On Learning Continuous Pairwise Markov Random Fields

Abhin Shah  Devavrat Shah  Gregory W. Wornell
Massachusetts Institute of Technology
{abhin, devavrat, gww}@mit.edu

Abstract

We consider learning a sparse pairwise Markov Random Field (MRF) with continuous valued variables from i.i.d samples. We adapt the algorithm of Vuffray et al. (2019) to this setting and provide finite-sample analysis revealing sample complexity scaling logarithmically with the number of variables, as in the discrete and Gaussian settings. Our approach is applicable to a large class of pairwise MRFs with continuous variables and also has desirable asymptotic properties, including consistency and normality under mild conditions. Further, we establish that the population version of the optimization criterion employed by Vuffray et al. (2019) can be interpreted as local maximum likelihood estimation (MLE). As part of our analysis, we introduce a robust variation of sparse linear regression à la Lasso, which may be of interest in its own right.

1 Introduction

1.1 Background

Markov random fields or undirected graphical models are an important class of statistical models and represent the conditional dependencies of a high dimensional probability distribution with a graph structure. There has been considerable interest in learning discrete MRFs in machine learning, statistics, and physics communities under different names (Chow and Liu, 1968; Abbeel et al., 2006; Negahban et al., 2012; Ackley et al., 1985; Sessak and Monasson, 2009). Bresler (2015) gave a simple greedy algorithm to learn arbitrary binary pairwise graphical models on p nodes and maximum node degree d with sample complexity \(O(\exp(\exp(\Theta(d))) \log p)\) and runtime \(\mathcal{O}(p^2)\). This improved upon the prior work of Bresler et al. (2013), with runtime \(\mathcal{O}(p^{d+2})\) by removing the dependence of d on the degree of the polynomial factor in runtime. Santhanam and Wainwright (2012) showed that only exponential dependence on d is required in the sample complexity and thus the doubly-exponential dependence on d of Bresler (2015) is provably suboptimal.

A recent work by Vuffray et al. (2019) learns t-wise MRFs over general discrete alphabets in a sample-efficient manner \(O(\exp(\Theta(d^{-1}) \log p))\) with runtime \(\mathcal{O}(p^t)\). The key to their proposal is a remarkable but seemingly mysterious objective function, the generalized interaction screening objective (GISO) which is an empirical average of an objective (that does not include a normalization factor) designed to screen an individual variable from its neighbors. While their approach can be formally extended to the continuous-valued setting, issues arise. First, as is, their work shows that, for the discrete setting, the condition for learning is satisfied by only the ‘edge’ parameters and their approach does not attempt to recover the ‘node’ parameters. Second, their condition for learning is cumbersome to verify as it is node-neighborhood-based and involves all the edges associated with the node.

In this work, we consider the problem of learning sparse pairwise MRFs from i.i.d. samples when the underlying random variables are continuous. The classical Gaussian graphical model is an example of this. There has been a long history of learning...
Gaussian MRFs, e.g. Graphical Lasso [Friedman et al., 2008] and associated recent developments e.g. [Misra et al., 2017; Kelner et al., 2019]. Another example is the following extension of the Ising model to the continuous case.

\[ f(x) \propto \exp \left( \sum_{i \in [p]} \theta^{(i)} x_i + \sum_{i \neq j \in [p]} \theta^{(ij)} x_i x_j \right), \tag{1} \]

where \( x = (x_1, \ldots, x_p) \) is a \( p \)-dimensional vector of continuous variables, \( x = (x_1, \ldots, x_p) \) is a realization of \( x \), and \( \theta^{(i)} \forall i \in [p], \theta^{(ij)} \forall i \neq j \in [p] \) are the parameters associated with the distribution.

Despite the progress on Gaussian graphical models, the overall progress for the generic continuous setting (including (1)) has been limited. In particular, the existing works for efficient learning require somewhat abstract, involved conditions that are hard to verify for e.g. incoherence (Yang et al., 2015; Tansey et al., 2015; Sun et al., 2013), dependency (Yang et al., 2013; Tansey et al., 2017; Sun et al., 2015), sparse eigenvalue (Yang et al., 2018), restricted strong convexity (Yuan et al., 2016; Suggala et al., 2017). The incoherence condition ensures that irrelevant variables do not exert an overly strong effect on the true neighboring variables, the dependency condition ensures that variables do not become overly dependent, the sparse eigenvalue condition and the restricted strong convexity imposes strong curvature condition on the objective function. Table 1 compares with the previous works on pairwise continuous MRFs with distribution of the form (1).

In summary, the key challenge that remains for continuous pairwise MRFs is finding a learning algorithm requiring (a) numbers of samples scaling as \( \exp(\Theta(d)) \) (in accordance with lower bound of Santhanam and Wainwright [2012]) and \( \log p \), (b) computation scaling as \( O(p^2) \), and (c) the underlying distribution to satisfy as few conditions as in the discrete setting.

### 1.2 Contributions

As the primary contribution of this work, we make progress towards the aforementioned challenge. Specifically, we provide desirable finite sample guarantees for learning continuous MRFs when the underlying distribution satisfies simple, easy to verify conditions (examples in Section 4.4). We summarize our contributions in the following two categories.

#### Finite Sample Guarantees.

We provide rigorous finite sample analysis for learning structure and parameters of continuous MRFs without the abstract

