Estimation Rates for Sparse Linear Cyclic Causal Models

Jan-Christian Hütter
Broad Institute of MIT and Harvard
jhuetter@broadinstitute.org

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
rigollet@math.mit.edu

Abstract

Causal models are fundamental tools to understand complex systems and predict the effect of interventions on such systems. However, despite an extensive literature in the population—or infinite-sample—case, where distributions are assumed to be known, little is known about the statistical rates of convergence of various methods, even for the simplest models. In this work, allowing for cycles, we study linear structural equations models with homoscedastic Gaussian noise and in the presence of interventions that make the model identifiable. More specifically, we present statistical rates of estimation for both the LLC estimator introduced by Hyttinen, Eberhardt and Hoyer and a novel two-step penalized maximum likelihood estimator. We establish asymptotic near minimax optimality for the maximum likelihood estimator over a class of sparse causal graphs in the case of near-optimally chosen interventions. Moreover, we find evidence for practical advantages of this estimator compared to LLC in synthetic numerical experiments.

1 INTRODUCTION

Directed graphical models (Pearl, 2009; Spirtes et al., 2000) provide a useful framework for interpretation, inference, and decision making in many areas of science such as biology, sociology, and environmental sciences (Friedman et al., 2000; Duncan, 1966; Keats and Hitt, 1988). Unlike their undirected counterparts that merely encode the structure of probabilistic dependence between random variables directed graphical models reveal causal effects that are the basis of scientific discovery (Pearl, 2009).

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

Most frequently, the model is assumed to be governed by a directed acyclic graph (DAG) $G = (\mathsf{V}, \mathsf{E})$, where $\mathsf{V} = \{X_1, \dots, X_p\}$ are the variables of an observed system and E is a set of edges such that there is no directed cycle in G. In such models, known as *Bayes networks* (Pearl, 2009), the variables follow a joint distribution that factorizes according to the graph G in the sense that node i is independent of other nodes conditionally on its parents. The absence of cycles allows for a direct interpretation of the causal structure between the variables $X_1, \dots, X_p$ whereby a directed edge corresponds to a causal effect. At the same time, most complex systems showcase feedback loops that can be both positive and negative, and the need to extend Bayes networks to allow for cycles was recognized long ago.

A large body of work focuses on learning Bayes networks from *observational* data, that is, data drawn independently from the joint distribution of $(X_1, \dots, X_p)$. Observational data is rather abundant but even in the acyclic cases, it is known to lead to a severe lack of identifiability: Such data, even in infinite abundance, can only yield an equivalence class—the Markov equivalence class—of DAGs that are all compatible with the conditional independence relation in the given data. While a DAG in the Markov equivalence class can already yield decisive scientific insight (Maathuis et al., 2009), searching over the space of DAGs is often computationally hard. Many algorithms have been proposed over the years such as the PC algorithm (Spirtes et al., 2000) and Greedy Equivalence search (Chickering, 2002) and max-min hill-climbing (Tsamardinos et al., 2006), but all of them rely on the notion of faithfulness of the distribution, *i.e.*, the assumption that all conditional dependence relations that could be compatible with the DAG G are actually fulfilled by the distribution of X. In fact, for consistency of these algorithms, one needs to assume that these dependencies observe a signal-to-noise ratio that allows to detect them with high probability (Kalisch and Bühlmann, 2007; Loh and

Bühlmann, 2014; van de Geer and Bühlmann, 2013). Extensions that allow certain kinds of cycles, (Richardson, 1996; Richardson and Spirtes, 1996; Schmidt and Murphy, 2009; Itani et al., 2010; Lacerda et al., 2008) have been proposed but at the expense of having an increased number of graphs in each equivalence class.

Recent breakneck advances in data collection processes such as the spread of A/B testing for online marketing or targeted gene editing with CRISPR-Cas9 are contributing to the proliferation of *interventional* data, the gold standard for causal inference. With unlimited interventions on any combination of nodes, learning a directed graphical model becomes a trivial task. However, exhaustively performing all interventions is a daunting and costly task and recent work has focused on finding a small number of interventions for several classes of DAGs (Shanmugam et al., 2015; Kocaoglu et al., 2017). For graphs with cycles, Hyttinen et al. (2012) have characterized the system of interventions necessary to learn a parametric linear structural equation model (SEM) (Bielby and Hauser, 1977; Bollen, 1989), in which all variables are real-valued and the causal relationships given by the edges E are linear. Formally, the special case of this model we consider here postulates that the following equation holds (in distribution) for observational samples from X:

$$X = B^* X + Z, \quad Z \sim \mathcal{N}(0, I), \tag{1.1}$$

where we exclude explicit self-loops by assuming that the diagonal of $B^* \in \mathbb{R}^{p\times p}$ is zero. By writing $X = (I - B^*)^{-1}Z$ and assuming that the corresponding inverse matrix exists, this allows us to handle underlying graphs that are cyclic. A more general form of this model, additionally allowing for latent confounders and unknown noise variances, has been extensively studied in Hyttinen et al. (2012). There, it is shown that if we have access to data from a sufficiently rich system of interventions, *i.e.*, if enough variables are randomized and are thus made independent of the influence of their parents encoded in B^*, then on a population level, B^* is identifiable by a method of moments type estimator that the authors call LLC (for linear, latent, causal).

