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 $\tilde{\mathcal{O}}(p^2)$.[1] This improved upon the prior work of Bresler et al. (2013), with runtime $\bar{\mathcal{O}}(p^{d+2})$,[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 ($O(\exp(\Theta(d^{t-1}))\log p)$) with runtime $\bar{\mathcal{O}}(p^t)$. 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.[3] 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

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] The $\tilde{\mathcal{O}}(\cdot)$ notation hides a factor $\text{poly}(\log p)$ as well as a constant (doubly-exponentially) depending on d.

[2] The $\bar{\mathcal{O}}(\cdot)$ notation hides a factor $\text{poly}(\log p)$ as well as a constant (exponentially) depending on d.

[3] For the discrete setup, learning edge parameters is sufficient since, knowing those, node parameters can be recovered using the conditional expectation function (which is a logistic function and can be inverted). However, the same is not straightforward in the continuous setup. We provide a rigorous way to tackle this (Section 3).

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

Work	Approach	Conditions	#samples
Yang et al. (2015)	ℓ_1 regularized node conditional log-likelihood	1. Incoherence condition 2. Dependency condition 3. Bounded moments of the variables 4. Local smoothness of the log-partition function 5. Conditional distribution lies in exponential family	$O(\text{poly}(d)\omega(p))$ s.t $\omega(p) = \bar{\omega}(p)\log p$ and $\bar{\omega}(p)$ is a density dependent function of p
Tansey et al. (2015)	Group lasso regularized node conditional log-likelihood		
Yang et al. (2018)	Node conditional pseudo-likelihood regularized by a nonconvex penalty	1. Sparse eigenvalue condition 2. Bounded moments of the variables 3. Local smoothness of the log-partition function 4. Conditional distribution lies in exponential family	$O(\text{poly}(d)\log p)$
Sun et al. (2015)	Penalized score matching objective	1. Incoherence condition 2. Dependency condition 3. Certain structural conditions	$O(\text{poly}(pd))$
Suggala et al. (2017)	ℓ_1 regularized node conditional log-likelihood	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	$O(\text{poly}(d)\log p)$
Yuan et al. (2016)	$\ell_{2,1}$ regularized node conditional log-likelihood	1. Restricted strong convexity 2. Bounded moment-generating function of variables	$O(\text{poly}(d)\log p)$
This work	Augmented GISO (Section 3)	1. Bounded domain of the variables 2. Conditional distribution lies in exponential family	$O(\exp(d)\log p)$ (Thm. 4.3-4.4)

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.

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

where $\mathbf{x} = (x_1, \cdots, x_p)$ is a p-dimensional vector of continuous variables, $\mathbf{x} = (x_1, \cdots, x_p)$ is a realization of $\mathbf{x}$, 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., 2015), dependency (Yang et al., 2015; Tansey et al., 2015; 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 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.

Result (pairwise)	Alphabet	Consistency (i.e. SLLN)	Normality (i.e. CLT)	#computations	#samples
Bresler et al. (2013)	Discrete	✓	×	$\bar{\mathcal{O}}(p^{d+2})$	$O(\exp(d)\log p)$
Bresler (2015)	Binary	✓	×	$\bar{\mathcal{O}}(p^2)$	$O(\exp(\exp(d))\log p)$
Klivans and Meka (2017)	Discrete	✓	×	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$
Vuffray et al. (2019)	Discrete	✓	×	$\bar{\mathcal{O}}(p^2)$	$O(\exp(d)\log p)$
This Work	Continuous	✓ (Thm. 4.2)	✓ (Thm. 4.2)	$\bar{\mathcal{O}}(p^2)$ (Thm. 4.3-4.4)	$O(\exp(d)\log p)$ (Thm. 4.3-4.4)

