Entropy Bounds for a Markov Random Subfield

Matthew G. Reyes
EECS Department
University of Michigan
Ann Arbor, MI 48109
mgreyes@umich.edu

David L. Neuhoff
EECS Department
University of Michigan
Ann Arbor, MI 48109
neuhoff@umich.edu

Abstract— **Given a Markov random field (MRF) X defined by potentials on a graph $G = (V, E)$, and given a subset $U \subset V$ of the sites on which X is defined, we prove, under a positive correlation constraint on the MRF, that the entropy of the subfield X_U is upper bounded by the entropy of an MRF defined on the subgraph induced by U with potentials taken directly from those assigned to U in G. To prove this we use exponential family representations of MRFs. We first show that the entropy of an MRF is monotone decreasing in the exponential parameters. We then use the Maximum Entropy principle and a well-known result from information geometry to show that the marginal entropy of X_U is upper bounded by the MRF on the induced subgraph with moments matching the marginal distribution. We then use the convexity of the log-partition function to show that to match the marginal moments on the induced subgraph, the exponential coordinates on the induced subgraph are component-wise greater than the corresponding parameter of the original exponential characterization. Our result follows from monotonicity.**

I. Introduction

A Markov random field (MRF) has a probability distribution p whose pairwise conditional independence relations can be depicted by an undirected graph $G = (V, E)$, where the nodes in V are the random variable indices and the edges in E represent direct dependencies between the random variables. Markov random fields (MRFs) are commonly used to model spatially distributed data such as images [8]. Their popularity is due in part to the fact that a typical MRF distribution can be expressed as a product of functions each defined over a relatively small subset of the random variables. Markov random fields, under various names, have been studied extensively from the points of view of statistical inference [2], [3], model selection [1], [10], and Belief Propagation [13].

In this paper we show that under a correlation[1] constraint on the MRF, the marginal entropy of a subfield X_U can be upper bounded by considering MRFs on the subgraph $G_U = (U, E_U)$[2], where $E_U \subset E$ is the set of edges both of whose endpoints are in U. In particular, we show that the MRF on G_U, where subsets of U have the same potentials as in the original MRF specification, has greater entropy than the marginal distribution on X_U. This problem is motivated by the success of a new lossy compression scheme for binary images [9]. There, binary images are modeled as Ising models and a subset $U \subset V$ of pixels consisting of a grid of evenly spaced rows and columns of the image are chosen. For a given image x, the pixel values x_U are losslessly encoded and the remaining pixel values $x_{V \setminus U}$ are estimated at the decoder. We are thus interested in the marginal entropy of the random variables X_U as this determines the lower bound to the rate of this encoding scheme. One reason for considering the subgraph G_U is that by removing edges we can make the problem more amenable to analysis or computational tools. It is intuitive that if we cut edges between U and the rest of the graph, the entropy of X_U should increase, as there are fewer constraints. However, we found no such result in the literature. The idea of modifying a graphical model to make a problem more manageable, however, has been used to find upper bounds to the log-partition function [12], and to perform approximate inference in MRFs [13].

The results in this paper use the representation of MRFs as members of exponential families of distributions, which have been extensively studied in statistics [3], and have become useful tools in analyzing graphical models [14]. A family of exponential distributions is specified by a vector statistic $t = (t_{ij})$ defined on the endpoints of the edges E of the graph.[3] That is, for a given image $\mathbf{x} = \{x_i : i \in V\}$ and each edge $\{i, j\} \in E$, the function $t_{ij} : \mathcal{X}_i \times \mathcal{X}_j \longrightarrow \mathbb{R}$ determines the contribution of the pair (x_i, x_j) to the probability of $\mathbf{x}$. 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 $\theta = (\theta_{ij})$ where for each edge $\{i, j\}$, θ_{ij} scales the sensitivity of the distribution $p(\mathbf{x}) = p(\mathbf{x}; \theta)$ to the function t_{ij}. Specifically, if X is an MRF based on t with exponential parameter θ, the probability of an image $\mathbf{x}$ is

