Maximum Pseudo-Likelihood Estimation Minimizes Conditional Description Length

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

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

***Abstract*—In this paper we discuss a method, which we call Minimum Conditional Description Length (MCDL), for estimating the parameters of a subset of sites within a Markov random field. We assume that the edges are known for the entire graph $G = (V, E)$. Then, for a subset $U \subset V$, we estimate the parameters for nodes and edges in U as well as for edges incident to a node in U, by finding the exponential parameter for that subset that yields the best compression conditioned on the values on the boundary ∂U. Our estimate is derived from a temporally stationary sequence of observations on the set U. We discuss how this method can also be applied to estimate a spatially invariant parameter from a single configuration, and in so doing, derive the Maximum Pseudo-Likelihood (MPL) estimate.**

I. Introduction

A Markov random field (MRF), also referred to as a Gibbs distribution, is a probability distribution on the colorings of an undirected 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 [20]. One of the primary research areas for MRFs is the problem of model selection or parameter estimation, where the objective may either be to determine the parameters for known edges [1], determine the edges of the graph [8], or jointly find the edges and the parameters for those edges [13]. Markov fields are a natural class of models for many types of data, including images and social networks. In images, it is natural to assume a set of edges, for instance, those connecting the 4 or 8 nearest neighbors. And for social networks, neighbor relations are known. With these two applications in mind, this paper focuses on the first model selection problem, that of determining the parameters on known edges.

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 site values at individual nodes and the endpoints of the edges E of the graph.[2] A particular MRF is indexed by an exponential parameter vector θ that scales the corresponding components of t in the probability of a configuration $\mathbf{x}$, which is given by

$$p(\mathbf{x};\theta) \quad = \quad \exp\{\langle \theta, t(\mathbf{x})\rangle - \Phi(\theta)\}, \tag{1}$$

where $\langle , \rangle$ denotes inner product and $\Phi(\theta)$ is the *log-partition function*.

In the *model selection* problem considered in this paper, the set of edges E is known, as well as the statistic t, and we have to determine the exponential parameter θ that weights the corresponding components of the statistic for nodes and edges. Generally, estimation is performed from a temporal sequence of observations $\mathbf{x}^{1:n} \triangleq \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, from which an estimate $\hat{\theta}^n$ is obtained. While it is often assumed that the $\mathbf{x}^{(i)}$ are independent to simplify analysis, in fact it is sufficient to assume that $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ is stationary, which is what we assume in this paper.

A popular criterion for estimating a parameter within a family of candidate models is *Maximum Likelihood* (ML), which seeks the parameter $\hat{\theta}^n$ which maximizes the probability $p(\mathbf{x}^{1:n}; \tilde{\theta})$ of the observed data over all parameter vectors $\tilde{\theta}$ indexing probability distributions within the specified class of probability distributions. For Markov fields, the ML criterion reduces to finding the exponential parameter $\tilde{\theta}$ such that the expected statistic $\tilde{\mu} \triangleq \mu(\tilde{\theta}) \triangleq \mathbb{E}_{\tilde{\theta}}[t(X)]$ under the MRF induced by $\tilde{\theta}$, referred to as the *moment* of the MRF, equals the *empirical moment* $\hat{\mu}^n$ of $\mathbf{x}^{1:n}$, which is the average value $\frac{1}{n}\sum_{i=1}^{n} t(\mathbf{x}^{(i)})$ of the statistic from the n observations [10]. For a tractable graph, such as a tree or one that can be clustered into a tree with only moderate numbers of nodes per cluster, the moments can be exactly and efficiently determined with Belief Propagation (BP), an iterative message passing algorithm. Thus, one can compute moments $\{\tilde{\mu}\}$ for a set of candidates $\{\tilde{\theta}\}$ and choose the one whose moment $\tilde{\mu}$ most closely matches the observed empirical moment $\hat{\mu}^n$. For a general graph, however, BP is intractable and thus the moment $\tilde{\mu}$ cannot be computed exactly. This intractability can circumvented by approximating the moment $\tilde{\mu}$, with either an approximate variant of BP [20], or by sampling the MRF's corresponding to candidate $\tilde{\theta}$, e.g. with Gibbs sampling [9], [10], [11], and selecting the $\tilde{\theta}$ whose empirical moment $\hat{\tilde{\mu}}$ most closely matches that of the observed data.

