tilde{X}_U$. When the moment-matching reduced MRF $p(G_U; X_U; \theta^*_U)$ is used as the coding distribution to encode X_U, the cross entropy is in fact the entropy $H(G_U; X_U; \theta^*_U)$ of the moment-matching reduced MRF [10].

For a tractable subset W, if configuration $\mathbf{x}_W$ is encoded conditioned on $\mathbf{x}_{\partial W}$ using coding distribution $p(G; X_W|\mathbf{x}_{\partial W}; \theta)$, then the average number of bits produced is $H(G; X_W|X_{\partial W}; \theta)$. Therefore, encoding $\mathbf{x}_W$ conditioned on $\mathbf{x}_{\partial W}$ is optimal, i.e., there is no redundancy.

In [10], Arithmetic Coding (AC) was proposed as the optimal encoder. Figure 1 illustrates the encoding of a line and a strip. The mathematical details of using AC in the encoding of an MRF are given in [8], [10], and [11], specifically in Chapter VI of [11].

3 Reduced Cutset Coding

In general, since the cutset U consists of disjoint lines, the entropy of the moment-matching reduced MRF on G_U is actually the sum $\sum_{L_i} H(G_{L_i}; X_{L_i}; \theta^*_{L_i})$ of the entropies of the reduced MRFs on the individual lines. Similarly, the conditional entropy of X_W given $X_{\partial W}$ is the sum $\sum_{S_i} H(G; X_{S_i}|X_{\partial S_i}; \theta)$ of the conditional entropies of the individuals strips given their respective boundaries.

In the present paper, we simplify this by considering vertically homogeneous parameters for the MRF, i.e., the components of the statistic t and the exponential parameter θ do not vary vertically within the image. Furthermore, focusing only on the middle $M' = (k' + 1)n_L + k'n_S \approx M/2$ rows of V, therefore excluding boundary effects, the image will be roughly stationary in the vertical direction. We let B_n be an $n \times N$ rectangular subset of sites.

The random field $X_{B_{n_L}}$ on a line is encoded with reduced MRF coding distribution $p(G_{n_L}; X_{B_{n_L}}; \theta^*_{B_{n_L}})$. Normalizing by the number of pixels, the per-row rate for encoding a line is then

$$\begin{aligned} \bar{R}^L_{n_L} &= \frac{1}{n_L} H(G; X_{B_{n_L}}; \theta || G_{n_L}; X_{B_{n_L}}; \theta^*_{B_{n_L}}) \\ &= \frac{1}{n_L} H(G_{B_{n_L}}; X_{B_{n_L}}; \theta^*_{B_{n_L}}). \end{aligned}$$

The random field $X_{B_{n_S}}$ on a strip is encoded conditioned on $X_{\partial B_{n_S}}$ with coding distribution $p(G; X_{B_{n_S}} | X_{\partial B_{n_S}}; \theta)$. The per-row rate for encoding a strip is then

$$\bar{R}^S_{n_S} \;=\; \frac{1}{n_S} H(G; X_{B_{n_S}} \mid X_{\partial B_{n_S}}; \theta).$$

We let $\bar{R}_{n_S,n_L}$ denote the total per-row rate of RCC with cutset parameters n_S and n_L, given by

$$\bar{R}(n_S, n_L) = \frac{(k+1)n_L}{(k+1)n_L + kn_S}\bar{R}^L_{n_L} + \frac{kn_S}{(k+1)n_L + kn_S}\bar{R}^S_{n_S}.$$

Assuming further that M' is very large relative to n_L and n_S, so that k is very large, this rate is well-approximated by

$$\bar{R}(n_S, n_L) \;\approx\; \frac{n_L}{n_L + n_S}\bar{R}^L_{n_L} + \frac{n_S}{n_L + n_S}\bar{R}^S_{n_S}. \tag{3}$$

We now see that the performance of RCC with cutset parameters n_S and n_L is characterized by the rates $\bar{R}^L_{n_L}$ and $\bar{R}^S_{n_S}$, and the fractions $\frac{n_L}{n_L+n_S}$ and $\frac{n_S}{n_L+n_S}$.

4 Moment-matching θ^*_U

Recall from the previous section that the cross-entropy $H(p(G; X_U; \theta)||p(G_U; X_U; \tilde{\theta}_U)$ between the marginal distribution $p(G; X_U; \theta)$ of subset X_U within an MRF on G with statistic t and a reduced MRF $p(G_U; X_U; \tilde{\theta}_U)$ on G_U with statistic t_U is minimized by the parameter θ^*_U such that the expected value $\mathbb{E}_{\theta^*_U}[t_U(X_U)]$ of the statistic t_U in the reduced MRF equals the expected value $(\mathbb{E}_\theta[t(X)])_U$ of the statistic t under the original MRF on the subset U, referred to as the *moment-matching* parameter. We will estimate θ^*_U from n observations $\mathbf{x}^{(1)}_U, \ldots, \mathbf{x}^{(n)}_U$ on U, by seeking an $\hat{\theta}^n_U$ that minimizes an empirical version of the cross entropy, at least approximately. First, some background.

We let $\Theta = \{\theta\}$ denote the set of parameter vectors for MRFs on G based on the statistic t. We restrict attention to the case where Θ is the subset of $\mathbb{R}^{|V|+|E|}$ of θ's with positive components. In this case, due to the openness of Θ, the family of MRFs based on t is said to be *regular* [14]. For parameter $\theta \in \Theta$, the function

$$\begin{aligned} \Lambda(\theta) \;&\triangleq\; \mathbb{E}_\theta[t(X)] \\ &\triangleq\; \mu \end{aligned}$$

maps θ to μ, the expected value of t under the MRF induced by θ, referred to as the *moment* of the MRF. The set $\mathcal{M} = \{\mu = \Lambda(\theta) : \theta \in \Theta\}$ is the set of *achievable moments* for MRFs on G based on t. We assume that the statistic t is *minimal* in that the components of t are affinely independent, meaning that the components of $t(\mathbf{x})$ do not sum to a constant for all configurations $\mathbf{x}$. In this case, the function $\Lambda(\cdot)$ is one-to-one [14]. Then, for $\mu \in \mathcal{M}$, the inverse function

$$\Lambda^{-1}(\mu) \;=\; \theta$$

is well-defined. Moreover, μ is a dual parameter to θ, in that the MRF $p(G; X; \theta)$ can alternatively be expressed as $p(G; X; \mu)$. For the MRF induced by parameter θ, the subvector of moments on the set U is given by

$$\Lambda_U(\theta) \;\; = \;\; \mu_U$$

which can be seen as the restriction of $\Lambda(\cdot)$ to the set U.