$$\begin{aligned} p(\mathbf{x}; \theta) &= \exp\{\sum_{\{i,j\} \in E} \theta_{ij} t_{ij}(x_i, x_j) - \Phi(\theta)\} \\ &= \exp\{\langle \theta, t(\mathbf{x}) \rangle - \Phi(\theta)\}, \end{aligned}$$

where

$$\Phi(\theta) = \log \left[\sum_{\mathbf{x} \in \mathcal{X}} \exp\{\langle \theta, t(\mathbf{x}) \rangle\} \right]$$

is the log-partition function. The set

$$\Theta = \{\theta \in \mathbb{R}_+^{|E|} \mid \Phi(\theta) < \infty\}$$

[1] We will make this notion explicit in Section II.

[2] G_U is called the *subgraph induced by* U.

[3] Properly, this is a *pairwise* MRF with no self-potentials. Generalizations to other MRFs are straightforward.

978-1-4244-4313-0/09/$25.00 ©2009 IEEE

is the set of *admissable* exponential parameters[4], while $\mathcal{F} = \{p(\cdot;\theta) \mid \theta \in \Theta\}$ is the family of all MRFs based on t.

The set Θ can be viewed as a coordinate system for $\mathcal{F}$, a particular element θ indexing an MRF $X \sim p(\cdot;\theta)$. For a subset of edges $D \subset E$, we can express an exponential parameter θ in partitioned form as $\theta = (\theta_D, \backslash\theta_D)$.[5] We can then define the new exponential parameter $\bar{\theta} = (\theta_D, 0)$. This amounts to removing edges from the graph, corresponding to the subgraph $\bar{G} = (V, D)$. Viewed in this way, studying exponential families allows us to consider a hierarchy of MRFs defined on subgraphs of the original graph [1]. This hierarchical perspective has been used in modelling [1] and to study Belief Propagation algorithms [14].

For each $\theta \in \Theta$, the expected value of the statistic t is the vector $\mathbb{E}_\theta[t] = \mu$, which is referred to as the moments of the MRF under θ. The set of all moments corresponding to MRFs based on t is

$$\mathcal{M} = \{\mu \in \mathbb{R}^{|E|} \mid \exists\theta \in \Theta, \mathbb{E}_\theta[t] = \mu\}.$$

As we will see, the set of moments $\mathcal{M}$ is also a coordinate system for $\mathcal{F}$. Thus an MRF based on t can be indexed by an exponential parameter θ as $p(\cdot;\theta)$ or the corresponding moment parameter μ as $p(\cdot;\mu)$. As with exponential parameters, moment parameters can be expressed in partitioned form $\mu = (\mu_D, \backslash\mu_D)$. In addition, we can specify an MRF based on t with mixed coordinate notation as $(\mu_D, \backslash\theta_D)$. *Information geometry* studies the structure of $\mathcal{F}$ as a statistical manifold parameterized by the dual coordinate systems Θ and $\mathcal{M}$ [2], [1], and is a powerful tool in the study of graphical models, particularly in learning the graphical structure of an MRF [7].

For a subset of sites U, we let t_U, θ_U, and μ_U be the components of t, θ, and μ, respectively, corresponding to edges in E_U. We say that a MRF on the subgraph G_U based on t_U is a *reduced MRF*. Such a reduced MRF can be parameterized by either θ_U or μ_U. The main result of this paper is that for a given θ and any subset of sites U, the reduced MRF on G_U indexed by exponential parameter θ_U has higher entropy than the marginal distribution of X_U under θ. The Maximum Entropy principle states that the reduced MRF on G_U with moment coordinate μ_U has greater entropy than the marginal distribution, thus providing a first upper bound to the marginal entropy. However, an exponential family is typically expressed in exponential coordinates and the mapping between exponential and moment coordinates is nontrivial [14]. Also, from a theoretical perspective, the new result based on preserving the exponential coordinates is complementary to the known result based on preserving moments. Indeed, our main result follows by showing that the entropy of the reduced MRF with exponential coordinate θ_U is an upper bound to that of the reduced MRF with moment coordinate μ_U.

