LA-UR-20-25018
Approved for public release; distribution is unlimited.

Title: Efficient Learning of Discrete Graphical Models

Author(s): Vuffray, Marc Denis
Misra, Sidhant
Lokhov, Andrey

Intended for: NeurIPS 2020, 2020-12-05 (Online, New Mexico, United States)
Report

Issued: 2021-01-29 (rev.1)

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Efficient Learning of Discrete Graphical Models

Marc Vuffray, Sidhant Misra, Andrey Y. Lokhov
Theoretical Division,
Los Alamos National Laboratory, USA
{vuffray, sidhant, lokhov}@lanl.gov

Abstract

Graphical models are useful tools for describing structured high-dimensional probability distributions. Development of efficient algorithms for learning graphical models with least amount of data remains an active research topic. Reconstruction of graphical models that describe the statistics of *discrete* variables is a particularly challenging problem, for which the maximum likelihood approach is intractable. In this work, we provide the first sample-efficient method based on the *Interaction Screening* framework that allows one to provably learn fully general discrete factor models with node-specific discrete alphabets and multi-body interactions, specified in an arbitrary basis. We identify a single condition related to model parametrization that leads to rigorous guarantees on the recovery of model structure and parameters in any error norm, and is readily verifiable for a large class of models. Importantly, our bounds make explicit distinction between parameters that are proper to the model and priors used as an input to the algorithm. Finally, we show that the Interaction Screening framework includes all models previously considered in the literature as special cases, and for which our analysis shows a systematic improvement in sample complexity.

1 Introduction

Representing and understanding the structure of direct correlations between distinct random variables with graphical models is a fundamental task that is essential to scientific and engineering endeavors. It is the first step towards an understanding of interactions between interleaved constituents of elaborated systems [11]; it is key for developing causal theories [6]; and it is at the core of automated decision making [8], cybersecurity [5] and artificial intelligence [19].

The problem of reconstruction of graphical models from samples traces back to the seminal work of [7] for tree-structured graphical models, and as of today is still at the center of attention of the learning community. For factor models defined over general hypergraphs, the learning problem is particularly challenging in graphical models over discrete variables, for which the maximum likelihood estimator is in general computationally intractable. One of the earlier tractable algorithms that has been suggested to provably reconstruct the structure of a subset of pairwise binary graphical models is based on inferring the sparsity pattern of the so-called regularized pseudo-likelihood estimator, equivalent to regularized logistic regression in the binary case [16]. However, additional assumptions required for this algorithm to succeed severely limit the set of pairwise binary models that can be learned [15]. After it was proven that reconstruction of any discrete graphical models with bounded degree can be done in polynomial time in the system size [4], Bresler showed that it is possible to bring the computational complexity down to quasi-quadratic in the number of variables for Ising models (pairwise graphical models over binary variables); however, the resulting algorithm has non-optimal sample requirements that are double-exponential in other model parameters [3]. The first computationally efficient reconstruction algorithm for sparse pairwise binary graphical models with a near-optimal sample complexity with respect to the information theoretic lower bound [17],

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

called RISE, was designed and analyzed in [18]. The algorithm RISE suggested in this work is based on the minimization of a novel local convex loss function, called the *Interaction Screening* objective, supplemented with an ℓ_1 penalty to promote sparsity. Even though it has been later shown in [13] that regularized pseudo-likelihood supplemented with a crucial post-processing step also leads to a structure estimator for pairwise binary models, strong numerical and theoretical evidence provided in that work demonstrated that RISE is superior in terms of worst-case sample complexity.

Algorithms for learning discrete graphical models beyond pairwise and binary alphabets have been proposed only recently in [9] and [12]. The method in [9] works for arbitrary models with bounded degrees, but being a generalization of Bresler's algorithm for Ising models [3], it suffers from similar prohibitive sample requirements growing double-exponentially in the strength of model parameters. The so-called SPARSITRON algorithm in [12] has the flavor of a stochastic first order method with multiplicative updates. It has a low computational complexity and is sample-efficient for structure recovery of two subclasses of discrete graphical models: multiwise graphical models over binary variables or pairwise models with general alphabets. A recent follow-up work [20] considered an $\ell_{2,1}$ constrained logistic regression, and showed that it provides a slight improvement of the sample complexity compared to [12] in the case of pairwise models over non-binary variables.

In this work, we propose a general framework for learning general discrete factor models expressed in an arbitrary parametric form. Our estimator termed GRISE is based on a significant generalization of the Interaction Screening method of [18, 13], previously introduced for pairwise binary models. Our primary insight lies in the identification of a single general condition related to model parameterization that is sufficient to obtain bounds on sample complexity. We show that this condition can be reduced to a set of local identifiability conditions that only depend on the size of the maximal clique of the factor graph and can be explicitly verified in an efficient way. We propose an iterative algorithm called SUPRISE which is based on GRISE and show that it can efficiently perform structure and parameter estimation for arbitrary graphical models. Existing results in the literature on this topic [18, 9, 12, 20] can be obtained as special cases of our general reconstruction results, which noticeably includes the challenging case of multi-body interactions defined over general discrete alphabets. Our theoretical guarantees can be expressed in any error norm, and explicitly includes distinction between bounds on the parameters of the underlying model and the prior parameters used in the optimization; as a result prior information that is not tight only has moderate effect on the sample complexity bounds. Finally, we also provide a fully parallelizable algorithmic formulation for the GRISE estimator and SUPRISE algorithm, and show that they have efficient run times of $\widetilde{O}(p^L)$ for a model of size p with L-order interactions, that includes the best-known $\widetilde{O}(p^2)$ scaling for pairwise models.

2 Problem formulation

In this Section, we formulate the general discrete graphical model selection problem that we consider and describe conditions that makes this problem well-posed.

2.1 Parameterized family of models

We consider positive joint probability distributions over p variables $\sigma_i \in \mathcal{A}_i$ for $i = 1, \ldots, p$. The set of variable indices i is referred to as vertices $\mathcal{V} = 1, \ldots, p$. Node-dependent alphabets $\mathcal{A}_i$ are assumed to be discrete and of size bounded by $q > 0$. Without loss of generality, the positive probability distribution over the p-dimensional vector $\underline{\sigma}$ can be expressed as

$$\mu(\underline{\sigma}) = \frac{1}{Z} \exp\left(\sum_{k \in \mathcal{K}} \theta_k^* f_k(\underline{\sigma}_k) \right), \tag{1}$$

where $\{f_k,\ k \in \mathcal{K}\}$ is a set of *basis functions* acting upon subsets of variables $\underline{\sigma}_k \subseteq \underline{\sigma}$ that specify a family of distributions and θ_k^* are parameters that specify a model within this family. The quantity Z denotes the partition function and serves as a normalization constant that enforces that the μ in (1) is a probability distribution. For $i \in \{1, \ldots, p\}$, let $\mathcal{K}_i \subseteq \mathcal{K}$ denote the set of factors corresponding to basis functions acting upon subsets $\underline{\sigma}_k$ that contain the variable σ_i and $|\mathcal{K}_i| = \mathbf{K}_i$.

Given any set of basis functions, we can locally center them by first defining for a given $i \in [p]$, the *local centering functions*

$$\phi_{ik}(\underline{\sigma}_{k\setminus i}) := \frac{1}{|\mathcal{A}_i|} \sum_{\sigma_i \in \mathcal{A}_i} f_k(\underline{\sigma}_k), \tag{2}$$

where $\underline{\sigma}_{k\setminus i}$ denotes the vector $\underline{\sigma}_k$ without σ_i, and define the *locally centered basis functions*,

$$g_{ik}(\underline{\sigma}_k) = f_k(\underline{\sigma}_k) - \phi_{ik}(\underline{\sigma}_{k\setminus i}). \tag{3}$$

As their name suggests, the locally centered basis functions sum to zero $\sum_{\sigma_i \in \mathcal{A}_i} g_{ik}(\underline{\sigma}_k) = 0$. To ensure the scales of the parameters are well defined, we assume that θ^*_k are chosen or rescaled such that all locally centered basis functions are *normalized* in the following sense:

$$\max_{\underline{\sigma}_k} |g_{ik}(\underline{\sigma}_k)| \leq 1, \tag{4}$$

for all vertices $i \in \mathcal{V}$ and basis factor $k \in \mathcal{K}_i$. This normalization can always be achieved by choosing bounded basis functions $|f_k(\underline{\sigma}_k)| \leq 1/2$. An important special case is when the basis functions are already centered, i.e. $g_{ik}(\underline{\sigma}_k) = f_k(\underline{\sigma}_k)$. In this case the basis functions are directly normalized $\max_{\underline{\sigma}_k} |f_k(\underline{\sigma}_k)| = 1$. Note that one of the reasons to define the normalization in (4) in terms of the centered functions g_k instead of f_k is to avoid spurious cases where the functions f_k have inflated magnitudes due to addition of constants $f_k \leftarrow f_k + C$. In Section A of the Supplementary Material, we show that the other important reason to employ centered functions is that degeneracy of the local parameterization with these functions translates to degeneracy of the global distribution in Eq. (1).

2.2 Model selection problem

For each $i \in [p]$, let $\mathcal{T}_i \subseteq \mathcal{K}_i$ denote the set of *target factors* that we aim at reconstructing accurately and let $\mathcal{R}_i = \mathcal{K}_i \setminus \mathcal{T}_i$ be the set of *residual factors* for which we do not need learning guarantees. The target and residual parameters are defined similarly as $\underline{\theta}^*_{\mathcal{T}_i} = \{\theta^*_k \mid k \in \mathcal{T}_i\}$ and $\underline{\theta}^*_{\mathcal{R}_i} = \{\theta^*_k \mid k \in \mathcal{R}_i\}$ respectively. Given independent samples from a model in the family in Section 2.1, the goal of the model selection problem is to reconstruct the target parameters of the model.

Definition 1 (Model Selection Problem)**.** *Given n i.i.d. samples $\underline{\sigma}^{(1)}, \ldots, \underline{\sigma}^{(n)}$ drawn from some distribution $\mu(\underline{\sigma})$ in Eq.* (1) *defined by $\underline{\theta}^*$, and prior information on $\underline{\theta}^*$ given in form of an upper bound on the ℓ_1-norm of the local sub-components*

$$\|\underline{\theta}^*_i\|_1 \leq \widehat{\gamma}, \tag{5}$$

and a local constraint set $\mathcal{Y}_i \subseteq \mathbb{R}^{\mathbf{K}_i}$ *for each $i \in [p]$ such that*

$$\underline{\theta}^*_i \in \mathcal{Y}_i, \tag{6}$$

compute estimates $\widehat{\underline{\theta}}$ of $\underline{\theta}^*$ such that the estimates of the target parameters satisfy

$$\|\widehat{\underline{\theta}}_{\mathcal{T}_i} - \underline{\theta}^*_{\mathcal{T}_i}\| \leq \frac{\alpha}{2}, \quad \forall i \in [p], \tag{7}$$

where $\|\cdot\|$ denotes some norm of interest with respect to which the error is measured.

The bound on the ℓ_1-norm in (5) is a natural generalization of the sparse case where $\underline{\theta}^*$ only has a small number of non-zero components; in the context of parameter estimation in graphical models, the setting of parameters bounded in the ℓ_1-norm has been previously considered in [12]. The constraint sets $\mathcal{Y}_i$ are used to encode any other side information that may be known about the model.

2.3 Sufficient conditions for well-posedness

We describe some conditions on the model in (1) that makes the model selection problem in Definition 1 well-posed. We first state the conditions formally.

Condition 1. *The model from which the samples are drawn in the model selection problem in Definition 1 satisfies the following:*

(C1) ***Local Learnability Condition for Graphical Models:*** *There exists a constant $\rho_i > 0$ such that for every vertex i and any vector in the perturbation set $\underline{x} \in \mathcal{X}_i \subseteq \mathbb{R}^{\mathbf{K}_i}$ defined as*

$$\mathcal{X}_i = \{\underline{x} = \underline{y}_1 - \underline{y}_2 \mid \underline{y}_1, \underline{y}_2 \in \mathcal{Y}_i, \|\underline{y}_1\|_1 \leq \widehat{\gamma}, \|\underline{y}_2\|_1 \leq \widehat{\gamma}\}, \tag{8}$$

the following holds:

$$\mathbb{E}\left[\left(\sum_{k \in \mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] \geq \rho_i \|\underline{x}_{\mathcal{T}_i}\|^2, \tag{9}$$

where $\underline{x}_{\mathcal{T}_i}$ denotes the components $k \in \mathcal{T}_i$ of x, and $\|\cdot\|$ is the norm used in Definition 1.

(C2) ***Finite Maximum Interaction Strength:*** *The following quantity γ is finite,*

$$\gamma = \max_{i \in \mathcal{V}} |\max_{\underline{\sigma}} \sum_{k \in \mathcal{K}_i} \theta^*_k g_{ik}(\underline{\sigma}_k)| < \infty. \tag{10}$$

Condition (C1) consists in satisfying the inequality in Eq. (9) involving a quadratic form $\underline{x}^\top \widetilde{I} \underline{x}$ where the matrix $\widetilde{I}$ has indices $k, k' \in \mathcal{K}_i$ and is explicitly defined as $\widetilde{I}_{k,k'} = \mathbb{E}\left[g_{ik}(\underline{\sigma}_k) g_{ik'}(\underline{\sigma}_{k'})\right]$. This matrix $\widetilde{I}$ is in fact related to the *conditional* Fisher information matrix.

