Estimating Structurally Similar Graphical Models

Saurabh Sihag and Ali Tajer, *Senior Member, IEEE*

***Abstract*—This paper considers the problem of estimating the structure of structurally similar graphical models in high dimensions. This problem is pertinent in multi-modal or multi-domain datasets that consist of multiple information domains, each modeled by one probabilistic graphical model (PGM), e.g., in brain network modeling using different neuroimaging modalities. Induced by an underlying shared causal source, the domains, and subsequently their associated PGMs, can have structural similarities. This paper focuses on Gaussian and Ising models and characterizes the information-theoretic sample complexity of estimating the structures of a pair of PGMs in the degree-bounded and edge-bounded subclasses. The PGMs are assumed to have p nodes with distinct and unknown structures. Their similarity is accounted for by assuming that a pre-specified set of q nodes form identical subgraphs in both PGMs. Necessary and sufficient conditions on the sample complexity for a bounded probability of error are characterized. The necessary conditions are information-theoretic (algorithm-independent), delineating the statistical difficulty of the problem. The sufficient conditions are based on deploying maximum likelihood decoders. While the specifics of the results vary across different subclasses and parameter regimes, one key observation is that in specific subclasses and regimes, the sample complexity varies with p and q according to to $\Theta(\log(p-q))$. For Ising models, a low complexity, online structure estimation (learning) algorithm based on multiplicative weights is also proposed. Numerical evaluations are also included to illustrate the interplay among different parameters on the sample complexity when the structurally similar graphs are recovered by a maximum likelihood-based graph decoder and the proposed online estimation algorithm.**

***Index Terms*—Structure learning, probabilistic graphical models, Gaussian, Ising, high-dimensional estimation.**

I. Introduction

Probabilistic graphical models (PGMs) are commonly used for capturing the conditional interdependence in probabilistic databases or random fields [1] and [2]. Each vertex of a PGM represents a random variable. The edge connectivity structure encodes the statistical dependence of these random variables, and the joint distribution of the random variables is fully characterized by the edge structure and parameters of the graph. PGMs have a growing list of applications as various technological, biological, and social systems are growing as complex systems of interconnected platforms in which highly structured data is constantly generated, communicated, and processed for various inferential and decision-making purposes.

Structurally similar networks. In some domains, the data is derived from multiple sources, owing to the proliferation of sensing technologies. Graphically modeling such databases, as a result, renders multiple PGMs, each corresponding to one data source. An example is brain network modeling using different neuroimaging modalities. The interplay between functional and structural connectivities of the brain can be leveraged to understand intrinsic brain functioning [3], [4] and relate it to different pathology and gender-related differences [5]. Another example is genomics, in which multiple genetic networks can be organized to form a multiplex network. Multiplex networks are often adopted to represent nodes that have similar inter-relationships in different contexts. For instance, a genetic network can be characterized by multiple gene expression measurements, essentially, a phylogenetic profile, a neighborhood in the interaction network, biological pathways involved, and a protein domain profile. Each type of measurement can form unique or similar links in different genes [6]. Multimodal data analysis strategies are also relevant in behavioral analysis [7]–[10]. For instance, the voting patterns of the members of the US Senate for various categories of bills can be modeled as a set of networks [7], where the graphical structures revealed common dependencies in voting patterns across different political affiliations. Emotion analysis frameworks that leverage different modalities such as language, visuals, and acoustics have also been investigated [8] and [9].

Structural similarity. Induced by a shared underlying physical or biological cause, the structures of graphical models derived from different sources are not always distinct, and often they bear similarities. For instance, different gene networks representing the same cancer subtypes share similar edges across all subtypes and have unique edges corresponding to each subtype [11]. Figure 1 illustrates the causal features and the two induced networks with structurally similar clusters. In the neuroimaging application, consider diffusion tensor imaging (DTI) and functional magnetic resonance imaging (fMRI) for brain imaging. DTI and fMRI images of a brain represent different structures of the network underlying the brain [12], and the conformity between the two images can be leveraged to assess an individual's cognitive health [13].

In this paper, we characterize the sample complexity of estimating the structure of structurally similar graphical models[1], in which the graph pair can share identical local structures accounting for their underlying similarity. We establish necessary

Saurabh Sihag is with the University of Pennsylvania, Philadelphia, PA 19104, USA (email: saurabh.sihag@pennmedicine.upenn.edu).

Ali Tajer is with Rensselaer Polytechnic Institute, Troy, NY 12180, USA (e-mail: tajer@ecse.rpi.edu).

This paper was presented in part at the 2019 IEEE International Symposium on Information Theory, the 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, and the 2019 Conference on Advances in Neural Information Processing Systems.

This research was supported in part by the U. S. National Science Foundation under Grants CAREER ECCS-1554482, ECCS-1933107, and by the IBM-RPI Artificial Intelligence Research Center.

[1]The problem of estimating the structures of graphical models is also referred to as model-based structure learning, especially in machine learning literature. In this paper, sometimes these terms are used interchangeably.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3207420

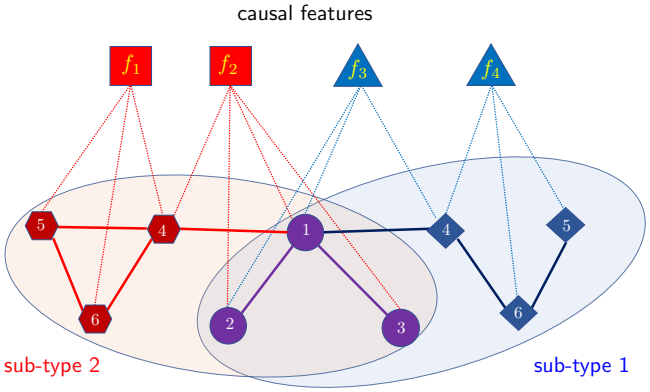

Fig. 1. Two graphs with partially similar structures. For both graphs, the causal features induce identical internal edge structures for purple nodes ($p = 6, q = 3$).

conditions (information-theoretic) and sufficient conditions on the sample complexity in high-dimensional Gaussian and Ising models under the bounded error probability criterion. The necessary conditions are algorithm-independent and establish the statistical difficulty of a problem. The sufficient conditions are characterized by adopting maximum likelihood (ML) decoders. While the problem of graphical model selection is NP-hard [14], the problem of learning Markov random fields (MRF) is well-posed and tractable strategies for inference have been studied in [15]–[17].

A. Related Literature

Information-theoretic analysis establishes the algorithm-independent guarantees on estimating the structures of graphical models. The existing information-theoretic studies on graphical models include those of [18]–[20], which analyze the sample complexity for selecting the model of a given graph in various subclasses of Ising models. Specifically, [18] establishes the necessary and sufficient conditions on the number of samples for the exact recovery of the Ising models under regimes characterized by a bounded degree and a bounded number of edges. These studies are generalized in [21] to establish necessary conditions for set-based graphical model selection, in which the graph estimator outputs a set of potentially true graphs instead of a unique graph. Necessary conditions for recovering girth-bounded graphs and path-restricted graphs are analyzed in [19]. The problem of graphical model selection for various subclasses of Ising models under the criterion of approximate recovery is investigated in [20], in which a certain number of missed edges or incorrectly included edges are tolerated in the estimated graph structure. Approximate recovery bounds on the sample complexity are characterized for Ising and Gaussian models without considering the effect of edge weights in [22]. The problems of structure recovery and inverse covariance matrix estimation for Gaussian models are studied in [23], where information-theoretic bounds on the sample complexity are delineated. Similarly, information-theoretic bounds are established for the class of power-law graphs in [24].

Algorithm-independent bounds on the sample complexity have also been investigated for inference tasks other than model selection in graphical models. The problem of detecting whether two Markov network structures are identical or different is investigated in [25], and its sample complexity is characterized. The problem of property testing for high-dimensional Ising models is investigated in [26], where information-theoretic limits for testing graph properties such as connectivity, cycle presence, and maximum clique size are established. In [27], the problem of density estimation using the data from the Ising model is analyzed, and the minimax rate of estimation is analyzed. Finally, active sampling strategies to detect the true Markov random field model are studied in [28]–[30].

The algorithmic aspects of high-dimensional model estimation in different subclasses of the Ising model are investigated in [15]–[17], [31]–[33]. Algorithms based on conditional independence testing for Ising models with correlation decay are studied in [31]. Estimating the structures of tree-structured Ising models is analyzed in [32]. Various algorithmic frameworks for Ising models with restrictions on the graph degree are proposed in [15]–[17], [33]. Specifically, Ising models with bounds on the average degree are studied in [33], where the correlation decay properties of Ising models are leveraged for the design of the algorithms. A greedy approach for estimating structures is studied in [16]. A convex optimization framework for structure estimation is analyzed in [17]. An online learning-based algorithm designed based on the principles of prediction with expert advice is studied in [15].

Estimating the structures of structurally similar graphical models, which is the problem that we focus on in this paper, is studied substantially from an algorithmic perspective [7], [11], [34]–[39]. Specifically, an empirical Bayes method is studied in [11] to identify interactions that are unique to each class and that are shared across all classes. Graphical Lasso-based algorithms are investigated in [34] and [35]–[37] for joint inference of Gaussian graphical models. An optimization-based approach to the joint estimation of the graph structures using discrete data is studied in [7]. A Bayesian approach to jointly estimate Gaussian graphical models is investigated in [38], where the models with shared structures are identified from the data groups, and their relative similarity is leveraged for inference. Approximately estimating the structures of partially similar Ising models has been studied from an information-theoretic perspective in [40], [41] where algorithm-independent bounds on the sample complexity are investigated.

B. Contributions

The existing studies on structure estimation (learning) in multiple graphical models focus on empirical or algorithmic frameworks for graph estimation or selection. Complementary to these studies, in this paper, we characterize the information-theoretic bounds on the sample complexity of jointly estimating the two partially similar graphs in the edge-bounded and degree-bounded subclasses of Ising models as well as Gaussian models. These results provide algorithm-independent necessary conditions on the sample complexity for an arbitrary degree of reliability in the recovered graphs for an exact

recovery criterion. Besides the information-theoretic bounds, we analyze an ML-based graph decoder for the exact recovery of partially similar graphs in different subclasses. The analysis of ML-based graph decoders establishes sufficient conditions on the sample complexity.

Based on the necessary and sufficient conditions, we investigate the variation in the number of samples necessary and sufficient for structure recovery with respect to structural similarity. We also analyze the asymptotic scaling behavior of the bounds on sample complexity. Our analysis reveals that for two graphs with p nodes and the structural similarity spanning $q < p$ nodes, the sample complexity varies with p and q according to $\log(p-q)$ in most regimes. This indicates how the sample complexity improves as the structural similarity increases. Our results, in its special cases when $p = q$, reduce to the known results for estimating a single graphical model in different subclasses of Ising models and Gaussian models established in [18] and [23], respectively. Although an ML decoder is optimal for structure estimation under the exact recovery criterion, it may be intractable to implement for general graphs. Therefore, we also propose a computationally efficient joint structure estimation algorithm for structurally similar Ising models that uses similar principles of the Hedge algorithm as in [15]. We also evaluate the performance of our algorithm and ML decoder in numerical experiments.

Besides the joint estimation results, to the best of our knowledge, there exist no parallel results on the sufficient conditions for recovering single graphs in certain classes of Gaussian models (Theorems 8, 9, and 10).

C. Connection to the Results on Single-graph Estimation

We investigate the connection between the sample complexities of jointly recovering two structurally similar graphs and recovering them independently. In the Gaussian models, we show that joint recovery strictly improves upon independent recovery. This is established by showing that our sufficient condition for joint recovery is strictly smaller than the necessary condition for independent recovery. Furthermore, noting that these conditions (even for independent recovery) are not tight (and quite loose in a wide range of settings), the actual performance gains are considerably higher than the gap between the pertinent sufficient and necessary conditions.

We note that the results are provided for the general non-asymptotic regimes of the graph size. By focusing on the extremities of graph similarity, we can recover the sample complexity of learning two identical single graphs. However, the converse is not valid. The analysis of single graphs provides no insight into the impact of structural similarity on sufficient or necessary conditions of joint structure learning. Specifically, one cannot predict the regimes in terms of graph similarity parameters or their presence in the sample complexity results from learning single graphs.

We conclude by noting that while there are close connections between the joint and single-graph recovery results, there are significant differences. Specifically, the two-graph estimation problem is an independent inference problem that has been investigated empirically, and the sample complexity results cannot be recovered from the known results for the single-graph setting.

II. Background and Problem Formulation

A. Markov Random Fields

Consider two sets of random variables $\mathbf{X}_1 \triangleq [X_1^1, \ldots, X_1^p]$ and $\mathbf{X}_2 \triangleq [X_2^1, \ldots, X_2^p]$ taking values in the set $\mathcal{X}^p$. Random variables $\mathbf{X}_1$ and $\mathbf{X}_2$ form Markov random fields with respect to two distinct undirected graphs $\mathcal{G}_1 \triangleq (V, E_1)$ and $\mathcal{G}_2 \triangleq (V, E_2)$, respectively. These graphs are specified over the same set of vertices $V \triangleq \{1, \ldots, p\}$, and the vertices in $\mathcal{G}_i$ are joined by the set of edges $E_i \subset V \times V$. The set of edges in E_i encodes the conditional independence among the random variables $\mathbf{X}_i$. The structures in both graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ (i.e., edge sets E_1 and E_2) are *unknown*. Finally, we denote the joint probability measure of $\mathbf{X}_i$ by $\mathbb{P}_i$.

Ising Markov Random Fields. In the Ising model, the random variables associated with vertices are binary, and $\mathbf{X}_i \in \{-1, 1\}^p$. The joint probability mass function (pmf) of the random variables $\mathbf{X}_i \triangleq [X_i^1, \ldots, X_i^p]$ associated with probability measure $\mathbb{P}_i$ and graph $\mathcal{G}_i$ is given by

$$p_i(\mathbf{x}_i) \triangleq \frac{1}{Z_i(\boldsymbol{\lambda}_i)} \exp\left(\sum_{(u,v)\in E_i} \lambda_i^{uv}\, x_i^u\, x_i^v \right) , \qquad (1)$$

where $[\boldsymbol{\lambda}_i]_{uv} \triangleq \lambda_i^{uv}$, and $Z_i(\boldsymbol{\lambda}_i)$ is the partition function given by

$$Z_i(\boldsymbol{\lambda}_i) \triangleq \sum_{\mathbf{x}_i \in \{-1,1\}^p} \exp\left(\sum_{(u,v)\in E_i} \lambda_i^{uv}\, x_i^u\, x_i^v \right) . \qquad (2)$$

Parameter $\lambda_i^{uv} \in \mathbb{R}^+$ specifies the inter-dependence between X_i^u and X_i^v conditioned on all other random variables associated with nodes $s \in V \backslash \{u, v\}$ in graph $\mathcal{G}_i$. As discussed in [18] and shown in our analyses, the sample complexity for recovering the structure of the graphs depends on the following two quantities, which in turn depend on the interaction parameters λ_i^{uv}.

Definition 1 (Minimum interaction)**.** *We define the minimum interaction constant as*

$$\lambda \triangleq \min_{i\in\{1,2\}} \min_{(u,v)\in E_i} \lambda_i^{uv} . \qquad (3)$$

This parameter captures the weakest interactions between any two interacting random variables. As illustrated in [18], in the asymptote of large or small values of λ, recovering the graph's structure from its samples becomes increasingly more difficult.

Definition 2 (Maximum neighborhood weight)**.** *We define the maximum neighborhood weight as*

$$\vartheta \triangleq \max_{i\in\{1,2\}} \max_{u\in V} \sum_{v:(u,v)\in E_i} \lambda_i^{uv} . \qquad (4)$$

Gauss-Markov Random Fields. In the Gauss-Markov random fields, random variables $\mathbf{X}_i = [X_i^1, \ldots, X_i^p]$ have a joint Gaussian distribution with the joint probability density

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3207420

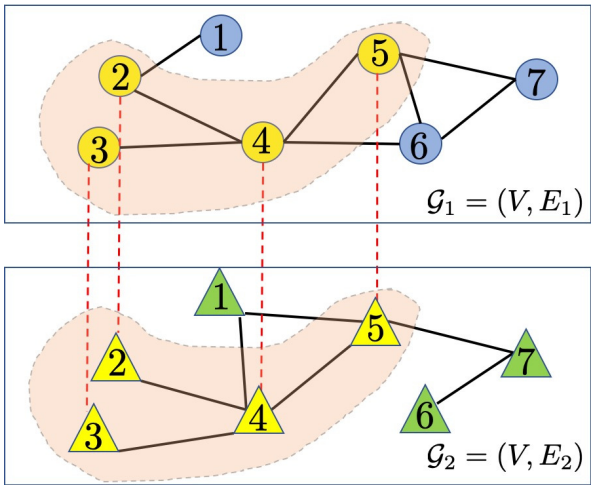

Fig. 2. Two graphs with partially similar structures. Yellow nodes in both graphs have identical internal edge structures ($p = 7, q = 4$).

function (pdf) associated with probability measure $\mathbb{P}_i$ and graph $\mathcal{G}_i$ given by

$$f_i(\mathbf{x}_i) \;=\; \frac{\det(\mathbf{P}_i)^{\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\mathbf{x}_i^{\mathsf{T}}\mathbf{P}_i\mathbf{x}_i\right) \;, \tag{5}$$

where $\mathbf{P}_i$ is the precision matrix associated with $\mathcal{G}_i$. The off-diagonal entries of $\mathbf{P}_i$ represent the edge structure of $\mathcal{G}_i$, i.e., the element at a coordinate (u, v) in $\mathbf{P}_i$, given by $\mathbf{P}_i(u, v)$, is non-zero if and only if $(u, v) \in E_i$. Structure recovery in the Gaussian model depends on the quantity formalized next, which reflects the scale-invariant minimum value of the elements in matrix $\mathbf{P}_i$.

Definition 3 (Minimum partial correlation)**.** *We define the minimum partial correlation factor as*

$$\rho \;\triangleq\; \min_{i\in\{1,2\}} \min_{(u,v)\in E_i} \frac{|\,\mathbf{P}_i(u,v)\,|}{\sqrt{\mathbf{P}_i(u,u)\mathbf{P}_i(v,v)}} \;. \tag{6}$$

B. Graph Similarity Models and Classes

Our objective of structure estimation is using the data (i.e., realizations of $\mathbf{X}_1$ and $\mathbf{X}_2$) to recover the unknown structures of $\mathcal{G}_1$ and $\mathcal{G}_2$. In this subsection, we formalize the similarity models and the classes of the Ising and Gaussian graphical models on which we will be focusing in this paper. Induced by an underlying shared system that generates both datasets, the two graphical models are assumed to have structural similarities. Specifically, it is assumed that they have identical structures in a pre-specified cluster of nodes denoted by $V_{\mathrm{s}} \subseteq V$. This means that the internal structures of the sub-graphs formed by nodes V_{s} are identical in both graphs. The rest of the two graphs may or may not have similarities. An example of a pair of structurally similar graphs is shown in Fig. 2. Graph similarity is formalized next.

Definition 4 (q-similar graphs)**.** *Graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be q-similar, for some $q \in [p]$, when the size of the shared cluster V_{s} is q, i.e., $|V_{\mathrm{s}}| = q$.*

Classes of Ising models. We denote the general class of Ising models by the minimum interaction constant λ and the maximum neighborhood weight ϑ by $\mathcal{I}(\lambda, \vartheta)$. Accordingly, we denote the class of pairs of q-similar Ising models by $\mathcal{I}_q(\lambda, \vartheta)$. In this paper, we focus on the following subclasses of $\mathcal{I}_q(\lambda, \vartheta)$.

- *Degree-bounded class* $\mathcal{I}_q^d(\lambda, \vartheta)$. This class contains all q-similar pairs of Ising graphical models $\mathcal{G}_1$ and $\mathcal{G}_2$, where each graph has the maximum degree d.
- *Edge-bounded class* $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$. This class contains all q-similar pairs of Ising graphical models $\mathcal{G}_1$ and $\mathcal{G}_2$, where each graph has at most k edges and the shared cluster V_{s}, has γk edges, where $\gamma \in [0, 1]$.

In $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$, we introduce γk to account for scenarios where not all k edges may be accommodated inside the shared region due to the restrictions imposed by the number of nodes q and let our analyses guide the appropriate regimes for γ in the sample complexity.

Classes of Gaussian models. We denote the class of q-similar Gaussian models with minimum partial correlation ρ by $\mathcal{G}_q(\rho)$. In this paper, we consider the following subclass of $\mathcal{G}_q(\rho)$.

- *Degree-bounded class* $\mathcal{G}_q^d(\rho)$. This class contains all q-similar pairs of Gaussian graphical models $\mathcal{G}_1$ and $\mathcal{G}_2$, where each graph has the maximum degree d.
- *Edge-bounded class* $\mathcal{G}_q^k(\rho)$. This class contains all q-similar pairs of Gaussian graphical models $\mathcal{G}_1$ and $\mathcal{G}_2$, where each graph has at most k edges.

C. Structure Estimation Criterion

For a given class $\mathcal{S}_q \in \{\mathcal{I}_q^d(\lambda, \vartheta), \mathcal{I}_{q,\gamma}^k(\lambda, \vartheta), \mathcal{G}_q^d(\rho), \mathcal{G}_q^k(\rho)\}$ the nature selects a pair of graphical models from $\mathcal{S}_q$. Our estimation objective consists of collecting n samples from both graphs, denoted by $\mathbf{x}_1^n$ and $\mathbf{x}_2^n$ and jointly forming estimates $\hat{E}_1$ and $\hat{E}_2$ corresponding to the structures E_1 and E_2, respectively. This process is formalized next.

Definition 5 (Graph decoding)**.** *We define a graph decoder ψ as a function that maps the data $(\mathbf{x}_1^n, \mathbf{x}_2^n)$ to a pair of graphs in $\mathcal{S}_q$, i.e.,*

$$\psi : \mathcal{X}^{n\times p} \times \mathcal{X}^{n\times p} \to \mathcal{S}_q \;, \tag{7}$$

where $\mathcal{X} = \{-1, +1\}$ for the Ising models and $\mathcal{X} = \mathbb{R}$ for the Gaussian models.

To capture the accuracy of a decoder ψ , we adopt the exact recovery criterion. Specifically, for a generic class of q-similar graph pairs $\mathcal{S}_q$, we define $\mathsf{P}(\mathcal{S}_q)$ as the maximal probability of error in the exact recovery of (E_1, E_2), i.e.,

$$\mathsf{P}(\mathcal{S}_q) \;\triangleq\; \max_{(\mathcal{G}_1,\mathcal{G}_2)\in\mathcal{S}_q} \mathbb{P}(\psi(\mathbf{x}_1^n, \mathbf{x}_2^n) \neq (E_1, E_2)) \;, \tag{8}$$

where the probability is evaluated with respect to both measures $\mathbb{P}_1$ and $\mathbb{P}_2$. We are interested in analyzing sample complexity, that is the number of samples n required for achieving a target decision reliability $\mathsf{P}(\mathcal{S}_q)$.

Definition 6 (Sample complexity)**.** *We define $n(q, \varepsilon)$ as the number of samples required for recovering a pair of q-similar graphs in the class $\mathcal{S}_q$ such that $\mathsf{P}(\mathcal{S}_q) \leq \varepsilon$.*

We are interested in analyzing the scaling behavior of $n(q,\varepsilon)$ with respect to q and ε, as well as parameters relevant to each specific model (i.e., λ,ϑ in Ising and ρ in Gaussian) and each subclass (i.e., graph size p, degree bound d, and edge bound k).

III. Main Results: Ising Models

In this section, we provide the necessary and sufficient conditions on the sample complexity $n(q,\varepsilon)$ for different models and their subclasses of Ising models. Specifically, we provide algorithm-independent necessary conditions on the sample complexity for recovering the structures with the desired reliability. These sample complexity analyses establish the performance benchmarks for any structure estimation algorithm, presented in Section III-A. Furthermore, we also provide the counterpart sufficient conditions by adopting maximum likelihood (ML) decoders. These results are provided in Section III-B.

A. Necessary Conditions for Ising Models

We start by providing the necessary condition on the sample complexity $n(q,\varepsilon)$ that any decoder requires to ensure that $\mathsf{P}(\mathcal{S}_q) \leq \varepsilon$ in Ising models. Throughout this section, we use the shorthand n for $n(q,\varepsilon)$. We provide the necessary conditions for the exact recovery and remark on the scaling behavior of the sample complexity with respect to the various parameters involved. We start with the degree-bounded subclass $\mathcal{I}_q^d(\lambda,\vartheta)$, for which, depending on the relative scaling behavior of λ with respect to the degree bound d we have different necessary conditions.

Theorem 1 (Necessary condition for class $\mathcal{I}_q^d(\lambda,\vartheta)$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_q^d(\lambda,\vartheta)$. Any graph decoder ψ that achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda,\vartheta)) \leq \varepsilon$, has the following sample size requirements:*

1) *In the regime $\lambda = O(1/d)$, the sample size n satisfies:*
   a) *if $\binom{q}{2} \leq \binom{p-q}{2}^2$,*
   $$n \geq \frac{(1-\varepsilon)}{4\lambda\tanh(\lambda)}\left(4\log\frac{p-q-1}{\sqrt{2}}-1\right), \quad (9)$$
   b) *if $\binom{q}{2} > \binom{p-q}{2}^2$,*
   $$n \geq \frac{(1-\varepsilon)}{4\lambda\tanh(\lambda)}\left(2\log\frac{q}{\sqrt{2}}-1\right). \quad (10)$$
2) *In the regimes $\lambda=\Theta(1)$ and $\lambda=\Theta(d)$, the sample size n satisfies:*
   a) *if $q+2\sqrt{\frac{q}{d}} \leq p$,*
   $$n \geq (1-\varepsilon)\cdot\frac{\mathrm{e}^{\vartheta-\lambda}}{8\vartheta}\left(2\log\frac{d(p-q)}{4}-1\right), \quad (11)$$
   b) *if $q+2\sqrt{\frac{q}{d}} > p$,*
   $$n \geq (1-\varepsilon)\cdot\frac{\mathrm{e}^{\vartheta-\lambda}}{8\vartheta}\left(\log\frac{dq}{4}-1\right). \quad (12)$$

Proof. See Section V-A. □

This theorem specifies how the sample complexity scales with respect to varying λ. Next, we discuss the impact of other parameters involved (i.e., p, q, and d) on the necessary conditions for the sample complexity. It is noteworthy that the conditions in (10) and (12) hold only when graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ have a high degree of similarity (i.e., $q \approx p$). For instance, when $p \geq 128$, the condition in (10) holds only when $q \geq 0.9p$. This is further amplified in (12) due to the additional role that d plays in making the gap between q and p smaller. Hence, except for the case of almost identical graphs, the necessary conditions on the sample complexity are specified by (9) and (11). We remark that the extreme cases are of less interest since, in these settings, the structure estimation objective reduces to the well-investigated structure estimation in a single graph. Motivated by this, for the rest of the discussions, unless stated explicitly, we focus on analyzing (9) and (11). We start by evaluating the impact of increasing similarity level q on the sample complexity.

