Appendix

Organization. In Appendix A, we provide an illustrative example to ease through the notations introduced in Section 2. In Appendix B, we derive the two forms of the conditional density used in Section 3 and provide lower and upper bounds on the conditional density. In Appendix C, we provide the proof of Theorem 4.1. In Appendix D, we provide the proof of Theorem 4.2. In Appendix E, we state the two key lemmas required in the proof of Theorem 4.3 and 4.4 — Lemma E.1 provides error bounds on edge parameter estimation using GRISE and Lemma E.2 provides error bounds on node parameter estimation using the three-step procedure from Section 3. In Appendix F, we provide the proof of Theorem 4.3 that relies on Lemma E.1. In Appendix G, we provide the proof of Theorem 4.4 that relies on Lemma E.1 and Lemma E.2. In Appendix H, we provide the proof of Proposition 4.1. In Appendix I, we state the two key propositions required in the proof of Lemma E.1 — Proposition I.1 bounds the gradient of the GISO and Proposition I.2 shows that the GISO obeys a restricted strong convexity like property. In Appendix J, we provide the proof of Lemma E.1. In Appendix K, we provide the proof of Proposition I.1 In Appendix L, we provide the proof of Proposition I.2. In Appendix M, we provide the *Generalized Interaction Screening* algorithm (Algorithm 1) and its computational complexity (Proposition M.1). In Appendix N, we present a robust variation of the sparse linear regression. In Appendix O, we state the two key propositions required in the proof of Lemma E.2 — Proposition O.1 provides guarantees for learning the conditional mean parameter vector and Proposition O.2 provides guarantees for learning the conditional canonical parameter vector. In Appendix P, we provide the proof of Lemma E.2. In Appendix Q, we provide the proof of Proposition O.1. In Appendix R, we discuss the theoretical properties of Algorithm 3 (used in the proof of Proposition O.2). In Appendix S, we provide the proof of Proposition O.2. In Appendix T, we discuss a few examples of distributions that naturally satisfy Condition 4.1. In Appendix U, we provide a few discussions.

A Notations via an example

In this appendix, we will provide an illustrative example to ease through the notations introduced in Section 2. Let $\forall i \in [p]$, $\mathcal{X}_i = [-b, b]$. Therefore $b_l = b_u = 2b$. Consider the density as shown below.

$$f_{\mathsf{x}}(\mathbf{x};\boldsymbol{\theta}^*) \propto \exp\Bigg(\sum_{i\in[p]}[\theta_1^{*(i)}x_i + \theta_2^{*(i)}x_i^2] + \sum_{i\in[p]}\sum_{j>i}[\theta_{1,1}^{*(ij)}x_ix_j + \theta_{1,2}^{*(ij)}x_ix_j^2 + \theta_{2,1}^{*(ij)}x_i^2x_j + \theta_{2,2}^{*(ij)}x_i^2x_j^2]\Bigg).$$

For this density, we have the following.

$$\begin{aligned}
g_i(x_i) &= \theta_1^{*(i)}x_i + \theta_2^{*(i)}x_i^2\\
g_{ij}(x_i,x_j) &= \theta_{1,1}^{*(ij)}x_ix_j + \theta_{1,2}^{*(ij)}x_ix_j^2 + \theta_{2,1}^{*(ij)}x_i^2x_j + \theta_{2,2}^{*(ij)}x_i^2x_j^2\\
\boldsymbol{\theta}^{*(i)} &= (\theta_1^{*(i)}, \theta_2^{*(i)})\\
\boldsymbol{\theta}^{*(ij)} &= (\theta_{1,1}^{*(ij)}, \theta_{1,2}^{*(ij)}, \theta_{2,1}^{*(ij)}, \theta_{2,2}^{*(ij)})\\
\boldsymbol{\phi}(x_i) &= (\boldsymbol{\phi}_1(x_i), \boldsymbol{\phi}_2(x_i)) = (x_i, x_i^2)\\
\boldsymbol{\psi}(x_i,x_j) &= (\boldsymbol{\psi}_{11}(x_i,x_j), \boldsymbol{\psi}_{12}(x_i,x_j), \boldsymbol{\psi}_{21}(x_i,x_j), \boldsymbol{\psi}_{22}(x_i,x_j)) = (x_ix_j, x_ix_j^2, x_i^2x_j, x_i^2x_j^2)\\
\phi_{\max} &= \max\{b, b^2\}\\
\bar{\phi}_{\max} &= \max\{1, 2b\}\\
\gamma &= \theta_{\max}(4d+2)\\
\varphi_{\max} &= 2\max\{b, b^4\}.
\end{aligned}$$

For node-wise notations, let us fix $i = 1$. Then, we have

$$\begin{aligned}
\boldsymbol{\vartheta}^{*(1)} &= (\theta_1^{*(1)}, \theta_2^{*(1)}, \theta_{1,1}^{*(12)}, \theta_{1,2}^{*(12)}, \theta_{2,1}^{*(12)}, \theta_{2,2}^{*(12)}, \cdots, \theta_{1,1}^{*(1p)}, \theta_{1,2}^{*(1p)}, \theta_{2,1}^{*(1p)}, \theta_{2,2}^{*(1p)})\\
\boldsymbol{\vartheta}_E^{*(1)} &= (\theta_{1,1}^{*(12)}, \theta_{1,2}^{*(12)}, \theta_{2,1}^{*(12)}, \theta_{2,2}^{*(12)}, \cdots, \theta_{1,1}^{*(1p)}, \theta_{1,2}^{*(1p)}, \theta_{2,1}^{*(1p)}, \theta_{2,2}^{*(1p)})\\
\boldsymbol{\phi}^{(1)}(x_1) &= (x_1, x_1^2 - b^2/3)\\
\boldsymbol{\psi}^{(1j)}(x_1,x_j) &= (x_1x_j, x_1x_j^2, (x_1^2 - b^2/3)x_j, (x_1^2 - b^2/3)x_j^2)
\end{aligned}$$

$$\boldsymbol{\varphi}^{(1)}(\mathbf{x}) = (x_1, x_1^2 - b^2/3, x_1x_2, x_1x_2^2, (x_1^2 - b^2/3)x_2, (x_1^2 - b^2/3)x_2^2, \cdots, x_1x_p, x_1x_p^2, (x_1^2 - b^2/3)x_p, (x_1^2 - b^2/3)x_p^2).$$

B Conditional density

In this appendix, we derive the two forms of the conditional density of x_i for $i \in [p]$ i.e., $f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)})$ used in Section 3. We further obtain lower and upper bounds on this conditional density.

B.1 Derivation of the two forms of conditional density

We will first derive the form of conditional density in (5). For any $i \in [p]$, the conditional density of node x_i given the values taken by all other nodes is obtained by applying Bayes' theorem to $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*)$ and is given by

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) = \frac{\exp\Big(\boldsymbol{\theta}^{*(i)^T}\boldsymbol{\phi}(x_i) + \sum_{j\in[p], j\neq i}\boldsymbol{\theta}^{*(ij)^T}\boldsymbol{\psi}(x_i, x_j)\Big)}{\int_{x_i\in\mathcal{X}_i}\exp\Big(\boldsymbol{\theta}^{*(i)^T}\boldsymbol{\phi}(x_i) + \sum_{j\in[p], j\neq i}\boldsymbol{\theta}^{*(ij)^T}\boldsymbol{\psi}(x_i, x_j)\Big)dx_i}, \tag{13}$$

where $\mathsf{x}_{-i} \coloneqq \mathbf{x} \setminus \mathsf{x}_i$ and $x_{-i} \coloneqq \mathbf{x} \setminus x_i$. Recall definition of locally centered basis functions in (3) and (4) from perspective of $i \in [p], j \in [p]\backslash\{i\}$. For $x \in \mathcal{X}_i$, $x' \in \mathcal{X}_j$

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

where $\mathcal{U}_{\mathcal{X}_i}(y)$ denotes the uniform density on $\mathcal{X}_i$. We can rewrite (13) as

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) = \frac{\exp\Big(\boldsymbol{\theta}^{*(i)^T}\boldsymbol{\phi}^{(i)}(x_i) + \sum_{j\in[p], j\neq i}\boldsymbol{\theta}^{*(ij)^T}\boldsymbol{\psi}^{(ij)}(x_i, x_j)\Big)}{\int_{x_i\in\mathcal{X}_i}\exp\Big(\boldsymbol{\theta}^{*(i)^T}\boldsymbol{\phi}^{(i)}(x_i) + \sum_{j\in[p], j\neq i}\boldsymbol{\theta}^{*(ij)^T}\boldsymbol{\psi}^{(ij)}(x_i, x_j)\Big)dx_i}.$$

Recalling notation of $\boldsymbol{\vartheta}^{*(i)}$ and $\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})$ from Section 2, this results in

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) = \frac{\exp\Big(\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})\Big)}{\int_{x_i\in\mathcal{X}_i}\exp\Big(\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})\Big)dx_i}. \tag{14}$$

We will now derive the form of conditional density in (9). Using the definition of Kronecker product, the conditional density in (13) can also be written as:

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) = \frac{\exp\Big(\sum_{r\in[k]}\theta_r^{*(i)}\phi_r(x_i) + \sum_{j\neq i}\sum_{r,s\in[k]}\theta_{r,s}^{*(ij)}\phi_r(x_i)\phi_s(x_j)\Big)}{\int_{x_i\in\mathcal{X}_i}\exp\Big(\sum_{r\in[k]}\theta_r^{*(i)}\phi_r(x_i) + \sum_{j\neq i}\sum_{r,s\in[k]}\theta_{r,s}^{*(ij)}\phi_r(x_i)\phi_s(x_j)\Big)dx_i}.$$

Recalling notation of $\boldsymbol{\lambda}^*(x_{-i})$ from Section 3, this results in

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) = \frac{\exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i})\boldsymbol{\phi}(x_i)\Big)}{\int_{x_i\in\mathcal{X}_i}\exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i})\boldsymbol{\phi}(x_i)\Big)dx_i}.$$

B.2 Bounds on conditional density

We will now provide lower and upper bounds on the conditional density of x_i for $i \in [p]$. Let us first bound the locally centered basis functions in (3) and (4). For any $i \in [p], r \in [k]$, let $\boldsymbol{\phi}_r^{(i)}(\cdot)$ denote the r^{th} element of $\boldsymbol{\phi}^{(i)}(\cdot)$. We have $\forall i \in [p], \forall r \in [k]$

$$\Big|\boldsymbol{\phi}_r^{(i)}(x_i)\Big| \overset{(a)}{\leq} \Big|\phi_r(x_i)\Big| + \Big|\int_{y_i\in\mathcal{X}_i}\phi_r(y_i)\mathcal{U}_{\mathcal{X}_i}(y_i)dy_i\Big| \overset{(b)}{\leq} |\phi_r(x_i)| + \int_{y_i\in\mathcal{X}_i}|\phi_r(y_i)|\mathcal{U}_{\mathcal{X}_i}(y_i)dy_i \overset{(c)}{\leq} 2\phi_{\max}.$$

where (a) follows by applying the triangle inequality, (b) follows because the absolute value of an integral is smaller than or equal to the integral of an absolute value and $\mathcal{U}_{\mathcal{X}_i}(\cdot)$ is strictly positive, and (c) follows because $|\phi_r(x)| \leq \phi_{\max} \; \forall r \in [k], x \in \cup_{i\in[p]}\mathcal{X}_i$ and the integral of $\mathcal{U}_{\mathcal{X}_i}(\cdot)$ over $\mathcal{X}_i$ is 1. Therefore,

$$\|\boldsymbol{\phi}^{(i)}(\cdot)\|_\infty \leq 2\phi_{\max}.$$

Similary,

$$\|\boldsymbol{\psi}^{(ij)}(\cdot)\|_\infty \leq 2\phi_{\max}^2.$$

Recall the definition of $\varphi_{\max}$. We now have

$$\|\boldsymbol{\varphi}^{(i)}(\mathbf{x})\|_\infty \leq \varphi_{\max}. \tag{15}$$

Also, recall that $\|\boldsymbol{\vartheta}^{*(i)}\|_1 \leq \gamma$. Using this and (15), we have

$$\exp\Big(-\gamma\varphi_{\max}\Big) \leq \exp\Big(\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big) \leq \exp\Big(\gamma\varphi_{\max}\Big). \tag{16}$$

As a result, we can lower and upper bound the conditional density in (14) as,

$$f_L \coloneqq \frac{\exp\Big(-2\gamma\varphi_{\max}\Big)}{b_u} \leq f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) \leq f_U \coloneqq \frac{\exp\Big(2\gamma\varphi_{\max}\Big)}{b_l}. \tag{17}$$

C Proof of Theorem 4.1

In this appendix, we prove Theorem 4.1. Consider $i \in [p]$. For any $\boldsymbol{\vartheta} \in \Lambda$, recall that the population version of GISO is given by

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

Also, recall that the parametric distribution $m_{\mathbf{x}}^{(i)}(\mathbf{x};\boldsymbol{\vartheta})$ under consideration has the following density:

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

and the density $u_{\mathbf{x}}^{(i)}(\mathbf{x})$ is given by:

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

We show that minimizing $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$ is equivalent to minimizing the KL-divergence between the distribution with density $u_{\mathbf{x}}^{(i)}(\cdot)$ and the distribution with density $m_{\mathbf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})$. In other words, we show that, at the population level, the GRISE is a "local" maximum likelihood estimate. We further show that the true parameter vector $\boldsymbol{\vartheta}^{*(i)}$ for $i \in [p]$ is a unique minimizer of $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$. We restate the Theorem below and then provide the proof.

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

$$\underset{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma}{\arg\min}\; D(u_{\mathbf{x}}^{(i)}(\cdot) \,\|\, m_{\mathbf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})) = \underset{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma}{\arg\min}\; \mathcal{S}^{(i)}(\boldsymbol{\vartheta}).$$

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

Proof of Theorem 4.1. We will first write $m_{\mathbf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})$ in terms of $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$. We have

$$m_{\mathbf{x}}^{(i)}(\mathbf{x};\boldsymbol{\vartheta}) = \frac{f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}^*)\exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)}{\int_{\mathbf{x}\in\mathcal{X}} f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}^*)\exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)d\mathbf{x}}$$

$$\overset{(a)}{=} \frac{f_{\mathsf{x}}(\mathbf{x};\boldsymbol{\theta}^*)\exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)}{\mathcal{S}^{(i)}(\boldsymbol{\vartheta})},$$

where (a) follows from definition of $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$.

Now let us write an alternative expression for $u_{\mathsf{x}}^{(i)}(\mathbf{x})$ which does not depend on x_i functionally. We have

$$\begin{aligned} u_{\mathsf{x}}^{(i)}(\mathbf{x}) &\overset{(a)}{\propto} f_{\mathsf{x}_{-i}}(x_{-i};\boldsymbol{\theta}^*) \times f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i}=x_{-i};\boldsymbol{\vartheta}^{*(i)}) \times \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big) \\ &\overset{(b)}{\propto} \frac{f_{\mathsf{x}_{-i}}(x_{-i};\boldsymbol{\theta}^*)}{\int_{x_i\in\mathcal{X}_i}\exp\Big(\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)dx_i}, \end{aligned} \tag{18}$$

where (a) follows from $f_{\mathsf{x}}(\cdot;\boldsymbol{\theta}^*) = f_{\mathsf{x}_{-i}}(\cdot;\boldsymbol{\theta}^*) \times f_{\mathsf{x}_i}(\cdot|\mathsf{x}_{-i}=x_{-i};\boldsymbol{\vartheta}^{*(i)})$ and (b) follows from (14).

We will now simplify the KL-divergence between $u_{\mathsf{x}}^{(i)}(\cdot)$ and $m_{\mathsf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})$. For any $l \in [k+k^2(p-1)]$, let $\boldsymbol{\vartheta}_l$ denote the l^{th} component of $\boldsymbol{\vartheta}$ and $\boldsymbol{\varphi}_l^{(i)}(\mathbf{x})$ denote the l^{th} component of $\boldsymbol{\varphi}^{(i)}(\mathbf{x})$.

$$\begin{aligned} &D(u_{\mathsf{x}}^{(i)}(\mathbf{x}) \parallel m_{\mathsf{x}}^{(i)}(\mathbf{x};\boldsymbol{\vartheta})) \\ &= \int_{\mathbf{x}\in\mathcal{X}} u_{\mathsf{x}}^{(i)}(\mathbf{x}) \log\bigg(\frac{u_{\mathsf{x}}^{(i)}(\mathbf{x})\mathcal{S}^{(i)}(\boldsymbol{\vartheta})}{f_{\mathsf{x}}(\mathbf{x};\boldsymbol{\theta}^*)\exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)}\bigg)d\mathbf{x} \\ &\overset{(a)}{=} \int_{\mathbf{x}\in\mathcal{X}} u_{\mathsf{x}}^{(i)}(\mathbf{x}) \log\bigg(\frac{u_{\mathsf{x}}^{(i)}(\mathbf{x})}{f_{\mathsf{x}}(\mathbf{x};\boldsymbol{\theta}^*)}\bigg)d\mathbf{x} + \int_{\mathbf{x}\in\mathcal{X}} u_{\mathsf{x}}^{(i)}(\mathbf{x}) \times \boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})d\mathbf{x} + \log\mathcal{S}^{(i)}(\boldsymbol{\vartheta}) \\ &= \int_{\mathbf{x}\in\mathcal{X}} u_{\mathsf{x}}^{(i)}(\mathbf{x}) \log\bigg(\frac{u_{\mathsf{x}}^{(i)}(\mathbf{x})}{f_{\mathsf{x}}(\mathbf{x};\boldsymbol{\theta}^*)}\bigg)d\mathbf{x} + \sum_l \Big[\boldsymbol{\vartheta}_l \int_{\mathbf{x}\in\mathcal{X}} u_{\mathsf{x}}^{(i)}(\mathbf{x}) \times \boldsymbol{\varphi}_l^{(i)}(\mathbf{x})d\mathbf{x}\Big] + \log\mathcal{S}^{(i)}(\boldsymbol{\vartheta}) \\ &\overset{(b)}{=} \int_{\mathbf{x}\in\mathcal{X}} u_{\mathsf{x}}^{(i)}(\mathbf{x}) \log\bigg(\frac{u_{\mathsf{x}}^{(i)}(\mathbf{x})}{f_{\mathsf{x}}(\mathbf{x};\boldsymbol{\theta}^*)}\bigg)d\mathbf{x} + \sum_l \Big[\boldsymbol{\vartheta}_l \int_{x_{-i}\in\prod_{j\neq i}\mathcal{X}_j} u_{\mathsf{x}}^{(i)}(\mathbf{x})\Big[\int_{x_i\in\mathcal{X}_i} \boldsymbol{\varphi}_l^{(i)}(\mathbf{x})dx_i\Big]dx_{-i}\Big] + \log\mathcal{S}^{(i)}(\boldsymbol{\vartheta}) \\ &\overset{(c)}{=} \int_{\mathbf{x}\in\mathcal{X}} u_{\mathsf{x}}^{(i)}(\mathbf{x}) \log\bigg(\frac{u_{\mathsf{x}}^{(i)}(\mathbf{x})}{f_{\mathsf{x}}(\mathbf{x};\boldsymbol{\theta}^*)}\bigg)d\mathbf{x} + \log\mathcal{S}^{(i)}(\boldsymbol{\vartheta}), \end{aligned}$$

where (a) follows because $\log(ab) = \log a + \log b$ and $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$ is a constant, (b) follows because $u_{\mathsf{x}}^{(i)}(\cdot)$ does not functionally depend on $x_i \in \mathcal{X}_i$ as shown in (18), and (c) follows because for any $l \in [k+k^2(p-1)]$ the basis function $\boldsymbol{\varphi}_l^{(i)}(\cdot)$ is locally centered from perspective of i i.e., $\int_{x_i\in\mathcal{X}_i}\boldsymbol{\varphi}_l^{(i)}(\mathbf{x})dx_i = 0$. Observing that the first term in the above equation is independent on $\boldsymbol{\vartheta}$, we can write

$$\underset{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma}{\arg\min}\; D(u_{\mathsf{x}}^{(i)}(\cdot) \parallel m_{\mathsf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})) = \underset{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma}{\arg\min}\; \log\mathcal{S}^{(i)}(\boldsymbol{\vartheta}) = \underset{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma}{\arg\min}\; \mathcal{S}^{(i)}(\boldsymbol{\vartheta}).$$

Further, the KL-divergence between $u_{\mathsf{x}}^{(i)}(\cdot)$ and $m_{\mathsf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})$ is minimized when $u_{\mathsf{x}}^{(i)}(\cdot) = m_{\mathsf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})$. Recall that the basis functions are such that the exponential family is minimal. Therefore, $u_{\mathsf{x}}^{(i)}(\cdot) = m_{\mathsf{x}}^{(i)}(\cdot;\boldsymbol{\vartheta})$ only when $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}^{*(i)}$. Thus,

$$\boldsymbol{\vartheta}^{*(i)} \in \underset{\boldsymbol{\vartheta}\in\Lambda:\|\boldsymbol{\vartheta}\|_1\leq\gamma}{\arg\min}\; \mathcal{S}^{(i)}(\boldsymbol{\vartheta}),$$

and it is a unique minimizer of $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$.

Similar analysis works for MRFs with discrete variables as well i.e., the setting considered in Vuffray et al. (2019). □

D Proof of Theorem 4.2

In this appendix, we prove Theorem 4.2. We will use the theory of M-estimation. In particular, we observe that $\hat{\boldsymbol{\vartheta}}_n^{(i)}$ is an M-estimator and invoke Theorem 4.1.1 and Theorem 4.1.3 of Amemiya (1985) for consistency and normality of M-estimators respectively. We restate the Theorem below and then provide the proof.

Theorem 4.2. *Given $i \in [p]$ and n independent samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ 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 $\boldsymbol{\Sigma}$.*

Proof of Theorem 4.2. **Consistency.** We will first show that the GRISE is a consistent estimator i.e., as $n \to \infty$, $\hat{\boldsymbol{\vartheta}}_n^{(i)} \xrightarrow{p} \boldsymbol{\vartheta}^{*(i)}$.

Recall (Amemiya, 1985, Theorem 4.1.1): Let $y_1, \cdots, y_n$ be i.i.d. samples of a random variable y. Let $q(y; \vartheta)$ be some function of y parameterized by $\vartheta \in \Theta$. Let ϑ^* be the true underlying parameter. Define

$$Q_n(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} q(y_i; \vartheta), \tag{19}$$

and

$$\hat{\vartheta}_n \in \arg\min_{\vartheta \in \Theta} Q_n(\vartheta). \tag{20}$$

The M-estimator $\hat{\vartheta}_n$ is consistent for ϑ^* i.e., $\hat{\vartheta}_n \xrightarrow{p} \vartheta^*$ as $n \to \infty$ if,

(a) Θ is compact,

(b) $Q_n(\vartheta)$ converges uniformly in probability to a non-stochastic function $Q(\vartheta)$,

(c) $Q(\vartheta)$ is continuous, and

(d) $Q(\vartheta)$ is uniquely minimzed at ϑ^*.

Comparing (6) and (7) with (19) and (20), we only need to show that the above regularity conditions (a)-(d) hold for $Q_n(\vartheta) \coloneqq \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})$ in order to prove that $\hat{\boldsymbol{\vartheta}}_n^{(i)} \xrightarrow{p} \boldsymbol{\vartheta}^{*(i)}$ as $n \to \infty$. We have the following:

(a) The parameter space Λ is bounded and closed. Therefore, we have compactness.

(b) Recall (Jennrich, 1969, Theorem 2): Let $y_1, \cdots, y_n$ be i.i.d. samples of a random variable y. Let $g(y; \vartheta)$ be a function of ϑ parameterized by $\vartheta \in \Theta$. Suppose (a) Θ is compact, (b) $g(\mathsf{y}, \vartheta)$ is continuous at each $\vartheta \in \Theta$ with probability one, (c) $g(\mathsf{y}, \vartheta)$ is dominated by a function $G(\mathsf{y})$ i.e., $|g(\mathsf{y}, \vartheta)| \leq G(\mathsf{y})$, and (d) $\mathbb{E}[G(\mathsf{y})] < \infty$. Then, $n^{-1} \sum_t g(y_t, \vartheta)$ converges uniformly in probability to $\mathbb{E}[g(\mathsf{y}, \vartheta)]$.

Using this theorem with $\mathsf{y} \coloneqq \mathbf{x}$, $y_t \coloneqq \mathbf{x}^{(t)}$, $\Theta \coloneqq \Lambda$, $g(\mathsf{y}, \vartheta) \coloneqq \exp\Big(- \boldsymbol{\vartheta}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)$, $G(\mathsf{y}) \coloneqq \exp(\gamma \varphi_{\max})$, we conclude that $\mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})$ converges to $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$ uniformly in probability.

(c) $\exp\Big(- \boldsymbol{\vartheta}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)$ is a continuous function of $\boldsymbol{\vartheta} \in \Lambda$. Therefore, we have continuity of $\mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})$ for all $\boldsymbol{\vartheta} \in \Lambda$. Further, $f_{\mathsf{x}}(\cdot; \boldsymbol{\theta}^*)$ does not functionally depend on $\boldsymbol{\vartheta}$. Therefore, we have continuity of $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$ for all $\boldsymbol{\vartheta} \in \Lambda$.

(d) From Theorem 4.1, $\boldsymbol{\vartheta}^{*(i)}$ is a unique minimizer of $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$.

Therefore, we have asymptotic consistency for GRISE.

Normality. We will now show that the GRISE is asymptotically normal i.e., $\sqrt{n}(\hat{\boldsymbol{\vartheta}}_n^{(i)} - \boldsymbol{\vartheta}^{*(i)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, B(\boldsymbol{\vartheta}^{*(i)})^{-1} A(\boldsymbol{\vartheta}^{*(i)}) B(\boldsymbol{\vartheta}^{*(i)})^{-1})$.

Recall (Amemiya, 1985, Theorem 4.1.3): Let $y_1, \cdots, y_n$ be i.i.d. samples of a random variable y. Let $q(y; \vartheta)$ be some function of y parameterized by $\vartheta \in \Theta$. Let ϑ^* be the true underlying parameter. Define

$$Q_n(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} q(y_i; \vartheta), \tag{21}$$

and

$$\hat{\vartheta}_n \in \arg\min_{\vartheta} Q_n(\vartheta). \tag{22}$$

The M-estimator $\hat{\vartheta}_n$ is normal for ϑ^* i.e., $\sqrt{n}(\hat{\vartheta}_n - \vartheta^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, B^{-1}(\vartheta^*)A(\vartheta^*)B^{-1}(\vartheta^*))$ if

(a) $\hat{\vartheta}_n$, the minimzer of $Q_n(\cdot)$, is consistent for ϑ^*,

(b) ϑ^* lies in the interior of the parameter space Θ,

(c) Q_n is twice continuously differentiable in an open and convex neighbourhood of ϑ^*,

(d) $\sqrt{n}\nabla Q_n(\vartheta)|_{\vartheta=\vartheta^*} \xrightarrow{d} \mathcal{N}(\mathbf{0}, A(\vartheta^*))$, and

(e) $\nabla^2 Q_n(\vartheta)|_{\vartheta=\hat{\vartheta}_n} \xrightarrow{p} B(\vartheta^*)$ with $B(\vartheta)$ finite, non-singular, and continuous at ϑ^*,

Comparing (6) and (7) with (21) and (22), we only need to show that the above regularity conditions (a)-(e) hold for $Q_n(\vartheta) \coloneqq \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})$ in order to prove that the GRISE is asymptotically normal. We have the following:

(a) We have already established that $\hat{\boldsymbol{\vartheta}}_n^{(i)}$ is consistent for $\boldsymbol{\vartheta}^{*(i)}$.

(b) We assume that none of the parameter is equal to the boundary values of $\theta_{\min_+}$ or $\theta_{\max}$. Therefore, $\boldsymbol{\vartheta}^{*(i)}$ lies in the interior of Λ.

(c) From (6), we have

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

For any $l \in [k + k^2(p-1)]$, let $\boldsymbol{\vartheta}_l$ denote the l^{th} component of $\boldsymbol{\vartheta}$ and $\boldsymbol{\varphi}_l^{(i)}(\mathbf{x}^{(t)})$ denote the l^{th} component of $\boldsymbol{\varphi}^{(i)}(\mathbf{x}^{(t)})$. For any $l_1, l_2 \in [k + k^2(p-1)]$, we have

$$\frac{\partial^2 \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_{l_1} \partial \boldsymbol{\vartheta}_{l_2}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x}^{(t)}) \boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x}^{(t)}) \exp\Big(- \boldsymbol{\vartheta}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x}^{(t)})\Big).$$

Thus, $\partial^2 \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}_{l_1} \partial \boldsymbol{\vartheta}_{l_2}$ exists. Using the continuity of $\boldsymbol{\varphi}^{(i)}(\cdot)$ and $\exp\Big(- \boldsymbol{\vartheta}^T \boldsymbol{\varphi}^{(i)}(\cdot)\Big)$, we see that $\partial^2 \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})/\partial \boldsymbol{\vartheta}_{l_1} \partial \boldsymbol{\vartheta}_{l_2}$ is continuous in an open and convex neighborhood of $\boldsymbol{\vartheta}^{*(i)}$.