An alternative method for making parameter estimation in MRFs tractable is Maximum *Pseudo-Likelihood* [3], which defines a different objective function that is tractable and hence can be solved exactly. Maximum Pseudo Likelihood (MPL) is based on the concept of a *Coding Method*, introduced by Besag [1]. Assuming a translation invariant statistic t as well as a translation invariant parameter θ, such that each site had the same conditional distribution conditioned on its *neighbors*, those sites connected to it by an edge, one chooses a subset $V_1 \subset V$ of sites such that no two sites in V_1 are neighbors in G. By the Markov property, the sites in V_1 are conditionally independent of one another conditioned on the sites in $V \setminus V_1$, permitting their conditional distribution to be

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

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

expressed as a product of single-site conditional probabilities. Thus, by conditioning on $\mathbf{x}_{V\setminus V_1}$, one can estimate θ through an analytically tractable objective function. MPL extends this idea by finding the parameter $\hat{\theta}^{MPL}$ that maximizes the *pseudo*-likelihood function

$$\mathrm{PL}(\mathbf{x};\tilde{\theta}) \quad = \quad \prod_{j=1}^{|V|} p(\mathbf{x}_j|\mathbf{x}_{V\setminus j};\tilde{\theta})$$

over candidate parameters $\tilde{\theta}$, or equivalently, the pseudo-log-likelihood function

$$\log \mathrm{PL}(\mathbf{x};\tilde{\theta}) \quad = \quad \sum_{j=1}^{|V|} \log p(\mathbf{x}_j|\mathbf{x}_{V\setminus j};\tilde{\theta}),$$

again assuming translation invariance, or spatial homogeneity, of t and θ. Again by the Markov property, these conditional probabilities simplify as conditional probabilities given the neighbors of each node. Much research has been done on MPL, and consistency of the MPL estimate $\hat{\theta}^{MPL}$ has been shown [12], [6]. An interpretation of MPL is that it finds the parameter $\hat{\theta}^{MPL}$ such that the induced conditional distributions of individual nodes best match the empirical conditional distributions of individual nodes.

The parameter estimation method proposed in the present paper, which we call Minimum Conditional Description Length (MCDL), can be understood as a generalization of Maximum Pseudo-Likelihood. Whereas the MPL method estimates a translation invariant parameter through observations $\mathbf{x}_{\bar{U}_1},\ldots,\mathbf{x}_{\bar{U}_n}$ of $n=|V|$ statistically identical subsets within a single observation $\mathbf{x}$, we propose MCDL as a method for estimating the parameter $\theta_{\bar{U}}$ within a single subset $\bar{U}$ from a sequence of observations $\mathbf{x}_{\bar{U}}^{(1)},\ldots,\mathbf{x}_{\bar{U}}^{(n)}$ on $\bar{U}$, where ∂U is the boundary or neighborhood of U and $\bar{U} = U \cup \partial U$ is the *closure* of U. We do not assume spatial homogeneity (translation invariance) of θ within G, but we do require temporal stationarity of $\mathbf{x}_{\bar{U}}^{(1)},\ldots,\mathbf{x}_{\bar{U}}^{(n)}$. Moreover, while in MPL the subsets U_j are single sites, here the only restriction we place on a subset U is that the *subgraph induced by* U, consisting of nodes and edges of G contained in U, be tractable with respect to BP.

The *Minimum Description Length* (MDL) [18] principle states essentially that the best model is that one that provides the best compression of the data. Since Markov fields are defined in terms of their conditional distributions, and since conditioning on the boundary of a subset renders the subfield within the subset conditionally independent of the subfield outside of the closure of the subset, MCDL is a natural extension of this for efficiently estimating the parameters $\theta_{\bar{U}}$ inducing the conditional distribution of X_U given $X_{\partial U}$. If subset U is tractable for BP, we can compute the conditional probability

$$p(\mathbf{x}_U^{(i)}|\mathbf{x}_{\partial U}^{(i)};\tilde{\theta}_{\bar{U}})$$

of a configuration $\mathbf{x}_U^{(i)}$ given the configuration $\mathbf{x}_{\partial U}^{(i)}$ on its boundary. Then, given a temporal sequence of configurations $\mathbf{x}_{\bar{U}}^{1:n} = (\mathbf{x}_{\bar{U}}^{(1)},\mathbf{x}_{\bar{U}}^{(2)},\ldots,\mathbf{x}_{\bar{U}}^{(n)})$ on the closure $\bar{U}$, we seek the parameter $\hat{\theta}_U = \hat{\theta}_{\bar{U}}^n$ that causes the conditional distribution of $\mathbf{x}_U$ given $\mathbf{x}_{\partial U}$ within the MRF modeled by $\hat{\theta}_{\bar{U}}$ to best approximate the empirical conditional distribution of the $(\mathbf{x}_U^{(i)} : 1 \le i \le n)$ conditioned on the corresponding values $(\mathbf{x}_{\partial U}^{(i)} : 1 \le i \le n)$ on the boundary. Thus for different candidate parameters $\tilde{\theta}_{\bar{U}}$ we compute the temporal average of the negative log likelihood