<table>
<thead>
<tr>
<th>Work</th>
<th>Approach</th>
<th>Conditions</th>
<th>#samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yang et al.</td>
<td>( \ell_1 ) regularized node conditional log-likelihood</td>
<td>1. Incoherence condition 2. Dependency condition 3. Bounded moments of the variables</td>
<td>( O(\text{poly}(d)\log p)) )</td>
</tr>
<tr>
<td>Tansey et al.</td>
<td>Group lasso regularized node conditional log-likelihood</td>
<td>4. Local smoothness of the log-partition function 5. Conditional distribution lies in exponential family</td>
<td>( \tilde{\omega}(p) \log p )</td>
</tr>
<tr>
<td>Sun et al.</td>
<td>Penalized score matching objective</td>
<td>1. Incoherence condition 2. Dependency condition 3. Certain structural conditions</td>
<td>( O(\text{poly}(d)) )</td>
</tr>
<tr>
<td>Suggala et al.</td>
<td>( \ell_1 ) regularized node conditional log-likelihood</td>
<td>1. Restricted strong convexity 2. Assumptions on gradient of the population loss 3. Bounded domain of the variables 4. Non-negative node parameters 5. Conditional distribution lies in exponential family</td>
<td>( O(\text{poly}(d)\log p)) )</td>
</tr>
<tr>
<td>Yuan et al.</td>
<td>( \ell_2,1 ) regularized node conditional log-likelihood</td>
<td>1. Restricted strong convexity 2. Bounded moment-generating function of variables</td>
<td>( O(\text{poly}(d)\log p)) )</td>
</tr>
<tr>
<td>This work</td>
<td>Augmented GISO (Section 3)</td>
<td>1. Bounded domain of the variables 2. Conditional distribution lies in exponential family</td>
<td>( O(\exp(\Theta(d)/\log p)) ) (Thm. 4.3)</td>
</tr>
</tbody>
</table>

Table 1: Comparison with existing works on pairwise continuous MRFs (beyond the Gaussian case) in terms of approach, conditions required and sample complexity: \( p \) is \# of variables, \( d \) is maximum node degree.
Table 2: Comparison with prior works on discrete MRFs in terms of asymptotic properties (consistency and normality), computational and sample complexities: \(p\) is \# of variables, \(d\) is maximum node degree.

<table>
<thead>
<tr>
<th>Result (pairwise)</th>
<th>Alphabet</th>
<th>Consistency (i.e. SLLN)</th>
<th>Normality (i.e. CLT)</th>
<th>#computations</th>
<th>#samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bresler et al. [2013]</td>
<td>Discrete</td>
<td>✓</td>
<td>(\mathcal{O}(p^{d+2}))</td>
<td>(O(\exp(d \log p)))</td>
<td></td>
</tr>
<tr>
<td>Bresler [2013]</td>
<td>Binary</td>
<td>✓</td>
<td>(\mathcal{O}(p^2))</td>
<td>(O(\exp(\exp(d)) \log p))</td>
<td></td>
</tr>
<tr>
<td>Klivans and Meka [2017]</td>
<td>Discrete</td>
<td>✓</td>
<td>(\mathcal{O}(p^2))</td>
<td>(O(\exp(d \log p)))</td>
<td></td>
</tr>
<tr>
<td>Vuffray et al. [2019]</td>
<td>Discrete</td>
<td>✓</td>
<td>(\mathcal{O}(p^2))</td>
<td>(O(\exp(d \log p)))</td>
<td></td>
</tr>
<tr>
<td>This Work</td>
<td>Continuous</td>
<td>✓</td>
<td>✓</td>
<td>(\mathcal{O}(p^2))</td>
<td>(O(\exp(d \log p)))</td>
</tr>
</tbody>
</table>

See Table 1 and 2 for a succinct comparison with prior works in the pairwise setting for continuous MRFs and discrete MRFs respectively.

### 1.3 Other related work

Discrete MRFs. After Bresler [2015] removed the dependence of maximum degree, \(d\), from the polynomial factor in the runtime (with sub-optimal sample complexity), Vuffray et al. [2016] achieved optimal sample complexity of \(O(\exp(\Theta(d)) \log p)\) for Ising models on \(p\) nodes but with runtime \(\mathcal{O}(p^4)\). Their work was the first to propose and analyze the interaction screening objective function. Hamilton et al. [2017] generalized the approach of Bresler [2015] for \(t\)-wise MRFs over general discrete alphabets but had non-optimal double-exponential dependence on \(d^{t-1}\). Klivans and Meka [2017] provided a multiplicative weight update algorithm (called the Sparsitron) for learning pairwise models over general discrete alphabets in time \(\mathcal{O}(p^2)\) with optimal sample complexity \(O(\exp(\Theta(d)) \log p)\) and \(t\)-wise MRFs over binary alphabets in time \(\mathcal{O}(p')\) with optimal sample complexity \(O(\exp(\Theta(d^{t-1}) \log p))\). Wu et al. [2018] considered an \(\ell_{2,1}\)-constrained logistic regression and improved the sample complexity of Klivans and Meka [2017] for pairwise models over general discrete alphabets in terms of dependence on alphabet size.

Gaussian MRFs. The problem of learning Gaussian MRFs is closely related to the problem of learning the sparsity pattern of the precision matrix of the underlying Gaussian distribution. Consider Gaussian MRFs on \(p\) nodes of maximum degree \(d\) and the minimum normalized edge strength \(\kappa\) (see Misra et al. [2017]). A popular approach, the Graphical Lasso [Friedman et al. 2008], recovered the sparsity pattern under the restricted eigenvalue and incoherence assumptions from \(O((d^2 + \kappa^{-2}) \log p)\) samples [Ravikumar et al. 2011] by \(\ell_1\) regularized log-likelihood estimator. The minimum required sample

---

4 Condition 1.1 effectively lower bounds the variance of a non-constant random variable and is an adequate condition to rule out certain singular distributions.
complexity was shown to be $\Omega(\log p/k^2)$ by Wang et al. (2010) via an information-theoretic lower bound. Misra et al. (2017) provided a multi-stage algorithm that learns the Gaussian MRFs with $O(d \log p / k^2)$ samples and takes time $O(p^{2d+1})$. A recent work by Kelner et al. (2019) proposed an algorithm with runtime $O(p^{d+1})$ that learns the sparsity pattern in $O(d \log p / k^2)$ samples. However, when the variables are positively associated, this algorithm achieves the optimal sample complexity of $O(\log p / k^2)$.

Continuous MRFs. Realizing that the normality assumption is restrictive, some researchers have recently proposed extensions to Gaussian MRFs that either learns transformations of the variables or learn the sufficient statistics functions. The non-paranormal (Liu et al. 2009) and the copula-based (Dobra et al. 2011) methods assumed that a monotone transformation Gaussianize the data. Rank-based estimators by Xue and Zou (2012) and Liu et al. (2012) used non-parametric approximations to the correlation matrix and then fit a Gaussian MRF.