In this paper, we present upper and lower bounds for the reconstruction of B^* in Frobenius norm for classes of sparse B^*, corresponding to graphs with bounded in-degree, using multiple observations for each intervention setup. We also provide upper bounds for the original LLC estimator with ℓ_1-penalization term as well as an ℓ_1-penalized maximum likelihood estimator, all under the simplifying assumption that the noise or disturbance variables Z are Gaussian, independent of each other, and have unit variance. Moreover, we provide numerical evidence that a non-convex ADMM type algorithm can be used to find a solution to this maximum likelihood problem, albeit without convergence guarantees.

1.1 RELATED WORK

It is known that several variants of the model (1.1) are identifiable from observational data, including nonlinear SEMs (Hoyer et al., 2009) or non-Gaussian noise (Shimizu et al., 2006). Linear SEMs with Gaussian noise can be identifiable under additional assumptions, for example when the components of the noise have equal variances and the underlying graph is a DAG (Loh and Bühlmann, 2014; Peters and Bühlmann, 2014), when the underlying graph is random and sparse (Abrahamsen and Rigollet, 2018), or when the noise variables fulfill certain additional identifiability conditions (Ghoshal and Honorio, 2018). Moreover, in the DAG case, lower bounds for general exponential family models are available (Ghoshal and Honorio, 2017). Similarly, structural assumptions that lead to identifiability from observational data also arise in Independent Component Analysis (Shimizu et al., 2006; Abrahamsen and Rigollet, 2018).

Moreover, many more approaches to dealing with cycles and/or interventions are known, such as convex regularizers in an exponential family model (Schmidt et al., 2007; Schmidt and Murphy, 2009), independence testing (Itani et al., 2010), Independent Component Analysis (Lacerda et al., 2008), noisy path queries (Bello and Honorio, 2018), and adapting Greedy Equivalence Search to handle interventional data (Hauser and Bühlmann, 2012; Wang et al., 2018). From the above, it seems that the linear Gaussian case is somewhat of a worst-case example for identifiability of the ground truth matrix, especially when allowing cycles, and thus warrants the investigation of controlled interventions to eliminate ambiguity, which is the main contribution of Hyttinen et al. (2012). Similar models have been considered for applications, for example in computational biology, see Cai et al. (2013), where identifiability is not provided by controlled experiments on the variance, but rather by a mean shift of some variables.

Our work extends the results in Hyttinen et al. (2012) by providing explicit upper bounds for their suggested method, as well as presenting an alternative estimator that leads to upper bounds independent of the conditioning of the experiments as explained in Section 3.3. In spirit, our results are similar to consistency guarantees obtained in van de Geer and Bühlmann (2013) and Wang et al. (2018), but we focus on the case where enough interventions are performed to identify the ground truth structure matrix B^*, alleviating the need for additional assumptions on B^*.

1.2 STRUCTURE OF THE PAPER

The rest of the paper is structured as follows: In Section 2, we give an overview of the linear structural equation model we consider and the main assumptions we make. In Section 3, we present lower bounds, upper bounds for LLC, and upper bounds for a two-step maximum likelihood estimator. In Section 4, we give an abbreviated version of numerical experiments on synthetic data. The full version is presented in Section A of the supplement, where we derive a non-convex variant of ADMM to solve part of the numerical optimization problem for the penalized maximum likelihood estimator and explore its performance on synthetic and semi-synthetic data. The proofs of the main results are deferred to Sections C – E in the supplement, and we collect extended notational conventions in Section B and general lemmas used in all the proofs in Section F. Section G contains a short argument for why experimental data is necessary given our assumption, and Section H provides a way of speeding up our numerical calculations.

1.3 NOTATION

We write $a \lesssim b$ for two quantities a and b if there exists an absolute constant $C > 0$ such that $a \leq Cb$, and similarly for $a \gtrsim b$. For two real numbers $a, b \in \mathbb{R}$, we write $a \wedge b$ for their minimum and $a \vee b$ their maximum, respectively. For a natural number p, we denote by $[p] = \{1, \dots, p\}$. Given a set S, we write $|S|$ for its cardinality.

For two matrices $A, B \in \mathbb{R}^{p_1 \times p_2}$, we abbreviate the ith row by $B_{i,:}$ and the ith column by $B_{:,i}$. Similarly, $B_{i,-j}$ denotes the ith row of B where the jth element is omitted. Further, $\|B\|_F$ denotes the Frobenius norm, $\|B\|_{\text{op}}$ the operator norm, and

$$\|B\|_1 = \sum_{i,j} |B_{i,j}|.$$

If A is a square invertible matrix, we denote by A^{-1} its inverse and by $A^{-\top}$ the transpose of A^{-1}. By $I \in \mathbb{R}^{p \times p}$, we denote the identity matrix.