For reduced MRFs on G_U based on statistic t_U, $\tilde{\Theta}_U$ denotes the associated set of exponential parameters. Now, consider the function

$$\tilde{\Lambda}_U(\tilde{\theta}_U) \;\; = \;\; \tilde{\mu}_U,$$

which maps a parameter $\tilde{\theta}_U \in \tilde{\Theta}_U$ to the corresponding moment $\tilde{\mu}_U$ for the reduced MRF $p(G_U; X_U; \tilde{\theta}_U)$ on G_U. Likewise, $\tilde{\mathcal{M}}_U = \{\tilde{\mu}_U = \tilde{\Lambda}_U(\tilde{\mu}_U) : \tilde{\theta}_U \in \tilde{\Theta}_U\}$ denotes the set of achievable moments for reduced MRFs on G_U. Since we have assumed that the statistic t for the original family of MRFs on G is minimal, the statistic t_U for the family of reduced MRFs on G_U is also minimal, and the inverse map $\tilde{\Lambda}_U^{-1}(\tilde{\mu}_U) = \tilde{\theta}_U$ is well-defined. Again, a reduced MRF $p(G_U; X_U; \tilde{\theta}_U)$ can also be parameterized as $p(G_U; X_U; \tilde{\mu}_U)$.

Given a parameter θ for an MRF $p(G; X; \theta)$, a subset U, and a sequence of observations $\mathbf{x}_U^{(1)}, \ldots, \mathbf{x}_U^{(n)}$ on U, we define the *empirical moment* of $p(G; X_U; \theta)$ as

$$\hat{\mu}_U^n \;\; \triangleq \;\; \frac{1}{n} \sum_{i=1}^{n} t_U(\mathbf{x}_U^{(i)}).$$

While $\mu_U = \Lambda_U(\theta)$ is always contained in $\tilde{\mathcal{M}}_U$, it is not necessarily the case that the empirical moment $\hat{\mu}_U^n$ is contained in $\tilde{\mathcal{M}}_U$. However, even if $\hat{\mu}_U^n$ is not in $\tilde{\mathcal{M}}_U$, $\hat{\mu}_U^n$ is still a limit point of $\tilde{\mathcal{M}}_U$ [14], meaning that for every $\epsilon > 0$, there is an ϵ-ball containing $\hat{\mu}_U^n$ that contains infinitely many points of $\tilde{\mathcal{M}}_U$. Moreover, as stated in the following proposition, as the number of observations n approaches ∞, not only is $\hat{\mu}_U^n$ in $\tilde{\mathcal{M}}_U$, but $\hat{\mu}_U^n$ converges to μ_U.

Proposition 4.1 *The empirical moment $\hat{\mu}_U^n$ converges in probability to μ_U, i.e., for any $\epsilon > 0$,*

$$\Pr\left(\left|\hat{\mu}_U^n - \mu_U\right| \leq \epsilon\right) \to 1, \;\; \textit{as } n \to \infty. \tag{4}$$

Proof To prove the proposition, one should recall that on a finite graph G, there does not exist a phase transition [5], and therefore, there is a unique MRF on G for the specified statistic t and exponential parameter θ. It follows that the sequence $\mathbf{x}_U^{(1)}, \ldots, \mathbf{x}_U^{(n)}, \ldots$ is not only stationary but also ergodic, from which the proposition follows [6]. This completes the proof. □

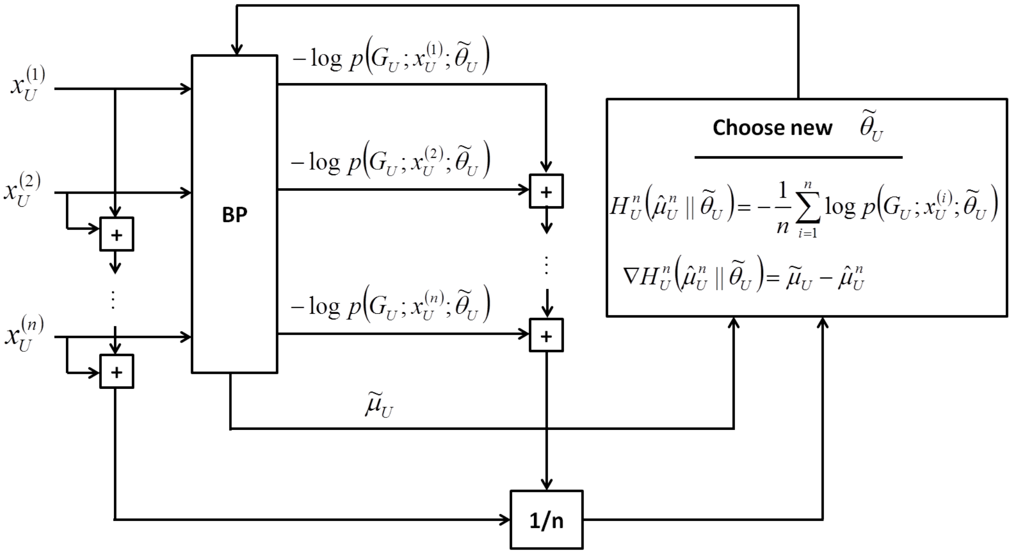

$x_U^{(1)}$
$x_U^{(2)}$
$x_U^{(n)}$
+
+
BP
$-\log p\left(G_U; x_U^{(1)}; \tilde{\theta}_U\right)$
$-\log p\left(G_U; x_U^{(2)}; \tilde{\theta}_U\right)$
$-\log p\left(G_U; x_U^{(n)}; \tilde{\theta}_U\right)$
+
+
Choose new $\tilde{\theta}_U$
$H_U^n\left(\hat{\mu}_U^n \,||\, \tilde{\theta}_U\right) = -\frac{1}{n}\sum_{i=1}^{n} \log p\left(G_U; x_U^{(i)}; \tilde{\theta}_U\right)$
$\nabla H_U^n\left(\hat{\mu}_U^n \,||\, \tilde{\theta}_U\right) = \tilde{\mu}_U - \hat{\mu}_U^n$
$\tilde{\mu}_U$
1/n

Figure 2: Block diagram for finding the moment-matching parameter θ_U^* for encoding X_U.

We now discuss the empirical version of cross entropy that we will minimize as a surrogate for cross entropy. From a sequence of observations $\mathbf{x}_U^{(1)}, \ldots, \mathbf{x}_U^{(n)}$, we define the *empirical cross entropy*

$$
\begin{aligned}
H_U^n(\hat{\mu}_U^n||\tilde{\theta}_U) &\triangleq -\frac{1}{n}\sum_{i=1}^{n} \log p(G_U; \mathbf{x}_U^{(i)}; \tilde{\theta}_U) \\
&= -\sum_{\mathbf{x}_U} f(\mathbf{x}_U : \mathbf{x}_U^{(1)}, \ldots, \mathbf{x}_U^{(n)}) \log p(G_U; \mathbf{x}_U; \tilde{\theta}_U)
\end{aligned}
$$

between the empirical distribution $f(\mathbf{x}_U : \mathbf{x}_U^{(1)}, \ldots, \mathbf{x}_U^{(n)})$ generated by $\mathbf{x}_U^{(1)}, \ldots, \mathbf{x}_U^{(n)}$ and the reduced MRF $p(G_U; X_U; \tilde{\theta}_U)$ induced by a candidate parameter $\tilde{\theta}_U$. That it makes sense to consider the empirical cross entropy to be a function of the empirical moment $\hat{\mu}_U^n$ is due to the proposition presented later. If U is a tractable subset, then the probabilities in the summation can be efficiently computed.