There have been some recent works on learning exponential family MRFs for the pairwise setting. The subclass where the node-conditional distributions arise from exponential families was looked at by Yang et al. (2015) and the necessary conditions for consistent joint distribution were derived. However, they considered only linear sufficient statistics and they needed incoherence and dependency conditions similar to the discrete setting analyzed by Wainwright et al. (2006); Jalali et al. (2011). Yang et al. (2018) studied the subclass with linear sufficient statistics for edge-wise functions and non-parametric node-wise functions with the requirement of sparse eigenvalue conditions on their loss function. Tansev et al. (2013) extended the approach by Yang et al. (2015) to vector-space MRFs and non-linear sufficient statistics but still required the incoherence and dependency conditions similar to Wainwright et al. (2006); Jalali et al. (2011). Sun et al. (2015) investigated infinite dimensional exponential family graphical models based on score matching loss. They assumed that the node and edge potentials lie in a reproducing kernel Hilbert space and needed incoherence and dependency conditions similar to Wainwright et al. (2006); Jalali et al. (2011). Yuan et al. (2016) explored the subclass where the node-wise and edge-wise statistics are linear combinations of two sets of pre-fixed basis functions. They proposed two maximum likelihood estimators under the restricted strong convexity assumption. Suggala et al. (2017) considered a semi-parametric version of the subclass where the node-conditional distributions arise from exponential families. However, they required restricted strong convexity and hard to verify assumptions on gradient of the population loss.

**Useful notations.** For any positive integer $n$, let $[n] := \{1, \cdots, n\}$. For a deterministic sequence $v_1, \cdots, v_n$, we let $\mathbf{v} := (v_1, \cdots, v_n)$. For a random sequence $v_1, \cdots, v_n$, we let $\mathbf{v} := (v_1, \cdots, v_n)$. Let $\mathbf{1}$ denote the indicator function. For a vector $\mathbf{v} \in \mathbb{R}^n$, we use $v_i$ to denote its $i$th coordinate and $v_{-i} \in \mathbb{R}^{n-1}$ to denote the vector after deleting the $i$th coordinate.

We denote the $\ell_p$ norm ($p \geq 1$) of a vector $\mathbf{v} \in \mathbb{R}^n$ by $||\mathbf{v}||_p := \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p}$ and its $\ell_\infty$ norm by $||\mathbf{v}||_\infty := \max_{i} |v_i|$. For a vector $\mathbf{v} \in \mathbb{R}^n$, we use $|\mathbf{v}|_0$ to denote the number of non-zero elements ($\ell_0$ norm) of $\mathbf{v}$. We denote the minimum of the absolute values of non-zero elements of a vector $\mathbf{v} \in \mathbb{R}^n$ by $|\mathbf{v}|_{\min} := \min_{v_i \neq 0} |v_i|$. For a matrix $\mathbf{V} \in \mathbb{R}^{m \times n}$, we denote the element in $i$th row and $j$th column by $V_{ij}$ and the max norm by $||\mathbf{V}||_{\max} := \max_{i,j} |V_{ij}|$. All logarithms are in base $e$.

**Organization.** The remainder of this paper is organized as follows. In Section 2, we formulate the problem setup which consists of pairwise MRFs, the particular parametric form of interest, modeling assumptions, objectives, and a few additional notations. Next, in Section 3, we describe our algorithm ‘Augmented GISO’ where we first learn the graph structure and edge parameters, and then learn the node parameters. In Section 4, we provide our main technical results including equivalence of the population version of GISO and KL Divergence (Theorem 4.1), asymptotic consistency and normality of the GISO (Theorem 4.2), the simplified sufficient condition for learning (Condition 4.1), finite sample guarantees for learning structure (Theorem 4.3) and parameters (Theorem 4.4), connection of the GISO to the surrogate likelihood (Proposition 4.1), and a few examples of distributions naturally satisfying Condition 4.1. In Section 5, we conclude and discuss some directions for future work. See supplementary for the organization of the Appendix.

## 2 Problem Formulation

In this Section, we formulate the problem of interest and also introduce a few additional notations. Appendix 4 provides an illustrative example to ease these notations.

**Pairwise MRF.** Let $\mathbf{x} = (x_1, \cdots, x_p)$ be a $p$-dimensional vector of continuous random variables such that each $x_i$ takes value in a real interval $X_i$ and let $\mathcal{X} = \prod_{i=1}^{p} X_i$. Let $\mathbf{x} = (x_1, \cdots, x_p) \in \mathcal{X}$ be a realization of $\mathbf{x}$. For any $i \in \{p\}$, let the length of the interval $X_i$ be upper (lower) bounded by a known constant $b_i (b_i)$. Consider an undirected graph $G = ([p], E)$ where the nodes correspond to the
random variables in \( \mathbf{x} \), and \( E \) denotes the edge set. The MRF corresponding to the graph \( G \) is the family of distributions that satisfy the global Markov property with respect to \( G \). According to the Hammersley-Clifford theorem [Hammersley and Clifford, 1971], any strictly positive distribution factorizes with respect to its cliques. Here, we consider the setting where the functions associated with cliques are non-trivial only for the nodes and the edges. This leads to the pairwise MRFs with respect to graph \( G \) with density as follows: with node potentials \( g_i : \mathcal{X}_i \to \mathbb{R} \), edge potentials \( g_{ij} : \mathcal{X}_i \times \mathcal{X}_j \to \mathbb{R} \),

\[
f_x(\mathbf{x}) \propto \exp \left( \sum_{i \in [p]} g_i(x_i) + \sum_{(i,j) \in E} g_{ij}(x_i, x_j) \right).
\]

**Parametric Form.** We consider potentials in parametric form. Specifically, let

\[
g_i(\cdot) = \boldsymbol{\theta}^{(i)\top} \phi(\cdot) \quad \text{and} \quad g_{ij}(\cdot, \cdot) = \boldsymbol{\theta}^{(ij)\top} \psi(\cdot, \cdot),
\]

where \( \boldsymbol{\theta}^{(i)} \in \mathbb{R}^{k_i} \) is the vector of parameters associated with the node \( i \), \( \boldsymbol{\theta}^{(ij)} \in \mathbb{R}^{k_{ij}} \) is the vector of parameters associated with the edge \((i,j)\), the map \( \phi : \mathcal{X} \to \mathbb{R}^{k_i} \) is a basis of the vector space of node potentials, and the map \( \psi : \mathcal{X}^2 \to \mathbb{R}^{k_{ij}} \) is a basis of the vector space of edge potentials. We assume that \( \psi(x, y) \) can be written as the Kronecker product of \( \phi(x) \) and \( \phi(y) \), i.e., \( \psi(x, y) = \phi(x) \otimes \phi(y) \). This is equivalent to the function space assumption common in the literature [Yang et al., 2015; 2018; Suggala et al., 2017; Tansey et al., 2017] that the conditional distribution of each node conditioned on all the other nodes has an exponential family form (see Yang et al. [2015] for details). Further, let the basis functions be such that the resulting exponential family is minimal\(^5\).