Corollary 1 (Sample complexity versus q)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_q^d(\lambda,\vartheta)$ with increasing similarity q. For any graph decoder ψ that achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda,\vartheta)) \leq \varepsilon$, the sample complexity scales with respect to q and p as $\Omega(\log(p-q))$.*

This corollary indicates that as the similarity level of the two graphs increases, the sample complexity requirement decreases. This signifies the gain of jointly recovering both structures instead of treating them in isolation, that is, for recovering the two structures separately, the sample complexity scales with $\Omega(\log p)$, while when recovering them jointly, it reduces to $\Omega(\log(p-q))$. Next, we evaluate the effect of increasing maximum degree d on the sample complexity.

Corollary 2 (Sample complexity versus d.)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_q^d(\lambda,\vartheta)$ with an increasing maximum degree d. Any graph decoder ψ that achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda,\vartheta)) \leq \varepsilon$, has the following sample size requirements:*

1) *In the regime $\lambda = O(1/d)$, the sample complexity scales with respect to d as $\Omega(d^2)$.*
2) *In the regimes $\lambda=\Theta(1)$ and $\lambda=\Theta(d)$, the sample complexity scales with respect to d exponentially.*

We note that the necessary condition on the sample complexity has a non-exponential behavior in d in the regime $\lambda = O(1/d)$. This observation is consistent with that for recovering single graphs in [18]. Next, we provide the counterpart necessary conditions for the edge-bounded class of pair of Ising models. For this purpose, we define

$$\bar{\gamma} \triangleq 1-\gamma. \quad (13)$$

Theorem 2 (Necessary condition for class $\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)$. Any graph decoder ψ that achieves $\mathsf{P}(\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)) \leq \varepsilon$, has the following sample size requirements:*

1) *In the regimes $k=o(p)$ or $\lambda = O(1/\sqrt{k})$, the sample size n satisfies:*

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3207420

a) *if* $\binom{q}{2} \leq \binom{p-q}{2}^2$,

$$n \geq \frac{(1-\varepsilon)}{4\lambda \tanh(\lambda)} \left(4 \log \frac{p-q-1}{\sqrt{2}} - 1\right), \tag{14}$$

b) *if* $\binom{q}{2} > \binom{p-q}{2}^2$,

$$n \geq \frac{(1-\varepsilon)}{4\lambda \tanh(\lambda)} \left(2 \log \frac{q-1}{\sqrt{2}} - 1\right). \tag{15}$$

2) *In the regime* $k = \Omega(p)$ *and* λ *satisfies either* $\lambda = \Theta(1)$ *or* $\Theta(\sqrt{k})$*, the sample size* n *satisfies:*

a) *if* $\gamma > 0.5$ *and* $2\lambda^2 k < \frac{\log^2(\gamma/\bar{\gamma})}{(1-2\sqrt{\gamma\bar{\gamma}})}$,

i) *if* $k \leq \frac{4\gamma}{\bar{\gamma}^2}$,

$$n \geq \frac{(1-\varepsilon)}{32 \exp(2\lambda) \sinh(\lambda)} \times \frac{\exp(\lambda\sqrt{k\gamma})}{\lambda\sqrt{k\gamma}} \cdot \left(\log \frac{k\gamma}{4} - 1\right), \tag{16}$$

ii) *if* $k > \frac{4\gamma}{\bar{\gamma}^2}$,

$$n \geq \frac{(1-\varepsilon)}{32 \exp(2\lambda) \sinh(\lambda)} \times \frac{\exp(\lambda\sqrt{k\gamma})}{\lambda\sqrt{k\gamma}} \cdot \left(2 \log \frac{k\bar{\gamma}}{4} - 1\right). \tag{17}$$

b) *if* $\gamma \leq 0.5$, $2\lambda^2 k \geq \frac{\log^2(\bar{\gamma}/\gamma)}{(1-2\sqrt{\gamma\bar{\gamma}})}$, *and* $k > \frac{4\gamma}{\bar{\gamma}^2}$,

$$n \geq \frac{(1-\varepsilon)}{32 \exp(2\lambda) \sinh(\lambda)} \times \frac{\exp(\lambda\sqrt{k\bar{\gamma}})}{\lambda\sqrt{k\bar{\gamma}}} \cdot \left(2 \log \frac{k\bar{\gamma}}{4} - 1\right). \tag{18}$$

Proof. See Section V-B. □

We observe that in the regime that either λ satisfies $\lambda = O(1/\sqrt{k})$ or k satisfies $k = o(p)$, the sample complexity necessary condition follows the same structure as in the regime $\lambda = O(1/d)$ in the degree-bounded class. In all other cases, the sample complexity is characterized by the maximum number of edges in the non-shared clusters, i.e., $\bar{\gamma}k$, and the number of edges in the shared cluster, i.e., γk. The results in (16), (17), and (18) provide the necessary conditions for different regimes of interest. However, the regimes listed in the theorem are not exhaustive. We remark that the regimes excluded from Theorem 2 when $k = \Omega(p)$ are either too restrictive or their corresponding sample complexity results are superseded by that in (14) and (15). For instance, the condition $k > \frac{4\gamma}{\bar{\gamma}^2}$ in (18) is satisfied by all settings with $k > 8$ and, therefore, the setting with the complementary condition $k \leq \frac{4\gamma}{\bar{\gamma}^2}$ when $\gamma \leq 0.5$ is too restrictive and not of interest. An exhaustive discussion of all possible regimes and their implications on the sample complexity is included in Section V-B. Similar to the observation in the regime $\lambda = O(1/\sqrt{k})$, we note that in the latter regimes, the sample complexity is dominated by the characteristics of the shared cluster (in this case, γk) only for extensively similar graphs such that $\gamma > 0.5$. However, for sufficiently large k such that $k > \frac{4\gamma}{\bar{\gamma}^2}$, the sample complexity in (17) includes logarithmic dependence on $\bar{\gamma}k$, although the exponential factor still dominates the sample complexity in γk. We note that except for the unlikely cases of extensively identical graphs with most edges in the shared cluster, the necessary condition on the sample complexity is captured by (18), which is of interest to understand the effect of structural similarity on sample complexity. Therefore, we focus our subsequent discussions on (18). Next, we further analyze (14) and (18), and start by evaluating the effect of similarity levels q and γ on the sample complexity.

Corollary 3 (Sample complexity versus q and γ.)**.** *Consider a pair of* q*-similar graphs* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *in class* $\mathcal{I}^k_{q,\gamma}(\lambda, \vartheta)$ *with increasing number of edges in the shared cluster* γk *and similarity level* q*. Any graph decoder* ψ *that achieves* $\mathsf{P}(\mathcal{I}^k_{q,\gamma}(\lambda, \vartheta)) \leq \varepsilon$ *has the following sample size requirements:*

1) *In the regime* $k = o(p)$*, the sample complexity scales with respect to* q *as* $\Omega(\log(p-q))$.
2) *In the regime* $k = \Omega(p)$*, the sample complexity scales with respect to* γ *as* $\Omega\left(\frac{\exp(\sqrt{k\bar{\gamma}})}{\sqrt{k\bar{\gamma}}} \log k\bar{\gamma})\right)$.

This indicates that for sparse graphs, the sample complexity is primarily characterized by q, and as the similarity level increases, the sample complexity decreases. For $k = \Omega(p)$, the sample complexity is characterized by $\bar{\gamma}k$, which is the maximum number of edges in the non-shared cluster in each graph, and as γ increases, the sample complexity decreases. These observations signify the gain of jointly recovering both structures instead of treating them in isolation, that is for recovering the two structures separately the sample complexity scales with $\Omega(\log p)$ and $\Omega(e^{\sqrt{k}}/\sqrt{k}\log k)$ in sparse and non sparse regimes, respectively, while when recovering them jointly, it reduces to $\Omega(\log(p-q))$ and $\Omega\left(\frac{e^{\sqrt{k\bar{\gamma}}}}{\sqrt{k\bar{\gamma}}} \log k\bar{\gamma})\right)$, respectively. Next, we evaluate the effect of an increasing number of edges k on the sample complexity.

Corollary 4 (Sample complexity versus k.)**.** *Consider a pair of* q*-similar graphs* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *in class* $\mathcal{I}^k_{q,\gamma}(\lambda, \vartheta)$ *with increasing number of edges* k*. Any graph decoder* ψ *that achieves* $\mathsf{P}(\mathcal{I}^k_{q,\gamma}(\lambda, \vartheta)) \leq \varepsilon$ *has the following sample size requirements:*

1) *In the regime* $\lambda = O(1/\sqrt{k})$*, the sample complexity scales with respect to* k *as* $\Theta(k)$.
2) *In the regimes* $\lambda = \Theta(1)$ *and* $\lambda = \Omega(\sqrt{k})$*, the sample complexity scales with respect to* k *as* $\Omega(\mathrm{e}^{\sqrt{k}}/\sqrt{k})$.

From Corollary 4, we note that the necessary condition on the sample complexity for jointly recovering the structures of the two graphs has a non-exponential behavior in k. This observation is consistent with that for recovering the structure of a single graph in this regime [18].

Remark 1. *We remark that the necessary conditions on the sample complexity in different regimes in Theorem 2 are independent of the maximum weight* ϑ*. However, this is an artifact of the fully-connected ensembles with* k *edges used for recovering the necessary conditions in the regimes* $\lambda = \Theta(1)$ *and* $\lambda = \Theta(\sqrt{k})$*, for which we have* $\vartheta = \lambda\sqrt{k}$ *in the ensemble constructions for the edge-bounded subclass.*

B. Sufficient Conditions for Ising Models

In this section, we provide sufficient conditions on the number of samples for model selection of different classes of Ising models. The sufficient conditions for exact recovery of graph models in or a generic class of q-similar graph pairs $\mathcal{S}_q$ are derived based on the large deviations analysis of an ML-based graph decoder given by

$$\Psi(\mathbf{x}_1^n, \mathbf{x}_2^n) = \arg \max_{(\mathcal{G}_1, \mathcal{G}_2) \in \mathcal{S}_q} f_{\mathcal{G}_1, \mathcal{G}_2}(\mathbf{x}_1^n, \mathbf{x}_2^n) \ , \tag{19}$$

where $f_{\mathcal{G}_1, \mathcal{G}_2}(\mathbf{x}_1^n, \mathbf{x}_2^n)$ is the joint pmf of the data samples $\mathbf{x}_1^n$ and $\mathbf{x}_2^n$. Note that since all q-similar graph pairs in a generic class $\mathcal{S}_q$ are equally likely to exist in nature, the ML decoder in (19) is equivalent to the maximum-a-posteriori rule, which is optimal for minimizing the probability of error in (8).

Theorem 3 (Class $\mathcal{I}_q^d(\lambda, \vartheta)$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_q^d(\lambda, \vartheta)$. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda, \vartheta)) \leq \varepsilon$, if the sample size n satisfies:*

1) *if* $(\upsilon - 1)^4(\upsilon + 1)^2 \geq \frac{32}{d}$,

$$n \geq 2c_1 d \ \log \frac{8d(p-q)^2(p+q)}{\varepsilon} \ , \tag{20}$$

2) *if* $(\upsilon - 1)^4(\upsilon + 1)^2 < \frac{32}{d}$,

$$n \geq c_1 d \ \log \frac{q^3 d}{2\varepsilon} \ , \tag{21}$$

where we have defined

$$c_1 \triangleq \frac{2(3\exp(2\vartheta) + 1)}{\sinh^2(\lambda/4)} \ , \qquad \text{and} \qquad \upsilon \triangleq \frac{p}{\sqrt{q}} \ . \tag{22}$$

Proof. See Section VI-A1. □

Examining the conditions in (20) and (21) implies that the condition of (21) will be satisfied when the two graphs are nearly identical for a wide range of feasible values of d. For instance, in this regime, when we have $d = 1$ and $p = 10$, we must have $q \geq 0.75p$ and for $d = 5$ and $p = 10$, we must have $q \geq 0.92p$. Therefore, except for the extreme cases, (20) provides the sample complexity for recovering two identical graphs jointly. We focus the rest of our analysis and discussions on the regime specified by (20). We start by evaluating the impact of structural similarity q on the sample complexity.

Corollary 5 (Sample complexity versus q.)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_q^d(\lambda, \vartheta)$ with increasing similarity level qin the regime characterized by* (20)*. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda, \vartheta)) \leq \varepsilon$, if the sample size n scales with respect to q and p as*

$$\Omega\left(\log d(p-q)^2(p+q)\right) \ . \tag{23}$$

We note that the sample complexity monotonically decreases in the similarity level q and its dependency on the edge parameters λ and ϑ is the same as that for joint recovery in [18]. This observation implies the gain in the sample complexity of jointly recovering the two structures jointly in contrast to recovering them independently, which has a sample complexity of $\Theta(\log dp)$ [18]. Next, we evaluate the effect of the degree d on the sample complexity.

Corollary 6 (Sample complexity versus d.)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_q^d(\lambda, \vartheta)$ with increasing maximum degree d. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda, \vartheta)) \leq \varepsilon$, if the number of samples satisfies the following conditions:*

1) *In the regime $\lambda = O(1/d)$, the sample complexity scales with respect to d as $\Omega(d^3 \log p)$.*
2) *In the regimes $\lambda = \Theta(1)$ and $\Theta(d)$, the sample complexity scales with respect to d as $\Omega(d\mathrm{e}^d \log p)$.*

From Corollary 6, we note that the sufficient condition on the sample complexity for joint structure estimation has a non-exponential scaling behavior in the regime $\lambda = O(1/d)$ and is consistent with that for the sample complexity of recovering a graph structure independently in this regime [18]. This implies that tractable algorithms with polynomial complexity in d exist for recovering q-similar graphs.

Theorem 4 (Class $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_q^k$. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)) \leq \varepsilon$, if for sufficiently large p, the sample size n satisfies:*

1) *if* $\log q \geq \frac{2(\bar{\gamma}k+1)}{\gamma k+1} \log p$ *and* $\gamma \geq \frac{2k+1}{3k}$,

$$n \ \geq \ \frac{c_1}{2}\left[6(\gamma k + 1)\log q + \log\frac{1}{\varepsilon}\right] \ , \tag{24}$$

2) *otherwise,*

$$n \ \geq \ c_1\left[6(\bar{\gamma} k + 1)\log p + \log\frac{1}{\varepsilon}\right] \ . \tag{25}$$

Proof. See Section VI-A2. □

Theorem 4 specifies the sufficient conditions on the sample complexity under different regimes of q and γ. It can be readily verified that the conditions in (24) hold only when the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are near identical, and most edges lie in the shared cluster V_{s}. Hence, we focus our subsequent discussions on the analysis of (25) to evaluate the sample complexity for cases except for near identical graphs. We start by evaluating the impact of the similarity metrics q and γ on the sample complexity.

Corollary 7 (Sample complexity versus γ.)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$ with increasing number of edges in the shared cluster γk and similarity level q. In the regime characterized by the conditions in* (24)*, there exists a graph decoder that achieves $\mathsf{P}(\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)) \leq \varepsilon$, if the sample complexity scales with respect to γ as $\Omega\left(\bar{\gamma} k \log p\right)$.*

Corollary 7 indicates that as the number of edges in the shared cluster γk increases, i.e., the proportion of total edges increases in V_{s}, the sample complexity decreases linearly with γk (recall that $\bar{\gamma} = 1 - \gamma$). This observation signifies the gain of jointly recovering the two graphs over treating them in isolation.

Corollary 8 (Sample complexity versus k.)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)$ with increasing number of maximum edges k. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)) \leq \varepsilon$ if the sample complexity satisfies the following requirements:*

1) *In the regime $\lambda = O(1/\sqrt{k})$, the sample complexity scales with respect to k as $\Omega(k^2)$.*
2) *In the regimes $\lambda = \Theta(1)$ and $\lambda = \Theta(\sqrt{k})$, the sample complexity scales with respect to k as $\Omega(k\mathrm{e}^{\vartheta})$.*

According to the definition of ϑ in (4), variations in k inevitably induce variations in ϑ as well. In the regime $\lambda = O(1/\sqrt{k})$, the effect of ϑ can be controlled because as we observe in Remark 1, for the worst case graph models that lead to the factor $\exp(\vartheta)$ in the sample complexity, we have $\vartheta \geq \lambda\sqrt{k}$. Therefore, in this regime, ϑ can be set to an arbitrary constant that will never be exceeded by its minimum feasible value for any combination of λ and k. On the other hand, in the regimes $\lambda = \Theta(1)$ and $\lambda = \Theta(\sqrt{k})$, the lower bound on ϑ increases with an increase in k and, therefore, the sample complexity scales exponentially in ϑ. This scaling behavior is consistent with the scaling behavior of the sample complexity for an extreme case of recovering a single graph independently [18].

Finally, we compare the results of Theorems 1, 2, 3, and 4 jointly in order to isolate the regimes under which we have the same scaling behavior of the sample complexity, i.e., the necessary and sufficient conditions scale at the same rate for both subclasses $\mathcal{I}^d_q(\lambda,\vartheta)$ and $\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)$. Note that for the class $\mathcal{I}^d_q(\lambda,\vartheta)$ with d fixed, the sample complexity depends on the similarity level q as the factor $\log(p-q)$ in the necessary conditions (see Corollary 1) and $\log((p-q)^2(p+q))$ in the sufficient conditions (see Corollary 5). Therefore, if q grows linearly with p, i.e., $q = \Theta(p)$, the necessary and sufficient conditions on the sample complexity have a similar asymptotic scaling behavior of $\Theta(\log p)$. On evaluating the dependence of sample complexity on d in Corollary 2 and Corollary 5, we observe that there is a mismatch of a factor of d between the sufficient and the necessary conditions. Next, for the class $\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)$, we observe that when γk is fixed and q scales linearly with p, the necessary conditions and the sufficient conditions have a similar scaling behavior of $\Theta(p)$. On the other hand, from Corollary 4 and Corollary 8, we observe that there is a mismatch of a factor of k between the necessary and the sufficient conditions on the sample complexity. By combining these observations, we can specify a specific regime for which the necessary conditions and the sufficient conditions meet, and the corresponding sample complexity is optimal.

Theorem 5 (Optimal Sample Complexity)**.** *When the maximum degree d and the maximum number of edges k are fixed, and in the regimes that satisfy $\lambda = O(1/p)$ and $q = \Theta(p)$, i.e., q increases linearly with p, the sample complexity of recovering graph models in the classes $\mathcal{I}^d_q(\lambda,\vartheta)$ and $\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)$ scales as $\Theta(p^2\log p)$ with growing graph size, p.*

The results that specify the bounds in necessary and sufficient conditions that have non-exponential scaling behavior for the classes $\mathcal{I}^d_q(\lambda,\vartheta)$ and $\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)$ are summarized in Table 1. Furthermore, we note that the two extreme cases of $q = 0$ and $q = p$ correspond to recovering two independent graphs and two identical graphs, respectively. For both these scenarios, the problem analyzed in this paper simplifies to the problem of structure estimation (learning) of one graph under an exact recovery criterion studied in [18]. In general, however, when we depart from these special cases, the results provided in this paper are distinct from the results in the existing literature. This observation is formalized in the following corollary.

Corollary 9 (Special cases for exact recovery)**.** *The necessary and sufficient conditions for the exact recovery of partially similar graphs in the subclasses $\mathcal{I}^d_q(\lambda,\vartheta)$ and $\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)$ in the extreme cases of $q = 0$ and $q = p$ subsume the existing results for the exact recovery of single graphs.*

We note that our results encompassing the necessary and sufficient conditions on the sample complexity are provided in non-asymptomatic regimes and, therefore, characterize the interplay between different graph parameters and similarity level q. Specifically, from Table 1, we observe that the factor $\log(p-q)$ appears in both necessary and sufficient conditions for the bounded degree class $\mathcal{I}^d_q$.

To theoretically establish that joint learning of q-similar Ising models is easier than recovering them independently, we would need to have the necessary conditions on the sample complexity for recovering graphs independently to be strictly larger than that for recovering q-similar graphs jointly for different q. However, such an analysis for Ising models is prohibited by a large mismatch in the terms that depend on edge parameters λ and ϑ in the two sets of conditions. For instance, for $d = 1$, the mismatch between the factor $\frac{1}{4\lambda\tanh(\lambda)}$ in the necessary condition and c_1 (defined in (22) in the sufficient condition in terms of ratio is at least 512 for any $\lambda > 0$. Similarly, for $d > 1$, the mismatch between the factor $\frac{\exp(\vartheta-\lambda)}{8\vartheta}$ and c_1 in terms of ratio scales at least as $471d^2$ for any $\lambda > 0$. This mismatch is highlighted in Table 1, where we see the sufficient conditions to scale at a factor d larger for class $\mathcal{I}^d_q$ and a factor k larger for class $\mathcal{I}^k_{q,\gamma}$ for certain regimes. Since the relative gain in terms of the ratio of sample complexities of jointly recovering q-similar graph models over independent graph recovery is at most 2, the mismatch between the different aforementioned factors in the case of Ising models is too large to be overcome by the gains offered by similarity level q in the sufficient conditions on the sample complexity for joint recovery. We remark that this is a limitation of the analysis techniques and is also present in the results for single graph recovery in both degree and edge bounded subclasses of Ising models [18]. However, these limitations are not as prominent in the analysis of Gaussian models, and we will discuss these aspects in Section IV.

We also remark that our numerical experiments for Ising models in Section 7 illustrate that learning q-similar graphs jointly indeed requires a significantly smaller number of samples than recovering the graphs independently for various settings. In our experiments, we start by using an ML decoder to recover sparse graphs. However, an ML decoder may be computationally intractable to implement due to the large

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3207420

TABLE I
SUMMARY OF MAIN RESULTS (NON-EXPONENTIAL SCALING) FOR EXACT RECOVERY OF ISING MODELS.

Graph Class	Parameters	Sample Complexity	
		Necessary	Sufficient (ML)
Bounded degree $\mathcal{I}_q^d$ $(p-q)^2 \gg q$	$\lambda = O\left(\frac{1}{d}\right)$	$\Theta(d^2 \log(p-q))$	$\Theta(d^3 \log((p-q)^2(p+q))$
	$\lambda = O\left(\frac{1}{p}\right)$ d fixed $q = \Theta(p)$	$\Theta(p^2 \log p)$	$\Theta(p^2 \log p))$
Bounded edge $\mathcal{I}_q^k$ $(p-q)^2 \gg q$ $\bar{\gamma}^2 \gg \gamma$	$\lambda = O\left(\frac{1}{\sqrt{k}}\right)$	$\Theta(k \log(p-q))$	$\Theta(\bar{\gamma} k^2 \log p)$
	$\lambda = O\left(\frac{1}{p}\right)$ k fixed $q = \Theta(p)$	$\Theta(p^2 \log p)$	$\Theta(p^2 \log p)$

computational complexity involved in the recovery of large graphs. In the next section, we discuss a computationally feasible structure estimation algorithm for jointly recovering similar graphs for Ising models that are valid for both degree-bounded and edge-bounded subclasses. This algorithm is subsequently applied to recover Ising models in our experiments in Section VII.

C. A Prediction-guided Algorithm for Jointly Recovering Ising Models

In this section, we discuss a computationally efficient structure estimation algorithm described in Algorithm 1 for recovering graph pairs jointly. The structure of the algorithm is motivated by a recent approach to structure estimation of MRFs in an online manner based on Hedge algorithm [15]. The Hedge algorithm uses multiplicative weight update rules for online estimation with expert advice in the context of multi-armed bandits [42]. We specify the steps involved in this algorithm and establish the sample complexity for perfectly recovering the graphs. For convenience in analysis, corresponding to each random variable $X_i^u \in \{-1, 1\}$, we define the Bernoulli random variable $B_i^u \triangleq \frac{1}{2}(1 - X_i^u)$ and instead of analyzing X_i^u we equivalently analyze B_i^u. The random instance of B_i^u based on j-th graph sample of $\mathcal{G}_i$ is given by $b_i^u(j)$ and computed as

$$b_i^u(j) = \frac{1}{2}(1 - x_i^u(j)) \ , \tag{26}$$

where $x_i^u(j)$ is the j-th random sample collected at node u in $\mathcal{G}_i$. The prediction-guided algorithm consists of two steps. Step 1 collects $n_{\mathrm{T}} < n$ samples to form multiple prediction-guided estimates for E_1 and E_2. Once the predictions are formed, we use the remaining $n_{\mathrm{M}} \triangleq n - n_{\mathrm{T}}$ samples to assess the risks associated with these predictions and to use the risk metrics for making a final decision for the structure estimates.

Step 1: Forming predictions of E_1 and E_2. This algorithm runs sequentially and collects the n_{T} samples one at a time, which are used to update a sequence of prediction-related decisions. The algorithm starts by considering that any pair of nodes in V can be potential neighbors. Each node acts as an *expert* and predicts the value of its neighbors. In the j^{th} iteration, at node u in $\mathcal{G}_i$ we form a prediction for B_i^u by aggregating the data samples provided by other nodes followed by a non-linear transformation according to

$$\hat{b}_i^u(j) = \sigma\Big(\sum_{v \neq u} w_i^{uv}(j) x_i^v(j)\Big) \ , \qquad \text{for } \ j \in \{1, \ldots, n_{\mathrm{T}}\} \ , \tag{27}$$

where $\{w_i^{uv}(j)\}$ are the weights to be selected properly as described below and σ is the standard sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. The main elements of this step are summarized below.

1) *Loss function.* To quantify the quality of the predictions, for every pair $u, v \in V$ we evaluate the pairwise loss function

$$\ell_i^{uv}(j) \ \triangleq \ \frac{1}{2}\Big(1 + [\hat{b}_i^u(j) - b_i^u(j)] x_i^v(j)\Big) \ . \tag{28}$$

2) *Predictor Update.* If nodes u and v lie in the set V_{s}, i.e., have a similar pairwise relationship in both graphs, we allow the transfer of loss functions between the graphs for updating the multiplicative weights associated with the pairwise relationship between nodes u and v in graph $\mathcal{G}_i$:

$$\kappa_i^{uv}(j+1) = \kappa_i^{uv}(j) \cdot \exp\left(\frac{\beta}{2}[\ell_1^{uv}(j) + \ell_2^{uv}(j)]\right) \ , \tag{29}$$

where β is an appropriately set hyper-parameter and $\kappa_i^{uv}(0) = \frac{1}{p-1}$. Otherwise, the updates follow the rule

$$\kappa_i^{uv}(j+1) = \kappa_i^{uv}(j) \cdot \exp\left(\beta \ell_i^{uv}(j)\right) \ . \tag{30}$$

We note that without the updates in (29), our algorithm reduces to estimating the two subgraphs independently using the algorithm in [15].