We are able to show this upper bound by using a property of Markov random fields that we prove in Section IV, that states that the entropy of a family of MRFs is monotone decreasing in the exponential parameters. This property, which we call Monotonicity, is a consequence of the convexity of the log-partition function and the fact that the entropy of an exponential distribution is the first-order Taylor series approximation to $\Phi(0)$ about the given exponential parameter θ. Convexity of Φ has been useful in performing inference in MRFs [14] and in exploiting the dual relation between Θ and $\mathcal{M}$ that we briefly describe in the next section [2]. We can directly applying Monotonicity by considering the exponential coordinate $\bar{\theta} = (\theta_U, 0)$. However, the vector $\bar{\theta}$ applies to the entire field, not just nodes in U.[6] This, then, does not provide a direct comparison with the marginal entropy of X_U. Hence Monotonicity alone is insufficient to prove our main result. Using the convexity of Φ, we can show that θ_U is component-wise smaller than the coordinates θ^*_U corresponding to moments μ_U for the reduced MRF on G_U, which by Monotonicity gives us our result.

The outline of the paper is as follows. In Section II we introduce and define all terms used. In Section III we formally set up the problem and give the main theorem. In Section IV we state and discuss Monotonicity. In Section V we state the Maximum Entropy Principle and the known upper bound based on preserving moments; we then argue that the exponential coordinates θ^*_U corresponding to the moments μ_U is component-wise greater than the coordinate θ_U, thus by Monotonicity establishing the main result. Proofs are only sketched.

II. Background

A random field is a finite collection of random variables $\mathbb{X} = (X_1, \cdots, X_N)$ indexed by V. For each $i \in V$, X_i takes a value x_i in discrete state space $\mathcal{X}_i$. For a subset of nodes $U \subset V$, the subfield $X_U = (X_i \mid i \in U)$ assumes a value $x_U = (x_i \mid i \in U)$ in state space $\mathcal{X}_U = \prod_{i \in U} \mathcal{X}_i$. An instantiation $\mathbf{x}$ of $\mathbb{X} = X_V$ is called an *image*, where $\mathcal{X} = \mathcal{X}_V$ is the space of all images. We let E be a subset of all pairs of nodes in V. Then $G = (V, E)$ is a graph with nodes V and undirected edges E. A random field X is said to be *Markov* with respect to G if for every pair of nodes i and j not joined by an edge, the random variables X_i and X_j are conditionally independent given the remaining variables $X_{V\backslash\{i,j\}}$ [10]. By the Hammersley-Clifford theorem [10], the probability distribution for a positive[7] MRF X can be expressed as the product of factors defined on subsets of the random variables, i.e., it can be expressed as an exponential family. In this paper we consider only positive MRFs. The entropy of a Markov random field based on t with exponential parameter θ is

$$\begin{aligned} H(X;\theta) &= -\sum_{x\in\mathcal{X}} p(x;\theta)\log p(x;\theta) \\ &= \Phi(\theta) - \mathbb{E}_\theta\left[t(X)\right]^T \theta. \end{aligned}$$

[4] The sign restriction on θ is required for Theorem 4.1.

[5] $\backslash\theta_D$ are the coordinates of θ not in θ_D.

[6] The graph corresponding to $\bar{\theta}$ is the subgraph G_U plus a number of isolated nodes.

[7] That is, $p(x) > 0$ for all $x \in \mathcal{X}$.

The statistic t is said to provide a *minimal representation* of $\mathcal{F}$ if the $\{t_{ij}\}$ are affinely independent. In this case, it is well-known that the mapping $p : \Theta \longrightarrow \mathcal{F}$ defined by $\theta \longmapsto p(\cdot;\theta)$ is one-to-one. It is clear that the map $p(\cdot;\theta)$ is infinitely differentiable. Therefore if Θ is an open subset of $\mathbb{R}^{|E|}$, then $\mathcal{F}$ is a statistical manifold with coordinate system Θ [1]. For two exponential parameter vectors θ_1 and θ_2, we say $\theta_2 \succeq \theta_1$ if $(\theta_2)_{ij} \geq (\theta_1)_{ij}$ for all $\{i,j\} \in E$ and $\theta_2 \succ \theta_1$ if $\theta_2 \succeq \theta_1$ and $(\theta_2)_{ij} > (\theta_1)_{ij}$ for some edge $\{i,j\} \in E$. For a given exponential parameter θ',