(d) For any $l \in [k + k^2(p-1)]$, define the following random variable:

$$x_{i,l} \coloneqq -\boldsymbol{\varphi}_l^{(i)}(\mathbf{x}) \exp\Big(- {\boldsymbol{\vartheta}^{*(i)}}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big).$$

The l^{th} component of the gradient of the GISO evaluated at $\boldsymbol{\vartheta}^{*(i)}$ is given by

$$\left.\frac{\partial \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}_l}\right|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{*(i)}} = \frac{1}{n}\sum_{t=1}^{n} -\boldsymbol{\varphi}_l^{(i)}(\mathbf{x}^{(t)}) \exp\Big(- {\boldsymbol{\vartheta}^{*(i)}}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x}^{(t)})\Big).$$

Each term in the above summation is distributed as the random variable $x_{i,l}$. The random variable $x_{i,l}$ has zero mean (see Lemma K.1). Using this and the multivariate central limit theorem (Van der Vaart, 2000), we have

$$\sqrt{n}\nabla \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{*(i)}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, A(\boldsymbol{\vartheta}^{*(i)})),$$

where $A(\boldsymbol{\vartheta}^{*(i)})$ is the covariance matrix of $\boldsymbol{\varphi}^{(i)}(\mathbf{x}) \exp\big(- {\boldsymbol{\vartheta}^{*(i)}}^T \boldsymbol{\varphi}^{(i)}(\mathbf{x})\big)$.

(e) We will first show that the following is true:

$$\nabla^2 \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}_n^{(i)}} \xrightarrow{p} \nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{*(i)}}. \tag{23}$$

To begin with, using the uniform law of large numbers (Jennrich, 1969, Theorem 2) for any $\boldsymbol{\vartheta} \in \Lambda$ results in

$$\nabla^2 \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}) \xrightarrow{p} \nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta}). \tag{24}$$

Using the consistency of $\hat{\boldsymbol{\vartheta}}_n^{(i)}$ and the continuous mapping theorem, we have

$$\nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}_n^{(i)}} \xrightarrow{p} \nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\boldsymbol{\vartheta}^{*(i)}}. \tag{25}$$

Let $l_1, l_2 \in [k + k^2(p-1)]$. From (24) and (25), for any $\epsilon > 0$, for any $\delta > 0$, there exists integers n_1, n_2 such that

$$\begin{aligned}
\mathbb{P}(|\big[\nabla^2 \mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)})\big]_{l_1,l_2} - \big[\nabla^2 \mathcal{S}^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)})\big]_{l_1,l_2}| > \epsilon/2) \leq \delta/2 \qquad & \text{if } n \geq n_1 \\
\mathbb{P}(|\big[\nabla^2 \mathcal{S}^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)})\big]_{l_1,l_2} - \big[\nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta}^{*(i)})\big]_{l_1,l_2}| > \epsilon/2) \leq \delta/2 \qquad & \text{if } n \geq n_2.
\end{aligned}$$

Now for $n \geq \max\{n_1, n_2\}$, we have

$$\begin{aligned}
&\mathbb{P}(|\big[\nabla^2 \mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)})\big]_{l_1,l_2} - \big[\nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta}^{*(i)})\big]_{l_1,l_2}| > \epsilon) \\
\leq &\mathbb{P}(|\big[\nabla^2 \mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)})\big]_{l_1,l_2} - \big[\nabla^2 \mathcal{S}^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)})\big]_{l_1,l_2}| > \epsilon/2) + \mathbb{P}(|\big[\nabla^2 \mathcal{S}^{(i)}(\hat{\boldsymbol{\vartheta}}_n^{(i)})\big]_{l_1,l_2} - \big[\nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta}^{*(i)})\big]_{l_1,l_2}| > \epsilon/2) \\
\leq &\delta/2 + \delta/2 = \delta.
\end{aligned}$$

Thus, we have (23). Using (10), we have

$$\begin{aligned}
\big[\nabla^2 \mathcal{S}^{(i)}(\boldsymbol{\vartheta}^{*(i)})\big]_{l_1,l_2} &= \mathbb{E}\bigg[\boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x})\boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)\bigg] \\
&\overset{(b)}{=} \mathbb{E}\bigg[\boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x})\boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)\bigg] - \mathbb{E}\bigg[\boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x})\bigg]\mathbb{E}\bigg[\boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)\bigg] \\
&= \text{cov}\bigg(\boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x}), \boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\Big)\bigg),
\end{aligned}$$

where (b) follows because $\mathbb{E}[\boldsymbol{\varphi}_{l}^{(i)}(\mathbf{x}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x}\Big)] = 0$ for any $l \in [k + k^2(p-1)]$ (see Lemma K.1). Therefore, we have

$$\nabla^2 \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta})|_{\boldsymbol{\vartheta}=\hat{\boldsymbol{\vartheta}}_n^{(i)}} \xrightarrow{p} B(\boldsymbol{\vartheta}^{*(i)}),$$

where $B(\boldsymbol{\vartheta}^{*(i)})$ is the cross-covariance matrix of $\boldsymbol{\varphi}^{(i)}(\mathbf{x})$ and $\boldsymbol{\varphi}^{(i)}(\mathbf{x}) \exp\big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\big)$. Finiteness and continuity of $\boldsymbol{\varphi}^{(i)}(\mathbf{x})$ and $\boldsymbol{\varphi}^{(i)}(\mathbf{x}) \exp\big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\big)$ implies the finiteness and continuity of $B(\boldsymbol{\vartheta}^{*(i)})$.

Therefore, under the assumption that the cross-covariance matrix of $\boldsymbol{\varphi}^{(i)}(\mathbf{x})$ and $\boldsymbol{\varphi}^{(i)}(\mathbf{x}) \exp\big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(\mathbf{x})\big)$ is invertible, and that none of the parameter is equal to the boundary values of $\theta_{\max}$ or $\theta_{\min_+}$, we have the asymptotic normality of GRISE i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\vartheta}}_n^{(i)} - \boldsymbol{\vartheta}^{*(i)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, B(\boldsymbol{\vartheta}^{*(i)})^{-1} A(\boldsymbol{\vartheta}^{*(i)}) B(\boldsymbol{\vartheta}^{*(i)})^{-1}).$$

□

E Supporting lemmas for Theorem 4.3 and 4.4

In this appendix, we will state the two key lemmas required in the proof of Theorem 4.3 and 4.4. Lemma E.1 provides error bounds on edge parameter estimation using GRISE and Lemma E.2 provides error bounds on node parameter estimation using the three-step procedure from Section 3. The proof of Theorem 4.3 is given in Appendix F and the proof of Theorem 4.4 is given in Appendix G. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$, $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$, and $c_1(\alpha)$ from Section 2. Also, define

$$c_3(\alpha) = \frac{2^{20}\pi^4 e^4 k^2 (d+1)^4 \gamma^4 \varphi_{\max}^4 (1+\gamma\varphi_{\max})^4 \exp(8\gamma\varphi_{\max})}{\kappa^4\alpha^8} = O\Bigg(\frac{\exp(\Theta(k^2 d))}{\kappa^4\alpha^8}\Bigg).$$

E.1 Error Bound on Edge Parameter Estimation with GRISE

The following lemma shows that, with enough samples, the parameters associated with the edge potentials can be recovered, within small error, with high probability using the GISO for continuous variables from Section 3.

Lemma E.1. *Let Condition 4.1 be satisfied. Given n independent samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ of $\mathbf{x}$, for each $i \in [p]$, let $\hat{\boldsymbol{\vartheta}}_{\epsilon}^{(i)}$ be an ϵ-optimal solution of (7). Let $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)} = (\hat{\theta}_{ij}, j \neq i, j \in [p])$ be its components corresponding to all possible $p-1$ edges associated with node i. Let $\alpha_1 > 0$ be the prescribed accuracy level. Then, for any $\delta \in (0,1)$,*

$$\|\boldsymbol{\vartheta}_E^{*(i)} - \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}\|_2 \leq \alpha_1, \qquad \forall i \in [p]$$

with probability at least $1-\delta$ as long as

$$n \geq c_1\big(\alpha_1\big)\log\bigg(\frac{2pk}{\sqrt{\delta}}\bigg) = \Omega\Bigg(\frac{\exp(\Theta(k^2 d))}{\kappa^2\alpha_1^4}\log\bigg(\frac{pk}{\sqrt{\delta}}\bigg)\Bigg).$$

The number of computations required scale as

$$c_3(\alpha_1) \times \log\bigg(\frac{2pk}{\sqrt{\delta}}\bigg) \times \log\big(2k^2 p\big) \times p^2 = \Omega\Bigg(\frac{\exp(\Theta(k^2 d))}{\kappa^4\alpha_1^8}\log^2\bigg(\frac{pk}{\sqrt{\delta}}\bigg)p^2\Bigg).$$

The proof of Lemma E.1 is given in Appendix J.

E.2 Error Bound on Node Parameter Estimation

The following lemma shows that, with enough samples, the parameters associated with the node potentials can be recovered, within small error, with high probability using the three-step procedure from Section 3.

Lemma E.2. *Let Condition 4.1 be satisfied. Given n independent samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ of $\mathbf{x}$, for each $i \in [p]$, let $\hat{\boldsymbol{\theta}}^{(i)}$ be an estimate of $\boldsymbol{\theta}^{*(i)}$ obtained using the three-step procedure from Section 3. Then, for any $\alpha_2 \in (0,1)$,*

$$\|\boldsymbol{\theta}^{*(i)} - \hat{\boldsymbol{\theta}}^{(i)}\|_\infty \leq \alpha_2, \qquad \forall i \in [p]$$

with probability at least $1-\alpha_2^4$ as long as

$$\begin{aligned} n &\geq \max\Bigg[c_1\Bigg(\min\bigg\{\frac{\theta_{\min_+}}{3}, \frac{\alpha_2}{2dk\phi_{\max}}\bigg\}\Bigg)\log\bigg(\frac{2^{5/2}pk}{\alpha_2^2}\bigg), c_2(\alpha_2)\Bigg] \\ &= \Omega\Bigg(\frac{\exp(\Theta\Big(k^2 d + d\log\big(\frac{dk}{\alpha_2 q^s}\big)\Big))}{\kappa^2\alpha_2^4} \times \log\bigg(\frac{pk}{\alpha_2^2}\bigg)\Bigg). \end{aligned}$$

The number of computations required scale as

$$c_3\Bigg(\min\bigg\{\frac{\theta_{\min_+}}{3}, \frac{\alpha_2}{2dk\phi_{\max}}\bigg\}\Bigg) \times \log\bigg(\frac{2^{5/2}pk}{\alpha_2^2}\bigg) \times \log\big(2k^2 p\big) \times p^2 = \Omega\Bigg(\frac{\exp(\Theta(k^2 d))}{\kappa^4\alpha_2^8}\log^2\bigg(\frac{pk}{\alpha_2^2}\bigg)p^2\Bigg).$$

The proof of Lemma E.2 is given in Appendix P.

F Proof of Theorem 4.3

In this appendix, we prove Theorem 4.3. See Appendix E.1 for the key lemma (Lemma E.1) required in the proof. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$, $\varphi_{\max} = 2\max\{\phi_{\max}, \phi^2_{\max}\}$ and $c_1(\alpha)$ from Section 2. We restate the Theorem below and then provide the proof.

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

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

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

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

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

Proof of Theorem 4.3. The graph $\hat{G} = ([p], \hat{E})$ is such that:

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

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

Let the number of samples satisfy

$$n \geq c_1\Big(\frac{\theta_{\min_+}}{3}\Big) \log\Big(\frac{2pk}{\sqrt{\delta}}\Big).$$

Recall that $\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon} \in \Lambda$ is an ϵ-optimal solution of GRISE and $\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,E}$ is the component of $\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon}$ associated with the edge potentials. Using Lemma E.1 with $\alpha_1 = \theta_{\min_+}/3$[8] and any $\delta \in (0,1)$, we have with probability at least $1-\delta$,

$$\begin{aligned} \|\boldsymbol{\vartheta}^{*(i)}_{E} - \hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,E}\|_2 &\leq \frac{\theta_{\min_+}}{3}, \qquad \forall i \in [p] \\ \implies \|\boldsymbol{\vartheta}^{*(i)}_{E} - \hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,E}\|_\infty &\overset{(a)}{\leq} \frac{\theta_{\min_+}}{3}, \qquad \forall i \in [p] \end{aligned} \tag{26}$$

where (a) follows because $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2$ for any vector $\mathbf{v}$.

From Section 2, we have $\|\boldsymbol{\vartheta}^{*(i)}\|_{\min_+} \geq \theta_{\min_+}$. This implies that $\|\boldsymbol{\vartheta}^{*(i)}_{E}\|_{\min_+} \geq \theta_{\min_+}$. Combining this with (26), we have with probability at least $1-\delta$,

$$\theta^{*(ij)}_{r,s} = 0 \iff |\hat{\theta}^{(ij)}_{r,s}| \leq \theta_{\min_+}/3, \qquad \forall i \in [p], \forall j \in [p] \setminus \{i\}, \forall r,s \in [k].$$

Therefore, with probability at least $1-\delta$, $E(\boldsymbol{\theta}^*) = \hat{E}$.

Further, from Lemma E.1, the number of computations required for generating $\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,E}$ scale as $\bar{\mathcal{O}}(p^2)$. Also, the number of computations required for generating $\hat{E}$ scale as $O(p^2)$. Therefore, the overall computational complexity is $\bar{\mathcal{O}}(p^2)$. □

[8] The threshold $\theta_{\min_+}/3$ could be replaced by any positive constant smaller than $\theta_{\min_+}/2$. Any threshold smaller than $\theta_{\min_+}/2$ sufficies as it ensures separation between the non-zero parameters and the zero parameters.

G Proof of Theorem 4.4

In this appendix, we prove Theorem 4.4. See Appendix E.1 and Appendix E.2 for two key lemmas (Lemma E.1 and Lemma E.2) required in the proof. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$, $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$ and $c_1(\alpha)$ from Section 2. We restate the Theorem below and then provide the proof.

Theorem 4.4. *Let Condition 4.1 be satisfied. Given n independent samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$ 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

$$\begin{aligned} n &\geq \max\left[c_1\left(\min\left\{\frac{\theta_{\min_+}}{3}, \alpha, \frac{\alpha}{2^{\frac{5}{4}} dk\phi_{\max}}\right\}\right) \log\left(\frac{8pk}{\alpha^2}\right), c_2\left(\frac{\alpha}{2^{\frac{1}{4}}}\right)\right], \\ &= \ \Omega\left(\frac{\exp\left(\Theta\left(k^2 d + d\log\left(\frac{dk}{\alpha q^s}\right)\right)\right)}{\kappa^2 \alpha^4} \times \log\left(\frac{pk}{\alpha^2}\right)\right). \end{aligned}$$

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

Proof of Theorem 4.4. Let the number of samples satisfy

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

For each $i \in [p], \hat{\boldsymbol{\theta}}^{(i)}$ is the estimate of node parameters obtained through robust Lasso. Using Lemma E.2 with $\alpha_2 = 2^{-\frac{1}{4}}\alpha$, the following holds with probability at least $1 - \alpha^4/2$,

$$\begin{aligned} &\|\boldsymbol{\theta}^{*(i)} - \hat{\boldsymbol{\theta}}^{(i)}\|_\infty \leq 2^{-\frac{1}{4}}\alpha, \qquad \forall i \in [p] \\ \implies &\|\boldsymbol{\theta}^{*(i)} - \hat{\boldsymbol{\theta}}^{(i)}\|_\infty \leq \alpha, \qquad \forall i \in [p] \end{aligned} \tag{27}$$

For each $i \in [p]$, $\hat{\boldsymbol{\vartheta}}_{\epsilon}^{(i)}$ is an ϵ-optimal solution of (7) and $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)} = (\hat{\theta}_{ij}, j \neq i, j \in [p])$ is the estimate of edge parameters associated with node i. Using Lemma E.1 with $\alpha_1 = \alpha$ and $\delta = \alpha^4/2$, the following holds with probability at least $1 - \alpha^4/2$,

$$\begin{aligned} &\|\boldsymbol{\vartheta}_E^{*(i)} - \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}\|_2 \leq \alpha, \qquad \forall i \in [p] \\ \implies &\|\boldsymbol{\vartheta}_E^{*(i)} - \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}\|_\infty \overset{(a)}{\leq} \alpha, \qquad \forall i \in [p] \end{aligned} \tag{28}$$

where (a) follows because $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_2$ for any vector $\mathbf{v}$.

Now $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}^*$ obtained after appropriately concatenating $\hat{\boldsymbol{\theta}}^{(i)}$ and $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}$ $\forall i \in [p]$. Combining (27) and (28), we have

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

with probability at least $1 - \alpha^4$. Further, combining the computations from Lemma E.1 and Lemma E.2, the total number of computations scale as $\bar{\mathcal{O}}(p^2)$. □

H GISO: Special instance of the penalized surrogate likelihood

In this appendix, we show that the GISO is a special case of the penalized surrogate likelihood introduced by Jeon and Lin (2006). In other words, we provide the proof of Proposition 4.1.

Consider nonparametric density estimation where densities are of the form $f_{\mathbf{x}}(\mathbf{x}) = e^{\eta(\mathbf{x})}/\int e^{\eta(\mathbf{x})}d\mathbf{x}$ from i.i.d samples $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}$. To circumvent the computational limitation of the exact likelihood-based functionals, Jeon and Lin (2006) proposed to minimize penalized surrogate likelihood. The surrogate likelihood is defined 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. As Proposition 4.1 establishes, GISO is a special case of the surrogate likelihood. We restate the Proposition below and then provide the proof.

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 $\mathcal{X}_i$.*

Proof of Proposition 4.1. Recall that the conditional density of x_i given $\mathsf{x}_{-i} = x_{-i}$ is as follows:

$$f_{\mathsf{x}_i}(x_i|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) \propto \exp\Big(\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})\Big).$$

For a given $\mathsf{x}_{-i} = x_{-i}$, estimation of the conditional density of x_i is equivalent to estimating $\boldsymbol{\vartheta}^{*(i)}$.

For any $\boldsymbol{\vartheta} \in \mathbb{R}^{k+k^2(p-1)}$, let us denote the surrogate likelihood associated with the conditional density of x_i by $\mathcal{L}_n^{(i)}(\boldsymbol{\vartheta})$. We have

$$\mathcal{L}_n^{(i)}(\boldsymbol{\vartheta}) = \frac{1}{n}\sum_{t=1}^{n} \exp\Big(-\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(\mathbf{x}^{(t)})\Big) + \int_{x_i \in \mathcal{X}_i} \rho(x_i) \times \Big(\boldsymbol{\vartheta}^T\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})\Big)dx_i, \tag{29}$$

Let $\rho(\cdot)$ be the uniform density over $\mathcal{X}_i$. Recall that the basis functions, $\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})$, are locally centered and their integral with respect to $\mathcal{U}_{\mathcal{X}_i}$ is 0. Therefore, (29) can be written as

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

□

As we see in the proof above, the equivalence between the GISO and the surrogate likelihood occurs only the integral in (29) is zero. As stated in Jeon and Lin (2006), $\rho(\cdot)$ can be chosen to be equal to any known density and the choice typically depends on mathematical simplicity. Therefore, this provides a motivation to locally center the basis functions to simplify the exposition.

This equivalence provides an association of the GISO to an estimator well-known in the literature and opens up avenues for future explorations.

I Supporting propositions for Lemma E.1

In this appendix, we will state the two key propositions required in the proof of Lemma E.1. Proposition I.1 bounds the gradient of the GISO and Proposition I.2 shows that the GISO obeys a restricted strong convexity like property. The proof of Lemma E.1 is given in Appendix J. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$ and $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$ from Section 2. For any $i \in [p]$, let $\nabla\mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)})$ denote the gradient of the GISO for node i evaluated at $\boldsymbol{\vartheta}^{*(i)}$.

I.1 Bounds on the gradient of the GISO

The following proposition shows that, with enough samples, the ℓ_∞-norm of the gradient of the GISO is bounded with high probability.

Proposition I.1. *Consider any $i \in [p]$. For any $\delta_1 \in (0,1)$, any $\epsilon_1 > 0$, the components of the gradient of the GISO are bounded from above as*

$$\|\nabla \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)})\|_\infty \leq \epsilon_1,$$

with probability at least $1 - \delta_1$ as long as

$$n > \frac{2\varphi_{\max}^2 \exp(2\gamma\varphi_{\max})}{\epsilon_1^2} \log\left(\frac{2p^2k^2}{\delta_1}\right) \;=\; \Omega\left(\frac{\exp(\Theta(k^2 d))}{\epsilon_1^2} \log\left(\frac{pk}{\sqrt{\delta_1}}\right)\right).$$

The proof of proposition I.1 is given in Appendix K.

I.2 Restricted Strong Convexity for GISO

Consider any $\boldsymbol{\vartheta} \in \Lambda$. Let $\Delta = \boldsymbol{\vartheta} - \boldsymbol{\vartheta}^{*(i)}$. Define the residual of the first-order Taylor expansion as

$$\delta\mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)}) = \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)} + \Delta) - \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)}) - \langle \nabla \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)}), \Delta\rangle. \tag{30}$$

Recall that $\boldsymbol{\vartheta}_E^{*(i)}$ denote the component of $\boldsymbol{\vartheta}^{*(i)}$ associated with the edge potentials. Let $\boldsymbol{\vartheta}_E$ denote the component of $\boldsymbol{\vartheta}$ associated with the edge potentials and let Δ_E denote the component of Δ associated with the edge potentials i.e., $\Delta_E = \boldsymbol{\vartheta}_E - \boldsymbol{\vartheta}_E^{*(i)}$.

The following proposition shows that, with enough samples, the GISO obeys a property analogous to the restricted strong convexity with high probability.

Proposition I.2. *Consider any $i \in [p]$. For any $\delta_2 \in (0,1)$, any $\epsilon_2 > 0$, the residual of the first-order Taylor expansion of the GISO satisfies*

$$\delta\mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)}) \geq \exp(-\gamma\varphi_{\max}) \frac{\frac{\kappa}{2\pi e(d+1)}\|\Delta_E\|_2^2 - \epsilon_2\|\Delta\|_1^2}{2 + \varphi_{\max}\|\Delta\|_1},$$

with probability at least $1 - \delta_2$ as long as

$$n > \frac{2\varphi_{\max}^4}{\epsilon_2^2} \log\left(\frac{2p^3k^4}{\delta_2}\right) \;=\; \Omega\left(\frac{1}{\epsilon_2^2} \log\left(\frac{p^3k^4}{\delta_2}\right)\right).$$

The proof of proposition I.2 is given in Appendix L.

J Proof of Lemma E.1

In this appendix, we prove Lemma E.1. See Appendix I.1 and Appendix I.2 for two key propositions (Proposition I.1 and Proposition I.2) required in the proof. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$, $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$ and $c_1(\alpha)$ from Section 2 and the definition of $c_3(\alpha)$ from Appendix E. Recall that $\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}$ is an ϵ-optimal solution of the GISO.

For any $i \in [p]$, let $\nabla \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)})$ denote the gradient of the GISO for node i evaluated at $\boldsymbol{\vartheta}^{*(i)}$. Define $\Delta = \hat{\boldsymbol{\vartheta}}_\epsilon^{(i)} - \boldsymbol{\vartheta}^{*(i)}$ and let Δ_E denote the component of Δ associated with the edge potentials i.e., $\Delta_E = \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)} - \boldsymbol{\vartheta}_E^{*(i)}$. Recall from (30) that $\delta\mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)})$ denotes the residual of the first-order Taylor expansion. We restate the Lemma below and then provide the proof.

Lemma E.1. *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). Let $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)} = (\hat{\theta}_{ij}, j \neq i, j \in [p])$ be its components corresponding to all possible $p-1$ edges associated with node i. Let $\alpha_1 > 0$ be the prescribed accuracy level. Then, for any $\delta \in (0,1)$,*

$$\|\boldsymbol{\vartheta}_E^{*(i)} - \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}\|_2 \leq \alpha_1, \qquad \forall i \in [p]$$

with probability at least $1 - \delta$ *as long as*

$$n \geq c_1(\alpha_1) \log\left(\frac{2pk}{\sqrt{\delta}}\right) \;=\; \Omega\Bigg(\frac{\exp(\Theta(k^2 d))}{\kappa^2 \alpha_1^4} \log\left(\frac{pk}{\sqrt{\delta}}\right)\Bigg).$$

The number of computations required scale as

$$c_3(\alpha_1) \times \log\left(\frac{2pk}{\sqrt{\delta}}\right) \times \log(2k^2 p) \times p^2 \;=\; \Omega\Bigg(\frac{\exp(\Theta(k^2 d))}{\kappa^4 \alpha_1^8} \log^2\left(\frac{pk}{\sqrt{\delta}}\right) p^2\Bigg).$$

Proof of Lemma E.1. Consider any $i \in [p]$. Let the number of samples satisfy

$$n \geq c_1(\alpha_1) \times \log\left(\frac{2pk}{\sqrt{\delta}}\right).$$

We have from (8)

$$\begin{aligned}
\epsilon &\geq \mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}) - \min_{\boldsymbol{\vartheta} \in \Lambda : \|\boldsymbol{\vartheta}\| \leq \gamma} \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}) \\
&\overset{(a)}{\geq} \mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}) - \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)}) \\
&\overset{(b)}{=} \langle \nabla \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)}), \Delta \rangle + \delta \mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)}) \\
&\geq -\|\nabla \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)})\|_\infty \|\Delta\|_1 + \delta \mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)}),
\end{aligned}$$

where (a) follows because $\boldsymbol{\vartheta}^{*(i)} \in \Lambda$ and $\|\boldsymbol{\vartheta}^{*(i)}\| \leq \gamma$ and (b) follows from (30). Using the union bound on Proposition I.1 and Proposition I.2 with $\delta_1 = \frac{\delta}{2}$ and $\delta_2 = \frac{\delta}{2}$ respectively, we have with probability at least $1 - \delta$,

$$\epsilon \geq -\epsilon_1 \|\Delta\|_1 + \exp(-\gamma \varphi_{\max}) \frac{\frac{\kappa}{2\pi e(d+1)} \|\Delta_E\|_2^2 - \epsilon_2 \|\Delta\|_1^2}{2 + \varphi_{\max} \|\Delta\|_1}.$$

This can be rearranged as

$$\|\Delta_E\|_2^2 \leq \frac{2\pi e(d+1)}{\kappa} \Bigg[\exp(\gamma \varphi_{\max}) \times \Big(\epsilon + \epsilon_1 \|\Delta\|_1\Big) \times \Big(2 + \varphi_{\max} \|\Delta\|_1\Big) + \epsilon_2 \|\Delta\|_1^2\Bigg].$$

Using $\|\boldsymbol{\vartheta}^{*(i)}\|_1 \leq \gamma$, $\|\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}\|_1 \leq \gamma$ and the triangle inequality, we see that $\|\Delta\|_1$ is bounded by 2γ. By choosing

$$\epsilon \leq \frac{\kappa \alpha_1^2 \exp(-\gamma \varphi_{\max})}{16\pi e(d+1)(1 + \varphi_{\max}\gamma)}, \epsilon_1 \leq \frac{\kappa \alpha_1^2 \exp(-\gamma \varphi_{\max})}{32\pi e(d+1)\gamma(1 + \varphi_{\max}\gamma)}, \epsilon_2 \leq \frac{\kappa \alpha_1^2}{16\pi e(d+1)\gamma^2},$$

and after some algebra, we obtain that

$$\|\Delta_E\|_2 \leq \alpha_1.$$

Using Proposition M.1, the number of computations required to compute $\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}$ scale as

$$\frac{k^2 \gamma^2 \varphi_{\max}^2 \exp(2\gamma \varphi_{\max}) np}{\epsilon^2} \times \log(2k^2 p).$$

Substituting for ϵ, n and observing that we need to compute the ϵ-optimal estimate for every node, the total number of computations scale as

$$c_3(\alpha_1) \times \log\left(\frac{2pk}{\sqrt{\delta}}\right) \times \log(2k^2 p) \times p^2.$$

□

K Proof of Proposition I.1

In this appendix, we prove Proposition I.1. However, before that, we will provide a supporting Lemma (Lemma K.1) wherein we show that the expected value of a random variable of interest is zero. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$ and $\varphi_{\max} = 2\max\{\phi_{\max}, \phi^2_{\max}\}$ from Section 2. Also, recall the definition of GISO from (6).