3) *Pseudo weights*: We introduce pseudo weights $\tilde{\kappa}_i^{uv}$ to accommodate for the setting when the neighborhood weight of a node u in graph $\mathcal{G}_i$ is strictly less than ϑ.
4) *Normalization of weights*: Note that the weight updates in (29) and (30) do not guarantee that $\sum_{v \in V} \kappa_i^{uv} \leq \vartheta$ for

any $u \in V$ in $\mathcal{G}_i$. Therefore, we introduce normalized weights w_i^{uv} which are evaluated as

$$w_i^{uv}(j+1) = \frac{\vartheta \kappa_i^{uv}(j+1)}{\sum\limits_{x \neq u} \left(\kappa_i^{ux}(j+1) + \tilde{\kappa}_i^{ux}(j+1)\right)} \ . \quad (31)$$

The processes above continue recursively until all the n_{T} samples are exhausted.

Step 2: Estimating E_1 and E_2. Finally, we collect additional n_{M} samples for all the nodes in V in graph $\mathcal{G}_i$ and based on these, we assign a risk metric to each predicted set E_i^j, $\forall j \in \{1, \ldots, n_{\mathrm{T}}\}$ according to

$$\varepsilon_i^j = \frac{1}{n_{\mathrm{M}}} \sum_{y=n_{\mathrm{T}}+1}^{n} \sum_{u \in V} \left[\hat{b}_i^u(j) - b_i^u(j)\right]^2 \ . \quad (32)$$

We select the predictions with the lowest empirical risks which is given by

$$\mathrm{s}_i = \underset{j \in \{1,\ldots,n_{\mathrm{T}}\}}{\operatorname{argmin}} \ \varepsilon_i^j \ , \quad (33)$$

for graph $\mathcal{G}_i$ and leverage the set of coefficients $\{w_i^{uv}(\mathrm{s}_i)\}$ to form the final estimates for E_1 and E_2 using a thresholding operation. Specifically, for graph $\mathcal{G}_i$, we form the prediction:

$$E_i^{\mathrm{s}_i} \ \triangleq \ \left\{(u,v) \ : \ w_i^{uv}(\mathrm{s}_i) \geq \frac{\lambda}{2}\right\} \ , \qquad \forall i \in \{1,2\}. \quad (34)$$

The steps of the algorithm are specified in Algorithm 1.

Algorithm 1 Estimating E_1 and E_2

Input $n = n_{\mathrm{T}} + n_{\mathrm{M}}$ pairs of data samples, $\beta = 1 - \sqrt{\log p / n_{\mathrm{T}}}$
Initialize $\kappa_i^{uv}(1) = 1/(p-1)$, $\tilde{\kappa}_i^{uv}(1) = 1/(p-1)$ and $w_i^{uv}(1) = 0$ for all $u \neq v$ and $i \in \{1,2\}$
for a new pair of data sample $j \in \{1, \ldots, n_{\mathrm{T}}\}$ **do**
  For each $u \in V$, compute $b_i^u(j)$ according to (26) for $i \in \{1,2\}$
  for each pair $u, v \in V$, $u \neq v$ **do**
    Compute losses $\ell_i^{uv}(j)$ according to (28)
    Update the weights $\tilde{\kappa}_i^{uv}(j+1) = \tilde{\kappa}_i^{uv}(j) \exp(\beta/2)$
    if $u, v \in V_{\mathrm{s}}$ **then**
      Update the weights $\kappa_i^{uv}(j+1)$ according to (29)
    else
      Update the weights $\kappa_i^{uv}(j+1)$ according to (30)
    end if
  end for
  for each pair $u \neq v$ **do**
    Compute $w_i^{uv}(j+1)$ according to (31)
  end for
  Compute empirical risks ε_i^k using n_{M} samples according to (32)
end for
Compute s_1 and s_2 according to (33)
Form estimates $E_1^{\mathrm{s}_1}$ and $E_2^{\mathrm{s}_2}$ according to (34)
return $E_1^{\mathrm{s}_1}$ and $E_1^{\mathrm{s}_2}$

The following theorem captures the sample complexity and the computational complexity of Algorithm 1.

Theorem 6 (Algorithm 1)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$. If the sample size $n = n_{\mathrm{T}} + n_{\mathrm{M}}$ satisfies $n_{\mathrm{T}} = O\left(\frac{\vartheta^2 \exp(\vartheta)}{\lambda^4} \log \frac{p}{\lambda\varepsilon}\right)$ and $n_{\mathrm{M}} = \Theta\left(\frac{\log(n_{\mathrm{T}}/\varepsilon)}{\rho\lambda^2}\right)$, Algorithm 1 achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda,\vartheta)) \leq 2\varepsilon$ for $\mathcal{I}_q^d(\lambda,\vartheta)$ and $\mathsf{P}(\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)) \leq 2\varepsilon$ for $\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)$. The run time of Algorithm 1 is $O(p^2 n)$.*

Proof. See Appendix A. □

Note that the mean loss function $(\ell_1^{uv} + \ell_2^{uv})/2$ is bounded in the range $[0,1]$. Therefore, the rules for updating weights in our algorithm satisfy the conditions for [15, Theorem 5] to hold, which allows us to leverage the regret bound on the Hedge algorithm in [42]. We remark that the results in Theorem 6 are in the asymptotic regime and do not capture the effect of structural similarity on the sample complexity. However, we note that Algorithm 1 achieves optimal asymptotic sample complexity for both degree-bounded and edge-bounded classes in the regime specified in Corollary 5. Specifically, in the regime, $\lambda = O(1/p)$ with the degree d fixed, Algorithm 1 achieves the asymptotic sample complexity of $O(p^2 \log p)$ which is the same as the optimal scaling behavior established for the regime in Corollary 5 when $q = \Theta(p)$. Similarly, Algorithm 1 achieves optimal sample complexity for $\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)$ in the regime $\lambda = O(1/p)$ with k fixed.

Our numerical evaluations in Section VII indicate a significant gain in performance when $q-$ similar graphs are learned jointly using Algorithm 1 in comparison to when they are learned independently using the algorithm in [15].

IV. Main Results: Gaussian Models

In this section, we provide the bounds on the sample complexity for the degree-bounded and edge-bounded subclasses of Gaussian models.

A. Necessary Conditions for Gaussian Models

We first provide necessary conditions on the sample complexity $n(q,\varepsilon)$ in order to ensure that $\mathsf{P}(\mathcal{S}_q) \leq \varepsilon$. We use the shorthand n for $n(q,\varepsilon)$ and provide the scaling behavior of the sample complexity in different regimes, illustrating the dependence of sample complexity on different graph parameters.

Theorem 7 (Class $\mathcal{G}_q^d(\rho)$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_q^d(\rho)$, for which $\rho \in [0, \frac{1}{2}]$. Any graph decoder ψ that achieves $\mathsf{P}(\mathcal{G}_q^d(\rho)) \leq \varepsilon$, has the following sample size requirements:*

1) *In the regime $\rho = \Theta\left(\frac{1}{d}\right)$, the sample size satisfies:*
   a) *if $\binom{q}{2} \leq \binom{p-q}{2}^2$,*
   $$n \geq \frac{(1-\varepsilon)}{4\rho^2}\left(4\log\frac{p-q-1}{\sqrt{2}} - 1\right) \ , \quad (35)$$
   b) *if $\binom{q}{2} > \binom{p-q}{2}^2$,*
   $$n \geq \frac{(1-\varepsilon)}{4\rho^2}\left(2\log\frac{q-1}{\sqrt{2}} - 1\right) \ . \quad (36)$$
2) *In the regime $\rho = \Theta(1)$, the sample size satisfies:*

a) *if* $q+\sqrt{qd}\leq p$,

$$n\geq\frac{1-\varepsilon}{\log\left(1+\frac{d\rho}{1-\rho}\right)}\left[2d\log\frac{p-q}{d}-1\right]\ , \quad (37)$$

b) *if* $q+\sqrt{qd}>p$,

$$n\geq\frac{1-\varepsilon}{\log\left(1+\frac{d\rho}{1-\rho}\right)}\left[d\log\frac{q}{d}-1\right]\ . \quad (38)$$

Proof. See Section V-C. □

Theorem 7 provides the necessary conditions on the sample complexity for different regimes of ρ. We note that the conditions on the similarity level q in (36) and (38) are satisfied only in graph pairs with extensive similarity, i.e., $q\approx p$, for which the problem of structure estimation approaches the extreme case of estimating two identical graphs. Therefore, we focus our discussions on the analysis of sample complexity based on (35) and (37). We start by evaluating the dependence of sample complexity on the similarity level q.

Corollary 10 (Sample complexity versus q)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_q^d(\rho)$ with the similarity level q. Any graph decoder that achieves $\mathsf{P}(\mathcal{G}_q^d(\rho))\leq\varepsilon$, the sample complexity scales with respect to q and p as $\Omega(\log(p-q))$.*

This corollary indicates that the sample complexity requirement decreases with an increase in the similarity level of the two graphs. Specifically, jointly recovering the two structures requires a sample complexity that scales with $\Omega(\log(p-q))$ in contrast to treating the two graphs independently for which the sample complexity scales as $\Omega(\log p)$ [23]. Next, we evaluate the effect of increasing degree d on the sample complexity.

Corollary 11 (Sample complexity versus d)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_q^d(\rho)$. For any graph decoder that achieves $\mathsf{P}(\mathcal{G}_q^d(\rho))\leq\varepsilon$, the sample complexity satisfies the following conditions:*

1) *In the regime $\rho=O\left(\frac{1}{d}\right)$, the sample complexity scales with respect to d as $\Omega(d^2)$.*
2) *In the regime $\rho=\Theta(1)$, the sample complexity scales with respect to d*

$$\Omega\left(\frac{d}{\log(1+d\rho)}\log\frac{p-q}{d}\right) \quad (39)$$

The existing results in [43] show that the sufficient condition on the sample complexity scales as $\Omega((d^2+\rho^{-2})\log p)$. This sample complexity is achievable via ℓ_1-regularized ML-based graphical model selection for single graphs. In the regime $\rho=\Theta(1/d)$, the sufficient condition matches with the scaling behavior of the necessary condition for single graphs up to constant factors, indicating the scaling behavior of necessary conditions is optimal in this regime. We conjecture that this observation extends to the results for joint model selection, i.e., the scaling behavior of the necessary conditions from Theorem 7 is optimal in the regime $\rho=\Theta(1/d)$. In the regime $\rho=\Theta(1)$, it is interesting to note that the lower bound on the sample complexity scales approximately as $\Theta\left(d^{\varphi}\log\frac{p-q}{d}\right)$ for some $\varphi<1$, which has a non-monotonic scaling behavior in d. We remark that the sample complexity in this regime is dominated by the graphs with densely-connected subgraphs of d nodes. This characteristic of the sample complexity implies that the sample complexity of estimating graph models with densely-connected subgraphs has a higher sample complexity than that for nearly fully-connected graphs. We also note that this observation is consistent with that of the sample complexity of recovering a single graph [23].

Theorem 8 (Class $\mathcal{G}_q^k(\rho)$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_q^k(\rho)$, for which $\rho\in[0,\frac{1}{2}]$. For any graph decoder ψ that achieves $\mathsf{P}(\mathcal{G}_q^k(\rho))\leq\varepsilon$, the sample size must satisfy the following requirements:*

1) *In the regime $\rho=\Theta(1/\sqrt{k})$, the sample size n must satisfy:*

a) *if* $\binom{q}{2}\leq\binom{p-q}{2}^2$,

$$n\geq\frac{(1-\varepsilon)}{4\rho^2}\left(4\log\frac{p-q-1}{\sqrt{2}}-1\right)\ , \quad (40)$$

b) *if* $\binom{q}{2}>\binom{p-q}{2}^2$,

$$n\geq\frac{(1-\varepsilon)}{4\rho^2}\left(2\log\frac{q-1}{\sqrt{2}}-1\right)\ . \quad (41)$$

2) *In the regime $\rho=\Theta(1)$, the sample size n must satisfy:*

a) *if* $q+\sqrt{q\tilde{k}}\leq p$,

$$n>\frac{(1-\varepsilon)\tilde{k}}{\log\left(1+\frac{\tilde{k}\rho}{1-\rho}\right)}\left(2\log\frac{p-q}{\tilde{k}}-1\right)\ , \quad (42)$$

b) *if* $q+\sqrt{q\tilde{k}}>p$,

$$n>\frac{(1-\varepsilon)\tilde{k}}{\log\left(1+\frac{\tilde{k}\rho}{1-\rho}\right)}\left(\log\frac{q}{\tilde{k}}-1\right)\ , \quad (43)$$

where

$$\tilde{k}\triangleq\lfloor\sqrt{k}\rfloor. \quad (44)$$

Proof. See Section V-C. □

Theorem 8 characterizes the necessary conditions on the sample complexity under two regimes of ρ. Note that the conditions on the similarity level q in (41) and (43) are satisfied only by graphs with extensive similarity, i.e., $q\approx p$. Therefore, we subsequently discuss the sample complexity based on the analysis of (40) and (42). It is readily observed that the sample complexity has a dependence on the similarity level q that is similar to that for the degree-bounded class, i.e., in terms of the factor $\log(p-q)$. We formalize this observation in the following Corollary.

Corollary 12 (Sample complexity versus q)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_q^k(\rho)$ with increasing similarity level q. For any graph decoder that achieves $\mathsf{P}(\mathcal{G}_q^k(\rho))\leq\varepsilon$, the sample complexity scales with respect to q and p as $\Omega(\log(p-q))$.*

Next, we evaluate the effect of increasing number of edges k on the sample complexity.

Corollary 13 (Sample complexity versus k)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in class $\mathcal{G}_q^k(\rho)$ with increasing similarity level q. For any graph decoder that achieves $\mathsf{P}(\mathcal{G}_q^k(\rho)) \leq \varepsilon$, as k increases, the sample complexity n must satisfy the following conditions:*

1) *In the regime $\rho = O(1/\tilde{k})$, the sample complexity scales with respect to k as $\Theta(k)$.*
2) *In the regime $\rho = \Theta(1)$, the sample complexity scales with respect to k as*

$$\Theta\left(\frac{\sqrt{k}}{\log(1+\rho\sqrt{k})}\log\frac{p-q}{\sqrt{k}}\right) . \tag{45}$$

Corollary 13 implies that the sample complexity has a linear scaling behavior in k in the regime $\rho = O(1/\tilde{k})$. In the regime, $\rho = \Theta(1)$, we observe that the sample complexity has a non-monotonic scaling behavior in k. We remark that the sample complexity in this context is recovered by the analysis of densely-connected subgraphs of $\tilde{k}$ nodes. Therefore, the necessary condition on the sample complexity implies that recovering graphs with a densely-connected subgraph up to a certain size dominates the sample complexity of recovering graphs that are nearly fully connected. This observation is consistent with our discussion in the context of degree-bounded Gaussian models.

B. Sufficient Conditions for Gaussian Models

In this section, we provide sufficient conditions on the sample complexity for a specifically constructed set of ensembles of graphs in classes $\mathcal{G}_q^d(\rho)$ and $\mathcal{G}_q^k(\rho)$. The sufficient conditions are established based on the large deviation analysis of the ML decoder similar to that in (19) for Gaussian models. We provide Lemma 1, which is instrumental to establishing sufficient conditions on the sample complexity. For this purpose, consider an arbitrary class of distinct Gaussian models indexed by $\mathcal{S} \triangleq \{1,\ldots,m\}$. For model $u \in \mathcal{S}$, denote the precision matrix by $\mathbf{P}[u]$, and define Λ_u as its log-likelihood function, i.e.,

$$\Lambda_u(\mathbf{x}) \triangleq \log\left(\frac{\sqrt{\det[\mathbf{P}[u]]}}{(2\pi)^{\frac{p}{2}}}\exp\left(\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{P}[u]\mathbf{x}\right)\right) . \tag{46}$$

Lemma 1. *Consider a Gaussian graphical model $\mathcal{G}$ in the set $\mathcal{S}$. For any $u, v \in \mathcal{S}$ we have*

$$\mathbb{P}[\Lambda_u(\mathbf{x}) \geq \Lambda_v(\mathbf{x})] \leq \left(\frac{\det[\mathbf{P}[u]]\cdot\det[\mathbf{P}[v]]}{\det^2\left[\frac{\mathbf{P}[u]+\mathbf{P}[v]}{2}\right]}\right)^{\frac{1}{4}} . \tag{47}$$

Proof. See Section VI-B. □

This lemma is pivotal for comparing the likelihoods of different models. For an unconstrained precision matrix $\mathbf{P}[u], \forall u \in \mathcal{S}$, it is infeasible to evaluate

$$\det\left[\frac{\mathbf{P}[u]+\mathbf{P}[v]}{2}\right] \tag{48}$$

for all possible graphs. Therefore, for our analysis, we focus on three specifically constructed ensembles of graphs in the degree and edge-bounded subclasses of Gaussian models, denoted by $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$, which enable us to gain intuition into the performance of ML decoder in graphs with different characteristics. Next, we provide the construction of subclasses that consist of cliques of different sizes. For this purpose, we define U_m as the subset of nodes in a clique of size $m \leq p$, where $U_m \subseteq V$. We also define the $p \times 1$ subclass $\mathbb{1}_{U_m}$ of dimension $p \times 1$, where its entries are given by

$$[\mathbb{1}_{U_m}]_i = \begin{cases} 1\,, & \text{if } i \in U_m \\ 0\,, & \text{if } i \in V\backslash U_m \end{cases} . \tag{49}$$

.

Restricted Subclass $\mathcal{A}_q$: The graphs in this subclass consist of an isolated clique U_2 consisting of 2 nodes that lie completely either in the shared part or in the non-shared part of the graphs. For a given parameter $a \geq 0$, the associated precision matrix is given by $\mathbf{P}_i = \mathbf{I} + a\mathbb{1}_{U_2}\mathbb{1}_{U_2}^\top$. There are $\binom{q}{2}$ graph pairs with the clique formed by U_2 in the shared part, and $\binom{p-q}{2}^2$ graph pairs with U_2 in the non-shared part. Furthermore, we have $\mathcal{A}_q \subseteq \mathcal{G}_q^d(\rho)$ and $\mathcal{A}_q \subseteq \mathcal{G}_q^k(\rho)$.

Restricted Subclass $\mathcal{B}_q$: The graphs in this subclass consist of a clique U_d formed by a set of d nodes and the associated precision matrix is given by $\mathbf{P}_i = \mathbf{I} + a\mathbb{1}_{U_d}\mathbb{1}_{U_d}^\top$. We assume that for each graph, the set U_d lies completely in either the shared part or the non-shared part of the graph. If U_d lies in the non-shared part, we have $\binom{q}{d}$ number of possible graph pairs. If U_d lies in the non-shared part, we have $\binom{p-q}{d}^2$ number of possible graph pairs. Furthermore, in this class, we must have $q = \Omega(d)$. Therefore, $\mathcal{B}_q \subseteq \mathcal{G}_q^d(\rho)$.

Restricted Subclass $\mathcal{C}_q$: The graphs in this subclass consist of a clique $U_{\tilde{k}}$ formed by a set of $\tilde{k}$ nodes and the associated precision matrix is given by $\mathbf{P}_i = \mathbf{I} + a\mathbb{1}_{U_{\tilde{k}}}\mathbb{1}_{U_{\tilde{k}}}^\top$. We assume that for each graph, the nodes spanning $U_{\tilde{k}}$ lie completely in either the shared part or the non-shared part of the graph. If $U_{\tilde{k}}$ lies in the non-shared part, we have $\binom{q}{\tilde{k}}$ number of possible graph pairs. If $U_{\tilde{k}}$ lies in the non-shared part, we have $\binom{p-q}{\tilde{k}}^2$ number of possible graph pairs. The properties of this restricted subclass of graphs dictate that $\mathcal{C}_q \subseteq \mathcal{G}_q^k(\rho)$.

Clearly, $\mathcal{A}_q$ represents the subclass of graph pairs with unknown isolated edges, and $\mathcal{B}_q$ and $\mathcal{C}_q$ represent the subclass of graph pairs with high connectivity in degree-bounded and edge-bounded subclasses. Thus, analyzing the subclasses $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$ provides the bounds on sample complexity for the subclasses that lie at the two extremes in terms of edge connectivity in the classes $\mathcal{G}_q^d(\rho)$ and $\mathcal{G}_q^k(\rho)$. Next, we provide sufficient conditions for the joint selection of graphs in the subclasses $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$.

Theorem 9 (Subclass $\mathcal{A}_q$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{A}_q$. There exists a graph decoder ψ that achieves $\mathsf{P}(\mathcal{A}_q) \leq \varepsilon$, if the sample size n satisfies:*

1) *if $q + \sqrt{q\sqrt{2}} \leq p$,*

$$n \geq \frac{1}{\log\frac{1}{(1-\rho^2)}}\left(4\log(p-q) + \log\frac{2}{\varepsilon}\right) , \tag{50}$$

2) *if* $q+\sqrt{q\sqrt{2}}>p$,

$$n \geq \frac{1}{\log \frac{1}{(1-\rho^2)}} \left(2\log q + \log \frac{2}{\varepsilon}\right) . \tag{51}$$

Proof. See Section VI-B1. □

Next, we analyze the variation in the bounds on sample complexity with respect to the structural similarity q. As observed previously, the regime for the sufficient condition in (51) is valid for almost identical graphs. Therefore, we focus our discussion on the analysis of the sufficient condition in (50).

Corollary 14 (Sufficient condition and q)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{A}_q$ with increasing similarity level q and graph size p. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{A}_q) \leq \varepsilon$, if the following conditions are satisfied:*

1) *In the regime $\rho = \Theta(1)$, the sample complexity scales with respect to q and p as $\Omega(\log(p-q))$.*
2) *In the regime $\rho = \Theta(1/p)$, the sample complexity scales with respect to q and p as $\Omega(p^2\log(p-q))$.*

From Corollary 14, we note that in both regimes, the sample complexity depends on q through $\log(p-q)$, which has a decreasing behavior with increasing similarity level q. This observation captures the gain in the sample complexity of jointly recovering the two graphs. We also remark that since we have $\mathcal{A}_q \subseteq \mathcal{G}_q^d(\rho)$, in the regime $\rho = \Theta(1/d)$, the sufficient condition on the sample complexity in Theorem 9 scales as $\Theta(d^2\log(p-q))$, which matches the scaling behavior of the necessary condition on the sample complexity in Theorem 7. This observation indicates that an ML decoder achieves the optimal sample complexity for recovering the q-similar graphs in the subclass $\mathcal{A}_q$.

Theorem 10 (Subclass $\mathcal{B}_q$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{B}_q$. There exists a graph decoder ψ that achieves $\mathsf{P}(\mathcal{B}_q) \leq \varepsilon$, if the sample size satisfies:*

$$n \geq c_2 \left(2d\log\frac{(p-q)\mathrm{e}}{d} + \log\frac{2}{\varepsilon}\right) , \quad \text{if } q+\sqrt{\frac{qd}{2\mathrm{e}}} \leq p , \tag{52}$$

$$n \geq c_2 \left(d\log\frac{q\mathrm{e}}{d} + \log\frac{2}{\varepsilon}\right) , \quad \text{if } q+\sqrt{\frac{qd}{2\mathrm{e}}} > p , \tag{53}$$

where we have defined

$$c_2 \triangleq \left[\log\left(1+\frac{d^2}{(1/\rho-1)(1/\rho-1+d)}\right]\right)^{-1} . \tag{54}$$

Proof. See Section VI-B2. □

Theorem 10 provides sufficient conditions on the sample complexity for different regimes of q. From a similar discussion as in the prior cases, we conclude that the regime in (53) is applicable to scenarios with an extensive similarity between the two graphs. Therefore, we focus our subsequent discussion on the analysis of sample complexity in the regime in (52).

Corollary 15 (Sufficient condition and q)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{B}_q$ with increasing similarity level q. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{B}_q) \leq \varepsilon$, if the sample complexity scales with respect to q as $\Omega(\log(p-q))$.*

From Corollary 15, we note that the sample complexity is characterized by $\log(p-q)$, which indicates the savings in sample complexity as the similarity level q increases. Next, we discuss the sample complexity with respect to the maximum degree d.

Corollary 16 (Sufficient condition and d)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{B}_q$ with increasing maximum degree d. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{B}_q) \leq \varepsilon$, if the following conditions are satisfied:*

1) *In the regime $\rho = (1/d)$, the sample complexity scales with respect to d as $\Omega(d\log(\frac{p-q}{d}))$.*
2) *In the regime $\rho = \Theta(1)$, the sample complexity scales with respect to d as $\Omega\Big(\frac{d}{\log(1+d\rho^2)}\log\frac{p-q}{d}\Big)$.*

We observe that in both regimes, the sample complexity depends on q through $\log(p-q)$, which captures the savings in the sample complexity as the similarity level q increases. Furthermore, we also note that the sufficient condition on the sample complexity for recovering graphs in $\mathcal{B}_q$ in the regime $\rho = \Theta(1/d)$ is dominated by the corresponding condition for the subclass $\mathcal{A}_q$ and the necessary condition (given by $\Omega(d^2\log(p-q))$) for the class $\mathcal{G}_q^d(\rho)$ in Theorem 7. This observation indicates that recovering the graphs in subclass $\mathcal{B}_q$ is easier than the graphs in the worst-case scenario for $\mathcal{G}_q^d(\rho)$ in this regime. On the other hand, in the regime $\rho = \Theta(1)$, the sufficient condition on the sample complexity matches the necessary condition on the sample complexity for graphs in $\mathcal{G}_q^d(\rho)$ up to constant factors (refer to Corollary 11). Moreover, we observe the non-monotonic scaling behavior of the sample complexity of an ML decoder with respect to d, which is consistent with the observations from the necessary conditions in Corollary 11.

Theorem 11 (Subclass $\mathcal{C}_q$)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{C}_q$. There exists a graph decoder ψ that achieves $\mathsf{P}(\mathcal{C}_q) \leq \varepsilon$, if the sample size satisfies:*

$$n \geq 2c_3\tilde{k}\log\frac{(p-q)\mathrm{e}}{\tilde{k}} + \log\frac{2}{\varepsilon} , \quad \text{if } q+\sqrt{\frac{q\tilde{k}}{2\mathrm{e}}} \leq p , \tag{55}$$

$$n \geq c_3\tilde{k}\log\frac{q\mathrm{e}}{\tilde{k}} + \log\frac{2}{\varepsilon} , \quad \text{if } q+\sqrt{\frac{q\tilde{k}}{2\mathrm{e}}} > p , \tag{56}$$

where we have defined

$$c_3 \triangleq \log^{-1}\left(1+\frac{\tilde{k}^2}{(1/\rho-1)(1/\rho-1+\tilde{k})}\right) , \tag{57}$$

and $\tilde{k} \triangleq \lfloor\sqrt{k}\rfloor$.