conditions common in literature (incoherence, dependency, sparse eigenvalue or restricted strong convexity). We require $\bar{\mathcal{O}}(p^2)$ computations and $O(\exp(d)\log p)$ samples, in-line with the prior works on discrete / Gaussian MRFs. We formally extend the approach of Vuffray et al. (2019) to the continuous setting to recover the 'edge' parameters and propose a novel algorithm for learning 'node' parameters through a robust variation of sparse linear regression (Lasso). Technically, this robust Lasso shows that even in the presence of arbitrary bounded additive noise, the Lasso estimator is 'prediction consistent' under mild assumptions (see Appendix N). Further, we simplify the sufficient conditions for learning of Vuffray et al. (2019) from node-neighborhood-based to edge-based (see Condition 4.1[4]). This is achieved through a novel argument that utilizes the structure of the weighted independent set induced by the MRF (see within Appendix L.3). We show that the new, easy-to-verify, sufficient condition is naturally satisfied by various settings including polynomial and harmonic sufficient statistics (see Section 4.4 for concrete examples). Thus, while most of the existing works focus on distributions of the form (1), our method is applicable to a large class of distributions beyond that.

Understanding GISO. We establish that minimizing the population version of GISO of Vuffray et al. (2019) is identical to minimizing an appropriate Kullback-Leibler (KL) divergence. This is true for MRFs with discrete as well as continuous-valued random variables. Using the equivalence of KL divergence and maximum likelihood, we can interpret minimizing the population version of GISO as "local" MLE. By observing that minimizing the GISO is equivalent to M-estimation, we obtain asymptotic consistency and normality for this method with mild conditions. Finally, we also draw connections between the GISO and the surrogate likelihood proposed by Jeon and Lin (2006) for log-density ANOVA model estimation (see Section 4.3 and Appendix H).

[4]Condition 4.1 effectively lower bounds the variance of a non-constant random variable and is an adequate condition to rule out certain singular distributions.

1.3 Other related work

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

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 $\bar{\mathcal{O}}(p^4)$. 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 $\bar{\mathcal{O}}(p^2)$ with optimal sample complexity ($O(\exp(\Theta(d))\log p)$) and t-wise MRFs over binary alphabets in time $\bar{\mathcal{O}}(p^t)$ 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 $\tilde{\kappa}$ (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 $O((d^2+\tilde{\kappa}^{-2})\log p)$ samples (Ravikumar et al., 2011) by ℓ_1 regularized log-likelihood estimator. The minimum required sample

complexity was shown to be $\Omega(\log p/\tilde{\kappa}^2)$ 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 $O(d \log p/\tilde{\kappa}^2)$ 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 $O(d \log p/\tilde{\kappa}^2)$ samples. However, when the variables are positively associated, this algorithm achieves the optimal sample complexity of $O(\log p/\tilde{\kappa}^2)$.

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. Tansey et al. (2015) 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 $[n] \coloneqq \{1, \cdots, n\}$. For a deterministic sequence $v_1, \cdots, v_n$, we let $\mathbf{v} \coloneqq (v_1, \cdots, v_n)$. For a random sequence $\mathsf{v}_1, \cdots, \mathsf{v}_n$, we let $\mathbf{v} \coloneqq (\mathsf{v}_1, \cdots, \mathsf{v}_n)$. Let $\mathbb{1}$ 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 ℓ_p norm ($p \geq 1$) of a vector $\mathbf{v} \in \mathbb{R}^n$ by $\|\mathbf{v}\|_p \coloneqq (\sum_{i=1}^n |v_i|^p)^{1/p}$ and its ℓ_∞ norm by $\|\mathbf{v}\|_\infty \coloneqq \max_i |v_i|$. For a vector $\mathbf{v} \in \mathbb{R}^n$, we use $\|\mathbf{v}\|_0$ to denote the number of non-zero elements (ℓ_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 $\|\mathbf{v}\|_{\min_+} \coloneqq \min_{i:v_i \neq 0} |v_i|$. 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 $\|\mathbf{V}\|_{\max} \coloneqq \max_{ij} |V_{ij}|$. 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 gurantees 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 A provides an illustrative example to ease through these notations.