$$H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} -\log p(\mathbf{x}_U^{(i)}|\mathbf{x}_{\partial U}^{(i)};\tilde{\theta}_{\bar{U}}) \qquad (2)$$

and select the $\tilde{\theta}_{\bar{U}}$ that minimizes $H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$. It is important to note that while $\tilde{\theta}_{\bar{U}}$ is properly the parameters for all nodes and edges within the closure $\bar{U}$ of U, the conditional distribution $p(X_U|\mathbf{x}_{\partial U};\tilde{\theta}_{\bar{U}})$ of X_U given $\mathbf{x}_{\partial U}$ depends only on the parameters for nodes and edges within U and for those edges connecting U to ∂U. It is in this more restricted sense that we use $\tilde{\theta}_{\bar{U}}$ throughout this paper.

This average negative log-likelihood can be interpreted as an empirical cross entropy between the true conditional distribution induced by $\theta_{\bar{U}}$ and the candidate parameter $\tilde{\theta}_{\bar{U}}$. Note that if $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$ were independent, this would be the negative log likelihood and this method would produce the ML estimate for $\theta_{\bar{U}}$. With an optimal encoder, for example Arithmetic Coding (AC) [22], for each i the number of bits produced in encoding $\mathbf{x}_U^{(i)}$ conditioned on $\mathbf{x}_{\partial U}^{(i)}$ will be within 1 or 2 bits of $-\log p(\mathbf{x}_U^{(i)}|\mathbf{x}_{\partial U}^{(i)};\tilde{\theta}_{\bar{U}})$. In other words, deriving the estimate $\hat{\theta}_{\bar{U}}^n$ as the parameter subvector that minimizes cross-entropy is essentially equivalent to estimating $\theta_{\bar{U}}$ as the parameter that minimizes coding rate when conditionally coding X_U given $X_{\partial U}$ with conditional coding distribution induced by $\tilde{\theta}_{\bar{U}}$. Indeed, it is straightforward to show that in the limit as the number of temporal samples n tends to infinity, the empirical average $\frac{1}{n}\sum_{i=1}^{n} -\log p(\mathbf{x}_U^{(i)}|\mathbf{x}_{\partial U}^{(i)};\tilde{\theta}_{\bar{U}})$ converges to

$$H(X_U|X_{\partial U};\theta) + D(p(X_U|X_{\partial U};\theta_{\bar{U}})||p(X_U|X_{\partial U};\tilde{\theta}_{\bar{U}}))$$

for a given candidate parameter $\tilde{\theta}$.

Ultimately, this method would be applied to different subsets $U_1,\ldots,U_k$, yielding estimates $\hat{\theta}_{\bar{U}_1},\ldots,\hat{\theta}_{\bar{U}_k}$ for the conditional distributions of $X_{U_1},\ldots,X_{U_k}$ given their respective boundaries. In order to produce an estimate $\hat{\theta}$ of the full parameter vector, we would need a way to enforce consistency of the $\hat{\theta}_{\bar{U}_1},\ldots,\hat{\theta}_{\bar{U}_k}$ on nodes and edges contained in multiple $\bar{U}_j$. At the moment we focus on estimating $\theta_{\bar{U}}$ for a single subset U.

To reiterate, one way in which MCDL differs from MPL is in the stationarity or homogeneity assumptions used to obtain the statistics for estimation. The setting in which MPL is generally applied assumes a translation invariant exponential parameter θ on a regular graph, in particular where the set of sites V form a lattice, and where an estimate of the global parameter θ is obtained from a single observation $\mathbf{x}$ on V. We do not require spatial homogeneity of the parameter, though we do require temporal stationarity and estimate the parameter for a single subset from a temporal sequence of observations

on that subset. In other words, whereas we are proposing to estimate the parameters $\theta_{\bar{U}}$ through n observations $\mathbf{x}_{\bar{U}}^{(1)}, \ldots, \mathbf{x}_{\bar{U}}^{(n)}$ on given subset U and its boundary, the MPL method estimates a translation invariant parameter θ through observations $\mathbf{x}_{\bar{U}_1}, \ldots, \mathbf{x}_{\bar{U}_n}$ on n statistically identical subsets $U_1, \ldots, U_n$ and their boundaries within a single observation $\mathbf{x}$.