Proof. See Section VI-B3. □

Theorem 11 provides sufficient conditions on the sample complexity in different regimes of q. We note that the regime

in (56) implies extensive similarity between the two graphs. For instance, when we have $p = 150$ and $k = 100$, the regime in (56) is satisfied only for $q \geq 0.9p$. Therefore, we focus our discussion on the analysis of sample complexity in the regime in (55), which covers cases other than the cases with extensive structural similarity.

Corollary 17 (Sufficient condition and q)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{C}_q$ with increasing similarity level q. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{C}_q) \leq \varepsilon$, if the sample complexity scales with respect to q and p as $\Omega(\log(p-q))$.*

We observe in Corollary 17 that the sample complexity is characterized by $\log(p-q)$, which quantifies the gains in sample complexity as the similarity level q increases. Next, we discuss the sample complexity with respect to the maximum number of edges k.

Corollary 18 (Sufficient condition and k)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{C}_q$ with increasing maximum number of edges k. There exists a graph decoder that achieves $\mathsf{P}(\mathcal{C}_q) \leq \varepsilon$, if the sample complexity satisfies:*

1) *In the regime $\rho = \Theta(1/\tilde{k})$, the sample complexity scales with respect to k as $\Omega(\tilde{k}\log(\frac{p-q}{\tilde{k}}))$, where $\tilde{k} \triangleq \lfloor\sqrt{k}\rfloor$.*
2) *In the regime $\rho = \Theta(1)$, the sample complexity scales with respect to k as $\Omega\Big(\frac{\tilde{k}}{\log(1+\tilde{k}\rho^2)}\log\frac{p-q}{\tilde{k}}\Big)$.*

From Corollary 18, in both regimes, we observe that the sample complexity depends on q through the factor $\log(p-q)$, which quantifies the savings in the sample complexity as the similarity level q increases. Furthermore, by comparing the sufficient conditions in the regime $\rho = \Theta(1/\tilde{k})$ in Corollary 18 with the corresponding sufficient conditions for subclass $\mathcal{A}_q$ and the necessary condition for class $\mathcal{G}_q^k(\rho)$ in Theorem 8, we note that recovering the graphs in subclass $\mathcal{C}_q$ is easier than the graphs in the worst case scenario for $\mathcal{G}_q^k(\rho)$ in this regime. On the other hand, in the regime $\rho = \Theta(1)$, the sufficient condition on the sample complexity matches the necessary condition on the sample complexity for graphs in $\mathcal{G}_q^k(\rho)$ up to constant factors (see Corollary 13).

Comparing the necessary and sufficient conditions in subclasses $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$ provides the following insights into the behavior of sample complexity:

1) As $\rho \to 0$ in the regimes $\lambda = \Theta(1/d)$, $\lambda = \Theta(1/p)$, or $\lambda = \Theta(1/\tilde{k})$, the connectivity of the graphs with isolated edges becomes *harder* to learn as compared to the fully-connected graphs.
2) For an invariant ρ and a regime characterized by an increase in p, d, or k at any rate, the connectivity of the graphs with densely-connected subgraphs becomes *harder* to learn as compared to the graphs with isolated edges.

The main results that specify the bounds on the sample complexity for the exact recovery of Gaussian models in classes $\mathcal{G}_q^d(\rho)$ and $\mathcal{G}_q^k(\rho)$ are summarized in Table 2. The sufficient conditions in Table 2 correspond to the sample complexity of the dominant subclass of Gaussian models among the restricted subclasses $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$.

From Table 2, we note that the factor $\log(p-q)$ appears in both necessary and sufficient conditions on the sample complexity of recovering q-similar graphs jointly.

C. Strict Improvement Compared with Single-graph Recovery

To establish that the joint graph recovery of q-similar graphs is indeed easier than independently recovering them, we compare the lower bound on the sample complexity (necessary conditions) for the single graph class and the upper bound on the sample complexity (sufficient conditions) for the class of q-similar graphs. In our analysis, we exclude the setting in which the two graphs are either almost identical (i.e., $q \to p$) or almost distinct (i.e., $q \to 0$). To this end, we focus on the regime $\max\{q, p-q\} < p^{1-2\varepsilon}$. We note that this regime is not too stringent. For instance, when $\varepsilon = 0.1$, q that satisfies $\max\{q, p-q\} < p^{0.8}$ lies in $q \in [p^{0.8}, p-p^{0.8}]$. For $p = 10000$, this range is $[1585, 8415]$.

Theorem 12 (Degree-bounded Subclass)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{B}_q$. For $\max\{q, p-q\} < p^{1-2\varepsilon}$ and $\rho > \frac{1}{d+1}$, the necessary condition on the sample complexity for recovering $\mathcal{G}_1$ (or $\mathcal{G}_2$) independently is strictly larger than the sufficient condition on the sample complexity for recovering $\mathcal{G}_1$ and $\mathcal{G}_2$ jointly.*

Proof. See Appendix B. □

Theorem 12 establishes that the sample complexity of jointly recovering the graphs is strictly smaller than that of recovering them independently for the graphs in subclass $\mathcal{B}_q$, thus, providing conclusive evidence that the joint structure learning of q-similar pair of graphs is easier than learning them independently. By setting $d = 1$, class $\mathcal{B}_q$ reduces to Class $\mathcal{A}_q$. Due to the similarity in constructions and results, the same line of analysis for $\mathcal{B}_q$ (with proper adjustments) applies to $\mathcal{A}_q$.

For the edge bounded subclass, $\mathcal{C}_q$, we note that the constructions of classes $\mathcal{C}_q$ and $\mathcal{B}_q$ are derived from the ensemble construction in Section V-C2. Therefore, the necessary conditions for joint graph learning in $\mathcal{C}_q$ follow directly from (238) for $m = \tilde{k}$ and the sufficient conditions follow from (209). We next formalize the comparison between the sufficient condition on the sample complexity for joint graph recovery and the necessary condition for independent graph recovery.

Theorem 13 (Edge-bounded Subclass)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the subclass $\mathcal{C}_q$. For $\max\{q, p-q\} < p^{1-2\varepsilon}$ and $\rho > \frac{1}{\tilde{k}+1}$, the necessary condition on the sample complexity for recovering $\mathcal{G}_1$ (or $\mathcal{G}_2$) independently is strictly larger than the sufficient condition on the sample complexity for recovering $\mathcal{G}_1$ and $\mathcal{G}_2$ jointly.*

Proof. The proof of Theorem 13 follows directly from the proof for subclass $\mathcal{B}_q$ in Appendix B by replacing d with $\tilde{k}$. For the edge bounded subclass, $\mathcal{C}_q$, we note that the construction of classes $\mathcal{C}_q$ and $\mathcal{B}_q$ is similar and stems from the ensemble construction in Section V-C2. Therefore, the necessary conditions for joint graph learning in $\mathcal{C}_q$ follow directly from (117) and (238) and the sufficient conditions are established in Section VI-B2 and Section VI-B3. Hence, the proof in

TABLE II
SUMMARY OF THE MAIN RESULTS (NON-EXPONENTIAL SCALING) FOR THE EXACT RECOVERY OF GAUSSIAN MODELS. SUFFICIENT CONDITIONS CORRESPOND TO THE DOMINANT SUBCLASS AMONG $\mathcal{A}_q, \mathcal{B}_q$, AND $\mathcal{C}_q$ IN THE SPECIFIED REGIMES.

Graph Class	Parameters	Sample Complexity for $\mathcal{A}_q, \mathcal{B}_q$, and $\mathcal{C}_q$	
		Necessary	Sufficient*
Degree-bounded $\mathcal{G}_q^d$	$\rho = \Theta\left(\frac{1}{d}\right)$	$\Theta(d^2 \log(p-q))$	$\Theta(d^2 \log(p-q))$
$(p-q)^2 \gg q$	$\rho = \Theta(1)$	$\Theta\left(\frac{d \log((p-q)/d)}{\log(1+\rho d)}\right)$	$\Theta\left(\frac{d \log((p-q)/d)}{\log(1+\rho^2 d)}\right)$
Edge-bounded $\mathcal{G}_q^k$	$\rho = \Theta\left(\frac{1}{\sqrt{k}}\right)$	$\Theta(k \log(p-q))$	$\Theta(k \log(p-q))$
$(p-q)^2 \gg q$	$\rho = \Theta(1)$	$\Theta\left(\frac{\tilde{k} \log((p-q)/\tilde{k})}{\log(1+\rho \tilde{k})}\right)$	$\Theta\left(\frac{\tilde{k} \log((p-q)/\tilde{k})}{\log(1+\rho^2 \tilde{k})}\right)$

Appendix B that establishes that the joint graph learning is strictly easier than single graph recovery for $\mathcal{B}_q$ readily extends to the class $\mathcal{C}_q$ by replacing d with $\tilde{k} = \lfloor k \rfloor$. □

To establish our results in Theorem 12 and Theorem 13, we assume that the single graph decoder is agnostic to the similarity in two graphs in terms of q. However, we note that the subclasses $\mathcal{A}_q$, $\mathcal{B}_q$, and $\mathcal{C}_q$ are structured to have the clique completely in either the shared part or the non-shared part of the graph. Hence, the size of the number of possible graphs in the corresponding class for single graphs is smaller than $\binom{p}{2}$ for $\mathcal{A}_q$, $\binom{p}{d}$ for $\mathcal{B}_q$ and $\binom{p}{\tilde{k}}$ for $\mathcal{C}_q$. In the following remark, we clarify that the gains offered by joint graph recovery are retained even if the clique in the construction of classes $\mathcal{B}_q$ and $\mathcal{C}_q$ can span across shared and non-shared parts of the graph.

Remark 2. *For a pair of q-similar graphs in the class $\mathcal{B}_q$ such that the clique U_d can span both shared and non-shared parts, the necessary condition on the sample complexity for recovering a single graph is strictly larger than the sufficient condition on the sample complexity for recovering q-similar graphs for sufficiently small ε.*

Additional analysis details supporting Remark 2 are available in Appendix C. Similar observations can be made for class $\mathcal{C}_q$.

V. PROOFS OF NECESSARY CONDITIONS

In this section, we provide the proof of the information-theoretic necessary conditions on the sample complexity of recovering graph pairs in different classes of Ising and Gaussian models. In general, we leverage Fano's Lemma for characterizing the necessary conditions, which are also used for other structure estimation purposes in [18], [19]. To formalize this, consider two q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ that belong to a generic class $\mathcal{S}_q$ of q-similar graphs that contains a total of M pairs of q-similar graphs. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be selected from $\mathcal{S}_q$ uniformly at random. We define ζ as a uniform random variable over the set $\{1, \ldots, M\}$ to denote the true model for the pair $\mathcal{G}_1$ and $\mathcal{G}_2$ from class $\mathcal{S}_q$. Define $\mathbb{Q}_i$ as the joint probability measure of $\mathbf{X}_1$ and $\mathbf{X}_2$ when $\zeta = i$. Also, we use the notations $\mathcal{G}_1^i \triangleq (V, E_1^i)$ and $\mathcal{G}_2^i \triangleq (V, E_2^i)$ to identify the pair of q-similar graph when $\zeta = i$. Hence, the KullbackLeibler (KL) divergence between two distinct models $\mathbb{Q}_i$ and $\mathbb{Q}_j$ is given by

$$D_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_i) \triangleq \int_{\mathbf{X}_1, \mathbf{X}_2} \log\left(\frac{\mathrm{d}\mathbb{Q}_i}{\mathrm{d}\mathbb{Q}_j}\right) \mathrm{d}\mathbb{Q}_i \ , \tag{58}$$

where $\frac{\mathrm{d}\mathbb{Q}_i}{\mathrm{d}\mathbb{Q}_j}$ is the Radon-Nikodym derivative of $\mathbb{Q}_i$ with respect to $\mathbb{Q}_j$. Accordingly, for each distinct pair $i, j \in \{1, \ldots, M\}$ we define the symmetricized KL divergence as:

$$S_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_j) \triangleq D_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_j) \ + D_{\mathrm{KL}}(\mathbb{Q}_j \| \mathbb{Q}_i) \ , \tag{59}$$

where for Ising models, $S_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_j)$ can be readily verified to be [18], [19]:

$$\begin{aligned} &S_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_j) \\ &= \sum_{r \in \{1,2\}} \left(\sum_{(u,v) \in E_r^i \setminus E_r^j} \lambda_r^{uv} (\mathbb{E}_i[X_r^u X_r^v] - \mathbb{E}_j[X_r^u X_r^v]) \right. \\ &\qquad \left. + \sum_{(u,v) \in E_r^j \setminus E_r^i} \lambda_r^{uv} (\mathbb{E}_j[X_r^u X_r^v] - \mathbb{E}_i[X_r^u X_r^v]) \right) \ , \end{aligned} \tag{60}$$

where $\mathbb{E}_j[X_r^u X_r^v]$ is the expected value of the random variable $X_r^u X_r^v$ in graph $\mathcal{G}_r^j$ for $u, v \in V$. Furthermore, we define $I(\zeta; \mathbf{X}_1, \mathbf{X}_2)$ as the mutual information between the random variable ζ (capturing the true pair model) and one pair of graph samples $(\mathbf{X}_1, \mathbf{X}_2)$. By using the convexity of KL divergence, it follows that

$$I(\zeta; \mathbf{X}_1, \mathbf{X}_2) \leq \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} S_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_i) \ . \tag{61}$$

For any class of graphs $\mathcal{S}_q$ with M number of total possible true graphs, we use the following variants of Fano's Lemma [44].

Lemma 2 (Fano's Lemma)**.** *For any class of graphs $\mathcal{S}_q$ with M members, if the number of samples is upper bounded by*

$$n \leq \frac{(1-\varepsilon)(\log M - 1) + \varepsilon}{I(\zeta; \mathbf{X}_1, \mathbf{X}_2)} \ , \tag{62}$$

for some $\varepsilon \in (0, 1)$, then the probability of forming erroneous estimates of the true pair of graphs $\mathsf{P}(\mathcal{S}_q)$ satisfies:

$$\mathsf{P}(\mathcal{S}_q) \geq \varepsilon \ . \tag{63}$$

Lemma 2 characterizes the upper bound on the number of samples for which $\mathsf{P}(\mathcal{S}_q)$ is guaranteed to be bounded away from 0 by an arbitrary finite value. Note that, in general, ε is intended to be small and closer to 0, and, therefore, for clarity in presentation, we will use a slightly looser condition on n. Specifically, for any decoder to satisfy $\mathsf{P}(\mathcal{S}_q) \leq \varepsilon$, we have

$$n \geq \frac{(1-\varepsilon)}{I(\zeta; \mathbf{X}_1, \mathbf{X}_2)}(\log M - 1) \ . \tag{64}$$

A. Ising Models: Degree-bounded Subclass

Next, we provide the constructions of different ensembles in the class $\mathcal{I}_q^d(\lambda, \vartheta)$ that enable us to recover the necessary conditions on sample complexity in Theorem 1.

1) Ensemble 1: Single-edge Graphs: We first consider an ensemble of the graph pairs such that each graph has exactly one edge. Clearly, the number of such pairs is

$$b_1 = \binom{p-q}{2}^2 + \binom{q}{2} \ . \tag{65}$$

In this ensemble, when the single edge connects two shared nodes, both graphs in the pair are identical. We remark that the graphs in this ensemble lie in both $\mathcal{I}_q^d(\lambda, \vartheta)$ and $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$. Note that in this ensemble, if we have $(u, v) \notin E_i$, then random variables X_i^u and X_i^v are independent. Also, for any pair of edges $(u, v) \in E_r^i$ and $(w, x) \in E_r^j$, we have $\mathbb{E}_i[X_r^u X_r^v] = \mathbb{E}_j[X_r^u X_r^v] = \tanh\lambda$ for $r \in \{1, 2\}$ and $i, j \in \{1, \ldots, M\}$. By leveraging these observations, for each pair of graphs, we have

$$S_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_j) = 4\lambda \tanh\lambda \ . \tag{66}$$

Based on (64) and (66), for any graph decoder whose maximal error in recovering any pair of graphs is less than ε, we must have

$$n \geq \frac{(1-\varepsilon)(\log b_1 - 1)}{4\lambda \tanh\lambda} \ . \tag{67}$$

It can be verified that

$$\binom{q}{2} \leq b_1 \ \leq \ 2\binom{q}{2} \ , \qquad \text{if } \binom{q}{2} \geq \binom{p-q}{2}^2 \ , \tag{68}$$

$$\binom{p-q}{2}^2 \leq b_1 \ \leq \ 2\binom{p-q}{2}^2 \ , \quad \text{if } \binom{q}{2} < \binom{p-q}{2}^2 \ . \tag{69}$$

Clearly, $b_1 = \Theta(q^2)$ in regime in (68) and $b_1 = \Theta((p-q)^2)$ in regime in (69), which clearly distinguish the scaling behavior in terms of q in the two regimes. For clarity, we further lower bound b_1 by $(q-1)^2/2$ in the regime in (68) and by $(p-q-1)^2/2$ in the regime in (69). Therefore, to accurately capture the effect of structural similarity on the sample complexity in the two regimes, we restate and simplify (67) as

$$n \geq \frac{1-\varepsilon}{4\lambda \tanh\lambda} \max\{A_1, A_2\} \ , \tag{70}$$

where

$$A_1 \ \triangleq \ 2\log\frac{q-1}{\sqrt{2}} - 1 \ , \ \text{ and } \ A_2 \ \triangleq \ 4\log\frac{p-q-1}{\sqrt{2}} - 1 \ . \tag{71}$$

2) Ensemble 2: $(d+1)$-vertex Fully-connected Subgraphs: For constructing this ensemble, we divide the shared vertices into a group of $(d+1)$ vertices, rendering $\lfloor q/(d+1) \rfloor$ such groups. We partition each of the two non-shared parts of the two graphs into groups of $(d+1)$ vertices. Hence, the total number of these groups is

$$\left\lfloor \frac{p-q}{d+1} \right\rfloor^2 + \left\lfloor \frac{q}{d+1} \right\rfloor \ . \tag{72}$$

The vertices within each group are fully-connected, and there is no inter-group edges. Note that the selection of the nodes and placing them into different groups has been arbitrary. We refer to the two graphs defined over E_1 and E_2 as the *base* graphs, and we denote them by $\mathcal{G}_1^{\mathrm{b}}$ and $\mathcal{G}_2^{\mathrm{b}}$, respectively. Next, we use this base graphs to construct an ensemble of graph pairs. Specifically, the ensemble includes all possible graph pairs $(\mathcal{G}_1, \mathcal{G}_2)$ such that $\mathcal{G}_i$ constructed by removing one edge of $\mathcal{G}_i^{\mathrm{b}}$. This ensemble of graphs lies in the class $\mathcal{I}_q^d(\lambda, \vartheta)$. By noting that the total number of fully-connected cliques in the base graphs is given in (72), it can be readily verified that the total number of graph pairs in this ensemble is given by

$$\left[\left\lfloor \frac{p-q}{d+1} \right\rfloor \binom{d+1}{2}\right]^2 + \left\lfloor \frac{q}{d+1} \right\rfloor \binom{d+1}{2} \ . \tag{73}$$

For the rest of analysis, we consider two regimes depending on the relative values of p and q.

1) **Regime 1:** In this regime we have either $(p-q) < 2(d+1)$ or $q < 2(d+1)$, resulting in $\left\lfloor \frac{p-q}{d+1} \right\rfloor = 1$ or $\left\lfloor \frac{q}{d+1} \right\rfloor = 1$, respectively. Hence, the number of graph pairs specified in (73) is lower bounded by
   a) if $q < 2(d+1)$ and $(p-q) \geq 2(d+1)$,
   $$b_2 \ \triangleq \ \left[\frac{(p-q)d}{4}\right]^2 + \frac{(d+1)d}{2} \ , \tag{74}$$
   b) if $q \geq 2(d+1)$ and $(p-q) < 2(d+1)$,
   $$b_3 \ \triangleq \ \left[\frac{(d+1)d}{2}\right]^2 + \frac{qd}{4} \ , \tag{75}$$
   c) if $q < 2(d+1)$ and $(p-q) < 2(d+1)$,
   $$b_4 \ \triangleq \ \left[\frac{(d+1)d}{2}\right]^2 + \frac{(d+1)d}{2} \ . \tag{76}$$
   We note that the regime (75) is characterized by the graphs being overly similar. For instance, for $d > 0.3p$, the regime in (75) corresponds to the case where $q > 0.6p$.
2) **Regime 2:** This is the complement of Regime 1, i.e., $(p-q) \geq 2(d+1)$ and $q \geq 2(d+1)$. In this regime, the number of graph pairs specified in (73) can be lower bounded by
   $$b_5 \ \triangleq \ \left[\frac{(p-q)d}{4}\right]^2 + \frac{qd}{4} \ , \tag{77}$$
   By using the bound $2(d+1) > q$ in (74), $2(d+1) > (p-q)$ in (75), and the corresponding lower bounds on $(d+1)$ in (76), we observe that the number of graphs

b_2, b_3 and b_4 in their respective regimes can be further lower bounded by a term equivalent to b_5, indicating that the analyses of the structure estimation algorithm in the regimes associated with (74), (75), (76), and (77) are equivalent. Therefore, in the subsequent analysis, we use the lower bound b_5 on the number of graphs in this ensemble.

By following the separation argument in [18, Lemma 2], we find that when $\lambda \geq 1/d$, for any pair of graphs in this ensemble we have

$$S_{\mathrm{KL}}(\mathbb{Q}_i \| \mathbb{Q}_j) \leq \frac{8\vartheta \exp(\lambda)}{\exp(\vartheta)} \ . \tag{78}$$

Therefore, using the number of graph pairs in different regimes and the divergence result in (78) in Fano's Lemma in (64), we conclude that for any graph decoder to have the error probability in recovering any pair of graphs in this ensemble less than ε, it is necessary that

$$n \geq \frac{\exp(\vartheta)(\log b_5 - 1)}{8\vartheta \exp(\lambda)} \ . \tag{79}$$

Clearly, the bound on sample complexity scales exponentially with at least λd as we have $\vartheta \geq \lambda d$, which reflects the difficulty in estimating the graphs with large edge weights. We simplify the condition in b_5 by characterizing the regimes for which $\frac{qd}{4}$ dominates $\left[\frac{(p-q)d}{4}\right]^2$ and vice-versa. Specifically, it can be verified that

$$\left[\frac{(p-q)d}{4}\right]^2 \leq b_5 \ \leq \ 2\left[\frac{(p-q)d}{4}\right]^2 \ , \quad \text{if } \ q + 2\sqrt{\frac{q}{d}} \leq p, \tag{80}$$

$$\frac{qd}{4} \leq b_5 \ \leq \ \frac{qd}{2} \ , \qquad \text{if } \ q + 2\sqrt{\frac{q}{d}} > p \ . \tag{81}$$

Clearly, we have $b_5 = \Theta((p-q)^2 d^2)$ in the regime in (80), and $b_5 = \Theta(qd)$ in the regime in (81). Therefore, for a better intuition into the effect of structure similarity on sample complexity, by leveraging (80) and (81), we simplify the condition in (79) to

$$n \geq \frac{\exp(\vartheta - \lambda)}{8\vartheta} \max\{A_3, A_4\} \ , \tag{82}$$

where

$$A_3 \ \triangleq \ \log \frac{qd}{4} - 1 \ , \quad \text{and} \quad A_4 \ \triangleq \ 2\log \frac{(p-q)d}{4} - 1 \ . \tag{83}$$

Finally, we combine the findings from the analysis of the sample complexities of recovering graphs in Ensemble 1 and Ensemble 2 to provide the necessary conditions on the sample complexity for recovering the graphs in $\mathcal{I}_q^d(\lambda, \vartheta)$.

Lemma 3 (Degree-bounded)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the class $\mathcal{I}_q^d(\lambda, \vartheta)$. Any graph decoder ψ that achieves $\mathsf{P}(\mathcal{I}_q^d(\lambda, \vartheta)) \leq \varepsilon$ must satisfy*

$$n \geq (1-\varepsilon) \max \Bigg\{ \frac{1}{4\lambda \tanh \lambda} \max\{A_1, A_2\}, \tag{84}$$

$$\frac{\exp(\vartheta)}{8\vartheta \exp(\lambda)} \max\{A_3, A_4\} \Bigg\} \ . \tag{85}$$

Next, we note that in the regime $\lambda = O(1/d)$, we have $\frac{1}{\lambda \tanh \lambda} = \Omega(d^2)$. Note that according to the definition of ϑ in (4), variations in d inevitably induce variations in ϑ as well since we have $\vartheta \geq \lambda d$. However, in the regime $\lambda = O(1/d)$, the effect of ϑ can be controlled because in this regime, ϑ can be set to an arbitrary constant that will never be exceeded by its minimum feasible value for any combination of λ and d. Therefore, the term $\frac{1}{4\lambda \tanh \lambda} \max\{A_1, A_2\}$ dominates the sample complexity when we have $\lambda = O(1/d)$ as the second term in (84) has only a logarithmic scaling behavior in $(p-q)$ or q. In contrast, when we have $\lambda = \Theta(1)$ or $\lambda = \Theta(d)$, the second term in (84) is characterized by an exponential scaling behavior in ϑ which dominates the scaling behavior of the first term. These observations complete the proof of Theorem 1.

B. Ising Models: Edge-bounded Subclass

The proof of the necessary conditions on the sample complexity for recovering graphs in $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$ follows the same template of application of Fano's Lemma as discussed in Section V-A. Therefore, in this section, we discuss the construction of ensembles in the class $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$.

Note that the construction of Ensemble 1 discussed in Section V-A consists of one edge per graph and, therefore, it is also valid for the class $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$. We provide the construction of another ensemble to cover the scenario of densely-connected graphs.