For any $l \in [k+k^2(p-1)]$, let $\boldsymbol{\vartheta}^{*(i)}_l$ denote the l^{th} component of $\boldsymbol{\vartheta}^{*(i)}$ and $\boldsymbol{\varphi}^{(i)}_l(x^{(t)}_i; x^{(t)}_{-i})$ denote the l^{th} component of $\boldsymbol{\varphi}^{(i)}(x^{(t)}_i; x^{(t)}_{-i})$. Define the following random variable:

$$\mathsf{x}_{i,l} \coloneqq -\boldsymbol{\varphi}^{(i)}_l(\mathsf{x}_i; \mathsf{x}_{-i}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T} \boldsymbol{\varphi}^{(i)}(\mathsf{x}_i; \mathsf{x}_{-i})\Big). \tag{31}$$

K.1 Supporting Lemma for Proposition I.1

The following Lemma shows that the expectation of the random variable $\mathsf{x}_{i,l}$ defined above is zero.

Lemma K.1. *For any $i \in [p]$ and $l \in [k + k^2(p-1)]$, we have*

$$\mathbb{E}[\mathsf{x}_{i,l}] = 0,$$

where the expectation is with respect to $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*)$.

Proof of Lemma K.1. Fix $i \in [p]$ and $l \in [k + k^2(p-1)]$. Using (31) and Bayes theorem, we have

$$\mathbb{E}[\mathsf{x}_{i,l}] = -\int_{\mathbf{x}\in\mathcal{X}} \boldsymbol{\varphi}^{(i)}_l(x_i; x_{-i}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T} \boldsymbol{\varphi}^{(i)}(x_i; x_{-i})\Big) f_{\mathsf{x}_i}(x_i | \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) f_{\mathsf{x}_{-i}}(x_{-i}; \boldsymbol{\theta}^*) d\mathbf{x}.$$

Using (14) results in

$$\mathbb{E}[\mathsf{x}_{i,l}] = \frac{-\int_{\mathbf{x}\in\mathcal{X}} \boldsymbol{\varphi}^{(i)}_l(x_i; x_{-i}) f_{\mathsf{x}_{-i}}(x_{-i}; \boldsymbol{\theta}^*) d\mathbf{x}}{\int_{x_i\in\mathcal{X}_i} \exp\Big(\boldsymbol{\vartheta}^{*(i)T} \boldsymbol{\varphi}^{(i)}(x_i; x_{-i})\Big) dx_i}.$$

Recall the fact that the basis functions are locally centered with respect to x_i i.e., $\int_{x_i\in\mathcal{X}_i} \boldsymbol{\varphi}^{(i)}_l(\mathbf{x}) dx_i = 0$. Therefore, $\mathbb{E}[\mathsf{x}_{i,l}] = 0$. □

K.2 Proof of Proposition I.1

We restate the Proposition below and then provide the proof.

Proposition I.1. *Consider any $i \in [p]$. For any $\delta_1 \in (0,1)$, any $\epsilon_1 > 0$, the components of the gradient of the GISO are bounded from above as*

$$\|\nabla \mathcal{S}^{(i)}_n(\boldsymbol{\vartheta}^{*(i)})\|_\infty \leq \epsilon_1,$$

with probability at least $1 - \delta_1$ as long as

$$n > \frac{2\varphi^2_{\max}\exp(2\gamma\varphi_{\max})}{\epsilon_1^2} \log\left(\frac{2p^2k^2}{\delta_1}\right) = \Omega\left(\frac{\exp(\Theta(k^2 d))}{\epsilon_1^2} \log\left(\frac{pk}{\sqrt{\delta_1}}\right)\right).$$

Proof of Proposition I.1. Fix $i \in [p]$ and $l \in [k + k^2(p-1)]$. We start by simplifying the gradient of the GISO evaluated at $\boldsymbol{\vartheta}^{*(i)}$. The l^{th} component of the gradient of the GISO evaluated at $\boldsymbol{\vartheta}^{*(i)}$ is given by

$$\frac{\partial \mathcal{S}^{(i)}_n(\boldsymbol{\vartheta}^{*(i)})}{\partial \boldsymbol{\vartheta}^{*(i)}_l} = \frac{1}{n}\sum_{t=1}^{n} -\boldsymbol{\varphi}^{(i)}_l(x^{(t)}_i; x^{(t)}_{-i}) \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T} \boldsymbol{\varphi}^{(i)}(x^{(t)}_i; x^{(t)}_{-i})\Big). \tag{32}$$

Each term in the above summation is distributed as the random variable $\mathsf{x}_{i,l}$. The random variable $\mathsf{x}_{i,l}$ has zero mean (Lemma K.1) and is bounded as follows:

$$\Big|\mathsf{x}_{i,l}\Big| = \Big|\boldsymbol{\varphi}^{(i)}_l(\mathsf{x}_i; \mathsf{x}_{-i})\Big| \times \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T} \boldsymbol{\varphi}^{(i)}(\mathsf{x}_i; \mathsf{x}_{-i})\Big) \overset{(a)}{\leq} \varphi_{\max}\exp(\gamma\varphi_{\max}),$$

where (a) follows from (15) and (16). Using the Hoeffding's inequality, we have

$$\mathbb{P}\bigg(\bigg|\frac{\partial \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}^{*(i)})}{\partial \boldsymbol{\vartheta}_l^{*(i)}}\bigg| > \epsilon_1\bigg) < 2\exp\bigg(-\frac{n\epsilon_1^2}{2\varphi_{\max}^2\exp(2\gamma\varphi_{\max})}\bigg). \tag{33}$$

The proof follows by using (33), the union bound over all $i \in [p]$ and $l \in [k + k^2(p-1)]$, and the fact that $k + k^2(p-1) \leq k^2p$. □

L Proof of Proposition I.2

In this appendix, we prove Proposition I.2. We start by introducing the notion of correlation between the locally centered basis functions and provide a supporting Lemma (Lemma L.1) wherein we will bound the deviation between the true correlation and the empirical correlation. Next, we provide the proof of Proposition I.2. Recall the definitions of $\gamma = \theta_{\max}(k + k^2d)$ and $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$ from Section 2.

For any $l \in [k + k^2(p-1)]$, let $\boldsymbol{\varphi}_l^{(i)}(x_i; x_{-i})$ denotes the l^{th} component of $\boldsymbol{\varphi}^{(i)}(x_i; x_{-i})$. For any $\boldsymbol{\vartheta} \in \Lambda$, let $\Delta = \boldsymbol{\vartheta} - \boldsymbol{\vartheta}^{*(i)}$. Let Δ_E denote the component of Δ associated with the edge potentials. Recall from (30) that $\delta\mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)})$ denotes the residual of the first-order Taylor expansion.

L.1 Correlation between locally centered basis functions

For any $l_1, l_2 \in [k + k^2(p-1)]$ let $H_{l_1 l_2}$ denote the correlation between $\boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x})$ and $\boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x})$ defined as

$$H_{l_1 l_2} = \mathbb{E}\Big[\boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x})\boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x})\Big], \tag{34}$$

and let $\mathbf{H} = [H_{l_1 l_2}] \in \mathbb{R}^{[k+k^2(p-1)]\times[k+k^2(p-1)]}$ be the corresponding correlation matrix. Similarly, we define $\hat{\mathbf{H}}$ based on the empirical estimates of the correlation i.e., $\hat{H}_{l_1 l_2} = \frac{1}{n}\sum_{t=1}^n \boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x}^{(t)})\boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x}^{(t)})$.

The following lemma bounds the deviation between the true correlation and the empirical correlation.

Lemma L.1. *Consider any $i \in [p]$ and $l_1, l_2 \in [k + k^2(p-1)]$. Then, we have for any $\epsilon_2 > 0$,*

$$|\hat{H}_{l_1 l_2} - H_{l_1 l_2}| < \epsilon_2,$$

with probability at least $1 - 2p^3k^4\exp\Big(-\frac{n\epsilon_2^2}{2\varphi_{\max}^4}\Big)$.

Proof of Lemma L.1. Fix $i \in [p]$ and $l_1, l_2 \in [k+k^2(p-1)]$. The random variable defined as $Y_{l_1 l_2} \coloneqq \boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x})\boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x})$ satisfies $|Y_{l_1 l_2}| \leq \varphi_{\max}^2$. Using the Hoeffding inequality we get

$$\mathbb{P}\Big(|\hat{H}_{l_1 l_2} - H_{l_1 l_2}| > \epsilon_2\Big) < 2\exp\bigg(-\frac{n\epsilon_2^2}{2\varphi_{\max}^4}\bigg).$$

The proof follows by using the union bound over all $i \in [p]$ and $l_1, l_2 \in [k + k^2(p-1)]$, and the fact that $k + k^2(p-1) \leq k^2p$. □

L.2 Supporting Lemma for Proposition I.2

The following Lemma provides a lower bound on the residual defined in (30) i.e., $\delta\mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)})$.

Lemma L.2. *Consider any $i \in [p]$. The residual of the first-order Taylor expansion of the* GISO *satisfies*

$$\delta\mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)}) \geq \exp(-\gamma\varphi_{\max})\frac{\Delta^T\hat{\mathbf{H}}\Delta}{2 + \varphi_{\max}\|\Delta\|_1}.$$

Proof of Lemma L.2. Fix any $i \in [p]$. Substituting (6) and (32) in (30), we have

$$\delta\mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)}) = \frac{1}{n}\sum_{t=1}^n \exp\Big(-\boldsymbol{\vartheta}^{*(i)^T}\boldsymbol{\varphi}^{(i)}(x_i^{(t)}; x_{-i}^{(t)})\Big) \times \Big(\exp\Big(-\Delta^T\boldsymbol{\varphi}^{(i)}(x_i^{(t)}; x_{-i}^{(t)})\Big) - 1 + \Delta^T\boldsymbol{\varphi}^{(i)}(x_i^{(t)}; x_{-i}^{(t)})\Big)$$

$$\overset{(a)}{\geq} \exp(-\gamma\varphi_{\max})\frac{1}{n}\sum_{t=1}^{n}\frac{\left(\Delta^T\boldsymbol{\varphi}^{(i)}(x_i^{(t)};x_{-i}^{(t)})\right)^2}{2+|\Delta^T\boldsymbol{\varphi}^{(i)}(x_i^{(t)};x_{-i}^{(t)})|}$$
$$\overset{(b)}{\geq} \exp(-\gamma\varphi_{\max})\frac{\Delta^T\hat{\mathbf{H}}\Delta}{2+\varphi_{\max}\|\Delta\|_1},$$

where (a) follows by using (16) and $e^{-z}-1+z \geq \frac{z^2}{2+|z|}$ $\forall z \in \mathbb{R}$ ($z = \Delta^T\boldsymbol{\varphi}^{(i)}(x_i^{(t)};x_{-i}^{(t)})$ is used here), and (b) follows by using (15), the defintion of $\hat{\mathbf{H}}$, and observing that $\forall\ t \in [n], |\Delta^T\boldsymbol{\varphi}^{(i)}(x_i^{(t)};x_{-i}^{(t)})| \leq \varphi_{\max}\|\Delta\|_1$. □

L.3 Proof of Proposition I.2

We restate the Proposition below and then provide the proof.

Proposition I.2. *Consider any $i \in [p]$. For any $\delta_2 \in (0,1)$, any $\epsilon_2 > 0$, the residual of the first-order Taylor expansion of the GISO satisfies*

$$\delta\mathcal{S}_n^{(i)}(\Delta,\boldsymbol{\vartheta}^{*(i)}) \geq \exp(-\gamma\varphi_{\max})\frac{\frac{\kappa}{2\pi e(d+1)}\|\Delta_E\|_2^2-\epsilon_2\|\Delta\|_1^2}{2+\varphi_{\max}\|\Delta\|_1},$$

with probability at least $1-\delta_2$ as long as

$$n > \frac{2\varphi_{\max}^4}{\epsilon_2^2}\log\left(\frac{2p^3k^4}{\delta_2}\right) = \Omega\left(\frac{1}{\epsilon_2^2}\log\left(\frac{p^3k^4}{\delta_2}\right)\right).$$

Proof of Proposition I.2. Consider any $i \in [p]$. Using Lemma L.2 we have

$$\delta\mathcal{S}_n^{(i)}(\Delta,\boldsymbol{\vartheta}^{*(i)}) \geq \exp(-\gamma\varphi_{\max})\frac{\Delta^T\hat{\mathbf{H}}\Delta}{2+\varphi_{\max}\|\Delta\|_1}$$
$$= \exp(-\gamma\varphi_{\max})\frac{\Delta^T\mathbf{H}\Delta+\Delta^T(\hat{\mathbf{H}}-\mathbf{H})\Delta}{2+\varphi_{\max}\|\Delta\|_1}.$$

Let the number of samples satisfy

$$n > \frac{2\varphi_{\max}^4}{\epsilon_2^2}\log\left(\frac{2p^3k^4}{\delta_2}\right).$$

Using Lemma L.1 and the triangle inequality, we have the following with probability at least $1-\delta_2$.

$$\delta\mathcal{S}_n^{(i)}(\Delta,\boldsymbol{\vartheta}^{*(i)}) \geq \exp(-\gamma\varphi_{\max})\frac{\Delta^T\mathbf{H}\Delta-\epsilon_2\|\Delta\|_1^2}{2+\varphi_{\max}\|\Delta\|_1}, \tag{35}$$

Now we will lower bound $\Delta^T\mathbf{H}\Delta$. First, let us unroll the vector Δ such that $\Delta^{(i)} \in \mathbb{R}^k$ is associated with $\boldsymbol{\phi}^{(i)}(x_i)$ and $\forall j \in [p]\setminus\{i\}$, $\Delta^{(ij)} \in \mathbb{R}^{k^2}$ is associated with $\boldsymbol{\psi}^{(ij)}(x_i,x_j)$. Recall that Δ_E is the component of Δ associated with the edge potentials i.e.,

$$\Delta_E = [\Delta^{(ij)} \in \mathbb{R}^{k^2} : j \in [p], j \neq i]. \tag{36}$$

Using (34) we have

$$\Delta^T\mathbf{H}\Delta = \mathbb{E}\left[\left(\Delta^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\right)^2\right] \overset{(a)}{\geq} \mathbb{V}\mathrm{ar}\big[\Delta^T\boldsymbol{\varphi}^{(i)}(\mathbf{x})\big], \tag{37}$$

where (a) follows from the fact that for any random variable $Z, \mathbb{E}[Z^2] \geq \mathbb{V}\mathrm{ar}[Z]$.

Now consider the graph $G_{-i}(\boldsymbol{\theta}^*)$ obtained from the graph $G(\boldsymbol{\theta}^*)$ by removing the node i and all the edges associated with it. We will next choose an independent set of the graph $G_{-i}(\boldsymbol{\theta}^*)$ with a special property. Let $r_1 \in [p]\setminus\{i\}$ be such that $\|\Delta^{(ir_1)}\|_2 \geq \|\Delta^{(ij)}\|_2$ $\forall j \in [p]\setminus\{i,r_1\}$. Let $r_2 \in [p]\setminus\{i,r_1,\mathcal{N}(r_1)\}$ be such that

$\|\Delta^{(ir_2)}\|_2 \geq \|\Delta^{(ij)}\|_2 \; \forall j \in [p] \setminus \{i, r_1, \mathcal{N}(r_1), r_2\}$, and so on. Denote by $m \geq p/(d+1)$ the total number of nodes selected in this manner, and let $\mathcal{R} = \{r_1, \cdots, r_m\}$. It is easy to see that $\mathcal{R}$ is independent set of the graph $G_{-i}(\boldsymbol{\theta}^*)$ with the following property:

$$\sum_{j \in \mathcal{R}} \|\Delta^{(ij)}\|_2^2 \geq \frac{1}{d+1} \sum_{j \in [p], j \neq i} \|\Delta^{(ij)}\|_2^2. \tag{38}$$

Let $\mathcal{R}^c = [p] \setminus \{i, \mathcal{R}\}$. Using the law of total variance and conditioning on $\mathcal{R}^c$, we can rewrite (37) as

$$\begin{aligned}
\Delta^T \mathbf{H} \Delta &\geq \mathbb{E}\Big[\mathbb{V}\text{ar}\big[\Delta^T \boldsymbol{\varphi}^{(i)}(\mathbf{x}) \big| x_i, x_{\mathcal{R}^c}\big]\Big] \\
&\overset{(a)}{=} \mathbb{E}\Big[\mathbb{V}\text{ar}\Big[\Delta^{(i)^T} \boldsymbol{\phi}^{(i)}(x_i) + \sum_{j \in [p], j \neq i} \Delta^{(ij)^T} \boldsymbol{\psi}^{(ij)}(x_i, x_j) \big| x_i, x_{\mathcal{R}^c}\Big]\Big] \\
&\overset{(b)}{=} \mathbb{E}\Big[\mathbb{V}\text{ar}\Big[\sum_{j \in \mathcal{R}} \Delta^{(ij)^T} \boldsymbol{\psi}^{(ij)}(x_i, x_j) \big| x_i, x_{\mathcal{R}^c}\Big]\Big] \\
&\overset{(c)}{=} \mathbb{E}\Big[\sum_{j \in \mathcal{R}} \mathbb{V}\text{ar}\Big[\Delta^{(ij)^T} \boldsymbol{\psi}^{(ij)}(x_i, x_j) \big| x_i, x_{\mathcal{R}^c}\Big]\Big] \\
&\overset{(d)}{=} \sum_{j \in \mathcal{R}} \mathbb{E}\Big[\mathbb{V}\text{ar}\Big[\Delta^{(ij)^T} \boldsymbol{\psi}^{(ij)}(x_i, x_j) \big| x_i, x_{\mathcal{R}^c}\Big]\Big] \\
&\overset{(e)}{=} \sum_{j \in \mathcal{R}} \mathbb{E}\Big[\mathbb{V}\text{ar}\Big[\Delta^{(ij)^T} \boldsymbol{\psi}^{(ij)}(x_i, x_j) \big| x_{\mathcal{N}(j)}\Big]\Big] \\
&\overset{(f)}{=} \sum_{j \in \mathcal{R}} \mathbb{E}\Big[\mathbb{V}\text{ar}\Big[\Delta^{(ij)^T} \boldsymbol{\psi}^{(ij)}(x_i, x_j) \big| x_{-j}\Big]\Big] \\
&\overset{(g)}{\geq} \frac{1}{2\pi e} \sum_{j \in \mathcal{R}} \mathbb{E}\Big[\exp\Big\{2h\Big[\Delta^{(ij)^T} \boldsymbol{\psi}^{(ij)}(x_i, x_j) \big| x_{-j}\Big]\Big\}\Big] \\
&\overset{(h)}{\geq} \frac{\kappa}{2\pi e} \sum_{j \in \mathcal{R}} \|\Delta^{(ij)}\|_2^2 \\
&\overset{(i)}{\geq} \frac{\kappa}{2\pi e(d+1)} \sum_{j \in [p], j \neq i} \|\Delta^{(ij)}\|_2^2 \\
&\overset{(j)}{=} \frac{\kappa}{2\pi e(d+1)} \|\Delta_E\|_2^2,
\end{aligned}$$

where (a) follows from the definition of $\boldsymbol{\varphi}^{(i)}(\mathbf{x})$ from Section 2, (b) follows because we have conditioned on x_i and $x_{\mathcal{R}^c}$ (note $(x_j)_{j \in \mathcal{R}^c}$ are constant given $x_{\mathcal{R}^c}$), (c) follows because $(x_j)_{j \in \mathcal{R}}$ are conditionally independent given $x_{\mathcal{R}^c}$ (note that $\mathcal{R}$ is an independent set in $G_{-i}(\boldsymbol{\theta}^*)$, i.e. there is no edge connecting two vertices in $\mathcal{R}$), (d) follows from linearity of expectation, (e) follows because $x_{\mathcal{N}(j)} \subseteq x_{\mathcal{R}^c} \cup x_i \; \forall j \in \mathcal{R}$, (f) follows from the global Markov property, (g) follows from monotonicity of expectation and Shannon's entropy inequality ($h(\cdot) \leq \log \sqrt{2\pi e \mathbb{V}\text{ar}(\cdot)}$), (h) follows from (12), (i) follows from (38) and (j) follows from (36).

Plugging this back in (35) we have

$$\delta \mathcal{S}_n^{(i)}(\Delta, \boldsymbol{\vartheta}^{*(i)}) \geq \exp(-\gamma \varphi_{\max}) \frac{\frac{\kappa}{2\pi e(d+1)} \|\Delta_E\|_2^2 - \epsilon_2 \|\Delta\|_1^2}{2 + \varphi_{\max} \|\Delta\|_1}.$$

□

M The Generalized Interaction Screening algorithm

In this appendix, we describe the *Generalized Interaction Screening* algorithm (Algorithm **??**) for the setup in Section 2 and also provide its computational complexity (Proposition M.1). Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$ and $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$ from Section 2.

M.1 The *Generalized Interaction Screening* algorithm

Vuffray et al. (2019) showed that an ϵ-optimal solution of GRISE could be obtained by first finding an ϵ-optimal solution of the unconstrained GRISE using a variation of the Entropic Descent Algorithm and then projecting the solution onto Λ. See Lemma 4 of Vuffray et al. (2019) for more details.

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

$$\mathcal{S}_n^{(i)}(\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,\text{unc}}) \leq \min_{\boldsymbol{\vartheta}:\|\boldsymbol{\vartheta}\|_1 \leq \gamma} \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}) + \epsilon. \tag{39}$$

The iterative Algorithm 1 outputs an ϵ-optimal solution of GRISE without constraints in (39). This algorithm is an application of the Entropic Descent Algorithm introduced in Beck and Teboulle (2003) to a reformulation of (7) as a minimization over the probability simplex.

Algorithm 1 Entropic Descent for unconstrained GRISE

```
Input: $k, p, \gamma, \varphi_{\max}, \mathcal{S}_n^{(i)}(\cdot), T$
Output: $\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,\text{unc}}$
Initialization:
    $w^{(1)}_{l,+} \leftarrow e/(2k^2(p-1)+2k+1), \forall l \in [k^2(p-1)+k]$
    $w^{(1)}_{l,-} \leftarrow e/(2k^2(p-1)+2k+1), \forall l \in [k^2(p-1)+k]$
    $y^{(1)} \leftarrow e/(2k^2(p-1)+2k+1)$
    $\eta^{(1)} \leftarrow \sqrt{\log(2k^2(p-1)+2k+1)}/2\gamma\varphi_{\max}\exp(\gamma\varphi_{\max})$
for $t = 1, \cdots, T$ do
    $\mathbf{w}^{(t)}_{+} = (w^{(t)}_{l,+} : l \in [k^2(p-1)+k])$
    $\mathbf{w}^{(t)}_{-} = (w^{(t)}_{l,-} : l \in [k^2(p-1)+k])$
    $v_l = \gamma \dfrac{\partial \mathcal{S}_n^{(i)}(\gamma(\mathbf{w}^{(t)}_{+} - \mathbf{w}^{(t)}_{-}))}{\partial \boldsymbol{\vartheta}_l}, \forall l \in [k^2(p-1)+k]$
    $x_{l,+} = w^{(t)}_{l,+}\exp(-\eta^t v_l), \forall l \in [k^2(p-1)+k]$
    $x_{l,-} = w^{(t)}_{l,-}\exp(\eta^t v_l), \forall l \in [k^2(p-1)+k]$
    $z = y^{(t)} + \sum_{l \in [k^2(p-1)+k]} (x_{l,+} + x_{l,-})$
    $w^{(t+1)}_{l,+} \leftarrow x_{l,+}/z, \forall l \in [k^2(p-1)+k]$
    $w^{(t+1)}_{l,-} \leftarrow x_{l,-}/z, \forall l \in [k^2(p-1)+k]$
    $y^{(t+1)} \leftarrow y^{(t)}/z$
    $\eta^{(t+1)} \leftarrow \eta^t \sqrt{t/t+1}$
$s = \arg\min_{s=1,\ldots,T} \mathcal{S}_n^{(i)}(\gamma(\mathbf{w}^{(s)}_{+} - \mathbf{w}^{(s)}_{-}))$
$\hat{\boldsymbol{\vartheta}}^{(i)}_{\epsilon,\text{unc}} \leftarrow \gamma(\mathbf{w}^{(s)}_{+} - \mathbf{w}^{(s)}_{-})$
```

M.2 Computational Complexity of Algorithm 1

The following proposition provides guarantees on the computational complexity of unconstrained GRISE.

Proposition M.1. *Let $\epsilon > 0$ be the optimality gap. Let the number of iterations satisfy*

$$T \geq \frac{\gamma^2\varphi_{\max}^2\exp(2\gamma\varphi_{\max})}{\epsilon^2} \times \log(2k^2(p-1)+2k+1) \;=\; \Omega\left(\frac{\exp(\Theta(k^2 d))}{\epsilon^2}\log(k^2 p)\right).$$

Then, Algorithm 1 is guaranteed to produce an ϵ-optimal solution of GRISE without constraints in (39) *with number of computations of the order*

$$\frac{k^2\gamma^2\varphi_{\max}^2\exp(2\gamma\varphi_{\max})np}{\epsilon^2} \times \log(2k^2(p-1)+2k+1) \;=\; \Omega\left(\frac{\exp(\Theta(k^2 d))}{\epsilon^2} np\log(k^2 p)\right).$$

Proof of Proposition M.1. We first show that the minimization of GRISE when $\Lambda = \mathbb{R}^{k^2(p-1)+k}$ (the unconstrained case) is equivalent to the following lifted minimization,

$$\min_{\boldsymbol{\vartheta}, \mathbf{w}_+, \mathbf{w}_-, y} \quad \mathcal{S}_n^{(i)}(\boldsymbol{\vartheta}) \tag{40}$$

$$\text{s.t.} \quad \boldsymbol{\vartheta} = \gamma(\mathbf{w}_+ - \mathbf{w}_-) \tag{41}$$

$$y + \sum_{l \in [k^2(p-1)+k]} (w_{l,+} + w_{l,-}) = 1 \tag{42}$$

$$y \geq 0, w_{l,+} \geq 0, w_{l,-} \geq 0, \forall l \in [k^2(p-1)+k], \tag{43}$$

where $\mathbf{w}_+ = (w_{l,+} : l \in [k^2(p-1)+k])$ and $\mathbf{w}_- = (w_{l,-} : l \in [k^2(p-1)+k])$.

We start by showing that for all $\boldsymbol{\vartheta} \in \mathbb{R}^{k^2(p-1)+k}$ such that $\|\boldsymbol{\vartheta}\|_1 \leq \gamma$, there exists $\mathbf{w}_+, \mathbf{w}_-, y$ satisfying constraints (41), (42), (43). This is easily done by choosing $\forall l \in [k^2(p-1)+k]$, $w_{l,+} = \max(\boldsymbol{\vartheta}_l/\gamma, 0)$, $w_{l,-} = \max(-\boldsymbol{\vartheta}_l/\gamma, 0)$ and $y = 1 - \|\boldsymbol{\vartheta}\|_1/\gamma$.

Next, we trivially see that for all $\boldsymbol{\vartheta}, \mathbf{w}_+, \mathbf{w}_-, y$ satisfying constraints (41), (42), (43), it implies that $\boldsymbol{\vartheta}$ also satisfies $\|\boldsymbol{\vartheta}\|_1 \leq \gamma$. Therefore, any $\boldsymbol{\vartheta}$ that is an ϵ-minimizer of (40) is also an ϵ-minimizer of (7) without constraints. The remainder of the proof is a straightforward application of the analysis of the Entropic Descent Algorithm in Beck and Teboulle (2003) to the above minimization where $\boldsymbol{\vartheta}$ has been replaced by $\mathbf{w}_+, \mathbf{w}_-, y$ using (41). □

The computational complexity of the projection step is usually insignificant compared to the computational complexity of Algorithm 1 provided in Proposition M.1.

N Robust LASSO

In this appendix, we present a robust variation of the sparse linear regression. More specifically, we show that even in the presence of bounded additive noise, the Lasso estimator is 'prediction consistent' under almost no assumptions at all.