$$\hat{\Phi}_{\theta'}(\theta) \triangleq \Phi(\theta') + \nabla\Phi(\theta')^T(\theta - \theta') \qquad (1)$$

is a hyperplane tangent to Φ at θ'. Also, $\hat{\Phi}_{\theta'}(\theta)$ is the first-order Taylor series approximation to $\Phi(\theta)$ about the parameter θ'. The gradient inequality for convex functions states that $\hat{\Phi}_{\theta'}(\theta) \leq \Phi(\theta)$ for all $\theta', \theta \in \Theta$, where the inequality is strict if and only if t is minimal for $\mathcal{F}$ [11].

It is straightforward to show that for $\theta \in \Theta$,

$$\nabla\Phi(\theta) = \mathbb{E}_\theta\left[t(X)\right]$$

and

$$\nabla^2\Phi(\theta) = \mathbb{E}_\theta\left[tt^T\right] - \mathbb{E}_\theta\left[t\right]\mathbb{E}_\theta\left[t\right]^T.$$

That $\nabla^2\Phi(\theta)$ is a covariance matrix implies that $\Phi(\cdot)$ is a convex function of θ, and if t is a minimal representation of $\mathcal{F}$, then Φ is *strictly convex* [3]. Here we make the assumption that the statistic t is *positively correlated* in the sense that for all $\theta \succeq 0$, $\mathrm{cov}_\theta\left[t_{ij}t_{kl}\right] \geq 0$, for all $\{i,j\},\{k,l\} \in E$. This condition is satisfied, for example, by the ferromagnetic Ising model [6]. In this case, for $\theta_2 \succeq \theta_1$, we have $\nabla\theta_2 \succeq \nabla\theta_1$, where again the inequalities are strict for minimal t.

The *mean parameter map* $\Lambda : \Theta \longrightarrow \mathcal{M}$ is defined as $\Lambda(\theta) = \nabla\Phi(\theta)$. It can be shown that if t is a minimal representation, then Λ is a bijection [3]. Hence, there exists a bijection between the set of mean parameters $\mathcal{M}$ and the manifold $\mathcal{F}$. Thus Θ and $\mathcal{M}$ are dual coordinate systems for $\mathcal{F}$.

Certain special subsets of the manifold of MRFs $\mathcal{F}$ can be specified by fixing a subset of the components of either the exponential parameter vector or the moment parameter vector. In particular, a subset $\mathcal{F}'$ defined by fixing a subset of exponential coordinates to a certain value and allowing the remaining coordinates to vary is called an *e-flat submanifold* [1]. Similarly, a subset $\mathcal{F}''$ defined by fixing a subset of the mean coordinates and varying the remaining components is called an *m-flat submanifold* [1]. In Section V we consider a fixed $\theta \in \Theta$ and the corresponding $\mu = \Lambda(\theta)$ indexing some initial MRF on G. For some $U \subset V$, we will then consider the e-flat submanifold

$$\mathcal{F}' = \{p_G(\cdot;\theta') \mid \theta' \in \Theta, \backslash\theta'_U = 0\}$$

of MRFs where the components of the parameter vectors corresponding to edges not contained in U are set to zero. We will also consider the m-flat submanifold

$$\mathcal{F}'' = \{p(\cdot;\mu'') \mid \mu'' \in \mathcal{M}, \mu''_U = \mu_U\}$$

of MRFs where the moment coordinates corresponding to edges in E_U are equal to the moments μ_U.

III. Setup and Statement of Main Results

In the succeeding sections we assume a given distribution p, indexed by either a fixed but arbitrary exponential vector $\theta \in \Theta$ or the corresponding moment vector $\Lambda(\theta) = \mu$.