1) Ensemble 3: Graphs with Cliques: Let m_1 be the largest integer such that $\gamma k \geq \binom{m_1}{2}$ and let m_2 be the largest integer such that $\bar{\gamma} k \geq \binom{m_2}{2}$, where $\bar{\gamma} = 1 - \gamma$. Clearly, m_1 and m_2 satisfy

$$\frac{\sqrt{\gamma k}}{2} \leq \lfloor \sqrt{\gamma k} \rfloor \leq m_1 \leq 2\sqrt{\gamma k} \tag{86}$$

and

$$\frac{\sqrt{\bar{\gamma} k}}{2} \leq \lfloor \sqrt{\bar{\gamma} k} \rfloor \leq m_2 \leq 2\sqrt{\bar{\gamma} k} \ . \tag{87}$$

We form a base pair of q-similar graphs, denoted by $\mathcal{G}_1^{\mathrm{b}}$ and $\mathcal{G}_2^{\mathrm{b}}$, by constructing a fully-connected clique of m_1 vertices in the shared cluster and a fully-connected clique of m_2 vertices in the non-shared cluster of each graph. Such selection of the nodes and placing them into different groups for the base pair is arbitrary. Next, we use this base pair of graphs to characterize an ensemble of pairs of q-similar graphs. Specifically, the ensemble includes all possible graphs pairs $\mathcal{G}_1, \mathcal{G}_2)$ such that the graph $\mathcal{G}_i$ is constructed from the base graph $\mathcal{G}_i^{\mathrm{b}}$ by removal of one edge. Considering the similarity between the two graphs, there are

$$b_6 \triangleq \binom{m_2}{2}^2 + \binom{m_1}{2} \tag{88}$$

possible pairs of graphs in this ensemble. By leveraging [18, Lemma 3], we have

$$S_{\mathrm{KL}}(\mathbb{Q}_i\|\mathbb{Q}_j) \leq \left(\frac{16\lambda m_1}{\exp(\lambda m_1)}+\frac{16\lambda m_2}{\exp(\lambda m_2)}\right)\exp(2\lambda)\sinh\lambda \tag{89}$$

$$\leq \left(\frac{32\lambda\sqrt{\gamma k}}{\exp(\lambda\sqrt{\gamma k}/2)}+\frac{32\lambda\sqrt{\bar{\gamma} k}}{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}\right)\exp(2\lambda)\sinh\lambda\ , \tag{90}$$

for any two distinct pair of graphs in this ensemble, where (90) follows from (89) by leveraging (86) and (87). Using Lemma 2, we get the necessary condition

$$n > \frac{(1-\varepsilon)(\log b_6-1)}{32\lambda\sqrt{k}\exp(2\lambda)\sinh(\lambda)} \times\left(\frac{\sqrt{\gamma}}{\exp(\lambda\sqrt{\gamma k}/2)}+\frac{\sqrt{\bar{\gamma}}}{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}\right)^{-1} \tag{91}$$

$$\geq \frac{(1-\varepsilon)b_7}{32\lambda\sqrt{k}\exp(2\lambda)\sinh(\lambda)}\ , \tag{92}$$

where

$$b_7 \triangleq \left[\log\left(\frac{\bar{\gamma}^2k^2}{16}+\frac{\gamma k}{4}\right)-1\right] \times\left(\frac{\sqrt{\gamma}}{\exp(\lambda\sqrt{\gamma k}/2)}+\frac{\sqrt{\bar{\gamma}}}{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}\right)^{-1}\ , \tag{93}$$

so that the error probability for the exact recovery is at most ε. To magnify the focus on the effect of shared structure in the graph, we relax the condition in (92) by investigating the regimes when the terms characterized by γk or $\bar{\gamma} k$ dominate the sample complexity. We make the following observations in different regimes regarding b_7.

1) **Regime 1**: In this regime, for $\bar{\gamma} > 0.5$, we have
   a) if $2\lambda^2 k \geq \frac{\log^2(\bar{\gamma}/\gamma)}{(1-2\sqrt{\gamma\bar{\gamma}})}$,
   $$\frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma}}} \geq \frac{\exp(\lambda\sqrt{\gamma k}/2)}{\sqrt{\gamma}}\ , \tag{94}$$
   b) otherwise,
   $$\frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma}}} < \frac{\exp(\lambda\sqrt{\gamma k}/2)}{\sqrt{\gamma}}\ . \tag{95}$$

   Furthermore,
   $$\frac{\bar{\gamma}^2k^2}{16} \geq \frac{\gamma k}{4}\ , \qquad \text{if } k > \frac{4\gamma}{\bar{\gamma}^2}\ , \tag{96}$$
   $$\frac{\bar{\gamma}^2k^2}{16} < \frac{\gamma k}{4}\ , \qquad \text{otherwise}\ . \tag{97}$$

   Note that on comparing the regimes in (96) and (97) for $k > 8$, the dominant regime is always (96). Therefore, we focus our subsequent discussion on (96). We also remark that the regime in (94) holds for a wide range of combinations of γ and k, except for the values of λ in the asymptote of $\lambda \to 0$. For instance, we have $2\lambda^2 k > 2.9$ when $\gamma = 0.05$ and $2\lambda^2 k > 1.509$ when $\gamma = 0.5$. In the asymptote of $\lambda \to 0$, the sample complexity scales as logarithmic factors in k and our analysis will reveal that in the regime $\lambda = O(1/\sqrt{k})$, the necessary conditions for recovering graph-pairs in Ensemble 1 have a linear dependence on k and, therefore, they dominate the sample complexity. Therefore, between (94) and (95), we focus our discussions only on (94). In the regime specified by (94) and (96), we have
   $$\left(\log\frac{\bar{\gamma}^2k^2}{16}-1\right)\frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma}}} \leq b_7 \tag{98}$$
   and
   $$b_7 \leq 2\left(\log\frac{\bar{\gamma}^2k^2}{8}-1\right)\frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma}}}\ . \tag{99}$$
   Therefore, in this regime, we have
   $$b_7 = \Theta\left(\log\bar{\gamma}^2k^2\cdot\frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma}}}\right)\ , \tag{100}$$
   and the overall sample complexity is dominated by $\bar{\gamma} k$, which specifies the maximum number of edges in the non-shared parts of the graphs. To place the emphasis on the dominating effect of non-shared part, in the regime specified jointly by the conditions $\bar{\gamma} > 0.5, k > \frac{4\gamma}{\bar{\gamma}^2}, 2\lambda^2 k \geq \frac{\log^2(\bar{\gamma}/\gamma)}{(1-2\sqrt{\gamma\bar{\gamma}})}$, we modify the necessary condition on the number of samples for any graph decoder to achieve a recovery error less than ε from (92) to
   $$n > \frac{(1-\varepsilon)}{32\lambda\exp(2\lambda)\sinh(\lambda)} \times A_5\ , \tag{101}$$
   where
   $$A_5 \triangleq \frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma} k}}\left(\log\frac{\bar{\gamma}^2k^2}{16}-1\right)\ , \tag{102}$$

2) **Regime 2**: In this regime, for $\gamma > 0.5$, we have
   a) if $2\lambda^2 k < \frac{\log^2(\gamma/\bar{\gamma})}{(1-2\sqrt{\gamma\bar{\gamma}})}$,
   $$\frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma}}} \geq \frac{\exp(\lambda\sqrt{\gamma k}/2)}{\sqrt{\gamma}}\ , \tag{103}$$
   b) otherwise,
   $$\frac{\exp(\lambda\sqrt{\bar{\gamma} k}/2)}{\sqrt{\bar{\gamma}}} < \frac{\exp(\lambda\sqrt{\gamma k}/2)}{\sqrt{\gamma}}\ . \tag{104}$$

   Furthermore,
   $$\frac{\bar{\gamma}^2k^2}{16} \geq \frac{\gamma k}{4}\ , \qquad \text{if } k > \frac{4\gamma}{\bar{\gamma}^2}\ , \tag{105}$$
   $$\frac{\bar{\gamma}^2k^2}{16} < \frac{\gamma k}{4}\ , \qquad \text{otherwise}\ . \tag{106}$$

   We remark that for moderate to large sized graphs (for instance, $k > 50$) the regime in (103) is applicable only for small values of λ. Since the sample complexity for small λ is dominated by that for ensemble 1, we focus our discussion on (104). The condition in (105) is satisfied for $k > 8$ when $\gamma \approx 0.5$. However, as γ becomes larger and gets closer to 1, the feasible values of k under this regime become larger. For instance, the regime in (105) implies that for $\gamma = 0.1$, we must have $k > 360$, and for $\gamma = 0.05$, we must have $k > 1520$. Using a similar line of arguments as in the Regime 1, for the regime specified jointly by the conditions

$\gamma > 0.5$ and $2\lambda^2 k \geq \frac{\log^2(\gamma/\bar{\gamma})}{(1-2\sqrt{\gamma\bar{\gamma}})}$, we modify the necessary condition on sample complexity from (92) to

$$n > \frac{(1-\varepsilon)}{32\lambda \exp(2\lambda)\sinh(\lambda)} \times A_6 \ , \quad \text{if } \ k \leq \frac{4\gamma}{\bar{\gamma}^2} \ , \tag{107}$$

where

$$A_6 \triangleq \frac{\exp(\lambda\sqrt{\gamma k}/2)}{\sqrt{\gamma k}} \left(\log \frac{\gamma k}{4} - 1 \right) \ , \tag{108}$$

and

$$n > \frac{(1-\varepsilon)}{32\lambda \exp(2\lambda)\sinh(\lambda)} \times A_7 \ , \quad \text{if } \ k > \frac{4\gamma}{\bar{\gamma}^2} \ , \tag{109}$$

where

$$A_7 \triangleq \frac{\exp(\lambda\sqrt{\gamma k}/2)}{\sqrt{\gamma k}} \left(\log \frac{\bar{\gamma}^2 k^2}{16} - 1 \right) \ . \tag{110}$$

We summarize the results from Ensemble 1 in Section V-A and Ensemble 3 to provide the necessary conditions for joint recovery of q-similar graphs in the edge-bounded subclass.

Lemma 4 (Edge-bounded)**.** *Consider a pair of q-similar graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in the class $\mathcal{I}^k_{q,\gamma}(\lambda, \vartheta)$. Any graph decoder ψ that achieves $\mathsf{P}(\mathcal{I}^k_{q,\gamma}(\lambda,\vartheta)) \leq \varepsilon$, must satisfy*

$$n \geq (1-\varepsilon) \max \left\{ \frac{1}{4\lambda \tanh \lambda} \max\{A_1, A_2\} \ , \right. \\ \left. \frac{1}{32\lambda \exp(2\lambda)\sinh(\lambda)} \max\{A_5, A_6, A_7\} \right\} \ . \tag{111}$$

Next, we note that in the regime $\lambda = O(1/\sqrt{k})$, we have $\frac{1}{\lambda \tanh \lambda} = \Omega(k)$. According to the definition of ϑ in (4), variations in k inevitably induce variations in ϑ as well. However, in the regime $\lambda = O(1/\sqrt{k})$, the effect of ϑ can be controlled because for the graph models in ensemble 3 that lead to the factor $\exp(\vartheta)$ in the sample complexity, we have $\vartheta \geq \lambda\sqrt{k}$. Therefore, in this regime, ϑ can be set to an arbitrary constant that will never be exceeded by its minimum feasible value for any combination of λ and k. Hence, $\frac{1}{4\lambda \tanh \lambda} \max\{A_1, A_2\}$ dominates the sample complexity in this regime. On the other hand, in the regimes $\lambda = \Theta(1)$ and $\lambda = \Theta(\sqrt{k})$, the lower bound on ϑ increases with an increase in k and, therefore, the second term in (111) grows exponentially and dominates the sample complexity. Specifically, when we have $\lambda = \Theta(1)$ or $\lambda = \Theta(\sqrt{k})$, the sample complexity has an exponential behavior in $\sqrt{k}$ as k increases. These observations complete the proof of Theorem 2.

C. Gaussian Models

To recover the necessary conditions for Gaussian models, we consider two simple ensembles for exact recovery of graph pairs in the class $\mathcal{G}^d_q(\rho)$ and apply Fano's Lemma.

1) Ensemble 1: Sparsely-connected Graphs: We consider an ensemble of graph pairs in which each graph consists of only one edge. Therefore, this ensemble of graphs lies in both the degree-bounded and the edge-bounded subclasses of q-similar Gaussian models. For our analysis, we consider two specific cases corresponding to whether the edge lies in the shared part or the non-shared part of a graph in a pair of q-similar graphs.

1) **Case 1**: We first consider the case where the edge lies in the shared cluster for both graphs. Therefore, the problem of exact recovery of the two graphs becomes equivalent to the problem of exact recovery of a single graph from $2n$ samples. We note that there are $\binom{q}{2}$ number of possible graph pairs in this scenario. Furthermore, using the entropy based bound in [23, Theorem 1], we have
$$I(\zeta; \mathbf{X}_1, \mathbf{X}_2) \leq 8\rho^2 \ . \tag{112}$$
By leveraging (112) and Lemma 2, we obtain that in order for the recovery error to be upper bounded by ε, we must have
$$n > \frac{(1-\varepsilon)}{8\rho^2} \left[\log \binom{q}{2} - 1 \right] \ , \tag{113}$$
for $\rho \in [0, 1/2]$.
2) **Case 2**: In this scenario, we assume that the edge lies in the non-shared part of the q-similar graph pair. Therefore, there are $\binom{p-q}{2}^2$ possible number of such graph pairs. By using (112), we get the necessary condition
$$n > \frac{(1-\varepsilon)}{8\rho^2} \left(2 \log \binom{p-q}{2} - 1 \right) \ , \tag{114}$$
for the recovery error to be upper bounded by ε.

Clearly, (113) and (114) lay emphasis on the sample complexity due to shared cluster and the non-shared clusters in the two graphs, respectively. Furthermore, the bound in (113) dominates that in (114) if we have $\log \binom{q}{2} \geq 2 \log \binom{p-q}{2}$. To further emphasize the effect of q on the sample complexity, we slightly relax the results in (113) and (114) and conclude that in order for the recovery error to be upper bounded by ε, the the number of samples must satisfy

$$n > \frac{(1-\varepsilon)}{4\rho^2} \max \left\{ \log \frac{q-1}{\sqrt{2}} - 1, 2 \log \frac{p-q-1}{\sqrt{2}} - 1 \right\} \ , \tag{115}$$

for $\rho \in (0, 1/2]$.

2) Ensemble 2: Densely-connected graphs: In this ensemble, we consider the graph pairs in which each graph consists of only one clique of m vertices. We assume that the clique can completely lie either in the shared part or in the non-shared part of the graph. This leads us to consider two cases.

1) **Case 1**: When the clique lies completely in the shared cluster of the two graphs, we have $\binom{q}{m}$ possible number of graph pairs. Furthermore, using the KL divergence based bound in [23, Theorem 1], we have
$$I(\zeta; \mathbf{X}_1, \mathbf{X}_2) \leq \log \left(1 + \frac{m\rho}{1-\rho} \right) - \frac{m\rho}{1+(m-1)\rho} \ . \tag{116}$$

Therefore, by using Lemma 2, we get the necessary condition that

$$n > (1-\varepsilon)\frac{\log\binom{q}{m}-1}{\log\left(1+\frac{m\rho}{1-\rho}\right)-\frac{m\rho}{1+(m-1)\rho}}\ , \qquad (117)$$

for achieving $\mathbb{P}(\mathcal{G}_q^d(\rho)) \leq \varepsilon$.

2) **Case 2**: When the clique lies completely in the non-shared clusters of the two graphs, we have $\binom{p-q}{m}^2$ number of possible graph pairs. Therefore, by leveraging (116) and Lemma 2, we get the necessary condition that n must satisfy

$$n > (1-\varepsilon)\frac{2\log\binom{p-q}{m}-1}{\log\left(1+\frac{m\rho}{1-\rho}\right)-\frac{m\rho}{1+(m-1)\rho}}\ , \qquad (118)$$

for achieving $\mathbb{P}(\mathcal{G}_q^d(\rho)) \leq \varepsilon$.

For the degree-bounded subclass, we set $m = d$. To recover the results for edge-bounded subclass, we set $m = \sqrt{k}$ in this ensemble. Finally, we summarize the results from the two ensembles to provide the necessary conditions for joint recovery of q-similar graphs in the degree-bounded and edge-bounded subclasses.

Lemma 5 (Degree-bounded)**.** *Consider a pair of q-similar graphs in the class $\mathcal{G}_q^d(\rho)$. Any graph decoder that achieves $\mathsf{P}(\mathcal{G}_q^d(\rho)) \leq \varepsilon$ must satisfy*

$$n \geq (1-\varepsilon)\max\{C_1, C_2\}\ , \qquad (119)$$

where

$$C_1 \triangleq \frac{1}{4\rho^2}\max\left\{\log\frac{q-1}{\sqrt{2}}-1, 2\log\frac{p-q-1}{\sqrt{2}}-1\right\}\ , \qquad (120)$$

$$C_2 \triangleq \frac{1}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} \times \max\left\{\log\binom{q}{d}-1, 2\log\binom{p-q}{d}-1\right\}\ . \qquad (121)$$

We note that in the regime $\rho = \Theta(1/d)$, we have $\rho d = \Theta(1)$ and, therefore, C_2 scales proportional to $d\max\{\log q/d, 2\log(p-q)/d\}$. On the other hand, in this regime, C_1 has a scaling behavior proportional to $d^2\max\{\log q, 2\log(p-q)\}$, which clearly dominates that of C_2. In the regime $\rho = \Theta(1)$, we have $\log\binom{q}{d} \geq d\log\frac{q}{d}$ and $\log\binom{p-q}{d} \geq d\log\frac{p-q}{d}$ and, therefore, C_2 dominates the sample complexity as C_1 has only a logarithmic scaling behavior in q or $(p-q)$. These observations complete the proof of Theorem 7. The necessary conditions on the sample complexity for $\mathcal{G}_q^k(\rho)$ are formalized below.

Lemma 6 (Edge-bounded)**.** *Consider a pair of q-similar graphs in the class $\mathcal{G}_q^k(\rho)$. Any graph decoder that achieves $\mathsf{P}(\mathcal{G}_q^k(\rho)) \leq \varepsilon$ must satisfy*

$$n \geq \frac{1-\varepsilon}{8\rho^2}\max\{C_1, C_3\}\ , \qquad (122)$$

where

$$C_3 \triangleq \frac{1}{\log\left(1+\frac{\tilde{k}\rho}{1-\rho}\right)-\frac{\tilde{k}\rho}{1+(\tilde{k}-1)\rho}} \times \max\left\{\log\binom{q}{\tilde{k}}-1, 2\log\binom{p-q}{\tilde{k}}-1\right\}\ . \qquad (123)$$

We note that in the regime $\rho = \Theta(1/\tilde{k})$, we have $\rho\tilde{k} = \Theta(1)$ and, therefore, C_1 dominates the sample complexity. On the other hand, in the regime $\rho = \Theta(1)$, we have $\log\binom{q}{\tilde{k}} \geq \tilde{k}\log\frac{q}{\tilde{k}}$ and $\log\binom{p-q}{\tilde{k}} \geq \tilde{k}\log\frac{p-q}{\tilde{k}}$ and, therefore, C_2 dominates the sample complexity. These observations complete the proof of Theorem 8.

VI. Proofs of Sufficient Conditions

To establish the sufficient conditions, we analyze the sample complexity of an ML decoder using the large deviations bound. We first provide the general setup for the analysis of an ML decoder for any generic class. Similarly to the proof of necessary conditions in Section V-A, we consider a subclass $\mathcal{S}_q$ of q-similar graphs that consists of $M = |\mathcal{S}_q|$ pair of q-similar graphs. The graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are selected from $\mathcal{S}_q$ uniformly at random and the random variable $\zeta \in \{1, \ldots, M\}$ denotes the true model. When $\zeta = i$, the pair of q-similar graphs are denoted by $\mathcal{G}_1^i \triangleq (V, E_1^i)$ and $\mathcal{G}_2^i = (V, E_2^i)$. Given the collections of graph samples $(\mathbf{x}_1^n, \mathbf{x}_2^n)$, the ML decoder decides on the true models according to the rule given by

$$\Psi(\mathbf{x}_1^n, \mathbf{x}_2^n) = \arg\max_{i\in\{1,\ldots,M\}} \ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n)\ , \qquad (124)$$

where $\ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n)$ is the log likelihood with respect to the model $i \in \{1, \ldots, M\}$ and is given by

$$\ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n) \triangleq \sum_{w=1}^{n} \log \mathrm{d}\mathbb{Q}_i(x_1(w), x_2(w))\ , \qquad (125)$$

where $x_u(w)$ is the w-th sample of $\mathbf{x}_u^n$. If the solution to (124) is not unique, we randomly select one. If the data $(\mathbf{x}_1^n, \mathbf{x}_2^n)$ is collected from a pair of graphs with true model $i \in \{1, \ldots, M\}$, the ML decoder fails to recover the true model only if there exists some other model $j \neq i$, for which, $\ell_j(\mathbf{x}_1^n, \mathbf{x}_2^n) \geq \ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n)$. Therefore, we have

$$\mathbb{P}[\Psi(\mathbf{x}_1^n, \mathbf{x}_2^n) \neq (E_1^i, E_2^i)] = \mathbb{P}\left(\bigcup_{j\in\{1,\ldots,M\}\setminus i} \ell_j(\mathbf{x}_1^n, \mathbf{x}_2^n) \geq \ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n)\right) \qquad (126)$$

$$\leq \sum_{j\in\{1,\ldots,M\}\setminus i} \mathbb{P}[\ell_j(\mathbf{x}_1^n, \mathbf{x}_2^n) \geq \ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n)]\ , \qquad (127)$$

where (127) follows from the union bound. In the proofs of sufficient conditions for all subclasses, we will upper bound the probabilities in (127) such that $\mathbb{P}[\Psi(\mathbf{x}_1^n, \mathbf{x}_2^n) \neq (E_1^i, E_2^i)]$ diminishes with an increase in the number of samples n. Since given the true model pair, the samples $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated independently for both graphs, we have

$$\mathrm{d}\mathbb{Q}_i(\mathbf{x}_1, \mathbf{x}_2) = p_1^i(\mathbf{x}_1)p_2^i(\mathbf{x}_2)\ , \qquad (128)$$

where p_r^i is the marginal probability measure of $\mathbf{X}_r$, for $r \in \{1, 2\}$, under model $i \in \{1, \ldots, M\}$. Next, we discuss the results for Ising models.

A. Ising Models

We start by providing the large deviations bound in Lemma 7, which provides the sufficient conditions for the probability of error of the ML decoder to vanish with an increase in the sample size n. For this purpose, we define $\boldsymbol{\lambda}_r^i \in \mathbb{R}^{\binom{p}{2}}$ as the vector of edge parameters associated with graph $\mathcal{G}_r^i$ when $\zeta = i \in \{1, \ldots, M\}$. Furthermore, we define $Z_r(\boldsymbol{\lambda}_r^i)$ as the partition function of $\mathcal{G}_r^i$ with parameter vector $\boldsymbol{\lambda}_r^i$ and the KL divergence between two Ising models with parameter vectors $\boldsymbol{\lambda}_r^i$ and $\boldsymbol{\lambda}_r^j$ is denoted by $D_{\mathrm{KL}}(\boldsymbol{\lambda}_r^i \| \boldsymbol{\lambda}_r^j)$ for $j \in \{1, \ldots, M\}$.

Lemma 7. *Given the i.i.d. graph samples $(\mathbf{x}_1^n, \mathbf{x}_2^n)$ from the model $i \in \{1, \ldots, M\}$, for any model $j \neq i$, we have*

$$\begin{aligned}&\mathbb{P}[\ell_j(\mathbf{x}_1^n, \mathbf{x}_2^n) \geq \ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n)] \\ &\quad \leq \exp\Big(-\frac{n}{2}(J(\boldsymbol{\lambda}_1^i \| \boldsymbol{\lambda}_1^j) + J(\boldsymbol{\lambda}_2^i \| \boldsymbol{\lambda}_2^j))\Big) , \end{aligned} \tag{129}$$

where we have defined

$$\begin{aligned}J(\boldsymbol{\lambda}_r^i \| \boldsymbol{\lambda}_r^j) \;\triangleq\; & D_{\mathrm{KL}}\left(\frac{\boldsymbol{\lambda}_r^i + \boldsymbol{\lambda}_r^j}{2} \parallel \boldsymbol{\lambda}_r^i\right) \\ & + D_{\mathsf{KL}}\left(\frac{\boldsymbol{\lambda}_r^i + \boldsymbol{\lambda}_r^j}{2} \parallel \boldsymbol{\lambda}_r^j\right) , \end{aligned} \tag{130}$$

for $r \in \{1, 2\}$.