2 MODEL AND ASSUMPTIONS

Before summarizing our explicit assumptions, we give a definition of observations under a linear cyclic structural equation model with and without interventions. We assume that a linear SEM on a random vector $X = (X_1, \dots, X_p)$ is given by a matrix $B^* \in \mathbb{R}^{p \times p}$ without self-cycles, *i.e.*, $B^* \in \mathcal{B}_0$ with

$$\mathcal{B}_0 := \{B \in \mathbb{R}^{p \times p} : B_{i,i} = 0, \text{ for all } i = 1, \dots, p\}.$$

That is, if $B_{i,j} \neq 0$ for $i \neq j$, then there is a linear causal dependence of X_i on X_j, or equivalently, an edge (j, i) in the directed graph associated with X. Without any intervention, each observation is an independent copy of $X = (I - B^*)^{-1} Z$, where Z can in principle be any noise variable. Since non-Gaussian noise can lead to identifiability from observational data through exploiting this particular property (Hoyer et al., 2009; Lacerda et al., 2008), we focus on Gaussian noise, and make the simplifying assumption that $Z \sim \mathcal{N}(0, I)$. In order to guarantee that $(I - B^*)^{-1}$ exists, we assume $\|B^*\|_{\text{op}} < 1$ which in particular allows us to write

$$X = \sum_{k=0}^{\infty} (B^*)^k Z,$$

and X can be interpreted as the steady state distribution of an auto-regressive process $\{x_t\}_{t \geq 0}$ governed by the dynamics

$$x_{t+1} = B^* x_t + Z, \quad x_0 = Z. \tag{2.1}$$

Hence, X is distributed according to $X \sim \mathcal{N}(0, \Sigma^*)$ with

$$\Sigma^* = (I - B^*)^{-1} (I - B^*)^{-\top}.$$

In order to obtain results in the high-dimensional regime $p \asymp n$, we additionally assume that the in-degree of B^* is bounded, resulting in a sparse matrix B^*. That is, if we denote the maximum in-degree of a matrix $B \in \mathbb{R}^{p \times p}$ by

$$d(B) = \max_{i \in [p]} |\{j : B_{i,j} \neq 0\}|,$$

then we assume $d(B^*) \ll p$.

Moreover, we assume that we have access to interventional, *a.k.a.* experimental, data, which is modeled as follows, keeping in line with the definition from Hyttinen et al. (2012). An experiment e is given by a partition

$$[p] = \mathcal{U}_e \,\dot\cup\, \mathcal{J}_e, \tag{2.2}$$

with associated projection matrices

$$(U_e)_{i,j} = \begin{cases} 1, & i = j \text{ and } i \in \mathcal{U}_e \\ 0, & \text{otherwise,} \end{cases}$$

$$(J_e)_{i,j} = \begin{cases} 1, & i = j \text{ and } i \in \mathcal{J}_e \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

In effect, all nodes in $\mathcal{J}_e$ are intervened on, *i.e.*, they are not influenced by their parents anymore. We assume that they follow a standard Gaussian distribution $\mathcal{N}(0, 1)$, leading to a random variable $X^e \sim \mathcal{N}(0, \Sigma^{*,e})$ corresponding to experiment e with covariance matrix

$$\Sigma^{*,e} = (I - U_e B^*)^{-1} (I - U_e B^*)^{-\top},$$

and inverse covariance matrix (concentration matrix)

$$\Theta^{*,e} = (\Sigma^{*,e})^{-1} = (I - U_e B^*)^\top (I - U_e B^*).$$

Hyttinen et al. (2012) provide the following criterion to identify B^* from interventional data associated with $\mathcal{E}$.

Definition 1 (Completely separating system)**.** *The set of experiments $\mathcal{E}$ is a* completely separating system *if for every $i \neq j \in [p]$, there exists $e \in \mathcal{E}$ such that $i \in \mathcal{J}_e$ and $j \in \mathcal{U}_e$.*

Note that Hyttinen et al. (2012) call the separation condition for a pair $(i, j) \in [p]^2$ the *pair condition*. They show that Definition 1 guarantees identifiability of B^* from observational data. Conversely, they show that if $\mathcal{E}$ is not separating, there exists a ground truth system that is not satisfied, albeit allowing a more general covariance structure on the noise terms (Z_k^e in Assumption A3 below) for the latter construction than we do.

We are now in a position to state our assumptions.

A1 (Structure matrix)**.** *For any two positive integers $d \leq p$ and $\eta \in (0, 1/2]$, let $\mathcal{B}(p, d, \eta)$ denote the set of sparse matrices defined by*

$$\mathcal{B}(p, d, \eta) := \{B \in \mathbb{R}^{p \times p} : B_{i,i} = 0 \, for \; i \in [p], \\ \|B\|_{\text{op}} \leq 1 - \eta, \; d(B) \leq d\},$$

and assume $B^* \in \mathcal{B}(p, d, \eta)$.