For a subset $U \subset V$ of nodes, there are two types of distributions that we consider on the subfield X_U and each type refers in some way to the joint distribution p. The first is the marginal distribution

$$p_G(x_U;\theta) = \sum_{x_{V\setminus U}} p(x_U, x_{V\setminus U}).$$

The subscript G on the distribution is to indicate the graph with respect to which the MRF is defined, with the understanding that the exponential parameter will have one component for each edge of G. The second class of distributions on X_U second are *reduced MRF* distributions in which we take X_U to be an MRF on the induced subgraph $G_U = (U, E_U)$ based on t_U. We let $\Theta(G_U)$ and $\mathcal{M}(G_U)$ be the sets of admissable exponential and moment parameter vectors, respectively, for the family of MRFs based on t_U. We note that if t is minimal for the family of MRFs on G, then the subvector t_U is minimal for the family of MRFs $\mathcal{F}(G_U)$ on G_U. Given the dual coordinates systems $\Theta(G_U)$ and $\mathcal{M}(G_U)$, there are two ways to specify a reduced MRF on X_U. We can do so by indicating either the exponential parameter vector θ_U or the moment parameter vector μ_U, which are related via $\Lambda(\theta_U) = \mu_U$. Specified in exponential coordinates $\theta' \in \Theta(G_U)$, the reduced MRF distribution has the form

$$p_{G_U}(x_U;\theta') = \exp\{\langle\theta', t_U(x)\rangle - \Phi_U(\theta')\},$$

where $\Phi_U(\cdot)$ is the log-partition function for reduced MRFs defined on G_U. Note that the subscript to p is G_U. For $\mu' \in \mathcal{M}(G_U)$, the reduced MRF on G_U is specified by $p_{G_U}(x_U;\mu')$.

As stated below, the main result of this paper is that for any subset of nodes $U \subset V$, the entropy of the reduced MRF on U with exponential parameter θ_U is an upper bound to the entropy of the marginal distribution on U.

Theorem 3.1: Let $G = (V, E)$ be a graph and $p = p(\cdot;\theta)$ be the distribution of a Markov random field X_V based on a positively correlated, minimal statistic t and exponential parameter θ. Let X_{G_U} denote a reduced MRF on G_U. Then, for any nonempty subset of nodes $U \subset V$,

$$H(X_{G_U};\theta_U) \geq H(X_{G_U};\mu_U) \geq H(X_U;\theta).$$

IV. Monotonicity of Entropy

In this section we state and discuss the proof that the entropy of a Markov random field $p = p_G(\cdot;\theta)$ is monotone decreasing in the exponential parameters. Monotonicity is a useful property that can be used to establish new relationships between exponential and moment parameters.

Theorem 4.1 (Monotonicity): Let $X \sim p(\cdot;\theta)$ be an MRF on a graph G based on positively correlated statistic t. Then, for $\theta_1, \theta_2 \in \Theta$, $\theta_1 \prec \theta_2$, we have that

$$H_G(X;\theta_1) \geq H_G(X;\theta_2),$$

where the inequality is strict if and only if t is minimal.

The monotonicity of entropy can be established by rewriting the entropy of an MRF with exponential parameter θ as

$$H_G(X;\theta) = \Phi(\theta) + \nabla\Phi(\theta)^T(0-\theta) = \hat{\Phi}_\theta(0).$$

Thus $H_G(X;\theta)$ equals the first-order Taylor series approximation for $\Phi(\theta')$ about the point θ evaluated at $\theta' = 0$. Now suppose that $\theta_2 \succ \theta_1$. By the gradient inequality of convex functions, $\hat{\Phi}_{\theta_2}(\theta_1)$ is an underestimator of $\Phi(\theta_1)$. Therefore, as illustrated in Figure 1, the hyperplane

$$\hat{\Phi}_{\theta_2,\theta_1}(\theta) \triangleq \hat{\Phi}_{\theta_2}(\theta_1) + \nabla\Phi(\theta_1)^T(\theta - \theta_1)$$