The conditional Fisher information matrix I with indices $k, k' \in \mathcal{K}_i$ is derived from the conditional distribution of σ_i given the remaining variables and reads,

$$I_{k,k'} = \mathbb{E}\left[g_{ik}(\underline{\sigma}_k) g_{ik'}(\underline{\sigma}_{k'})\right] - \mathbb{E}_{(\underline{\sigma}_{\setminus i})}\left[\mathbb{E}_{(\sigma_i|\underline{\sigma}_{\setminus i})}\left[g_{ik}(\underline{\sigma}_k)\right] \mathbb{E}_{(\sigma_i|\underline{\sigma}_{\setminus i})}\left[g_{ik'}(\underline{\sigma}_{k'})\right]\right]. \tag{11}$$

We immediately see that the matrix $\widetilde{I}$ dominates the conditional Fisher information matrix in the positive semi-definite sense, that is $\underline{x}^\top \widetilde{I}(\theta^*) \underline{x} \geq \underline{x}^\top I(\theta^*) \underline{x}$ for all $\underline{x} \in \mathbb{R}^{\mathbf{K}_i}$. Therefore, Condition *(C1)* is satisfied whenever the conditional Fisher information matrix is non-singular in the parameter subspace $\underline{x}_{\mathcal{T}_i}$ that we care to reconstruct and which is compatible with our priors, i.e. for $\underline{x} \in \mathcal{X}_i$. We would like to add that the conditional Fisher information matrix is a natural quantity to consider in this problem as we deliberately focus on using conditional statistics rather than global ones in order to bypass the intractability of the global log-likelihood approach. We are strongly convinced that it should appear in any analysis that entails conditional statistics.

Condition *(C2)* is required to ensure that the model can be recovered with finitely many samples. For many special cases, such as the Ising model, the minimum number of samples required to estimate the parameters must grow exponentially with the maximum interaction strength [17]. A more detailed discussion about well-posedness and Conditions *(C1)* and *(C2)* can be found in Section A of the Supplementary Material.

Conditions *(C1)* and *(C2)* differ from the concepts in [18] called restricted strong convexity property and bound on the interaction strength, respectively, in a subtle but critical manner. Conditions *(C1)* can be identified with restricted strong convexity only when the ℓ_2-norm is used in Eq. (9). We will see later that the notion of restricted strong convexity is not required for the ℓ_∞-norm that appears to be a natural metric for which the local learnability condition can be verified for general models. Moreover, for general models it remains unclear whether the restricted strong convexity holds for values of ρ_i that are independent of the problem dimension p. Condition *(C2)* is a weaker assumption than the bound on the interaction strength from [17] for it does not require an extra assumption on the maximum degree of the graphical model.

3 Generalized interaction screening

In this Section, we introduce the algorithm that efficiently solves the model selection problem in Definition 1 and provides rigorous guarantees on its reconstruction error and computational complexity.

3.1 Generalized regularized interaction screening estimator

We propose a generalization of the estimator RISE, first introduced in [18] for pairwise binary graphical models, in order to reconstruct general discrete graphical models defined in (1). The

generalized interaction screening objective (GISO) is defined for each vertex u separately and is given by

$$\mathcal{S}_n(\underline{\theta}_u) = \frac{1}{n}\sum_{t=1}^{n} \exp\left(-\sum_{k\in\mathcal{K}_u} \theta_k g_{uk}(\underline{\sigma}_k^{(t)})\right), \tag{12}$$

where $\underline{\sigma}^{(1)}, \ldots, \underline{\sigma}^{(n)}$ are n i.i.d samples drawn from $\mu(\underline{\sigma})$ in Eq. (1), $\underline{\theta}_u := (\underline{\theta}_k)_{k\in\mathcal{K}_u}$ is the vector of parameters associated with the factors in $\mathcal{K}_u$ and the locally centered basis functions g_{uk} are defined as in Eq. (3). The GISO retains the main feature of the interaction screening objective (ISO) in [18]: it is proportional to the inverse of the factor in $\mu(\underline{\sigma})$, except for the additional centering terms ϕ_{uk}. The GISO is a convex function of $\underline{\theta}_u$ and retains the "screening" property of the original ISO. The GISO is used to define the generalized regularized interaction screening estimator (GRISE) for the parameters given by

$$\widehat{\underline{\theta}}_u = \underset{\underline{\theta}_u\in\mathcal{Y}_u:\|\underline{\theta}_u\|_1\le\widehat{\gamma}}{\operatorname{argmin}} \mathcal{S}_n(\underline{\theta}_u), \tag{13}$$

where $\widehat{\gamma}$ and $\mathcal{Y}_u$ are the prior information available on $\underline{\theta}_u^*$ as defined in (5) and (6).

3.2 Error bound on parameter estimation with GRISE

We now state our first main result regarding the theoretical guarantees on the parameters reconstructed by GRISE. We call $\widehat{\underline{\theta}}_u$ an ϵ-optimal solution of (13) if

$$\mathcal{S}_n(\widehat{\underline{\theta}}_u) \le \min_{\underline{\theta}_u\in\mathcal{Y}_u:\|\underline{\theta}_u\|_1\le\widehat{\gamma}} \mathcal{S}_n(\underline{\theta}_u) + \epsilon. \tag{14}$$

Theorem 1 (Error Bound on GRISE Estimates)**.** *Let $\underline{\sigma}^{(1)}, \ldots, \underline{\sigma}^{(n)}$ be i.i.d. samples drawn according to $\mu(\underline{\sigma})$ in* (1)*. For some node $u \in \mathcal{V}$, assume that the model satisfies Condition 1 for some norm $\|\cdot\|$ and some constraint set $\mathcal{Y}_u$, and let $\alpha > 0$ be the prescribed accuracy level. If the number of samples satisfies*

$$n \ge 2^{14}\frac{\widehat{\gamma}^2(1+\widehat{\gamma})^2 e^{4\gamma}}{\alpha^4\rho_u^2}\log(\frac{4\mathbf{K}_u^2}{\delta}), \tag{15}$$

then, with probability at least $1-\delta$, any estimate that is an ϵ-minimizer of GRISE*, with $\epsilon \le (\rho_u\alpha^2 e^{-\gamma})/(20(1+\widehat{\gamma}))$, satisfies $\|\widehat{\underline{\theta}}_{\mathcal{T}_u} - \underline{\theta}^*_{\mathcal{T}_u}\| \le \frac{\alpha}{2}$.*

The proof of Theorem 1 can be found in Section B of the Supplementary Material.

The computational complexity of finding an ϵ-optimal solution of GRISE for a trivial constraint set $\mathcal{Y}_u = \mathbb{R}^{\mathbf{K}_u}$ is $C\frac{c_g n\mathbf{K}_u}{\epsilon^2}\ln(1+\mathbf{K}_u)$, where c_g is an upper-bound on the computational complexity of evaluating any $g_{ik}(\underline{\sigma}_k)$ for $k \in \mathcal{K}_i$, and C is a universal constant independent of all the parameters of the problem, see Proposition 5 in Section C of the Supplementary Material. For a certain class of constraint sets $\mathcal{Y}_u$, which we term parametrically complete, the problem can be solved in two steps: first, finding a solution to an unconstrained problem, and then projecting onto this set. Note, however, that in general the problem of finding ϵ-optimal solutions to constrained GRISE can still be difficult since the constraint set $\mathcal{Y}_u$ can be arbitrarily complicated.

Definition 2. *The constraint set $\mathcal{Y}_u$ is called a parametrically complete set if for all $\underline{\theta}_u \in \mathbb{R}^{|\mathbf{K}_u|}$, there exists $\widehat{\underline{\theta}}_u \in \mathcal{Y}_u$ such that for all $\underline{\sigma}_u$, we have*

$$\sum_{k\in\mathcal{K}_u} \underline{\theta}_k g_{uk}(\underline{\sigma}_k) = \sum_{k\in\mathcal{K}_u} \widehat{\underline{\theta}}_k g_{uk}(\underline{\sigma}_k). \tag{16}$$

Any $\widehat{\underline{\theta}}_k \in \mathcal{Y}_u$ satisfying (16) *is called an* equi-cost projection *of $\underline{\theta}_u$ onto $\mathcal{Y}_u$ and is denoted by*

$$\widehat{\underline{\theta}}_u \in \mathcal{P}_{\mathcal{Y}_u}(\underline{\theta}_u). \tag{17}$$

The computational complexity of finding of an ϵ-optimal solution of GRISE with parametrically complete set is $C\frac{c_g n\mathbf{K}_u}{\epsilon^2}\ln(1+\mathbf{K}_u) + \mathcal{C}(\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{\text{unc}}))$, where $\mathcal{C}(\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{\text{unc}}))$ denotes the computational complexity of the projection step, see Theorem 3 in Section C of the Supplementary Material.

As we will see, for many graphical models it is often possible to explicitly construct parametrically complete sets for which the computational complexity of the projection step $\mathcal{C}(\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{\text{unc}}))$ is insignificant compared to the computational complexity of unconstrained GRISE.

4 Structure identification and parameter estimation

In this Section we show that the structure of graphical models, which is the collection of maximal subsets of variables that are associated through basis functions, as well as the associated parameters, can be efficiently recovered. The key elements are twofold. First, we prove that for maximal cliques, the Local Learnability Condition (LLC) in *(C1)* can be easily verified and yields a LLC constant independent of the system size. Second, we leverage this property to design an efficient structure and parameter learning algorithm coined SUPRISE that requires iterative calls of GRISE.

4.1 The structure of graphical models

The structure plays a central role in graphical model learning for it contains all the information about the conditional independence or Markov property of the distribution $\mu(\underline{\sigma})$ from Eq (1). In order to reach the definition of the structure presented in Eq. (21), we have to introduce graph theoretic concepts specific to graphical models.

The factor graph associated with the model *family* is a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{K}, \mathcal{E})$ with vertex set $\mathcal{V}$, factor set $\mathcal{K}$ and edges connecting factors and vertices,

$$\mathcal{E} = \{(i,k) \subseteq \mathcal{V} \times \mathcal{K} \mid \sigma_i \in \underline{\sigma}_k\}\,. \tag{18}$$

We see from Eq. (18) that the edge (i,k) exists when the variable σ_i associated with the vertex i is an argument of the basis function $f_k(\underline{\sigma}_k)$ associated with the factor k. Note that this definition only depends on the set of basis functions $\mathcal{K}$ and does not refer to a particular choice of model within the family. The factor graph $\mathcal{G}^* = (\mathcal{V}, \mathcal{K}^*, \mathcal{E}^*)$ associated with a *model*, as defined in Eq. (1), is the induced subgraph of $\mathcal{G}$ obtained from the vertex set $\mathcal{V}$ and factor subset $\mathcal{K}^* = \{k \in \mathcal{K} \mid \theta_k^* \neq 0\}$. We also use the shorthand notation $\mathcal{G}^* = \mathcal{G}\left[(\mathcal{V}, \mathcal{K}^*)\right]$ to denote an induced subgraph of $\mathcal{G}$.

We define the neighbors of a factor k as the set of vertices linked by an edge to k and denote it by $\partial k = \{i \in \mathcal{V} \mid (i,k) \in \mathcal{E}\}$. The largest factor neighborhood size $L = \max_{k\in\mathcal{K}} |\partial k|$ is called the interaction order. Families of graphical models with $L \leq 2$ are referred to as pairwise models as opposed to the generic denomination of L-wise models when L is expected to be arbitrary.

The set of maximal factors of a graph is the set of factors whose neighborhood is not strictly contained in the neighborhood of another factor,

$$\mathcal{M}_{\text{fac}}\left(\mathcal{G}\right) = \{k \in \mathcal{K} \mid \nexists k' \in \mathcal{K} \text{ s.t } \partial k \subset \partial k'\}\,. \tag{19}$$

Notice that multiple maximal factors may have the same neighborhood. This motivates the definition of the set of maximal cliques which is contained in the powerset $P(\mathcal{V})$ and consists of all neighborhoods of maximal factors,

$$\mathcal{M}_{\text{cli}}\left(\mathcal{G}\right) = \{c \in P(\mathcal{V}) \mid \exists k \in \mathcal{M}_{\text{fac}}\left(\mathcal{G}\right) \text{ s.t. } c = \partial k\}\,. \tag{20}$$

The set of factors whose neighborhoods are the same maximal clique $c \in \mathcal{M}_{\text{cli}}\left(\mathcal{G}\right)$ is called the span of the clique defined as $[c]_{\text{sp}} = \{k \in \mathcal{M}_{\text{fac}}\left(\mathcal{G}\right) \mid c = \partial k\}$. Finally, the structure $\mathbb{S}$ of a graphical model is the set of maximal cliques associated with the factor graph of the *model*,

$$\mathbb{S}(\mathcal{G}^*) = \mathcal{M}_{\text{cli}}\left(\mathcal{G}^*\right)\,. \tag{21}$$

We would like to stress that the structure of a model is different from the set of maximal cliques of the *family* of graphical models $\mathcal{M}_{\text{cli}}\left(\mathcal{G}\right)$ as the former is constructed with the set of factors associated with non-zero parameters while the latter consists of all potential maximal factors.

4.2 From local learnability condition to nonsingular parametrization of cliques

We show that the learning problem of reconstructing maximal cliques is well-posed in general and especially for non-degenerate globally centered basis functions. To this end, we demonstrate that the LLC in *(C1)* is automatically satisfied whenever the target sets $\mathcal{T}_i$ consist of factors corresponding to maximal cliques of the graphical model family. Importantly, we prove that the LLC constant ρ_i does not depend on the dimension of the model for the $\ell_{\infty,2}$-norm but rather relies on the Nonsingular Parametrization of Clique (NPC) by the basis functions. Similarly, we also guarantee that the LLC holds for the ℓ_2-norm in the case of pairwise colorable models.