A few examples of basis functions in-line with these assumptions are: (1) Polynomial basis with \( \phi(x) = (x^r : r \in [k_i]) \), \( \psi(x, y) = (x^r y^s : r, s \in [k_i]) \); (2) Harmonic basis with \( \phi(x) = (\sin(rx) ; \cos(rx) : r \in [k_i]) \), \( \psi(x, y) = (\sin(rx + sy) ; \cos(rx + sy) : r, s \in [k_i]) \).\(^6\)

For any \( r \in [k_i] \), let \( \phi_r(x) \) denote the \( r \)th element of \( \phi(x) \) and let \( \theta_r^{(i)} \) be the corresponding element of \( \boldsymbol{\theta}^{(i)} \). For any \( r, s \in [k_i] \), let \( \psi_{rs}(x, y) \) denote that element of \( \psi(x, y) \) which is the product of \( \phi_r(x) \) and \( \phi_s(y) \), i.e., \( \psi_{rs}(x, y) = \phi_r(x) \phi_s(y) \). Let \( \theta_{rs}^{(ij)} \) be element of \( \boldsymbol{\theta}^{(ij)} \) corresponding to \( \psi_{rs}(x, y) \). We also assume that \( \forall r \in [k_i], \forall x \in \cup_{i \in [p]} \mathcal{X}_i, \| \phi_r(x) \| \leq \phi_{\max}, \phi_r(x) \) is differentiable in \( x \), and \( |d\phi_r(x)/dx| \leq \phi_{\max} \).

Summarizing, the distribution of focus is

\[
f_x(x; \boldsymbol{\theta}) \propto \exp \left( \sum_{i \in [p]} \theta^{(i)\top} \phi(x_i) + \sum_{i \in [p], j \neq i} \theta^{(ij)\top} \psi(x_i, x_j) \right),
\]

where \( \boldsymbol{\theta} := (\theta^{(i)} \in \mathbb{R}^k : i \in [p]) \cup (\theta^{(ij)} \in \mathbb{R}^{k_{ij}} : i \in [p], j > i) \) is the parameter vector associated with the distribution. For any \( i \in [p], j > i \), define \( \theta^{(ij)} = \theta^{(ji)} \) i.e., both \( \theta^{(ij)} \) and \( \theta^{(ji)} \) denote the parameter vector associated with the edge \((i,j)\).

Let the true parameter vector and the true distribution of interest be denoted by \( \theta^* \) and \( f_x(x; \theta^*) \) respectively. We assume a known upper (lower) bound on the maximum (minimum) absolute value of all non-zero parameter in \( \theta^* \), i.e., \( \| \theta^* \|_{\infty} \leq \theta_{\max}, \| \theta^* \|_{\min} \geq \theta_{\min}. \)

Suppose we are given additional structure. Define

\[
E(\theta^*) = \{ (i, j) : i < j \in [p], \| \theta^{(ij)} \|_2 > 0 \}.
\]

Consider the graph \( G(\theta^*) \) = \( (p, E(\theta^*)) \) such that \( f_x(x; \theta^*) \) is Markov with respect to \( G(\theta^*) \). Let the max-degree of any node of \( G(\theta^*) \) be at most \( d \). For any node \( i \in [p] \), let the neighborhood of node \( i \) be denoted as \( \mathcal{N}^i = \{ j : (i, j) \in E(\theta^*) \} \cup \{ j : (j, i) \in E(\theta^*) \} \).

The learning tasks of interest are as follows:

**Problem 2.1. (Structure Recovery).** Given \( n \) independent samples of \( x \) i.e., \( x^{(1)} \ldots x^{(n)} \) obtained from \( f_x(x; \theta^*) \), produce a graph \( \hat{G} \), such that \( \hat{G} = G(\hat{\theta}^*) \).

**Problem 2.2. (Parameter Recovery).** Given \( n \) independent samples of \( x \) i.e., \( x^{(1)} \ldots x^{(n)} \) obtained from \( f_x(x; \theta^*) \) and \( \alpha > 0 \), compute an estimate \( \hat{\theta} \) of \( \theta^* \) such that

\[
\| \theta^* - \hat{\theta} \|_\infty \leq \alpha.
\]

**Additional Notations.** For every node \( i \in [p] \), define \( \theta_{(i)} := (\theta^{(i)} \in \mathbb{R}^{k_i}; \theta^{(ij)} \in \mathbb{R}^{k_{ij}} : j \in [p], j \neq i) \in \mathbb{R}^{k_i+k_{ij}(p-1)} \) to be the weight vector associated with node \( i \) that consists of all the true parameters involving node \( i \). Define \( \Lambda = \{ \hat{\theta} \in \mathbb{R}^{k_i+k_{ij}(p-1)} : \| \hat{\theta} \|_{\min} \geq \theta_{\min}, \| \hat{\theta} \|_{\infty} \leq \theta_{\max} \} \). Then under our formulation, \( \theta^{(i)} \) is fixed for any \( i \in [p] \). Define \( \theta_{(i)} := (\theta^{(ij)} \in \mathbb{R}^{k_{ij}} : j \in [p], j \neq i) \in \mathbb{R}^{k_{ij}(p-1)} \) to be the component of \( \theta^{(i)} \) associated with the edge parameters.