N.1 Setup

Suppose that $\mathsf{v}_1, \cdots, \mathsf{v}_{\tilde{p}}$ (where $\tilde{p} \geq 1$) are (possibly dependent) random variables, and suppose $\tilde{c}_1$ is a constant such that $|\mathsf{v}_r| \leq \tilde{c}_1$ almost surely for each $r \in [\tilde{p}]$. Let

$$\mathsf{y} = \sum_{r=1}^{\tilde{p}} \beta_r^* \mathsf{v}_r + \tilde{\eta} + \tilde{\epsilon},$$

where $\tilde{\eta}$ is bounded noise with $|\tilde{\eta}| \leq \tilde{\eta}_0$, $\tilde{\epsilon}$ is *sub-Gaussian* noise with mean 0 and variance proxy $\tilde{\sigma}^2$, and $\tilde{\epsilon}$ is independent of the v_r's and $\tilde{\eta}$. Define $\boldsymbol{\beta}^* := (\beta_1^*, \cdots, \beta_{\tilde{p}}^*)$. We also have the 'sparsity' condition that $\|\boldsymbol{\beta}^*\|_1 \leq \tilde{c}_2$. Here $\beta_1^*, \cdots, \beta_{\tilde{p}}^*, \tilde{c}_2$, and $\tilde{\sigma}$ are unknown constants.

N.2 Data

Let $\boldsymbol{\mathsf{v}}$ denote the random vector $(\mathsf{v}_1, \cdots, \mathsf{v}_{\tilde{p}})$. Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be n i.i.d copies of $\boldsymbol{\mathsf{v}}$ and let $\mathbf{y} := (y_1, \cdots, y_n)$ denote the corresponding true values of y. Let $\mathbf{V}$ be a $n \times \tilde{p}$ matrix such that the j^{th} row is $\mathbf{v}_j$.

Suppose that our task is to predict y given the value of $\boldsymbol{\mathsf{v}}$. If the parameter vector $\boldsymbol{\beta}^*$ was known, then the predictor of y, of interest, based on $\boldsymbol{\mathsf{v}}$ would be $\hat{\mathsf{y}} := \sum_{r=1}^{\tilde{p}} \beta_r^* \mathsf{v}_r$. However, $\boldsymbol{\beta}^*$ is unknown, and we need to estimate it from the data $(\mathbf{V}, \mathbf{y})$. Let $\tilde{\boldsymbol{\beta}}$ be the output of Algorithm 2. Let $\hat{\mathbf{y}} := (\hat{y}_1, \cdots, \hat{y}_n)$ where

$$\hat{y}_j = \boldsymbol{\beta}^{*T} \mathbf{v}_j. \tag{44}$$

Let $\tilde{\mathbf{y}} := (\tilde{y}_1, \cdots, \tilde{y}_n)$ where

$$\tilde{y}_j = \tilde{\boldsymbol{\beta}}^T \mathbf{v}_j. \tag{45}$$

The step 3 of Algorithm 2 below can be solved using Coordinate Descent.

Algorithm 2 Robust LASSO

```
Input: $\mathbf{V}, \mathbf{y}, \tilde{c}_2$
Output: $\tilde{\boldsymbol{\beta}}$
$\tilde{\boldsymbol{\beta}} \leftarrow \arg\min_{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|_1 \leq \tilde{c}_2} [\mathbf{y} - \mathbf{V} \cdot \boldsymbol{\beta}]^T [\mathbf{y} - \mathbf{V} \cdot \boldsymbol{\beta}]$
```

N.3 Prediction error

Definition N.1. *The 'mean square prediction error' of any estimator $\tilde{\boldsymbol{\beta}} := (\tilde{\beta}_1, \cdots, \tilde{\beta}_{\tilde{p}})$ is defined as the expected squared error in estimating $\hat{y}$ using $\tilde{\boldsymbol{\beta}}$, that is,*

$$MSPE(\tilde{\boldsymbol{\beta}}) := \mathbb{E}_{\mathsf{v}}(\hat{y} - \tilde{y})^2,$$

where $\tilde{y} := \sum_{r=1}^{\tilde{p}} \tilde{\beta}_r \mathsf{v}_r$.

Definition N.2. *The 'estimated mean square prediction error' of any estimator $\tilde{\boldsymbol{\beta}} := (\tilde{\beta}_1, \cdots, \tilde{\beta}_{\tilde{p}})$ is defined*

$$\widehat{MSPE}(\tilde{\boldsymbol{\beta}}) := \frac{1}{n} \sum_{j \in [n]} (\hat{y}_j - \tilde{y}_j)^2.$$

The following Lemma shows that the Lasso estimator of Algorithm 2 is 'prediction consistent' even in presence of bounded noise if $\tilde{c}_2$ is correctly chosen and $n \gg \tilde{p}$.

Lemma N.1. *Let $\tilde{\boldsymbol{\beta}}$ be the ouput of Algorithm 2. Then,*

$$\mathbb{E}[\widehat{MSPE}(\tilde{\boldsymbol{\beta}})] \leq 4\tilde{\eta}_0^2 + 4\tilde{c}_1\tilde{c}_2\tilde{\sigma}\sqrt{\frac{2\log 2\tilde{p}}{n}},$$

$$MSPE(\tilde{\boldsymbol{\beta}}) \leq 4\tilde{\eta}_0^2 + 4\tilde{c}_1\tilde{c}_2\tilde{\sigma}\sqrt{\frac{2\log 2\tilde{p}}{n}} + 8\tilde{c}_1^2\tilde{c}_2^2\sqrt{\frac{2\log(2\tilde{p}^2)}{n}}.$$

Proof of Lemma N.1. $\hat{\mathbf{y}}$ is the vector of the best predictions of $\mathbf{y}$ based on $\mathbf{V}$ and $\tilde{\mathbf{y}}$ is the vector of predictions of $\mathbf{y}$ using Algorithm 2. Let $\mathbf{v}^{(j)}$ denote the j^{th} column of $\mathbf{V}$ $\forall j \in [\tilde{p}]$. Let $v_{h,r}$ denote the h^{th} element of $\mathbf{v}^{(r)}$ $\forall h \in [n], r \in [\tilde{p}]$. Let $\tilde{\eta}_h$ $(\tilde{\epsilon}_h)$ denote the bounded (sub-Gaussian) noise associated with y_h $\forall h \in [n]$.

Define the set

$$\mathcal{Y} := \big\{ \sum_{r=1}^{\tilde{p}} \tilde{\beta}_r \mathbf{v}^{(r)} : \sum_{r=1}^{\tilde{p}} |\tilde{\beta}_r| \leq \tilde{c}_2 \big\}.$$

Note that $\mathcal{Y}$ is a compact and convex subset of $\mathbb{R}^n$. By definition, $\tilde{\mathbf{y}}$ is the projection of $\mathbf{y}$ on to the set $\mathcal{Y}$. Because $\mathcal{Y}$ is convex and $\hat{\mathbf{y}} \in \mathcal{Y}$, we have from the Pythagorean theorem for projection onto a convex set,

$$(\hat{\mathbf{y}} - \tilde{\mathbf{y}})^T(\mathbf{y} - \tilde{\mathbf{y}}) \leq 0.$$

Adding and subtracting $\hat{\mathbf{y}}$ in the second term we get,

$$\begin{aligned}
\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 &\leq (\mathbf{y} - \hat{\mathbf{y}})^T(\tilde{\mathbf{y}} - \hat{\mathbf{y}}) \\
&\overset{(a)}{=} \sum_{h=1}^{n} \big(\tilde{\eta}_h + \tilde{\epsilon}_h\big)\big(\sum_{r=1}^{\tilde{p}} (\tilde{\beta}_r - \beta_r^*) v_{h,r}\big) \\
&= \sum_{h=1}^{n} \tilde{\eta}_h \big(\sum_{r=1}^{\tilde{p}} (\tilde{\beta}_r - \beta_r^*) v_{h,r}\big) + \sum_{h=1}^{n} \tilde{\epsilon}_h \big(\sum_{r=1}^{\tilde{p}} (\tilde{\beta}_r - \beta_r^*) v_{h,r}\big) \\
&= \sum_{h=1}^{n} \tilde{\eta}_h \big(\sum_{r=1}^{\tilde{p}} (\tilde{\beta}_r - \beta_r^*) v_{h,r}\big) + \sum_{r=1}^{\tilde{p}} (\tilde{\beta}_r - \beta_r^*) \big(\sum_{h=1}^{n} \tilde{\epsilon}_h v_{h,r}, \big) \qquad (46)
\end{aligned}$$

where (a) follows from (44), (45) and definitions of y and $\hat{y}$.

Let us first focus on only the first term in (46).

$$\begin{aligned}\sum_{h=1}^{n}\tilde{\eta}_h\big(\sum_{r=1}^{\tilde{p}}(\tilde{\beta}_r-\beta^*_r)v_{h,r}\big) &\overset{(a)}{\leq}\sum_{h=1}^{n}|\tilde{\eta}_h|\big|\big(\sum_{r=1}^{\tilde{p}}(\tilde{\beta}_r-\beta^*_r)v_{h,r}\big)\big| \\ &\overset{(b)}{\leq}\tilde{\eta}_0\sum_{h=1}^{n}\Big|\sum_{r=1}^{\tilde{p}}(\tilde{\beta}_r-\beta^*_r)v_{h,r},\Big| \qquad (47)\end{aligned}$$

where (a) follows from the triangle inequality and (b) follows because $|\tilde{\eta}_h|\leq\tilde{\eta}_0\ \forall h\in[n]$. Notice that $\sum_{h=1}^{n}\big|\sum_{r=1}^{\tilde{p}}(\tilde{\beta}_r-\beta^*_r)v_{h,r}\big|$ is the ℓ_1 norm of the vector $\hat{\mathbf{y}}-\tilde{\mathbf{y}}$.

Let us now focus on the second term in (46). Using the facts that $\|\boldsymbol{\beta}^*\|_1\leq\tilde{c}_2$ and $\|\tilde{\boldsymbol{\beta}}\|_1\leq\tilde{c}_2$, we have

$$\sum_{r=1}^{\tilde{p}}(\tilde{\beta}_r-\beta^*_r)\big(\sum_{h=1}^{n}\tilde{\epsilon}_h v_{h,r}\big)\leq 2\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r|, \qquad (48)$$

where

$$u_r:=\sum_{h=1}^{n}\tilde{\epsilon}_h v_{h,r}.$$

Now plugging back the upper bounds from (47) and (48) in (46) we get,

$$\begin{aligned}\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2^2&\leq\tilde{\eta}_0\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_1+2\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r| \\ &\overset{(a)}{\leq}\tilde{\eta}_0\sqrt{n}\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2+2\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r| \\ &\overset{(b)}{\leq}2\max\big\{\tilde{\eta}_0\sqrt{n}\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2,2\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r|\big\}\end{aligned}$$

where (a) follows from the fact that $\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_1\leq\sqrt{n}\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2$ and (b) follows from the fact $a+b\leq 2\max\{a,b\}$ for any $a,b\geq 0$. Looking at the two cases separately, we have

$$\begin{aligned}&\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2^2\leq 2\tilde{\eta}_0\sqrt{n}\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2, &\qquad&\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2^2\leq 4\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r| \\ &\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2\leq 2\tilde{\eta}_0\sqrt{n}, &&\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2^2\leq 4\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r|\end{aligned}$$

Combining the two cases, we have

$$\begin{aligned}\|\hat{\mathbf{y}}-\tilde{\mathbf{y}}\|_2^2&\leq\max\{4\tilde{\eta}_0^2 n,4\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r|\} \\ &\overset{(a)}{\leq}4\tilde{\eta}_0^2 n+4\tilde{c}_2\max_{1\leq r\leq\tilde{p}}|u_r|, \qquad (49)\end{aligned}$$

where (a) follows from the fact $\max\{a,b\}\leq a+b$ for any $a,b\geq 0$.

Let $\mathcal{F}$ be the sigma-algebra generated by $(v_{h,r})_{1\leq h\leq n,1\leq r\leq\tilde{p}}$. Let $\mathbb{E}^{\mathcal{F}}$ denote the conditional expectation given $\mathcal{F}$. Conditional on $\mathcal{F}$, u_r is *sub-Gaussian* with variance proxy $\tilde{\sigma}^2\Big(\sum_{h=1}^{n}v_{h,r}^2\Big)$. Since $v_{h,r}\leq\tilde{c}_1$ almost surely for all h,r, it follows from the *maximal inequality* of the *sub-Gaussian* random variables (see Lemma 4 in Chatterjee (2013)) that

$$\mathbb{E}^{\mathcal{F}}\big(\max_{1\leq r\leq\tilde{p}}|u_r|\big)\leq\tilde{c}_1\tilde{\sigma}\sqrt{2n\log(2\tilde{p})}.$$

Since the right-hand-side is non-random, taking expectation on both sides with respect to $\mathcal{F}$ result in,

$$\mathbb{E}\big(\max_{1\leq r\leq\tilde{p}}|u_r|\big)\leq\tilde{c}_1\tilde{\sigma}\sqrt{2n\log(2\tilde{p})}. \qquad (50)$$

Taking expectation on both sides in (49) and using (50), we get

$$\mathbb{E}(\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2) \leq 4\tilde{\eta}_0^2 n + 4\tilde{c}_1\tilde{c}_2\tilde{\sigma}\sqrt{2n\log(2\tilde{p})}. \tag{51}$$

Dividing both sides by n results in

$$\mathbb{E}[\widehat{\text{MSPE}}(\tilde{\boldsymbol{\beta}})] \leq 4\tilde{\eta}_0^2 + 4\tilde{c}_1\tilde{c}_2\tilde{\sigma}\sqrt{\frac{2\log 2\tilde{p}}{n}}.$$

Recall that $\tilde{\boldsymbol{\beta}}$ is computed using the data $\mathbf{V}$ and $\mathbf{y}$, and is therefore independent of v and y. Using definitions of $\tilde{y}$ and $\hat{y}$, we have

$$\mathbb{E}^{\mathcal{F}}(\hat{y} - \tilde{y})^2 = \sum_{r,s=1}^{p} (\beta_r^* - \tilde{\beta}_r)(\beta_s^* - \tilde{\beta}_s)\mathbb{E}(\mathsf{v}_r\mathsf{v}_s).$$

We also have

$$\frac{1}{n}\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 = \frac{1}{n}\sum_{h=1}^{n}\sum_{r,s=1}^{p} (\beta_r^* - \tilde{\beta}_r)(\beta_s^* - \tilde{\beta}_s)v_{h,r}\mathsf{v}_{h,s}.$$

Therefore by defining

$$\mathsf{u}_{r,s} = \mathbb{E}(\mathsf{v}_r\mathsf{v}_s) - \frac{1}{n}\sum_{h=1}^{n} v_{h,r}\mathsf{v}_{h,s},$$

we have

$$\mathbb{E}^{\mathcal{F}}(\hat{y} - \tilde{y})^2 - \frac{1}{n}\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 = \sum_{r,s=1}^{p} (\beta_r^* - \tilde{\beta}_r)(\beta_s^* - \tilde{\beta}_s)\mathsf{u}_{r,s} \overset{(a)}{\leq} 4\tilde{c}_2^2 \max_{1\leq r,s\leq \tilde{p}} |\mathsf{u}_{r,s}|, \tag{52}$$

where (a) follows from the facts that $\|\boldsymbol{\beta}^*\|_1 \leq \tilde{c}_2$ and $\|\tilde{\boldsymbol{\beta}}\|_1 \leq \tilde{c}_2$. Recall that $|\mathsf{v}_r| \leq \tilde{c}_1\ \forall r \in [\tilde{p}]$. Using the triangle inequality, we have $\mathbb{E}(\mathsf{v}_r\mathsf{v}_s) - v_{h,r}\mathsf{v}_{h,s} \leq 2\tilde{c}_1^2$ for all h, r, and s. It follows by Hoeffding's inequality (see Lemma 5 in Chatterjee (2013)) that for any $\varsigma \in \mathbb{R}$,

$$\mathbb{E}(e^{\varsigma \mathsf{u}_{r,s}}) \leq e^{2\varsigma^2\tilde{c}_1^4/n}.$$

Again by the *maximal inequality* of the *sub-Gaussian* random variables (see Lemma 4 in Chatterjee (2013)) we have,

$$\mathbb{E}(\max_{1\leq r,s\leq \tilde{p}} |\mathsf{u}_{r,s}|) \leq 2\tilde{c}_1^2\sqrt{\frac{2\log(2\tilde{p}^2)}{n}}. \tag{53}$$

Taking expectation on both sides in (52) and plugging in (51) and (53), we get

$$\mathbb{E}(\hat{y} - \tilde{y})^2 \leq 4\tilde{\eta}_0^2 + 4\tilde{c}_1\tilde{c}_2\tilde{\sigma}\sqrt{\frac{2\log 2\tilde{p}}{n}} + 8\tilde{c}_1^2\tilde{c}_2^2\sqrt{\frac{2\log(2\tilde{p}^2)}{n}},$$

and this completes the proof. □

O Supporting propositions for Lemma E.2

In this appendix, we will state the key propositions required in the proof of Lemma E.2. Proposition O.1 provides guarantees for learning the conditional mean parameter vector and Proposition O.2 provides guarantees for learning the conditional canonical parameter vector. The proof of Lemma E.2 is given in Appendix P. Recall from Section 3 that $\boldsymbol{\lambda}^*(x_{-i})$ denotes the conditional canonical parameter vector and $\boldsymbol{\mu}^*(x_{-i})$ denotes the conditional mean parameter vector of the conditional density $f_{\mathsf{x}_i}(\cdot|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)})$. Recall the definition of q^s from Section 2.

O.1 Learning conditional mean parameter vector

The first step of the algorithm for recovering node parameters from Section 3 provides an estimate of the conditional mean parameter vector. The following proposition shows that, with enough samples and an estimate of the graph structure, we can learn the conditional mean parameter vector such that the ℓ_∞ error is small with high probability.

Proposition O.1. *Suppose we have an estimate $\hat{G}$ of $G(\boldsymbol{\theta}^*)$ such that for any $\delta_4 \in (0,1)$, $\hat{G} = G(\boldsymbol{\theta}^*)$ with probability at least $1-\delta_4$. Given n independent samples $\mathbf{x}^{(1)} \cdots, \mathbf{x}^{(n)}$ of $\mathbf{x}$, consider $x_{-i}^{(z)}$ where z is chosen randomly from $\{1, \cdots, n\}$. There exists an alogrithm that produces an estimate $\hat{\boldsymbol{\mu}}(x_{-i}^{(z)})$ of $\boldsymbol{\mu}^*(x_{-i}^{(z)})$ such that for any $\epsilon_4 \in (0,1)$,*

$$\|\boldsymbol{\mu}^*(x_{-i}^{(z)}) - \hat{\boldsymbol{\mu}}(x_{-i}^{(z)})\|_\infty \leq \epsilon_4 \qquad \forall i \in [p],$$

with probability at least $1 - \delta_4 - k\epsilon_4^2/4$ as long as

$$n \geq \left(\frac{2^{9d+17} b_u^{2d} k^{4d} d^{2d+1} \theta_{\max}^{2d} \phi_{\max}^{4d+4} \bar{\phi}_{\max}^{2d}}{\epsilon_4^{4d+8}} \right) \log\left(\frac{2^{5.5} b_u k^2 d \theta_{\max} \phi_{\max}^2 \bar{\phi}_{\max}}{\epsilon_4^2} \right).$$

The number of computations required scale as

$$\frac{2^{18d+17} b_u^{4d} k^{8d+1} d^{4d+1} \theta_{\max}^{4d} \phi_{\max}^{8d+4} \bar{\phi}_{\max}^{4d}}{\epsilon_4^{8d+8}} \times p.$$

The proof of proposition O.1 is given in Appendix Q.

O.2 Learning canonical parameter vector

The second step of the algorithm for recovering node parameters from Section 3 is to obtain an estimate of the canonical parameter vector given an estimate of the mean parameter vector. We exploit the conjugate duality between the canonical and mean parameters and run a projected gradient descent algorithm for this purpose.

We will describe the algorithm using a generic setup in this section and then apply it to the current setting in the proof of Lemma E.2 in Appendix P.

O.2.1 Setup for the projected gradient descent algorithm

Let $\mathcal{X}_0$ be a real interval such that its length is upper (lower) bounded by b_u (b_l). Suppose that w is a random variable that takes value in $\mathcal{X}_0$ with probability density function as follows,

$$f_{\mathsf{w}}(w; \boldsymbol{\rho}^*) \propto \exp(\boldsymbol{\rho}^{*T} \boldsymbol{\phi}(w)), \tag{54}$$

where the parameter vector $\boldsymbol{\rho}^* := (\rho_1^*, \cdots, \rho_k^*)$ is unknown and is such that $\|\boldsymbol{\rho}^*\|_\infty \leq \rho_{\max}$. Let $\mathcal{P} := \{\boldsymbol{\rho} \in \mathbb{R}^k : \|\boldsymbol{\rho}\|_\infty \leq \rho_{\max}\}$. Let $\boldsymbol{\upsilon}^* := (\upsilon_1^*, \cdots, \upsilon_k^*)$ denote the mean parameter vector of $f_{\mathsf{w}}(w; \boldsymbol{\rho}^*)$ and let $\hat{\boldsymbol{\upsilon}}$ be an estimate of $\boldsymbol{\upsilon}^*$ such that, we have $\|\boldsymbol{\upsilon}^* - \hat{\boldsymbol{\upsilon}}\|_\infty \leq \epsilon_5$ with probability at least $1 - \delta_5$ for any $\epsilon_5 > 0$, and any $\delta_5 \in (0,1)$. The goal is to estimate the parameter vector $\boldsymbol{\rho}^*$ using the projected gradient descent algorithm (Boyd and Vandenberghe, 2014; Bubeck, 2015).

O.2.2 The projected gradient descent algorithm

Let $\mathcal{U}_{\mathcal{X}_0}$ denote the uniform distribution on $\mathcal{X}_0$. Algorithm 3 is a subroutine that is used in the projected gradient descent algorithm. This subroutine is a Markov chain and it provides an estimate of the mean parameters of an exponential family distribution of the form (54) when the underlying canonical parameters are known. See Appendix R for discussion on the theoretical properties of this subroutine.

Algorithm 3 Metropolized random walk (MRW)

```
Input: $\boldsymbol{\rho}, \mathcal{X}_0, \tau_1, \tau_2, w_{(0)}$
Output: $\hat{\boldsymbol{\nu}}(\boldsymbol{\rho})$
for $m = 1, \cdots, \tau_2$ do
    for $r = 0, \cdots, \tau_1$ do
        Proposal step: Draw $z_{(r+1)} \sim \mathcal{U}_{\mathcal{X}_0}$
        Accept-reject step:
            Compute $\alpha_{(r+1)} \leftarrow \min \left\{ 1, \frac{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(z_{(r+1)}))}{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w_{(r)}))} \right\}$
            With probability $\alpha_{(r+1)}$ accept the proposal: $w_{(r+1)} \leftarrow z_{(r+1)}$
            With probability $1 - \alpha_{(r+1)}$ reject the proposal: $w_{(r+1)} \leftarrow w_{(r)}$
    $\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}) \leftarrow \hat{\boldsymbol{\nu}}(\boldsymbol{\rho}) + \boldsymbol{\phi}(w_{(\tau_1+1)})$
$\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}) \leftarrow \frac{1}{\tau_2} \hat{\boldsymbol{\nu}}(\boldsymbol{\rho})$
```

Algorithm 4 Projected Gradient Descent

```
Input: $\xi, \mathcal{X}_0, \tau_1, \tau_2, \tau_3, w_{(0)}, \boldsymbol{\rho}^{(0)}, \hat{\boldsymbol{\upsilon}}$
Output: $\hat{\boldsymbol{\rho}}$
for $r = 0, \cdots, \tau_3$ do
    $\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}^{(r)}) \leftarrow MRW(\boldsymbol{\rho}^{(r)}, k, \mathcal{X}_0, \tau_1, \tau_2, w_{(0)})$
    $\boldsymbol{\rho}^{(r+1)} \leftarrow \arg\min_{\boldsymbol{\rho} \in \mathcal{P}} \|\boldsymbol{\rho}^{(r)} - \xi[\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}^{(r)}) - \hat{\boldsymbol{\upsilon}}] - \boldsymbol{\rho}\|$
$\hat{\boldsymbol{\rho}} \leftarrow \boldsymbol{\rho}^{(\tau_3+1)}$
```

O.2.3 Guarantees on the output of the projected gradient descent algorithm

The following Proposition shows that running sufficient iterations of the projected gradient descent (Algorithm 4) results in an estimate, $\hat{\boldsymbol{\rho}}$, of the parameter vector, $\boldsymbol{\rho}^*$, such that the ℓ_2 error is small with high probability.

Define $\bar{c}_1 := q^s, \bar{c}_2 := 2k\phi_{\max}^2$, $\bar{c}_3 := \frac{4k\epsilon_5(\epsilon_5 + 2\bar{c}_2\rho_{\max} + 2\phi_{\max})}{\bar{c}_1\bar{c}_2}$.

Proposition O.2. *Let $\epsilon_6 > 0$. Let $\hat{\boldsymbol{\rho}}$ denote the output of Algorithm 4 with $\xi = 1/\bar{c}_2$, $\tau_1 = 8b_l^{-2} \exp(12k\rho_{\max}\phi_{\max}) \left[\log \frac{4\phi_{\max}\sqrt{b_u}}{\epsilon_9\sqrt{b_l}} + k\rho_{\max}\phi_{\max} \right]$, $\tau_2 = \frac{8\phi_{\max}^2}{\epsilon_5^2} \log\left(\frac{2k\tau_3}{\delta_5}\right)$, $\tau_3 = \frac{\bar{c}_2}{\bar{c}_1} \log\left(\frac{k\rho_{\max}^2}{\epsilon_6^2 - \bar{c}_3}\right)$, $w_{(0)} = 0$, $\boldsymbol{\rho}^{(0)} = (0, \cdots, 0)$ and $\hat{\boldsymbol{\upsilon}} = (\hat{\upsilon}_1, \cdots, \hat{\upsilon}_k)$. Then,*

$$\|\boldsymbol{\rho}^* - \hat{\boldsymbol{\rho}}\|_2 \leq \epsilon_6,$$

with probability at least $1 - 2\delta_5$.

The proof of proposition O.2 is given in Appendix S.

P Proof of Lemma E.2

In this appendix, we prove Lemma E.2. See Appendix O.1 and Appendix O.2 for two key propositions required in the proof. Recall from Section 3 that $\boldsymbol{\lambda}^*(x_{-i})$ denotes the conditional canonical parameter vector and $\boldsymbol{\mu}^*(x_{-i})$ denotes the conditional mean parameter vector of the conditional density $f_{\mathsf{x}_i}(\cdot|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)})$. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$, $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$, $c_1(\alpha)$, and $c_2(\alpha)$ from Section 2 and the definition of $c_3(\alpha)$ from Appendix E. We restate the Lemma below and then provide the proof.

Lemma E.2. *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 $\hat{\boldsymbol{\theta}}^{(i)}$ be an estimate of $\boldsymbol{\theta}^{*(i)}$ obtained using the three-step procedure from Section 3. Then, for any $\alpha_2 \in (0, 1)$,*

$$\|\boldsymbol{\theta}^{*(i)} - \hat{\boldsymbol{\theta}}^{(i)}\|_\infty \leq \alpha_2, \qquad \forall i \in [p]$$

with probability at least $1 - \alpha_2^4$ as long as

$$n \geq \max\left[c_1\left(\min\left\{ \frac{\theta_{\min_+}}{3}, \frac{\alpha_2}{2dk\phi_{\max}} \right\} \right) \log\left(\frac{2^{5/2}pk}{\alpha_2^2} \right), c_2(\alpha_2) \right]$$

$$= \Omega\Bigg(\frac{\exp(\Theta\Big(k^2 d + d\log\big(\frac{dk}{\alpha_2 q^s}\big)\Big))}{\kappa^2\alpha_2^4} \times \log\Big(\frac{pk}{\alpha_2^2}\Big)\Bigg).$$

The number of computations required scale as

$$c_3\Big(\min\Big\{\frac{\theta_{\min_+}}{3}, \frac{\alpha_2}{2dk\phi_{\max}}\Big\}\Big) \times \log\Big(\frac{2^{5/2}pk}{\alpha_2^2}\Big) \times \log(2k^2 p) \times p^2 \;=\; \Omega\Bigg(\frac{\exp(\Theta(k^2 d))}{\kappa^4\alpha_2^8}\log^2\Big(\frac{pk}{\alpha_2^2}\Big)p^2\Bigg).$$

Proof of Lemma E.2. Let the number of samples satisfy

$$n \geq \max\Big[c_1\Big(\min\Big\{\frac{\theta_{\min_+}}{3}, \frac{\alpha_2}{2dk\phi_{\max}}\Big\}\Big)\log\Big(\frac{2^{\frac{5}{2}}pk}{\alpha_2^2}\Big), c_2(\alpha_2)\Big].$$

Using Theorem 4.3 with $\delta = \alpha_2^4/8$, we know the true neighborhood $\mathcal{N}(i)$ $\forall i \in [p]$, with probability at least $1 - \alpha_2^4/8$. Throughout the remainder of the proof, we will condition on the event that we know the true neighborhood for every node.