We introduce globally centered basis functions defined for any factor $k \in \mathcal{K}$ through the inclusion–exclusion formula,

$$h_k(\underline{\sigma}_k) = f_k(\underline{\sigma}_k) + \sum_{r \in P(\partial k) \setminus \emptyset} \frac{(-1)^{|r|}}{|\mathcal{A}_r|} \sum_{\underline{\sigma}_r} f_k(\underline{\sigma}_k), \tag{22}$$

where $\mathcal{A}_r = \bigotimes_{j \in r} \mathcal{A}_j$. It is straightforward to see that globally centered functions sum partially to zero for any variables, i.e. $\sum_{\sigma_i \in \mathcal{A}_i} h_k(\underline{\sigma}_k) = 0$ for all $i \in \partial k$. It is worth noticing that when the functions f_k are already globally centered, we have $f = g = h$. We would also like to point out that unlike locally centered functions g_{ik}, globally centered functions cannot in general be interchanged with functions f_k without modifying conditional distributions. However they play an important role in determining the independence of basis functions around cliques through the Nonsingular Parametrization of Cliques (NPC) constant introduced below. Given a perturbation set $\mathcal{X}_i$, as defined in Eq. (8), the NPC constant is defined through the following minimization,

$$\rho_i^{\mathrm{NPC}} = \min_{\substack{c \in \mathcal{M}_{\mathrm{cli}}(\mathcal{G}) \\ c \ni i}} \min_{\substack{\|\underline{x}_c\|_2 = 1 \\ \underline{x}_c \in \mathcal{X}_i^c}} \mathbb{E}_{(\sigma_i)} \left[\sum_{\underline{\sigma}_{c \setminus i} \in \mathcal{A}_{c \setminus i}} \left(\sum_{k \in [c]_{\mathrm{sp}}} x_k h_k(\underline{\sigma}_k) \right)^2 \right], \tag{23}$$

where the vector $\underline{x}_c \in \mathbb{R}^{|[c]_{\mathrm{sp}}|}$ belongs to $\mathcal{X}_i^c$, the projection of the constraint $\mathcal{X}_i \subseteq \mathbb{R}^{\mathbf{K}_i}$ to the components $k \in [c]_{\mathrm{sp}}$ and the expectation is with respect to the marginal distribution of σ_i. Note that NPC constant only depends on L and not on the size of the system, and can be explicitly computed in time $O(\mathbf{K})$. A detailed discussion can be found in Section D of the Supplementary Material. The importance of the NPC constant is highlighted by the following proposition that guarantees that the LLC is satisfied for maximal factors in $\ell_{\infty,2}$-norm as long as $\rho_i^{\mathrm{NPC}} > 0$.

Proposition 1 (LLC in $\ell_{\infty,2}$-norm)**.** *For a specific vertex $i \in \mathcal{V}$, let the target set be maximal factors with i as neighbor, i.e. $\mathcal{T}_i = \{k \in \mathcal{M}_{\mathrm{fac}}(\mathcal{G}) \mid \partial k \ni i\}$. For vectors in the perturbation set $\underline{x} \in \mathcal{X}_i \subseteq \mathbb{R}^{\mathbf{K}_i}$, define the $\ell_{\infty,2}$-norm over components that are maximal factors as $\|\underline{x}_{\mathcal{T}_i}\|_{\infty,2} = \max_{\substack{c \in \mathcal{M}_{\mathrm{cli}}(\mathcal{G}) \\ c \ni i}} \sqrt{\sum_{k \in [c]_{\mathrm{sp}}} x_k^2}$. Then for discrete graphical models with maximum alphabet size q, interaction order L and models with finite maximum interaction strength γ as defined in Eq.* (10)*, the Local Learnability Condition* (C1) *is satisfied whenever the Nonsingular Parameterization of Cliques constant ρ_i^{NPC} is nonzero and we have,*

$$\mathbb{E}\left[\left(\sum_{k \in \mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k) \right)^2 \right] \geq \rho_i^{\mathrm{NPC}} \left(\frac{\exp(-2\gamma)}{q} \right)^{L-1} \|\underline{x}_{\mathcal{T}_i}\|_{\infty,2}^2. \tag{24}$$

Proposition 1 guarantees that the LLC can be satisfied uniformly in the size p of the model whenever $\rho_i^{\mathrm{NPC}} > 0$. The proof of Proposition 1 can be found in Section D of the Supplementary Material.

For family of models whose factors involve at most $L = 2$ variables, the so-called pairwise models, we can show that the LLC conditions for maximal factors also holds for the ℓ_2-norm. This LLC conditions depends on the vertex chromatic number χ of the *model* factor graph. We recall that a vertex coloring of a graph $\mathcal{G}^* = (\mathcal{V}, \mathcal{K}^*, \mathcal{E}^*)$ is a partition $\{S_r\}_{r \in \mathbb{N}} \in P(\mathcal{V})$ of the vertex set such that no two vertices with the same color are connected to the same factor node, i.e. $i, j \in S_r \Rightarrow \nexists k \in \mathcal{K}^*$ s.t. $i, j \in \partial k$. The chromatic number is the cardinality of the smallest graph coloring.

Proposition 2 (LLC in ℓ_2-norm for pairwise models)**.** *For a specific vertex $i \in \mathcal{V}$, let the target set be maximal factors with i as neighbor, i.e. $\mathcal{T}_i = \{k \in \mathcal{M}_{\mathrm{fac}}(\mathcal{G}) \mid \partial k \ni i\}$. For $\underline{x} \in \mathcal{X}_i \subseteq \mathbb{R}^{\mathbf{K}_i}$, define the ℓ_2-norm over components that are maximal factors $\|\underline{x}_{\mathcal{T}_i}\|_2 = \sqrt{\sum_{k \in \mathcal{T}_i} x_k^2}$. Then for discrete pairwise graphical models with maximum alphabet size q and models with chromatic number χ and finite maximum interaction strength γ as defined in Eq.* (10)*, the Local Learnability Condition* (C1) *is satisfied whenever the NPC constant ρ_i^{NPC} is nonzero and we have,*

$$\mathbb{E}\left[\left(\sum_{k \in \mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k) \right)^2 \right] \geq \frac{\rho_i^{\mathrm{NPC}}}{\chi} \frac{\exp(-2\gamma)}{q} \|\underline{x}_{\mathcal{T}_i}\|_2^2. \tag{25}$$

The reader will find the proof of Proposition 2 in Section D of the Supplementary Material.

4.3 Structure unveiling and parameter reconstruction with interaction screening estimation

Algorithm 1: Structure Unveiling and Parameter Reconstruction with Interaction Screening Estimation (SUPRISE)

```
// Step 1: Initialization of set of considered factors
```
$\mathcal{K}^0 \leftarrow \mathcal{K}$;

for $t = 0, \ldots, L-1$ **do**

```
    // Step 2: Reconstruct maximal factors bigger than L − t
```
  Construct the induced sub-graph: $\mathcal{G}^t \leftarrow \mathcal{G}\left[(\mathcal{V}, \mathcal{K}^t)\right]$;

  for $u \in \mathcal{V}$ **do**

    Set target factors: $\mathcal{T}_u^t \leftarrow \{k \in \mathcal{M}_{\text{fac}}\left(\mathcal{G}^t\right) \mid \partial k \ni u\}$;

    Set residual factors: $\mathcal{R}_u^t \leftarrow \mathcal{K}_u^t \setminus \mathcal{T}_u^t$;

    Estimate $\widehat{\underline{\theta}}_u^t$ using GRISE with accuracy at least $\epsilon = \rho_{\text{NPC}}\alpha^2 \exp(-\gamma(2L-1))/(20(1+\widehat{\gamma})q^{L-1})$ on the model defined by $\mathcal{K}_u^t, \mathcal{T}_u^t, \mathcal{R}_u^t$ and constraint set $\mathcal{Y}_u$;

  end

```
    // Step 3: Identify max cliques associated with zero parameters
```
  Initialize set of removable factors: $\mathcal{N}^t \leftarrow \emptyset$;

  for $c \in \mathcal{M}_{\text{cli}}\left(\mathcal{G}^t\right)$ **do**

    Compute average reconstruction: $\widehat{\underline{\theta}}_c^{\text{avg}(t)} \leftarrow \left\{|c|^{-1} \sum_{u \in c} (\widehat{\underline{\theta}}_u^t)_k \mid k \in [c]_{\text{sp}}\right\}$;

    if $\|\widehat{\underline{\theta}}_c^{\text{avg}(t)}\|_2 < \alpha/2$ **then**

      Update set of removable factors: $\mathcal{N}^t \leftarrow \mathcal{N}^t \cup [c]_{\text{sp}}$;

    end

  end

  Update considered factors: $\mathcal{K}^{t+1} \leftarrow \mathcal{K}^t \setminus \mathcal{N}^t$;

end

```
// Step 4: output structure and non-zero parameters of maximal factors
```
return $\widehat{\mathbb{S}} = \mathcal{M}_{\text{cli}}\left(\mathcal{G}\left[(\mathcal{V}, \mathcal{K}^L)\right]\right)$ and $\widehat{\underline{\theta}}_{\mathcal{M}} = \left\{\widehat{\theta}_k^{\text{avg}(L-1)} \mid k \in \mathcal{M}_{\text{fac}}\left(\mathcal{G}\left[(\mathcal{V}, \mathcal{K}^L)\right]\right)\right\}$;

Suppose that we know $\alpha > 0$, a lower-bound on the minimal intensity of the parameters associated with the structure in the sense that $\alpha \leq \min_{c \in \mathbb{S}(\mathcal{G}^*)} \sqrt{\sum_{k \in [c]_{\text{sp}}} \theta_k^{*2}}$. Then we can recover the structure and parameters associated with maximal factors of any graphical models using Algorithm 1, coined SUPRISE for Structure Unveiling and Parameter Reconstruction with Interaction Screening Estimation. SUPRISE that implements an iterative use of GRISE is shown to have a sample complexity logarithmic in the system size for models with non-zero NPC constants. Our second main result is the following Theorem 2, proved in Section D of the Supplementary Material, which provides guarantees on SUPRISE.

Theorem 2 (**Reconstruction and Estimation Guarantees for SUPRISE**). *Let $\mu(\underline{\sigma})$ in* (1) *be the probability distribution of a discrete graphical model with maximum alphabet size q, interaction order L, finite maximum interaction strength γ and smallest Nonsingular Parameterization of Cliques constant greater than zero, i.e. $\rho_{\text{NPC}} = \min_{u \in \mathcal{V}} \rho_u^{\text{NPC}} > 0$. Let $\underline{\sigma}^{(1)}, \ldots, \underline{\sigma}^{(n)}$ be i.i.d. samples drawn according to $\mu(\underline{\sigma})$ and assume that*

$$n \geq 2^{14} q^{2(L-1)} \frac{\widehat{\gamma}^2 (1+\widehat{\gamma})^2 e^{4\gamma L}}{\alpha^4 \rho_{\text{NPC}}^2} \log\left(\frac{4pL\mathbf{K}^2}{\delta}\right), \tag{26}$$

where $\mathbf{K} = \max_{u \in \mathcal{V}} \mathbf{K}_u$ is the maximal number of basis functions in which a variable can appear and $\widehat{\gamma} \geq \max_{u \in \mathcal{V}} \|\underline{\theta}_u^*\|_1$ is our ℓ_1-prior on the components of the parameters. Then the structure of the general graphical model is perfectly recovered using Algorithm 1, i.e. $\widehat{\mathbb{S}} = \mathbb{S}$, with probability $1-\delta$. In addition, the parameters associated with maximal factors are reconstructed with precision $\max_{c \in \mathbb{S}} \sum_{k \in [c]_{\text{sp}}} \left(\widehat{\theta}_k - \theta_k^*\right)^2 \leq \alpha^2/4$ for general models and with $\sum_{c \in \mathbb{S}} \sum_{k \in [c]_{\text{sp}}} \left(\widehat{\theta}_k - \theta_k^*\right)^2 \leq \chi^2 \alpha^2/4$ for pairwise models with chromatic number χ.

Table 1: Sample complexity and computational complexity of SUPRISE over special cases.

Model name	Inter. order	Alphabet size	Recovery type	Sample complexity	Algo. complexity
Ising	2	2	structure	$O\left(\alpha^{-4}e^{8\gamma}\log p\right)$	$\widetilde{O}(p^2)$
Ising	2	2	ℓ_2-parameter	$O\left(\chi^2\alpha^{-4}e^{8\gamma}\log p\right)$	$\widetilde{O}(p^2)$
Binary	L	2	structure	$O\left(\alpha^{-4}4^L e^{4\gamma L}L\log p\right)$	$\widetilde{O}(p^L)$
Pairwise	2	q	structure	$O\left(\alpha^{-4}q^4e^{12\gamma}\log(pq)\right)$	$\widetilde{O}(p^2)$
Pairwise	2	q	ℓ_2-parameter	$O\left(\chi^2\alpha^{-4}q^4e^{12\gamma}\log(pq)\right)$	$\widetilde{O}(p^2)$
General	L	q	structure	$O\left(\alpha^{-4}q^{2L}e^{4\gamma(L+1)}L\log(pq)\right)$	$\widetilde{O}(p^L)$

The total computational complexity scales as $\widetilde{\mathcal{O}}(p\mathbf{K})$, for fixed L, α, γ, $\widehat{\gamma}$ and δ, if the constraint sets $\mathcal{Y}_u$ are parametrically complete.

As an application of Theorem 2, we show the sample and computational complexity of recovering parameter values and the structure of some well-known special cases in Table 1. The parameter α appearing in Table 1 is the precision to which parameters are recovered in the considered norm, χ is the chromatic number of the graph, L is the interaction order, q is the maximum alphabet size, γ is the maximum interaction strength and p is the number of variables. At this point, it is instructive to compare our sample complexity requirements to existing results. A direct application of bounds of [12] and [20] to the case of pairwise multi-alphabet models that we consider below yields $O(\exp(14\gamma))$ dependence, whereas SUPRISE has a complexity that scales as $O(\exp(12\gamma))$. In the case of binary L-wise models, while [12] shows the $O(\exp(O(\gamma L)))$ scaling, SUPRISE enjoys a sample complexity $O(\exp(4\gamma L))$. The algorithm of [9] recovers a subclass of general graphical models with bounded degree, but has a sub-optimal double-exponential scaling in γ, while SUPRISE leads to recovery of arbitrary discrete graphical models with a single-exponential dependence in γ and needs no bound on the degree. In terms of the computational complexity, SUPRISE achieves the efficient scaling $\widetilde{O}(p^L)$ for models with the maximum interaction order L, which matches the best-known $\widetilde{O}(p^2)$ scaling for pairwise models [12, 20]. In summary, SUPRISE generalizes, improves and extends the existing results in the literature. The proofs for special cases can be found in Section E of the Supplementary Material.