Proof. Let $R \triangleq \ell_j(\mathbf{x}_1^n, \mathbf{x}_2^n) - \ell_i(\mathbf{x}_1^n, \mathbf{x}_2^n)$. Then, using Chernoff's bound, we have

$$\mathbb{P}(R \geq 0) \leq \inf_{s>0} \mathbb{E}_i[\exp(sR)] . \tag{131}$$

Note that

$$\begin{aligned}&\mathbb{E}_i[\exp(sR)] \\ &= \sum_{\mathbf{x}_1^n, \mathbf{x}_2^n} \exp\Big(\sum_{w=1}^n s\ell_j(\mathbf{x}_1(w), \mathbf{x}_2(w)) - s\ell_i(\mathbf{x}_1(w), \mathbf{x}_2(w)))\Big) \\ &\quad \times \prod_{m=1}^n \mathrm{d}\mathbb{Q}_i(\mathbf{x}_1(m), \mathbf{x}_2(m)) . \end{aligned} \tag{132}$$

Then, using

$$\ell_i(\mathbf{x}_1(w), \mathbf{x}_2(w)) = \log \mathrm{d}\mathbb{Q}_i(\mathbf{x}_1(w), \mathbf{x}_2(w)) , \tag{133}$$

from (132) we have

$$\begin{aligned}&\mathbb{E}_i[\exp(sR)] \\ &\sum_{\mathbf{x}_1^n, \mathbf{x}_2^n} \prod_{w=1}^n [\mathrm{d}\mathbb{Q}_j(\mathbf{x}_1(w), \mathbf{x}_2(w))]^s \times [\mathrm{d}\mathbb{Q}_i(\mathbf{x}_1(w), \mathbf{x}_2(w))]^{1-s} , \end{aligned} \tag{134}$$

$$= \left(\sum_{\mathbf{x}_1, \mathbf{x}_2} [\mathrm{d}\mathbb{Q}_j(\mathbf{x}_1, \mathbf{x}_2)]^s [\mathrm{d}\mathbb{Q}_i(\mathbf{x}_1, \mathbf{x}_2)]^{1-s}\right)^n . \tag{135}$$

Using (128) and (135), we have

$$\begin{aligned}&\mathbb{E}_i[\exp(sR)] \\ &= \left(\sum_{\mathbf{x}_1} [p_1^j(\mathbf{x}_1)]^s [p_1^i(\mathbf{x}_1)]^{1-s} \sum_{\mathbf{x}_2} [p_1^j(\mathbf{x}_2)]^s [p_1^i(\mathbf{x}_2)]^{1-s}\right)^n . \end{aligned} \tag{136}$$

From (131), note that by setting $s = 1/2$, we always have

$$\mathbb{P}(R \geq 0) \leq \mathbb{E}_i\left[\exp\left(\frac{R}{2}\right)\right] . \tag{137}$$

Therefore, for $s = 1/2$, by using the expansions of p_1^i and p_1^j specified in (1), it can be readily verified that

$$\sum_{\mathbf{x}_1 \in \{-1,1\}^p} \left[[p_1^j(\mathbf{x}_1)]^{1/2} [p_1^i(\mathbf{x}_1)]^{1/2}\right] = \frac{Z_1\left(\frac{\boldsymbol{\lambda}_1^i + \boldsymbol{\lambda}_1^j}{2}\right)}{(Z_1(\boldsymbol{\lambda}_1^i) Z_1(\boldsymbol{\lambda}_1^j))^{1/2}} , \tag{138}$$

$$= \exp\left(-\frac{J(\boldsymbol{\lambda}_1^i \| \boldsymbol{\lambda}_1^j)}{2}\right) , \tag{139}$$

where $J(\boldsymbol{\lambda}_1^i \| \boldsymbol{\lambda}_1^j)$ is defined in (130). Following a similar analysis as in (138) and (139) for $\mathcal{G}_2$, and by setting $s = 1/2$ in (135), we obtain

$$\mathbb{E}_i[\exp(R/2)] = \exp\Big(-\frac{n}{2}(J(\boldsymbol{\lambda}_1^i \| \boldsymbol{\lambda}_1^j) + J(\boldsymbol{\lambda}_2^i \| \boldsymbol{\lambda}_2^j))\Big) . \tag{140}$$

From (131) and (140), the proof of Lemma 7 is completed. □

Next, we leverage [18, Lemma 4] to find a lower bound on the divergence $J(\boldsymbol{\lambda}_1^i \| \boldsymbol{\lambda}_1^j) + J(\boldsymbol{\lambda}_2^i \| \boldsymbol{\lambda}_2^j))$ in terms of the edge mismatch between the models i and j for $i, j \in \{1, \ldots, M\}$. For $r \in \{1, 2\}$, define $T(\boldsymbol{\lambda}_r^i, \boldsymbol{\lambda}_r^j)$ as the matching number of the graph whose edges are given by the set

$$E_r^i \triangle E_r^j \;\triangleq\; (E_r^i \backslash E_r^j) \cup (E_r^i \backslash E_r^j) . \tag{141}$$

We refer to $|E_r^i \triangle E_r^j|$ as *edit distance* between the models $\mathcal{G}_r^i$ and $\mathcal{G}_r^j$ for $i, j \in \{1, \ldots, M\}$. Then, using [18, Lemma 4], we have

$$J(\boldsymbol{\lambda}_1^i \| \boldsymbol{\lambda}_1^j) + J(\boldsymbol{\lambda}_2^j \| \boldsymbol{\lambda}_2^j) \geq \frac{T(\boldsymbol{\lambda}_1^i, \boldsymbol{\lambda}_1^j) + T(\boldsymbol{\lambda}_2^i, \boldsymbol{\lambda}_2^j)}{3\exp(2\vartheta) + 1} \sinh^2\left(\frac{\lambda}{4}\right) , \tag{142}$$

where ϑ is the maximum neighborhood weight. We use Lemma 7 and (142) to characterize the sufficient conditions for recovery of graphs in $\mathcal{I}_q^d(\lambda, \vartheta)$ in Section VI-A1 and $\mathcal{I}_{q,\gamma}^k(\lambda, \vartheta)$ in Section VI-A2.

1) Proof of Theorem 3: Consider models i and j in the class $\mathcal{I}_q^d(\lambda, \vartheta)$ such that the non-shared parts of the graphs $\mathcal{G}_1^i$ and $\mathcal{G}_1^j$ have an edit distance e_1, graphs $\mathcal{G}_2^i$ and $\mathcal{G}_2^j$ have an edit distance e_2, and the shared parts for the two models have an edit distance e_{s}. In this case, we have $|E_1^i \triangle E_1^j| = e_1 + e_{\mathsf{s}}$ and $|E_2^i \triangle E_2^j| = e_2 + e_{\mathsf{s}}$. Since the maximum degree of the graphs is bounded by d, we have

$$T(\boldsymbol{\lambda}_1^i, \boldsymbol{\lambda}_1^j) \geq \frac{e_1 + e_{\mathsf{s}}}{4d} , \quad \text{and} \quad T(\boldsymbol{\lambda}_2^i, \boldsymbol{\lambda}_2^j) \geq \frac{e_2 + e_{\mathsf{s}}}{4d} . \tag{143}$$

Furthermore, the shared part of the graphs can have at most $dq/2$ edges and, therefore, e_{s} lies between 0 and qd. Without loss of generality, we assume that $(\mathcal{G}_1^i, \mathcal{G}_2^i)$ are the true models. For each $e_{\mathsf{s}} \in \{0, 1, \ldots, qd\}$, we have at most $\binom{\binom{q}{2}}{e_{\mathsf{s}}}$ models in $\mathcal{I}_q^d(\lambda, \vartheta)$ that have a mismatch of e_{s} edges in the shared part from that in the true model. Also, in general the non-shared

part of the graph can have at most $d(p-q)$ edges (when each edge in the non-shared part is between a node from the non-shared part and a node from the shared part of the graph) and, therefore, e_1 and e_2 lie between 0 and $2d(p-q)$. Therefore, there can be at most

$$\binom{\binom{p-q}{2}+\binom{p-q}{1}\binom{q}{1}}{\ell} \tag{144}$$

number of models in $\mathcal{I}_q^d(\lambda,\vartheta)$ that have a mismatch in ℓ edges from the true model i in the non-shared part. Using (127), the large deviations bound in Lemma 7, and (142) we have

$$\begin{aligned}
&\mathbb{P}[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^i,E_2^i)]\\
&\leq \left(1+\sum_{e_1=1}^{2d(p-q)}B(e_1)\right)^2\left(1+\sum_{e_{\mathsf{s}}=1}^{qd}A(e_{\mathsf{s}})\right)-1 \qquad (145)\\
&=\left(\sum_{e_1=1}^{2d(p-q)}B(e_1)\right)^2+2\sum_{e_1=1}^{2d(p-q)}B(e_1)+\sum_{e_{\mathsf{s}}=1}^{qd}A(e_{\mathsf{s}})\\
&\quad+\left(\sum_{e_1=1}^{2d(p-q)}B(e_1)\right)^2\sum_{e_{\mathsf{s}}=1}^{qd}A(e_{\mathsf{s}})+2\sum_{e_1=1}^{2d(p-q)}B(e_1)\sum_{e_{\mathsf{s}}=1}^{qd}A(e_{\mathsf{s}})\,, \qquad (146)
\end{aligned}$$

where we have defined

$$A(e_{\mathsf{s}}) \triangleq \binom{\binom{q}{2}}{e_{\mathsf{s}}}\exp\left(-n\frac{e_{\mathsf{s}}/(2d)}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\right), \tag{147}$$

and

$$\begin{aligned}
&B(e_1)\\
&\triangleq \binom{\binom{p-q}{2}+\binom{p-q}{1}\binom{q}{1}}{e_1}\exp\left(\frac{-ne_1/(4d)}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\right). \qquad (148)
\end{aligned}$$

If we have

$$\sum_{e_{\mathsf{s}}=1}^{qd}A(e_{\mathsf{s}})\leq\frac{\varepsilon}{4}\,, \tag{149}$$

and

$$\sum_{e_1=1}^{2d(p-q)}B(e_1)\leq\frac{\varepsilon}{4}\,, \tag{150}$$

then the probability $\mathbb{P}[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^i,E_2^i)]$ is strictly less than $\varepsilon\in(0,1)$. To ensure that (149) is satisfied, we obtain

$$\begin{aligned}
\sum_{e_{\mathsf{s}}=1}^{qd}A(e_{\mathsf{s}})\leq \max_{e_{\mathsf{s}}\in\{1,\ldots,qd\}}\exp\Bigg(&\log qd+\log\binom{\binom{q}{2}}{e_{\mathsf{s}}}\\
&-n\frac{e_{\mathsf{s}}/(2d)}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\Bigg)\,, \qquad (151)
\end{aligned}$$

which is less than $\varepsilon/4$ if

$$n\geq\frac{2d(3\exp(2\vartheta)+1)}{\sinh^2(\lambda/4)}\left(2\log q+\log qd+\log\frac{1}{\varepsilon}\right). \tag{152}$$

Equation (152) provides one half of the sufficient conditions in Theorem 3. To ensure that (150) is satisfied, we obtain

$$\begin{aligned}
\sum_{e_1=1}^{2d(p-q)}B(e_1)\leq \max_{e_1\in\{1,\ldots,2(p-q)d\}}\Bigg\{\exp\Bigg(&\log 2d(p-q)\\
&+\log\binom{\binom{p-q}{2}\binom{p-q}{1}\binom{q}{1}}{e_1}\\
&-n\frac{e_1/(4d)}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\Bigg)\Bigg\}\,, \qquad (153)
\end{aligned}$$

which is less than $\varepsilon/4$ if

$$\begin{aligned}
n>\frac{4d(3\exp(2\vartheta)+1)}{\sinh^2(\lambda/4)}\Big(&\log 8d+\log(p-q)\\
&+\log\left(\binom{p-q}{2}+(p-q)q\right)+\log\frac{1}{\varepsilon}\Big)\,. \qquad (154)
\end{aligned}$$

We remark that (152) and (154) place emphasis on the sample complexity driven by the shared cluster and non-shared clusters, respectively. Furthermore, the sufficient condition on the sample complexity in (152) dominates that in (154) when we have

$$\frac{q^3d}{2}\geq\Big(4d(p-q)(p^2-q^2)\Big)^2\,, \tag{155}$$

which simplifies to

$$\frac{32}{d}\geq(1+\upsilon)^2(\upsilon-1)^4\,, \tag{156}$$

where $\upsilon=\frac{p}{\sqrt{q}}$. This observation is formalized as follows. The ML decoder achieves $\mathbb{P}(\mathcal{S}_q)\leq\varepsilon$, if we have

1) if $\frac{32}{d}\geq(1+\upsilon)^2(\upsilon-1)^4$,

$$n\geq\frac{2d(3\exp(2\vartheta)+1)}{\sinh^2(\lambda/4)}\left(2\log q+\log qd+\log\frac{1}{\varepsilon}\right), \tag{157}$$

2) otherwise,

$$\begin{aligned}
n\geq\frac{4d(3\exp(2\vartheta)+1)}{\sinh^2(\lambda/4)}\Big(&\log 8d+\log\frac{p-q}{\varepsilon}\\
&+\log\left(\binom{p-q}{2}+(p-q)q\right)\Big)\,. \qquad (158)
\end{aligned}$$

2) Proof of Theorem 4: Consider the models i and j in the class $\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)$ such that the non-shared parts of the graphs $\mathcal{G}_1^i$ and $\mathcal{G}_1^j$ have an edit distance e_1, that of $\mathcal{G}_2^i$ and $\mathcal{G}_2^j$ have an edit distance e_2, and the shared part of the two models have an edit distance e_{s}. Therefore, $e_u\in\{0,\ldots,2\bar{\gamma}k\}$, for $u\in\{1,2\}$, and $e_{\mathsf{s}}\in\{0,\ldots,2\gamma k\}$. Without loss of generality, we assume $\zeta=i$ to be the true model. By using notion of vertex cover, Next, we provide an upper bound on the total number of models in $\mathcal{I}_{q,\gamma}^k(\lambda,\vartheta)$ that satisfies $|E_1^i\Delta E_1^j|=e_1+e_{\mathsf{s}}$ and $|E_2^i\Delta E_2^j|=e_2+e_{\mathsf{s}}$. Note that the vertex cover of a set of edges specifies a set of nodes such that each edge is incident on at least one node in the vertex cover. Furthermore, the nodes spanned by the maximal matching of a given graph also form its vertex cover. Using the upper bound on the number of graph models with a given edit distance in [18, Section V-D], we conclude that there are at most $2^{2\bar{\gamma}k}p^{2(e_1+e_2)(\bar{\gamma}k+1)}\times 2^{\gamma k}q^{2e_{\mathsf{s}}(\gamma k+1)}$ number of q-similar graph

pairs that differ in e_1 edges in the non-shared part of $\mathcal{G}_1^i$, e_2 edges in the non-shared part of $\mathcal{G}_2^j$, and e_{s} edges in the shared part of model i. Using (127), the large deviations bound in Lemma 7, and (142) we obtain

$$\begin{aligned}&\mathbb{P}[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^i,E_2^i)]\\&\leq\sum_{e_1=0}^{2\bar{\gamma}k}\sum_{e_2=0}^{2\bar{\gamma}k}\sum_{e_{\mathsf{s}}=0}^{2\gamma k}2^{2\bar{\gamma}k+\gamma k}\Big(p^{2(e_1+e_2)(\bar{\gamma}k+1)}\\&\quad\times q^{2e_{\mathsf{s}}(\gamma k+1)}\times\exp\left(-n\frac{e_1+e_2+2e_{\mathsf{s}}}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\right)-1\Big)\qquad(159)\\&=2^{2\bar{\gamma}k+\gamma k}\left(\left(1+\sum_{e_1=1}^{2\bar{\gamma}k}C(e_1)\right)^2\times\left(1+\sum_{e_{\mathsf{s}}=1}^{2\gamma k}D(e_{\mathsf{s}})\right)-1\right),\qquad(160)\end{aligned}$$

where

$$C(e_1)\triangleq p^{2e_1(\bar{\gamma}k+1)}\exp\Big(-n\frac{e_1}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\Big),\qquad(161)$$

and

$$D(e_{\mathsf{s}})\triangleq q^{2e_{\mathsf{s}}(\gamma k+1)}\exp\Big(-n\frac{2e_{\mathsf{s}}}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\Big).\qquad(162)$$

We simplify (160) to

$$\begin{aligned}&\mathbb{P}[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^i,E_2^i)]\\&\leq 2^{2\bar{\gamma}k+\gamma k}\left(\left(\sum_{e_1=1}^{2\bar{\gamma}k}C(e_1)\right)^2+2\sum_{e_1=1}^{2\bar{\gamma}k}C(e_1)+\sum_{e_{\mathsf{s}}=1}^{2\gamma k}D(e_{\mathsf{s}})\right.\\&\quad\left.+\left(\sum_{e_1=1}^{2\bar{\gamma}k}C(e_1)\right)^2\sum_{e_{\mathsf{s}}=1}^{2\gamma k}D(e_{\mathsf{s}})+2\sum_{e_{\mathsf{s}}=1}^{2\gamma k}D(e_{\mathsf{s}})\sum_{e_1=1}^{2\bar{\gamma}k}C(e_1)\right).\qquad(163)\end{aligned}$$

If we have

$$2^{2\bar{\gamma}k+\gamma k}\sum_{e_1=1}^{2\bar{\gamma}k}C(e_1)\leq\frac{\varepsilon}{4}\quad\text{and}\quad 2^{2\bar{\gamma}k+\gamma k}\sum_{e_{\mathsf{s}}=1}^{2\gamma k}D(e_{\mathsf{s}})\leq\frac{\varepsilon}{4},\qquad(164)$$

then the probability $\mathbb{P}[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^i,E_2^i)]$ is strictly less than $\varepsilon\in(0,1)$. To ensure that the first part of (164) is satisfied, we obtain

$$\begin{aligned}&2^{2\bar{\gamma}k+\gamma k}\sum_{e_1=1}^{2\bar{\gamma}k}C(e_1)\\&\leq\max_{e_1\in\{1,\ldots,2\bar{\gamma}k\}}\Big\{\exp\Big((2\bar{\gamma}k+\gamma k)+\log(2\bar{\gamma}k)\\&\qquad+2e_1(\bar{\gamma}k+1)\log p\\&\qquad-n\frac{e_1}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\Big)\Big\},\qquad(165)\end{aligned}$$

which is less than $\varepsilon/4$ if

$$n\geq\frac{3\exp(2\vartheta)+1}{\sinh^2(\lambda/4)}\Big((2\bar{\gamma}k+\gamma k)+\log 8\bar{\gamma}k+2(\bar{\gamma}k+1)\log p+\log\frac{1}{\varepsilon}\Big).\qquad(166)$$

To ensure that second part of (164) is satisfied, we obtain

$$\begin{aligned}&2^{2\bar{\gamma}k+\gamma k}\sum_{e_{\mathsf{s}}=1}^{2\gamma k}D(e_{\mathsf{s}})\\&\leq\max_{e_{\mathsf{s}}\in\{1,\ldots,\gamma k\}}\Big\{\exp\Big((2\bar{\gamma}k+\gamma k)+\log(2\lfloor\gamma k\rfloor)\\&\qquad+2e_{\mathsf{s}}(\gamma k+1)\log q\\&\qquad-n\frac{2e_{\mathsf{s}}}{3\exp(2\vartheta)+1}\sinh^2\left(\frac{\lambda}{4}\right)\Big)\Big\},\qquad(167)\end{aligned}$$

which is less than $\varepsilon/4$ if

$$n\geq\frac{3\exp(2\vartheta)+1}{2\sinh^2(\lambda/4)}\Big((2\bar{\gamma}k+\gamma k)+\log(8\gamma k)+2(\gamma k+1)\log q+\log\frac{1}{\varepsilon}\Big).\qquad(168)$$

We remark that (166) and (168) place emphasis on the sample complexity due to the shared and non-shared parts, respectively. This is noted by the fact that the bound in (168) dominates that in (166) if we have $\log q\geq\frac{2(\bar{\gamma}k+1)}{\gamma k+1}\log p$, which is feasible only if we have $2(\bar{\gamma}k+1)\leq\gamma k+1$. Furthermore, when the asymptotic scaling behavior for large graphs is in focus, we can further simplify (166) and (168) to include only the terms that dominate the sample complexity. Specifically, in (166), for sufficiently large p, the term $2(\bar{\gamma}k+1)\log p$ dominates the terms $(2\bar{\gamma}k+\gamma k)$ and $\log 8\bar{\gamma}k$, and specifies the asymptotic scaling behavior of the sufficient condition. To emphasize upon this behavior in the results, we relax the bound in (166) to

$$n\geq\frac{3\exp(2\vartheta)+1}{\sinh^2(\lambda/4)}\Big(6(\bar{\gamma}k+1)\log p+\log\frac{1}{\varepsilon}\Big).\qquad(169)$$

Similarly, we note that in (168), for sufficiently large q, the term $2(\gamma k+1)\log q$ dominates the term $(2\bar{\gamma}k+\gamma k+\log 8\gamma k)$ and characterizes the asymptotic scaling behavior. Therefore, to lay emphasis upon this behavior in the results, we modify the bound in (168) to

$$n\geq\frac{3\exp(2\vartheta)+1}{2\sinh^2(\lambda/4)}\Big(6(\gamma k+1)\log q+\log\frac{1}{\varepsilon}\Big).\qquad(170)$$

The results in (169) and (170) complete the sufficient conditions in Theorem 4.

B. Gaussian Models

To establish the sufficient conditions for the subclasses of Gaussian models, we first establish the large deviations bound on the ML decoder for recovering a single graph in Lemma 1 and generalize it to the recovery of q-similar graph pairs. We first restate Lemma 1 below.

Lemma 8. *Consider a Gaussian graphical model $\mathcal{G}$ in the set $\mathcal{S}$. Then, for any $u,v\in\mathcal{S}$, we have*

$$\mathbb{P}[\Lambda_u(\mathbf{x})\geq\Lambda_v(\mathbf{x})]\leq K(\mathbf{P}[u],\mathbf{P}[v]),\qquad(171)$$

where we have defined

$$K(\mathbf{P}([u],\mathbf{P}[v])\triangleq\left(\frac{\det[\mathbf{P}[u]]\cdot\det[\mathbf{P}[v]]}{\det^2\left[\frac{\mathbf{P}[u]+\mathbf{P}[v]}{2}\right]}\right)^{\frac{1}{4}}.\qquad(172)$$

Proof. Using Chernoff's bound we have

$$\mathbb{P}[\Lambda_u(\mathbf{x}) - \Lambda_v(\mathbf{x}) \geq 0] \leq \inf_{s>0} \mathbb{E}_v[\exp(sQ)] \,, \tag{173}$$

where the expectation is with respect to model v and we have defined $Q \triangleq \Lambda_u(\mathbf{x}) - \Lambda_v(\mathbf{x})$, where $\Lambda_u(\mathbf{x})$ is defined (46). We define f^u as the Gaussian pdf of $\mathbf{x}$ under model u that is characterized by the inverse covariance matrix $\mathbf{P}[u]$, for $u \in \{1, \ldots, |\mathcal{S}|\}$. Furthermore, we have

$$\mathbb{E}_i[\exp(sQ)] = \int [f^u(\mathbf{x})]^s [f^v(\mathbf{x})]^{1-s} \mathrm{d}\mathbf{x} \,. \tag{174}$$

By setting $s = 1/2$ in (174), we get

$$\begin{aligned} &\int [f^u(\mathbf{x})]^{1/2} [f^v(\mathbf{x})]^{1/2} \mathrm{d}\mathbf{x} \\ &= \int \frac{1}{\sqrt{(2\pi)^p \sqrt{\mathsf{det}(\mathbf{P}^{-1}[u])\mathsf{det}(\mathbf{P}^{-1}[v])}}} \\ &\qquad \times \exp\left(\frac{1}{2}\mathbf{x}^{\mathrm{T}} \left(\frac{\mathbf{P}[u] + \mathbf{P}[v]}{2}\right)\mathbf{x}\right) \mathrm{d}\mathbf{x} \qquad (175) \\ &= \frac{\left(\mathsf{det}\left[\left(\frac{\mathbf{P}[u]+\mathbf{P}[v]}{2}\right)^{-1}\right]\right)^{1/2}}{\left(\mathsf{det}(\mathbf{P}^{-1}[u])\mathsf{det}(\mathbf{P}^{-1}[v])\right)^{1/4}} \,. \qquad (176) \end{aligned}$$

The statement of Lemma 8 follows from (173) and (176). □

1) Proof of Theorem 9: We leverage the fact that the ensemble $\mathcal{A}_q$ is the generalization of an ensemble for single Gaussian models in [23] to a setting of two q-similar graphs. We first restate the description of the relevant ensemble from [23] here. *Ensemble $\mathcal{A}$*: This ensemble is characterized by a set of single graphs. Any graph $\mathcal{G} = (V, E)$ in $\mathcal{A}$ consists of one isolated edge between a pair of nodes given by $U_2 \subset V$ and a clique of d nodes in the set $U_d \subset V$ such that $U_d \cap U_2 = \phi$. The set U_d is fixed and known to the graph decoder and therefore, the structure estimation problem reduces to estimating the set U_2. The inverse covariance matrix $\mathbf{P}$ for $\mathcal{G}$ is characterized by a parameter $a > 0$ such that

$$\mathbf{P} = \mathbf{I} + a\mathbb{1}_{U_2}\mathbb{1}_{U_2}^{\mathrm{T}} + a\mathbb{1}_{U_d}\mathbb{1}_{U_d}^{\mathrm{T}} \,. \tag{177}$$

Lemma 9 provides the sufficient conditions on the model selection of single graphs in the classes $\mathcal{A}$, which would be instrumental in the proof of the results in Theorem 9.

Lemma 9. *Consider a graph $\mathcal{G}$ in the class $\mathcal{A}$. If the sample size n satisfies*

$$n \geq 2\frac{\log\binom{p-d}{2} + \log\frac{1}{\varepsilon}}{\frac{1}{\log(1-\rho^2)}} \,, \tag{178}$$

then there exists a graph decoder $\Phi : \mathbb{R}^{n\times p} \to \mathcal{A}$ that achieves $\mathsf{P}(\mathcal{A}) \leq \varepsilon$.