Let us define

$$\epsilon_4 = \frac{\alpha_2^2 q^s}{2^7 k^2 d\theta_{\max}\phi_{\max}}, \epsilon_6 = \frac{\alpha_2}{2}, \bar{c}_3 = \frac{\alpha_2^2}{8}.$$

Consider $x_{-i}^{(z)}$ where z is chosen uniformly at random from $[n]$. Using Proposition O.1 with $\delta_4 = \alpha_2^4/8$, the estimate $\hat{\boldsymbol{\mu}}(x_{-i}^{(z)})$ is such that

$$\|\boldsymbol{\mu}^*(x_{-i}^{(z)}) - \hat{\boldsymbol{\mu}}(x_{-i}^{(z)})\|_\infty \leq \epsilon_4 \qquad \forall i \in [p],$$

with probability at least $1 - \alpha_2^4/8 - k\epsilon_4^2/4$. This puts us in a position to use Proposition O.2.

Observe that $\boldsymbol{\lambda}^*(x_{-i})$ is such that $\|\boldsymbol{\lambda}^*(x_{-i})\|_\infty \leq \rho_{\max} = 2kd\theta_{\max}\phi_{\max}$ $\forall x_{-i} \in \Pi_{j\in[p]\setminus\{i\}}\mathcal{X}_j$. Using Proposition O.2 with $\hat{\boldsymbol{\upsilon}} = \hat{\boldsymbol{\mu}}(x_{-i}^{(z)})$, $\epsilon_5 = \epsilon_4$ and $\delta_5 = \alpha_2^4/8 + k\epsilon_4^2/4$, the estimate $\hat{\boldsymbol{\lambda}}(x_{-i}^{(z)})$ is such that

$$\begin{aligned} &\|\boldsymbol{\lambda}^*(x_{-i}^{(z)}) - \hat{\boldsymbol{\lambda}}(x_{-i}^{(z)})\|_2 \leq \epsilon_6 \qquad \forall i \in [p] \\ \implies &\|\boldsymbol{\lambda}^*(x_{-i}^{(z)}) - \hat{\boldsymbol{\lambda}}(x_{-i}^{(z)})\|_\infty \leq \epsilon_6 \qquad \forall i \in [p], \end{aligned} \tag{55}$$

with probability at least $1-\alpha_2^4/4-k\epsilon_4^2/2$. Plugging in the value of ϵ_4 and assuming that $(q^s)^2 \leq 2^{13}k^4d^2\theta_{\max}^2\phi_{\max}^2$, it is easy to see that $k\epsilon_4^2/2 \leq \alpha_2^4/4$.

Let $\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)} \in \Lambda$ be an ϵ-optimal solution of GRISE and let $\hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}$ be the component of $\hat{\boldsymbol{\vartheta}}_\epsilon^{(i)}$ associated with the edge potentials. Using Lemma E.1 with $\alpha_1 = \alpha_2/2dk\phi_{\max}$ and $\delta = \alpha_2^4/4$, we have

$$\begin{aligned} &\|\boldsymbol{\vartheta}_E^{*(i)} - \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}\|_2 \leq \frac{\alpha_2}{2dk\phi_{\max}}, \qquad \forall i \in [p] \\ \implies &\|\boldsymbol{\vartheta}_E^{*(i)} - \hat{\boldsymbol{\vartheta}}_{\epsilon,E}^{(i)}\|_\infty \leq \frac{\alpha_2}{2dk\phi_{\max}}, \qquad \forall i \in [p], \end{aligned} \tag{56}$$

with probability at least $1 - \alpha_2^4/4$.

For any $r \in [k]$ and $i \in [p]$, let $\lambda_r^*(x_{-i}^{(z)})$ denote the r^{th} element of $\boldsymbol{\lambda}^*(x_{-i}^{(z)})$. We have, for any $r \in [k]$ and $i \in [p]$,

$$\begin{aligned} \theta_r^{*(i)} &= \lambda_r^*(x_{-i}^{(z)}) - \sum_{j\neq i}\sum_{s\in[k]}\theta_{r,s}^{*(ij)}\phi_s(x_j^{(z)}) \\ &\overset{(a)}{=} \lambda_r^*(x_{-i}^{(z)}) - \sum_{j\in\mathcal{N}(i)}\sum_{s\in[k]}\theta_{r,s}^{*(ij)}\phi_s(x_j^{(z)}), \end{aligned} \tag{57}$$

where (a) follows because $\forall r, s \in [k], j \notin \mathcal{N}(i)$, $\theta^{*(ij)}_{r,s} = 0$. Let $\hat{\lambda}_r(x^{(z)}_{-i})$ denote the r^{th} element of $\hat{\boldsymbol{\lambda}}(x^{(z)}_{-i})$. Define the estimate, $\hat{\theta}^{(i)}_r$, as follows:

$$\hat{\theta}^{(i)}_r \coloneqq \hat{\lambda}_r(x^{(z)}_{-i}) - \sum_{j \in \mathcal{N}(i)} \sum_{s \in [k]} \hat{\theta}^{(ij)}_{r,s} \phi_s(x^{(z)}_j). \tag{58}$$

Combining (57) and (58), the following holds $\forall i \in [p], \forall r \in [k]$ with probability at least $1 - \alpha_2^4$:

$$\begin{aligned}\Big|\theta^{*(i)}_r - \hat{\theta}^{(i)}_r\Big| &= \Big|\lambda^*_r(x^{(z)}_{-i}) - \hat{\lambda}_r(x^{(z)}_{-i}) - \sum_{j \in \mathcal{N}(i)} \sum_{s \in [k]} \Big(\theta^{*(ij)}_{r,s} - \hat{\theta}^{(ij)}_{r,s}\Big) \phi_s(x^{(z)}_j)\Big| \\ &\overset{(a)}{\leq} \Big|\lambda^*_r(x^{(z)}_{-i}) - \hat{\lambda}_r(x^{(z)}_{-i})\Big| + \sum_{j \in \mathcal{N}(i)} \sum_{s \in [k]} \Big|\theta^{*(ij)}_{r,s} - \hat{\theta}^{(ij)}_{r,s}\Big| \phi_s(x^{(z)}_j) \\ &\overset{(b)}{\leq} \epsilon_6 + \frac{\alpha_2}{2dk\phi_{\max}} \sum_{j \in \mathcal{N}(i)} \sum_{s \in [k]} \phi_s(x^{(z)}_j) \\ &\overset{(c)}{\leq} \frac{\alpha_2}{2} + \frac{\alpha_2}{2} = \alpha_2,\end{aligned}$$

where (a) follows from the triangle inequality, (b) follows from (55) and (56), and (c) follows because $\|\boldsymbol{\phi}(x_j)\|_\infty \leq \phi_{\max}$ for any $x_j \in \Pi_{j \in [p]} \mathcal{X}_j$, $|\mathcal{N}(i)| \leq d$ and $\epsilon_6 = \alpha_2/2$.

The key computational steps are estimating $\hat{\boldsymbol{\vartheta}}^{(i)}_\epsilon$ and $\mathcal{N}(i)$ for every node. Using Lemma E.1 with $\alpha_1 = \alpha_2/2dk\phi_{\max}$, $\delta = \alpha_2^4/4$ and Theorem 4.3 with $\delta = \alpha_2^4/4$, the computational complexity scales as

$$c_3\Big(\min\Big\{\frac{\theta_{\min_+}}{3}, \frac{\alpha_2}{2dk\phi_{\max}}\Big\}\Big) \times \log\bigg(\frac{2^{\frac{5}{2}}pk}{\alpha_2^2}\bigg) \times \log(2k^2p) \times p^2.$$

□

Q Proof of Proposition O.1

In this appendix, we prove Proposition O.1. We begin by showing Lipschitzness of the conditional mean parameters (Lemma Q.1) and then express the problem of learning the conditional mean parameters as a sparse linear regression (Lemma Q.2). This will put us in a position to prove Proposition O.1. Recall from Section 3 that $\boldsymbol{\lambda}^*(x_{-i})$ denotes the conditional canonical parameter vector and $\boldsymbol{\mu}^*(x_{-i})$ denotes the conditional mean parameter vector of the conditional density $f_{\mathsf{x}_i}(\cdot|\mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)})$ in (9). For any $j \in [k]$, the j^{th} element of the conditional mean parameter vector is given by

$$\mu^*_j(x_{-i}) = \frac{\displaystyle\int_{x_i \in \mathcal{X}_i} \phi_j(x_i) \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i})\boldsymbol{\phi}(x_i)\Big) dx_i}{\displaystyle\int_{x_i \in \mathcal{X}_i} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i})\boldsymbol{\phi}(x_i)\Big) dx_i}. \tag{59}$$

Define $L_1 \coloneqq 2k^2\theta_{\max}\phi^2_{\max}\bar{\phi}_{\max}$.

Q.1 Lipschitzness of conditional mean parameters

The following Lemma shows that $\forall i \in [p]$, the conditional mean parameters associated with the conditional density of node x_i given the values taken by all the other nodes ($\mathsf{x}_{-i} = x_{-i}$) are Lipschitz functions of x_m $\forall m \in [p] \setminus \{i\}$.

Lemma Q.1. *For any $i \in [p]$, $j \in [k]$, $m \in [p] \setminus \{i\}$ and $x_{-i} \in \Pi_{j \in [p]\setminus\{i\}} \mathcal{X}_j$, $\mu^*_j(x_{-i})$ is a L_1 Lipschitz function of x_m.*

Proof of Lemma Q.1. Fix any $i \in [p]$ and $j \in [k]$. Consider any $m \in [p] \setminus \{i\}$. Differentiating both sides of (59) with respect to x_m and applying the quotient rule gives us,

$$\frac{\partial \mu_j^*(x_{-i})}{\partial x_m} = \frac{\frac{\partial}{\partial x_m} \int_{x_i \in \mathcal{X}_i} \phi_j(x_i) \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i}{\int_{x_i \in \mathcal{X}_i} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i} - \frac{\Big(\int_{x_i \in \mathcal{X}_i} \phi_j(x_i) \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i\Big)\Big(\frac{\partial}{\partial x_m} \int_{x_i \in \mathcal{X}_i} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i\Big)}{\Big(\int_{x_i \in \mathcal{X}_i} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i\Big)^2}.$$

Observe that $\phi_j(x_i) \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big)$ and $\frac{\partial}{\partial x_m} \phi_j(x_i) \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big)$ are analytic functions, and therefore are also continuous functions of x_i and x_m. We can apply the Leibniz integral rule to interchange the integral and partial differential operators. This results in

$$\begin{aligned}\frac{\partial \mu_j^*(x_{-i})}{\partial x_m} &= \frac{\int_{x_i \in \mathcal{X}_i} \phi_j(x_i) \frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)}{\partial x_m} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i}{\int_{x_i \in \mathcal{X}_i} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i} \\ &\quad - \frac{\Big(\int_{x_i \in \mathcal{X}_i} \phi_j(x_i) \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i\Big)\Big(\int_{x_i \in \mathcal{X}_i} \frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)}{\partial x_m} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i\Big)}{\Big(\int_{x_i \in \mathcal{X}_i} \exp\Big(\boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)\Big) dx_i\Big)^2} \\ &= \mathbb{E}\bigg(\phi_j(\mathsf{x}_i) \times \frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m} \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg) \\ &\quad - \mathbb{E}\bigg(\phi_j(\mathsf{x}_i) \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg) \times \mathbb{E}\bigg(\frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m} \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg).\end{aligned}$$

Using the triangle inequality we have,

$$\begin{aligned}\bigg|\frac{\partial \mu_j^*(x_{-i})}{\partial x_m}\bigg| &\leq \bigg|\mathbb{E}\bigg(\phi_j(\mathsf{x}_i) \times \frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m} \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg)\bigg| \\ &\quad + \bigg|\mathbb{E}\bigg(\phi_j(\mathsf{x}_i) \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg) \times \mathbb{E}\bigg(\frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m} \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg)\bigg| \\ &\overset{(a)}{=} \bigg|\mathbb{E}\bigg(\phi_j(\mathsf{x}_i) \times \frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m} \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg)\bigg| \\ &\quad + \bigg|\mathbb{E}\bigg(\phi_j(\mathsf{x}_i) \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg)\bigg| \times \bigg|\mathbb{E}\bigg(\frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m} \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg)\bigg| \\ &\overset{(b)}{\leq} \mathbb{E}\bigg(\Big|\phi_j(\mathsf{x}_i)\Big| \times \bigg|\frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m}\bigg| \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg) \\ &\quad + \mathbb{E}\bigg(\Big|\phi_j(\mathsf{x}_i)\Big| \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg) \times \mathbb{E}\bigg(\bigg|\frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(\mathsf{x}_i)}{\partial x_m}\bigg| \bigg| \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}\bigg),\end{aligned}$$

where (a) follows because for any a, b, we have $|ab| = |a||b|$ and (b) follows because the absolute value of an integral is smaller than or equal to the integral of an absolute value.

We will now upper bound $\frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)}{\partial x_m}$ as follows:

$$\frac{\partial \boldsymbol{\lambda}^{*T}(x_{-i}) \boldsymbol{\phi}(x_i)}{\partial x_m} \overset{(a)}{=} \frac{\partial \Big[\sum_{r \in [k]} \big(\theta_r^{*(i)} + \sum_{j \neq i} \sum_{s \in [k]} \theta_{r,s}^{*(ij)} \phi_s(x_j)\big) \phi_r(x_i)\Big]}{\partial x_m}$$

$$= \sum_{r\in[k]} \bigg(\sum_{s\in[k]} \theta^{*(im)}_{r,s} \times \frac{d\phi_s(x_m)}{dx_m} \bigg) \phi_r(x_i)$$

$$\overset{(b)}{\leq} k^2 \phi_{\max} \bar{\phi}_{\max} \theta_{\max}, \tag{60}$$

where (a) follows from the definition of $\boldsymbol{\lambda}^{*T}(x_{-i})$ and $\boldsymbol{\phi}(x_i)$ and (b) follows because $\phi_r(x_i) \leq \phi_{\max}$ $\forall r \in [k], \forall x_i \in \mathcal{X}_i$ and $\frac{d\phi_s(x_m)}{dx_m} \leq \bar{\phi}_{\max}$ $s \in [k], \forall x_m \in \mathcal{X}_m$.

Using (60) along with the fact that $|\phi_j(x_i)| \leq \phi_{\max}$ $\forall j \in [k], \forall x_i \in \mathcal{X}_i$, we can further upper bound $\Big|\partial\mu^*_j(x_{-i})/\partial x_m\Big|$ as

$$\left|\frac{\partial\mu^*_j(x_{-i})}{\partial x_m}\right| \leq k^2\theta_{\max}\phi_{\max}\bar{\phi}_{\max} \times \phi_{\max} + k^2\theta_{\max}\phi_{\max}\bar{\phi}_{\max} \times \phi_{\max} = L_1.$$

As a result, we have $\|\nabla\boldsymbol{\mu}^*(x_{-i})\|_\infty \leq L_1$ and this concludes the proof. □

Q.2 Learning conditional mean parameters as a sparse linear regression

The following Lemma shows that learning the conditional mean parameters $\boldsymbol{\mu}^*(x_{-i})$ as a function of x_{-i} using an estimate of the graph structure is equivalent to solving a sparse linear regression problem.

Lemma Q.2. *Suppose we have an estimate $\hat{G}$ of $G(\boldsymbol{\theta}^*)$ such that for any $\delta_4 \in (0,1)$, $\hat{G} = G(\boldsymbol{\theta}^*)$ with probability at least $1-\delta_4$. Let t be a parameter and $\tilde{p}$ be such that $\tilde{p} \leq \big(b_u/t\big)^d$. The following holds with probability at least $1-\delta_4$. For every $i \in [p], j \in [k]$, $x_{-i} \in \Pi_{j\in[p]\setminus\{i\}}\mathcal{X}_j$, we can write $\mu^*_j(x_{-i})$ as the following sparse linear regression:*

$$\mu^*_j(x_{-i}) = {\boldsymbol{\Psi}^{(j)}}^T\mathbf{b} + \bar{\eta},$$

where $\boldsymbol{\Psi}^{(j)} \in \mathbb{R}^{\tilde{p}}$ is the unknown parameter vector and $\mathbf{b} \in \mathbb{R}^{\tilde{p}}$ is the covariate vector and it is a function of x_{-i}. Further, we also have $|\bar{\eta}| \leq L_1 dt$, $\|\mathbf{b}\|_\infty \leq 1$ and $\|\boldsymbol{\Psi}^{(j)}\|_1 \leq \phi_{\max}\big(b_u/t\big)^d$.

Proof of Lemma Q.2. For mathematical simplicity, $\forall i \in [p]$ let the interval $\mathcal{X}_i = \mathcal{X}_b = [0,b]$ where b is such that $b_l \leq b \leq b_u$. Divide the interval $\mathcal{X}_b$ into non-overlapping intervals of length t. For the sake of simplicity, we assume that b/t is an integer. Let us enumerate the resulting b/t intervals as the set of integers $\mathcal{I} \coloneqq \{1,\cdots,b/t\}$. For any $x \in \mathcal{X}_b$, $\exists \zeta \in \mathcal{I}$ s.t $x \in ((\zeta-1)t, \zeta t]$ and this allows us to define a map $\mathcal{M}: \mathcal{X}_b \to \mathcal{I}$ s.t $\mathcal{M}(x) = \zeta t$. Similarly, for any $\mathbf{x} \coloneqq (x_j : j \in \mathcal{J}) \in \mathcal{X}_b^{|\mathcal{J}|}$ where $\mathcal{J}$ is any subset of $[p]$, we have the mapping $\mathcal{M}(\mathbf{x}) = \boldsymbol{\zeta}t$ where $\boldsymbol{\zeta} \coloneqq (\zeta_j : j \in \mathcal{J})$ is such that $\zeta_j = \mathcal{M}(x_j)/t$. Now for any $x \in \mathcal{X}_b$, consider a binary mapping $\mathcal{W}: \mathcal{X}_b \to \{0,1\}^{\mathcal{I}}$ defined as $\mathcal{W}(x) = (w_j(x) : j \in \mathcal{I})$ such that $w_{\mathcal{M}(x)/t}(x) = 1$ and $w_j(x) = 0$ $\forall j \in \mathcal{I}\setminus\{\mathcal{M}(x)/t\}$.

Let us condition on the event that $\hat{G} = G(\boldsymbol{\theta}^*)$. Therefore, we know the true neighborhood $\mathcal{N}(i)$ $\forall i \in [p]$ with probability at least $1-\delta_4$ (because $\hat{G} = G(\boldsymbol{\theta}^*)$ with probability at least $1-\delta_4$). Using the Markov property of the graph $G(\boldsymbol{\theta}^*)$, we know that the conditional density (and therefore the conditional mean parameters) of a node x_i given the values taken by the rest of nodes depend only on the values taken by the neighbors of x_i. Therefore, we have

$$\mu^*_j(x_{-i}) = \mu^*_j(x_{\mathcal{N}(i)}), \tag{61}$$

where $x_{\mathcal{N}(i)}$ denotes the values taken by the neighbors of x_i. Using the fact that max-degree of any node in $G(\boldsymbol{\theta}^*)$ is at-most d and Lemma Q.1, we can write for any $j \in [k]$

$$\Big|\mu^*_j(x_{\mathcal{N}(i)}) - \mu^*_j\big(\mathcal{M}(x_{\mathcal{N}(i)})\big)\Big| \leq L_1\sqrt{d}\Big\|x_{\mathcal{N}(i)} - \mathcal{M}(x_{\mathcal{N}(i)})\Big\|_2$$

$$\overset{(a)}{\leq} L_1 dt, \tag{62}$$

where (a) follows because $\forall m \in [p]$, $|x_m - \mathcal{M}(x_m)| \leq t$ and cardinality of $\mathcal{N}(i)$ is no more than d. Now using the binary mapping $\mathcal{W}$ defined above, we can expand $\mu_j^*\big(\mathcal{M}(x_{\mathcal{N}(i)})\big)$ as

$$\mu_j^*\big(\mathcal{M}(x_{\mathcal{N}(i)})\big) = \sum_{k_1 \in \mathcal{I}} \cdots \sum_{k_{|\mathcal{N}(i)|} \in \mathcal{I}} \Big(\prod_{m=1}^{|\mathcal{N}(i)|} w_{k_m}(x_{\mathcal{N}(i)_m}) \Big) \mu_j^*\big(k_1 t, \cdots, k_{|\mathcal{N}(i)|} t\big), \tag{63}$$

where $\mathcal{N}(i)_m$ denotes the m^{th} element of $\mathcal{N}(i)$. Observe that, $\prod_{m=1}^{|\mathcal{N}(i)|} w_{k_m}(x_{\mathcal{N}(i)_m}) = 1$ only when $k_m t = \mathcal{M}(x_{\mathcal{N}(i)_m})$ $\forall m \in [|\mathcal{N}(i)|]$.

Combining (61), (62) and (63) we have the following regression problem:

$$\mu_j^*(x_{-i}) = \mathbf{\Psi}^{(j)^T} \mathbf{b} + \bar{\eta},$$

where $\mathbf{\Psi}^{(j)} := \Big(\mu_j^*\big(k_1 t, \cdots, k_{|\mathcal{N}(i)|} t\big) : k_r \in \mathcal{I} \;\; \forall r \in [|\mathcal{N}(i)|]\Big) \in \mathbb{R}^{\tilde{p}}$, $\tilde{p} = \big(b/t\big)^{\mathcal{N}(i)}$, $\mathbf{b} = \Big(\prod_{m=1}^{|\mathcal{N}(i)|} w_{k_m}(x_{\mathcal{N}(i)_m}) : k_r \in \mathcal{I} \;\; \forall r \in [|\mathcal{N}(i)|]\Big) \in \{0,1\}^{\tilde{p}}$, and $\bar{\eta}$ is such that $|\bar{\eta}| \leq L_1 dt$. Observe that $\|\mathbf{b}\|_\infty \leq 1$. Using the fact that cardinality of $\mathcal{N}(i)$ is no more than d, we have

$$\tilde{p} \leq \left(\frac{b}{t}\right)^d \leq \left(\frac{b_u}{t}\right)^d.$$

Using the fact that the conditional mean parameters are upper bounded by $\phi_{\max}$, we have the following sparsity condition:

$$\left\| \mathbf{\Psi}^{(j)} \right\|_1 \leq \phi_{\max} \left(\frac{b_u}{t}\right)^d.$$

□

Q.3 Proof of Proposition O.1

We restate the Proposition and then provide the proof.

Proposition O.2. *Let $\epsilon_6 > 0$. Let $\hat{\boldsymbol{\rho}}$ denote the output of Algorithm 4 with $\xi = 1/\bar{c}_2$, $\tau_1 = 8 b_l^{-2} \exp(12 k \rho_{\max} \phi_{\max}) \Big[\log \frac{4\phi_{\max}\sqrt{b_u}}{\epsilon_9 \sqrt{b_l}} + k\rho_{\max}\phi_{\max}\Big]$, $\tau_2 = \frac{8\phi_{\max}^2}{\epsilon_5^2} \log\big(\frac{2k\tau_3}{\delta_5}\big)$, $\tau_3 = \frac{\bar{c}_2}{\bar{c}_1} \log\Big(\frac{k\rho_{\max}^2}{\epsilon_6^2 - \bar{c}_3}\Big)$, $w_{(0)} = 0$, $\boldsymbol{\rho}^{(0)} = (0, \cdots, 0)$ and $\hat{\boldsymbol{\upsilon}} = (\hat{\upsilon}_1, \cdots, \hat{\upsilon}_k)$. Then,*

$$\|\boldsymbol{\rho}^* - \hat{\boldsymbol{\rho}}\|_2 \leq \epsilon_6,$$

with probability at least $1 - 2\delta_5$.

Proof of Proposition O.1. Let us condition on the event that $\hat{G} = G(\boldsymbol{\theta}^*)$. The following holds with probability at least $1 - \delta_4$. From Lemma Q.2, for a parameter t and for every $i \in [p], j \in [k]$, $x_{-i} \in \Pi_{j \in [p]\setminus\{i\}} \mathcal{X}_j$, we have

$$\mu_j^*(x_{-i}) = \mathbf{\Psi}^{(j)^T} \mathbf{b} + \bar{\eta},$$

where $\mathbf{\Psi}^{(j)} \in \mathbb{R}^{\tilde{p}}$ is an unknown parameter vector and $\mathbf{b} \in \mathbb{R}^{\tilde{p}}$, a function of x_{-i}, is the covariate vector. Further, we also have $\tilde{p} = \big(b_u/t\big)^d$, $|\bar{\eta}| \leq L_1 dt$, $\|\mathbf{b}\|_\infty \leq 1$ and $\|\mathbf{\Psi}^{(j)}\|_1 \leq \phi_{\max}\big(b_u/t\big)^d$.

Suppose $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}$ are the n independent samples of $\mathbf{x}$. We tranform these to obtain the corresponding covariate vectors $\mathbf{b}^{(1)}, \cdots, \mathbf{b}^{(n)}$ where $\mathbf{b}^{(l)} = \Big(\prod_{m=1}^{|\mathcal{N}(i)|} w_{k_m}(x_{\mathcal{N}(i)_m}^{(l)}) : k_r \in \mathcal{I} \;\; \forall r \in [|\mathcal{N}(i)|]\Big)$. Let $\mathbf{B}$ be a $n \times \tilde{p}$ matrix such that l^{th} row of $\mathbf{B}$ is $\mathbf{b}^{(l)}$. We also obtain the vector $\bar{\boldsymbol{\mu}}_j(x_{-i}) := (\mu_j^{(r)}(x_{-i}) : r \in [n])$ where $\mu_j^{(r)}(x_{-i}) = \phi_j(x_i^{(r)})$. Letting $\tilde{\epsilon} = \phi_j(x_i) - \mu_j^*(x_{-i})$, we see that $\tilde{\epsilon}$ is bounded *sub-Gaussian* random variable with zero mean and variance proxy $\tilde{\sigma}^2 = 4\phi_{\max}^2$ (follows from Hoeffding's lemma).

Let $\hat{\mathbf{\Psi}}^{(j)}$ be the output of algorithm 2 with inputs $\mathbf{V} = \mathbf{B}, \mathbf{y} = \bar{\boldsymbol{\mu}}_j(x_{-i})$ and $\tilde{c}_2 = \phi_{\max}\big(b_u/t\big)^d$. Using Lemma N.1 with $\tilde{\eta} = L_1 dt$, $\tilde{c}_1 = 1$, and $\tilde{\sigma} = 2\phi_{\max}$ we have

$$\mathbb{E}[\widehat{\text{MSPE}}(\hat{\mathbf{\Psi}}^{(j)})] \leq 4L_1^2 d^2 t^2 + 8\phi_{\max}^2 \bigg(\frac{b_u}{t}\bigg)^d \sqrt{\frac{2\log 2\tilde{p}}{n}}.$$

Using the upper bound on $\tilde{p}$ results in

$$\mathbb{E}[\widehat{\text{MSPE}}(\hat{\mathbf{\Psi}}^{(j)})] \leq 4L_1^2 d^2 t^2 + 8\phi_{\max}^2 \bigg(\frac{b_u}{t}\bigg)^d \sqrt{\frac{2d}{n}\log\frac{2^{1/d} b_u}{t}}.$$

As $d \geq 1$, we have $2^{1/2d} \leq 2$. Choosing the parameter $t = \frac{\epsilon_4^2}{8\sqrt{2}L_1 d}$ and plugging in $L_1 = 2k^2\theta_{\max}\phi_{\max}^2\bar{\phi}_{\max}$ and n, we have

$$\mathbb{E}[\widehat{\text{MSPE}}(\hat{\mathbf{\Psi}}^{(j)})] \leq \frac{\epsilon_4^4}{16}. \tag{64}$$

Consider $x_{-i}^{(z)}$ where z is chosen uniformly at random from $[n]$. For the prediction $\hat{\mu}_j(x_{-i}^{(z)})$, we transform $x_{-i}^{(z)}$ to obtain the corresponding covariate vector $\mathbf{b} = \Big(\prod_{m=1}^{|\mathcal{N}(i)|} w_{k_m}(x_{\mathcal{N}(i)_m}^{(z)}) : k_r \in \mathcal{I} \;\; \forall r \in [|\mathcal{N}(i)|]\Big)$ and take its dot product with $\hat{\mathbf{\Psi}}^{(j)}$ as follows:

$$\hat{\mu}_j(x_{-i}^{(z)}) = \hat{\mathbf{\Psi}}^{(j)^T}\mathbf{b}.$$

Using Markov's inequality, we have

$$\mathbb{P}(|\mathbf{\Psi}^{(j)^T}\mathbf{b} - \hat{\mathbf{\Psi}}^{(j)^T}\mathbf{b}|^2 \geq \frac{\epsilon_4^2}{4}) \leq \frac{4\mathbb{E}[(\mathbf{\Psi}^{(j)^T}\mathbf{b} - \hat{\mathbf{\Psi}}^{(j)^T}\mathbf{b})^2]}{\epsilon_4^2} \overset{(a)}{=} \frac{4\mathbb{E}[\widehat{\text{MSPE}}(\hat{\mathbf{\Psi}}^{(j)})]}{\epsilon_4^2} \overset{(b)}{\leq} \frac{\epsilon_4^2}{4},$$

where (a) follows from Definition N.2 and (b) follows from (64). Therefore, we have $|\mathbf{\Psi}^{(j)^T}\mathbf{b} - \hat{\mathbf{\Psi}}^{(j)^T}\mathbf{b}| \leq \frac{\epsilon_4}{2}$ with probability at least $1 - \frac{\epsilon_4^2}{4}$.