5 Conclusion and future work

A key result of our paper is the existence of a computationally efficient algorithm that is able to recover arbitrary discrete graphical models with multi-body interactions. This result is a particular case of the general framework that we have introduced, which considers arbitrary model parametrization and makes distinction between the bounds on the parameters of the underlying model and the prior parameters. The computational complexity $\widetilde{O}(p^L)$ that we achieve is believed to be efficient for this problem [12]. In terms of sample complexity, the information-theoretic bounds for recovery of general discrete graphical models are unknown. In the case of binary pairwise models, the sample complexity bounds resulting from our general analysis are near-optimal with respect to known information-theoretic lower bounds [17]. It would be interesting to see if the $1/\alpha^4$ factor in our sample complexity bounds can be improved to $1/\alpha^2$ using an ℓ_1-norm penalty rather than an ℓ_1-norm constraint, as it has been shown for the particular case of Ising models [13, 18].

Other open questions left for future exploration include the possibility to extend the analysis to the case of graphical models with nonlinear parametrizations like in [10], and to graphical models with continuous variables. It is particularly interesting to see whether the computationally efficient and nearly sample-optimal method introduced in the present work could be useful for designing efficient learning algorithms that can improve the state-of-the-art in the well-studied case of Gaussian graphical models, for which it has been recently shown that the information-theoretic lower bound on sample complexity is tight [14].

Broader impact

We believe that this work, as presented here, does not present any foreseeable societal consequence.

Acknowledgments

Research presented in this article was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project numbers 20190059DR, 20190195ER, 20190351ER, and 20210078DR.

References

[1] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167 – 175, 2003.

[2] A. Ben-Tal, T. Margalit, and A. Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12(1):79–108, 2001.

[3] Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 771–782. ACM, 2015.

[4] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of Markov random fields from samples: Some observations and algorithms. *SIAM Journal on Computing*, 42(2):563–578, 2013.

[5] Anna L Buczak and Erhan Guven. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176, 2016.

[6] Rafael Chaves, Christian Majenz, and David Gross. Information–theoretic implications of quantum causal structures. *Nature communications*, 6:5766, 2015.

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

[8] Anthony Costa Constantinou, Norman Fenton, William Marsh, and Lukasz Radlinski. From complex questionnaire and interviewing data to intelligent bayesian network models for medical decision support. *Artificial intelligence in medicine*, 67:75–93, 2016.

[9] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of markov random fields, and their algorithmic applications. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2463–2472. Curran Associates, Inc., 2017.

[10] Abhijith J., Andrey Lokhov, Sidhant Misra, and Marc Vuffray. Learning of discrete graphical models with neural networks. *arXiv e-prints*, pages arXiv–2006, 2020.

[11] Ronald Jansen, Haiyuan Yu, Dov Greenbaum, Yuval Kluger, Nevan J. Krogan, Sambath Chung, Andrew Emili, Michael Snyder, Jack F. Greenblatt, and Mark Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, 2003.

[12] A. Klivans and R. Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354, Oct 2017.

[13] Andrey Y Lokhov, Marc Vuffray, Sidhant Misra, and Michael Chertkov. Optimal structure and parameter learning of Ising models. *Science advances*, 4(3):e1700791, 2018.

[14] Sidhant Misra, Marc Vuffray, and Andrey Y Lokhov. Information theoretic optimal learning of gaussian graphical models. *The 33rd Annual Conference on Learning Theory (COLT 2020)*, 2020.

[15] Andrea Montanari and Jose A. Pereira. Which graphical models are difficult to learn? In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1303–1311. Curran Associates, Inc., 2009.

[16] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using $\ell 1$-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 06 2010.

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

[18] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2595–2603. Curran Associates, Inc., 2016.

[19] Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627, 2013.

[20] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, pages 8071–8081, 2019.

Supplementary Material

This document contains supplementary materials for the paper "Efficient Learning of Discrete Graphical Models". Section A contains a detailed discussion about Conditions *(C1)* and *(C2)*. In Section B, the reader can find the proofs for the error bound on GRISE estimates. Section C contains the proofs relatives to the computational complexity of GRISE and its efficient implementation. In Section D, the reader can find the proofs in connection with the NPC constant and relative to structure and parameter estimation with SUPRISE. Finally, Section E contains the proofs for the applications of SUPRISE for structure and parameter estimation of iconic special cases.

A About well-posedness and local learnability conditions

A.1 About condition *(C1)*

To illustrate further why Condition *(C1)* is required, we look at a case where the local constraint set is trivial, i.e. $\mathcal{Y}_i = \mathbb{R}^{\mathbf{K}_i}$ and we consider a model that violates Condition *(C1)*. This implies that there exists a sequence $\underline{x}_n \in \mathcal{X}_i$ such that $\underline{x}_n^\top \widetilde{I}(\theta^*)\underline{x}_n / \|\underline{x}_{\mathcal{T}_i}\|^2 < \rho_n$ with $\rho_n \to 0$. In the limit, we can find a vector $\underline{x}$ such that $\underline{x}^\top \widetilde{I}(\theta^*)\underline{x} = 0$ and $\|\underline{x}_{\mathcal{T}_i}\| = 1$. In other words, it implies that for this model there exists a vertex i and a perturbation vector $\underline{x} \in \mathcal{X}_i$ such that $\underline{x}_{\mathcal{T}_i} \neq 0$ and for which $\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] = 0$. Since the probability distribution in Eq. (1) is positive, it further implies that for all configurations $\underline{\sigma}$ we have the functional equality $\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k) = 0$. This enables us to locally reparameterize the distribution:

$$\exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* f_k(\underline{\sigma}_k)\right) = \exp\left(c(\underline{x}, \underline{\sigma}_{\setminus i}) + \sum_{k\in\mathcal{K}_i} (\theta_k^* - x_k)\, f_k(\underline{\sigma}_k)\right), \tag{27}$$

where $c(\underline{x}, \underline{\sigma}_{\setminus i}) = \sum_{k\in\mathcal{K}_i} x_k \phi_{ik}(\underline{\sigma}_{k\setminus i})$ is a sum of locally centered functions that does not involve the variable σ_i. At this point, we should distinguish between the two cases when the basis functions f_k are centered or not. When the basis functions are centered, i.e. $f_k = g_k$, the residual in Eq. (27) is identically zero, $c(\underline{x}, \underline{\sigma}_{\setminus i}) = 0$. Therefore, the probability distribution of the model in (1) can be reparameterized entirely with $\theta_k^* \to \theta_k^* - x_k$ for $k \in \mathcal{K}_i$. It implies, as $\underline{x}_{\mathcal{T}_i} \neq 0$, that there exists two parameterization of the same models with different target parameters and the model selection problem as stated in Definition 1 is ill-posed. In the case when the basis functions are not centered, i.e. $f_k \neq g_k$, it may not be possible to reparameterized the whole distribution of the model. However, the conditional probability distribution $\mathbb{P}(\sigma_i \mid \underline{\sigma}_{\setminus i})$ can be reparameterized as it is proportional to $\exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* f_k(\underline{\sigma}_k)\right)$ and $\exp\left(\sum_{k\in\mathcal{K}_i} (\theta_k^* - x_k)\, f_k(\underline{\sigma}_k)\right)$ thanks to Eq. (27). Thereby, even if the model is uniquely parameterized with $\underline{\theta}_{\mathcal{T}_i}$, local methods based on independent neighborhood reconstructions using conditional distributions will fail at selecting a unique model as shown in the following example.

Example 1. *Consider two family of models over two binary variables $\sigma, s \in \{-1, 1\}$, parameterized by $\underline{\theta}$ and $\underline{\eta}$,*

$$\mu_{\underline{\theta}}(\sigma, s) \propto \exp\left(\theta_1 \sigma(s-1) + \theta_2 s(\sigma - 1) + \theta_3(\sigma + s)\right), \tag{28}$$

and

$$\mu_{\underline{\eta}}(\sigma, s) \propto \exp\left(\eta_1 \sigma s + \eta_2 \sigma + \eta_3 s\right). \tag{29}$$

Both models are equivalent through the invertible mapping $\eta_1 = \theta_1 + \theta_2$, $\eta_2 = \theta_3 - \theta_1$ and $\eta_3 = \theta_3 - \theta_2$. However the model in Eq. (29) *that has centered basis functions satisfies* (C1) *from Condition 1, while the model in Eq.* (28) *that has non-centered basis functions does not. This implies that the parameters $\underline{\theta}$ cannot be recovered by looking independently at conditional distributions as they are degenerate in this basis,*

$$\mathbb{P}(\sigma \mid s) \propto \exp\left((\theta_1 + \theta_2)\sigma s + (\theta_3 - \theta_1)\sigma\right), \tag{30}$$

$$\mathbb{P}(s \mid \sigma) \propto \exp\left((\theta_1 + \theta_2)\sigma s + (\theta_3 - \theta_2) s\right). \tag{31}$$

Indeed the change of parameters $\theta_1 \to \theta_1 + \epsilon$, $\theta_2 \to \theta_2 - \epsilon$ and $\theta_3 \to \theta_3 + \epsilon$ leaves the conditional distribution (30) *unchanged while the change of parameters $\theta_1 \to \theta_1 - \epsilon$, $\theta_2 \to \theta_2 + \epsilon$ and*

$\theta_3 \to \theta_3 - \epsilon$ *leaves the conditional distribution* (31) *unaffected. Note that there does not exist a change of parameters that leaves both* (30) *and* (31) *unchanged. This is in agreement with the fact that the model in Eq.* (28) *is uniquely parameterized and can be in principle recovered by looking* jointly *at both conditional distributions.*

For the specific models that we considered in Section E, the basis functions are always centered, which implies that failure to satisfy *(C1)* means that the model selection problem is ill-posed.

A.2 About condition *(C2)*

The bound on the interaction strength in *(C2)* translates directly into a uniform bound on the conditional probabilities of the models as shown in the following lemma.

Lemma 1 (Lower-Bounded Conditional Probabilities)**.** *Consider a graphical model with bounded maximum interaction strength of* $\gamma = \max_{i\in\mathcal{V}} |\max_{\underline{\sigma}} \sum_{k\in\mathcal{K}_i} \theta_k^* g_{ik}(\underline{\sigma}_k)|$*. Then for any two disjoint subsets of vertices* $A, B \subseteq \mathcal{V}$ *the conditional probability of* $\underline{\sigma}_A$ *given* $\underline{\sigma}_B$ *is bounded from below,*

$$\mathbb{P}(\underline{\sigma}_A \mid \underline{\sigma}_B) \geq \prod_{i\in A} \frac{\exp(-2\gamma)}{|\mathcal{A}_i|}, \tag{32}$$

where $|\mathcal{A}_i|$ *is the alphabet size of* σ_i*.*

Proof of Lemma 1: Lower-bound on conditional probabilities. We start by bounding the conditional probability of one variable σ_i given the rest $\underline{\sigma}_{\setminus i}$. This is given by the following expression,

$$\mathbb{P}(\sigma_i \mid \underline{\sigma}_{\setminus i}) = \frac{\exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* f_k(\underline{\sigma}_k)\right)}{\sum_{\sigma_i\in\mathcal{A}_i} \exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* f_k(\underline{\sigma}_k)\right)}, \tag{33}$$

$$= \frac{\exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* \left(g_{ik}(\underline{\sigma}_k) + \phi_{ik}(\sigma_{k\setminus i})\right)\right)}{\sum_{\sigma_i\in\mathcal{A}_i} \exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* \left(g_{ik}(\underline{\sigma}_k) + \phi_{ik}(\sigma_{k\setminus i})\right)\right)}, \tag{34}$$

$$= \frac{\exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* g_{ik}(\underline{\sigma}_k)\right)}{\sum_{\sigma_i\in\mathcal{A}_i} \exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* g_{ik}(\underline{\sigma}_k)\right)}, \tag{35}$$

as the centering functions $\phi_{ik}(\sigma_{k\setminus i})$ are independent of σ_i. The last expression can be simply bounded away from zero,

$$\mathbb{P}(\sigma_i \mid \underline{\sigma}_{\setminus i}) = \frac{\exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* g_{ik}(\underline{\sigma}_k)\right)}{\sum_{\sigma_i\in\mathcal{A}_i} \exp\left(\sum_{k\in\mathcal{K}_i} \theta_k^* g_{ik}(\underline{\sigma}_k)\right)} \geq \frac{\exp(-2\gamma)}{|\mathcal{A}_i|}, \tag{36}$$

using $\gamma \geq |\sum_{k\in\mathcal{K}_i} \theta_k^* g_{ik}(\underline{\sigma}_k)|$.

Now we consider the conditional probability of one variable σ_i given a subset of variable $\underline{\sigma}_B$ where $B \subseteq \mathcal{V}$ and $i \notin B$. Denote the complementary set of i and B by $S = \mathcal{V} \setminus (\{i\} \cup B)$. Then using the chain rule and the inequality from Eq. (36) we find the following lower-bound,

$$\mathbb{P}(\sigma_i \mid \underline{\sigma}_B) = \sum_{\underline{\sigma}_S} \mathbb{P}(\sigma_i, \underline{\sigma}_S \mid \underline{\sigma}_B) \tag{37}$$

$$= \sum_{\underline{\sigma}_S} \mathbb{P}(\sigma_i \mid \underline{\sigma}_S, \underline{\sigma}_B)\mathbb{P}(\underline{\sigma}_S \mid \underline{\sigma}_B) \tag{38}$$

$$= \sum_{\underline{\sigma}_S} \mathbb{P}(\sigma_i \mid \underline{\sigma}_{\setminus i})\mathbb{P}(\underline{\sigma}_S \mid \underline{\sigma}_B) \tag{39}$$

$$\geq \frac{\exp(-2\gamma)}{|\mathcal{A}_i|}. \tag{40}$$

Finally we consider the conditional probability of a set of variable $\underline{\sigma}_A$ given another set $\underline{\sigma}_B$ where $A \cap B = \emptyset$. Denote the vertices in A by $\{1, 2, \ldots, |A|\}$. Then using the chain rule and the inequality

from Eq. (36), we obtain the desired result,

$$\mathbb{P}(\underline{\sigma}_A \mid \underline{\sigma}_B) = \prod_{j=1,\ldots,|A|} \mathbb{P}(\sigma_j, \mid \underline{\sigma}_B, \sigma_{j+1}, \ldots, \sigma_{|A|}), \tag{41}$$

$$\geq \prod_{j=1,\ldots,|A|} \frac{\exp(-2\gamma)}{|\mathcal{A}_j|}. \tag{42}$$

□

B Proofs of GRISE estimation error bound

Proposition 3 (Gradient Concentration for GRISE)**.** *For some node $u \in V$, let $n > \frac{2e^{2\gamma}}{\epsilon_1^2}\log(\frac{2\mathbf{K}_u}{\delta_1})$, then with probability at least $1-\delta_1$ the components of the gradient of the GISO are bounded from above as*

$$\|\nabla \mathcal{S}_n(\underline{\theta}_u^*)\|_\infty < \epsilon_1. \tag{43}$$

Define the residual of the first order Taylor expansion as

$$\delta\mathcal{S}_n(\Delta, \underline{\theta}_u^*) = \mathcal{S}_n(\underline{\theta}_u^* + \Delta) - \mathcal{S}_n(\underline{\theta}_u^*) - \langle \nabla \mathcal{S}_n(\underline{\theta}_u^*), \Delta\rangle. \tag{44}$$

Proposition 4 (Restricted Strong Convexity for GRISE)**.** *For some node $u \in \mathcal{V}$, let $n > \frac{2}{\epsilon_2^2}\log\left(\frac{2\mathbf{K}_u^2}{\delta_2}\right)$ and assume that Condition 1 holds for some norm $\|\cdot\|$. Then, with probability at least $1-\delta_2$ the error of the first order Taylor expansion of the GISO satisfies*

$$\delta\mathcal{S}_n(\Delta, \underline{\theta}_u^*) \geq \exp(-\gamma)\frac{\rho_u\|\Delta_{\mathcal{T}_u}\|^2 - \epsilon_2\|\Delta\|_1^2}{2+\|\Delta\|_1}. \tag{45}$$

for all $\Delta \in \mathcal{X}_u \subseteq \mathbb{R}^{\mathbf{K}_u}$.