**Definition 2.1. (Locally centered basis functions).** For \( i \in [p], j \in [p] \backslash \{i\} \), define locally centered basis functions as follows: for \( x \in \mathcal{X}_i, x' \in \mathcal{X}_j \),

\[
\phi^{(i)}(x) := \phi(x) - \int_{y \in \mathcal{X}_i} \phi(y) \mu_{\mathcal{X}_i}(y) dy,
\]

\( ^5 \) See Appendix 3 for a brief discussion on minimality of the exponential family.

\( ^6 \) \( \psi(x, y) \) can be written as \( \phi(x) \otimes \phi(y) \) using the sum formulae for sine and cosine.
On Learning Continuous Pairwise Markov Random Fields

\[
\psi^{(ij)}(x, x') := \psi(x, x') - \int_{y \in \mathcal{X}_i} \psi(y, x') d\mathcal{U}_Y(y) dy.
\]

where \( \mathcal{U}_Y(y) \) denotes the uniform density on \( \mathcal{X}_i \). For any \( i \in [p], j \in [p] \setminus \{i\} \), the locally centered basis functions \( \phi^{(ij)}(\cdot) \) and \( \psi^{(ij)}(\cdot, \cdot) \) integrate to zero with respect to \( \mathcal{U}_Y(y) \). This is motivated by the connection of the GISO to the penalized surrogate likelihood (See Appendix B).

Define \( \varphi^{(i)}(x) := (\phi^{(i)}(x_i) \in \mathbb{R}^k, \psi^{(ij)}(x_i, x_j) \in \mathbb{R}^{k^2}, j \in [p], j \neq i) \in \mathbb{R}^{k+k^2(p-1)} \) to be the vector of all locally centered basis functions involving node \( i \). We may also utilize notation \( \varphi^{(i)}(x) = \varphi^{(i)}(x_i, x_{-i}) \). Similarly, we define \( \varphi^{(i)}(x) \) when \( x = x_i \). Define

\[
\gamma := \theta_{\text{max}}(k + k^2d),
\]

\[
\varphi_{\text{max}} := 2 \max(\phi_{\text{max}}, \phi^2_{\text{max}}).
\]

Let \( q^* := q^*(k, b_i, \theta_{\text{max}}, \theta_{\text{min}} + \gamma, \varphi_{\text{max}}, \tilde{\varphi}_{\text{max}}) \) denote the smallest possible eigenvalue of the Fisher information matrix of any single-variable exponential family distribution with sufficient statistics \( \phi^{(i)}(x) \), with length of the support upper (lower) bounded by \( b_u(b_l) \) and with absolute value of all non-zero parameters bounded above (below) by \( \theta_{\text{max}}(\theta_{\text{min}}) \).

Let

\[
c_1(\alpha) = \frac{212\pi^2 e^2(d + 1)\gamma^2 \varphi_{\text{max}}^2 (1 + \gamma \varphi_{\text{max}})^2 \exp(4\gamma \varphi_{\text{max}})}{k^2 \alpha^4},
\]

\[
c_2(\alpha) = \frac{237d + 734d^2 k^{12}d + 16 \gamma d + 9 \gamma d + 8 \gamma d + 12 \gamma d^2}{\alpha^2 d^2 + 16(q^*)^2d^2 + 16 \gamma d + 12 \gamma d^2}.
\]

Observe that

\[
c_1(\alpha) = O\left(\frac{\exp(\Theta(k^2d))}{k^2 \alpha^4}\right),
\]

\[
c_2(\alpha) = O\left(\frac{\exp(\Theta(k^2d))}{\alpha^2 d^4 + 16(q^*)^2d^4 + 16 \gamma d + 12 \gamma d^2}\right).
\]

Let \( A(\varphi^{(i)}(x)) \) be the covariance matrix of \( \varphi^{(i)}(x) \) \( \exp(-\varphi^{(i)}(x)^T \varphi^{(i)}(x)) \) and \( B(\varphi^{(i)}(x)) \) be the cross-covariance matrix of \( \varphi^{(i)}(x) \) and \( \varphi^{(i)}(x) \) \( \exp(-\varphi^{(i)}(x)^T \varphi^{(i)}(x)) \), where \( x \) is distributed as per \( f_\lambda(x; \theta^*) \).

### 3 Algorithm

Our algorithm, ‘Augmented GISO’ has two parts: First, it recovers graph structure, i.e., edges \( E(\theta^*) \) and associated edge parameters, \( \theta^{(ij)}, i \neq j \in [p] \). This is achieved through the Generalized Regularized Interaction Screening Estimator (GRISE) of Vuffray et al. (2019) by extending the definition of GISO for continuous variables in a straightforward manner. This, however, does not recover node parameters \( \theta^{(i)}, i \in [p] \). Second, we transform the problem of learning node parameters as solving a sparse linear regression. Subsequently, using a robust variation of the classical Lasso (Tibshirani, 1996; Efron et al., 2004) and knowledge of the learned edge parameters, we recover node parameters.

#### Learning Edge Parameters

Given \( f_\lambda(x; \theta^*) \), the conditional density of \( x_i \) reduces to

\[
f_{\lambda}(x_i | x_{-i} = x_{-i}; \theta^{(i)}) \propto \exp\left(\varphi^{(i)}(x_i; x_{-i})\right).
\]

See Appendix B.1 for the derivation of (5). This form of conditional density inspired an unusual local or node specific objective GISO (Vuffray et al., 2019).