A2 (Interventions)**.** *Let $\mathcal{E}$ be a set of experiments with associated partitions $\{(\mathcal{U}_e, \mathcal{J}_e)\}_{e \in \mathcal{E}}$ and projection matrices $\{(U_e, J_e)\}_{e \in \mathcal{E}}$ as in (2.2) and (2.3), respectively. Assume that $\mathcal{E}$ is* separating *in the sense of Definition 1.*

A3 (Noise)**.** *Assume $n \in \mathbb{N}$ is divisible by $E := |\mathcal{E}|$, set $n_e = n/E$ for $e \in \mathcal{E}$, and for $k \in [n_e], e \in \mathcal{E}$, denote by $Z_k^e \sim \mathcal{N}(0, I)$ i.i.d. Gaussian random vectors. Then, we assume that we have access to observations of the form $X_k^e = (I - U_e B^*)^{-1} Z_k^e$.*

A few remarks are in order.

A1. The bound $\|B^*\|_{\text{op}} \leq 1 - \eta$ guarantees invertibility of $I - UB^*$ for any projection matrix U and convergence of the process (2.1).

A2. As mentioned, this is the same assumption under which Hyttinen et al. (2012) show identifiability of B^* under more general assumptions than the ones presented here, in particular allowing more general noise variances and hidden variables. Note that their proof of necessity of this assumption does not exactly match our assumption because our noise variances are restricted, so in principle, identifiability from observational data could be possible under a weaker condition. However, we give evidence in Section G that at least observational data alone is not sufficient to recover a general B^*.

Intuitively, the fact that $\mathcal{E}$ is separating guarantees that B^* can be recovered from submatrices of $\{\Sigma^{*,e}\}_{e \in \mathcal{E}}$ via solving a system of linear equations, a fact that is made more precise in Section 3.2. Since we are interested in recovering B^* under otherwise minimal assumptions on B^*, this is the case we consider for the theoretical contributions of this paper. We do however investigate the behavior of the two estimators considered in Section 3 with respect to a violation of this assumption numerically in Section 4.

A3. The assumption of Gaussian noise is not critical for our analysis, and in fact all our proofs extend readily to sub-Gaussian noise. Similarly, the assumption $n_e = n/E$ can be replaced by $n_e \asymp n/E$, that is, the number of observations in all experiments is comparable. Next, the assumption $\mathbb{E}[Z_k^e] = 0$ can be relaxed to an unknown mean by estimating the means of the individual experiments and subtracting them off, incurring only higher-order error terms with respect to n. On the other hand, the assumption that, $\mathbb{E}[(Z_k^e)^2] = 1$ might be restrictive in practice. We conjecture that it might be relaxed while maintaining many of the guarantees we give in Section 3, but due to the notational burden associated with incorporating these additional factors into the estimation, we chose to leave this topic as the subject of future research. Note that while the assumption of equal variances implies identifiability from observational data in the case where B^* is assumed to be acyclic (Loh and Bühlmann, 2014; Peters and Bühlmann, 2014), it does not in the cyclic case, see Section G. Hence, the assumptions as presented still lead to a class rich enough to require controlled experiments to estimate B^*. Moreover, contrary to the approach in (Loh and Bühlmann, 2014; Peters and Bühlmann, 2014), we do not explicitly exploit the fact that the variance is known by sorting the variables, likely rendering the estimators considered here robust in the case where the variances have to be estimated as well.

Remark 2. *It was shown in Hyttinen et al. (2013) that the minimum number of experiments necessary to obtain a completely separating system is of the order $\log(p)$, which can be seen by a simple binary coding argument. Hence, if we are able to pick the experiments in the most parsimonious way possible, $E = O(\log(p))$ only contributes a logarithmic factor to any of the rates presented in Section 3.*

3 MAIN RESULTS

3.1 LOWER BOUNDS

First, we give lower bounds for the estimation of matrices $B^* \in \mathcal{B}(p, d, \eta)$. This information-theoretic result sets a benchmark for any method employed in this model. To that end, let κ denote the *redundancy* of the experiments $\mathcal{E}$. It is defined as the maximum number of experiments that separate two variables,

$$\kappa = \kappa(\mathcal{E}) = \max_{i \neq j \in [p]} |\{e \in \mathcal{E} : i \in \mathcal{U}_e, j \in \mathcal{J}_e\}|.$$

Theorem 3. *There exists a constant $c > 0$ such that if $d \leq p/4$ and*

$$n \geq pdE^2 \log\left(1 + \frac{p}{4d}\right),$$

then, for any estimator $\hat{B}$, there exists $B^* \in \mathcal{B}(p, d, \eta)$ such that

$$\|\hat{B} - B^*\|_F^2 \geq c\frac{pdE}{\kappa n} \log\left(1 + \frac{p}{4d}\right) \tag{3.1}$$

with constant probability.

The proof of Theorem 3 is deferred to Section C of the supplement. We remark that there is a mismatch in the lower bound and the range of n for which it is effective that is of order E. In the case of a minimal system of completely separating interventions, by Remark 2, this mismatch is of order $\log(p)$.

3.2 UPPER BOUNDS FOR THE LLC ESTIMATOR

Next, we give bounds on the performance of the LLC estimator introduced in Hyttinen et al. (2012). We briefly summarize the algorithm below, which can be seen as a moment estimator for B^*.

3.2.1 The LLC estimator

Denote by $b_i^* \in \mathbb{R}^{p-1}$ the ith row of B^*, where we omit the ith entry, which is assumed to be zero since $B^* \in \mathcal{B}$. Formally, $b_i^* = (P_i B_{i,:}^\top) = B_{i,-i}^\top$, where $P_i \colon \mathbb{R}^p \to \mathbb{R}^{p-1}$ denotes the projection operator that omits the ith coordinate.