We first prove Theorem 1 before proving the propositions.

Proof of Theorem 1: Error Bound on GRISE. For some node $u \in \mathcal{V}$, let $n \geq \frac{2^{14}\widehat{\gamma}^2(1+\widehat{\gamma})^2 e^{4\gamma}}{\alpha^4\rho_u^2}\log(\frac{4\mathbf{K}_u^2}{\delta})$. As the estimate $\widehat{\underline{\theta}}_u$ is an ϵ-optimal point of the GISO and $\underline{\theta}_u^*$ lies in the constraint set from Eq. (6), we find that for $\Delta = \widehat{\underline{\theta}}_u - \underline{\theta}_u^*$

$$\epsilon \geq \mathcal{S}_n(\widehat{\underline{\theta}}_u) - \mathcal{S}_n(\underline{\theta}_u^*) \tag{46}$$

$$= \langle \nabla \mathcal{S}_n(\underline{\theta}_u^*), \Delta\rangle + \delta\mathcal{S}_n(\Delta, \underline{\theta}_u^*) \tag{47}$$

$$\geq -\|\nabla \mathcal{S}_n(\underline{\theta}_u^*)\|_\infty\|\Delta\|_1 + \delta\mathcal{S}_n(\Delta, \underline{\theta}_u^*). \tag{48}$$

Using the union bound on Proposition 3 and Proposition 4 with $\delta_1 = \delta_2 = \frac{\delta}{2}$ and

$$\epsilon \leq \frac{\rho_u\alpha^2 e^{-\gamma}}{20(1+\gamma)}, \epsilon_1 = \frac{\rho_u\alpha^2 e^{-\gamma}}{40\widehat{\gamma}(1+\widehat{\gamma})}, \epsilon_2 = \frac{\rho_u\alpha^2}{80\widehat{\gamma}^2}, \tag{49}$$

we can express the inequality as

$$\epsilon \geq -\epsilon_1\|\Delta\|_1 + e^{-\gamma}\frac{\rho_u\|\Delta_{\mathcal{T}_u}\|^2 - \epsilon_2\|\Delta\|_1^2}{2+\|\Delta\|_1}. \tag{50}$$

Since by assumptions $\|\underline{\theta}_u^*\|_1 \leq \gamma$ and $\|\widehat{\underline{\theta}}_u\|_1 \leq \widehat{\gamma}$ for $\gamma \leq \widehat{\gamma}$ as the estimate is an ϵ-optimal point of the ℓ_1-*constrained* GISO, the error $\|\Delta\|_1$ is bounded by $2\widehat{\gamma}$. By choosing

and after some algebra, we obtain that

$$\|\Delta_{\mathcal{T}_u}\| \leq \frac{\alpha}{2}. \tag{51}$$

□

B.1 Gradient concentration

The components of the gradient of the GISO is given by

$$\frac{\partial}{\partial \theta_k} \mathcal{S}_n(\underline{\theta}_u^*) = \frac{1}{n} \sum_{t=1}^{n} -g_{uk}(\underline{\sigma}_k^{(t)}) \exp\left(- \sum_{l \in \mathcal{K}_u} \theta_l^* g_{ul}(\underline{\sigma}_l^{(t)}) \right). \tag{52}$$

Each term in the summation above is distributed as the random variable

$$X_{uk} = -g_{uk}(\underline{\sigma}_k) \exp\left(- \sum_{l \in \mathcal{K}_u} \theta_k^* g_{ul}(\underline{\sigma}_l) \right) \quad \forall k \in \mathcal{K}_u. \tag{53}$$

Lemma 2. *For any $u \in \mathcal{V}$ and $k \in \mathcal{K}_u$, we have*

$$\mathbb{E}\left[X_{uk}\right] = 0. \tag{54}$$

Proof. Simple computation. □

Proof of Proposition 3: Gradient Concentration for GRISE. The random variable X_{uk} is bounded as

$$|X_{uk}| = |g_{uk}(\underline{\sigma}_k)| \exp\left(- \sum_{k \in \mathcal{K}_u} \theta_k^* g_{uk}(\underline{\sigma}_k) \right) \leq \exp(\gamma). \tag{55}$$

Using Lemma 2 and the Hoeffding inequality, we get

$$\mathbb{P}\left(\left| \frac{\partial}{\partial \theta_k} \mathcal{S}_n(\underline{\theta}_u^*) \right| > \epsilon_1 \right) < 2 \exp\left(-\frac{n\epsilon_1^2}{2e^{2\gamma}} \right). \tag{56}$$

The proof follows by using (56) and the union bound over all $k \in \mathcal{K}_u$. □

B.2 Restricted strong convexity

We make use of the following deterministic functional inequality derived in [18].

Lemma 3. *The following inequality holds for all $z \in \mathbb{R}$.*

$$e^{-z} - 1 + z \geq \frac{z^2}{2 + |z|}. \tag{57}$$

Proof of Lemma 3. Note that the inequality is true for $z = 0$ and the first derivative of the difference is positive for $z > 0$ and negative for $z < 0$. □

Let $H_{k_1 k_2}$ denote the correlation between g_{k_1} and g_{k_2} defined as

$$H_{k_1 k_2} = \mathbb{E}\left[g_{uk_1}(\underline{\sigma}_{k_1}) g_{uk_2}(\underline{\sigma}_{k_2}) \right], \tag{58}$$

and let $H = [H_{k_1 k_2}] \in \mathbb{R}^{|\mathcal{K}_u| \times |\mathcal{K}_u|}$ be the corresponding matrix. We define $\hat{H}$ similarly based on the empirical estimates of the correlation $\hat{H}_{k_1 k_2} = \frac{1}{n} \sum_{t=1}^{n} g_{uk_1}(\underline{\sigma}_{k_1}^{(t)}) g_{uk_2}(\underline{\sigma}_{k_2}^{(t)})$. The following lemma bounds the deviation between the above two quantities.

Lemma 4. *Choose some node $u \in \mathcal{V}$. With probability at least $1 - 2\mathbf{K}_u^2 \exp\left(-\frac{n\epsilon_2^2}{2}\right)$, we have*

$$|\hat{H}_{k_1 k_2} - H_{k_1 k_2}| < \epsilon_2, \tag{59}$$

for all $k_1, k_2 \in \mathcal{K}_u$.

Proof of Lemma 4. Fix $k_1, k_2 \in \mathcal{K}_u$. Then the random variable defined as $Y_{k_1 k_2} = g_{uk_1}(\underline{\sigma}_{k_1}) g_{uk_2}(\underline{\sigma}_{k_2})$ satisfies $|Y_{k_1 k_2}| \leq 1$. Using the Hoeffding inequality we get

$$\mathbb{P}\left(|\hat{H}_{k_1 k_2} - H_{k_1 k_2}| > \epsilon_2 \right) < 2 \exp\left(-\frac{n\epsilon_2^2}{2} \right). \tag{60}$$

The proof follows by using the union bound over $k_1, k_2 \in \mathcal{K}_u$. □

Lemma 5. *The residual of the first order Taylor expansion of the* GISO *satisfies*

$$\delta\mathcal{S}_n(\Delta, \underline{\theta}_u^*) \geq \exp(-\gamma)\frac{\Delta^T \hat{H}\Delta}{2 + \|\Delta\|_1}. \tag{61}$$

Proof of Lemma 5. Using Lemma 3 we have

$$\delta\mathcal{S}_n(\Delta, \underline{\theta}_u^*) = \frac{1}{n}\sum_{t=1}^{n}\exp\left(-\sum_{k\in\mathcal{K}_u}\theta_k^* g_{uk}(\underline{\sigma}_k^{(t)})\right)\times \tag{62}$$

$$\left(\exp\left(-\sum_{k\in\mathcal{K}_u}\Delta_k g_{uk}(\underline{\sigma}_k^{(t)})\right) - 1 + \sum_{k\in\mathcal{K}_u}\Delta_k g_{uk}(\underline{\sigma}_k^{(t)})\right) \tag{63}$$

$$\geq \exp(-\gamma)\frac{\Delta^T \hat{H}\Delta}{2 + |\sum_{k\in\mathcal{K}_u}\Delta_k g_{uk}(\underline{\sigma}_k^{(t)})|}. \tag{64}$$

The proof follows by observing that $|\sum_{k\in\mathcal{K}_u}\Delta_k g_{uk}(\underline{\sigma}_k^{(t)})| \leq \|\Delta\|_1$. □

We are now in a position to complete the proof of Proposition 4.

Proof of Proposition 4: Restricted Strong Convexity for GRISE. Using Lemma 5 we have

$$\delta\mathcal{S}_n(\Delta, \underline{\theta}_u^*) \geq \exp(-\gamma)\frac{\Delta^T \hat{H}\Delta}{2 + \|\Delta\|_1} \tag{65}$$

$$= \exp(-\gamma)\frac{\Delta^T H\Delta + \Delta^T(\hat{H} - H)\Delta}{2 + \|\Delta\|_1} \tag{66}$$

$$\overset{(a)}{\geq} \exp(-\gamma)\frac{\Delta^T H\Delta - \epsilon_2\|\Delta\|_1^2}{2 + \|\Delta\|_1} \tag{67}$$

$$\overset{(b)}{\geq} \exp(-\gamma)\frac{\rho_u\|\Delta_{\mathcal{T}_u}\|^2 - \epsilon_2\|\Delta\|_1^2}{2 + \|\Delta\|_1}. \tag{68}$$

where (a) follows from Lemma 4 and (b) follows from Condition 1 as

$$\Delta^T H\Delta = \mathbb{E}\left[\left(\sum_{k\in K_u}\Delta_k g_{uk}(\underline{\sigma}_k)\right)^2\right]. \tag{69}$$

□

C Efficient implementation of GRISE and its computational complexity

The iterative Algorithm 2 takes as input a number of steps T and output an ϵ-optimal solution of GRISE without constraints in Eq. (74). This algorithm is an application of the Entropic Descent Algorithm introduced by [1] to reformulation of Eq. (13) as a minimization over the probability simplex. Note that there exist other efficient iterative methods for minimizing the GISO, such as the mirror gradient descent of [2]. The following proposition provides guarantees on the computational complexity of unconstrained GRISE.

Proposition 5 (Computational Complexity for Unconstrained GRISE)**.** *Let $1 \geq \epsilon > 0$ be the optimality gap and $T \geq 6\epsilon^{-2}\ln(2\mathbf{K}_u + 1)$ be the maximum number of iterations. Then Algorithm 2 is guaranteed to produce an ϵ-optimal solution of* GRISE *without a constraint set $\mathcal{Y}_u$ with a number of operation less than $C\frac{c_g n\mathbf{K}_u}{\epsilon^2}\ln(1 + \mathbf{K}_u)$, where c_g is an upper bound on the computational complexity of evaluating any $g_{ik}(\underline{\sigma}_k)$ for $k \in \mathcal{K}_i$ and C is a universal constant that is independent of all parameters of the problem.*

Algorithm 2: Entropic Descent for unconstrained GRISE

```
// Step 1: Initialization
```
$x^1_{k,+} \leftarrow 1/(2\mathbf{K}_u+1), x^1_{k,-} \leftarrow 1/(2\mathbf{K}_u+1), \forall k \in \mathcal{K}_u;$

$y^1 \leftarrow 1/(2\mathbf{K}_u+1), \eta^1 \leftarrow \sqrt{2\ln(2\mathbf{K}_u+1)};$

```
// Step 2: Entropic Descent Steps
```
for $t = 1, \ldots, T$ **do**

```
// Gradient Update:
```
$v_k = \frac{\partial}{\partial \theta_k}\mathcal{S}(\widehat{\gamma}(\underline{x}^t_+ - \underline{x}^t_-))/\mathcal{S}(\widehat{\gamma}(\underline{x}^t_+ - \underline{x}^t_-));$

$w^+_k = x^t_{k,+}\exp(-\eta^t v_k), w^-_k = x^t_{k,-}\exp(\eta^t v_k).;$

```
// Projection Step:
```
$z = y^t + \sum_{k\in\mathcal{K}_u}(w^+_k + w^-_k);$

$x^{t+1}_{k,+} \leftarrow \frac{w^+_k}{z}, x^{t+1}_{k,-} \leftarrow \frac{w^-_k}{z};$