Pairwise MRF. Let $\mathbf{x} = (\mathsf{x}_1, \cdots, \mathsf{x}_p)$ be a $p-$dimensional vector of continuous random variables such that each x_i takes value in a real interval $\mathcal{X}_i$ and let $\mathcal{X} = \prod_{i=1}^p \mathcal{X}_i$. 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 $\mathcal{X}_i$ be upper (lower) bounded by a known constant b_u (b_l). 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 $g_{ij} : \mathcal{X}_i \times \mathcal{X}_j \to \mathbb{R}$,

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

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

$$g_i(\cdot) = \boldsymbol{\theta}^{(i)^T}\boldsymbol{\phi}(\cdot) \qquad \text{and} \qquad g_{ij}(\cdot,\cdot) = \boldsymbol{\theta}^{(ij)^T}\boldsymbol{\psi}(\cdot,\cdot),$$

where $\boldsymbol{\theta}^{(i)} \in \mathbb{R}^k$ is the vector of parameters associated with the node i, $\boldsymbol{\theta}^{(ij)} \in \mathbb{R}^{k^2}$ is the vector of parameters associated with the edge (i,j), the map $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^k$ is a basis of the vector space of node potentials, and the map $\boldsymbol{\psi} : \mathbb{R}^2 \to \mathbb{R}^{k^2}$ is a basis of the vector space of edge potentials. We assume that $\boldsymbol{\psi}(x,y)$ can be written as the Kronecker product of $\boldsymbol{\phi}(x)$ and $\boldsymbol{\phi}(y)$ i.e., $\boldsymbol{\psi}(x,y) = \boldsymbol{\phi}(x) \otimes \boldsymbol{\phi}(y)$. This is equivalent to the function space assumption common in the literature (Yang et al., 2015, 2018; Suggala et al., 2017; Tansey et al., 2015) 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 $\boldsymbol{\phi}(x) = (x^r : r \in [k])$, $\boldsymbol{\psi}(x,y) = (x^r y^s : r,s \in [k])$; (2) Harmonic basis with $\boldsymbol{\phi}(x) = (\sin(rx); \cos(rx) : r \in [k])$, $\boldsymbol{\psi}(x,y) = (\sin(rx+sy); \cos(rx+sy) : r,s \in [k])$.[6]

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

[5] See Appendix U.3 for a brief discussion on minimality of the exponential family.

[6] $\boldsymbol{\psi}(x,y)$ can be written as $\boldsymbol{\phi}(x) \otimes \boldsymbol{\phi}(y)$ using the sum formulae for sine and cosine.

Summarizing, the distribution of focus is

$$f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}) \propto \exp\Big(\sum_{i\in[p]} \boldsymbol{\theta}^{(i)^T}\boldsymbol{\phi}(x_i) + \sum_{i\in[p],j>i} \boldsymbol{\theta}^{(ij)^T}\boldsymbol{\psi}(x_i,x_j)\Big), \tag{2}$$

where $\boldsymbol{\theta} := \big(\boldsymbol{\theta}^{(i)} \in \mathbb{R}^k : i \in [p]; \boldsymbol{\theta}^{(ij)} \in \mathbb{R}^{k^2} : i \in [p], j > i\big) \in \mathbb{R}^{kp + \frac{k^2 p(p-1)}{2}}$ is the parameter vector associated with the distribution. For any $i \in [p], i > j$, define $\boldsymbol{\theta}^{(ij)} = \boldsymbol{\theta}^{(ji)}$ i.e., both $\boldsymbol{\theta}^{(ij)}$ and $\boldsymbol{\theta}^{(ji)}$ denote the parameter vector associated with the edge (i,j).