**Definition 3.1** (GISO). Given \( n \) samples \( x^{(1)} \ldots x^{(n)} \) of \( x \) and \( i \in [p] \), the GISO maps \( \vartheta \in \mathbb{R}^{k+k^2(p-1)} \) to \( S^{(i)}(\vartheta) \in \mathbb{R} \) defined as

\[
S^{(i)}(\vartheta) = \frac{1}{n} \sum_{t=1}^{n} \exp\left(-\vartheta^T \varphi^{(i)}(x^{(i)}_t)\right).
\]

Since the maximum node degree in \( G(\theta^*) \) is \( d \) and \( \|\theta^*\|_{\infty} \leq \theta_{\text{max}} \), we have \( \|\theta^{(i)}\|_1 \leq \gamma = \theta_{\text{max}}(k + k^2d) \) for any \( i \in [p] \). The GRISE produces an estimate of \( \theta^{(i)} \) for each \( i \in [p] \) by solving a separate optimization problem as

\[
\hat{\vartheta}^{(i)}(\lambda) \in \arg\min_{\vartheta \in \Lambda; \|\vartheta\|_{\infty} \leq \gamma} S^{(i)}(\vartheta).
\]

For \( \epsilon > 0 \), \( \hat{\vartheta}^{(i)}(\lambda) \) is an \( \epsilon \)-optimal solution of GRISE for \( i \in [p] \) if

\[
S^{(i)}(\hat{\vartheta}^{(i)}(\lambda)) \leq S^{(i)}(\hat{\vartheta}^{(i)}(\lambda)) + \epsilon.
\]

The (14) is a convex minimization problem and has an efficient implementation for finding an \( \epsilon \)-optimal solution. Appendix B.2 describes such an implementation for completeness borrowing from Vuffray et al. (2019).

Now, given such an \( \epsilon \)-optimal solution \( \hat{\vartheta}^{(i)}(\lambda) \) for GRISE corresponding to \( i \in [p] \), let \( \hat{\vartheta}^{(i)}(\lambda, j) = (\hat{\vartheta}^{(i)}(\lambda), j \neq i, j \in [p]) \) be its components corresponding to all possible \( p - 1 \) edges associated with node \( i \). Then, we declare \( \hat{\vartheta}^{(i)}(\lambda) \) as the edge parameters associated with \( i \) for each \( i \in [p] \). These edge parameters can be used to recover the graph structure as shown in Theorem 4.2.

#### Learning Node Parameters

As we shall argue in Theorems 4.1.1.2 for each \( i \in [p] \), the exact solution of GRISE is consistent, i.e., \( \hat{\vartheta}^{(i)}(\lambda, p) \rightarrow \theta^{(i)} \) in large sample limit — as well as it is normal, i.e., appropriately normalized \( \hat{\vartheta}^{(i)}(\lambda) - \theta^{(i)} \) obeys Central Limit Theorem in the large sample limit. While these are remarkable asymptotic results, they do not provide non-asymptotic
or finite sample error bounds. We will be able to provide finite sample error bounds for edge parameters learned from an $\epsilon$-optimal solution of GRISE, i.e., $\|\hat{\theta}^{(i)}_{E,\epsilon} - \theta^*_E\|_\infty$ is small. But to achieve the same for node parameters, we need additional processing.

This is the purpose of the method described next.

To that end, let us consider any $i \in [p]$. Given access to $\hat{\theta}^{(i)}_E$ (precisely, access to $\hat{\theta}^{(i)}_{E,\epsilon}$), we wish to identify node parameters $\theta^*(i) = (\theta^*(r) : r \in [k])$. Now the conditional density of $x_i \in \mathcal{X}$ when given $x_{-i} = x_{-i} \in \prod_{j \neq i} \mathcal{X}_j$, can be written as

$$f_{x_i}(x_i|x_{-i} = x_{-i}; \theta^*(i)) \propto \exp \left( \lambda^*(x_{-i}) \phi(x_i) \right),$$

where $\lambda^*(x_{-i}) := (\theta^*_r + \sum_{j \neq i} \sum_{s \in [k]} \theta^*_r \phi_s(x_j) : r \in [k])$ is the canonical parameter vector of the density in (1). See Appendix B for the derivation of (1). Let $\mu^*(x_{-i}) = \mathbb{E}[\phi(x_i)|x_{-i} = x_{-i}] \in \mathbb{R}^k$.

Now if we know $\lambda^*(x_{-i})$, and since we know $\hat{\theta}^{(i)}_{E,\epsilon}$, we can recover $(\theta^*_r : r \in [k])$. However, learning $\lambda^*(x_{-i})$ from samples is not straightforward. By duality of exponential family, in principle, if we know $\mu^*(x_{-i})$, we can recover $\lambda^*(x_{-i})$. Now learning $\mu^*(x_{-i})$ can be viewed as a traditional regression problem: features $Z = x_{-i}$, label $Y = \phi(x_i)$, regression function $\mathbb{E}[Y|Z] = \mu^*(x_{-i})$ and indeed samples $x^{(1)}, \ldots, x^{(n)}$ of $x$ provide samples of $Y, Z$ as defined here. Therefore, in principle, we can learn the regression function. As it turns out, the regression function $\mu^*(\cdot) : \mathbb{R}^{p-1} \to \mathbb{R}^k$ is Lipschitz and hence we can approximately linearize it leading to a sparse linear regression problem. Therefore, by utilizing Lasso on appropriately linearized problem, we can (approximately) learn $\mu^*(x_{-i})$, which in turn leads to $\lambda^*(x_{-i})$ and hence learning $(\theta^*_r : r \in [k])$ as desired. This is summarized as a three-step procedure:

Consider $x^{(z)}_{-i}$ where $z$ is chosen uniformly at random from $[n]$.

1. Express learning $\mu^*(\cdot)$ as a sparse linear regression problem (Details in Appendix B.2). Use robust variation of Lasso (Details in Appendix B) to obtain an estimate $(\hat{\mu}(x^{(z)}_{-i}))$ of $\mu^*(x^{(z)}_{-i})$ (Details in Appendix B.1).

2. Use $\hat{\mu}(x^{(z)}_{-i})$, and the conjugate duality between the canonical parameters and the mean parameters to learn an estimate $(\hat{\lambda}(x^{(z)}_{-i}))$ of $\lambda^*(x^{(z)}_{-i})$ (Details in Appendix B.2).

3. Use the estimates of the edge parameters i.e., $\hat{\theta}^{(i)}_{E,\epsilon}$ and $\hat{\lambda}(x^{(z)}_{-i})$ to learn an estimate $(\hat{\theta}^{(i)}(\cdot))$ of the node parameters $(\theta^*(i))$ (Summarized in Appendix E.2).

4 Analysis and Main results

4.1 Understanding GRISE: “Local” MLE, M-estimation, Consistency, Normality

For a given $i \in [p]$, we establish a surprising connection between the population version of GRISE and Maximum Likelihood Estimate (MLE) for a specific parametric distribution in an exponential family which varies across $i$. That is, for each $i \in [p]$, GRISE is a “local” MLE at the population level. Further, observing that minimizing the GISO is equivalent to M-estimation allows us to import asymptotic theory of M-estimation to establish consistency and normality of GRISE under mild conditions.