$y^{t+1} \leftarrow \frac{y^t}{z};$

```
// Step Size Update:
```
$\eta^{t+1} \leftarrow \eta^t\sqrt{\frac{t}{t+1}};$

end

```
// Step 3:
```
$s = \operatorname{argmin}_{t=1,\ldots,T}\mathcal{S}(\widehat{\gamma}(\underline{x}^t_+ - \underline{x}^t_-));$

return $\widehat{\underline{\theta}}_u = \widehat{\gamma}(\underline{x}^s_+ - \underline{x}^s_-);$

Proof of Proposition 5: Computational complexity of unconstrained GRISE. We start by showing that the minimization of GRISE in Eq. (13) in the unconstrained case where $\mathcal{Y}_u = \mathbb{R}^{\mathbf{K}_u}$ is equivalent to the following lifted minimization on the logarithm of GRISE,

$$\min_{\underline{\theta}_u, \underline{x}^+, \underline{x}^-, y} \quad \log \mathcal{S}_n(\underline{\theta}_u) \tag{70}$$

$$\text{s.t.} \quad \underline{\theta}_u = \widehat{\gamma}(\underline{x}^+ - \underline{x}^-) \tag{71}$$

$$y + \sum_k (x^+_k + x^-_k) = 1 \tag{72}$$

$$y \geq 0, x^+_k \geq 0, x^-_k \geq 0, \forall k \in K_u. \tag{73}$$

We first show that for all $\underline{\theta}_u \in \mathbb{R}^{\mathbf{K}_u}$ such that $\|\underline{\theta}_u\|_1 \leq \widehat{\gamma}$, there exists $\underline{x}^+, \underline{x}^-, y$ satisfying constraints (71), (72), (73). This is easily done by choosing $x^+_k = \max(\theta_k/\widehat{\gamma}, 0)$, $x^-_k = \max(-\theta_k/\widehat{\gamma}, 0)$ and $y = 1 - \|\underline{\theta}_u\|_1/\widehat{\gamma}$. Second, we trivially see that for all $\underline{\theta}_u, \underline{x}^+, \underline{x}^-, y$ satisfying constraints (71), (72), (73), it implies that $\underline{\theta}_u$ also satisfies $\|\underline{\theta}_u\|_1 \leq \widehat{\gamma}$. Therefore, any $\underline{\theta}^{\min}_u$ that is an argmin of Eq. (70) is also an argmin of Eq. (13) without constraint set $\mathcal{Y}_u$. Moreover, if we find $\underline{\theta}^{\epsilon}_u$ such that $\log \mathcal{S}_n(\underline{\theta}^{\epsilon}_u) - \log \mathcal{S}_n(\underline{\theta}^{\min}_u) \leq \epsilon/\sqrt{3}$, we obtain an ϵ-minimizer of Eq. (13) without constraint set $\mathcal{Y}_u$. Indeed, since $\epsilon/\sqrt{3} \leq \log(1+\epsilon)$ for $1 \geq \epsilon > 0$, we have that $\mathcal{S}_n(\underline{\theta}^{\epsilon}_u) - \mathcal{S}_n(\underline{\theta}^{\min}_u) \leq \mathcal{S}_n(\underline{\theta}^{\min\,\epsilon}_u)\epsilon \leq \epsilon$ as $\mathcal{S}_n(\underline{\theta}^{\min}_u) \leq \mathcal{S}_n(0) = 1$. The remainder of the proof is a straightforward application of the analysis of the Entropic Descent Algorithm in [1, Th. 5.1] to the above minimization where $\underline{\theta}_u$ has been replaced by $\underline{x}^+, \underline{x}^-, y$ using Eq. (71). In this analysis we use the fact that the logarithm of GRISE remains a convex function as it is a sum of exponential functions and also that the gradient of our objective function is bounded uniformly by $\|\nabla \log \mathcal{S}_n(\underline{\theta}_u)\|_\infty = \|\nabla \mathcal{S}_n(\underline{\theta}_u)/\mathcal{S}_n(\underline{\theta}_u)\|_\infty \leq 1$ as $|g_k(\underline{\sigma}_k)| \leq 1$. Note that the computational complexity of the gradient evaluation is proportional to $n\mathbf{K}_u c_g$. This is because for each sample, one has to first compute an exponential containing $\mathbf{K}_u$ terms $g_k(\underline{\sigma}_k)$ with an evaluation cost of c_g and then multiply the exponential by the factor $-g_k(\underline{\sigma}_k)$ corresponding to each of the $\mathbf{K}_u$ components of the gradient. □

When the constraint set $\mathcal{Y}_u$ is parametrically complete, an ϵ-optimal solution to (13) can be found by first solving the unconstrained version of GRISE and then performing an equi-cost projection onto

$\mathcal{Y}_u$. We define an ϵ-optimal solution to the unconstrained GRISE problem as

$$\mathcal{S}_n(\widehat{\underline{\theta}}_u^{unc}) \leq \min_{\underline{\theta}_u : \|\underline{\theta}_u\|_1 \leq \widehat{\gamma}} \mathcal{S}_n(\underline{\theta}_u) + \epsilon. \tag{74}$$

Lemma 6. *Let $\widehat{\underline{\theta}}_u^{unc}$ be an ϵ-optimal solution of the unconstrained* GRISE *problem in Eq.* (74)*. Then an equi-cost projection of $\widehat{\underline{\theta}}_u^{unc}$ is an ϵ-optimal solution of the constrained* GRISE *problem,*

$$\mathcal{S}_n(\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{unc})) \leq \min_{\underline{\theta}_u \in \mathcal{Y}_u : \|\underline{\theta}_u\|_1 \leq \widehat{\gamma}} \mathcal{S}_n(\underline{\theta}_u) + \epsilon. \tag{75}$$

Proof of Lemma 6. Since unconstrained GRISE is a relaxation of GRISE with the constraints $\underline{\theta}_u \in \mathcal{Y}_u$, we must have

$$\mathcal{S}_n(\widehat{\underline{\theta}}_u^{\text{unc}}) \leq \min_{\underline{\theta}_u : \|\underline{\theta}_u\|_1 \leq \widehat{\gamma}} \mathcal{S}_n(\underline{\theta}_u) + \epsilon \leq \min_{\underline{\theta}_u \in \mathcal{Y}_u : \|\underline{\theta}_u\|_1 \leq \widehat{\gamma}} \mathcal{S}_n(\underline{\theta}_u) + \epsilon. \tag{76}$$

Since, $\mathcal{Y}_u$ is parametrically complete, by definition,

$$\mathcal{S}_n\left(\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{\text{unc}})\right) = \mathcal{S}_n(\widehat{\underline{\theta}}_u^{\text{unc}}). \tag{77}$$

The estimates $\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{\text{unc}})$ are feasible for the constrained GRISE problem, completing the proof. □

Lemma 6 implies that the computational complexity of GRISE for parametrically complete cases is the sum of the computational complexity of the unconstrained GRISE and the projection step.

Algorithm 3: Computing GRISE estimates for parametrically complete constraints

```
// Step 1: Solve unconstrained GRISE
```

1 Use Algorithm 2 to obtain solutions $\widehat{\underline{\theta}}_u^{\text{unc}}$ to the unconstrained GRISE ;

```
// Step 2: Perform projection step
```

2 Project $\widehat{\underline{\theta}}_u^{\text{unc}}$ onto $\mathcal{Y}_u$ to obtain the final estimates;

3 $\widehat{\underline{\theta}}_u = \mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{\text{unc}})$;

4 **return** $\widehat{\underline{\theta}}_u$;

Algorithm 3 is an implementation of GRISE for parametrically complete cases. Its computational complexity is obtained easily by combining Lemma 6 and Proposition 5.

Theorem 3 (Computational Complexity for GRISE with P.C. Constraints)**.** *Let $\mathcal{Y}_u$ be a parametrically complete set and let $1 \geq \epsilon > 0$ be given. Then Algorithm 3 computes an ϵ-optimal solution to* GRISE *with a number of operations bounded by $C\frac{c_g n \mathbf{K}_u}{\epsilon^2}\ln(1+\mathbf{K}_u) + \mathcal{C}(\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{unc}))$, where c_g is an upper bound on the computational complexity of evaluating any $g_{ik}(\underline{\sigma}_k)$ for $k \in \mathcal{K}_i$ and where $\mathcal{C}(\mathcal{P}_{\mathcal{Y}_u}(\widehat{\underline{\theta}}_u^{unc}))$ denotes the computational complexity of the projection step.*

D Proofs & algorithms for structure and parameter estimation

D.1 Dimension independence and easier computation of NPC constants

We recall the definition of the NPC constant,

$$\rho_i^{\text{NPC}} = \min_{\substack{c \in \mathcal{M}_{\text{cli}}(\mathcal{G}) \\ c \ni i}} \min_{\substack{\|\underline{x}_c\|_2 = 1 \\ \underline{x}_c \in \mathcal{X}_i^c}} \mathbb{E}_{(\sigma_i)}\left[\sum_{\underline{\sigma}_{c\setminus i} \in \mathcal{A}_{c\setminus i}} \left(\sum_{k \in [c]_{\text{sp}}} x_k h_k(\underline{\sigma}_k)\right)^2\right]. \tag{78}$$

In order to give an intuition for the intricate formula in Eq. (78), let us define for maximal cliques $c \in \mathcal{M}_{\text{cli}}(\mathcal{G})$, their clique parameterization matrix G^c, with indices being maximal factors $k, k' \in [c]_{\text{sp}}$.

The clique parameterization matrix is obtained by summing over variable $\underline{\sigma}_c$ the globally centered basis functions,

$$G^c_{k,k'} = \sum_{\underline{\sigma}_c \in \mathcal{A}_c} h_k(\underline{\sigma}_c) h_{k'}(\underline{\sigma}_c). \tag{79}$$

Note that the clique parametertization matrices are positive semi-definite matrices by construction. Bounding the expectation over σ_i using Lemma 1, we see that the NPC constant is linked to the smallest eigenvalue of the clique parameterization matrix,

$$\rho_i^{\mathrm{NPC}} \geq \frac{\exp(-2\gamma)}{q_i} \lambda_{\min}(G^c). \tag{80}$$

Clique parametertization matrices have a typical size of $\mathcal{O}(q^L \times q^L)$ since variables in a clique can take up to $\mathcal{O}(q^L)$ different configurations. Therefore, Eq. (80) emphasizes that that the NPC constant does not depend on the dimension of the model p but rather on local properties of the parameterization of the family.

D.2 Proofs for local learnability condition from nonsingular parametrization of cliques

Proof of Proposition 1: LLC in $\ell_{\infty,2}$-norm. For a given vertex $i \in \mathcal{V}$, let $\underline{x} \in \mathcal{X}_i \subseteq \mathbb{R}^{\mathbf{K}_i}$ be a vector in the perturbation set. First, suppose that $\{i\}$ is not a maximal clique and choose any maximal clique $c \in \mathcal{M}_{\mathrm{cli}}(\mathcal{G})$ that contains the vertex i and let the set $S = c \setminus \{i\}$ be the set of nodes in the clique without i. The expression characterizing the LLC can be evaluated conditioning the expectation over S. Denoting the marginal and the conditional probability distribution used to compute expectation by subscripts, we find,

$$\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] = \mathbb{E}_{(\underline{\sigma}_{\setminus S})}\left[\mathbb{E}_{(\underline{\sigma}_S|\underline{\sigma}_{\setminus S})}\left[\left(\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right]\right], \tag{81}$$

$$\geq \mathbb{E}_{(\underline{\sigma}_{\setminus S})}\left[\prod_{j\in S}\frac{\exp(-2\gamma)}{|\mathcal{A}_j|}\sum_{\underline{\sigma}_S\in\mathcal{A}_S}\left(\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right], \tag{82}$$

where in the last line, we bounded the probability of $\mathbb{P}(\underline{\sigma}_S \mid \underline{\sigma}_{\setminus S})$ using Lemma 1. We want to rewrite the sum over $\underline{\sigma}_S$ in Eq. (82) using globally centered functions h_k for factors $k \in [c]_{\mathrm{sp}}$ instead of locally centered functions g_{ik}. Using definitions of locally centered functions in Eq. (3) and globally centered functions in Eq. (22), we see that $g_{ik}(\underline{\sigma}_k) = h_k(\underline{\sigma}_k) + R_{ik}(\underline{\sigma}_k)$, where

$$R_{ik}(\underline{\sigma}_k) = -\sum_{\substack{r\in P(\partial k)\setminus\emptyset \\ r\neq\{i\}}} \frac{(-1)^{|r|}}{|\mathcal{A}_r|}\sum_{\underline{\sigma}_r} f_k(\underline{\sigma}_k). \tag{83}$$

The sum in Eq. (82) can be expanded into the four contributions,

$$\sum_{\underline{\sigma}_S\in\mathcal{A}_S}\left(\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2 = \sum_{\underline{\sigma}_S\in\mathcal{A}_S}\left(\sum_{k\in[c]_{\mathrm{sp}}} x_k h_k(\underline{\sigma}_c)\right)^2 \tag{84}$$

$$+ \sum_{k\in[c]_{\mathrm{sp}}}\sum_{l\in\mathcal{K}_i\setminus[c]_{\mathrm{sp}}} x_k x_l \sum_{\underline{\sigma}_S\in\mathcal{A}_S} h_k(\underline{\sigma}_c) g_{il}(\underline{\sigma}_l) \tag{85}$$

$$+ \sum_{k\in[c]_{\mathrm{sp}}}\sum_{l\in[c]_{\mathrm{sp}}} x_k x_l \sum_{\underline{\sigma}_S\in\mathcal{A}_S} h_k(\underline{\sigma}_c) R_{il}(\underline{\sigma}_c) \tag{86}$$

$$+ \sum_{\underline{\sigma}_S\in\mathcal{A}_S}\left(\sum_{k\in\mathcal{K}_i\setminus[c]_{\mathrm{sp}}} x_k g_{ik}(\underline{\sigma}_k) + \sum_{k\in[c]_{\mathrm{sp}}} x_k R_{ik}(\underline{\sigma}_c)\right)^2. \tag{87}$$

We start by evaluating the contribution from terms in Eq. (85). For $k \in [c]_{\mathrm{sp}}$ and $l \in \mathcal{K}_i \setminus [c]_{\mathrm{sp}}$, there exists at least one node $u \in c$ such that $u \neq i$ and $u \notin \partial l$. Summing over the variable σ_u cancels the

expression,

$$\sum_{\sigma_u \in \mathcal{A}_u} h_k(\underline{\sigma}_c) g_{il}(\underline{\sigma}_l) = 0, \tag{88}$$

as $h_k(\underline{\sigma}_c)$ is globally centered and $g_{il}(\underline{\sigma}_l)$ does not depend on σ_u.