lies beneath the hyperplane $\hat{\Phi}_{\theta_1}(\theta)$ for all θ. In particular, $\hat{\Phi}_{\theta_2,\theta_1}(0) < \hat{\Phi}_{\theta_1}(0) = H(X;\theta_1)$. Note also that $\hat{\Phi}_{\theta_2,\theta_1}$ is a hyperplane parallel to $\hat{\Phi}_{\theta_1}$, shifted down by $\Phi(\theta_1) - \hat{\Phi}_{\theta_2}(\theta_1)$. Since $\theta_2 \succ \theta_1$, the convexity of Φ implies that $\nabla\Phi(\theta_2) \succeq \nabla\Phi(\theta_1)$. Since $\hat{\Phi}_{\theta_2}$ and $\hat{\Phi}_{\theta_2,\theta_1}$ intersect at θ_1, this means that as we move from θ_1 to $\theta' = 0$, $\hat{\Phi}_{\theta_2}$ decreases no slower than $\hat{\Phi}_{\theta_2,\theta_1}$. This implies that $H_G(X;\theta_2) = \hat{\Phi}_{\theta_2}(0) \leq \hat{\Phi}_{\theta_2,\theta_1}(0) \leq H_G(X;\theta_1)$, which concludes the proof of Theorem 4.1.

An immediate application of Monotonicity is that we can upper bound the entropy of a Markov random field by setting some of the exponential parameters to zero.

Corollary 4.2: Let $\theta \in \Theta$ be an exponential parameter vector for an MRF on $G = (V,E)$, let U be a subset of V, and let $\bar{\theta} = (\theta_U, 0)$ be the coordinate vector obtained by setting the components for edges outside of U to zero. Then,

$$H_G(X;\theta) \leq H_G(X;\bar{\theta})$$

where the inequality is strict if and only if t is minimal.

Note that under $\bar{\theta}$, X_U and $X_{V\setminus U}$ are independent, $X_{V\setminus U}$ is uniformly distributed on $\mathcal{X}_{V\setminus U}$ and $H_G(X_U;\bar{\theta}) = H_{G_U}(X_U;\theta_U)$. It follows that $H_G(X;\bar{\theta}) = H_{G_U}(X_U;\theta_U) + |V \setminus U|\log_2|\mathcal{X}|$, hence we cannot say anything about the relationship between $H_G(X_U;\theta)$ and $H_{G_U}(X_U;\theta_U)$.

V. Upper Bounds: proof of Theorem 3.1

In this section we fix an exponential parameter θ which induces an MRF $p_G(X;\theta)$ on the graph G and the corresponding moment parameter $\mu = \Lambda(\theta)$. In the last section we used Monotonicity to argue that removing edges from a graph increases the entropy of an MRF. We prove Theorem 3.1 in three steps. In the first, we show that $\mu_U \in \mathcal{M}(G_U)$, in the second, we show that $H_{G_U}(X_U;\mu_U) \geq H_G(X_U;\theta)$, and in the third, we show that $H_{G_U}(X_U;\theta_U) \geq H_{G_U}(X_U;\mu_U)$. Consider the e-flat submanifold $\mathcal{F}'$ and m-flat submanifold $\mathcal{F}''$ introduced in Section II. An MRF $p_G(\cdot;\theta') \in \mathcal{F}'$ has an exponential parameter whose components corresponding to edges not contained in U are zero.

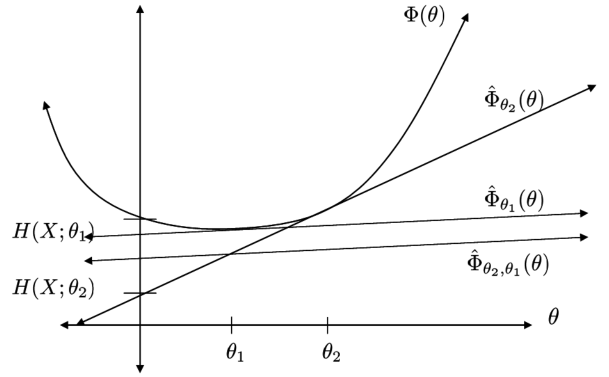
$\Phi(\theta)$
$\hat{\Phi}_{\theta_2}(\theta)$
$\hat{\Phi}_{\theta_1}(\theta)$
$H(X;\theta_1)$
$\hat{\Phi}_{\theta_2,\theta_1}(\theta)$
$H(X;\theta_2)$
θ
θ_1
θ_2

Fig. 1. The entropy of a Markov random field can be expressed as a Taylor series approximation of a convex function $\Phi(\cdot)$.