LLC is motivated by the observation that on the population level, each b_i^* satisfies a linear system $T_i^* b_i^* = t_i^*$, where $T_i^* \in \mathbb{R}^{m_i \times (p-1)}$ and $t_i^* \in \mathbb{R}^{m_i}$ for some $m_i \geq 1$ are defined as follows. For $i = 1, \dots, p$, define the matrix T_i^* and the column vector t_i^* row by row. For each experiment e such that $i \in \mathcal{U}_e$ and each $j \in \mathcal{J}_e$, add a row to T_i^* and to t_i^*, say with index $\ell = \ell(e, j)$, that is of the form

$$(T_i^*)_{\ell,:} = \mathfrak{e}_j^\top \Sigma^{*,e} P_i^\top, \qquad (t_i^*)_\ell = \Sigma_{j,i}^{*,e}$$

where $\mathfrak{e}_j$ is the jth canonical vector of $\mathbb{R}^p$. To better visualize $(T_i^*)_{\ell,:}$, one may rearrange the indices so that $\mathcal{J}_e = \{1, \dots, |\mathcal{J}_e|\}$, in which case we have

$$(T_i^*)_{\ell,:} = \begin{bmatrix} 0 & \dots & 1 & \dots & 0 & \Sigma_{j,\mathcal{U}_e \setminus \{i\}}^{*,e} \end{bmatrix},$$

where "1" appears in the jth coordinate. Let m_i denote the total number of such rows obtained by scanning through all experiments e such that $i \in \mathcal{U}_e$ and j such that $j \in \mathcal{J}_e$.

When $\mathcal{E}$ is a completely separating system, $T_i^* b_i = t_i^*$ has the unique solution $b_i^* = (B_{i,-i}^*)^\top$, (Hyttinen et al., 2012). The LLC estimator is obtained by substituting $\Sigma^{*,e}$ in the above definitions with its empirical counterpart $\hat{\Sigma}^e$ defined by

$$\hat{\Sigma}^e = \frac{1}{n_e} \sum_{k=1}^{n_e} X_k^e (X_k^e)^\top,$$

except for where the variances are known exactly due to the fact that an intervention is performed. This leads to a linear system of the form $\hat{T}_i b_i = \hat{t}_i$. Rather than solving the linear system exactly, the LLC estimator is obtained by minimizing a penalized least squares problem to promote sparsity in the resulting estimate:

$$\hat{b}_i = \operatorname*{argmin}_{b \in \mathbb{R}^{p-1}} \|\hat{T}_i b - \hat{t}_i\|_2^2 + \lambda \|b\|_1, \quad i = 1, \dots, p,$$

where $\lambda > 0$ is a tuning parameter. The solutions to the above problems are assembled into the LLC estimator $\hat{B}_{\text{llc}}$ by setting

$$(\hat{B}_{\text{llc}})_{i,-i} = \hat{b}_i^\top, \quad (\hat{B}_{\text{llc}})_{i,i} = 0, \quad i \in [p]. \tag{3.2}$$

3.2.2 Statistical performance

The upper bounds we give for the performance of LLC depend on additional constants that are not directly controlled for an arbitrary $B^* \in \mathcal{B}(p, d, \eta)$. Loosely speaking, they pertain to the conditioning of the ℓ^1-regularized least squares problems that are solved to obtain $\hat{B}_{\text{llc}}$. These constants are defined as follows. Denote by

$$\mathcal{C}(d) := \{v \in \mathbb{R}^p : \text{for all } S \subseteq [p] \text{ with } |S| \leq d, \\ \|v_{S^c}\|_1 \leq 3\|v_S\|_1\}.$$

Then, define

$$\rho(d) = \min_{i \in [p]} \inf_{v \in \mathcal{C}(d), v \neq 0} \frac{\|T_i^* v\|_2}{\|v\|_2},$$

$$R(d) = \max_{i \in [p]} \sup_{\substack{v \in \mathbb{R}^p, v \neq 0, \\ |\operatorname{supp}(v)| \leq d}} \frac{\|T_i^* v\|_2}{\|v\|_2},$$

$$\tilde{R} = \max_{i \in [p]} \max_{j \in [p]} \sum_{k \in [p]} |(T_i^*)_{k,j}|.$$

We are now in a position to state the first rate of convergence for the LLC estimator.

Theorem 4 (Rates for LLC estimator)**.** *Let assumptions A1 – A3 hold and fix $\delta \in (0,1)$. Assume further that*

$$n \gtrsim \left(1 \vee \frac{p^2}{\tilde{R}^2\eta^4} \vee \frac{pd}{(R(d)+1)^2\eta^4\rho(d)^4}\right) E\log(e\kappa p/\delta).$$

Then LLC estimator $\hat{B}_{\text{llc}}$ defined in (3.2) with λ chosen such that

$$\lambda \asymp \tilde{R}\sqrt{\frac{E\log(e\kappa p/\delta)}{n}},$$

satisfies

$$\|\hat{B}_{\text{llc}} - B^*\|_F^2 \lesssim \frac{\tilde{R}^2}{\rho(d)^4\eta^4}\frac{pdE\log(e\kappa p/\delta)}{n}, \quad (3.3)$$

with probability at least $1-\delta$.