Now, our estimate for the moment-matching parameter θ_U^* will be the $\hat{\theta}_U^n$ that minimizes this empirical cross entropy, at least approximately. It is well-known that $\Phi_U(\tilde{\theta}_U)$ is convex in $\tilde{\theta}_U$, and, as follows from the following theorem, so is the empirical cross-entropy $H_U^n(\hat{\mu}_U^n||\tilde{\theta}_U)$. If, as we have assumed, the components of t_U are affinely independent, then $\Phi_U(\tilde{\theta}_U)$ and hence $H_U^n(\hat{\mu}_U^n||\tilde{\theta}_U)$ is strictly convex. Therefore, either $H_U^n(\hat{\mu}_U^n||\tilde{\theta}_U)$ has a unique minimum at a $\tilde{\theta}_U$ at which the gradient of $H_U^n(\hat{\mu}_U^n||\tilde{\theta}_U)$ is zero, or since $\tilde{\Theta}_U$ is open, $H_U^n(\hat{\mu}_U^n||\tilde{\theta}_U)$ does not have a minimum but approaches an infimum at a limit point of $\tilde{\Theta}_U$. Moreover, from the following theorem and the fact that for any $\hat{\mu}_U^n$ there exists $\tilde{\theta}_U$ such that $\tilde{\Lambda}_U(\tilde{\theta}_U)$ is arbitrarily close to $\hat{\mu}_U^n$, we can find $\tilde{\theta}_U$ such that the gradient is arbitrarily small and such $\tilde{\theta}_U$ must come arbitrarily close to attaining the infimum of $H_U^n(\hat{\mu}_U^n||\tilde{\theta}_U)$. In either case, our "moment-matching" estimate $\hat{\theta}_U^n$ will be a $\tilde{\theta}_U$ that induces a very small gradient.

Proposition 4.2

$$\begin{aligned} H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U) &= \Phi_U(\tilde{\theta}_U) - \langle \hat{\mu}^n_U, \tilde{\theta}_U \rangle \\ \nabla H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U) &= \tilde{\mu}_U - \hat{\mu}^n_U \\ &= \tilde{\Lambda}_U(\tilde{\theta}_U) - \hat{\mu}^n_U, \end{aligned}$$

where the gradient is with respect to $\tilde{\theta}_U$*.*

Proof Using relation (2) for the reduced MRF on G_U with parameter $\tilde{\theta}_U$, we get

$$\begin{aligned} H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U) &= \frac{1}{n}\sum_{i=1}^{n}\left[\Phi_U(\tilde{\theta}_U) - \langle t(\mathbf{x}_U^{(i)}), \tilde{\theta}_U\rangle\right] \\ &= \Phi_U(\tilde{\theta}_U) - \left\langle \sum_{i=1}^{n} t(\mathbf{x}_U^{(i)}), \tilde{\theta}_U \right\rangle \\ &= \Phi_U(\tilde{\theta}_U) - \left\langle \hat{\mu}^n_U, \tilde{\theta}_U \right\rangle \end{aligned}$$

It is well-known that $\nabla\Phi_U(\tilde{\theta}_U) = \tilde{\mu}_U$ [14]. Then, taking the gradient of $H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U)$ yields

$$\begin{aligned} \nabla H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U) &= \nabla\frac{1}{n}\sum_{i=1}^{n}\left[\Phi_U(\tilde{\theta}_U) - \langle t_U(\mathbf{x}_U^{(i)}), \tilde{\theta}_U\rangle\right] \\ &= \nabla\Phi_U(\tilde{\theta}_U) - \nabla\frac{1}{n}\sum_{i=1}^{n}\left\langle t_U(\mathbf{x}_U^{(i)}), \tilde{\theta}_U\right\rangle \\ &= \nabla\Phi_U(\tilde{\theta}_U) - \nabla\left\langle\frac{1}{n}\sum_{i=1}^{n} t_U(\mathbf{x}_U^{(i)}), \tilde{\theta}_U\right\rangle \\ &= \tilde{\mu}_U - \frac{1}{n}\sum_{i=1}^{n} t_U(\mathbf{x}_U^{(i)}) \\ &= \tilde{\mu}_U - \hat{\mu}^n_U. \end{aligned}$$

This completes the proof. □

We now describe how a gradient descent algorithm can be used to find an estimate $\hat{\theta}^n_U$ of θ^*_U at which the gradient of $H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U)$ is arbitrarily small. From the sequence $\mathbf{x}_U^{(1)}, \ldots, \mathbf{x}_U^{(n)}$, we first compute the empirical moment $\hat{\mu}^n_U = \frac{1}{n}\sum_{i=1}^{n} t_U(\mathbf{x}_U^{(i)})$. Then, given a candidate parameter $\tilde{\theta}_U$, use Belief Propagation to compute the negative log-likelihood $-\log p(G_U; \mathbf{x}_U^{(i)}; \tilde{\theta}_U)$ of the configuration $\mathbf{x}^{(i)}$ under the reduced MRF $p(G_U; X_U; \tilde{\theta}_U)$, for each $i = 1, \ldots, n$. Additionally, we compute the moment $\tilde{\mu}_U$ of the reduced MRF induced by the candidate parameter $\tilde{\theta}_U$, which like the probabilities, can be computed due to tractability of U. We then compute the objective function

$H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U) = -\frac{1}{n}\sum_{i=1}^n \log p(G_U; \mathbf{x}^{(i)}_U; \tilde{\theta}_U)$ and the gradient $\nabla H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U) = \tilde{\mu}_U - \hat{\mu}^n_U$. Finally, given a tolerance ϵ_μ, if $\|\nabla H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U)\| < \epsilon_\mu$, the algorithm terminates and we set $\hat{\theta}^n_U = \tilde{\theta}_U$ which corresponds to the estimated moment $\hat{\hat{\mu}}^n_U = \tilde{\Lambda}_U(\hat{\theta}^n_U)$ at which the algorithm is terminated. Note that by Proposition 4.2, the estimated moment $\hat{\hat{\mu}}^n_U$ is within ϵ_μ of $\hat{\mu}^n_U$. If $\|\nabla H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U)\| \geq \epsilon_\mu$, we determine a new candidate parameter $\tilde{\theta}_U$ using a standard gradient descent method [3] and repeat the above steps. This is illustrated in Figure 2.