Further, the following holds with probability at least $1 - \frac{\epsilon_4^2}{4}$:

$$\begin{aligned} |\mu_j^*(x_{-i}^{(z)}) - \hat{\mu}_j(x_{-i}^{(z)})| &= |\mathbf{\Psi}^{(j)^T}\mathbf{b} + \bar{\eta} - \hat{\mathbf{\Psi}}^{(j)^T}\mathbf{b}| \\ &\overset{(a)}{\leq} |\mathbf{\Psi}^{(j)^T}\mathbf{b} - \hat{\mathbf{\Psi}}^{(j)^T}\mathbf{b}| + |\bar{\eta}| \\ &\overset{(b)}{\leq} \frac{\epsilon_4}{2} + L_1 dt \overset{(c)}{\leq} \frac{\epsilon_4}{2} + \frac{\epsilon_4^2}{8\sqrt{2}} \overset{(d)}{\leq} \epsilon_4, \end{aligned} \tag{65}$$

where (a) follows from the triangle inequality, (b) follows because $|\bar{\eta}| \leq L_1 dt$ and $|\mathbf{\Psi}^{(j)^T}\mathbf{b} - \hat{\mathbf{\Psi}}^{(j)^T}\mathbf{b}| \leq \frac{\epsilon_4}{2}$, (c) follows by plugging in the value of t and L_1, and (d) follows because $\epsilon_4 \leq 1$.

As (65) holds $\forall j \in [k]$, the proof follows by using the union bound over all $j \in [k]$.

Solving the sparse linear regression takes number of computations that scale as $\tilde{p}^2 \times n$ (see Efron et al. (2004) for details). There are k such sparse linear regression problems for each node. Substituting for $\tilde{p}$, t, and n, the total number of computations required scale as

$$\bigg(\frac{2^{18d+17} b_u^{4d} k^{8d+1} d^{4d+1} \theta_{\max}^{4d} \phi_{\max}^{8d+4} \bar{\phi}_{\max}^{4d}}{\epsilon_4^{8d+8}} \times p\bigg) \log\bigg(\frac{2^{5.5} b_u k^2 d\theta_{\max}\phi_{\max}^2\bar{\phi}_{\max}}{\epsilon_4^2}\bigg).$$

The log term is dominated by the preceding term. □

R Analysis of Algorithm 3

In this appendix, we discuss the theoretical properties of Algorithm 3. These will be used in the proof of Proposition O.2. Recall that we design a Markov chain in Algorithm 3 that estimates the mean parameter vector of an exponential family distribution whose canonical parameters are known. The sufficient statistic vector of this exponential family distribution is the basis vector $\boldsymbol{\phi}(\cdot)$. We design this Markov chain using a zeroth-order Metropolized random walk algorithm. We will provide an upper bound on the mixing time of this Markov chain (Lemma R.2) and provide error bounds on the estimate of the mean parameter vector computed using the samples obtained from the Markov chain (Lemma R.3).

R.1 Setup: The exponential family distribution

Let $\mathcal{X}_0$ be a real interval such that its length is upper (lower) bounded by known constant b_u (b_l). Suppose that w is a random variable that takes value in $\mathcal{X}_0$ with probability density function as follows,

$$f_{\mathsf{w}}(w; \boldsymbol{\rho}) \propto \exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w)), \tag{66}$$

where $\boldsymbol{\rho} \coloneqq (\rho_1, \cdots, \rho_k)$ is the canonical parameter vector of the density in (66) and it is such that $\|\boldsymbol{\rho}\|_\infty \leq \rho_{\max}$. Let the cumulative distribution function of w be denoted by $F_{\mathsf{w}}(\cdot; \boldsymbol{\rho})$. Let $\boldsymbol{\nu}(\boldsymbol{\rho}) = \mathbb{E}_{\mathsf{w}}[\boldsymbol{\phi}(\mathsf{w})] \in \mathbb{R}^k$ be the mean parameter vector of the density in (66), i.e., $\boldsymbol{\nu}(\boldsymbol{\rho}) = (\nu_1, \cdots, \nu_k)$ such that

$$\nu_j \coloneqq \int_{w \in \mathcal{X}_0} \phi_j(w) f_{\mathsf{w}}(w; \boldsymbol{\rho}) dw. \tag{67}$$

We aim to estimate $\boldsymbol{\nu}(\boldsymbol{\rho})$ for a given parameter vector $\boldsymbol{\rho}$ using Algorithm 3. Let the estimated vector of mean parameters be denoted by $\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}) \coloneqq (\hat{\nu}_1, \cdots, \hat{\nu}_k)$. Let $Z(\boldsymbol{\rho})$ be the partition function of $f_{\mathsf{w}}(\cdot; \boldsymbol{\rho})$ i.e.,

$$Z(\boldsymbol{\rho}) = \int_{w \in \mathcal{X}_0} \exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w)) dw. \tag{68}$$

R.2 Bounds on the probability density function

Let us define $\mathcal{H}(\cdot) \coloneqq \exp(|\boldsymbol{\rho}^T \boldsymbol{\phi}(\cdot)|)$ and $\mathcal{H}_{\max} \coloneqq \exp(k\rho_{\max}\phi_{\max})$. We have $\forall w \in \mathcal{X}_0$,

$$\mathcal{H}^{-1}(w) \leq \exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w)) \leq \mathcal{H}(w). \tag{69}$$

Bounding the density function defined in (66) using (69) results in

$$\frac{1}{b_u \mathcal{H}^2(w)} \leq f_{\mathsf{w}}(w; \boldsymbol{\rho}) \leq \frac{\mathcal{H}^2(w)}{b_l}. \tag{70}$$

Let us also upper bound $\mathcal{H}(\cdot)$. We have $\forall w \in \mathcal{X}_0$,

$$\mathcal{H}(w) \overset{(a)}{\leq} \exp(\sum_{j=1}^{k} |\rho_j \phi_j(w)|) \overset{(b)}{\leq} \exp(\rho_{\max} \sum_{j=1}^{k} |\phi_j(w)|) \overset{(c)}{\leq} \exp(k\rho_{\max}\phi_{\max}) = \mathcal{H}_{\max}, \tag{71}$$

where (a) follows from the triangle inequality, (b) follows because $|\rho_j| \leq \rho_{\max}$ $\forall j \in [k]$, and (c) follows because $|\phi_j(w)| \leq \phi_{\max}$ $\forall j \in [k]$ and $\forall w \in \mathcal{X}_0$.

R.3 Mixing time of the Markov chain in Algorithm 3

We set up an irreducible, aperiodic, time-homogeneous, discrete-time Markov chain, whose stationary distribution is equal to $F_{\mathsf{w}}(w; \boldsymbol{\rho})$, using a zeroth-order Metropolized random walk algorithm (Hastings, 1970; Metropolis et al., 1953). The Markov chain is defined on a measurable state space $(\mathcal{X}_0, \mathcal{B}(\mathcal{X}_0))$ with a transition kernel $\mathcal{K} : \mathcal{X}_0 \times \mathcal{B}(\mathcal{X}_0) \to \mathbb{R}_+$ where $\mathcal{B}(\mathcal{X}_0)$ denotes the $\sigma-$algebra of $\mathcal{X}_0$.

R.3.1 Total variation distance

Definition R.1. *Let Q_1 be a distribution with density q_1 and Q_2 be a distribution with density q_2 defined on a measureable state space $(\mathcal{X}_0, \mathcal{B}(\mathcal{X}_0))$. The total variation distance of Q_1 and Q_2 is defined as*

$$\|Q_1 - Q_2\|_{TV} = \sup_{A \in \mathcal{B}(\mathcal{X}_0)} |Q_1(A) - Q_2(A)|.$$

The following Lemma shows that if the total variation distance between two distributions on the same domain is small, then $\forall j \in [k]$, the difference between the expected value of $\phi_j(\cdot)$ with respect to the two distributions is also small.

Lemma R.1. *Let Q_1 and Q_2 be two different distributions of the random variable w defined on $\mathcal{X}_0$. Let $\|Q_1 - Q_2\|_{TV} \leq \epsilon_7$ for any $\epsilon_7 > 0$. Then,*

$$\Big\|\mathbb{E}_{Q_1}[\boldsymbol{\phi}(\mathsf{w})] - \mathbb{E}_{Q_2}[\boldsymbol{\phi}(\mathsf{w})]\Big\|_\infty \leq 2\epsilon_7\phi_{\max}.$$

Proof of Lemma R.1. We will use the following relationship between the total variation distance and the ℓ_1 norm in the proof:

$$\|Q_1 - Q_2\|_{TV} = \frac{1}{2}\int_{\mathcal{X}_0} |q_1(w) - q_2(w)| dw. \tag{72}$$

For any $j \in [k]$, we have,

$$\begin{aligned}\Big|\mathbb{E}_{Q_1}[\phi_j(\mathsf{w})] - \mathbb{E}_{Q_2}[\phi_j(\mathsf{w})]\Big| &\overset{(a)}{=} \Big|\int_{\mathcal{X}_0} \phi_j(w)[q_1(w) - q_2(w)]dw\Big| \\ &\overset{(b)}{\leq} \int_{\mathcal{X}_0} |\phi_j(w)||q_1(w) - q_2(w)|dw \\ &\overset{(c)}{\leq} \phi_{\max}\int_{\mathcal{X}_0} |q_1(w) - q_2(w)|dw \\ &\overset{(d)}{=} 2\phi_{\max}\|Q_1 - Q_2\|_{TV} \\ &\leq 2\epsilon_7\phi_{\max},\end{aligned}$$

where (a) follows from the definition of expectation, (b) follows because the absolute value of integral is less than integral of absolute value, (c) follows because $|\phi_j(w)| \leq \phi_{\max}$ $\forall j \in [k]$, and (d) follows from (72). □

R.3.2 Definitions

Definition R.2. *Given a distribution F_0 with density f_0 on the current state of a Markov chain, the transition operator $\mathcal{T}(F_0)$ gives the distribution of the next state of the chain. Mathematically, we have*

$$\mathcal{T}(F_0)(A) = \int_{\mathcal{X}_0} \mathcal{K}(w, A)f_0(w)dw, \text{ for any } A \in \mathcal{B}(\mathcal{X}_0). \tag{73}$$

Definition R.3. *The mixing time of a Markov chain, with initial distribution F_0 and transition operator $\mathcal{T}$, in a total variation distance sense with respect to its stationary distribution F_{w}, is defined as*

$$\tau(\epsilon) = \inf\Big\{r \in \mathbb{N} \;\; s.t \;\; \big\|\mathcal{T}^{(r)}(F_0) - F_{\mathsf{w}}\big\|_{TV} \leq \epsilon\Big\},$$

where ϵ is an error tolerance and $\mathcal{T}^{(r)}$ stands for r-step transition operator.

Definition R.4. *The conductance of the Markov chain with transition operator $\mathcal{T}$ and stationary distribution F_{w} (with density $f_{\mathsf{w}}(w)$) is defined as*

$$\varphi := \min_{0 < F_{\mathsf{w}}(A) \leq \frac{1}{2}} \frac{\int_A \mathcal{T}(\delta_w)(A^c)f_{\mathsf{w}}(w)dw}{F_{\mathsf{w}}(A)},$$

where $\mathcal{T}(\delta_w)$ is obtained by applying the transition operator to a Dirac distribution concentrated on w.

R.3.3 Upper bound on the mixing time

Recall that $\mathcal{U}_{\mathcal{X}_0}$ denotes the uniform distribution on $\mathcal{X}_0$. We let the initial distribution of the Markov chain be $\mathcal{U}_{\mathcal{X}_0}$. We run independent copies of the Markov chain and use the samples obtained after the mixing time in each copy to compute $\hat{\boldsymbol{\nu}}$. In Algorithm 3, τ_1 is the number of iterations of the Markov chain and τ_2 denotes the number of independent copies of the Markov chain used. The following Lemma gives an upper bound on the mixing time of the Markov chain defined in Algorithm 3.

Lemma R.2. *Let the mixing time of the Markov chain defined in Algorithm 3 be denoted by $\tau_M(\epsilon_8)$ where $\epsilon_8 > 0$ is the error tolerance. Then,*

$$\tau_M(\epsilon_8) \leq 8 b_l^{-2} \exp(12 k \rho_{\max} \phi_{\max}) \bigg[\log \frac{\sqrt{b_u}}{\epsilon_8 \sqrt{b_l}} + k \rho_{\max} \phi_{\max} \bigg].$$

Proof of Lemma R.2. We will control the mixing time of the Markov chain via worst-case conductance bounds. This method was introduced for discrete space Markov chains by Jerrum and Sinclair Jerrum and Sinclair (1988) and then extended to the continuous space Markov chains by Lovász and Simonovits (Lovász and Simonovits, 1993); see Vempala (2005) for a detailed discussion on the continuous space setting.

For any initial distribution F_0 and stationary distribution F_w of a Markov chain, define $c_0 \coloneqq \sup_A \frac{F_0(A)}{F_w(A)}$. Lovász and Simonovits (1993) proved that,

$$\big|\big| \mathcal{T}^{(r)}(F_0) - F_w \big|\big|_{TV} \leq \sqrt{c_0} \exp^{-r\varphi^2/2}.$$

Therefore to upper bound the total variation distance by ϵ_8, it is sufficient to have

$$\sqrt{c_0} \exp^{-r\varphi^2/2} \leq \epsilon_8.$$

This can be rewritten as

$$r \geq \frac{2}{\varphi^2} \log \frac{\sqrt{c_0}}{\epsilon_8}.$$

Therefore, after $r = \frac{2}{\varphi^2} \log \frac{\sqrt{c_0}}{\epsilon_8}$ steps of the Markov chain, the total variation distance is less than ϵ_8 and $\tau_M(\epsilon_8) \leq \frac{2}{\varphi^2} \log \frac{\sqrt{c_0}}{\epsilon_8}$. In order to upper bound the mixing time, we need upper bound the constant c_0 and lower bound the conductance φ.

We will first upper bound c_0. We have the initial distribution to be uniform on $\mathcal{X}_0$. Therefore,

$$c_0 = \sup_A \frac{\mathcal{U}_{\mathcal{X}_0}(A)}{F_w(A)} \overset{(a)}{\leq} \sup_A \frac{\int_A \frac{1}{b_l} dw}{\int_A \frac{1}{b_u \mathcal{H}^2(w)} dw} \overset{(b)}{\leq} \frac{b_u}{b_l} \mathcal{H}^2_{\max}, \tag{74}$$

where (a) follows from the lower bound in (70) and because the length of $\mathcal{X}_0$ is lower bounded by b_l and (b) from (71).

Let us now lower bound φ. From the Definition R.4 we have,

$$\begin{aligned}
\varphi &= \min_{0 < \int_A f_w(w;\boldsymbol{\rho}) dw \leq \frac{1}{2}} \frac{\int_A \mathcal{T}(\delta_w)(A^c) f_w(w;\boldsymbol{\rho}) dw}{\int_A f_w(w;\boldsymbol{\rho}) dw} \\
&\overset{(a)}{=} \min_{0 < \int_A f_w(w;\boldsymbol{\rho}) dw \leq \frac{1}{2}} \frac{\int_A \mathcal{T}(\delta_w)(A^c) \exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w)) dw}{\int_A \exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w)) dw} \\
&\overset{(b)}{\geq} \min_{0 < \int_A f_w(w;\boldsymbol{\rho}) dw \leq \frac{1}{2}} \frac{\int_A \mathcal{T}(\delta_w)(A^c) \mathcal{H}^{-1}(w) dw}{\int_A \mathcal{H}(w) dw} \\
&\overset{(c)}{\geq} \frac{1}{\mathcal{H}^2_{\max}} \min_{0 < \int_A f_w(w;\boldsymbol{\rho}) dw \leq \frac{1}{2}} \frac{\int_A \mathcal{T}(\delta_w)(A^c) dw}{\int_A dw} \\
&\overset{(d)}{\geq} \frac{1}{\mathcal{H}^2_{\max}} \min_{0 < \int_A f_w(w;\boldsymbol{\rho}) dw \leq \frac{1}{2}} \frac{\int_A \big(\int_{\mathcal{X}_0} \mathcal{K}(w, A^c) \delta_w(w) dw \big) dw}{\int_A dw} \\
&= \frac{1}{\mathcal{H}^2_{\max}} \min_{0 < \int_A f_w(w;\boldsymbol{\rho}) dw \leq \frac{1}{2}} \frac{\int_A \mathcal{K}(w, A^c) dw}{\int_A dw} \\
&= \frac{1}{\mathcal{H}^2_{\max}} \min_{0 < \int_A f_w(w;\boldsymbol{\rho}) dw \leq \frac{1}{2}} \frac{\int_A \int_{A^c} \mathcal{K}(w, dy) dy dw}{\int_A dw},
\end{aligned}$$

where (a) follows by canceling out $Z(\boldsymbol{\rho})$ in the numerator and the denominator, (b) follows from (69), (c) follows from (71), and (d) follows from (73).

Recall from Algorithm 3 that we make a transition from the current state w to the next state y with probability $\mathcal{K}(w, dy) = \min\left\{1, \frac{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(y))}{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w))}\right\}$. Therefore,

$$\varphi \geq \frac{1}{\mathcal{H}_{\max}^2} \min_{0<\int_A f_{\mathbf{w}}(w;\boldsymbol{\rho})dw \leq \frac{1}{2}} \frac{\int_A \int_{A^c} \min\left\{1, \frac{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(y))}{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w))}\right\} dy dw}{\int_A dw}.$$

Using (69) and observing that $\mathcal{H}_{\max}^{-2} \leq 1$, we have $\min\left\{1, \frac{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(y))}{\exp(\boldsymbol{\rho}^T \boldsymbol{\phi}(w))}\right\} \geq \frac{1}{\mathcal{H}_{\max}^2}$. This results in,

$$\varphi \geq \frac{1}{\mathcal{H}_{\max}^4} \min_{0<\int_A f_{\mathbf{w}}(w;\boldsymbol{\rho})dw \leq \frac{1}{2}} \frac{\int_A \int_{A^c} dy dw}{\int_A dw} = \frac{1}{\mathcal{H}_{\max}^4} \min_{0<\int_A f_{\mathbf{w}}(w;\boldsymbol{\rho})dw \leq \frac{1}{2}} \int_{A^c} dw. \tag{75}$$

We have $\int_A f_{\mathbf{w}}(w;\boldsymbol{\rho})dw \leq \frac{1}{2}$. This can be rewritten as,

$$\int_{A^c} f_{\mathbf{w}}(w;\boldsymbol{\rho})dw \geq \frac{1}{2} \implies \int_{A^c} dw \overset{(a)}{\geq} \frac{b_l}{2\mathcal{H}_{\max}^2}, \tag{76}$$

where (a) follows from the upper bound in (70). Using (76) in (75), we have

$$\varphi \geq \frac{b_l}{2\mathcal{H}_{\max}^6}. \tag{77}$$

Now using (74) and (77) to bound the mixing time, we have

$$\tau_M(\epsilon_8) \leq \frac{8\mathcal{H}_{\max}^{12}}{b_l^2} \log \frac{\mathcal{H}_{\max}\sqrt{b_u}}{\epsilon_8\sqrt{b_l}}.$$

Using the upper bound of $\mathcal{H}_{\max}$ from (71), we have

$$\begin{aligned}\tau_M(\epsilon_8) &\leq \frac{8\exp(12k\rho_{\max}\phi_{\max})}{b_l^2} \log \frac{\sqrt{b_u}\exp(k\rho_{\max}\phi_{\max})}{\epsilon_8\sqrt{b_l}} \\ &= 8b_l^{-2}\exp(12k\rho_{\max}\phi_{\max})\left[\log\frac{\sqrt{b_u}}{\epsilon_8\sqrt{b_l}} + k\rho_{\max}\phi_{\max}\right].\end{aligned}$$

□

R.3.4 Guarantees on the output of Algorithm 3

The following Lemma shows that the estimate, obtained from Algorithm 3, of the mean parameter vector, is such that the ℓ_∞ error is small with high probability.

Lemma R.3. *Let $\epsilon_9 > 0$ and $\delta_9 \in (0,1)$. Let $\hat{\boldsymbol{\nu}}(\boldsymbol{\rho})$ be the output of Algorithm 3 with $w_{(0)} = 0$, $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_k)$, $\tau_1 = 8b_l^{-2}\exp(12k\rho_{\max}\phi_{\max})\left[\log\frac{4\phi_{\max}\sqrt{b_u}}{\epsilon_9\sqrt{b_l}} + k\rho_{\max}\phi_{\max}\right]$, and $\tau_2 = \frac{8\phi_{\max}^2}{\epsilon_9^2}\log\left(\frac{2}{\delta_9}\right)$. Then,*

$$\|\boldsymbol{\nu}(\boldsymbol{\rho}) - \hat{\boldsymbol{\nu}}(\boldsymbol{\rho})\|_\infty \leq \epsilon_9,$$

with probability at least $1 - k\delta_9$.

Proof of Lemma R.3. The distribution of the Markov chain in Algorithm 3 after $\tau_1 + 1$ steps is $\mathcal{T}^{(\tau_1+1)}(\mathcal{U}_{\mathcal{X}_0})$ where $\mathcal{U}_{\mathcal{X}_0}$ denotes the initial uniform distribution. Let $\boldsymbol{\nu}^M(\boldsymbol{\rho}) := (\nu_1^M, \cdots, \nu_k^M)$ be the vector such that ν_j^M is the expected value of $\phi_j(\cdot)$ with respect to the distribution $\mathcal{T}^{(\tau_1+1)}(\mathcal{U}_{\mathcal{X}_0})$. Using Lemma R.2, we have $\tau_1 \geq \tau_{\mathrm{M}}(\frac{\epsilon_9}{4\phi_{\max}})$. Therefore,

$$||\mathcal{T}^{\tau_1+1}(\mathcal{U}_{\mathcal{X}_0}) - F_{\mathbf{w}}||_{TV} \leq \frac{\epsilon_9}{4\phi_{\max}}.$$

From Lemma R.1, we have

$$\|\boldsymbol{\nu}(\boldsymbol{\rho}) - \boldsymbol{\nu}^M(\boldsymbol{\rho})\|_\infty \leq \frac{\epsilon_9}{2}. \tag{78}$$

$\hat{\boldsymbol{\nu}}(\boldsymbol{\rho})$ is computed using the samples obtained from the distribution $\mathcal{T}^{(\tau_1+1)}(\mathcal{U}_{\mathcal{X}_0})$. Using Hoeffding's inequality, we have $\forall j \in [k]$

$$\mathbb{P}(|\hat{\nu}_j - \nu_j^M| \geq t_0) \leq 2\exp(\frac{-\tau_2 t_0^2}{2\phi_{\max}^2}).$$

Therefore when $\tau_2 \geq \frac{2\phi_{\max}^2}{t_0^2}\log\left(\frac{2}{\delta_9}\right)$, we have $|\hat{\nu}_j - \nu_j^M| \leq t_0$ with probability at least $1 - \delta_9$.

Using the union bound $\forall j \in [k]$, when $\tau_2 \geq \frac{8\phi_{\max}^2}{\epsilon_9^2}\log\left(\frac{2}{\delta_9}\right)$, we have

$$\|\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}) - \boldsymbol{\nu}^M(\boldsymbol{\rho})\|_\infty \leq \frac{\epsilon_9}{2}, \tag{79}$$

with probability at least $1 - k\delta_9$.

Combining (78) and (79) by triangle inequality, we have

$$\|\boldsymbol{\nu}(\boldsymbol{\rho}) - \hat{\boldsymbol{\nu}}(\boldsymbol{\rho})\|_\infty = \|\boldsymbol{\nu}(\boldsymbol{\rho}) - \boldsymbol{\nu}^M(\boldsymbol{\rho}) + \boldsymbol{\nu}^M(\boldsymbol{\rho}) - \hat{\boldsymbol{\nu}}(\boldsymbol{\rho})\|_\infty \leq \|\boldsymbol{\nu}(\boldsymbol{\rho}) - \boldsymbol{\nu}^M(\boldsymbol{\rho})\|_\infty + \|\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}) - \boldsymbol{\nu}^M(\boldsymbol{\rho})\|_\infty \leq \epsilon_9.$$

□

S Proof of Proposition O.2

In this appendix, we prove Proposition O.2. First, we will prove the strict convexity of the log partition function of interest (Lemma S.1). Next, we express the conjugate duality between the mean parameters and the canonical parameters. Next, we argue the need of the projected gradient descent algorithm. Finally, we provide the proof of Proposition O.2.

Recall the setup for the projected gradient descent algorithm from Appendix O.2.1. Specifically, recall the definitions of $\boldsymbol{\rho}^*$, $f_{\mathsf{w}}(w; \boldsymbol{\rho}^*)$, $\mathcal{P}$, $\rho_{\max}$, and $\boldsymbol{\upsilon}^*$. Also, $\hat{\boldsymbol{\upsilon}}$ is an estimate of $\boldsymbol{\upsilon}^*$ such that, with probability at least $1 - \delta_5$, we have $\|\boldsymbol{\upsilon}^* - \hat{\boldsymbol{\upsilon}}\|_\infty \leq \epsilon_5$. Further, recall the setup from Appendix R.1. Specifically, for any $\boldsymbol{\rho} \in \mathcal{P}$, recall the definitions of $f_{\mathsf{w}}(w; \boldsymbol{\rho})$, $\boldsymbol{\nu}(\boldsymbol{\rho})$, and $Z(\boldsymbol{\rho})$ from (66), (67), and (68) respectively. Recall the definition of q^s from Section 2.

S.1 Convexity of the log partition function

Let $\Phi(\boldsymbol{\rho})$ be the log partition function of $f_{\mathsf{w}}(w; \boldsymbol{\rho})$. Because $f_{\mathsf{w}}(w; \boldsymbol{\rho})$ is an exponential family density, $\nabla\Phi(\boldsymbol{\rho}) = \boldsymbol{\nu}(\boldsymbol{\rho})$; see Wainwright and Jordan (2008) for details. The following Lemma shows that $\Phi(\boldsymbol{\rho})$ is a strictly convex function of $\boldsymbol{\rho}$.

Lemma S.1. $\Phi(\boldsymbol{\rho})$ *is a strictly convex function of* $\boldsymbol{\rho}$*.*

Proof of Lemma S.1. For any non-zero $\mathbf{e} \in \mathbb{R}^k$, $\mathbf{e}^T\boldsymbol{\phi}(w)$ is not a constant with respect to w. Therefore,

$$\begin{aligned}
0 &\overset{(a)}{<} \mathbb{V}\text{ar}\Big(\mathbf{e}^T\boldsymbol{\phi}(w)\Big) \\
&= \text{cov}\Big(\mathbf{e}^T\boldsymbol{\phi}(w), \mathbf{e}^T\boldsymbol{\phi}(w)\Big) \\
&= \sum_{j=1}^{k}\sum_{r=1}^{k} e_j e_r \times \text{cov}(\phi_j(w), \phi_r(w)) \\
&\overset{(b)}{=} \sum_{j=1}^{k}\sum_{r=1}^{k} e_j e_r [\nabla^2\Phi(\boldsymbol{\rho})]_{j,r}
\end{aligned}$$

$$= \mathbf{e}^T \nabla^2 \Phi(\boldsymbol{\rho}) \mathbf{e}.$$

where (a) follows because the variance of a non-constant random variable is strictly positive and (b) follows because for any regular exponential family the Hessian of the log partition function is the covariance matrix of the associated sufficient statistic vector; see Wainwright and Jordan (2008) for details.