The contribution from Eq. (86) is also null. To see this we expand the sums using the formula for the reminder in Eq. (83),

$$\sum_{\underline{\sigma}_S \in \mathcal{A}_S} h_k(\underline{\sigma}_c) R_{il}(\underline{\sigma}_c) = - \sum_{\substack{r \in P(\partial l) \setminus \emptyset \\ r \neq \{i\}}} \frac{(-1)^{|r|}}{|\mathcal{A}_r|} \sum_{\underline{\sigma}_S \in \mathcal{A}_S} h_k(\underline{\sigma}_c) \sum_{\underline{\sigma}_r} f_l(\underline{\sigma}_c) = 0, \tag{89}$$

where the sum over $\underline{\sigma}_S$ vanishes as $h_k(\underline{\sigma}_c)$ depends on $\underline{\sigma}_r \neq \sigma_i$ while $\sum_{\underline{\sigma}_r} f_l(\underline{\sigma}_c)$ does not. As the contribution from Eq.(87) is non-negative, we can lower-bound Eq. (82) by the following expression,

$$\mathbb{E}\left[\left(\sum_{k \in \mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] \geq \prod_{j \in S} \frac{\exp(-2\gamma)}{|\mathcal{A}_j|} \mathbb{E}_{(\underline{\sigma}_{\setminus S})}\left[\sum_{\underline{\sigma}_S \in \mathcal{A}_S}\left(\sum_{k \in [c]_{\text{sp}}} x_k h_k(\underline{\sigma}_c)\right)^2\right], \tag{90}$$

$$\geq \prod_{j \in S} \frac{\exp(-2\gamma)}{|\mathcal{A}_j|} \rho_i^{\text{NPC}} \sum_{k \in [c]_{\text{sp}}} x_k^2, \tag{91}$$

where in the last line we have recognized the definition of the NPC constant from Eq. (23). Since (91) holds for any $c \in \mathcal{M}_{\text{cli}}(\mathcal{G})$ that contains the vertex i, the Local Learnability Condition is satisfied for a weighted $\ell_{\infty,2}$-norm with LLC constant equal to ρ^{NPC},

$$\mathbb{E}\left[\left(\sum_{k \in \mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] \geq \rho_i^{\text{NPC}} \|\underline{x}_{\mathcal{T}_i}\|^2_{w(\infty,2)}, \tag{92}$$

where the weighted $\ell_{\infty,2}$-norm is defined as follows,

$$\|\underline{x}_{\mathcal{T}_i}\|_{w(\infty,2)} = \max_{\substack{c \ni i \\ c \in \mathcal{M}_{\text{cli}}(\mathcal{G})}} \sqrt{\prod_{j \in c \setminus \{i\}} \frac{\exp(-2\gamma)}{|\mathcal{A}_j|} \sum_{k \in [c]_{\text{sp}}} x_k^2}. \tag{93}$$

As the weighted $\ell_{\infty,2}$-norm in Eq. (93) is lower-bounded by the $\ell_{\infty,2}$-norm,

$$\|\underline{x}_{\mathcal{T}_i}\|^2_{w(\infty,2)} \geq \left(\frac{\exp(-2\gamma)}{q}\right)^{L-1} \|\underline{x}_{\mathcal{T}_i}\|^2_{\infty,2}, \tag{94}$$

we have that the LLC is also satisfied for the $\ell_{\infty,2}$-norm with LLC constant equal to $\rho_i^{\text{NPC}} \left(\frac{\exp(-2\gamma)}{q}\right)^{L-1}$.

When $\{i\}$ is a maximal clique, then $\mathcal{K}_i = [\{i\}]_{\text{sp}}$ and it straightforward to see that

$$\mathbb{E}\left[\left(\sum_{k \in \mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] = \mathbb{E}_{(\sigma_i)}\left[\left(\sum_{k \in [\{i\}]_{\text{sp}}} \underline{x}_k h_k(\sigma_i)\right)^2\right] \tag{95}$$

$$\geq \rho_i^{\text{NPC}} \sum_{k \in [\{i\}]_{\text{sp}}} x_k^2. \tag{96}$$

□

Lemma 7. *Let $\sigma \in \mathcal{A}$, be a discrete random variable with probability distribution $p(\sigma)$. Consider $x_\sigma \in \mathbb{R}$, a function defined over σ that is centered, i.e. $\sum_{\sigma \in \mathcal{A}} x_\sigma = 0$. The variance of the function x_σ is lower-bounded by,*

$$\text{Var}[x_\sigma] \geq p_{min} \sum_{\sigma \in \mathcal{A}} x_\sigma^2, \tag{97}$$

where $p_{min} = \min_{\sigma \in \mathcal{A}} p(\sigma)$.

Proof. The proof goes as follows,

$$\mathrm{Var}\left[x_\sigma\right] = \sum_{\sigma\in\mathcal{A}} p(\sigma)\left(x_\sigma - \sum_{\sigma'\in\mathcal{A}} p(\sigma')x_{\sigma'}\right)^2 \tag{98}$$

$$\geq p_{\min}\sum_{\sigma\in\mathcal{A}}\left(x_\sigma - \sum_{\sigma'\in\mathcal{A}} p(\sigma')x_{\sigma'}\right)^2 \tag{99}$$

$$= p_{\min}\sum_{\sigma\in\mathcal{A}}\left(x_\sigma^2 - 2x_\sigma\sum_{\sigma'\in\mathcal{A}} p(\sigma')x_{\sigma'} + \left(\sum_{\sigma'\in\mathcal{A}} p(\sigma')x_{\sigma'}\right)^2\right), \tag{100}$$

$$\geq p_{\min}\sum_{\sigma\in\mathcal{A}} x_\sigma^2, \tag{101}$$

where in the last line we used that $\sum_{\sigma\in\mathcal{A}} x_\sigma = 0$ and $\left(\sum_{\sigma'\in\mathcal{A}} p(\sigma')x_{\sigma'}\right)^2 \geq 0$. □

Proof of Proposition 2: LLC in ℓ_2-norm for pairwise models. For a given vertex $i \in \mathcal{V}$, let $\underline{x} \in \mathcal{X}_i \subseteq \mathbb{R}^{\mathbf{K}_i}$ be a vector in the perturbation set. When $\{i\}$ is a maximal clique, then $\mathcal{K}_i = [\{i\}]_{\mathrm{sp}}$ and we immediately see that

$$\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] = \mathbb{E}_{(\sigma_i)}\left[\left(\sum_{k\in[\{i\}]_{\mathrm{sp}}} \underline{x}_k h_k(\sigma_i)\right)^2\right] \tag{102}$$

$$\geq \rho_i^{\mathrm{NPC}} \sum_{k\in[\{i\}]_{\mathrm{sp}}} x_k^2. \tag{103}$$

Now suppose that $\{i\}$ is not a maximal clique, i.e. there exists $j \in \mathcal{V}$ such that $\{i,j\} \in \mathcal{M}_{\mathrm{cli}}(\mathcal{G}^*)$. The expectation that arises in the LLC is lower-bounded by its variance,

$$\mathbb{E}\left[\left(\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right)^2\right] \geq \mathrm{Var}\left[\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right]. \tag{104}$$

Let $\{S_r\}_{r=1,\ldots,\chi}$ be a minimal coloring of the graph $\mathcal{G}^*$. For a given color r, define the set $C_r = S_r \setminus \{i\}$ and apply the law of total variance on the right-hand side of Eq. (104), conditioning on $\underline{\sigma}_{\setminus C_r}$,

$$\mathrm{Var}\left[\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right] \geq \mathbb{E}_{(\underline{\sigma}_{\setminus C_r})}\left[\mathrm{Var}_{(\underline{\sigma}_{C_r}|\underline{\sigma}_{\setminus C_r})}\left[\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right]\right], \tag{105}$$

where the marginal and the conditional probability distribution used to compute expectation and variance respectively are indicated by subscripts. As the variance on the right-hand side of Eq. (105) is conditioned on $\underline{\sigma}_{\setminus C_r}$, only basis functions involving a pair (σ_i, σ_j) with $j \in C_r$ are giving a non-zero contribution to the conditional variance,

$$\mathrm{Var}_{(\underline{\sigma}_{C_r}|\underline{\sigma}_{\setminus C_r})}\left[\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right] = \mathrm{Var}_{(\underline{\sigma}_{C_r}|\underline{\sigma}_{\setminus C_r})}\left[\sum_{j\in C_r}\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k g_{ik}(\sigma_i,\sigma_j)\right]. \tag{106}$$

We can rewrite the locally centered functions with respect to globally centered functions using their definitions found in Eq. (3) and in Eq. (22),

$$g_{ik}(\sigma_i,\sigma_j) = h_k(\sigma_i,\sigma_j) - \frac{1}{|\mathcal{A}_i|}\sum_{\sigma_j} f_k(\sigma_i,\sigma_j) + \frac{1}{|\mathcal{A}_i||\mathcal{A}_j|}\sum_{\sigma_i,\sigma_j} f_k(\sigma_i,\sigma_j). \tag{107}$$

We see from Eq. (107) that the difference between locally and globally centered functions only depends on the variable σ_i. This means that we can interchange locally centered functions with

globally centered functions in the right-hand side of Eq. (106) as the variance is conditioned on σ_i,

$$\mathrm{Var}_{(\underline{\sigma}_{C_r}|\underline{\sigma}_{\setminus C_r})}\left[\sum_{j\in C_r}\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k g_{ik}(\sigma_i,\sigma_j)\right] = \mathrm{Var}_{(\underline{\sigma}_{C_r}|\underline{\sigma}_{\setminus C_r})}\left[\sum_{j\in C_r}\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k h_k(\sigma_i,\sigma_j)\right]. \tag{108}$$

Since $\{S_r\}_{r=1,\dots,\chi}$ is a vertex coloring, by definition all nodes $j \in C_r$ having the same color are not sharing a factor node, i.e. $\forall j_1, j_2 \in C_r$, $\nexists k \in \mathcal{K}^*$ such that $j_1, j_2 \in \partial k$. This implies that variables σ_j with $j \in C_r$ are independent conditioned on the remaining variables $\underline{\sigma}_{\setminus C_r}$ and the variance in Eq. (108) can be rewritten,

$$\mathrm{Var}_{(\underline{\sigma}_{C_r}|\underline{\sigma}_{\setminus C_r})}\left[\sum_{j\in C_r}\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k h_k(\sigma_i,\sigma_j)\right] = \sum_{j\in C_r}\mathrm{Var}_{(\underline{\sigma}_j|\underline{\sigma}_{\setminus C_r})}\left[\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k h_k(\sigma_i,\sigma_j)\right]. \tag{109}$$

The right-hand side of Eq. (109) is centered with respect to σ_j and we can apply Lemma 7 and Lemma 1 to find a lower-bound that is only dependant on the random variable σ_i,

$$\sum_{j\in C_r}\mathrm{Var}_{(\underline{\sigma}_j|\underline{\sigma}_{\setminus C_r})}\left[\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k h_k(\sigma_i,\sigma_j)\right] \geq \frac{\exp(-2\gamma)}{q}\sum_{j\in C_r}\sum_{\sigma_j\in\mathcal{A}_j}\left(\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k h_k(\sigma_i,\sigma_j)\right)^2. \tag{110}$$

Plugging back the results derived in Eq. (108), Eq. (109), and Eq. (110) into the initial inequality in Eq. (105), we find,

$$\mathrm{Var}\left[\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right] \geq \mathbb{E}_{(\sigma_i)}\left[\frac{\exp(-2\gamma)}{q}\sum_{j\in C_r}\sum_{\sigma_j\in\mathcal{A}_j}\left(\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k h_k(\sigma_i,\sigma_j)\right)^2\right], \tag{111}$$

$$\geq \frac{\exp(-2\gamma)}{q}\rho_i^{\mathrm{NPC}}\sum_{j\in C_r}\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k^2, \tag{112}$$

where in Eq. (112) we used the definition of the NPC constant in Eq. (23) to bound the quadratic form involving $\underline{x}$.

Finally, we average the inequality described by Eq. (112) over the different colors and hence possible conditioning sets C_r to conclude the proof,

$$\mathrm{Var}\left[\sum_{k\in\mathcal{K}_i} x_k g_{ik}(\underline{\sigma}_k)\right] \geq \frac{\exp(-2\gamma)}{q}\frac{\rho_i^{\mathrm{NPC}}}{\chi}\sum_{r=1,\dots,\chi}\sum_{j\in C_r}\sum_{k\in[\{i,j\}]_{\mathrm{sp}}} x_k^2, \tag{113}$$

$$= \frac{\exp(-2\gamma)}{q}\frac{\rho_i^{\mathrm{NPC}}}{\chi}\sum_{\substack{k\in\mathcal{M}_{\mathrm{fac}}(\mathcal{G})\\ \partial k\ni i}} x_k^2. \tag{114}$$

□

D.3 Proofs of estimation guarantees for the SUPRISE algorithm

Proof of Theorem 2: Reconstruction and Estimation Guarantees for SUPRISE. As the NPC constant is non-zero $\rho_u^{\mathrm{NPC}} > 0$ for all nodes $u \in \mathcal{V}$, we apply Proposition 1 in conjunction with Theorem 1 to find that for each step $t \in \{0,\dots,L-1\}$ and with probability at least $1-\delta/(pL)$, GRISE around a node $u \in \mathcal{V}$ recovers the parameters in each maximal clique $c \in \mathcal{M}_{\mathrm{cli}}\left(\mathcal{G}\left[(\mathcal{V},\mathcal{K}^t)\right]\right)$ that contains u with precision $\sum_{k\in[c]_{\mathrm{sp}}}(\underline{\theta}_k - \underline{\widehat{\theta}}_k)^2 \leq (\alpha/2)^2$. Therefore, at each step $t \in \{0,\dots,L-1\}$ and with probability at least $1-\delta/L$, the factor removal procedure is guaranteed to remove all factors of size $L-t$ that are not present in the graph if all factors of size bigger than $L-t$ were correctly removed in the previous steps. Since there are at most L removal steps, it implies that the overall procedure discovers all maximal cliques with probability at least $1-\delta$.