In partition form, these exponential vectors can be expressed as $\theta' = (\theta'_U, 0)$. An MRF $p_G(\cdot;\mu'') \in \mathcal{F}''$ is parameterized by a moment vector whose coordinates for edges inside U are equal to the corresponding coordinates in the moment vector μ parameterizing the original MRF. In partition form, these moment vectors can be expressed as $\mu'' = (\mu_U, \backslash\mu''_U)$.

It is not immediately clear whether a given vector $\mu_U \in \mathbb{R}^{|E_U|}$ is a valid moment coordinate for MRFs on G_U based on t_U. In the information geometry literature, the e-flat submanifold $\mathcal{F}'$ and the m-flat submanifold $\mathcal{F}''$ are what are known as *orthogonal* submanifolds [2]. It follows that $\mathcal{F}'$ and $\mathcal{F}''$ intersect uniquely at an MRF p^*_G with mixed coordinate $(\mu_U, \backslash\theta^*_U = 0)$ and exponential coordinate $\theta^* = (\theta^*_U, 0)$. This is the m-projection of $p_G(\cdot;\theta)$ onto the subgraph $\bar{G} = (V, E \setminus E_U)$ [1]. Though $p^*_G = p_G(\cdot;\theta^*)$ is defined on the original graph G, the subfields X_U and $X_{V\setminus U}$ are independent under p^*_G, so that

$$p_G(X_U;\theta^*) = p_{G_U}(X_U;\theta^*_U)$$

and since t_U is minimal, this shows that $\mu_U \in \mathcal{M}(G_U)$.

Since μ_U is a valid moment parameter for reduced MRFs on G_U, for the second step we need to show that $H_{G_U}(X_U;\mu_U)$ is an upper bound to $H_G(X_U;\theta)$. To do this we can use the well-known maximum entropy principle for exponential families [5], a slight variation of which is given below.

Proposition 5.1 (Maximum Entropy): Let $U \subset V$ be a subset of nodes, let $\mu_U \in \mathcal{M}(G_U)$, and let $\mathcal{P}_{\mu_U}$ be the set of probability distributions on X_U satisfying

$$\mathbb{E}_p\left[t_U(X)\right] = \mu_U.$$

If $p \in \mathcal{P}_{\mu_U}$, then

$$H_p(X_U) \leq H_{G_U}(X_U;\mu_U),$$

with equality if and only if $p = p_{G_U}(X_U;\mu_U)$.

As the final step we show that the exponential coordinate vector θ^*_U for the moment-matching MRF on G_U is component-wise larger than the subvector θ_U of the original exponential coordinate. We do this by showing that

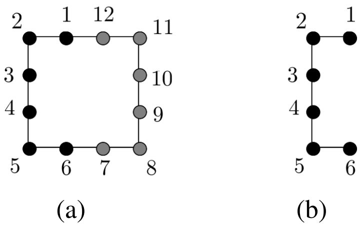

Fig. 2. (a) Original cycle on which Ising model defined. Subset U indicated in black. (b) Induced subgraph G_U on which upper bound $H_{G_U}(X_U;\theta_U)$ is based.

$\nabla\Phi(\theta^*_U) \succeq \nabla\Phi(\theta_U)$. Since t is minimal, Φ is strictly convex, hence the gradient $\nabla\Phi$ is strictly monotone. It follows that $\theta^*_U \succeq \theta_U$. This is summarized in the following theorem.

Theorem 5.2: Let $G = (V, E)$ be an undirected graph, let t be a positively correlated, minimal statistic MRFs on G, and let $\theta \in \Theta$ be an exponential parameter for MRFs on G with corresponding mean parameter vector μ. For a subset of nodes $U \subset V$, let θ^*_U be an exponential parameter vector for MRFs on G_U with corresponding moment vector μ_U. Then

$$\theta^*_U \succ \theta_U.$$

By Monotonicity of entropy, this implies that $H_{G_U}(X_U;\theta_U)$ is greater than $H_{G_U}(X_U;\mu_U)$, concluding the proof of Thm. 3.1.