The proof is deferred to Section D of the supplement. It uses standard arguments for the LASSO, together with perturbation results for regression with noisy design from Loh and Wainwright (2011) in Lemma 8 to handle the presence of noise in the matrices $\hat{T}_i$.

Remark 5. *Unfortunately, it is not clear whether the factors $\rho(d)$, $R(d)$, $\tilde{R}$ stay bounded with increasing p, d, and E, uniformly over all possible ground truth matrices $B^* \in \mathcal{B}(p,d,\eta)$. Hence, even though the explicit dependence on p, d, and E in the upper bounds (3.3) matches the lower bounds (3.1), we can not claim this rate to be (near) minimax optimal.*

Remark 6. *Comparing the definitions of $\rho(d)$ and $R(d)$, one might prefer an alternative definition of the former of the form*

$$\tilde{\rho}(d) := \min_{i\in[p]} \inf_{\substack{v\in\mathbb{R}^p,\, v\neq 0,\\ |\operatorname{supp} v|\leq d}} \frac{\|T_i^* v\|_2}{\|v\|_2}.$$

In fact, these two quantities are related, albeit for different values of d, see Section 8 in Bellec et al. (2018). We choose $\rho(d)$ instead of $\tilde{\rho}(d)$ for the sake of a simpler presentation.

In order to address the issues raised in the previous remark, we next give a penalized maximum likelihood estimator.

3.3 UPPER BOUNDS FOR TWO-STEP PENALIZED LIKELIHOOD

3.3.1 Two-step maximum likelihood estimator

One shortcoming in the rate for LLC for large n in Theorem 4 are the constants $\rho(d)$ and $\tilde{R}$ which might actually grow with p, see Remark 5. Moreover, as a moment estimator, it does not naturally behave well with respect to model misspecification. This motivates a different estimator based on a penalized maximum likelihood approach.

Recall that the negative log-likelihood of a multivariate Gaussian with empirical covariance matrix $\hat{\Sigma}$ and precision matrix Θ is given by.

$$\ell(\Theta,\hat{\Sigma}) = \mathsf{Tr}(\hat{\Sigma}\Theta) - \log\det(\Theta)$$

Thus, the negative log-likelihood for the whole model is proportional to

$$\mathcal{L}(B) = \mathcal{L}(B,\hat{\Sigma}^1,\ldots,\hat{\Sigma}^E) = \sum_{e\in\mathcal{E}} \ell(\Theta^e(B),\hat{\Sigma}^e),$$

where $\Theta^e(B) = (I-U_eB)^\top(I-U_eB)$, and

$$\hat{\Sigma}^e = \frac{1}{n_e}\sum_{k=1}^{n_e} X_k^e(X_k^e)^\top = \frac{E}{n}\sum_{k=1}^{n_e} X_k^e(X_k^e)^\top.$$

In order to exploit sparsity in the underlying matrix B^*, we need to penalize $\mathcal{L}(B)$ before minimizing it. However, due to the non-linear dependence of Σ^e on B, a vanilla ℓ_1-penalization term might not yield desirable statistical rates. To overcome this limitation, we propose a two-step estimation procedure. First, an initial guess $\hat{B}_{\text{init}}$ is produced using a penalization acting on the scale of the concentration matrices. This initial guess is subsequently refined to $\hat{B}$ as the solution to the ℓ_1-penalized log-likelihood restricted to a small ball around $\hat{B}_{\text{init}}$.

In the first step, we employ penalization with a term resembling a graphical lasso penalty for each experiment,

$$\text{pen}_{\text{init}}(B) = \text{pen}_{\text{init},\lambda_{\text{init}}}(B) = \lambda_{\text{init}}\sum_{e\in\mathcal{E}}\|\Theta^e(B)\|_1,$$

leading to the penalized log-likelihood

$$\begin{aligned}\mathcal{T}_{\text{init}}(B) &= \mathcal{T}_{\text{init},\lambda_{\text{init}}}(B,\hat{\Sigma}^1,\ldots,\hat{\Sigma}^E)\\ &= \mathcal{L}(B,\hat{\Sigma}^1,\ldots,\hat{\Sigma}^E) + \text{pen}_{\text{init},\lambda_{\text{init}}}(B) \quad (3.4)\end{aligned}$$

The initialization estimator is then given by

$$\hat{B}_{\text{init}} \in \operatorname*{argmin}_{B\in\mathcal{B}_0} \mathcal{T}_{\text{init}}(B). \quad (3.5)$$

Note that this is not a convex optimization problem due to the fact that B enters the log-likelihood term quadratically and the penalty term linearly, which means it might be hard to solve in general. However, we do give a local optimization algorithm in Section 4 that attempts to find a local minimum for (3.4).