Proposition 4.3 *The estimate $\hat{\theta}^n_U$ is consistent, i.e., for any $\epsilon > 0$,*

$$\Pr\left(\left|\hat{\theta}^n_U - \theta^*_U\right| \leq \epsilon\right) \to 1, \ \textit{as } n \to \infty. \tag{5}$$

Proof Let $B(\theta^*_U, \epsilon_\theta)$ be the ϵ_θ-ball centered at θ^*_U. Assume without loss of generality that $B(\theta^*_U, \epsilon_\theta) \subset \tilde{\Theta}_U$. Then, let ϵ_μ be the largest tolerance around μ_U such that the ϵ_μ-ball $B(\mu_U, \epsilon_\mu)$ centered at μ_U is contained in $\tilde{\Lambda}_U(B(\theta^*_U, \epsilon_\theta))$. It follows that

$$\begin{aligned}\Pr\left(\left|\hat{\theta}^n_U - \theta^*_U\right| \leq \epsilon_\theta\right) &= \Pr\left(\hat{\hat{\mu}}^n_U \in \tilde{\Lambda}_U(B(\theta^*_U, \epsilon_\theta))\right) \\ &\geq \Pr\left(\hat{\hat{\mu}}^n_U \in B(\mu_U, \epsilon_\mu)\right) \\ &= \Pr\left(\left|\hat{\hat{\mu}}^n_U - \mu_U\right| \leq \epsilon_\mu\right).\end{aligned}$$

Now let $\epsilon'_\mu = \epsilon_\mu/2$ be the tolerance on $\|\nabla H^n_U(\hat{\mu}^n_U||\tilde{\theta}_U)\|$ in the gradient descent algorithm. This means that $|\hat{\hat{\mu}}^n_U - \hat{\mu}^n_U| \leq \epsilon'_\mu$, which in turn implies that

$$\Pr\left(\left|\hat{\hat{\mu}}^n_U - \mu_U\right| \leq \epsilon_\mu\right) = \Pr\left(\left|\hat{\mu}^n_U - \mu_U\right| \leq \epsilon'_\mu\right).$$

Using Proposition 4.1, we can now say that for an arbitrary tolerance $\delta > 0$, there exists N such that if the number of observations n is greater than or equal to N, then

$$\begin{aligned}\Pr\left(\left|\hat{\theta}^n_U - \theta^*_U\right| \leq \epsilon_\theta\right) &\geq \Pr\left(\left|\hat{\mu}^n_U - \mu_U\right| \leq \epsilon'_\mu\right) \\ &\geq 1 - \delta.\end{aligned}$$

This completes the proof. □

5 Tradeoffs between Lines and Strips

The following proposition shows that, as intuited earlier, strip rate increases with strip width.

Proposition 5.1

$$\bar{R}^S_{n+1} > \bar{R}^S_n.$$

Lemma 5.2 *Let r_1 denote the first row of rectangular region B_n of sites of height n. Then,*

$$H(G; X_{r_1}|X_{\partial B_n}; \theta) \quad < \quad H(G; X_{r_1}|X_{\partial B_{n+1}}; \theta). \tag{6}$$

Proof Note B_{n+1} consists of B_n and an additional row r_{n+1}, which is part of the boundary of B_n. By the Markov property, $H(G; X_{r_1}|X_{\partial B_n}; \theta) = H(G; X_{r_1}|X_{\partial B_{n+1}}, X_{r_{n+1}}; \theta)$. That is, conditioning on ∂B_{n+1} and r_{n+1} is the same as conditioning on ∂B_n. Finally, $H(G; X_{r_1}|X_{\partial B_{n+1}}, X_{r_{n+1}}; \theta) < H(G; X_{r_1}|X_{\partial B_{n+1}}; \theta)$ as the left side has more conditioning. In summary

$$\begin{aligned} H(G; X_{r_1}|X_{\partial B_n}; \theta) &= H(G; X_{r_1}|X_{\partial B_{n+1}}, X_{r_{n+1}}; \theta) \\ &< H(G; X_{r_1}|X_{\partial B_{n+1}}; \theta). \end{aligned}$$

This completes the proof of Lemma 5.2. □

We continue with the proof of Proposition 5.1.

Proof By direct calculation we have for a strip of height $n+1$ that

$$\begin{aligned} \bar{R}^S_{n+1} &= \frac{1}{(n+1)} H(G; X_{B_{n+1}}|X_{\partial B_{n+1}}; \theta) \\ &= \frac{1}{(n+1)} H(G; X_{B_n}|X_{\partial B_n}; \theta) + \frac{1}{(n+1)} H(G; X_{r_1}|X_{\partial B_{n+1}}; \theta), \end{aligned} \tag{7}$$

and for a strip of height n,

$$\begin{aligned} \bar{R}^S_n &= \frac{1}{n} H(G; X_{B_n}|X_{\partial B_n}; \theta) \\ &= \frac{n+1}{n} \frac{1}{(n+1)} H(G; X_{B_n}|X_{\partial B_n}; \theta) \\ &= \frac{1}{(n+1)} H(G; X_{B_n}|X_{\partial B_n}; \theta) + \frac{1}{n(n+1)} H(G; X_{B_n}|X_{\partial B_n}; \theta) \\ &= \frac{1}{(n+1)} H(G; X_{B_n}|X_{\partial B_n}; \theta) + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(n+1)} H(G; X_{r_i}|X_{\partial B_{n-i+1}}; \theta) \\ &< \frac{1}{(n+1)} H(G; X_{B_n}|X_{\partial B_n}; \theta) + \frac{1}{(n+1)} H(G; X_{r_1}|X_{\partial B_{n+1}}; \theta) \\ &= \bar{R}^S_{n+1} \end{aligned}$$

by (7) and Lemma 5.2. This completes the proof. □

Likewise, the next proposition shows that, as supposed earlier, line rate decreases with line width.

Proposition 5.3

$$\bar{R}^L_{n+1} \quad < \quad \bar{R}^L_n.$$

Proof First we note that reducing $X_{B_{n+1}}$ to $\tilde{X}_{B_{n+1}}$ by matching moments and further reducing the X_{B_n} marginal of $\tilde{X}_{B_{n+1}}$ to $\tilde{X}_{B_n}$ by matching moments results in the same reduced MRF on G_{B_n} as would reducing the original X_{B_n} to $\tilde{X}_{B_n}$ by matching moments. Let θ^*_n be the moment matching parameter for $\tilde{X}_{B_n}$.

$$\begin{aligned}
\bar{R}^l_{n+1} &= \frac{1}{n+1} H(G_{B_{n+1}}; X_{B_{n+1}}; \theta^*_{n+1}) \\
&= \frac{1}{n+1} \left[H(G_{B_{n+1}}; X_{B_n}; \theta^*_{n+1}) + H(G_{B_{n+1}}; X_{r_{n+1}} | X_{B_n}; \theta^*_{n+1}) \right] \\
&< \frac{1}{n+1} \left[H(G_{B_{n+1}}; X_{B_n}; \theta^*_{n+1}) + \frac{1}{n} H(G_{B_{n+1}}; X_{B_n}; \theta^*_{n+1}) \right] \\
&= \frac{1}{n} H(G_{B_{n+1}}; X_{B_n}; \theta^*_{n+1}) \\
&< \frac{1}{n} H(G_{B_n}; X_{B_n}; \theta^*_n) \\
&= \bar{R}^L_n,
\end{aligned}$$

where the second inequality is from the maximum entropy property of MRFs. This completes the proof □

Proposition 5.4 *For all strip widths n_S and line widths n_L,*

$$\bar{R}^L_{n_L} \quad > \quad \bar{R}^S_{n_S}.$$

Proof We prove the proposition by cases: $n_S = n_L$, $n_S > n_L$, and $n_S < n_L$.