The proposed MCDL algorithm also differs from MPL in that it allows larger subsets U rather than single sites, and more conceptually, in the formulation of the objective function. We now digress for a moment to think about MPL in the context of these other two differences. A common remark in the literature is that while the pseudo-likelihood function is tractable it is viewed as an approximation to the (chain rule decomposition of) the true likelihood function $p(\mathbf{x}; \tilde{\theta})$ of the observed data. However, in the translation invariant setting of MPL analysis, rather than attempt to approximate the likelihood function, instead consider the MCDL objective function, the cross entropy

$$H^n(\tilde{\theta}) \quad \triangleq \quad \frac{1}{n} \sum_{i=1}^{n} - \log p(\mathbf{x}_i | \mathbf{x}_{\partial i}; \tilde{\theta}) \tag{3}$$

between the empirical conditional distributions of single sites and the single site conditional distributions induced by a candidate parameter $\tilde{\theta}$. Mathematically, we have the same objective function for a candidate parameter $\tilde{\theta}$. However, viewed through the lens of MCDL, this function now yields the parameter that achieves minimal conditional description length of a site conditioned on its neighbors, without recourse to anything 'pseudo' or approximate. Indeed, in the limit of a large lattice of sites V, Equation (3) above tends to

$$H^-(X; \theta) + D(p(X_0 | X_{\partial 0}; \theta) || p(X_0 | X_{\partial 0}; \tilde{\theta})), \tag{4}$$

where $H^-(X; \theta)$ is the *erasure entropy* [19], given by

$$H^-(X; \theta) = H(X_0 | X_{\partial 0}; \theta),$$

which is the information lost if X_0 is erased from X, or in other words, the minimal amount of information needed to describe it conditioned on the values of its neighbors. It should be noted that (4) is not the number of bits from a lossless code of X, as clearly $H^-(X) < H(X; \theta)$. Nonetheless, through the MCDL paradigm, the MPL estimate can be interpreted as minimizing the empirical coding rate of $\{\mathbf{x}_{U_i}\}$ conditioned on the values $\{\mathbf{x}_{\partial U_i}\}$ rather than as an approximation of the likelihood function. Since Markov/Gibbs fields are specified in terms of their local characteristics, i.e., their conditional distributions, it makes perfect sense that MPL would yield a consistent estimate of θ.

Moreover, casting MPL as a conditional description length problem, one can generalize from considering conditional distributions of single nodes to considering conditional distributions of larger subsets U_i. Then for an MRF induced by a translation invariant parameter θ, the objective function to be minimized is now

$$\frac{1}{n} \sum_{i=1}^{n} - \log p(\mathbf{x}_{U_i} | \mathbf{x}_{\partial U_i}; \tilde{\theta}).$$

As opposed to subsets U_i of size 1, using larger subsets will reduced the number of samples n, so in that sense could potentially have an adverse affect on convergence and therefore the accuracy of $\hat{\theta}^n$. On the other hand, as the subsets U_i become larger, the effect of conditioning is reduced relative to the inter-site interactions within the U_i and as a result the local characteristics within a U_i conditioned on its boundary ∂U_i will more closely approximate the local characteristics of the full distribution. In other words, it is worth examining the tradeoffs involved in using larger subsets. Moreover, considering larger subsets U_i allows for greater flexibility in the invariance required for this method to provide good estimates. For example, instead of requiring site invariance of the statistic and parameter, one could simply assume row invariance of the statistic and parameter in which case the subsets U_i would be different rows of the lattice.

We now return to MCDL and consider the task of showing that the estimate $\hat{\theta}_{\bar{U}}^n$ of $\theta_{\bar{U}}$ is consistent, that is, that $\hat{\theta}_{\bar{U}}^n \to \theta_{\bar{U}}$ as $n \to \infty$. A reasonable course of action would be to mimic as closely as possible the proofs of consistency of the MPL estimate [12], [6]. The only difference it seems is that in the MPL regime, the $X_{U_1}, \ldots, X_{U_n}$ are independent conditioned on their respective boundaries, whereas in our case the $X_U^{(1)}, \ldots, X_U^{(n)}$ are not independent conditioned on the boundaries. Both problems have the same objective function, however, so it remains to be seen just how much tweaking is required to extend the MPL results to the present paradigm.

In the rest of this paper, Section II provides background on MRFs. Section III discusses the use of BP in lossless coding, Section IV presents our algorithm for estimating the parameter within a subset, and Section V discusses an example where we apply MCDL to both temporally stationary observations on a single subset as well as spatially observations on multiple subsets of a single configuration.