Proof. For class $\mathcal{A}$, there are $\binom{p-d}{2}$ number of possible models. Assuming that any of the possible models can be uniformly selected to be the true model, we denote the random variable for selection of the true model by κ, which lies in the set $\{1, \ldots, \binom{p-d}{2}\}$. Furthermore, we denote the graph model associated with $\kappa = i$ by $\mathcal{G}^i \triangleq (V, E^i)$ which has an inverse covariance matrix $\mathbf{P}[i]$. For $\mathcal{G}^i$, we denote the pair of nodes connected by the isolated edge by U_2^i. Without loss of generality, we assume that $\mathcal{G}^u$ is the true model, for $u \in \{1, \ldots, \binom{p-d}{2}\}$. Using Lemma 8 and the union bound, we have

$$\mathbb{P}[\Phi(\mathbf{x}^n) \neq E^u)] \leq \sum_{v \in \{1,\ldots,\binom{p}{2}\}\setminus u} \mathbb{P}[\Lambda_v(\mathbf{x}^n) \geq \Lambda_u(\mathbf{x}^n)] \,. \tag{179}$$

Since U_d is fixed and known, and $U_2^i \cap U_d = \phi$, the graph decoder can estimate the unknown structure by samples collected only from the nodes $V\setminus U_d$. In this scenario, the reduced inverse covariance matrix formed by the nodes in the set $V\setminus U_d$ for model i is given by

$$\tilde{\mathbf{P}}[i] \triangleq \mathbf{I} + a\mathbb{1}_{U_2^i}\mathbb{1}_{U_2^i}^{\mathrm{T}} \,. \tag{180}$$

It follows that we have $\mathsf{det}([\tilde{\mathbf{P}}^{-1}[i]]) = \frac{1}{1+2a}, \forall i \in \{1, \ldots, \binom{p-d}{2}\}$ and

$$\mathsf{det}\left[\left(\frac{\tilde{\mathbf{P}}[u] + \tilde{\mathbf{P}}[v]}{2}\right)^{-1}\right] \leq \left(\frac{1}{1+a}\right)^2 \,, \tag{181}$$

$\forall v \in \left\{1, \ldots, \binom{p-d}{2}\right\} \setminus u$. Using Lemma 1, (179), and (181), we have

$$\begin{aligned} \mathbb{P}[\Phi(\mathbf{x}^n) \neq E^u)] &\leq \binom{p-d}{2}\left(\frac{1+2a}{(1+a)^2}\right)^{n/2} \\ &\leq \exp\left(\log\binom{p-d}{2} + 0.5n\log(1-\rho^2)\right) \,, \end{aligned} \tag{182}$$

where the second inequality in (182) follows from $\frac{a}{1+a} \geq \rho$. Therefore, the condition
$\mathbb{P}[\Phi(\mathbf{x}^n) \neq E^u)] \leq \varepsilon$ is satisfied if we have

$$n \geq 2\frac{\log\binom{p-d}{2} + \log\frac{1}{\varepsilon}}{\log\frac{1}{1-\rho^2}} \,. \tag{183}$$

□

The proof of Theorem 9 leverages different elements of Lemma 9. The graphs in ensemble $\mathcal{A}_q$ consist of only an isolated edge, either in the shared part or the non-shared part for each graph. Therefore, there are

$$c_1 \triangleq \binom{q}{2} + \binom{p-q}{2}^2 \tag{184}$$

number of possible graph pairs in $\mathcal{A}_q$. Assuming that any of the possible models can be selected uniformly to be the true graph pair, the random variable ζ denotes the selection of the true graph pairs from the set $\{1, \ldots, c_1\}$. When $\zeta = i$, the true graphs are given by $\mathcal{G}_1^i \triangleq (V, E_1^i)$ and $\mathcal{G}_2^i \triangleq (V, E_2^i)$ and the inverse covariance matrix associated with $\mathcal{G}_r^i$ is $\mathbf{P}_r[i]$ for $r \in \{1, 2\}$. Without loss of generality, we assume that $(\mathcal{G}_1^u, \mathcal{G}_2^u)$ is the true graph pair. Using the union bound in (126)

and the technical arguments similar to those in (131)- (136) we have

$$\begin{aligned}\mathbb{P}\left[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^u,E_2^u)\right]&\leq\sum_{j\in\{1,\ldots,c_1\}\setminus u}\mathbb{P}[\ell_j(\mathbf{x}_1^n,\mathbf{x}_2^n)\geq\ell_u(\mathbf{x}_1^n,\mathbf{x}_2^n)]\\&\leq\sum_{j\in\{1,\ldots,c_1\}\setminus u}\left(\int[f_1^j(\mathbf{x}_1)]^{\frac{1}{2}}[f_1^u(\mathbf{x}_1)]^{\frac{1}{2}}\mathrm{d}\mathbf{x}_1\times\int[f_1^j(\mathbf{x}_2)]^{\frac{1}{2}}[f_1^u(\mathbf{x}_2)]^{\frac{1}{2}}\,\mathrm{d}\mathbf{x}_2\right)^n.\end{aligned}\tag{185}$$

Using (176), we obtain

$$\mathbb{P}\left[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^u,E_2^u)\right]\leq\sum_{j\in\{1,\ldots,c_1\}\setminus u}\left(K(\mathbf{P}_1[j],\mathbf{P}_1[u])\times K(\mathbf{P}_2[j],\mathbf{P}_2[u])\right)^n.\tag{186}$$

From the proof of Lemma 8, we leverage the result

$$K(\mathbf{P}_1[j],\mathbf{P}_1[u])\leq\left(1-\rho^2\right)^{\frac{1}{2}},\tag{187}$$

for graph models with a single edge to update (186) to

$$\mathbb{P}\left[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^u,E_2^u)\right]\leq\exp\left(\log c_1+n\log(1-\rho^2)\right).\tag{188}$$

Therefore, the condition $\mathbb{P}\left[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^u,E_2^u)\right]\leq\varepsilon$ is satisfied if we have

$$n\geq\frac{1}{\log\frac{1}{(1-\rho^2)}}\log\frac{c_1}{\varepsilon}.\tag{189}$$

It can be readily verified that

1) if $\binom{q}{2}\geq\binom{p-q}{2}^2$,

$$\binom{q}{2}\leq c_1\;\leq\;2\binom{q}{2}<q^2,\tag{190}$$

2) if $\binom{q}{2}\leq\binom{p-q}{2}^2$,

$$\binom{p-q}{2}^2\leq c_1\;\leq\;2\binom{p-q}{2}^2<(p-q)^4.\tag{191}$$

Therefore, $c_1=\Theta(q^2)$ in the regime in (190) and $c_1=\Theta((p-q)^2)$ in the regime in (191). For clarity in presentation and to place emphasis on the effect of structural similarity on the sample complexity in the two regimes, we relax (189) as

$$n\geq\frac{2}{\log\frac{1}{(1-\rho^2)}}\times\left(\max\left\{\log q,2\log(p-q)\right\}+\log\frac{1}{\varepsilon}\right).\tag{192}$$

2) Proof of Theorem 10: We leverage the fact that ensemble $\mathcal{B}_q$ is the generalization of an ensemble in [23] to a setting consisting of two q-similar graphs. We denote the relevant ensemble in [23] by $\mathcal{B}$ whose description is as follows.

Ensemble $\mathcal{B}$: This ensemble is characterized by a set of single single graphs, where each graph consists of one clique of size d. For a graph $\mathcal{G}$ with a clique formed by nodes in the set U_d, its inverse covariance matrix is given by

$$\mathbf{P}=\mathbf{I}+a\mathbb{1}_{U_d}\mathbb{1}_{U_d}^{\mathrm{T}},\tag{193}$$

for $a>0$. For ensemble $\mathcal{B}$, there are $\binom{p}{d}$ total number of possible models. Lemma 10 provides the sufficient conditions on the model selection of single graphs in the classes $\mathcal{B}$, which would be instrumental in the proof of the results in Theorem 10.

Lemma 10. *Consider a graph $\mathcal{G}_1$ in the class $\mathcal{B}$. If the sample size n satisfies*

$$n\geq2\frac{\log\binom{p}{d}+\log\frac{1}{\varepsilon}}{\log\frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}},\tag{194}$$

then there exists a graph decoder $\Phi:\mathbb{R}^{n\times p}\to\mathcal{B}$ that achieves $\mathsf{P}(\mathcal{B})\leq\varepsilon$.

Proof. Assuming that any model in $\mathcal{B}$ can be selected uniformly to be the true model, we denote the random variable for selection of the true model by ζ whose support lies in the set $\{1,\ldots,\binom{p}{d}\}$. We use the same definitions as in the proof of Lemma 8 for $\mathcal{G}^i$ and its inverse covariance matrix $\mathbf{P}[i]$ when $\zeta=i$. For graph $\mathcal{G}^i$, we denote the set of d nodes that form the clique by U_d^i and, therefore, we have

$$\mathsf{det}([\mathbf{P}[i]]^{-1})=\frac{1}{1+da},\tag{195}$$

and

$$\mathsf{det}\left[\left(\frac{\mathbf{P}[i]+\mathbf{P}[j]}{2}\right)^{-1}\right]\leq\left(\frac{1}{1+da/2}\right)^2,\tag{196}$$

$\forall j\in\{1,\ldots,\binom{p}{d}\}\setminus u$. Using Lemma 8, (195) and (196), we have

$$\begin{aligned}\mathbb{P}[\Phi(\mathbf{x}^n)\neq E^u)]&\leq\binom{p}{d}\left(\frac{1+da}{(1+da/2)^2}\right)^{n/2}&(197)\\&\leq\exp(\log\binom{p}{d}-0.5n\log\frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}),&(198)\end{aligned}$$

where (198) follows from (197) by using $\frac{a}{1+a}\geq\rho$. The condition $\mathbb{P}[\Phi(\mathbf{x}^n)\neq E^u)]\leq\varepsilon$ is satisfied if we have

$$n\geq2\frac{\log\binom{p}{d}+\log\frac{1}{\varepsilon}}{\log\frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}.\tag{199}$$

□

The proof of Theorem 10 leverages Lemma 10. The total number of graph pairs in $\mathcal{B}_q$ is given by

$$c_2\;\triangleq\;\binom{q}{d}+\binom{p-q}{d}^2.\tag{200}$$

Therefore, under the assumption that any of the possible graph pairs can be selected as the true models uniformly at random and using similar arguments as in Section VI-B1, we obtain

$$\mathbb{P}\left[\Psi(\mathbf{x}_1^n,\mathbf{x}_2^n)\neq(E_1^u,E_2^u)\right]\leq\sum_{j\in\{1,\ldots,c_2\}\setminus u}\left(K(\mathbf{P}_1[j],\mathbf{P}_1[u])\times K(\mathbf{P}_2[j],\mathbf{P}_2[u])\right)^n.\tag{201}$$

Furthermore, by leveraging the following result for single graphs from $\mathcal{B}$

$$K(\mathbf{P}_r[j], \mathbf{P}_r[u]) \leq \left(\frac{1+da}{(1+da/2)^2}\right)^{\frac{1}{2}}, \tag{202}$$

$$\leq \frac{(1-\rho)(1+(d-1)\rho}{(d^2-d+1)\rho^2+(d-2)\rho+1}, \tag{203}$$

we update (201) to

$$\mathbb{P}\left[\Psi(\mathbf{x}_1^n, \mathbf{x}_2^n) \neq (E_1^u, E_2^u)\right] \leq \exp\left(\log c_2 - n \log \frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}\right). \tag{204}$$

Therefore, the condition $\mathbb{P}\left[\Psi(\mathbf{x}_1^n, \mathbf{x}_2^n) \neq (E_1^u, E_2^u)\right] \leq \varepsilon$ is satisfied if we have

$$n \geq \frac{\log c_2/\varepsilon}{\log \frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}. \tag{205}$$

In the context of c_2, it readily follows that

1) if $\binom{q}{d} \geq \binom{p-q}{d}^2$,

$$\binom{q}{d} \leq c_2 \leq 2\binom{q}{d} < 2\left(\frac{q\mathrm{e}}{d}\right)^d, \tag{206}$$

2) if $\binom{q}{d} \leq \binom{p-q}{d}^2$,

$$\binom{p-q}{d}^2 \leq c_2 \leq 2\binom{p-q}{d}^2 < 2\left(\frac{(p-q)\mathrm{e}}{d}\right)^{2d}, \tag{207}$$

where the upper bounds in (206) and (207) follow from the inequality $\binom{z}{y} < (z\mathrm{e}/y)^y$, for any pair of positive integers $z > y$. Clearly, we have $c_2 = \Theta(\binom{q}{d})$ in the regime in (206) and $c_2 = \Theta(\binom{p-q}{d})$ in the regime in (207). Therefore, for clarity and to emphasize on the effect of structural similarity on the sample complexity, we relax the bound in (205) to

$$n \geq \frac{1}{\log \frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}} \times \left(\max\left\{2d \log \frac{(p-q)\mathrm{e}}{d}, d \log \frac{q\mathrm{e}}{d}\right\} + \log \frac{2}{\varepsilon}\right). \tag{208}$$

3) Proof of Theorem 11: The proof of Theorem 11 follows the same line of analysis as that of Theorem 10. Firstly, we create an ensemble of single graphs $\mathcal{C}$ which consists of graphs with cliques of size $\tilde{k}$. Note that ensemble $\mathcal{C}$ is similar to the design of $\mathcal{B}$ with the size of the cliques being the only difference between the two ensembles. The ensemble $\mathcal{C}_q$ is a generalization of $\mathcal{C}$ to the setting with q-similar graph pairs and consists of

$$c_3 \triangleq \binom{q}{\tilde{k}} + \binom{p-q}{\tilde{k}}^2, \tag{209}$$

total number of graph pairs. Due to the equivalence in the design of ensembles, we can leverage the results of Lemma 10 and follow the same line of analysis to recover the result that the condition $\mathbb{P}\left[\Psi(\mathbf{x}_1^n, \mathbf{x}_2^n) \neq (E_1^u, E_2^u)\right] \leq \varepsilon$ is satisfied for the ensemble $\mathcal{C}_q$ if we have

$$n \geq \frac{\log c_3/\varepsilon}{\log \frac{(\tilde{k}^2-\tilde{k}+1)\rho^2+(\tilde{k}-2)\rho+1}{(1-\rho)(1+(\tilde{k}-1)\rho)}}. \tag{210}$$

Next, we relax the bound in (210) using similar technical arguments followed in (206) and (207) to obtain the condition

$$n \geq \frac{1}{\log \frac{(\tilde{k}^2-\tilde{k}+1)\rho^2+(\tilde{k}-2)\rho+1}{(1-\rho)(1+(\tilde{k}-1)\rho)}} \times \left(\max\left\{2\tilde{k} \log \frac{(p-q)\mathrm{e}}{\tilde{k}}, \tilde{k} \log \frac{q\mathrm{e}}{\tilde{k}}\right\} + \log \frac{2}{\varepsilon}\right). \tag{211}$$

VII. Numerical Evaluations

In this section, we illustrate the effect of $\eta = q/p$, which quantifies the structural similarity, on the performance of an ML based graph decoder and Algorithm 1.

A. Joint Structure Estimation via ML Decoding

In general, for any class $\mathcal{S}_q$, the graph decoder that minimizes $\mathsf{P}(\mathcal{S}_q)$ is the ML decoder given in (19). Since the implementation of (19) requires a search over all the possible graph pairs in a class, it becomes computationally intractable as the graph size p increases. Therefore, we evaluate this graph decoder over a restricted ensemble of Ising models for which the implementation is feasible.

We consider an ensemble that is characterized by many isolated edges. We assume that for a graph with p total nodes out of which q nodes lie in the shared cluster such that $\eta = \frac{q}{p}$, there are α isolated edges with $\lfloor \eta\alpha \rfloor$ edges in the shared cluster. Each graph is constructed in the following manner. We randomly divide the non-shared cluster with $p-q$ nodes in $(p-q)/2$ pair of nodes and randomly connect $\alpha - \lfloor \eta\alpha \rfloor$ pairs. The edge structure in the shared cluster is constructed in a similar manner.

Under joint recovery, the data from both graphs are processed jointly to estimate the edge structure. Under independent recovery, the structures of the graphs are learned independently. Figure 3 illustrates the effect of structural similarity on the performance of the graph decoder. Clearly, as η increases, the graph decoder that jointly processes the data requires a smaller number of samples to achieve the same performance as a graph decoder that learns the graph structures independently. For the results in Fig. 3, we set $p = 100$, $\alpha = 20$, and $\lambda = 1$. The performance of the graph decoders is evaluated over 1500 trials. Next, we keep the number of edges k and degree d fixed as we evaluate the error probability for increasing the number of nodes p. For the results in Fig. 4, we set $\alpha = 20$, $\lambda = 0.4$ and $\eta = 0.5$, and evaluate the error probability for the graph decoder based on 40 samples from each graph. We observe that the error probability monotonically increases as the graph size increases, indicating that the structure estimation problem becomes more difficult, as implied by Corollary 1.

This article has been accepted for publication in IEEE Transactions on Information Theory. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3207420

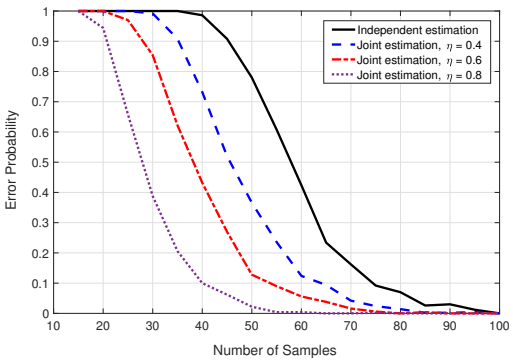

Fig. 3. Joint recovery versus independent recovery for varying η

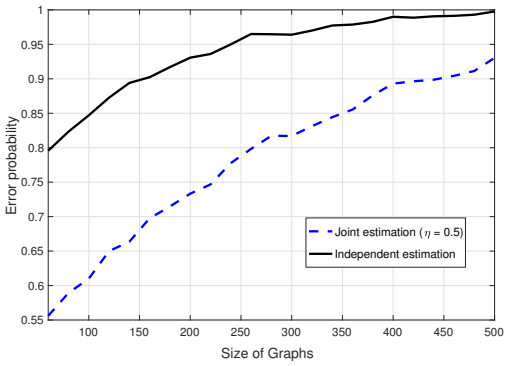

Fig. 4. Error probability versus size of graphs p after 40 samples from each graph

B. Joint Structure Estimation using Algorithm 1

We study the performance of structure estimation of graphs with loops which are, in general, infeasible to be learned by an ML decoder. For this purpose, we generate an ensemble of graphs of size $p = 20$, where the nodes are randomly divided into groups of size 4 and each group is connected in a ring, followed by random single-edge connections among different groups, and each node in a group connected to at most one other node outside its own group. Therefore, the maximum degree of a node in this ensemble is 3.

Figure 5 illustrates the comparison of the mean performance of Algorithm 1 for recovering graph pairs with different structural similarities against recovering them independently using the algorithm in [15] over 1000 random instances of graph pairs. The probability of error corresponds to the event that the true graph pair was not recovered exactly in any of the iterations when the online estimation algorithm was run up to a horizon indicated on the x-axis.

Clearly, our algorithm outperforms the independent structure estimation algorithm for $\eta = 0.25, 0.5$ and 1. When $\eta = 1$, the graph pairs are identical and, therefore, Algorithm 1 is equivalent to processing the data $\mathbf{x}_1^n$ and $\mathbf{x}_2^n$ in parallel with two processing units processing one graph sample each in every iteration with an exchange of pairwise loss functions between the two. This indicates that Algorithm 1 performs better by processing two graph samples in every iteration up to a horizon n_{T} compared to an approach that sequentially processes one graph sample up to a horizon n_{T}.

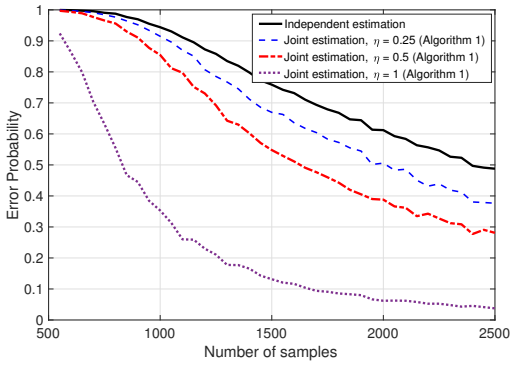

Fig. 5. Error probability versus horizon (n_{T}) or the number of samples for each graph.

VIII. Conclusions

In this paper, we have considered the problem of structure estimation of partially similar graphs in various subclasses of Ising models and Gaussian models. Due to the partial similarity in structure, any inference about the structure of one graph provides side information about the structure of the other graph. Under the criterion of exact recovery of the structure of the graphs, we have characterized necessary and sufficient conditions on the sample complexity of joint model selection for various subclasses of Ising models and Gaussian models. The sufficient conditions are based on the analysis of an ML decoder which is optimal for exact recovery. We have analyzed variation in sample complexity with respect to structural similarity. We have also studied the scaling behavior of the sample complexity in different regimes. Our analysis has also revealed the regimes in which the asymptotic scaling behavior of the necessary and sufficient conditions coincide, thus establishing optimal sample complexity. Moreover, for different subclasses of Gaussian models, our theoretical results enable us to conclusively establish that jointly recovering q-similar graphs is easier than recovering the graphs independently.

Appendix A
Proof of Theorem 6

We start by noting that the Sparsitron algorithm proposed in [15] for estimating a sparse generalized linear model (GLM) was shown to enable structure estimation of a single Ising model due to certain properties of the random variables associated with a degree-bounded Ising model. Here, we will build upon the principles adopted in [15] to first propose Algorithm 2 for estimating two sparse GLMs jointly and characterize its performance. Then, we will leverage the performance of Algorithm 2 and the properties of Ising models to complete the proof of Theorem 6.

Algorithm 2 Estimating two GLMs jointly

Input β, $\mathcal{R}$, data samples $(\mathbf{c}_1^T, \mathbf{d}_1^T)$ and $(\mathbf{c}_2^T, \mathbf{d}_2^T)$
initialize $w_i^0 = \mathbf{1}_a/a$ for $i \in \{1,2\}$
for a new pair of data sample $j \in \{1,\dots,T\}$ **do**
  Compute $\mathbf{h}_i^j$ using (216)
  Compute losses $\boldsymbol{\ell}_i^j$ for $i \in \{1,2\}$ according to (215)
  for $t \in \{1,\dots,a\}$ **do**
    Update the weights $w_i^j(t)$ according to (213)
  end for
end for

A. Joint Estimation of Sparse GLMs

Define g_1 and g_2 as two pdfs in the space $[-1,1]^a \times \{0,1\}$ and (C_i, D_i) as the random variables whose joint pdf is g_i, i.e., $(C_i, D_i) \sim g_i$, where $C_i \in [-1,1]^a$ and $D_i \in \{0,1\}$. We assume that the pair C_i and D_i satisfies the property

$$\mathbb{E}[D_i|C_i] = \tilde{\sigma}(\mathbf{r}_i \cdot C_i) \ , \ \text{for } i \in \{1,2\} \ , \tag{212}$$

where $\tilde{\sigma} : \mathbb{R} \to [0,1]$ is a non-decreasing 1-Lipschitz function, and $\mathbf{r}_i \triangleq [r_i^1, \dots, r_i^a]$ is a vector of weights such that $\|\mathbf{r}_i\|_1 \leq \vartheta$ for $i \in \{1,2\}$ for $\vartheta > 0$. Furthermore, we assume that the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ are partially similar, i.e., $r_1^i = r_2^i$, $\forall i \in \mathcal{R}$, where $\mathcal{R} \subseteq \{1,\dots,a\}$ is the set of indices at which the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ have identical entries. In this scenario, the objective is to learn the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ from the random samples from g_1 and g_2. The collection of T independent and identically distributed (i.i.d.) samples from g_i is denoted by $(\mathbf{c}_i^T, \mathbf{d}_i^T)$, where $\mathbf{c}_i^T \triangleq [c_i^1, \dots, c_i^T]$ and $\mathbf{d}_i^T \triangleq [d_i^1, \dots, d_i^T]$. Furthermore, corresponding to the model g_i, we denote the weight associated with the t-th element in $\mathbf{r}_i$ in the j-th iteration by $w_i^t(j)$ and its update rule is given by

$$w_i^j(t) = w_i^0(t) \prod_{u=1}^{j} \exp(\beta L_i^u(t)) \ , \tag{213}$$

where

$$L_i^u(t) \triangleq \mathbb{1}_{t\in\mathcal{R}} \frac{(\ell_1^u(t) + \ell_2^u(t))}{2} + (1 - \mathbb{1}_{t\in\mathcal{R}})\ell_i^u(t) \ , \tag{214}$$

and $\mathbb{1}_{\{.\}}$ is an indicator function and the structure of the loss function $\ell_1^u(t)$ is discussed next. We denote the local loss function for model g_i evaluated at the $j-$th iteration by $\boldsymbol{\ell}_i^j \triangleq [\ell_i^j(1), \dots, \ell_i^j(a)]$, where $\ell_i^j(t)$ is defined as

$$\boldsymbol{\ell}_i^j \triangleq \frac{1}{2}(\mathbf{1}_a + (\tilde{\sigma}(\vartheta \mathbf{h}_i^j \cdot c_i^j) - d_i^j)c_i^j) \ , \tag{215}$$

$\mathbf{1}_a$ is an $a \times 1$ vector of all $1'$s and $\mathbf{h}_i^j$ is obtained by normalizing the vector $\mathbf{w}_i^j \triangleq [w_i^j(1), \dots, w_i^j(a)]$ using

$$\mathbf{h}_i^j = \frac{\mathbf{w}_i^{j-1}}{\|\mathbf{w}_i^{j-1}\|_1} \ . \tag{216}$$

In this scenario, the steps to jointly learn $\mathbf{r}_1$ and $\mathbf{r}_2$ are provided in Algorithm 2 which builds upon the principles of Hedge algorithm in [42]. Theorem 14 provides the sample complexity of Algorithm 2.

Theorem 14. *Given $T = O(\vartheta^2(\log(a/\delta\varepsilon)/\varepsilon^2)$ number of i.i.d. samples from g_1 and g_2, Algorithm 2 forms estimates $\hat{r}_1$ and $\hat{r}_2$ such that with probability at least $1-\delta$, we have*

$$\mathbb{E}_{g_1,g_2}[(\tilde{\sigma}(\hat{r}_i \cdot \mathbf{c}_i) - \tilde{\sigma}(r_i \cdot \mathbf{c}_i))^2] \leq \varepsilon \ , \ \textit{for } i \in \{1,2\}. \tag{217}$$

Proof. We start by presenting a result similar to [42, Theorem 5], which establishes that the overall regret of an online estimation framework given by Algorithm 2 is upper bounded by the regret of the best expert with addition of terms that scale as $O(\sqrt{T\log a}) + \log a$. This result is formalized in the next lemma.

Lemma 11. *Given T data samples, the overall regret corresponding to estimating the GLM for g_i in Algorithm 2 is bounded as*

$$\sum_{j=1}^{T} \mathbf{h}_i^j \cdot \mathbf{L}_i^j \leq \min_{t\in\{1,\dots,a\}} \sum_{j=1}^{T} L_i^j(t) + O(\sqrt{T\log a}) + \log a \ , \tag{218}$$

where

$$\mathbf{L}_i^j \triangleq [L_i^j(1), \dots, L_i^j(a)]^{\mathsf{T}} \quad \textit{and} \quad \mathbf{h}_i^j \triangleq [h_i^j(1), \dots, h_i^j(a)] \ , \tag{219}$$

such that $\|\mathbf{h}_i^j\|_1 = 1$ and $h_i^j(t) \geq 0, \forall t \in \{1,\dots,a\}$.