First assume $n_S = n_L = n$. Then,

$$\begin{aligned}
\bar{R}^S_{n_S} &= \frac{1}{n} H(G; X_{B_n} | X_{\partial B_n}; \theta) \\
&\leq \frac{1}{n} H(G; X_{B_n}; \theta) \\
&< \frac{1}{n} H(G_{B_n}; X_{B_n}; \theta^*_n) \qquad (8) \\
&= \bar{R}^L_{n_L},
\end{aligned}$$

where (8) follows from the maximum entropy property of MRFs. Next, assume $n_S > n_L$. Then,

$$\begin{aligned}
\bar{R}^S_{n_S} &= \frac{1}{n_S} H(G; X_{B_n} | X_{\partial B_n}; \theta) \\
&\leq \frac{1}{n_S} H(G; X_{B_n}; \theta) \\
&< \frac{1}{n_S} H(G_{B_n}; X_{B_{n_S}}; \theta^*_{n_S}) \qquad (9) \\
&= \bar{R}^L_{n_S} \\
&< \bar{R}^L_{n_L},
\end{aligned}$$

where (9) follows from the maximum entropy property of MRFs. Finally, assume $n_S < n_L$. Then,

$$\begin{aligned}
\bar{R}^S_{n_S} &< \bar{R}^S_{n_L} \\
&= \frac{1}{n_L} H(G; X_{B_{n_L}} | X_{\partial B_{n_L}}; \theta) \\
&\leq \frac{1}{n_L} H(G; X_{B_{n_L}}; \theta) \\
&< \frac{1}{n_L} H(G_{B_{n_L}}; X_{B_{n_L}}; \theta^*_{n_L}) \qquad (10) \\
&= \bar{R}^L_{n_L},
\end{aligned}$$

where (10) follows from the maximum entropy property of MRFs. This completes the proof. □

Together these three propositions indicate that $\bar{R}^L_{n_L}$ and $\bar{R}^S_{n_S}$ always behave as in Figure 3 (a), which as discussed in the next section, plots them for a specific case. They also illustrate the potential tradeoffs between line width n_L and strip width n_S. Specifically, by increasing n_L the line rate $\bar{R}^L_{n_L}$ decreases, though the fraction $\frac{n_L}{n_S+n_L}$ of pixels encoded at the higher rate increases, while increasing n_S increases the fraction $\frac{n_S}{n_L+n_S}$ of pixels encoded at the lower rate, though the strip rate $\bar{R}^S_{n_S}$ increases.

In addition to considering the effect of n_S and n_L on rate, we can look at their influence on the rate redundancy $\Delta(n_S, n_L) \overset{\Delta}{=} \frac{1}{|V|} D(X_U || \tilde{X}_U)$, which is entirely due to encoding the lines independently and as moment-matching reduced MRFs. We use the shorthand notation $\tilde{X}_{B_{n_L}}$ to indicate the moment-matching reduced MRF on B_{n_L} and $D(X_{B_{n_L}} || \tilde{X}_{B_{n_L}})$ to denote the divergence between the marginal and moment-matching reduced MRF distributions for $X_{B_{n_L}}$.

Proposition 5.5 *The per-row rate redundancy due to coding $X_U \sim p(G; X_U; \theta)$ as a reduced MRF $X_U \sim p(G_U; X_U : \theta^*_U)$ is*

$$\bar{\Delta}(n_S, n_L) = \frac{n_S}{n_S + n_L} I(X_{r_1}; X_{r_{-n_S}}) + \frac{n_L}{n_S + n_L} D(X_{B_{n_L}} || \tilde{X}_{B_{n_L}}),$$

where r_1 is the 1st row of a line, and r_{-n_S} is the last row of the previous line.

Proof To prove the proposition, consider a joint distribution $p(x_1, \ldots, x_N)$ on N variables, where we have in mind each variable representing one of the $N = k+1$ lines. By approximating $p(x_1, \ldots, x_N)$ with $\tilde{p}(x_1, \ldots, x_N) = \prod_{i=1}^{N} \tilde{p}(x_i)$ we can see that the divergence between p and $\tilde{p}$ is

$$\begin{aligned}
D(p||\tilde{p}) &= \sum_{x_1,\ldots,x_N} p(x_1, \ldots, x_N) \log \frac{p(x_1, \ldots, x_N)}{\tilde{p}(x_1) \cdots \tilde{p}(x_N)} \\
&= -\sum_{x_1,\ldots,x_N} p(x_1, \ldots, x_N) \log \tilde{p}(x_1) \cdots \tilde{p}(x_N) - H(X_1, \ldots, X_N) \\
&= \sum_{i=1}^{N} \sum_{x_i} -p(x_i) \log \tilde{p}(x_i) - H(X_1, \ldots, X_N) \\
&= \sum_{i=1}^{N} [H(X_i) + D(p(X_i)||\tilde{p}(X_i))] - H(X_1, \ldots, X_N) \\
&= \sum_{i=1}^{N} [H(X_i) - H(X_i|X_{i-1}, \ldots, X_1) + D(p(X_i)||\tilde{p}(X_i))] \\
&= \sum_{i=2}^{N} I(X_i; X_{i-1}) + \sum_{i=1}^{N} D(p(X_i)||\tilde{p}(X_i)).
\end{aligned}$$

Applying the stationarity assumption, weighting the last two terms by the (approximate) fractions in (3), and substituting $N = k+1$ and $X_i = X_{B_{n_L}}$ yields

$$\bar{\Delta}(n_S, n_L) = \frac{n_S}{n_S + n_L} I(X_{B_{n_L}}; X_{B_{n_L}, -n_S}) + \frac{n_L}{n_S + n_L} D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}}),$$

where $I(X_{B_{n_L}}; X_{B_{n_L}, -n_S})$ is the mutual information between two $n_L \times N$ rectangular blocks of sites separated by a $n_S \times N$ rectangular block of sites. To finish the proof, it suffices to consider $I(X_1, X_2; Y_1, Y_2)$ where $X_1 - X_2 - Y_1 - Y_2$ form a Markov Chain. In this case,

$$\begin{aligned}
I(X_1, X_2; Y_1, Y_2) &= H(Y_1, Y_2) - H(Y_1, Y_2|X_1, X_2) \\
&= H(Y_1) + H(Y_2|Y_1) - H(Y_1|X_1, X_2) - H(Y_2|Y_1, X_1, X_2) \\
&= H(Y_1) + H(Y_2|Y_1) - H(Y_1|X_2) - H(Y_2|Y_1) \\
&= H(Y_1) - H(Y_1|X_2) \\
&= I(Y_1; X_2).
\end{aligned}$$

Making the appropriate substitutions yields

$$\bar{\Delta}(n_S, n_L) = \frac{n_S}{n_S + n_L} I(X_{r_1}; X_{r_{-n_S}}) + \frac{n_L}{n_S + n_L} D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}}),$$

where $I(X_{r_1}; X_{r_{-n_S}})$ is the mutual information between the 1st row of a line and the last row of the previous line. This completes the proof. □

This proposition shows specifically how the redundancy of RCC has two components: a correlation redundancy $I(X_{r_1}; X_{r_{-n_S}})$ due to encoding the lines independently of one another, and a distribution redundancy $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$ due to approximating the lines as moment matching reduced MRFs.