II. Graphs and Markov Random Fields

At each site $i \in V$ there is random variable X_i assuming values in alphabet $\mathcal{X}_i$. For a given configuration $\mathbf{x} = \{x_i : i \in V\}$, 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 $\mathbf{x}$, and similarly for $t_i : \mathcal{X}_i \longrightarrow \mathbb{R}$. We say that $X = (X_i, i \in V)$ is an MRF based on t. The entire family of MRFs based on t is generated by introducing an exponential parameter $\theta = (\theta_i, i \in V; \theta_{ij}, \{i, j\} \in E)$ where for each node i, and neighbor $j \in \partial i$, θ_i and θ_{ij} scale the sensitivity of the distribution $p(\mathbf{x}) = p(\mathbf{x}; \theta)$ to the functions t_i and t_{ij}, respectively.

The conditional probability of a configuration $\mathbf{x}_U$ on subset $U \subset V$ given the values $\mathbf{x}_W$ on another subset $W \subset V$ is denoted $p(\mathbf{x}_U | \mathbf{x}_W; \theta)$. It is straightforward to check that $p(\mathbf{x}_U | \mathbf{x}_{\partial U}; \theta) = p(\mathbf{x}_U | \mathbf{x}_{V \setminus U}; \theta)$ for all U, $\mathbf{x}_U$, and $\mathbf{x}_{\partial U}$.

This is the *Markov Property*. The conditional distributions of random subfield X_U given a specific configuration $\mathbf{x}_{\partial U}$, or on the random subfield $X_{\partial U}$, are denoted $p(X_U|\mathbf{x}_{\partial U};\theta)$ and $p(X_U|X_{\partial U};\theta)$, respectively. Likewise $H(X_U|\mathbf{x}_{\partial U};\theta)$ and $H(X_U|X_{\partial U};\tilde{\theta})$ are the respective conditional entropies of X_U given a specific configuration $\mathbf{x}_{\partial U}$ or the random subfield $X_{\partial U}$.

It is straightforward to show the following.

Proposition 2.1:

$$p(\mathbf{x}_U|\mathbf{x}_{\partial U};\tilde{\theta}_{\bar{U}}) \;=\; \exp\{\langle t_{\bar{U}}(\mathbf{x}_U,\mathbf{x}_{\partial U}),\tilde{\theta}_{\bar{U}}\rangle - \Phi_{U|\mathbf{x}_{\partial U}}(\tilde{\theta}_{\bar{U}}\},$$

where

$$\Phi_{U|\mathbf{x}_{\partial U}}(\tilde{\theta}_{\bar{U}}) \;=\; \log\left[\sum_{\mathbf{x}'_U}\exp\{\langle t_{\bar{U}}(\mathbf{x}'_U,\mathbf{x}_{\partial U}),\tilde{\theta}_{\bar{U}}\rangle\}\right]$$

is the log partition function for the conditional distribution of X_U with boundary condition $\mathbf{x}_{\partial U}$. Note that the statistic $t_{\bar{U}}(\mathbf{x}'_U,\mathbf{x}_{\partial U})$ includes all components of t at least one argument of which is contained in U. Thus $p(\mathbf{x}_U|\mathbf{x}_{\partial U};\tilde{\theta}_{\bar{U}})$ does not depend on $\tilde{\theta}_{\partial U}$.

III. Belief Propagation and Minimum Description

In general, ones uses Belief Propagation (BP) [20] to compute $p(\mathbf{x};\theta)$ for a configuration $\mathbf{x}$. Since the inner product $\langle t(\mathbf{x}),\theta\rangle$ can be computed directly, BP is used to (indirectly) compute the normalizing constant, the log-partition function $\Phi(\theta)$. If G has no cycles, then $p(\mathbf{x};\theta)$ can be computed with complexity linear in the number of nodes in V. If G has cycles, one can compute $p(\mathbf{x};\theta)$ by grouping subsets of V into supernodes such that the new graph is acyclic [20]. 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, for example no more than 10. A subset U is said to be tractable if the subgraph induced by U is tractable. For tractable subset U, $p(\mathbf{x}_U|\mathbf{x}_{\partial U};\theta)$ can be computed for given configurations $\mathbf{x}_U$ and $\mathbf{x}_{\partial U}$. Specifically, the conditional probability distribution $p(X_U|\mathbf{x}_{\partial U};\theta)$ of X_U given the configuration $\mathbf{x}_U$ on ∂U can be computed exactly and efficiently.