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

Suppose we are given additional structure. Define

$$E(\boldsymbol{\theta}^*) = \{(i,j) : i < j \in [p], \|\boldsymbol{\theta}^{*(ij)}\|_0 > 0\}.$$

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

The learning tasks of interest are as follows:

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

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

$$\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_\infty \le \alpha.$$

Additional Notations. For every node $i \in [p]$, define $\boldsymbol{\vartheta}^{*(i)} := \big(\boldsymbol{\theta}^{*(i)} \in \mathbb{R}^k; \boldsymbol{\theta}^{*(ij)} \in \mathbb{R}^{k^2} : j \in [p], j \ne i\big) \in \mathbb{R}^{k+k^2(p-1)}$ to be the weight vector associated with node i that consists of all the true parameters involving node i. Define $\Lambda = \{\boldsymbol{\vartheta} \in \mathbb{R}^{k+k^2(p-1)} : \|\boldsymbol{\vartheta}\|_{\min_+} \ge \theta_{\min_+}, \|\boldsymbol{\vartheta}\|_\infty \le \theta_{\max}\}$. Then under our formulation, $\boldsymbol{\vartheta}^{*(i)} \in \Lambda$ for any $i \in [p]$. Define $\boldsymbol{\vartheta}_E^{*(i)} := \big(\boldsymbol{\theta}^{*(ij)} \in \mathbb{R}^{k^2} : j \in [p], j \ne i\big) \in \mathbb{R}^{k^2(p-1)}$ to be the component of $\boldsymbol{\vartheta}^{*(i)}$ 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 $x \in \mathcal{X}_i$, $x' \in \mathcal{X}_j$*

$$\boldsymbol{\phi}^{(i)}(x) := \boldsymbol{\phi}(x) - \int_{y\in\mathcal{X}_i} \boldsymbol{\phi}(y)\mathcal{U}_{\mathcal{X}_i}(y)dy, \tag{3}$$

$$\boldsymbol{\psi}^{(ij)}(x,x') \coloneqq \boldsymbol{\psi}(x,x') - \int_{y\in\mathcal{X}_i} \boldsymbol{\psi}(y,x')\mathcal{U}_{\mathcal{X}_i}(y)dy. \quad (4)$$

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

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

$$\gamma \coloneqq \theta_{\max}(k + k^2 d),$$
$$\varphi_{\max} \coloneqq 2\max\{\phi_{\max}, \phi_{\max}^2\}.$$

Let $q^s \coloneqq q^s(k, b_l, b_u, \theta_{\max}, \theta_{\min_+}, \phi_{\max}, \bar{\phi}_{\max})$ denote the smallest possible eigenvalue of the Fisher information matrix of any single-variable exponential family distribution with sufficient statistics $\boldsymbol{\phi}(\cdot)$, 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 $\theta_{\max}(\theta_{\min_+})$. Let

$$c_1(\alpha) = \frac{2^{12}\pi^2 e^2 (d+1)^2 \gamma^2 \varphi_{\max}^2 (1+\gamma\varphi_{\max})^2 \exp(4\gamma\varphi_{\max})}{\kappa^2\alpha^4},$$

$$c_2(\alpha) = \frac{2^{37d+73} b_u^{2d} k^{12d+16} d^{6d+9} \theta_{\max}^{6d+8} \phi_{\max}^{8d+12} \bar{\phi}_{\max}^{2d}}{\alpha^{8d+16}(q^s)^{4d+8}}.$$

Observe that

$$c_1(\alpha) = O\left(\frac{\exp(\Theta(k^2 d))}{\kappa^2\alpha^4}\right), c_2(\alpha) = O\left(\left(\frac{kd}{\alpha q^s}\right)^{\Theta(d)}\right).$$

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

3 Algorithm

Our algorithm, 'Augmented GISO' has two parts: First, it recovers graph structure, i.e. edges $E(\boldsymbol{\theta}^*)$ and associated edge parameters, $\boldsymbol{\theta}^{*(ij)}, i \neq j \in [p]$. 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 $\boldsymbol{\theta}^{*(i)}, i \in [p]$. 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 $f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}^*)$, for any $i \in [p]$, the conditional density of x_i reduces to

$$f_{x_i}(x_i|x_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) \propto \exp\Big(\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})\Big) \quad (5)$$

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

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

$$\mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}) = \frac{1}{n}\sum_{t=1}^{n}\exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x}^{(t)})\Big). \quad (6)$$

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

$$\hat{\boldsymbol{\vartheta}}_n^{(i)} \in \underset{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma}{\arg\min}\ \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}). \quad (7)$$

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

$$\mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}) \leq \mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)}) + \epsilon. \quad (8)$$

The (7) is a convex minimization problem and has an efficient implementation for finding an ϵ-optimal solution. Appendix M describes such an implementation for completeness borrowing from Vuffray et al. (2019).