Consider $i \in [p]$. For any $\theta \in \Lambda$, the population version of GISO as defined in (1) is given by

$$S^{(i)}(\theta) := \mathbb{E} [ \exp \left( - \theta^T \varphi^{(i)}(x) \right) ].$$

Consider the distribution over $X$ with density given by

$$u^{(i)}_x = f_x(x; \theta^*) \propto \exp \left( - \theta^*^T \varphi^{(i)}(x) \right).$$

Define a parametric distribution over $X$ parameterized by $\theta \in \Lambda$ with density given by

$$u^{(i)}_x(x; \theta) \propto f_x(x; \theta^*) \propto \exp \left( - \theta^T \varphi^{(i)}(x) \right).$$

The following result argues that the MLE for parametric class induced by (11) coincides with the minimizer of the population version of GISO as defined in (11). This provides an intuitively pleasing connection of the GISO in terms of the KL-divergence. Proof can be found in Appendix C.

**Theorem 4.1.** Consider $i \in [p]$. Then, with $D(\cdot || \cdot)$ representing KL-divergence,

$$\arg \min_{\theta \in \Lambda} D(u^{(i)}_x(x; \theta) || u^{(i)}_x(x; \theta)) = \arg \min_{\theta \in \Lambda \| \theta \|_i \leq \gamma} S^{(i)}(\theta).$$

Further, the true parameter $\theta^{(i)}$ for $i \in [p]$ is a unique minimizer of $S^{(i)}(\theta)$.

Even though at the population level, GRISE is equivalent to MLE for parametric class induced by (11), the link between the finite-sample GRISE and the finite-sample MLE is missing. However, observe that minimizing the finite-sample GISO as defined...
in (1) is equivalent to M-estimation. This results in the following consistency and normality property of GRISE. Proof can be found in Appendix B.

**Theorem 4.2.** Given \( i \in [p] \) and \( n \) independent samples \( \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \) of \( \mathbf{x} \), let \( \hat{\theta}^{(i)} \) be a solution of (1). Then, as \( n \to \infty \), \( \hat{\theta}^{(i)} \overset{p}{\to} \theta^{(i)} \). Further, under the assumptions that \( B(\theta^{(i)}) \) is invertible, and that none of the true parameter is equal to the boundary values of \( \theta_{\max} \) or \( \theta_{\min} \), we have \( \vartheta_n(\hat{\theta}^{(i)} - \theta^{(i)}) \overset{d}{\to} N(0, B(\theta^{(i)})^{-1} A(\theta^{(i)}) B(\theta^{(i)})^{-1}) \) where \( N(\mu, \Sigma) \) represents multi-variate Gaussian with mean \( \mu \) and covariance \( \Sigma \).

See Appendix C for a brief discussion on invertibility of \( B(\theta^{(i)})^{-1} \). We emphasize that \( B(\theta^{(i)})^{-1} A(\theta^{(i)}) B(\theta^{(i)})^{-1} \) need not be equal to the inverse of the corresponding Fisher information matrix. See Appendix D for a counterexample. Thus, \( \hat{\theta}^{(i)} \) is asymptotically only normal and not efficient.

### 4.2 Finite Sample Guarantees

While Theorem 4.2 talks about asymptotic consistency and normality, it does not provide finite-sample error bounds. In this section, we provide the finite-sample error bounds which require the following additional condition.

**Condition 4.1.** Let \( \hat{\theta}, \hat{\theta} \in \mathbb{R}^{kp + \frac{k^2}{2}(p-1)} \) be feasible weight vectors associated with the distribution in (1) i.e., they have an upper (lower) bound on the maximum (minimum) absolute value of all non-zero parameters. There exists a constant \( \kappa > 0 \) such that for any \( i \neq j \in [p] \)

\[
\mathbb{E} \left[ \exp \left\{ -2h \left( (\hat{\theta}^{(ij)} - \hat{\theta}^{(ij)})^T \psi^{(ij)}(x_i, x_j) \right) \right\} \right] \\
\geq \kappa^2 \| \theta^{(ij)} - \hat{\theta}^{(ij)} \|^2_2.
\]

(12)

Here \( h(\cdot|x_{-j}) \) represents conditional differential entropy conditioned on \( x_{-j} \).

Under condition 4.1, we obtain the following structural recovery result whose proof is in Appendix E.

**Theorem 4.3.** Let Condition 4.1 be satisfied. Given \( n \) independent samples \( \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \) of \( \mathbf{x} \), for each \( i \in [p] \), let \( \hat{\theta}^{(i)} \) be an \( \epsilon \)-optimal solution of (1) and \( \hat{\theta}^{(i)}_{r,E} \) be the associated edge parameters. Let

\[
\hat{E} = \left\{ (i, j) : i < j \in [p], \sum_{r,s \in [k]} \mathbb{I}\{ | \hat{\theta}^{(ij)}_{r,s} | > \theta_{\min} / 3 \} > 0 \right\}.
\]

Let \( \hat{G} = ([p], \hat{E}) \). Then for any \( \delta \in (0, 1) \), \( G(\theta^*) = \hat{G} \) with probability at least \( 1 - \delta \) as long as

\[
n \geq n \cdot \left( \frac{\theta_{\min}^2}{3} \right) \log \left( \frac{2pk}{\delta} \right) = \Omega \left( \frac{\exp(\Theta(k^2 d))}{\kappa^2} \log \left( \frac{pk}{\sqrt{\delta}} \right) \right).
\]

The number of computations required scale as \( \tilde{O}(p^2) \).

Now we state our result about parameter recovery whose proof can be found in Appendix F.