VI. Example

Consider the twelve-node cycle with site set $V = \{1,\cdots,12\}$ and edge set $E = \{\{i,i+1\} \mid i = 1,\cdots,12\}$, $(12+1 \equiv 1)$ shown in Figure 2. For each $i \in V$, X_i takes values in $\{-1,1\}$. For each edge $\{i,i+1\}$ the statistic is $t_{i,i+1} = x_i x_{i+1}$ and the exponential parameter $\theta_{ij} = \theta$. This is an example of an Ising model [4]. Letting U be the nodes $\{1,\cdots,6\}$, it is straightforward to compute the marginal entropy $H_G(X_U;\theta)$ and the moments μ_U, e.g., using [4]. As we can see from Fig 3, the upper bound $H_{G_U}(X_U;\mu_u)$ is very close to the true value $H_G(X_U;\theta)$ for the entire range of θ. The upper bound $H_{G_U}(X_U;\theta)$ is close for low and high values of θ. For low values, the coupling between the nodes is weak, so not much is lost by separating U from the rest of the graph. For high values of θ, the coupling is strong enough that distributions on both the cycle and chain are becoming more concentrated on the all -1s and all 1s configurations. For intermediate values of θ, the upper bounds are more loose, especially the one obtained by preserving exponential rather than moment parameters.

VII. Discussion

We have shown that for an MRF satisfying a positive correlation condition, preserving the exponential coordinates on an induced subgraph provides an upper bound to the marginal entropy of the corresponding random variables. This result is complementary to a known result on preserving the moment coordinates. To do this, we have shown the new result that the entropy of a positive correlation MRF is monotone decreasing in the exponential parameters.

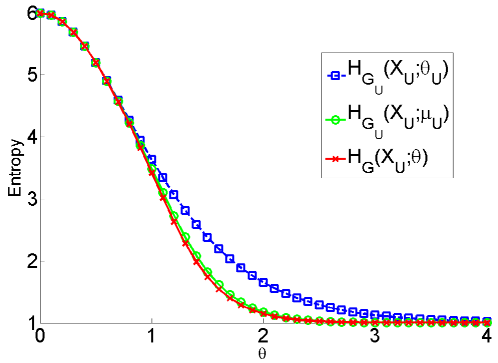

Fig. 3. Plot of $H_G(X_U;\theta)$, $H_{G_U}(X_U;\theta_U)$ and $H_{G_U}(X_U;\mu_U)$ for example shown in Fig 2.

This monotonicity result takes advantage of the convexity of the log-partition function, illustrating a new way in which this important quantity can be used to shed light on information theoretic properties of the family of MRFs based on a given statistic.

Acknowledgment

The authors would like to thank Martin Wainwright for helpful discussion.

References

[1] S. Amari, "Information Geometry on Hierarchy of Probability Distributions," *IEEE Trans. Info. Theory*, vol. 47, no. 5, July 2001.

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

[3] O.E. Barndorff-Nielson, *Information and Exponential Families*, Chichester, U.K.: Wiley, 1978.

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

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

[6] R.B. Griffiths, "Correlations in Ising Ferromagnets I.", *Jrnl. Math Physics*, 8 (478), 1967.

[7] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," *IEEE Trans. PAMI*, vol. 19, no. 4, pp. 28-41, Mar. 1997.

[8] 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.

[9] M. G. Reyes, X. Zhao, D. L. Neuhoff, T. N. Pappas, "Lossy Compression of Bilevel Images Based on Markov Random Fields," *ICIP*, San Antonio, TX, September 2007.

[10] S. Lauritzen, *Graphical Models*, Oxford University Press, 1990.

[11] R. T. Rockafeller, *Convex Analysis*, Princeton U. Press, Princeton, NJ.

[12] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky, "A New Class of Upper Bounds on the Log Partition Function," *IEEE Trans. Info. Theory*, vol. 51, no. 7, July 2005.

[13] M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," *IEEE Trans. Info. Theory*, vol. 49, no. 5, May 2003.

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