Now, given such an ϵ-optimal solution $\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}$ for GRISE corresponding to $i \in [p]$, let $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)} = (\hat{\boldsymbol{\theta}}^{(ij)}, j \neq i, j \in [p])$ be its components corresponding to all possible $p-1$ edges associated with node i. Then, we declare $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}$ 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 4.3.

Learning Node Parameters. As we shall argue in Theorems 4.1-4.2, for each $i \in [p]$, the exact solution of GRISE is consistent, i.e. $\hat{\boldsymbol{\vartheta}}_n^{(i)} \xrightarrow{p} \boldsymbol{\vartheta}^{*(i)}$ in large sample limit — as well as it is normal, i.e. appropriately normalized $\hat{\boldsymbol{\vartheta}}_n^{(i)} - \boldsymbol{\vartheta}^{*(i)}$ 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 ϵ-optimal solution of GRISE, i.e. $\|\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,E} - \boldsymbol{\vartheta}^{*(i)}_E\|_\infty$ is small. But to achieve the same for node parameters, we need additional processing [7]. This is the purpose of the method described next.

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

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) \propto \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i})\boldsymbol{\phi}(x_i)\Big), \quad (9)$$

where $\boldsymbol{\lambda}^*(x_{-i}) \coloneqq (\theta^{*(i)}_r + \sum_{j\neq i}\sum_{s\in[k]} \theta^{*(ij)}_{r,s}\phi_s(x_j) : r \in [k])$ is the canonical parameter vector of the density in (9). See Appendix B.1 for the derivation of (9). Let $\boldsymbol{\mu}^*(x_{-i}) = \mathbb{E}[\boldsymbol{\phi}(\mathsf{x}_i)|\mathsf{x}_{-i} = x_{-i}] \in \mathbb{R}^k$.

Now if we know $\boldsymbol{\lambda}^*(x_{-i})$, and since we know $\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,E} \approx \boldsymbol{\vartheta}^{*(i)}_E$, we can recover $(\theta^{*(i)}_r : r \in [k])$. However, learning $\boldsymbol{\lambda}^*(x_{-i})$ from samples is not straightforward. By duality of exponential family, in principle, if we know $\boldsymbol{\mu}^*(x_{-i})$, we can recover $\boldsymbol{\lambda}^*(x_{-i})$. Now learning $\boldsymbol{\mu}^*(x_{-i})$ can be viewed as a traditional regression problem: features $Z = \mathsf{x}_{-i}$, label $Y = \boldsymbol{\phi}(\mathsf{x}_i)$, regression function $\mathbb{E}[Y|Z] = \boldsymbol{\mu}^*(\mathsf{x}_{-i})$ and indeed samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ of $\mathbf{x}$ provides samples of Y, Z as defined here. Therefore, in principle, we can learn the regression function. As it turns out, the regression function $\boldsymbol{\mu}^*(\cdot) : \mathbb{R}^{p-1} \to \mathbb{R}^k$ 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 $\boldsymbol{\mu}^*(x_{-i})$, which in turn leads to $\boldsymbol{\lambda}^*(x_{-i})$ and hence learning $(\theta^{*(i)}_r : r \in [k])$ 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 $\boldsymbol{\mu}^*(\cdot)$ as a sparse linear regression problem (Details in Appendix Q.2). Use robust variation of Lasso (Details in Appendix N) to obtain an estimate $(\hat{\boldsymbol{\mu}}(x^{(z)}_{-i}))$ of $\boldsymbol{\mu}^*(x^{(z)}_{-i})$ (Details in Appendix O.1).
2. Use $\hat{\boldsymbol{\mu}}(x^{(z)}_{-i})$, and the conjugate duality between the canonical parameters and the mean parameters to learn an estimate $(\hat{\boldsymbol{\lambda}}(x^{(z)}_{-i}))$ of $\boldsymbol{\lambda}^*(x^{(z)}_{-i})$ (Details in Appendix O.2).
3. Use the estimates of the edge parameters i.e., $\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,E}$ and $\hat{\boldsymbol{\lambda}}(x^{(z)}_{-i})$ to learn an estimate $(\hat{\boldsymbol{\theta}}^{(i)})$ of the node parameters $(\boldsymbol{\theta}^{*(i)})$ (Summarized in Appendix E.2).