Proposition 5.6 *$I(X_{r_1}; X_{r_{-n_S}})$ is decreasing in n_S.*

Proof We let $X_{r_{i,1}}$ denote the 1st row of the i-th line and $X_{r_{i-1,n_L}}$ and $X_{r_{i-1,n_L-1}}$ denote, respectively, the n_L-th and (n_L-1)-st lines of the $(i-1)$-st line.

$$\begin{aligned}
I(X_{r_1}; X_{r_{-n_S}}) &= H(G; r_{i,1}; \theta) - H(G; X_{r_{i,1}}|X_{r_{i-1,n_L}}; \theta) \\
&= H(G; r_{i,1}; \theta) - H(G; X_{r_{i,1}}|X_{r_{i-1,n_L}}, X_{r_{i-1,n_L-1}}; \theta) \qquad (11) \\
&> H(G; r_{i,1}; \theta) - H(G; X_{r_{i,1}}|X_{r_{i-1,n_L-1}}; \theta) \qquad (12) \\
&= I(X_{r_1}; X_{r_{-(n_S+1)}}),
\end{aligned}$$

where (11) is due to the Markov property and (12) is due to removing conditioning. This completes the proof. □

To analyze the distribution redundancy, we let $\tilde{\tilde{X}}_{B_n}$ be the marginal distribution of $X_{B_{n-1}}$ as a subset of the moment-matching reduced MRF $\tilde{X}_{B_n}$on B_n. More generally, X_{B_n} decorated with k "tildes" indicates the marginal distribution of $X_{B_{n-k+1}}$ as a subset of the moment-matching reduced MRF $\tilde{X}_{B_n}$ on B_n. Moreover, we let θ^*_n be shorthand for $\theta^*_{B_n}$. We then have the following recursive expression for the distribution redundancy.

Proposition 5.7

$$\begin{aligned}
D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}}) &= D(X_{B_{n_L-1}}||\tilde{X}_{B_{n_L-1}}) - D(\tilde{\tilde{X}}_{B_{n_L}}||\tilde{X}_{B_{n_L-1}}) \\
&\quad + H(G_{B_{n_L}}; r_{n_L}|r_{n_L-1}; \theta^*_{n_L}) - H(G; r_{n_L}|r_{n_L-1}; \theta),
\end{aligned}$$

where $D(\tilde{\tilde{X}}_{B_{n_L}}||\tilde{X}_{B_{n_L-1}})$ is the divergence between the marginal distribution of $X_{B_{n_L-1}}$ as a subfield of $\tilde{X}_{B_{n_L}}$ and the reduced MRF $\tilde{X}_{B_{n_L-1}}$ on B_{n_L-1}, and where $H(\cdot; r_n|r_{n-1}; \cdot)$ is the conditional entropy of row r_n condition on row r_{n-1} for the specified graph and parameter vector.

Proof We prove the proposition by using the fact that the divergence $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$ between the marginal distribution of $X_{B_{n_L}}$ and the reduced MRF for $X_{B_{n_L}}$ can be expressed as the difference between the entropy of the latter and that of the former. Specifically,

$$\begin{aligned}
D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}}) &= H(G_{B_{n_L}};X_{B_{n_L}};\theta^*_{n_L}) - H(G;X_{B_{n_L}};\theta) \\
&= H(G_{B_{n_L}};X_{B_{n_L-1}};\theta^*_{n_L}) - H(G;X_{B_{n_L-1}};\theta) \\
&\quad +H(G_{B_{n_L}};r_{n_L}|r_{n_L-1};\theta^*_{n_L}) - H(G;r_{n_L}|r_{n_L-1};\theta) \\
&= H(G_{B_{n_L-1}};X_{B_{n_L-1}};\theta^*_{n_L-1}) - D(\tilde{\tilde{X}}_{B_{n_L}}||\tilde{X}_{B_{n_L-1}}) \\
&\quad -H(G;X_{B_{n_L-1}};\theta) + H(G_{B_{n_L}};r_{n_L}|r_{n_L-1};\theta^*_{n_L}) - H(G;r_{n_L}|r_{n_L-1};\theta) \\
&= H(G_{B_{n_L-1}};X_{B_{n_L-1}};\theta^*_{n_L-1}) - H(G;X_{B_{n_L-1}};\theta) \\
&\quad -D(\tilde{\tilde{X}}_{B_{n_L}}||\tilde{X}_{B_{n_L-1}}) + H(G_{B_{n_L}};r_{n_L}|r_{n_L-1};\theta^*_{n_L}) - H(G;r_{n_L}|r_{n_L-1};\theta) \\
&= D(X_{B_{n_L-1}}||\tilde{X}_{B_{n_L-1}}) - D(\tilde{\tilde{X}}_{B_{n_L}}||\tilde{X}_{B_{n_L-1}}) \\
&\quad +H(G_{B_{n_L}};r_{n_L}|r_{n_L-1};\theta^*_{n_L}) - H(G;r_{n_L}|r_{n_L-1};\theta).
\end{aligned}$$

This completes the proof. □

Furthermore, the divergence $D(\tilde{\tilde{X}}_{B_{n_L-1}}||\tilde{X}_{B_{n_L-1}})$ has the following recursive relationship.