In the second step, this estimator is refined by employing a different regularization term,

$$\mathrm{pen}_{\mathrm{loc}}(B) = \mathrm{pen}_{\mathrm{loc},\lambda_{\mathrm{loc}}}(B) = \lambda_{\mathrm{loc}}\|B\|_1,$$

$$\begin{aligned}\mathcal{T}_{\mathrm{loc}}(B) &= \mathcal{T}_{\mathrm{loc},\lambda_{\mathrm{loc}}}(B, \hat\Sigma^1, \dots, \hat\Sigma^E)\\ &= \mathcal{L}(B, \hat\Sigma^1, \dots, \hat\Sigma^E) + \mathrm{pen}_{\mathrm{loc},\lambda_{\mathrm{loc}}}(B), \end{aligned} \tag{3.6}$$

and the estimator is given by

$$\hat B_{\mathrm{loc}} \in \underset{\substack{B\in\mathcal{B}_0\\ \|B-\hat B_{\mathrm{init}}\|_F \le R_{\mathrm{loc}}}}{\operatorname{argmin}} \mathcal{T}_{\mathrm{loc}}(B), \tag{3.7}$$

with a suitably chosen localization parameter $R_{\mathrm{loc}} > 0$.

The loss function (3.6) is again non-convex and hence hard to optimize, but local optimization algorithms seem to produce good results, see Section 4.

3.3.2 Statistical performance

Assuming we have access to the global minima $\hat B_{\mathrm{init}}$ and $\hat B_{\mathrm{loc}}$, we show the following rates for $\hat B_{\mathrm{loc}}$:

Theorem 7. *Under assumptions A1 – A3, if*

$$n \gtrsim \left(E^2 \vee \frac{1}{\eta^4} \vee p^2\right) \frac{p^2(d+1)^2 E^3}{\eta^4} \log(epE/\delta)$$

and the parameters for the estimators $\hat B_{\mathrm{init}}$ and $\hat B_{\mathrm{loc}}$ are chosen such that

$$\begin{aligned} R_{\mathrm{loc}} &\asymp \frac{1}{\sqrt{E}} \wedge \eta \wedge \frac{1}{\sqrt{p}},\\ \lambda_{\mathrm{init}} &\asymp \sqrt{\frac{E\log(epE/\delta)}{n}}, \quad \textit{and}\\ \lambda_{\mathrm{loc}} &\asymp \sqrt{\frac{E^2\log(epE/\delta)}{n}} \end{aligned}$$

then

$$\|\hat B_{\mathrm{loc}} - B^*\|_F^2 \lesssim \frac{p(d+1)E^2}{\eta^8\, n} \log(pE/\delta), \tag{3.8}$$

with probability at least $1-\delta$.

The proof is deferred to Section E of the supplement. It is based on the one hand on restricted convexity properties of the Gaussian log-likelihood function that were developed in the context of convex optimization problems for estimation of sparse concentration matrices in Rothman et al. (2008) and Loh and Wainwright (2013), see also Negahban et al. (2012), and on the other to new structural results on the difference $\Theta^e(B) - \Theta^{*,e}$ between concentration matrices expressed in terms of $B - B^*$; see Lemma 10.

Note that the upper bound (3.8) is worse by a factor of E and a log factor than the lower bound (3.1) in Theorem 3. However, the completely separating system $\mathcal{E}$ can be chosen to be as small as $E \asymp \log(p)$, see Hyttinen et al. (2013) and Remark 2, in which case this eventual rate is almost minimax optimal up to logarithmic terms.

We also note that the requirement on n in Theorem 7 of $n \gtrsim (p^2 \vee E^2)p^2 d^2 E^3 \log(Ep)$ is much larger than the regime at which (3.8) becomes less than 1, $n \gtrsim pdE^2\log(pE)$. It is unclear whether these are due to inefficiencies in our proof technique or shortcomings of the particular estimator in question.

4 NUMERICAL EXPERIMENTS

Recall that we want to find solutions to the two regularized maximum likelihood problems (3.5) and (3.7). Both problems are non-convex and there is no obvious strategy for how to find global minima. However, since they are continuous, we can empirically study the performance of optimization algorithms designed for convex problems, hoping to obtain at least local minima. In Sections A.1 and A.2 of the supplement, we describe how candidate solutions for both (3.5) and (3.7) can be found efficiently by using a nonlinear version of the Alternating Direction Method of Multipliers (ADMM) (Gabay and Mercier, 1976; Glowinski and Marroco, 1975; Boyd et al., 2011) and an augmented Lagrangian method (Nocedal and Wright, 2006), respectively.