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 encoded conditioned on $\mathbf{x}_{\partial U}$ using coding distribution $p(X_U|\mathbf{x}_{\partial U};\tilde{\theta}_{\bar{U}})$, then the average number of bits produced is

$$H(X_U|X_{\partial U};\theta_{\bar{U}}||X_U|X_{\partial U};\tilde{\theta}_{\bar{U}}) \stackrel{\Delta}{=} H(X_U|X_{\partial U};\theta_{\bar{U}}) + D(p(X_U|X_{\partial U};\theta_{\bar{U}})||p(X_U|X_{\partial U};\tilde{\theta}_{\bar{U}}))$$

where $D(p(X_U|X_{\partial U};\theta_{\bar{U}})||p(X_U|X_{\partial U};\tilde{\theta}_{\bar{U}}))$ is the *divergence* between $p(X_U|X_{\partial U};\theta_{\bar{U}})$ and $p(X_U|X_{\partial U};\tilde{\theta}_{\bar{U}})$ and is the *redundancy* in the code [7]. Clearly, then, the true parameter $\theta_{\bar{U}}$ eliminates the redundancy and achieves the minimal conditional description length. In [15], Arithmetic Coding (AC) was proposed as the optimal encoder and details on the use of AC in the encoding of an MRF are given in [14], [15], and [16].

IV. Model Selection by Conditioning

We now discuss the MCDL method for estimating the parameters $\theta_{\bar{U}}$ of a subset $\bar{U}$. For tractable subset U, recall that we can exactly compute the probabilities $\{p(\mathbf{x}_U^{(i)}|\mathbf{x}_{\partial U}^{(i)};\tilde{\theta}_{\bar{U}})\}$ since U is chosen to be tractable with respect to Belief Propagation. If $\tilde{\theta}_{\bar{U}}$ is the parameter for the conditional distribution used to encode $\mathbf{x}_U^{(i)}$ given $\mathbf{x}_{\partial U}^{(i)}$, then the codeword length for $\mathbf{x}_U^{(i)}$ conditioned on $\mathbf{x}_{\partial U}^{(i)}$ is approximately

$$-\log p(\mathbf{x}_U^{(i)} \mid \mathbf{x}_{\partial U}^{(i)};\tilde{\theta}_{\bar{U}})$$

To form the estimate $\hat{\theta}_{\bar{U}}^n$ from observations $(\mathbf{x}_U^{(1)},\mathbf{x}_{\partial U}^{(1)}),\ldots,(\mathbf{x}_U^{(n)},\mathbf{x}_{\partial U}^{(n)})$, we use BP to compute the empirical cross entropy given in (2) for a candidate parameter $\tilde{\theta}_{\bar{U}}$ and then seek to minimize $H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$ over $\tilde{\theta}_{\bar{U}}$. The following is straightforward to show.

Proposition 4.1:

$$H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}}) \;=\; \frac{1}{n}\sum_{i=1}^n \Phi_{U|\mathbf{x}_{\partial U}^{(i)}}(\tilde{\theta}_{\bar{U}}) - \langle\hat{\mu}_{\bar{U}}^n,\tilde{\theta}_{\bar{U}}\rangle$$

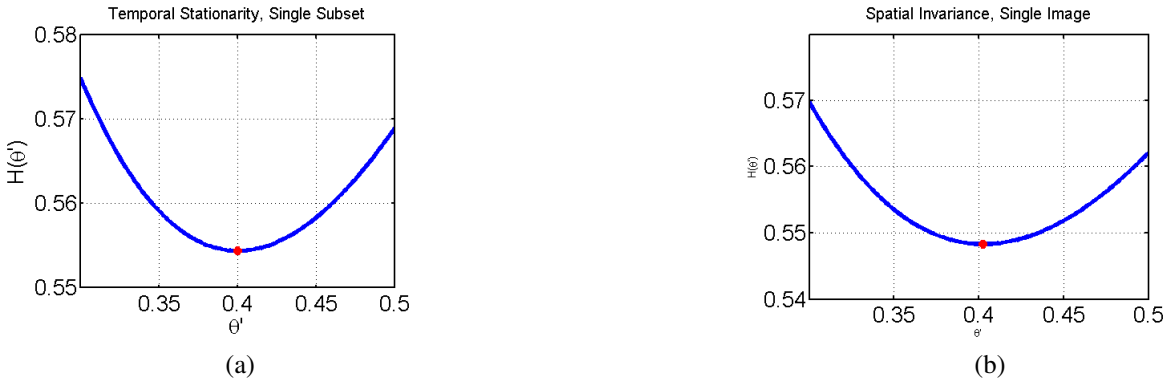
$$\nabla H_U^n(\tilde{\theta}_{\bar{U}}) \;=\; \frac{1}{n}\sum_{i=1}^n \tilde{\mu}_{\bar{U}|\mathbf{x}_{\partial U}^{(i)}} - \hat{\mu}_{\bar{U}}^n$$

where $\hat{\mu}_{\bar{U}}^n = \frac{1}{n}\sum_{i=1}^n t_{\bar{U}}(\mathbf{x}_U^{(i)},\mathbf{x}_{\partial U}^{(i)})$ is the empirical moment for the subset $\bar{U}$ given $\{\mathbf{x}_{\partial U}^{(i)}\}$ and $\tilde{\mu}_{\bar{U}|\mathbf{x}_{\partial U}^{(i)}}$ is the conditional moment for $p(X_U|\mathbf{x}_{\partial U}^{(i)};\tilde{\theta}_{\bar{U}})$.