Thus, $\nabla^2 \Phi(\boldsymbol{\rho})$ is a positive definite matrix and this is a sufficient condition for strict convexity of $\Phi(\boldsymbol{\rho})$. □

S.2 Conjugate Duality

Expressing the relationship between the canonical and mean parameters via conjugate duality (Bresler et al., 2014; Wainwright and Jordan, 2008), we know that for each $\boldsymbol{\upsilon}$ in the set of realizable mean parameters, there is a unique $\boldsymbol{\rho}(\boldsymbol{\upsilon}) \in \mathcal{P}$ satisfying the dual matching condition $\boldsymbol{\nu}(\boldsymbol{\rho}(\boldsymbol{\upsilon})) = \boldsymbol{\upsilon}$. The backward mapping of the mean parameters to the canonical parameters ($\boldsymbol{\upsilon} \mapsto \boldsymbol{\rho}(\boldsymbol{\upsilon})$) is given by,

$$\boldsymbol{\rho}(\boldsymbol{\upsilon}) = \arg\max_{\boldsymbol{\rho} \in \mathcal{P}} \Big\{ \langle \boldsymbol{\upsilon}, \boldsymbol{\rho} \rangle - \Phi(\boldsymbol{\rho}) \Big\}. \tag{80}$$

Defining $\Omega(\boldsymbol{\rho}, \boldsymbol{\upsilon}) := \Phi(\boldsymbol{\rho}) - \langle \boldsymbol{\upsilon}, \boldsymbol{\rho} \rangle$, we can rewrite (80) as

$$\boldsymbol{\rho}(\boldsymbol{\upsilon}) = \arg\min_{\boldsymbol{\rho} \in \mathcal{P}} \big\{ \Omega(\boldsymbol{\rho}, \boldsymbol{\upsilon}) \big\}. \tag{81}$$

For any $\boldsymbol{\rho} \in \mathcal{P}$, let $q(\boldsymbol{\rho})$ denote the smallest eigenvalue of the Hessian of the log partition function with canonical parameter $\boldsymbol{\rho}$. Recall that q^s denotes the minimum of $q(\boldsymbol{\rho})$ over all possible $\boldsymbol{\rho} \in \mathcal{P}$.

Lemma S.2. *$\Omega(\boldsymbol{\rho}, \boldsymbol{\upsilon})$ is a q^s strongly convex function of $\boldsymbol{\rho}$ and a $2k\phi_{\max}^2$ smooth function of $\boldsymbol{\rho}$.*

Proof of Lemma S.2. Observe that $\nabla^2 \Omega(\boldsymbol{\rho}, \boldsymbol{\upsilon}) = \nabla^2 \Phi(\boldsymbol{\rho})$. Therefore $\Omega(\boldsymbol{\rho}, \boldsymbol{\upsilon})$ being a q^s strongly convex function of $\boldsymbol{\rho}$ and a $2k\phi_{\max}^2$ smooth function of $\boldsymbol{\rho}$ is equivalent to $\Phi(\boldsymbol{\rho})$ being a q^s strongly convex function of $\boldsymbol{\rho}$ and a $2k\phi_{\max}^2$ smooth function of $\boldsymbol{\rho}$.

We will first show the strong convexity of $\Phi(\boldsymbol{\rho})$. Consider any $\mathbf{e} \in \mathbb{R}^k$ such that $\|\mathbf{e}\|_2 = 1$. We have

$$q(\boldsymbol{\rho}) = \inf_{\mathbf{e} : \|\mathbf{e}\|_2 \leq 1} \mathbf{e}^T \nabla^2 \Phi(\boldsymbol{\rho}) \mathbf{e}.$$

Using Lemma S.1 we know that $q(\boldsymbol{\rho}) > 0$ for any $\boldsymbol{\rho} \in \mathcal{P}$. Observe that $[\nabla^2 \Phi(\boldsymbol{\rho})]_{j,r} = \mathrm{cov}(\phi_j(w), \phi_r(w))$, and is a continuous function of $\boldsymbol{\rho}$,$\forall j, r \in [k]$. Now $q(\boldsymbol{\rho})$ is a linear combination of $[\nabla^2 \Phi(\boldsymbol{\rho})]_{j,r}$ $\forall j, r \in [k]$. Therefore $q(\boldsymbol{\rho})$ is also a continuous function of $\boldsymbol{\rho}$. Using the continuity of $q(\boldsymbol{\rho})$ and compactness of $\mathcal{P}$, we apply the extreme value theorem and conclude that the function $q(\boldsymbol{\rho})$ will attain its minimum value of

$$q^s = \inf_{\boldsymbol{\rho} \in \mathcal{P}} q(\boldsymbol{\rho}),$$

and that this value is positive. Now using the fact that $\nabla^2 \Phi(\boldsymbol{\rho})$ is a symmetric matrix and the Courant-Fischer theorem, we conclude that the minimum possible eigenvalue of $\nabla^2 \Phi(\boldsymbol{\rho})$ for any $\boldsymbol{\rho} \in \mathcal{P}$ is greater than or equal to q^s. Thus, the smallest possible eigenvalue of the Hessian of the log partition function is uniformly lower bounded. As a result, $\Phi(\boldsymbol{\rho})$ and $\Omega(\boldsymbol{\rho}, \boldsymbol{\upsilon})$ are q^s-strongly convex.

We will now show the smoothness of $\Phi(\boldsymbol{\rho})$. From the Gershgorin circle theorem, we know that the largest eigenvalue of any matrix is upper bounded by the largest absolute row sum or column sum. Applying this, we see that the largest eigenvalue of $\nabla^2 \Phi(\boldsymbol{\rho})$ is upper bounded by $\max_{1 \leq r \leq k} \sum_{j=1}^{k} |[\nabla^2 \Phi(\boldsymbol{\rho})]_{j,r}|$. Now

$$\begin{aligned}
\max_{1 \leq r \leq k} \sum_{j=1}^{k} |[\nabla^2 \Phi(\boldsymbol{\rho})]_{j,r}| &= \max_{1 \leq r \leq k} \sum_{j=1}^{k} |\mathrm{cov}(\phi_j(w), \phi_r(w))| \\
&\overset{(a)}{\leq} \max_{1 \leq r \leq k} \sum_{j=1}^{k} 2\phi_{\max}^2 \\
&\leq 2k\phi_{\max}^2,
\end{aligned}$$

where (a) follows from the triangle inequality and because $|\phi_j(w)| \leq \phi_{\max}$ $\forall j \in [k]$.

Now because the largest eigenvalue of the Hessian matrix of the log partition function is uniformly upper bounded by $2k\phi_{\max}^2$, $\Phi(\boldsymbol{\rho})$ and $\Omega(\boldsymbol{\rho}, \boldsymbol{\upsilon})$ are $2k\phi_{\max}^2$ smooth function of $\boldsymbol{\rho}$. □

S.3 Why projected gradient descent algorithm?

From Lemma S.2, we see that there is a unique minimum in (81). In other words, when the mean parameter in (81) is the true mean parameter of (54) i.e., $\boldsymbol{\upsilon} = \boldsymbol{\upsilon}^*$, then the unique minima in (81) is $\boldsymbol{\rho}^*$. Therefore, in principle, we can estimate $\boldsymbol{\rho}^*$ using a projected gradient descent algorithm.

In each step of this algorithm, we need access to $\boldsymbol{\nu}(\boldsymbol{\rho})$ for the estimate $\boldsymbol{\rho}$. However, we don't have access to $\boldsymbol{\upsilon}^*$ and $\boldsymbol{\nu}(\boldsymbol{\rho})$. Instead, we have access to $\hat{\boldsymbol{\upsilon}}$ and $\hat{\boldsymbol{\nu}}(\boldsymbol{\rho})$ (from Algorithm 3). Therefore, we can estimate the parameter vector $\boldsymbol{\rho}^*$ using the projected gradient descent in Algorithm 4.

S.4 Proof of Proposition O.2

We restate the Proposition below and then provide the proof.

Proposition O.2. *Let* $\epsilon_6 > 0$. *Let* $\hat{\boldsymbol{\rho}}$ *denote the output of Algorithm 4 with* $\xi = 1/\bar{c}_2$, $\tau_1 = 8b_l^{-2}\exp(12k\rho_{\max}\phi_{\max})\left[\log\frac{4\phi_{\max}\sqrt{b_u}}{\epsilon_9\sqrt{b_l}} + k\rho_{\max}\phi_{\max}\right]$, $\tau_2 = \frac{8\phi_{\max}^2}{\epsilon_5^2}\log\left(\frac{2k\tau_3}{\delta_5}\right)$, $\tau_3 = \frac{\bar{c}_2}{\bar{c}_1}\log\left(\frac{k\rho_{\max}^2}{\epsilon_6^2-\bar{c}_3}\right)$, $w_{(0)} = 0$, $\boldsymbol{\rho}^{(0)} = (0,\cdots,0)$ *and* $\hat{\boldsymbol{\upsilon}} = (\hat{\upsilon}_1,\cdots,\hat{\upsilon}_k)$. *Then,*

$$\|\boldsymbol{\rho}^* - \hat{\boldsymbol{\rho}}\|_2 \leq \epsilon_6,$$

with probability at least $1 - 2\delta_5$.

Proof of Lemma O.2. The projection of $\tilde{\boldsymbol{\rho}}$, onto a set $\mathcal{P}$ is defined as

$$\Pi_{\mathcal{P}}(\tilde{\boldsymbol{\rho}}) \coloneqq \operatorname*{arg\,min}_{\boldsymbol{\rho}\in\mathcal{P}} \|\boldsymbol{\rho} - \tilde{\boldsymbol{\rho}}\|.$$

If we had access to $\boldsymbol{\upsilon}^*$ and $\boldsymbol{\nu}(\boldsymbol{\rho})$, the iterates of the projected gradient descent algorithm could be rewritten as

$$\boldsymbol{\rho}^{(r+1)} = \boldsymbol{\rho}^{(r)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r)}),$$

where $\gamma_{\mathcal{P}}(\boldsymbol{\rho})$ is the gradient mapping and is defined as $\gamma_{\mathcal{P}}(\boldsymbol{\rho}) \coloneqq \frac{1}{\xi}(\boldsymbol{\rho} - \boldsymbol{\rho}^\dagger)$ with $\boldsymbol{\rho}^\dagger \coloneqq \Pi_{\mathcal{P}}(\boldsymbol{\rho} - \xi[\boldsymbol{\nu}(\boldsymbol{\rho}) - \boldsymbol{\upsilon}^*])$. See Bubeck (2015) for more details. Because we are using the respective estimates $\hat{\boldsymbol{\upsilon}}$ and $\hat{\boldsymbol{\nu}}(\boldsymbol{\rho})$, the iterates of the projected gradient descent algorithm are as follows:

$$\boldsymbol{\rho}^{(r+1)} = \boldsymbol{\rho}^{(r)} - \xi\hat{\gamma}_{\mathcal{P}}(\boldsymbol{\rho}^{(r)}),$$

where $\hat{\gamma}_{\mathcal{P}}(\boldsymbol{\rho}) \coloneqq \frac{1}{\xi}(\boldsymbol{\rho} - \boldsymbol{\rho}^{\dagger\dagger})$ with $\boldsymbol{\rho}^{\dagger\dagger} \coloneqq \Pi_{\mathcal{P}}(\boldsymbol{\rho} - \xi[\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}) - \hat{\boldsymbol{\upsilon}}])$.

Using Lemma R.3, we have

$$\|\boldsymbol{\nu}(\boldsymbol{\rho}) - \hat{\boldsymbol{\nu}}(\boldsymbol{\rho})\|_\infty \leq \epsilon_5,$$

with probability at least $1 - \delta_5/\tau_3$.

Recall the setup from Appendix O.2.1. Let us condition on the events that $\|\boldsymbol{\upsilon}^* - \hat{\boldsymbol{\upsilon}}\|_\infty \leq \epsilon_5$ and that, for each of the τ_3 steps of Algorithm 4, $\|\boldsymbol{\nu}(\boldsymbol{\rho}) - \hat{\boldsymbol{\nu}}(\boldsymbol{\rho})\|_\infty \leq \epsilon_5$. These events simultaneously hold with probability at least $1 - 2\delta_5$.

Now for any $r \leq \tau_3 + 1$ the following hold with probability at least $1 - 2\delta_5$:

$$\begin{aligned}
\|\boldsymbol{\rho}^{(r)} - \boldsymbol{\rho}^*\|_2 &= \|\boldsymbol{\rho}^{(r-1)} - \xi\hat{\gamma}_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2 \\
&= \|\boldsymbol{\rho}^{(r-1)} - \xi[\hat{\gamma}_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) + \gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)})] - \boldsymbol{\rho}^*\|_2 \\
&= \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \xi[\hat{\gamma}_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)})] - \boldsymbol{\rho}^*\|_2 \\
&\overset{(a)}{\leq} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2 + \xi\|\hat{\gamma}_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)})\|_2 \\
&\overset{(b)}{\leq} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2 + \xi\|\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\nu}(\boldsymbol{\rho}^{(r-1)}) + \boldsymbol{\upsilon}^* - \hat{\boldsymbol{\upsilon}}\|_2
\end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{\le} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2 + \xi\|\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\nu}(\boldsymbol{\rho}^{(r-1)})\|_2 + \xi\|\boldsymbol{\upsilon}^* - \hat{\boldsymbol{\upsilon}}\|_2 \\
&\overset{(d)}{\le} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2 + \xi\sqrt{k}\|\hat{\boldsymbol{\nu}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\nu}(\boldsymbol{\rho}^{(r-1)})\|_\infty + \xi\sqrt{k}\|\boldsymbol{\upsilon}^* - \hat{\boldsymbol{\upsilon}}\|_\infty \\
&\overset{(e)}{\le} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2 + 2\xi\sqrt{k}\epsilon_5,
\end{aligned} \tag{82}$$

where (a) follows from the triangle inequality, (b) follows from the definitions of $\gamma_{\mathcal{P}}(\boldsymbol{\rho})$ and $\hat{\gamma}_{\mathcal{P}}(\boldsymbol{\rho})$ and because the projection onto a convex set is non-expansive i.e., $\|\Pi_{\mathcal{P}}(\tilde{\boldsymbol{\rho}}) - \Pi_{\mathcal{P}}(\bar{\boldsymbol{\rho}})\| \le \|\tilde{\boldsymbol{\rho}} - \bar{\boldsymbol{\rho}}\|$, (c) follows from the triangle inequality, (d) follows because $\forall\ \mathbf{v} \in \mathbb{R}^k$, $\|\mathbf{v}\|_2 \le \sqrt{k}\|\mathbf{v}\|_\infty$, and (e) follows because of the conditioning.

Squaring both sides of (82) the following hold with probability at least $1 - 2\delta_5$:

$$\begin{aligned}
\|\boldsymbol{\rho}^{(r)} - \boldsymbol{\rho}^*\|_2^2 &\le \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2^2 + 4\xi^2 k\epsilon_5^2 + 4\xi\sqrt{k}\epsilon_5\|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2 \\
&\overset{(a)}{\le} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2^2 + 4\xi^2 k\epsilon_5^2 + 4\xi\sqrt{k}\epsilon_5\Big[\|\boldsymbol{\rho}^{(r-1)} - \boldsymbol{\rho}^*\|_2 + \xi\|\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)})\|_2\Big] \\
&\overset{(b)}{\le} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2^2 + 4\xi^2 k\epsilon_5^2 + 4\xi\sqrt{k}\epsilon_5\Big[\|\boldsymbol{\rho}^{(r-1)} - \boldsymbol{\rho}^*\|_2 + \xi\|\boldsymbol{\nu}(\boldsymbol{\rho}) - \boldsymbol{\upsilon}^*\|_2\Big] \\
&\overset{(c)}{\le} \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2^2 + 4\xi^2 k\epsilon_5^2 + 8\xi k\epsilon_5(\rho_{\max} + \xi\phi_{\max}),
\end{aligned}$$

where (a) follows from the triangle inequality, (b) follows by using the non-expansive property to observe that $\|\gamma_{\mathcal{P}}(\boldsymbol{\rho})\|_2 \le \|\boldsymbol{\nu}(\boldsymbol{\rho}) - \boldsymbol{\upsilon}^*\|_2$, and (c) follows because $\|\boldsymbol{\rho}^{(r-1)} - \boldsymbol{\rho}^*\|_2 \le 2\sqrt{k}\rho_{\max}$ and $\|\boldsymbol{\nu}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\upsilon}^*\|_2 \le 2\sqrt{k}\phi_{\max}$. Letting $\Upsilon(\xi) := 4\xi^2 k\epsilon_5^2 + 8\xi k\epsilon_5(\rho_{\max} + \xi\phi_{\max})$, the following hold with probability at least $1 - 2\delta_5$:

$$\begin{aligned}
\|\boldsymbol{\rho}^{(r)} - \boldsymbol{\rho}^*\|_2^2 &\le \|\boldsymbol{\rho}^{(r-1)} - \xi\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}) - \boldsymbol{\rho}^*\|_2^2 + \Upsilon(\xi) \\
&\overset{(a)}{=} \|\boldsymbol{\rho}^{(r-1)} - \boldsymbol{\rho}^*\|_2^2 + \xi^2\|\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)})\|_2^2 - 2\xi\Big\langle\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)}), \boldsymbol{\rho}^{(r-1)} - \boldsymbol{\rho}^*\Big\rangle + \Upsilon(\xi),
\end{aligned} \tag{83}$$

where (a) follows from the fact that for any two vectors $\mathbf{f}_1, \mathbf{f}_2$, $\|\mathbf{f}_1 - \mathbf{f}_2\|_2^2 = \|\mathbf{f}_1\|_2^2 + \|\mathbf{f}_2\|_2^2 - 2\langle\mathbf{f}_1, \mathbf{f}_2\rangle$.

For a twice differentiable, $\bar{c}_1$ strongly convex and $\bar{c}_2$ smooth function $\Omega(\boldsymbol{\rho})$, we have, for any $\boldsymbol{\rho} \in \mathcal{P}$

$$\langle\gamma_{\mathcal{P}}(\boldsymbol{\rho}), \boldsymbol{\rho} - \boldsymbol{\rho}^*\rangle \ge \frac{\bar{c}_1}{2}\|\boldsymbol{\rho} - \boldsymbol{\rho}^*\|_2^2 + \frac{1}{2\bar{c}_2}\|\gamma_{\mathcal{P}}(\boldsymbol{\rho})\|_2^2, \tag{84}$$

where $\boldsymbol{\rho}^*$ is the minimizer of $\Omega(\boldsymbol{\rho})$. See Bubeck (2015) for more details. Using (84) in (83), the following hold with probability at least $1 - 2\delta_5$:

$$\|\boldsymbol{\rho}^{(r)} - \boldsymbol{\rho}^*\|_2^2 \le (1 - \xi\bar{c}_1)\|\boldsymbol{\rho}^{(r-1)} - \boldsymbol{\rho}^*\|_2^2 + \bigg(\xi^2 - \frac{\xi}{\bar{c}_2}\bigg)\|\gamma_{\mathcal{P}}(\boldsymbol{\rho}^{(r-1)})\|_2^2 + \Upsilon(\xi).$$

Substituting $\xi = \frac{1}{\bar{c}_2}$, the following hold with probability at least $1 - 2\delta_5$:

$$\|\boldsymbol{\rho}^{(r)} - \boldsymbol{\rho}^*\|_2^2 \le \bigg(1 - \frac{\bar{c}_1}{\bar{c}_2}\bigg)\|\boldsymbol{\rho}^{(r-1)} - \boldsymbol{\rho}^*\|_2^2 + \Upsilon\bigg(\frac{1}{\bar{c}_2}\bigg).$$

Unrolling the recurrence gives, we have the following with probability at least $1 - 2\delta_5$:

$$\begin{aligned}
\|\boldsymbol{\rho}^{(r)} - \boldsymbol{\rho}^*\|_2^2 &\le \bigg(1 - \frac{\bar{c}_1}{\bar{c}_2}\bigg)^r\|\boldsymbol{\rho}^{(0)} - \boldsymbol{\rho}^*\|_2^2 + \sum_{j=0}^{r-1}\bigg(1 - \frac{\bar{c}_1}{\bar{c}_2}\bigg)^j\Upsilon\bigg(\frac{1}{\bar{c}_2}\bigg) \\
&\le \bigg(1 - \frac{\bar{c}_1}{\bar{c}_2}\bigg)^r\|\boldsymbol{\rho}^{(0)} - \boldsymbol{\rho}^*\|_2^2 + \sum_{j=0}^{\infty}\bigg(1 - \frac{\bar{c}_1}{\bar{c}_2}\bigg)^j\Upsilon\bigg(\frac{1}{\bar{c}_2}\bigg) \\
&\overset{(a)}{=} \bigg(1 - \frac{\bar{c}_1}{\bar{c}_2}\bigg)^r\|\boldsymbol{\rho}^*\|_2^2 + \bar{c}_3 \\
&\overset{(b)}{\le} \exp\Big(\frac{-\bar{c}_1 r}{\bar{c}_2}\Big)\|\boldsymbol{\rho}^*\|_2^2 + \bar{c}_3,
\end{aligned}$$

where (a) follows by observing that $\frac{\bar{c}_2}{\bar{c}_1}\Upsilon\big(\frac{1}{\bar{c}_2}\big) = \bar{c}_3$ and $\boldsymbol{\rho}^{(0)} = (0, \cdots, 0)$, and (b) follows because for any $y \in \mathbb{R}$, $1 - y \leq e^{-y}$.

A sufficient condition for $\|\boldsymbol{\rho}^{(r)} - \boldsymbol{\rho}^*\|_2 \leq \epsilon_6$ with probability at least $1 - 2\delta_5$ is

$$\exp\big(\frac{-\bar{c}_1 r}{\bar{c}_2}\big)\|\boldsymbol{\rho}^*\|_2^2 + \bar{c}_3 \leq \epsilon_6^2.$$

Rearranging gives us,

$$\exp\big(\frac{\bar{c}_1 r}{\bar{c}_2}\big) \geq \frac{\|\boldsymbol{\rho}^*\|_2^2}{\epsilon_6^2 - \bar{c}_3}.$$

Taking logarithm on both sides, we have

$$r \geq \frac{\bar{c}_2}{\bar{c}_1} \log\left(\frac{\|\boldsymbol{\rho}^*\|_2^2}{\epsilon_6^2 - \bar{c}_3}\right).$$

Observe that $\|\boldsymbol{\rho}^*\|_2^2 \leq k\rho_{\max}^2$. Therefore, after τ_3 steps, we have $\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}^*\|_2 \leq \epsilon_6$ with probability at least $1 - 2\delta_5$ and this completes the proof. □

T Examples of distributions

In this appendix, we discuss the examples of distributions from Section 4 that satisfy the Condition 4.1. We also discuss a few other examples. Recall the definitions of $\gamma = \theta_{\max}(k + k^2 d)$ and $\varphi_{\max} = 2\max\{\phi_{\max}, \phi_{\max}^2\}$ from Section 3. Also recall the definitions of $f_L \coloneqq \exp\big(-2\gamma\varphi_{\max}\big)/b_u$ and $f_U \coloneqq \exp\big(2\gamma\varphi_{\max}\big)/b_l$ from Appendix B.

T.1 Example 1

The following distribution with polynomial sufficient statistics is a special case of density in (2) with $\boldsymbol{\phi}(x) = x$ and $k = 1$. Let $\forall i \in [p]$, $\mathcal{X}_i = [-b, b]$. Therefore $b_l = b_u = 2b$, $\phi_{\max} = b$ and $\bar{\phi}_{\max} = 1$. The density, in this case, is given by

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

For this density, we see that $\gamma = \theta_{\max}(d+1)$ and $\varphi_{\max} = 2\max\{b, b^2\}$. Let us first lower bound the conditional entropy of x_j given x_{-j}.

$$\begin{aligned} h\Big(\mathsf{x}_j \Big| \mathsf{x}_{-j}\Big) &= -\int_{\mathbf{x} \in \mathcal{X}} f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*) \log f_{\mathsf{x}_j}(x_j | \mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) d\mathbf{x} \\ &\overset{(a)}{\geq} -\int_{\mathbf{x} \in \mathcal{X}} f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*) \log(f_U) d\mathbf{x} \\ &\overset{(b)}{=} -\log f_U, \end{aligned} \tag{85}$$

where (a) follows from (17) with $f_U = \exp(4\theta_{\max}(d+1)\max\{b, b^2\})/2b$ and (b) follows because the integral of any density function over its entire domain is 1.

Observing that $\int_{x_i \in \mathcal{X}_i} x_i x_j \mathcal{U}_{\mathcal{X}_i}(x_i) dx_i = 0$, the left-hand-side of Condition 4.1 can be written and simplified as follows:

$$\begin{aligned} \mathbb{E}\Big[\exp\Big\{2h\Big((\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})\mathsf{x}_i\mathsf{x}_j \Big| \mathsf{x}_{-j}\Big)\Big\}\Big] &\overset{(a)}{=} \mathbb{E}\Big[\exp\Big\{2h\Big(\mathsf{x}_j \Big| \mathsf{x}_{-j}\Big) + 2\log\Big|(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})\mathsf{x}_i\Big|\Big\}\Big] \\ &\overset{(b)}{\geq} \mathbb{E}\Big[\exp\Big\{-2\log f_U + 2\log\Big|(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})\mathsf{x}_i\Big|\Big\}\Big] \\ &= \frac{(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})^2}{f_U^2}\mathbb{E}\Big[\mathsf{x}_i^2\Big] \end{aligned}$$

$$\begin{aligned}
&\overset{(c)}{=} \frac{(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})^2}{f_U^2} \mathbb{E}\Big[\mathbb{E}\Big[\mathsf{x}_i^2 | \mathsf{x}_{-i}\Big]\Big] \\
&\overset{(d)}{=} \frac{(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})^2}{f_U^2} \mathbb{E}\Big[\int_{x_i \in \mathcal{X}_i} x_i^2 f_{\mathsf{x}_i}(x_i | \mathsf{x}_{-i} = x_{-i}; \boldsymbol{\vartheta}^{*(i)}) dx_i\Big] \\
&\overset{(e)}{\geq} \frac{f_L(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})^2}{f_U^2} \mathbb{E}\Big[\int_{x_i \in \mathcal{X}_i} x_i^2 dx_i\Big] \\
&\geq \frac{2b^3 f_L}{3 f_U^2} (\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})^2,
\end{aligned}$$

where (a) follows because for a constant a, $h(aX) = h(X) + \log|a|$, (b) follows from (85), (c) follows from the law of total expectation, (d) follows from the definition of conditional expectation, and (e) follows from (17).

Substituting for f_L and f_U, we see this density satisfies Condition 4.1 with $\kappa = \frac{4b^4}{3}\exp(-12\theta_{\max}(d+1)\max\{b, b^2\})$.

T.2 Example 2

The following distribution with harmonic sufficient statistics is a special case of density in (2) with $\boldsymbol{\phi}(x) = \Big(\sin\big(\pi x/b\big), \cos\big(\pi x/b\big)\Big)$ and $k = 2$. Let $\forall i \in [p]$, $\mathcal{X}_i = [-b, b]$. Therefore $b_l = b_u = 2b$, $\phi_{\max} = 1$, and $\bar{\phi}_{\max} = \pi/b$. The density in this case is given by

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

For this density, we see that $\gamma = \theta_{\max}(4d+2)$ and $\varphi_{\max} = 2$. Let $\mathsf{y}_j = \sin\Big(\frac{\pi \mathsf{x}_j}{b} + z\Big)$ where z is a constant with respect to x_j. Then, the conditional density of y_j given x_{-j} can be obtained using the change of variables technique to be as follows:

$$f_{\mathsf{y}_j}(y_j | \mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) = \begin{cases} \dfrac{b\Big[f_{\mathsf{x}_j}\big(\frac{b}{\pi}[\sin^{-1} y_j - z]\big|x_{-j}; \boldsymbol{\vartheta}^{*(j)}\big) + f_{\mathsf{x}_j}\big(-b - \frac{b}{\pi}[\sin^{-1} y_j + z]\big|x_{-j}; \boldsymbol{\vartheta}^{*(j)}\big)\Big]}{\pi\sqrt{1 - y_j^2}} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } y_j \in [-1, 0] \\ \dfrac{b\Big[f_{\mathsf{x}_j}\big(\frac{b}{\pi}[\sin^{-1} y_j - z]\big|x_{-j}; \boldsymbol{\vartheta}^{*(j)}\big) + f_{\mathsf{x}_j}\big(b - \frac{b}{\pi}[\sin^{-1} y_j + z]\big|x_{-j}; \boldsymbol{\vartheta}^{*(j)}\big)\Big]}{\pi\sqrt{1 - y_j^2}} \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{if } y_j \in [0, 1]. \end{cases}$$

Using (17), we can bound the above conditional density as:

$$f_{\mathsf{y}_j}(y_j | \mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) \leq \frac{2bf_U}{\pi\sqrt{1 - y_j^2}}, \tag{86}$$

where $f_U = \exp(4\theta_{\max}(4d+2))/2b$.