Proposition 5.8

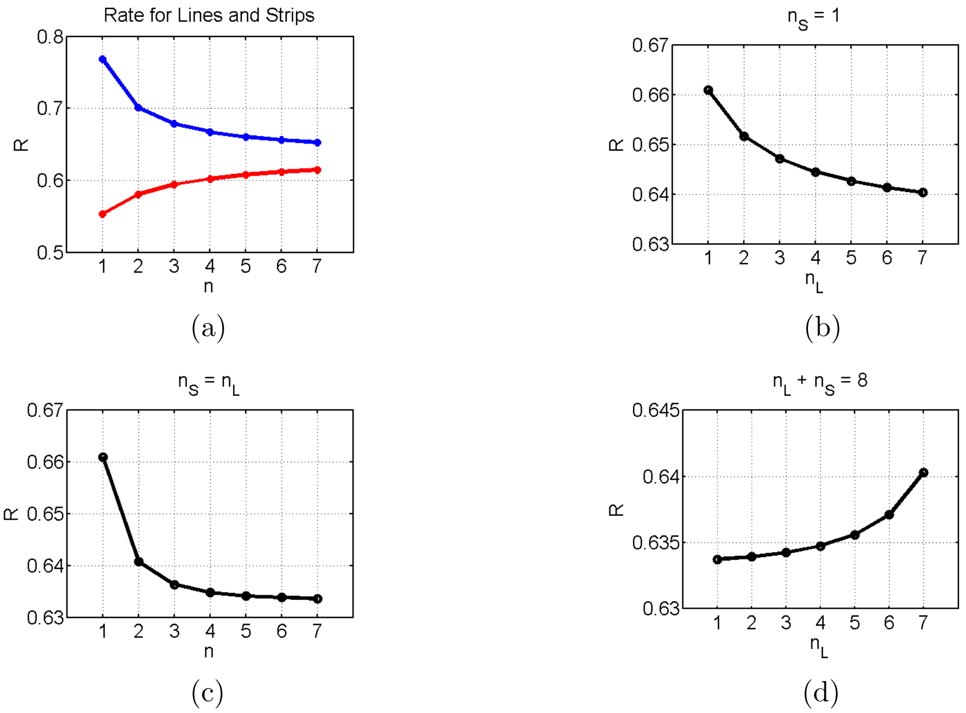
$$\begin{aligned}
D(\tilde{\tilde{X}}_{B_{n-k+1}}||\tilde{X}_{B_{n-k}}) &= D(\tilde{\tilde{\tilde{X}}}_{B_{n-k+1}}||\tilde{X}_{B_{n-k-1}}) - D(\tilde{\tilde{X}}_{B_{n-k}}||\tilde{X}_{B_{n-k-1}}) \\
&\quad +H(G_{B_{n-k}};r_{n-k}|r_{n-k-1};\theta^*_{n-k}) - H(G_{B_{n-k+1}};r_{n-k}|r_{n-k-1};\theta^*_{n-k+1})
\end{aligned}$$

where $D(\tilde{\tilde{\tilde{X}}}_{B_{n-k+1}}||\tilde{X}_{B_{n-k-1}})$ *is the divergence between the marginal distribution of* $X_{B_{n-k-1}}$ *as a subfield of* $\tilde{X}_{B_{n-k+1}}$ *and the reduced MRF* $\tilde{X}_{B_{n-k-1}}$ *on* B_{n-k-1}.

Proof

$$\begin{aligned}
D(\tilde{\tilde{X}}_{B_{n-k+1}}||\tilde{X}_{B_{n-k}}) &= H(G_{B_{n-k}};X_{B_{n-k}};\theta^*_{n-k}) - H(G_{B_{n-k+1}};X_{B_{n-k}};\theta^*_{n-k+1}) \\
&= H(G_{B_{n-k}};X_{B_{n-k-1}};\theta^*_{n-k}) - H(G_{B_{n-k+1}};X_{B_{n-k-1}};\theta^*_{n-k+1}) \\
&\quad +H(G_{B_{n-k}};r_{n-k}|r_{n-k-1};\theta^*_{n-k}) - H(G_{B_{n-k+1}};r_{n-k}|r_{n-k-1};\theta^*_{n-k+1}) \\
&= H(G_{B_{n-k-1}};X_{B_{n-k-1}};\theta^*_{n-k-1}) - D(\tilde{\tilde{X}}_{n-k}||\tilde{X}_{n-k-1}) \\
&\quad -H(G_{B_{n-k+1}};X_{B_{n-k-1}};\theta^*_{n-k+1}) + H(G_{B_{n-k}};r_{n-k}|r_{n-k-1};\theta^*_{n-k}) \\
&\quad -H(G_{B_{n-k+1}};r_{n-k}|r_{n-k-1};\theta^*_{n-k+1}) \\
&= D(\tilde{\tilde{\tilde{X}}}_{n-k+1}||\tilde{X}_{n-k-1}) - D(\tilde{\tilde{X}}_{n-k}||\tilde{X}_{n-k-1}) \\
&\quad +H(G_{B_{n-k}};r_{n-k}|r_{n-k-1};\theta^*_{n-k}) - H(G_{B_{n-k+1}};r_{n-k}|r_{n-k-1};\theta^*_{n-k+1}).
\end{aligned}$$

This completes the proof. □



(a)

(b)

(c)

(d)

Figure 3: Rate (a) for lines (blue) and strips (red); (b) as a function of n_L for $n_S = 1$; (c) as a function of $n = n_S = n_L$; and (d) for $n_S + n_L = 8$.

Intuitively we would expect the term $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$ to decrease in n_L, as this divergence is zero when $n_L = M$ and indeed we conjecture that this is the case. At the very least, we expect $\frac{1}{n_L}D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$ to decrease in n_L.

We now consider the effects of changing n_S and n_L on redundancy, as expressed in Proposition 5.5. Increasing n_S decreases distribution redundancy through the factor $\frac{n_L}{n_S+n_L}$. It is not so clear what happens to the correlation redundancy, as increasing n_S increases the fraction $\frac{n_S}{n_S+n_L}$, while decreasing the information $I(X_{r_1}; X_{r_{-n_S}})$. However, if we keep n_S and n_L proportional to one another, as n_S increases, the fraction stays the same, the correlation redundancy decreases, and assuming the conjecture, so too does distribution redundancy.

Similarly, increasing n_L decreases the correlation redundancy through the factor $\frac{n_S}{n_S+n_L}$. Even assuming the above conjecture, it is not clear what happens to the distribution redundancy, as increasing n_L increases the fraction $\frac{n_L}{n_S+n_L}$, while decreasing the divergence $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$. However, as mentioned above, if n_S and n_L increase proportionally to one another, then the fraction stays the same and both the correlation and distribution redundancies decrease in n_L.

The complexity of this coding scheme can be expressed as

$$C_{n_S,n_L} \quad = \quad \frac{n_S}{n_S+n_L}|\mathcal{X}|^{n_S}c_S + \frac{n_L}{n_S+n_L}|\mathcal{X}|^{n_L}c_L$$

where $|\mathcal{X}|$ denotes the number of elements of $\mathcal{X}$, and c_S and c_L are factors relating the complexity of encoding a strip versus a line. For example, numerical simulations show that for $n_S = n_L$, the run-time involved in encoding a strip is a little higher than that for a line, which is due to additional operations for conditioning on the boundary of a strip. However, the difference becomes negligible as n_S and n_L become larger. As a result, the complexity C_{n_S,n_L} is dominated by $\max\{n_S, n_L\}$. Given a constraint $\max\{n_S, n_L\} \leq n^*$ on the maximum exponent in the complexity, since both Proposition 5.6 and our conjecture indicate choosing n_S and n_L each to be as large as possible, we propose setting $n_S = n_L$.