It is well-known that $\Phi_{U|\mathbf{x}_{\partial U}^{(i)}}(\tilde{\theta}_{\bar{U}})$ is convex in $\tilde{\theta}_{\bar{U}}$ for each i, and as such so is $H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$. If the components of t are affinely independent, then $H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$ is strictly convex and thus has a unique minimum. Note that we are able to compute $\tilde{\mu}_{\bar{U}|\mathbf{x}_{\partial U}^{(i)}}$ with BP because U was chosen to be tractable. We can therefore apply a gradient descent algorithm to minimize $H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$ and obtain the estimate

$$\hat{\theta}_{\bar{U}}^n \;=\; \operatorname*{argmin}_{\tilde{\theta}_{\bar{U}}} H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$$

The MCDL algorithm for estimating the parameter subvector $\theta_{\bar{U}}$ within a subset U can be summarized as follows. Given $(\mathbf{x}_U^{(1)},\mathbf{x}_{\partial U}^{(1)}),\ldots,(\mathbf{x}_U^{(n)},\mathbf{x}_{\partial U}^{(n)})$, we initially compute the empirical moment $\hat{\mu}_{\bar{U}}^n = \frac{1}{n}\sum_{i=1}^n t_{\bar{U}}(\mathbf{x}_U^{(i)},\mathbf{x}_{\partial U}^{(i)})$. Then for a candidate parameter $\tilde{\theta}_{\bar{U}}$, we compute $H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$ and the $\{\tilde{\mu}_{\bar{U}|\mathbf{x}_{\partial U}^{(i)}}\}$ using BP. We then compute the gradient $\nabla H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}}) = \frac{1}{n}\sum_{i=1}^n \tilde{\mu}_{\bar{U}|\mathbf{x}_{\partial U}^{(i)}} - \hat{\mu}_{\bar{U}}^n$. Using a standard search, we select a new $\tilde{\theta}_{\bar{U}}$, and continue this process until a desired threshold for the norm of $\nabla H_{\bar{U}}^n(\tilde{\theta}_{\bar{U}})$ is attained [4].



(a) (b)

Fig. 1. True parameter is $\theta = .4$. Minimizing θ' is indicated in red in each case. Plot of empirical cross entropy for (a) temporally stationary sequence on a single subset, and (b) spatially invariant parameter on multiple subsets.

V. Example: Homogeneous Ising Model

We experimented with a (spatially) homogeneous Ising model, with edge parameter $\theta_{ij} = .4$ and node parameter $\theta_i = 0$ on a 200×200 square grid of sites, where each interior site is connected to its four nearest neighbors. The results are show in Figure 1. In (a), we consider a single subset U that is the middle row of the grid. The boundary ∂U consists of the row above and the row below. We generated a sequence of $n = 198$ configurations on G and computed

$$-\frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{x}_U^{(i)} | \mathbf{x}_{\partial U}^{(i)}; \theta') \tag{5}$$

for 161 evenly spaced θ' values ranging from .3 to .5 (granularity .00125). We found the minimizing θ' to be the true parameter value of .4.

In (b), we consider a single configuration $\mathbf{x}$ on G, and let $U_1, \ldots, U_n$ be the $n = 198$ rows of G with both an upper and lower boundary row. We computed

$$\frac{1}{n}\sum_{i=1}^{n} \log p(\mathbf{x}_{U_i} | \mathbf{x}_{\partial U_i}; \theta') \tag{6}$$

for the same 161 θ' values. In this case, the minimizing θ' to be .4025.

VI. Discussion

In this paper we have elaborated on the concept inherent in Maximum Pseudo-Likelihood, namely, that of using conditioning to simplify the task of parameter estimation, and have posed the problem as one of Minimum Conditional Description Length. The specific setting we have considered differs from the typical setting of MPL in that we have in mind temporal rather than spatial invariance, and we have here focused only on estimation of parameters within a single subset. Relaxing the spatial invariance assumption broadens the class of graphs and accompanying parameters to which we can apply this method. However, by requiring temporal stationarity we have imposed a new set of restrictions. More substantively, though, we feel that framing the problem as one of minimizing conditional description length is very natural given that Markov/Gibbs fields are specified by their conditional distributions. This leads to the same MPL estimate when applied to a single configuration generated by a spatially invariant parameter, and as such, we feel that the Minimum Conditional Description Length perspective places the Maximum Pseudo-Likelihood estimate on a firmer theoretical footing.