**Theorem 4.4.** Let Condition 4.1 be satisfied. Given \( n \) independent samples \( \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \) of \( \mathbf{x} \), for each \( i \in [p] \), let \( \hat{\theta}^{(i)} \) be an \( \epsilon \)-optimal solution of (1) and \( \hat{\theta}^{(i)}_{E} \) be the associated edge parameters. Let \( \hat{\theta} \in \mathbb{R}^k, i \in [p] \) be estimates of node parameters obtained through the three-step procedure involving robust Lasso. Let \( \hat{\theta} = (\hat{\theta}^{(i)}; \hat{\theta}^{(i)}_{E} : i \in [p]) \in \mathbb{R}^{kp + \frac{k^2}{2}(p-1)} \) be their appropriate concatenation. Then, for any \( \alpha \in (0, 1) \)

\[
\| \hat{\theta} - \theta^* \|_\infty \leq \alpha,
\]

with probability at least \( 1 - \alpha^4 \) as long as

\[
n \geq \max \left\{ c_1 \left( \min \left\{ \frac{\theta_{\min}^4}{3}, \alpha, \frac{\alpha}{2 + dk_{\max}} \right\} \right) \log \left( \frac{pk}{\alpha^2} \right), c_2 \left( \frac{\alpha}{2 + \frac{dp_{\max}}{k^2}} \right), \right\}
\]

\[
\geq \Omega \left( \frac{\exp(\Theta(k^2 d + 4 \log \frac{d}{k^2 \alpha^4})))}{\kappa^2 \alpha^4} \times \log \left( \frac{pk}{\alpha^2} \right) \right).
\]

The number of computations required scale as \( \tilde{O}(p^2) \).

### 4.3 Connections to surrogate likelihood.

To circumvent the computational limitation of exact likelihood-based functionals in nonparametric density estimation, Jean and Lin (2006) proposed to minimize the surrogate likelihood. Let \( \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \) be \( n \) independent samples of \( \mathbf{x} \) where \( \mathbf{x} \in \mathcal{X} \). For densities of the form \( f_{\mathbf{x}}(\mathbf{x}) \propto e^{\eta(\mathbf{x})} \), the surrogate likelihood is as follows:

\[
L_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} \exp \left( - \eta(\mathbf{x}^{(i)}) \right) + \int_{\mathcal{X}} \rho(\mathbf{x}) \times \eta(\mathbf{x}) d\mathbf{x},
\]

where \( \rho(\cdot) \) is some known probability density function on \( \mathcal{X} \). The following proposition shows that the GISO is a special case of the surrogate likelihood. Proof can be found in Appendix G.

**Proposition 4.1.** For any \( i \in [p] \), the GISO is equivalent to the surrogate likelihood associated with the conditional density of \( x_i \) when \( \rho(\cdot) \) is the uniform density on \( \mathcal{X}_i \).
4.4 Examples

The following are a few examples where the Condition 4.1 is naturally satisfied (subject to problem setup) as explained in Appendix T. Therefore, these distributions are learnable consistently, have asymptotic Gaussian-like behavior (under the assumptions in Theorem 4.2), and have finite sample guarantees. This is in contrast to most prior works for the continuous setting where there are difficult to verify conditions (even for these examples) such as incoherence, dependency, sparse eigenvalue, and restricted strong convexity.

A. Polynomial (linear) sufficient statistics i.e., $\phi(x) = x$ and $k = 1$.

$$f_x(x; \theta^*) \propto \exp \left( \sum_{i \in [p]} \theta_s^{(i)} x_i + \sum_{i \in [p], j > i} \sum_{\theta^{(ij)} x_i x_j} \right).$$

B. Harmonic sufficient statistics i.e., $\phi(x) = (\sin(\pi x/b), \cos(\pi x/b))$ and $k = 2$.

$$f_x(x; \theta^*) \propto \exp \left( \sum_{i \in [p]} \left[ \theta_1^{(i)} \sin \frac{\pi x_i}{b} + \theta_2^{(i)} \cos \frac{\pi x_i}{b} \right] + \sum_{i \in [p], j > i} \left[ \theta_1^{(ij)} \sin \frac{\pi (x_i + x_j)}{b} + \theta_2^{(ij)} \cos \frac{\pi (x_i + x_j)}{b} \right] \right).$$

5 Conclusion

We provide rigorous finite sample analysis for learning structure and parameters of continuous MRFs without the abstract conditions of incoherence, dependency, sparse eigenvalue or restricted strong convexity that are common in literature. We provide easy-to-verify sufficient condition for learning that is naturally satisfied for polynomial and harmonic sufficient statistics. Our methodology requires $O(p^2)$ computations and $O(\exp(d) \log p)$ samples similar to the discrete and Gaussian settings. Additionally, we propose a robust variation of Lasso by showing that even in the presence of bounded additive noise, the Lasso estimator is ‘prediction consistent’ under mild assumptions.

We also establish that minimizing the population version of GISO (Vuffray et al., 2019) is equivalent to finding MLE of a certain related parametric distribution. We provide asymptotic consistency and normality of the estimator under mild conditions. Further, we show that the GISO is equivalent to the surrogate likelihood proposed by Jeon and Lin (2006).

A natural extension of the pairwise setup is the $t$-wise MRF with continuous variables. The approach and the objective function introduced by Vuffray et al. (2019) naturally extend for such a setting allowing to learn $t$-wise MRFs with general discrete variables as explained in that work. We believe that our results for continuous setting, in a similar vein, extend for $t$-wise MRFs as well and it is an important direction for immediate future work. We also believe that the connection of the GISO to KL-divergence could be used to remove the bounded random variables assumption of our work. Another important direction is to leverage the asymptotic normality of the estimator established in our work to construct data-driven explicit confidence intervals for learned parameters of MRF.

Acknowledgements

This work was supported, in part, by NSF under Grant Nos. CCF-171761, CMMI-1462158, and CNS-1523546, the MIT-IBM Watson Lab under Agreement No. W1771646, and the MIT-KACST project.

We would like to thank Andrey Y. Lokhov, Marc Vuffray, and Sidhant Misra for pointing to us the possibility of using GRISE for finite sample analysis of learning continuous graphical models during the MIFODS Workshop on Graphical models, Exchangeable models and Graphons organized at MIT in summer of 2019. We would also like to thank the anonymous referees of NeurIPS 2020 for pointing out a bug in the earlier version of Theorem 4.2.
On Learning Continuous Pairwise Markov Random Fields

References


Abhin Shah, Devavrat Shah, Gregory W. Wornell