6 Example: Homogeneous Ising Model

We simulated a homogeneous Ising model with edge parameter $\theta_{ij} = 0.4$ and node parameter $\theta_i = 0$ using Gibbs sampling. To encode the lines with line width n_L, we approximate the moment-matching parameter $\theta^*_{n_L}$ by minimizing the empirical cross entropy

$$H^{nK}_{n_L}(\tilde{\theta}_{n_L}) \quad = \quad \frac{1}{nK}\sum_{L_i}\sum_{j=1}^{n} -\log p(G_{L_i}; \mathbf{x}^{(j)}_{L_i}; \tilde{\theta}_{n_L}).$$

Note that even for a homogeneous MRF, the moment-matching parameter for a subset U will in general not be homogeneous.

The line rate $\bar{R}^L_{n_L}$ is approximated by

$$\hat{R}^L_{n_L} \quad = \quad \frac{1}{nK}\sum_{L_i}\sum_{j=1}^{n} -\log p(G_{L_i}; \mathbf{x}^{(j)}_{L_i}; \theta^*_{n_L}).$$

Similarly, $\bar{R}^S_{n_S}$ is approximated by

$$\hat{R}^S_{n_S} \quad = \quad \frac{1}{nK}\sum_{S_i}\sum_{j=1}^{n} -\log p(G; \mathbf{x}^{(j)}_{S_i}|\mathbf{x}^{(j)}_{\partial S_i}; \theta).$$

Figure 3(a) shows $\hat{R}^L_{n_L}$ and $\hat{R}^L_{n_S}$. As predicted by Propositions 5.1, 5.3, and 5.4, $\hat{R}^S_{n_S}$ is increasing in n_S, $\hat{R}^L_{n_L}$ is decreasing in n_L, and $\hat{R}^S_{n_S} < \hat{R}^L_{n_L}$ for all n_S, n_L. We computed $\hat{R}_{n_S,n_L}$ from $\hat{R}^L_{n_L}$ and $\hat{R}^L_{n_S}$ using (3), and as seen in Figure 3(b), we found that $\hat{R}_{n_S,n_L}$ decreases as n_L increases for constant n_S. We also found, see Figure 3(c), that $\hat{R}_{n_S,n_L}$ decreases with n increasing when $n = n_L = n_S$, which is consistent with the earlier discussion that presumed the conjecture. Finally, we found that if one holds the sum $n_L + n_S$ constant, then the rate $\hat{R}_{n_S,n_L}$ is minimized when $n_L = 1$. This indicates that the information $I(X_{r_1}; X_{r_{-n_S}})$ decreases with n_S faster than the divergence $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$ decreases with n_L. Though not apparent in the Figure, we found that $\hat{R}_{7,7} < \hat{R}_{7,1}$, an improvement over our earlier paper [10] which focused exclusively on $n_L = 1$. However, the improvement is nominal, so therefore, at least for this particular value of θ_{ij}, does not justify the significantly increased complexity.

7 Concluding Remarks

In this paper we have addressed the topic of tradeoffs in the choice of the width n_L and spacing n_S of the cutset components in Reduced Cutset Coding of Markov random fields. We have provided analysis from the perspective of the rate of this scheme in terms of the rates for encoding lines and strips and the relative contributions of each to the overall rate. We have shown that the rate for encoding lines with the moment-matching reduced MRF decreases with n_L, and that the rate for encoding strips increases with n_S, and on the basis of just these results one might conclude that large n_L and small n_S would provide an optimal combination. However, we also show that for all combinations of n_L and n_S, the rate for encoding lines is strictly greater than the rate for encoding strips. Moreover, the fraction $\frac{n_L}{n_S+n_L}$ of sites encoded at the larger rate obviously increases with n_L, while the fraction $\frac{n_S}{n_S+n_L}$ of sites encoded at the smaller rate obviously decreases with n_S.

Additionally, we have analyzed the problem from the perspective of the redundancy in the code, showing that this redundancy decomposes into a distribution redundancy due to approximating the lines as moment-matching reduced MRFs, and a correlation redundancy due to independent coding of the lines. We show that the correlation redundancy is decreasing in n_S and provide analysis of the distribution redundancy and conjecture that it is decreasing in n_L. Indeed, numerical experiments with an Ising model corroborate this conjecture. Moreover, if we let n_L be the height of the original image, then clearly the divergence $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}}) = 0$, and at least offhand, there is no reason to suspect that this divergence is non-monotonic in n_L. Naturally, though, further analysis of $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$ remain to be done, and at least at the moment, we suspect that the recursive relations for $D(X_{B_{n_L}}||\tilde{X}_{B_{n_L}})$ will be useful in proving our conjecture.

While for general row-invariant statistics t and exponential parameters θ it is not clear what the best choices of n_L and n_S should be, our numerical experiments with a uniform Ising model with parameters $\theta_{ij} = 0.4, \theta_i = 0$ suggest that letting n_S and n_L both be as large as possible achieves a lower rate. However, since the decrease in rate over a large n_S and $n_L = 1$ is in the fourth decimal place (in terms of per-site rate), the greatly increased complexity in encoding lines with large n_L does not seem worth it. However, more work remains to be done in understanding how differences in parameter values affect these tradeoffs. And more generally, beyond the Ising model, we would like to understand how the apparent tradeoffs between n_S and n_L vary with θ for different types of statistic t. Previous work of the authors [9, 11, 12] has looked at the relationship between *positively correlated* statistics t and quantities of interest and it will be interesting to see if such statistics can be shown to have significant consequences for RCC.

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.

[3] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

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

[5] H.O. Georgii, “Gibbs Measures and Phase Transitions,” De Gruyter, New York, 1988.

[6] G. Grimmett and D. Stirzaker, “Probability and Random Processes,” Oxford, 2001.

[7] I. Kontoyiannis, “Pattern Matching and Lossy Data Compression on Random Fields,” *IEEE Tr. Inform. Thy.*, v. 49, pp. 1047–1051, April 2003.

[8] M.G. Reyes and D.L. Neuhoff, *Arithmetic Compression of Markov Random Fields*, Seoul, Korea, ISIT 2009.

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

[10] M.G. Reyes and D.L. Neuhoff, *Lossless Reduced Cutset Coding of Markov Random Fields*, Snowbird, UT, DCC 2010.

[11] M.G. Reyes, *Cutset Based Processing and Compression of Markov Random Fields*, Ph.D. thesis, University of Michigan, April 2011.

[12] M.G. Reyes, “Covariance and Entropy in Markov Random Fields,” ITA, San Diego, 2013.

[13] M. G. Reyes, D. L. Neuhoff, T. N. Pappas, “Lossy Cutset Coding of Bilevel Images Based on Markov Random Fields,” *IEEE Trans. Img. Proc.*, vol. 23, pp. 1652-1665, April 2014.

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

[15] I. H. Whitten, R. M. Neal, and J. G. Cleary, *Arithmetic Coding For Data Compression*, Comm. of the ACM, vol. 30, pp. 520-540, June 1987.