As we mentioned in the Introduction, though this method can be applied to obtain estimates $\hat{\theta}^n_{U_1}, \ldots, \hat{\theta}^n_{U_k}$ for the parameters within different subsets, there is potential inconsistency of these estimates for nodes and edges contained within the intersection of these subsets. While resolving this, for example through alternating direction method of multipliers [5], remains to be done, we still believe there is value in the notion of taking a large intractable Markov random field and decomposing it into tractable conditional random fields, on which good parameter estimates can be obtained efficiently and in which exact inference and prediction can be performed with respect to these parameters, conditioned on the boundaries of these subsets. Indeed, it was shown in [21] that if the MRF is on an intractable graph, such that suboptimal inference and prediction will be performed with respect to whatever parameters are available, then there can be benefits to incorrectly estimating the parameters. In our case, good estimates would be obtained on each tractable conditional random field, and exact inference could be performed with respect to these parameters, but they may not yield a consistent estimate of the global parameter.

Additionally, the MCDL method for parameter estimation introduced in this paper is complementary with our previous work in using cutsets to simplify the processing, in particular the compression, of intractable MRFs [15], [16], [17]. In these works, there is an initial lossless compression of a cutset of sites, followed by either estimation or optimal lossless conditional compression of the remaining sites given the values on the cutset. If a fixed cutset was to be used in one of these algorithms, then one could simply estimate the parameters of the tractable conditional subfields that would be estimated or compressed given the values on their boundaries.

REFERENCES

[1] J. Besag, "Spatial Interaction and the Statistial Analysis of Lattice Systems," *J. of Roy. Stat. Soc. B*, vol. 36, pp. 192-235, March 1974.

[2] J. Besag, "Statistical Analysis of Non-lattice Data," *J. of Roy. Stat. Soc. D*, vol. 24, pp. 179-195, Sept. 1975.

[3] J. Besag, "Efficiency of pseudo-likelihood estimation for simple Gaussian fields," *Biometrika*, Vol. 64, pp 616-618, 1977.

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

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed Optimization and Statistical Leaning via the Alternating Direction Method of Multipliers*, Foundations in Trends in Machine Learning, Vol. 3, pp. 1-122, 2011.

[6] F. Comets, "On Consistency of a Class of Estimators for Exponential Families of Markov Random Fields on the Lattice," *The Annals of Statistics*, Vol. 20, pp. 455-468, March 1992.

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

[8] I. Csiszar and Z. Talata, "Consistent Estimation of the Basic Neighborhood of Markov Random Fields," *The Annals of Statistics*, Vol. 34, pp. 123-145, February 2006.

[9] 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.

[10] C. J. Geyer and E. A. Thompson, "Constrained Monte Carlo Maximum Likelihood for Dependent Data," *Journal of the Royal Statistical Society B*, Vol. 54, 654-699, 1992.

[11] C. J. Geyer, "On the Convergence of Monte Carlo Maximum Likelihood Calculations," *Journal of the Royal Statistical Society B*, Vol. 56, 261-274, 1994.

[12] B. Gidas, "Consistency of maximum likelihood and pseudolikelihood estimators for Gibbs distributions,", *Stochastic Differential Equations with Applications to Electronic/Computer Engineering, Control Theory, and Operations Research* (W. Fleming and P. L. Lions, eds.), 1-17, Springer, Berlin.

[13] S. D. Pietra, V. D. Pietra, and J. Lafferty, "Inducing Features of Random Fields," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, No. 4, April 1997.

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

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

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

[17] M.G. Reyes and D.L. Neuhoff, *Cutset Width and Spacing for Reduced Cutset Coding of Markov Random Fields*, submitted to ISIT 2016 (also in Deep Blue repository, University of Michigan).

[18] J. Rissanen, "Modeling by Shortest Data Description", *Automatica*, Vol. 14, pp. 465-471, September 1978.

[19] S. Verdu and T. Weissman, "The Information Lost in Erasures", IEEE Tran. Info. Thy., Vol. 54, No. 11, November 2008.

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

[21] M. J. Wainwright, "Estimating the 'Wrong' Graphical Model: Benefits in the Computation Limited Setting", Journal of Machine Learning Research, Vol. 7, pp. 1829-1859, September 2006.

[22] 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.