[7] This happens because only the edge parameters show up in a restricted strong convexity like property obeyed by the GISO (see Proposition I.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 minimzing 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 $\boldsymbol{\vartheta} \in \Lambda$, the population version of GISO as defined in (6) is given by

$$\mathcal{S}^{(i)}(\boldsymbol{\vartheta}) \coloneqq \mathbb{E}\bigg[\exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)\bigg]. \quad (10)$$

Consider the distribution over $\mathcal{X}$ with density given by

$$u^{(i)}_{\mathbf{x}}(\mathbf{x}) \propto f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*) \times \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big).$$

Define a parametric distribution over $\mathcal{X}$ parameterized by $\boldsymbol{\vartheta} \in \Lambda$ with density given by

$$m^{(i)}_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\vartheta}) \propto f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*) \times \exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big). \quad (11)$$

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 (10). 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,*

$$\operatorname*{arg\,min}_{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma} D(u^{(i)}_{\mathbf{x}}(\cdot) \,\|\, m^{(i)}_{\mathbf{x}}(\cdot; \boldsymbol{\vartheta})) = \operatorname*{arg\,min}_{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma} \mathcal{S}^{(i)}(\boldsymbol{\vartheta}).$$

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

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 (6) is equivalent to M-estimation. This results in the following consistency and normality property of GRISE. Proof can be found in Appendix D.

Theorem 4.2. *Given $i \in [p]$ and n independent samples $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ of $\mathbf{x}$, let $\hat{\boldsymbol{\vartheta}}_n^{(i)}$ be a solution of (7). Then, as $n \to \infty$, $\hat{\boldsymbol{\vartheta}}_n^{(i)} \xrightarrow{p} \boldsymbol{\vartheta}^{*(i)}$. Further, under the assumptions that $B(\boldsymbol{\vartheta}^{*(i)})$ is invertible, and that none of the true parameter is equal to the boundary values of $\theta_{\max}$ or $\theta_{\min_+}$, we have $\sqrt{n}(\hat{\boldsymbol{\vartheta}}_n^{(i)} - \boldsymbol{\vartheta}^{*(i)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, B(\boldsymbol{\vartheta}^{*(i)})^{-1} A(\boldsymbol{\vartheta}^{*(i)}) B(\boldsymbol{\vartheta}^{*(i)})^{-1})$ where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents multi-variate Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.*

See Appendix U.1 for a brief discussion on invertibility of $B(\boldsymbol{\vartheta}^{*(i)})$. We emphasize that $B(\boldsymbol{\vartheta}^{*(i)})^{-1} A(\boldsymbol{\vartheta}^{*(i)}) B(\boldsymbol{\vartheta}^{*(i)})^{-1}$ need not be equal to the inverse of the corresponding Fisher information matrix. See Appendix U.2 for a counterexample. Thus, $\hat{\boldsymbol{\vartheta}}_n^{(i)}$ 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 $\bar{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}} \in \mathbb{R}^{kp + \frac{k^2 p(p-1)}{2}}$ be feasible weight vectors associated with the distribution in (2) 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 $i \neq j \in [p]$*

$$\mathbb{E}\left[\exp\left\{ 2h\left((\bar{\boldsymbol{\theta}}^{(ij)} - \tilde{\boldsymbol{\theta}}^{(ij)})^T \boldsymbol{\psi}^{(ij)}(x_i, x_j) \middle| x_{-j} \right) \right\} \right] \geq \kappa \|\bar{\boldsymbol{\theta}}^{(ij)} - \tilde{\boldsymbol{\theta}}^{(ij)}\|_2^2. \tag{12}$$

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 F.