Proof. We note that

$$\sum_{t=1}^{a} w_i^j(t) = \sum_{t=1}^{a} w_i^{j-1}(t) \exp(\beta L_i^j(t)) \ . \tag{220}$$

Since, we have $L_i^j(t) \in [0,1]$, and from the convexity argument in [42], we get

$$\exp(\beta L_i^j(t)) \leq 1 - (1 - \exp(\beta))L_i^j(t) \ . \tag{221}$$

Therefore, it readily follows that

$$\sum_{t=1}^{a} w_i^j(t) \leq \sum_{t=1}^{a} w_i^{j-1}(t)(1 - (1 - \exp(\beta))\mathbf{h}_i^j \cdot \mathbf{L}_i^j) \ . \tag{222}$$

For $j = T$ and by repeating the steps (220) and (222), we have

$$\sum_{t=1}^{a} w_i^T(t) \leq \sum_{t=1}^{a} w_i^0(t) \prod_{j=1}^{T} (1 - (1 - \exp(\beta))\mathbf{h}_i^j \cdot \mathbf{L}_i^j) \ . \tag{223}$$

By using $\sum_{t=1}^{a} w_i^0(t) = 1$ and the property $1 + x \leq \exp(x), \forall x$, we get

$$\sum_{t=1}^{a} w_i^T(t) \leq \exp(-(1 - \exp(\beta)) \sum_{j=1}^{T} \mathbf{h}_i^j \cdot \mathbf{L}_i^j) \ . \tag{224}$$

The overall regret of the Algorithm 2 is given by $\sum_{j=1}^{T} \mathbf{h}_i^j \cdot \mathbf{L}_i^j$ and from (224), we have

$$\sum_{j=1}^{T} \mathbf{h}_i^j \cdot \mathbf{L}_i^j \leq \frac{-\log(\sum_{t=1}^{a} w_i^T(t))}{1 - \exp(\beta)} \ . \tag{225}$$

Therefore, we have established that the sequence of loss functions for joint estimation of the two GLMs satisfies the same property as the loss function for estimating a single GLM in [15]. Subsequent arguments in Lemma 4 and Lemma 5 in [42] complete the proof of Lemma 11. □

The proof of Theorem 14 leverages Lemma 11 and the subsequent analysis follows the same line of analysis as in the proof of [15, Theorem 3.1]. We will leverage Lemma 11 to characterize T for prediction of $\mathbf{r}_i$ next. Corresponding to g_i, we define the random variable

$$V_i^j \triangleq (\mathbf{h}_i^j - \mathbf{r}_i/\vartheta) \cdot \mathbf{L}_i^j \ , \tag{226}$$

such that, $V_i^j \in [-1, 1]$. Based on V_i^j, we define another sequence of random variables

$$Z_i^j = V_i^j - \mathbb{E}[V_i^j | (\mathbf{c}_1^{j-1}, \mathbf{d}_1^{j-1}), (\mathbf{c}_2^{j-1}, \mathbf{d}_2^{j-1})] \ . \tag{227}$$

Then, we have $Z_i^j \in [-2, 2]$. Note that using Azuma's inequality on martingales with bounded differences, we find that the following event holds with probability at least $1 - \delta$,

$$\sum_{j=1}^{T} \mathbb{E}[V_i^j | ((\mathbf{c}_1^{j-1}, \mathbf{d}_1^{j-1}), (\mathbf{c}_2^{j-1}, \mathbf{d}_2^{j-1})] \leq \sum_{j=1}^{T} V_i^j + O(T \log(1/\delta)) \ . \tag{228}$$

Furthermore, note that

$$\mathbb{E}[V_i^j | ((\mathbf{c}_1^{j-1}, \mathbf{d}_1^{j-1}), (\mathbf{c}_2^{j-1}, \mathbf{d}_2^{j-1})] = \frac{1}{\vartheta} \mathbb{E}[(\vartheta \mathbf{h}_i^j - \mathbf{r}_i) \cdot \mathbf{L}_i^j] \tag{229}$$

and

$$\mathbb{E}[V_i^j] \geq \frac{1}{4\vartheta} \mathbb{E}[\tilde{\sigma}(\vartheta \mathbf{h}_i^j \cdot \mathbf{c}_i) - \tilde{\sigma}(\mathbf{r}_i \cdot \mathbf{c}_i))^2] \tag{230}$$

where (230) follows from the inequality that $\forall a, b \in \mathbb{R}, (a - b)(\sigma(a) - \sigma(b)) \geq (\sigma(a) - \sigma(b))^2$ and that the lower bound corresponds to indices with identical values in $\mathbf{r}_1$ and $\mathbf{r}_2$. Then, it follows from (219), (228), and (230) that with probability at least $1 - \delta$, we have

$$\frac{1}{4\vartheta} \sum_{j=1}^{T} \mathbb{E}[\tilde{\sigma}(\vartheta \mathbf{h}_i^j \cdot \mathbf{c}_i) - \tilde{\sigma}(\mathbf{r}_i \cdot \mathbf{c}_i))^2] \leq \min_{t \in \{1, \ldots, a\}} \sum_{j=1}^{T} L_i^j(t) - \sum_{j=1}^{T} (\mathbf{r}_i/\vartheta) \cdot \mathbf{L}_i^j + O(\sqrt{T \log a}) + \log a + O(T \log(1/\delta)) \ . \tag{231}$$

Clearly, when $\|\mathbf{r}_i\|_1 = \vartheta$, we have that

$$\min_{t \in \{1, \ldots, a\}} \sum_{j=1}^{T} L_i^j(t) - \sum_{j=1}^{T} (\mathbf{r}_i/\vartheta) \cdot \mathbf{L}_i^j \leq 0 \ . \tag{232}$$

When we have $\|\mathbf{r}_i\|_1 < \vartheta$, we can augment $\mathbf{r}_i$ with a pseudo vector $\tilde{\mathbf{r}}_i$ such that $\|[\mathbf{r}_i, \tilde{\mathbf{r}}_i]\|_1 = \vartheta$ and the random vector $\mathbf{c}_i$ with an additional element that corresponds to 0 such that $\tilde{\mathbf{r}}_i$ corresponds to the weight associated with 0 and proceed further [15]. This also motivates the inclusion of auxiliary weights $\tilde{\kappa}_i^{uv}$ in Algorithm 1. Next, we note that with probability $1 - \delta$, we have

$$\frac{1}{4\vartheta} \sum_{j=1}^{T} \mathbb{E}[\tilde{\sigma}(\vartheta \mathbf{h}_i^j \cdot \mathbf{c}_i) - \tilde{\sigma}(\mathbf{r}_i \cdot \mathbf{c}_i))^2] = O(\sqrt{T \log a}) + O(\log a) + O(T \log(1/\delta)) \ . \tag{233}$$

Therefore, for $T = O(\vartheta^2 \log(a/\delta)/\varepsilon^2$, we must have that with probability $1 - \delta$,

$$\min_{j \in \{1, \ldots, T\}} \mathbb{E}[\tilde{\sigma}(\vartheta \mathbf{h}_i^j \cdot \mathbf{c}_i) - \tilde{\sigma}(\mathbf{r}_i \cdot \mathbf{c}_i))^2] \leq \varepsilon \ . \tag{234}$$

□

B. Joint Estimation of Ising Models

To complete the proof of Theorem 6, we note that

$$\mathbb{E}[B_i^u] = \frac{1}{1 + \exp(2\lambda \sum\limits_{\{v:(u,v) \in E_i\}} \lambda_i^{uv} X_i^u X_i^v)} \ . \tag{235}$$

Therefore, every node $u \in V$ can determine its neighborhood in $\mathcal{G}_1$ and $\mathcal{G}_2$ using Algorithm 2 by setting $\tilde{\sigma}$ to be a sigmoid function $\sigma(x) = \frac{1}{1+\exp(-x)}$, $a = p - 1$, and $D_i = B_i^u$ in $\mathcal{G}_i$. In this scenario, we have the following lemma in the context of Ising models that is equivalent to Theorem 14.

Lemma 12. *For a node u in $\mathcal{G}_i$, given $n_{\mathrm{T}} = O\left(\frac{\vartheta^2}{\alpha \varepsilon^2} \log \frac{p}{\varepsilon}\right)$ number of pairs of samples from nodes in $\mathcal{G}_1$ and $\mathcal{G}_2$, Algorithm 1 produces at least one set of weights $\{w_i^{uv}(j)\}$ for $j \in \{1, \ldots, n_{\mathrm{T}}\}$ such that with probability at least $1 - \frac{\alpha}{p^2}$,*

$$\mathbb{E}\left[\sigma\left(-2 \sum_{\{v:(u,v) \in E_i^j\}} w_i^{uv}(j) X_i^v\right) - \sigma\left(-2 \sum_{\{v:(u,v) \in E_i\}} \lambda_i^{uv} X_i^v\right)\right] \leq \varepsilon \ , \quad \forall \varepsilon > 0. \tag{236}$$

Subsequently, the statement of the Theorem 6 follows from Lemma 12 and [15, Lemma 4.3].

Appendix B
Proof of Theorem 12

In this class based on (205), the sufficient condition for recovering two q-similar graphs is

$$n \geq \frac{1}{\log \frac{(d^2-d+1)\rho^2 + (d-2)\rho + 1}{(1-\rho)(1+(d-1)\rho)}} \log\left(\frac{\binom{q}{d} + \binom{p-q}{d}^2}{\varepsilon}\right) \ . \tag{237}$$

In parallel, for recovering two graphs independently, we have

$$n > (1 - \varepsilon) \frac{2 \log \binom{p}{d} - 1}{\log\left(1 + \frac{d\rho}{1-\rho}\right) - \frac{d\rho}{1+(d-1)\rho}} \ , \tag{238}$$

which is the result established in (Wang and Wainwright, 2010) and can also be recovered from our results by setting $q = 0$ in (238) and $m = d$ in the ensemble construction for dense graphs in Section V-C2. To establish the desired

result, in our analysis we exclude the setting in which the two graphs are either almost identical (i.e., $q \to p$) or almost distinct (i.e., $q \to 0$). To this end, we focus on the regime $\max\{q, p-q\} < p^{1-2\varepsilon}$. We note that this regime is not too stringent. For instance, when $\varepsilon = 0.1$, q that satisfies $\max\{q, p-q\} < p^{0.8}$ lies in $q \in [p^{0.8}, p-p^{0.8}]$. For $p = 10000$, this range is $[1585, 8415]$.

We show that for any target error rate ε, as long as $\max\{q, p-q\} < p^{1-2\varepsilon}$, for $\rho > \frac{1}{d+1}$, we have

$$(1-\varepsilon)\frac{2\log\binom{p}{d}-1}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} > \frac{1}{\log\frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}\log\left(\frac{\binom{q}{d}+\binom{p-q}{d}^2}{\varepsilon}\right) , \quad (239)$$

which is equivalent to

$$\frac{\log\frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} > \frac{1}{(1-\varepsilon)(2\log\binom{p}{d}-1)}\log\left(\frac{\binom{q}{d}+\binom{p-q}{d}^2}{\varepsilon}\right) . \quad (240)$$

By noting that $\binom{q}{d} < \binom{q}{d}^2$ and leveraging the combinatorial inequalities

$$\left(\frac{p}{d}\right)^d \le \binom{p}{d} \le \left(\frac{p\mathrm{e}}{d}\right)^d , \quad (241)$$

we prove the following inequality, which is stronger than (240) and implies (240):

$$\frac{\log\frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} > \frac{1}{(1-\varepsilon)(2d\log p/d-1)}\left(\log\left(\frac{2r}{d}\right)^{2d}-\log\varepsilon\right) , \quad (242)$$

where we have defined $r \triangleq \max\{q, p-q\}\mathrm{e}$. Next, we show that for $\rho > \frac{1}{d+1}$ we have

$$\frac{\log\frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} > 1 , \quad (243)$$

and for $r < p^{1-2\varepsilon}$ we have

$$\frac{1}{(1-\varepsilon)(2d\log p/d-1)}\left(\log\left(\frac{2r}{d}\right)^{2d}-\log\varepsilon\right) < 1 . \quad (244)$$

To show the inequality in (262), we start by noting that for any ρ and d, we have

$$[\rho(d+1)-1]^2 \ge 0 . \quad (245)$$

This is equivalent to

$$\rho^2d^2 \ge -2\rho^2d+2\rho d-\rho^2+2\rho-1 . \quad (246)$$

Adding $3\rho^2d^2$ to both sides results in

$$4\rho^2d^2 \ge [(3d+1)\rho-1][(d-1)\rho+1] , \quad (247)$$

or equivalently in

$$\frac{\rho^2d^2}{(1-\rho)[(d-1)\rho+1]} \ge \frac{(3d+1)\rho-1}{4(1-\rho)} . \quad (248)$$

Subsequently

$$1+\frac{\rho^2d^2}{(1-\rho)[(d-1)\rho+1]} \ge \frac{3}{4}\left(1+\frac{d\rho}{1-\rho}\right) . \quad (249)$$

By noting that $\frac{1}{\sqrt{2}} < \frac{3}{4}$, (249) implies that

$$1+\frac{\rho^2d^2}{(1-\rho)[(d-1)\rho+1]} \ge \frac{1}{\sqrt{2}}\left(1+\frac{d\rho}{1-\rho}\right) , \quad (250)$$

or equivalently

$$\log\left(1+\frac{\rho^2d^2}{(1-\rho)[(d-1)\rho+1]}\right) \ge \log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{1}{2} . \quad (251)$$

Next, we note that for $\rho > \frac{1}{d+1}$, we have

$$\frac{d\rho}{1+(d-1)\rho} > \frac{1}{2} . \quad (252)$$

Hence, (251) and (252) indicate that

$$\log\left(1+\frac{\rho^2d^2}{(1-\rho)[(d-1)\rho+1]}\right) > \log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho} , \quad (253)$$

which proves that for $\rho > \frac{1}{d+1}$, the inequality in (262) holds. Next, we show that for $r < \rho^{1-2\varepsilon}$ and $\log p > \frac{1}{2d\varepsilon}\left(2+\log\frac{1}{\varepsilon}\right)$, (244) holds as well. To show this, we start by noting that

$$\log p > \frac{1}{2d\varepsilon}\left(2+\log\frac{1}{\varepsilon}\right) , \quad (254)$$

implies that

$$2d\varepsilon\log p > 2+\log\frac{1}{\varepsilon}-(\varepsilon+\varepsilon\log d) . \quad (255)$$

By expanding ε as $\varepsilon = (1-\varepsilon)-(1-2\varepsilon)$ and leveraging (255) we obtain

$$2d\left[\log\frac{p^{1-\varepsilon}}{p^{1-2\varepsilon}}+\log\frac{d}{d^{1-\varepsilon}}\right] > 2-\varepsilon+\log\frac{1}{\varepsilon} . \quad (256)$$

By noting that $r < p^{1-2\varepsilon}$, from (256), we get

$$2d\left[\log\frac{p^{1-\varepsilon}d}{rd^{1-\varepsilon}}\right] > 2-\varepsilon+\log\frac{1}{\varepsilon} , \quad (257)$$

which is equivalent to

$$2d(1-\varepsilon)\log\frac{p}{d}-(1-\varepsilon) > 1+2d\log\frac{r}{d}-\log\varepsilon = \log\left(2\left(\frac{r}{d}\right)^{2d}\right)+\log\frac{1}{\varepsilon} , \quad (258)$$

thus, establishing the inequality in (244).

APPENDIX C
GAINS OVER INDEPENDENT RECOVERY FOR SHARED CLIQUE IN GAUSSIAN MODELS

We expand our analysis for gains in the Gaussian models to consider a class of q-similar graphs in which the d-clique, U_d is not restricted to lie wholly within the shared part or the non-shared part. Therefore, this class supercedes the class $\mathcal{B}_q$.

Class $\mathcal{D}_q$: We assume that any graph in this sub-class consists of a clique of d nodes. Therefore, there are $\sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'}^2$ number of q-similar graph pairs in this subclass.

Note that this subclass has significantly higher number of possible q-similar graph pairs than the original class $\mathcal{B}_q$. Since the d-clique can also lie at the interface of shared and non-shared parts for $\mathcal{D}_q$, a sub-class of single graphs with $\binom{p}{d}$ number of graphs (i.e., the single graph class in [23]) is apt for comparison of sample complexities to establish gains due to structural similarity. The sufficient condition for sample complexity for jointly recovering two q-similar graphs in $\mathcal{D}_q$ is given by

$$n \geq \frac{1}{\log \frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}} \log \left(\frac{\sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'}^2}{\varepsilon} \right) . \tag{259}$$

To prove theoretical gains of joint recovery, we need to establish

$$(1-\varepsilon)\frac{2\log\binom{p}{d}-1}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} > \frac{1}{\log \frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}} \log \left(\frac{\sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'}^2}{\varepsilon} \right) , \tag{260}$$

which is equivalent to

$$\frac{\log \frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} > \frac{1}{(1-\varepsilon)(2\log\binom{p}{d}-1)} \log \left(\frac{\sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'}^2}{\varepsilon} \right) . \tag{261}$$

For the left hand side of (261), we note that the following holds for any $\rho > \frac{1}{d+1}$ from our previous analysis in Appendix B:

$$\frac{\log \frac{(d^2-d+1)\rho^2+(d-2)\rho+1}{(1-\rho)(1+(d-1)\rho)}}{\log\left(1+\frac{d\rho}{1-\rho}\right)-\frac{d\rho}{1+(d-1)\rho}} > 1 . \tag{262}$$

Therefore, we need to establish that

$$\frac{1}{(1-\varepsilon)(2\log\binom{p}{d}-1)} \log \left(\frac{\sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'}^2}{\varepsilon} \right) < 1 . \tag{263}$$

By re-arranging the terms in (263), we note that (263) is equivalent to

$$\log \frac{\binom{p}{d}^{2(1-\varepsilon)}}{\sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'}^2} > 1-\varepsilon-\log\varepsilon . \tag{264}$$

From Vandermonde's identity, we have

$$\binom{p}{d} = \sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'} , \tag{265}$$

and therefore,

$$\binom{p}{d}^2 = \left(\sum_{d'=0}^{d} \binom{q}{d'}\binom{p-q}{d-d'} \right)^2 > \sum_{d'=0}^{d} \binom{q}{d'}^2\binom{p-q}{d-d'}^2 . \tag{266}$$

Define

$$A_{d'} \triangleq \binom{q}{d'}\binom{p-q}{d-d'}^2 \quad \text{and} \quad B_{d'} \triangleq \binom{q}{d'}^2\binom{p-q}{d-d'}^2 . \tag{267}$$

Therefore,

$$\frac{B_{d'}}{A_{d'}} = \binom{q}{d'} \geq q, \forall d' > 0, d' < d . \tag{268}$$

From (266) and (267), we note that (263) is satisfied if we have

$$\log\left(\frac{\sum_{d'=0}^{d} B_{d'}}{\sum_{d'=0}^{d} A_{d'}}\right) + \log \frac{1}{\binom{p}{d}^{2\varepsilon}} > 1-\varepsilon-\log\varepsilon . \tag{269}$$

Further, from (268), we have $B_{d'} > qA_{d'}$ for all $0 < d' < d$. Therefore, (263) is satisfied if we have

$$\log((d-2)q+2) + \log \frac{1}{\binom{p}{d}^{2\varepsilon}} > 1-\varepsilon-\log\varepsilon . \tag{270}$$

We note that for sufficiently small ε, (270) is always satisfied. For instance, if $\varepsilon = 1/p$, $p = 1000$ and $q \in [300, 1000]$, (270) is always satisfied for any $d > 8$ and for $d > 27$ for $q \in [100, 300)$. Therefore, we are able to establish gains for joint graph recovery for class $\mathcal{D}_q$ which supercedes class $\mathcal{B}_q$ and has the corresponding class of single graphs as considered in [23].

REFERENCES

[1] S. L. Lauritzen, *Graphical Models*. Clarendon Press, May 1996, vol. 17.

[2] J. Pearl, *Causality: Models, Reasoning, and Inference.*, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[3] F. Battiston, J. Guillon, M. Chavez, V. Latora, and F. de Vico Fallani, "Multiplex core–periphery organization of the human connectome," *Journal of the Royal Society Interface*, vol. 15, no. 146, p. 20180514, 2018.

[4] T. Simas, M. Chavez, P. R. Rodriguez, and A. Diaz-Guilera, "An algebraic topological method for multimodal brain networks comparisons," *Frontiers in psychology*, vol. 6, p. 904, 2015.

[5] A. Nebli and I. Rekik, "Adversarial brain multiplex prediction from a single brain network with application to gender fingerprinting," *Medical Image Analysis*, p. 101843, Jan. 2021.

[6] B. Lee, S. Zhang, A. Poleksic, and L. Xie, "Heterogeneous multi-layered network model for omics data integration and analysis," *Frontiers in Genetics*, vol. 10, p. 1381, 2020.

[7] J. Guo, J. Cheng, E. Levina, G. Michailidis, and J. Zhu, "Estimating heterogeneous graphical models for discrete data with an application to roll call voting," *The Annals of Applied Statistics*, vol. 9, no. 2, pp. 821 – 848, Jun. 2015.

[8] R. Saqur and K. Narasimhan, "Multimodal graph networks for compositional generalization in visual question answering," *Proc. Advances in Neural Information Processing Systems*, vol. 33, 2020.

[9] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, Jul. 2018, pp. 2236–2246.

[10] J. Banks, "Multimodal, multiplex, multispatial: A network model of the self," *New Media and Society*, vol. 19, no. 3, pp. 419–438, Mar. 2017.

[11] X. Chen, F. J. Slack, and H. Zhao, "Joint analysis of expression profiles from multiple cancers improves the identification of microRNA–gene interactions," *Bioinformatics*, vol. 29, no. 17, pp. 2137–2145, 2013.

[12] C. J. Honey, R. Kötter, M. Breakspear, and O. Sporns, "Network structure of cerebral cortex shapes functional connectivity on multiple time scales," *Proceedings of the National Academy of Sciences*, vol. 104, no. 24, pp. 10 240–10 245, 2007.

[13] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, "The brain's default network: anatomy, function, and relevance to disease." 2008.

[14] D. M. Chickering, "Learning Bayesian networks is NP-complete," *Learning from Data*, vol. 112, pp. 121–130, 1996.

[15] A. Klivans and R. Meka, "Learning graphical models using multiplicative weights," in *Proc. Annual Symposium on Foundations of Computer Science*, Berkeley, CA, Oct. 2017, pp. 343–354.

[16] G. Bresler, "Efficiently learning Ising models on arbitrary graphs," in *Proc. Annual ACM Symposium on Theory of Computing*, Jun. 2015, pp. 771–782.

[17] M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov, "Interaction screening: Efficient and sample-optimal learning of Ising models," in *Proc. Advances in Neural Information Processing Systems*, Barcelona, Spain, Dec. 2016, pp. 2595–2603.

[18] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions." *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4117–4134, May 2012.

[19] R. Tandon, K. Shanmugam, P. K. Ravikumar, and A. G. Dimakis, "On the information-theoretic limits of learning Ising models," in *Proc. Advances in Neural Information Processing Systems*, Montreal, Canada, Dec. 2014, pp. 2303–2311.

[20] J. Scarlett and V. Cevher, "On the difficulty of selecting Ising models with approximate recovery," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 4, pp. 625–638, Dec. 2016.

[21] D. Vats and J. M. Moura, "Necessary conditions for consistent set-based graphical model selection," in *Proc. IEEE International Symposium on Information Theory*, Saint-Petersburg, Russia, Jul. 2011, pp. 303–307.

[22] A. K. Das, P. Netrapalli, S. Sanghavi, and S. Vishwanath, "Learning Markov graphs up to edit distance," in *Proc. IEEE International Symposium on Information Theory*, Cambridge, MA, Jul. 2012, pp. 2731–2735.

[23] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. IEEE International Symposium on Information Theory*, Austin, Texas, Jun. 2010.

[24] R. Tandon and P. Ravikumar, "On the difficulty of learning power law graphical models," in *Proc. IEEE International Symposium on Information Theory*, Istanbul, Turkey, Jul. 2013, pp. 2493–2497.

[25] A. Gangrade, B. Nazer, and V. Saligrama, "Lower bounds for two-sample structural change detection in Ising and Gaussian models." in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Oct. 2017, pp. 1016–1025.

[26] M. Neykov and H. Liu, "Property testing in high dimensional Ising models," *Annals of Statistics*, vol. 47, no. 5, pp. 2472–2503, 10 2019.

[27] L. Devroye, A. Mehrabian, and T. Reddad, "The minimax learning rate of Normal and Ising undirected graphical models," *Electronic Journal of Statistics*, vol. 14, no. 1, pp. 2338–2361, Jun. 2020.

[28] J. Heydari, A. Tajer, and H. V. Poor, "Quickest detection of gauss-markov random fields," in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2015, pp. 808–814.

[29] ——, "Quickest detection of Markov networks," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 1341–1345.

[30] A. Tajer, J. Heydari, and H. V. Poor, "Active sampling for the quickest detection of markov networks," *IEEE Transactions on Information Theory*, vol. 68, no. 4, pp. 2479–2508, 2021.

[31] R. Wu, R. Srikant, and J. Ni, "Learning loosely connected Markov random fields," *Stochastic Systems*, vol. 3, no. 2, pp. 362–404, 2013.

[32] G. Bresler and M. Karzand, "Learning a tree-structured Ising model in order to make predictions," *Annals of Statistics*, vol. 48, no. 2, pp. 713–737, Apr. 2020.

[33] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional structure estimation in Ising models: Local separation criterion," *The Annals of Statistics*, vol. 40, no. 3, pp. 1346–1375, Jun. 2012.

[34] J. Fang, L. S. Dongdong, S. Charles, Z. Xu, V. D. Calhoun, and Y.-P. Wang, "Joint sparse canonical correlation analysis for detecting differential imaging genetics modules," *Bioinformatics*, vol. 32, no. 15, pp. 3480–3488, 2016.

[35] P. Danaher, P. Wang, and D. M. Witten, "The joint graphical lasso for inverse covariance estimation across multiple classes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 2, pp. 373–397, Mar. 2014.

[36] S. Yang, Z. Lu, X. Shen, P. Wonka, and J. Ye, "Fused multiple graphical lasso," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 916–943, 2015.

[37] K. Mohan, P. London, M. Fazel, D. Witten, and S.-I. Lee, "Node-based learning of multiple Gaussian graphical models," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 445–488, 2014.

[38] C. B. Peterson, F. C. Stingo, and M. Vannucci, "Bayesian inference of multiple Gaussian graphical models," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 159–174, 2015.

[39] H. Qiu, F. Han, H. Liu, and B. Caffo, "Joint estimation of multiple graphical models from high-dimensional time series," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 2, pp. 487–504, 2016.

[40] S. Sihag and A. Tajer, "Structure learning with side information: Sample complexity," in *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2019, pp. 14 380–14 390.

[41] ——, "Approximate recovery of Ising models with side information," in *Proc. IEEE International Symposium on Information Theory*, Paris, France, Jul. 2020, pp. 1319–1324.

[42] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119 – 139, Apr. 1997.

[43] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory.*, 2nd ed. John Wiley and Sons, 2012.

Saurabh Sihag Saurabh Sihag (Member, IEEE) received the B.Tech. and M.Tech. degrees in electrical engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 2016, and the Ph.D. degree in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, USA, in 2020. He was the recipient of the 2021 Charles M. Close '62 Doctoral Prize by the Department of Electrical, Computer and Systems Engineering at Rensselaer Polytechnic Institute. He is currently a postdoctoral research fellow with the University of Pennsylvania, Philadelphia, PA, USA, where he was the Clinical Research in ALS and related disorders for Therapeutic Development (CReATe) Consortium scholar in 2021. His research interests include statistical signal processing, information theory, high-dimensional statistics, machine learning, and network neuroscience.

Ali Tajer (S'05, M'10, SM'15) received the B.Sc. and M.Sc. degrees in Electrical Engineering from Sharif University of Technology in 2002 and 2004, respectively. During 2007-2010 he was with Columbia University where he received the M.A degree in Statistics and the Ph.D. degree in Electrical Engineering, and during 2010-2012 he was with Princeton University as a Postdoctoral Research Associate. He is currently an Associate Professor of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute. His research interests include mathematical statistics, statistical signal processing, and network information theory, with applications in wireless communications and power grids. His recent publications include an edited book entitled Advanced Data Analytics for Power Systems (Cambridge University Press, 2021). He has received an NSF CAREER award in 2016 and AFRL Faculty Fellowship in 2019. He is currently serving as an Associate Editor for the IEEE Transactions on Information Theory and for the IEEE Transactions on Signal Processing. In the past he has served as an an Editor for the IEEE Transactions on Communications, a Guest Editor for the IEEE Signal Processing Magazine, an Editor for the IEEE Transactions on Smart Grid, an Editor for the IET Transactions on Smart Grid, and as the Guest Editor-in-Chief for the IEEE Transactions on Smart Grid Special Issue on Theory of Complex Systems with Applications to Smart Grid Operations.