□

E Application to special cases

In this section, we show how to apply Theorem 2 in order to derive the sample and computational complexity of reconstructing graphical models for some common basis functions.

E.1 Binary models on the monomial basis

In this subsection, we consider general models on binary alphabet $\mathcal{A}_i = \{-1, 1\}$. Let the factors be all nonempty subsets of $\{1, \dots, p\} = \mathcal{V}$ of size at most L,

$$\mathcal{K} = \{k \subseteq \mathcal{V} \mid |k| \leq L\}. \tag{115}$$

The set $\mathcal{K}$ contains all potential subsets of variable of size at most L. The parameterization uses the monomial basis given by $f_k(\underline{\sigma}_k) = \prod_{j \in k} \sigma_j$ with $k \in \mathcal{K}$. Note that the monomial basis functions are already globally centered $f_k \equiv g_k \equiv h_k$. The probability distribution for this model is expressed as

$$\mu_{\text{binary}}(\underline{\sigma}) = \frac{1}{Z} \exp\left(\sum_{k \in \mathcal{K}} \theta^*_k \prod_{j \in k} \sigma_j\right). \tag{116}$$

When $L \leq 2$, the model in Eq. (116) is pairwise and it is referred as the Ising Model.

For each maximal clique there exists exactly one maximal factors in its span. Therefore, the NPC constant as defined in Eq. (23) is $\rho_{\text{NPC}} = 1$ since for any clique c we have,

$$\sum_{\underline{\sigma}_{c \setminus u} \in \mathcal{A}_{c \setminus u}} \left(\sum_{k \in [c]_{\text{sp}}} x_k h_k(\underline{\sigma}_k)\right)^2 = x_k^2 \sum_{\underline{\sigma} \in \{-1,1\}^{|c|-1}} \left(\prod_{i=1}^{|c|} \sigma_i\right)^2 = 2^{|c|-1} x_k^2, \tag{117}$$

and the minimum is achieved for cliques of size one. As every node is involved in at most $\mathbf{K} \leq p^{L-1}$ factor functions, the structure of binary models can be recovered as a corollary of Theorem 2.

Corollary 1 (**Structure recovery for binary graphical models**). *Let $\underline{\sigma}^{(1)}, \dots, \underline{\sigma}^{(n)}$ be i.i.d. samples drawn according to $\mu(\underline{\sigma})$ in* (116) *and let $\alpha \leq \min_{k \in \mathbb{S}(\mathcal{G}^*)} |\theta^*_k|$ be the intensity of the smallest non-zero parameter. If*

$$n \geq 2^{12} 4^L \frac{\widehat{\gamma}^2 (1 + \widehat{\gamma})^2 e^{4\gamma L}}{\alpha^4} \log\left(\frac{4Lp^{2L-1}}{\delta}\right), \tag{118}$$

then the structure of the binary graphical model is perfectly recovered using Algorithm 1, i.e. $\widehat{\mathbb{S}} = \mathbb{S}(\underline{\theta}^*)$, with probability $1 - \delta$. Moreover the total computational complexity scales as $\widetilde{\mathcal{O}}(p^L)$, for fixed L, α, γ, $\widehat{\gamma}$ and δ.

For pairwise Ising models that are χ colorable, we have also guarantees on the ℓ_2-norm reconstruction by SUPRISE of pairwise parameters.

Corollary 2 (ℓ_2-parameter estimation for Ising models). *Let $\underline{\sigma}^{(1)}, \dots, \underline{\sigma}^{(n)}$ be i.i.d. samples drawn according to $\mu(\underline{\sigma})$ in* (116) *for $L = 2$ and let $\alpha > 0$ be the prescribed estimation accuracy. If*

$$n \geq 2^{16} \frac{\widehat{\gamma}^2 (1 + \widehat{\gamma})^2 \chi^2 e^{8\gamma}}{\alpha^4} \log\left(\frac{8p^3}{\delta}\right), \tag{119}$$

then, with probability at least $1 - \delta$, the parameters are estimated by Algorithm 1 with the error

$$\sqrt{\sum_{i,j \in V} |\widehat{\underline{\theta}}_{ij} - \underline{\theta}^*_{ij}|^2} \leq \frac{\alpha}{2}. \tag{120}$$

The computational complexity of obtaining these estimates is $\widetilde{\mathcal{O}}(p^2)$ for fixed χ, α, γ, $\widehat{\gamma}$ and δ.

As graphs with bounded degree d have a chromatic number at most $d + 1 \geq \chi$, Corollary 2 recovers the ℓ_2-guarantees for sparse graphs recovery of [18] albeit with slightly worse dependence with respect to γ and α. The worse γ dependence is an artifact of the general analysis presented in this paper. For models over binary variables one can improve the $e^{8\gamma}$ dependence to $e^{6\gamma}$ using Berstein's

inequality in Proposition 3 instead of Hoeffding's inequality. However, the worse α dependence seems to be more fundamental. It is caused by the replacement of the ℓ_1-penalty used in [18] by an ℓ_1-constraint.

For graphs with unbounded vertex degree but low chromatic number, such as star graphs or bipartite graphs, Corollary 2 shows that the parameters of the corresponding Ising model can be fully recovered with a bounded ℓ_2-error using a number of samples that is only logarithmic in the model size p.

E.2 L-wise models with arbitrary alphabets on the indicator basis

In this subsection, we consider L-wise graphical models over variables taking values in arbitrary alphabet $\mathcal{A}_i$ of size q_i, parametrized with indicator-type functions. The building block of the set of basis functions is the centered univariate indicator function defined as

$$\Phi_{s_i,\sigma_i} = \begin{cases} 1 - \frac{1}{q_i}, & \text{if } \sigma_i = s_i, \\ -\frac{1}{q_i}, & \text{otherwise,} \end{cases} \tag{121}$$

where $s_i, \sigma_i \in \mathcal{A}_i$ are prescribed letters of the alphabet. The univariate indicator functions in Eq. (121) are centered Kronecker delta functions and possess similar properties such as symmetry $\Phi_{s_i,\sigma_i} = \Phi_{\sigma_i,s_i}$ and contraction under a summation,

$$\sum_{\tau_i \in \mathcal{A}_i} \Phi_{\tau_i,s_i}\Phi_{\tau_i,\sigma_i} = \Phi_{s_i,\sigma_i}. \tag{122}$$

The set of factors $\mathcal{K}$ are pairs associating elements of $R = \{r \in P(\mathcal{V}) \mid |r| \leq L\}$ which are subsets of variable of size at most L with an alphabet configuration in $\mathcal{A}_r = \bigotimes_{i\in r} \mathcal{A}_i$,

$$\mathcal{K} = \{(r, \underline{s}_r) \mid r \in R, \underline{s}_r \in \mathcal{A}_r\}. \tag{123}$$

In what follows, we slightly abuse the notation of factors and parameters by shortening $(r, \underline{s}_r) \equiv \underline{s}_r$. With these notations, the indicator basis functions are constructed as $f_{\underline{s}_r}(\underline{\sigma}_r) = \prod_{i\in r} \Phi_{s_i,\sigma_i}$. Note that the indicator basis functions are globally centered i.e. $f_{\underline{s}_r} \equiv g_{\underline{s}_r} \equiv h_{\underline{s}_r}$. The probability distribution of an L-wise graphical model with arbitrary alphabet is defined as follows,

$$\mu_{\text{general}}(\underline{\sigma}) = \frac{1}{Z} \exp\left(\sum_{r\in R} \sum_{\underline{s}_r \in \mathcal{A}_r} \theta^*_{\underline{s}_r} \prod_{i\in r} \Phi_{s_i,\sigma_i} \right). \tag{124}$$

The family of distribution in Eq. (124) is not uniquely parameterized by the parameters $\underline{\theta}^*$. To see this, we introduce the linear application $\mathcal{P}_r$ acting on arrays $\theta_{\underline{s}_r}$ as follows,

$$[\mathcal{P}_r\underline{\theta}]_{\underline{\sigma}_r} = \sum_{\underline{s}_r \in \mathcal{A}_r} \theta_{\underline{s}_r} \prod_{i\in r} \Phi_{s_i,\sigma_i}. \tag{125}$$

Using the contraction property from Eq. (122), it is easy to see that $\mathcal{P}_r$ is a projector, i.e $\mathcal{P}_r^2 = \mathcal{P}_r$. It is also straightforward to verify that $\mathcal{P}_r\underline{\theta}$ is always a globally centered array and if $\underline{\theta}$ is already globally centered then $\mathcal{P}_r\underline{\theta} = \underline{\theta}$. Therefore, the applications $\mathcal{P}_r$ are projectors on the space of array $\theta_{\underline{s}_r}$ which are globally centered, i.e. $\sum_{s_i} \theta_{\underline{s}_r} = 0$ for all $i \in r$. We lift the parametrization degeneracy in Eq. (124) by imposing that parameters $\underline{\theta}^*$ are in the range of the projector $\mathcal{P}_r$. We thus require that the parameters satisfy the following linear constraints at each vertex $u \in \mathcal{V}$,

$$\mathcal{Y}_u = \bigcap_{\substack{r\in R \\ r\ni u}} \left\{ \underline{\theta}_u \in \mathbb{R}^{\mathbf{K}_u} \;\middle|\; \forall i \in r, \sum_{s_i \in \mathcal{A}_i} \theta_{\underline{s}_r} = 0 \right\}. \tag{126}$$

The constraint set in Eq. (126) is parametrically complete according to Definition 2 as we explicitly exhibited the equi-cost projection $\{\mathcal{P}_r\}_{r\in R}$ onto it. The computational complexity of this projection is no more than $\mathcal{O}(p^{L-1}q^L)$.

As the constraint set in Eq. (126) forms a linear subspace, the perturbation set is simply $\mathcal{X}_u = \mathcal{Y}_u \cap B_1(2\widehat{\gamma})$, the intersection of the constraint set with the ℓ_1-ball of radius $2\widehat{\gamma}$. Maximal cliques

are subset of vertices and hence are also elements of R. Therefore, the NPC constant as defined in Eq. (23) is bounded by $\rho_{\mathrm{NPC}} \geq \exp(-2\gamma)/q$ since for each clique we have,

$$\mathbb{E}_{(\sigma_i)}\left[\sum_{\underline{\sigma}_{c\setminus i}\in\mathcal{A}_{c\setminus i}}\left(\sum_{k\in[c]_{\mathrm{sp}}} x_k h_k(\underline{\sigma}_k)\right)^2\right] \geq \frac{\exp(-2\gamma)}{q_i}\sum_{\underline{\sigma}_c\in\mathcal{A}_c}\left([\mathcal{P}_c\underline{x}]_{\underline{\sigma}_c}\right)^2, \tag{127}$$

$$= \frac{\exp(-2\gamma)}{q_i}\sum_{\underline{\sigma}_c\in\mathcal{A}_c} x^2_{\underline{\sigma}_c}, \tag{128}$$

as $\underline{x} \in \mathcal{X}_u$ is globally centered and thus is in the range of the projector $\mathcal{P}_c$. Every node is involved in at most $\mathbf{K} \leq p^{L-1}q^L$ factor functions and the structure of L-wise models with arbitrary alphabets can be recovered as a corollary of Theorem 2.

Corollary 3 (**Structure recovery for L-wise graphical models**). *Let $\underline{\sigma}^{(1)},\ldots,\underline{\sigma}^{(n)}$ be i.i.d. samples drawn according to $\mu(\underline{\sigma})$ in* (124) *and let $\alpha \leq \min_{c\in\mathbb{S}(\mathcal{G}^*)}\sqrt{\sum_{\underline{s}_c\in\mathcal{A}_c}{\theta^*_{\underline{s}_c}}^2}$ be the intensity of the smallest non-zero parameter. If*

$$n \geq 2^{14}q^{2L}\frac{\widehat{\gamma}^2(1+\widehat{\gamma})^2 e^{4\gamma(L+1)}}{\alpha^4}\log\left(\frac{4Lq^{2L}p^{2L-1}}{\delta}\right), \tag{129}$$

then the structure of the L-wise graphical model with arbitrary alphabets is perfectly recovered using Algorithm 1, i.e. $\widehat{\mathbb{S}} = \mathbb{S}(\underline{\theta}^*)$, with probability $1-\delta$. Moreover the total computational complexity scales as $\widetilde{\mathcal{O}}(p^L)$, for fixed L, q, α, γ, $\widehat{\gamma}$ and δ.

For pairwise models with arbitrary alphabet that are χ colorable, we have also guarantees on the ℓ_2-norm reconstruction by SUPRISE of pairwise parameters.

Corollary 4 (ℓ_2-parameter estimation for pairwise models). *Let $\underline{\sigma}^{(1)},\ldots,\underline{\sigma}^{(n)}$ be i.i.d. samples drawn according to $\mu(\underline{\sigma})$ in* (124) *for $L=2$ and let $\alpha > 0$ be the prescribed estimation accuracy. If*

$$n \geq 2^{14}q^4\frac{\widehat{\gamma}^2(1+\widehat{\gamma})^2\chi^2 e^{12\gamma}}{\alpha^4}\log\left(\frac{8q^4p^3}{\delta}\right), \tag{130}$$

then, with probability at least $1-\delta$, the parameters are estimated by Algorithm 1 with the error

$$\sqrt{\sum_{i,j\in V}\sum_{s_i\in\mathcal{A}_i,s_j\in\mathcal{A}_j}|\widehat{\underline{\theta}}_{s_i,s_j}-\underline{\theta}^*_{s_i,s_j}|^2} \leq \frac{\alpha}{2}. \tag{131}$$

The computational complexity of obtaining these estimates is $\widetilde{\mathcal{O}}(p^2)$ for fixed χ, q, α, γ, $\widehat{\gamma}$ and δ.