Let us now lower bound the conditional entropy of y_j given $\mathsf{x}_{-j} = x_{-j}$.

$$\begin{aligned}
h\Big(\mathsf{y}_j \Big| \mathsf{x}_{-j} = x_{-j}\Big) &= -\int_{y_j=-1}^{y_j=1} f_{\mathsf{y}_j}(y_j | \mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) \log f_{\mathsf{y}_j}(y_j | \mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) dy_j \\
&\overset{(a)}{\geq} -\int_{y_j=-1}^{y_j=1} \frac{2bf_U}{\pi\sqrt{1 - y_j^2}} \log\frac{2bf_U}{\pi\sqrt{1 - y_j^2}} dy_j \\
&= -\frac{2bf_U}{\pi}\Big[\log\frac{2bf_U}{\pi}\int_{y_j=-1}^{y_j=1} \frac{1}{\sqrt{1 - y_j^2}} dy_j - \int_{y_j=-1}^{y_j=1} \frac{1}{\sqrt{1 - y_j^2}} \log\sqrt{1 - y_j^2} dy_j\Big]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(b)}{=} -\frac{2bf_U}{\pi}\left[\pi \log \frac{2bf_U}{\pi} + \pi \log 2\right]\\
&= -2bf_U \log \frac{4bf_U}{\pi},
\end{aligned}
\tag{87}
$$

where (a) follows from (86) and (b) follows from standard definite integrals. Now, we are in a position to lower bound the conditional entropy of y_j given x_{-j}.

$$
\begin{aligned}
h\Big(\mathsf{y}_j \Big| \mathsf{x}_{-j}\Big) &= \int_{x_{-j}\in\prod_{r\neq j}\mathcal{X}_r} f_{\mathsf{x}_{-j}}(x_{-j};\boldsymbol{\theta}^*) h\Big(\mathsf{y}_j \Big| \mathsf{x}_{-j} = x_{-j}\Big) dx_{-j}\\
&\overset{(a)}{\geq} -2bf_U \log \frac{4bf_U}{\pi},
\end{aligned}
\tag{88}
$$

where (a) follows from (87) and because the integral of any density function over its entire domain is 1.

Observe that $\int_{x_i\in\mathcal{X}_i} \sin\big(\frac{\pi(x_i+x_j)}{b}\big)\mathcal{U}_{\mathcal{X}_i}(x_i)dx_i = \int_{x_i\in\mathcal{X}_i} \cos\big(\frac{\pi(x_i+x_j)}{b}\big)\mathcal{U}_{\mathcal{X}_i}(x_i)dx_i = 0$. Letting $\bar{\theta}_1^{(ij)} - \tilde{\theta}_1^{(ij)} = \alpha$ and $\bar{\theta}_2^{(ij)} - \tilde{\theta}_2^{(ij)} = \beta$, the left-hand-side of Condition 4.1 can be written and simplified as follows:

$$
\begin{aligned}
&\mathbb{E}\bigg[\exp\bigg\{2h\Big(\alpha \sin\big(\frac{\pi(\mathsf{x}_i+\mathsf{x}_j)}{b}\big) + \beta\cos\big(\frac{\pi(\mathsf{x}_i+\mathsf{x}_j)}{b}\big)\Big|\mathsf{x}_{-j}\Big)\bigg\}\bigg]\\
&\overset{(a)}{=} \mathbb{E}\bigg[\exp\bigg\{2h\Big(\sqrt{\alpha^2+\beta^2}\sin\Big(\frac{\pi(\mathsf{x}_i+\mathsf{x}_j)}{b} - \tan^{-1}\frac{\beta}{\alpha}\Big)\Big|\mathsf{x}_{-j}\Big)\bigg\}\bigg]\\
&\overset{(b)}{=} \mathbb{E}\bigg[\exp\bigg\{2h\Big(\sin\Big(\frac{\pi(\mathsf{x}_i+\mathsf{x}_j)}{b} - \tan^{-1}\frac{\beta}{\alpha}\Big)\Big|\mathsf{x}_{-j}\Big) + 2\log\Big|\sqrt{\alpha^2+\beta^2}\Big|\bigg\}\bigg]\\
&\overset{(c)}{\geq} \mathbb{E}\bigg[\exp\bigg\{-4bf_U\log\frac{4bf_U}{\pi} + \log\Big|\alpha^2+\beta^2\Big|\bigg\}\bigg]\\
&\overset{(d)}{=} \Big(\frac{\pi}{4bf_U}\Big)^{4bf_U} \times \Big[(\bar{\theta}_1^{(ij)} - \tilde{\theta}_1^{(ij)})^2 + (\bar{\theta}_2^{(ij)} - \tilde{\theta}_2^{(ij)})^2\Big],
\end{aligned}
$$

where (a) follows from standard trigonometric identities, (b) follows because for a constant a, $h(aX) = h(X) + \log|a|$, (c) follows from (88) with $z = \pi\mathsf{x}_i/b - \tan^{-1}\beta/\alpha$, and (d) follows by substituting for α and β.

Substituting for f_L and f_U, we see this density satisfies Condition 4.1 with $\kappa = \Big(\frac{\pi\exp(-4\theta_{\max}(4d+2))}{2}\Big)^{2\exp(4\theta_{\max}(4d+2))}$.

T.3 Example 3

The following distribution with polynomial sufficient statistics is a special case of density in (2) with $\boldsymbol{\phi}(x) = (x, x^2)$, $k = 2$ and with the assumption that the parameters associated with $x_i x_j^2$ and $x_i^2 x_j^2$ are zero $\forall i \in [p], j > i$. Let $\forall i \in [p]$, $\mathcal{X}_i = [-b, b]$. Therefore $b_l = b_u = 2b, \phi_{\max} = \max\{b, b^2\}$, and $\bar{\phi}_{\max} = \max\{1, 2b\}$. The density in this case is given by

$$
f_{\mathbf{x}}(\mathbf{x};\boldsymbol{\theta}^*) \propto \exp\bigg(\sum_{i\in[p]}[\theta_1^{*(i)}x_i + \theta_2^{*(i)}x_i^2] + \sum_{i\in[p]}\sum_{j>i}[\theta_{1,1}^{*(ij)}x_i x_j + \theta_{2,1}^{*(ij)}x_i^2 x_j]\bigg).
$$

For this density, we see that $\gamma = \theta_{\max}(4d+2)$ and $\varphi_{\max} = 2\max\{b, b^4\}$. As in Appendix T.1, we have the following lower bound on the conditional entropy of x_j given x_{-j}

$$
h\Big(\mathsf{x}_j\Big|\mathsf{x}_{-j}\Big) \geq -\log f_U,
\tag{89}
$$

where $f_U = \exp(4\theta_{\max}(4d+2)\max\{b, b^4\})/2b$.

Observing that $\int_{x_i\in\mathcal{X}_i} x_i x_j \mathcal{U}_{\mathcal{X}_i}(x_i)dx_i = 0$ and $\int_{x_i\in\mathcal{X}_i} x_i^2 x_j \mathcal{U}_{\mathcal{X}_i}(x_i)dx_i = \frac{b^2}{3}x_j$, the left-hand-side of Condition 4.1 can be written and simplified as follows:

$$
\mathbb{E}\bigg[\exp\bigg\{2h\Big((\bar{\theta}_{1,1}^{(ij)} - \tilde{\theta}_{2,1}^{(ij)})\mathsf{x}_i\mathsf{x}_j + (\bar{\theta}_{2,1}^{(ij)} - \tilde{\theta}_{2,1}^{(ij)})\Big(\mathsf{x}_i^2\mathsf{x}_j - \frac{b^2}{3}\mathsf{x}_j\Big)\Big|\mathsf{x}_{-j}\Big)\bigg\}\bigg]
$$

$$
\begin{aligned}
&= \mathbb{E}\bigg[\exp\bigg\{2h\bigg(\bigg[(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})\mathsf{x}_i + (\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})\Big(\mathsf{x}_i^2 - \frac{b^2}{3}\Big)\bigg]\mathsf{x}_j \bigg| \mathsf{x}_{-j}\bigg)\bigg\}\bigg] \\
&\overset{(a)}{=} \mathbb{E}\bigg[\exp\bigg\{2h\bigg(\mathsf{x}_j \bigg| \mathsf{x}_{-j}\bigg) + 2\log\bigg|(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})\mathsf{x}_i + (\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})\Big(\mathsf{x}_i^2 - \frac{b^2}{3}\Big)\bigg|\bigg\}\bigg] \\
&\overset{(b)}{\geq} \mathbb{E}\bigg[\exp\bigg\{-2\log f_U + 2\log\bigg|(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})\mathsf{x}_i + (\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})\Big(\mathsf{x}_i^2 - \frac{b^2}{3}\Big)\bigg|\bigg\}\bigg] \\
&= \frac{1}{f_U^2}\mathbb{E}\bigg[\bigg((\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})\mathsf{x}_i + (\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})\Big(\mathsf{x}_i^2 - \frac{b^2}{3}\Big)\bigg)^2\bigg] \\
&\overset{(c)}{=} \frac{1}{f_U^2}\mathbb{E}\bigg[\mathbb{E}\bigg[\bigg((\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})\mathsf{x}_i + (\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})\Big(\mathsf{x}_i^2 - \frac{b^2}{3}\Big)\bigg)^2 \bigg| \mathsf{x}_{-i}\bigg]\bigg] \\
&\overset{(d)}{\geq} \frac{f_L}{f_U^2}\mathbb{E}\bigg[(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})^2\int_{x_i\in\mathcal{X}_i} x_i^2 dx_i + (\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})^2\bigg(\int_{x_i\in\mathcal{X}_i}\Big[x_i^4 + \frac{b^4}{9} - \frac{2b^2}{3}x_i^2\Big]dx_i\bigg) \\
&\qquad + 2(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})(\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})\bigg(\int_{x_i\in\mathcal{X}_i}\Big[x_i^3 - \frac{b^2}{3}x_i\Big]dx_i\bigg)\bigg] \\
&= \frac{f_L}{f_U^2}\bigg[\frac{2b^3}{3}(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})^2 + \bigg(\frac{2b^5}{5} + \frac{2b^5}{9} - \frac{4b^5}{9}\bigg)(\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})^2\bigg] \\
&= \frac{f_L}{f_U^2}\bigg[\frac{2b^3}{3}(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})^2 + \frac{8b^5}{45}(\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})^2\bigg] \\
&\geq \frac{8f_L b^3\min\{45/12, b^2\}}{45 f_U^2}\bigg[(\bar{\theta}^{(ij)}_{1,1} - \tilde{\theta}^{(ij)}_{2,1})^2 + (\bar{\theta}^{(ij)}_{2,1} - \tilde{\theta}^{(ij)}_{2,1})^2\bigg],
\end{aligned}
$$

where (a) follows because for a constant a, $h(aX) = h(X) + \log|a|$, (b) follows from (89), (c) follows from the law of total expectation, and (d) follows from the definition of conditional expectation and (17).

Substituting for f_L and f_U, we see this density satisfies Condition 4.1 with $\kappa = \frac{16b^4\min\{45/12, b^2\}}{45}\exp(-12\theta_{\max}(4d+2)\max\{b, b^4\})$.

T.4 Example 4

The following distribution with polynomial sufficient statistics is a special case of density in (2) with $\boldsymbol{\phi}(x) = (x, x^2)$, $k = 2$ and with the assumption that the parameters associated with $x_i x_j$, $x_i^2 x_j$, $x_i x_j^2$ and $x_i^2 x_j^2$ are same $\forall i \in [p], j > i$. Let $\forall i \in [p]$, $\mathcal{X}_i = [-b, b]$. Therefore $b_l = b_u = 2b$, $\phi_{\max} = \max\{b, b^2\}$, and $\bar{\phi}_{\max} = \max\{1, 2b\}$. The density in this case is given by

$$
f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*) \propto \exp\bigg(\sum_{i\in[p]}[\theta_1^{*(i)}x_i + \theta_2^{*(i)}x_i^2] + \sum_{\substack{i\in[p]\\ j>i}}\theta^{*(ij)}(x_i + x_i^2)(x_j + x_j^2)\bigg).
$$

For this density, we see that $\gamma = \theta_{\max}(4d+2)$ and $\varphi_{\max} = 2\max\{b, b^4\}$. Let $\mathsf{y}_j = \mathsf{x}_j + \mathsf{x}_j^2$. It is easy to obtain the range of y_j as follows:

$$
\mathsf{y}_j \in \mathcal{Y} := \begin{cases} [-1/4, b + b^2] & \text{if} \quad b \geq 1/2 \\ [b - b^2, b + b^2] & \text{if} \quad b < 1/2. \end{cases}
$$

We obtain the conditional density of y_j given x_{-j} using the change of variables technique and upper bound it using (17) as follows:

$$
f_{\mathsf{y}_j}(y_j|\mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) \leq \frac{2f_U}{\sqrt{1 + 4y_j}}, \tag{90}
$$

where $f_U = \exp(4\theta_{\max}(4d+2)\max\{b, b^4\})/2b$.

We will now lower bound the conditional entropy of y_j given $\mathsf{x}_{-j} = x_{-j}$. In the first scenario, let $b \geq 1/2$.

$$
h\bigg(\mathsf{y}_j \bigg| \mathsf{x}_{-j} = x_{-j}\bigg) = -\int_{y_j=-1/4}^{y_j=b+b^2} f_{\mathsf{y}_j}(y_j|\mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)})\log f_{\mathsf{y}_j}(y_j|\mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)})dy_j
$$

$$
\begin{aligned}
&\overset{(a)}{\geq} -\int_{y_j=-1/4}^{y_j=b+b^2} \frac{2f_U}{\sqrt{1+4y_j}} \log \frac{2f_U}{\sqrt{1+4y_j}} dy_j \\
&\overset{(b)}{=} -f_U\sqrt{1+4y_j} \log \frac{2f_U e}{\sqrt{1+4y_j}}\Bigg|_{-1/4}^{b+b^2} \\
&= -(1+2b) f_U \log \frac{2f_U e}{1+2b}, \qquad (91)
\end{aligned}
$$

where (a) follows from (90) and (b) follows from standard indefinite integrals. Now, we will lower bound the conditional entropy of y_j given $\mathsf{x}_{-j} = x_{-j}$ and $b < 1/2$.

$$
\begin{aligned}
h\bigg(\mathsf{y}_j\bigg|\mathsf{x}_{-j} = x_{-j}\bigg) &= -\int_{y_j=b-b^2}^{y_j=b+b^2} f_{\mathsf{y}_j}(y_j|\mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) \log f_{\mathsf{y}_j}(y_j|\mathsf{x}_{-j} = x_{-j}; \boldsymbol{\vartheta}^{*(j)}) dy_j \\
&\overset{(a)}{\geq} -\int_{y_j=b-b^2}^{y_j=b+b^2} \frac{2f_U}{\sqrt{1+4y_j}} \log \frac{2f_U}{\sqrt{1+4y_j}} dy_j \\
&\overset{(b)}{=} -f_U\sqrt{1+4y_j} \log \frac{2f_U e}{\sqrt{1+4y_j}}\Bigg|_{b-b^2}^{b+b^2} \\
&= -(1+2b) f_U \log \frac{2f_U e}{1+2b} + (1-2b) f_U \log \frac{2f_U e}{1-2b} \\
&\overset{(c)}{\geq} -(1+2b) f_U \log \frac{2f_U e}{1+2b}, \qquad (92)
\end{aligned}
$$

where (a) follows from (90), (b) follows from standard indefinite integrals and (c) follows because $(1-2b)\log\frac{2f_U e}{1-2b} > 0$ when $b < 1/2$. Now, we are in a position to lower bound the conditional entropy of y_j given x_{-j}.

$$
\begin{aligned}
h\bigg(\mathsf{y}_j\bigg|\mathsf{x}_{-j}\bigg) &\int_{x_{-j}\in\prod_{r\neq j}\mathcal{X}_r} f_{\mathsf{x}_{-j}}(x_{-j}; \boldsymbol{\theta}^*) h\bigg(\mathsf{y}_j\bigg|\mathsf{x}_{-j} = x_{-j}\bigg) dx_{-j} \\
&\overset{(a)}{\geq} -(1+2b) f_U \log \frac{2f_U e}{1+2b}, \qquad (93)
\end{aligned}
$$

where (a) follows from (91), (92), and because the integral of any density function over its entire domain is 1.

Observing that $\int_{x_i\in\mathcal{X}_i} x_i \mathcal{U}_{\mathcal{X}_i}(x_i) dx_i = 0$ and $\int_{x_i\in\mathcal{X}_i} x_i^2 \mathcal{U}_{\mathcal{X}_i}(x_i) dx_i = \frac{b^2}{3}$, the left-hand-side of Condition 4.1 can be written and simplified as follows:

$$
\begin{aligned}
&\mathbb{E}\bigg[\exp\bigg\{2h\bigg(\Big(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)}\Big)\Big(\mathsf{x}_i + \mathsf{x}_i^2 - \frac{b^2}{3}\Big)\Big(\mathsf{x}_j + \mathsf{x}_j^2\Big)\bigg|\mathsf{x}_{-j}\bigg)\bigg\}\bigg] \\
&\quad \overset{(a)}{=} \mathbb{E}\bigg[\exp\bigg\{2h\bigg(\mathsf{x}_j + \mathsf{x}_j^2\bigg|\mathsf{x}_{-j}\bigg) + 2\log\bigg|\Big(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)}\Big)\Big(\mathsf{x}_i + \mathsf{x}_i^2 - \frac{b^2}{3}\Big)\bigg|\bigg\}\bigg] \\
&\quad \overset{(b)}{\geq} \mathbb{E}\bigg[\exp\bigg\{-2(1+2b) f_U \log \frac{2f_U e}{1+2b} + 2\log\bigg|\Big(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)}\Big)\Big(\mathsf{x}_i + \mathsf{x}_i^2 - \frac{b^2}{3}\Big)\bigg|\bigg\}\bigg] \\
&\quad = \Big(\frac{1+2b}{2f_U e}\Big)^{2f_U(1+2b)} \mathbb{E}\bigg[\Big(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)}\Big)^2\Big(\mathsf{x}_i + \mathsf{x}_i^2 - \frac{b^2}{3}\Big)^2\bigg] \\
&\quad \overset{(c)}{=} f_L(\bar{\theta}^{(ij)} - \tilde{\theta}^{(ij)})^2 \Big(\frac{2b^3}{3} + \frac{8b^5}{45}\Big)\Big(\frac{1+2b}{2f_U e}\Big)^{2f_U(1+2b)},
\end{aligned}
$$

where (a) follows because for a constant a, $h(aX) = h(X) + \log|a|$, (b) follows from (93), (c) follows from steps similar to the ones in Appendix T.3.

Substituting for f_L and f_U, we see this density satisfies Condition 4.1 with $\kappa = \frac{e(15b+4b^3)}{45(1+2b)}\Big(\frac{b(1+2b)\exp(-4\theta_{\max}(4d+2)\max\{b,b^4\})}{e}\Big)^{\frac{1+2b}{b}\exp(4\theta_{\max}(4d+2)\max\{b,b^4\})+1}$.

U Discussions

In this appendix, we discuss the invertibility of the cross-covariance matrix $B(\boldsymbol{\vartheta}^{*(i)})$ via a simple example as well as explicitly show that the matrix $B(\boldsymbol{\vartheta}^{*(i)})^{-1}A(\boldsymbol{\vartheta}^{*(i)})B(\boldsymbol{\vartheta}^{*(i)})^{-1}$ need not be equal to the inverse of the Fisher information matrix of $\mathbf{x}$. This concludes that even though the estimator $\hat{\boldsymbol{\vartheta}}_n^{(i)}$ is asymptotically normal, it is not asymptotically efficient. Finally, we also provide a brief discussion on the assumption of the minimality of the exponential family.

U.1 Invertibility of the cross-covariance matrix

We will look at the special case of $\boldsymbol{\phi}(x) = x$ and $k = 1$ and show that the cross-covariance matrix of $\boldsymbol{\varphi}^{(i)}(\mathbf{x})$ and $\boldsymbol{\varphi}^{(i)}(\mathbf{x}) \exp\big(- \boldsymbol{\vartheta}^{*(i)^T} \boldsymbol{\varphi}^{(i)}(\mathbf{x})\big)$ i.e., $B(\boldsymbol{\vartheta}^{*(i)})$ is invertible $\forall i \in [p]$ when $p = 2$. Let $\mathcal{X}_1 = [-b, b]$ and $\mathcal{X}_2 = [-b, b]$.

The density in this special case is as follows.

$$f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta}^*) \propto \exp\bigg(\theta^{*(1)}x_1 + \theta^{*(2)}x_2 + \theta^{*(12)}x_1x_2\bigg). \tag{94}$$

It is easy to see that the basis functions x_1, x_2, and x_1x_2 are already locally centered. Therefore, we have

$$\begin{aligned} \boldsymbol{\varphi}^{(1)}(\mathbf{x}) = (x_1, x_1x_2) \qquad &\text{and} \qquad \boldsymbol{\varphi}^{(2)}(\mathbf{x}) = (x_2, x_1x_2) \\ \boldsymbol{\vartheta}^{*(1)} = (\theta^{*(1)}, \theta^{*(12)}) \qquad &\text{and} \qquad \boldsymbol{\vartheta}^{*(2)} = (\theta^{*(2)}, \theta^{*(12)}). \end{aligned}$$

Then, the cross-covariance matrices $B(\boldsymbol{\vartheta}^{*(1)})$ and $B(\boldsymbol{\vartheta}^{*(2)})$ are

$$B(\boldsymbol{\vartheta}^{*(1)}) = \begin{bmatrix} \mathbb{E}[x_1^2 \exp{(-\theta^{*(1)}x_1 - \theta^{*(12)}x_1x_2)}] & \mathbb{E}[x_1^2x_2 \exp{(-\theta^{*(1)}x_1 - \theta^{*(12)}x_1x_2)}] \\ \mathbb{E}[x_1^2x_2 \exp{(-\theta^{*(1)}x_1 - \theta^{*(12)}x_1x_2)}] & \mathbb{E}[x_1^2x_2^2 \exp{(-\theta^{*(1)}x_1 - \theta^{*(12)}x_1x_2)}] \end{bmatrix}$$

and

$$B(\boldsymbol{\vartheta}^{*(2)}) = \begin{bmatrix} \mathbb{E}[x_2^2 \exp{(-\theta^{*(2)}x_2 - \theta^{*(12)}x_1x_2)}] & \mathbb{E}[x_2^2x_1 \exp{(-\theta^{*(2)}x_2 - \theta^{*(12)}x_1x_2)}] \\ \mathbb{E}[x_2^2x_1 \exp{(-\theta^{*(2)}x_2 - \theta^{*(12)}x_1x_2)}] & \mathbb{E}[x_2^2x_1^2 \exp{(-\theta^{*(2)}x_2 - \theta^{*(12)}x_1x_2)}] \end{bmatrix}.$$

Using the Cauchy-Schwarz inequality, for random variables M and N, we have

$$[\mathbb{E}(MN)]^2 \leq \mathbb{E}(M^2)\mathbb{E}(N^2),$$

with equality only if M and N are linearly dependent. Using the Cauchy-Schwarz inequality with $M = x_1 \exp(-0.5\theta^{*(1)}x_1 - 0.5\theta^{*(12)}x_1x_2)$, $N = x_1x_2 \exp(-0.5\theta^{*(1)}x_1 - 0.5\theta^{*(12)}x_1x_2)$ and observing that M and N are not linearly dependent (because x_2 is a random variable), we have invertibility of $B(\boldsymbol{\vartheta}^{*(1)})$. Similarly, using the Cauchy-Schwarz inequality with $M = x_2 \exp(-0.5\theta^{*(2)}x_2 - 0.5\theta^{*(12)}x_1x_2)$, $N = x_2x_1 \exp(-0.5\theta^{*(2)}x_2 - 0.5\theta^{*(12)}x_1x_2)$ and observing that M and N are not linearly dependent (because x_1 is a random variable), we have invertibility of $B(\boldsymbol{\vartheta}^{*(2)})$.

U.2 Fisher information matrix

Let $J(\boldsymbol{\vartheta}^{*(i)})$ denote the Fisher information matrix of $\mathbf{x}$ with respect to node i. For any $l \in [k + k^2(p-1)]$, let $\boldsymbol{\varphi}_l^{(i)}(\mathbf{x})$ denote the l^{th} component of $\boldsymbol{\varphi}^{(i)}(\mathbf{x})$. Using the fact that for any regular exponential family the Hessian of the log partition function is the covariance matrix of the associated sufficient statistic vector, we have the Fisher information matrix

$$\big[J(\boldsymbol{\vartheta}^{*(i)})\big]_{l_1,l_2} = \text{Cov}\big(\boldsymbol{\varphi}_{l_1}^{(i)}(\mathbf{x}), \boldsymbol{\varphi}_{l_2}^{(i)}(\mathbf{x})\big).$$

Consider the density in (94) with $b = 1, \theta^{*(1)} = \theta^{*(2)} = 0$ and $\theta^{*(12)} = 1$. We will evaluate the matrices $B(\boldsymbol{\vartheta}^{*(1)})$, $A(\boldsymbol{\vartheta}^{*(1)})$, and $J(\boldsymbol{\vartheta}^{*(1)})$. We have

$$B(\boldsymbol{\vartheta}^{*(1)}) = \begin{bmatrix} \mathbb{E}[x_1^2 \exp{(-x_1x_2)}] & \mathbb{E}[x_1^2x_2 \exp{(-x_1x_2)}] \\ \mathbb{E}[x_1^2x_2 \exp{(-x_1x_2)}] & \mathbb{E}[x_1^2x_2^2 \exp{(-x_1x_2)}] \end{bmatrix} = \begin{bmatrix} \frac{1}{3\text{Shi}(1)} & 0 \\ 0 & \frac{1}{9\text{Shi}(1)} \end{bmatrix}$$

$$A(\boldsymbol{\vartheta}^{*(1)}) = \begin{bmatrix} \mathbb{E}[x_1^2 \exp{(-2x_1x_2)}] & \mathbb{E}[x_1^2 x_2 \exp{(-2x_1x_2)}] \\ \mathbb{E}[x_1^2 x_2 \exp{(-2x_1x_2)}] & \mathbb{E}[x_1^2 x_2^2 \exp{(-2x_1x_2)}] \end{bmatrix} = \begin{bmatrix} \frac{1}{e\mathrm{Shi}(1)} & 0 \\ 0 & \frac{2\mathrm{Shi}(1)+2/e-e}{\mathrm{Shi}(1)} \end{bmatrix}$$

$$J(\boldsymbol{\vartheta}^{*(1)}) = \begin{bmatrix} \mathrm{Cov}(x_1, x_1) & \mathrm{Cov}(x_1, x_1x_2) \\ \mathrm{Cov}(x_1, x_1x_2) & \mathrm{Cov}(x_1x_2, x_1x_2) \end{bmatrix} = \begin{bmatrix} \frac{1}{e\mathrm{Shi}(1)} & 0 \\ 0 & \frac{2\mathrm{Shi}(1)+2/e-e}{\mathrm{Shi}(1)} - \left[\frac{\sinh(1)}{\mathrm{Shi}(1)} - 1\right]^2 \end{bmatrix},$$

where sinh is the hyperbolic sine function and Shi is the hyperbolic sine integral function. Plugging in the values of $\mathrm{Shi}(1), \sinh(1)$, and e, we have (upto two decimals)

$$B(\boldsymbol{\vartheta}^{*(1)})^{-1} A(\boldsymbol{\vartheta}^{*(1)}) B(\boldsymbol{\vartheta}^{*(1)})^{-1} = \begin{bmatrix} 3.50 & 0 \\ 0 & 11.30 \end{bmatrix} \neq J^{-1}(\boldsymbol{\vartheta}^{*(1)}) = \begin{bmatrix} 3.007 & 0 \\ 0 & 8.90 \end{bmatrix}.$$

U.3 Discussion on minimality of the exponential family

Minimality of the exponential family is used in Theorem 4.1 to ensure a unique minimizer of $\mathcal{S}^{(i)}(\boldsymbol{\vartheta})$. This uniqueness effectively leads to the consistency and the normality of $\hat{\boldsymbol{\vartheta}}_n^{(i)}$ via Theorem 4.2. Even though a milder condition on the exponential family might suffice, assuming minimality of the exponential family does not seem very restrictive in the continuous setting.

Consider $f_{x_1,x_2}(x_1, x_2) \propto \exp{(\theta x_1 x_2)}$ with $x_1, x_2 \in [-1, 1]$ as an example. $f_{\mathbf{x}}(\mathbf{x})$ is minimal as there does not exist a non-zero θ such that $\theta x_1 x_2$ is a constant (almost everywhere). The same also seems to hold more generally.