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 $\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}$ be an ϵ-optimal solution of (7) and $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}$ be the associated edge parameters. Let*

$$\hat{E} = \left\{ (i,j) : i < j \in [p], \left(\sum_{r,s \in [k]} \mathbb{1}\{ |\hat{\theta}_{r,s}^{(ij)}| > \theta_{\min_+}/3 \} \right) > 0 \right\}.$$

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

$$n \geq c_1 \left(\frac{\theta_{\min_+}}{3} \right) \log \left(\frac{2pk}{\sqrt{\delta}} \right) = \Omega\left(\frac{\exp(\Theta(k^2 d))}{\kappa^2} \log\left(\frac{pk}{\sqrt{\delta}} \right) \right).$$

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

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

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 $\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}$ be an ϵ-optimal solution of (7) and $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)} \in \mathbb{R}^{k^2(p-1)}$ be the associated edge parameters. Let $\hat{\boldsymbol{\theta}}^{(i)} \in \mathbb{R}^k, i \in [p]$ be estimates of node parameters obtained through the three-step procedure involving robust Lasso. Let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}^{(i)}; \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)} : i \in [p]) \in \mathbb{R}^{kp + \frac{k^2 p(p-1)}{2}}$ be their appropriate concatenation. Then, for any $\alpha \in (0,1)$*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \alpha,$$

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

$$n \geq \max\left[c_1\left(\min\left\{ \frac{\theta_{\min_+}}{3}, \alpha, \frac{\alpha}{2^{\frac{5}{4}} dk \phi_{\max}} \right\} \right) \log\left(\frac{8pk}{\alpha^2} \right), c_2\left(\frac{\alpha}{2^{\frac{1}{4}}} \right) \right],$$

$$= \Omega\left(\frac{\exp\left(\Theta\left(k^2 d + d \log\left(\frac{dk}{\alpha q^s} \right) \right) \right)}{\kappa^2 \alpha^4} \times \log\left(\frac{pk}{\alpha^2} \right) \right).$$

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

4.3 Connections to surrogate likelihood.

To circumvent the computational limitation of exact likelihood-based functionals in nonparametric density estimation, Jeon 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:

$$\mathcal{L}_n(\eta) = \frac{1}{n} \sum_{t=1}^{n} \exp\Big(-\eta(\mathbf{x}^{(t)}) \Big) + \int_{\mathbf{x}} \rho(\mathbf{x}) \times \eta(\mathbf{x}) d\mathbf{x},$$

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 H.

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., $\boldsymbol{\phi}(x) = x$ and $k = 1$.

$$f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}^*) \propto \exp\Bigg(\sum_{i\in[p]} \theta^{*(i)} x_i + \sum_{i\in[p]}\sum_{j>i} \theta^{(ij)} x_i x_j\Bigg).$$

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

$$f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}^*) \propto \exp\Bigg(\sum_{i\in[p]} \Big[\theta_1^{*(i)} \sin\frac{\pi x_i}{b} + \theta_2^{*(i)} \cos\frac{\pi x_i}{b}\Big] +$$

$$\sum_{i\in[p],j>i} \Big[\theta_1^{*(ij)} \sin\frac{\pi(x_i+x_j)}{b} + \theta_2^{*(ij)} \cos\frac{\pi(x_i+x_j)}{b}\Big]\Bigg).$$

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 $\bar{\mathcal{O}}(p^2)$ 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.

References

Abbeel, P., Koller, D., and Ng, A. Y. (2006). Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.*, 7:1743–1788.

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Beck, A. and Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175.

Boyd, S. P. and Vandenberghe, L. (2014). *Convex Optimization*. Cambridge University Press.

Bresler, G. (2015). Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 771–782.

Bresler, G., Gamarnik, D., and Shah, D. (2014). Hardness of parameter estimation in graphical models. In *Advances in Neural Information Processing Systems*, pages 1062–1070.

Bresler, G., Mossel, E., and Sly, A. (2013). Reconstruction of markov random fields from samples: Some observations and algorithms. *SIAM J. Comput.*, 42(2):563–578.

Bubeck, S. (2015). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357.

Chatterjee, S. (2013). Assumptionless consistency of the lasso.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467.

Dobra, A., Lenkoski, A., et al. (2011). Copula gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *Annals of statistics*, 32(2):407–499.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Hamilton, L., Koehler, F., and Moitra, A. (2017). Information theoretic properties of markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems*, pages 2463–2472.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. (2011). On learning discrete graphical models using group-sparse regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 378–387.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.*, 40(2):633–643.

Jeon, Y. and Lin, Y. (2006). An effective method for high-dimensional log-density anova estimation, with application to nonparametric graphical model building. *Statistica Sinica*, pages 353–374.

Jerrum, M. and Sinclair, A. (1988). Conductance and the rapid mixing property for markov chains: the approximation of permanent resolved. In *Proceedings of the twentieth annual ACM symposium on Theory of computing*, pages 235–244. ACM.

Kelner, J., Koehler, F., Meka, R., and Moitra, A. (2019). Learning some popular gaussian graphical models without condition number bounds.

Klivans, A. R. and Meka, R. (2017). Learning graphical models using multiplicative weights. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 343–354.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., et al. (2012). High-dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326.

Liu, H., Lafferty, J. D., and Wasserman, L. A. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328.

Lovász, L. and Simonovits, M. (1993). Random walks in a convex body and an improved volume algorithm. *Random Struct. Algorithms*, 4(4):359–412.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Misra, S., Vuffray, M., and Lokhov, A. Y. (2017). Information theoretic optimal learning of gaussian graphical models. *arXiv preprint arXiv:1703.04886*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., Yu, B., et al. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Information Theory*, 58(7):4117–4134.

Sessak, V. and Monasson, R. (2009). Small-correlation expansions for the inverse ising problem. *Journal of Physics A: Mathematical and Theoretical*, 42(5):055001.

Suggala, A. S., Kolar, M., and Ravikumar, P. (2017). The expxorcist: Nonparametric graphical models via conditional exponential densities. In *Advances in Neural Information Processing Systems*, pages 4446–4456.

Sun, S., Kolar, M., and Xu, J. (2015). Learning structured densities via infinite dimensional exponential families. In *Advances in Neural Information Processing Systems*, pages 2287–2295.

Tansey, W., Padilla, O. H. M., Suggala, A. S., and Ravikumar, P. (2015). Vector-space markov random fields via exponential families. In *International Conference on Machine Learning*, pages 684–692.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vempala, S. (2005). Geometric random walks: A survey. *Combinatorial and Computational Geometry*, pages 573–612.

Vuffray, M., Misra, S., and Lokhov, A. Y. (2019). Efficient learning of discrete graphical models. *CoRR*, abs/1902.00600.

Vuffray, M., Misra, S., Lokhov, A. Y., and Chertkov, M. (2016). Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

Wainwright, M. J., Ravikumar, P., and Lafferty, J. D. (2006). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems*, pages 1465–1472.

Wang, W., Wainwright, M. J., and Ramchandran, K. (2010). Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE.

Wu, S., Sanghavi, S., and Dimakis, A. G. (2018). Sparse logistic regression learns all discrete pairwise graphical models. *CoRR*, abs/1810.11905.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.*, 16:3813–3847.

Yang, Z., Ning, Y., and Liu, H. (2018). On semiparametric exponential family graphical models. *J. Mach. Learn. Res.*, 19:57:1–57:59.

Yuan, X., Li, P., Zhang, T., Liu, Q., and Liu, G. (2016). Learning additive exponential family graphical models via $\ell_{2,1}$-norm regularized m-estimation. In *Advances in Neural Information Processing Systems*, pages 4367–4375.