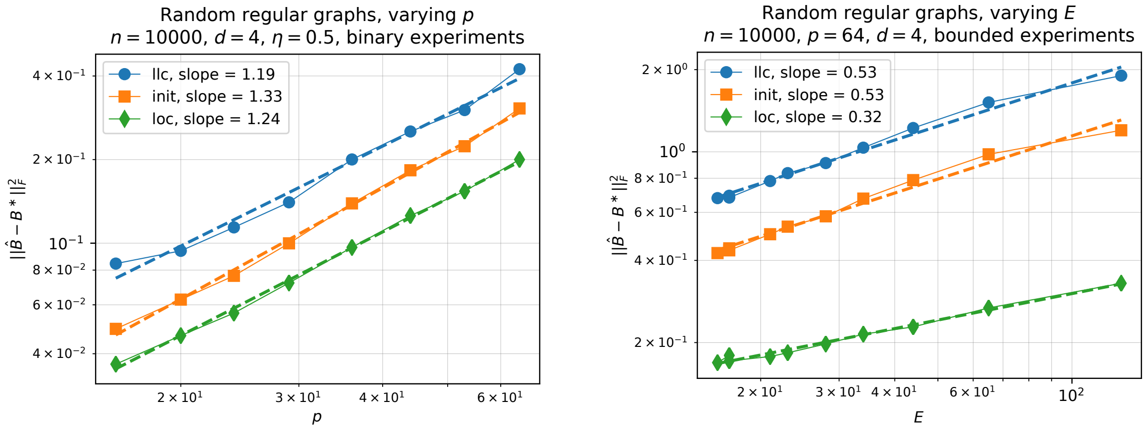
Here, we report results from experiments with synthetic data generated using (directed) random regular graphs to gauge the performance of the maximum likelihood procedure, comparing it to the LLC algorithm (Hyttinen et al., 2012). Further details on the experiments and more experiments on synthetic and semi-synthetic data involving graphs comprised of disconnected cliques and a small gene regulatory network from Cai et al. (2013) can be found in the full version of the Numerical Experiments, Section A of the supplement.

4.1 EXPERIMENTAL SETUP

Data generation The ground truth graphs are generated by first obtaining the (directed) adjacency matrix $B_{\mathrm{adj}} \in \{0,1\}^{p\times p}$, a matrix $B_{\mathrm{val}} \in \mathbb{R}^{p\times p}$ containing edge values, and finally setting B^* to be the Hadamard product of the two, normalized to have operator norm $1-\eta = 0.5$,

$$\tilde B = B_{\mathrm{adj}} \odot B_{\mathrm{val}}, \quad B^* = \frac{(1-\eta)}{\|\tilde B\|_{\mathrm{op}}}\tilde B.$$

Here, B_{val} consists of independent standard Gaussian entries, and B_{adj} is the adjacency matrix of a regular



(a) Varying p

(b) Varying E

Figure 1: Experiments for random regular graphs, varying one parameter while keeping the other ones fixed. "llc" refers to $\hat{B}_{\text{llc}}$, "init" to $\hat{B}_{\text{init}}$, "loc" to $\hat{B}_{\text{loc}}$.

random graph, where $\operatorname{supp}((B_{\text{adj}})_{i,:})$ is constructed by sampling d times uniformly at random without replacement from $\{1, \ldots, p\} \setminus \{i\}$ and all elements in the support are assigned the value 1.

Choice of λ: To keep the comparison simple, we use an oracle choice of λ_{init}, λ_{loc} and R_{loc}. For the first two, this means choosing them such that $\|\hat{B}_{\text{init}} - B^*\|_F$ and $\|\hat{B}_{\text{loc}} - B^*\|_F$ is minimal. For R_{loc}, we choose $R_{\text{loc}} = 2\|\hat{B}_{\text{init}} - B^*\|_F$. In practice, both parameters could be chosen by cross-validation.

Initialization of optimization algorithm: We initialize the calculation of $\hat{B}_{\text{init}}$ for the largest value of λ_{init} with the all zeros matrix and then warm-start the calculation with the output of the calculation for the next larger value of λ_{init}. The calculation of $\hat{B}_{\text{loc}}$ is initialized with the output of $\hat{B}_{\text{init}}$. We further investigate the dependence on the initialization value in the full version of the numerical experiments, Section A of the supplement.

Systems of interventions: We consider two choices for the experiments $\mathcal{E}$. The first one, which we call *binary*, consists of separating the nodes with a bisection approach similar to the construction given in Dickson (1969) that leads to $E = O(\log p)$. The second one, which we call *bounded*, is given by Cai (1984) and produces experiments whose sizes $|\mathcal{J}_e|$ are bounded by k. In this case, $E = O(n/k)$.

Repetitions: All errors are averaged over 32 random repetitions of sampling B^* and the observations X_k^e.

4.2 RESULTS

In Figure 1, we collect comparisons for the estimation rates of $\hat{B}_{\text{llc}}$, $\hat{B}_{\text{init}}$, and $\hat{B}_{\text{loc}}$, varying p and E, respectively, where in the varying p case, we consider binary experiments. The varying E case is given by bounded experiments with a varying bound on the size k of the experiments which, of course, governs the total number E of experiments needed for separation. In all cases, we performed linear regression on the log-transformed values to arrive at an estimate of the polynomial dependence of the error rate on the parameters, indicated by a dashed line.

In Figure 1(a), we observe a scaling with respect to p that is slightly worse than guaranteed by our theorems and could be due to the presence of log factors. In Figure 1(b), we observe that the scaling with respect to E when increasing the number of experiments appears to be better than predicted by our theory: about $E^{1/2}$ for $\hat{B}_{\text{llc}}$ and $\hat{B}_{\text{init}}$, about $E^{1/3}$ for $\hat{B}_{\text{loc}}$.

Further experiments with varying n and d are reported in Figure 2 of Section A of the supplement.

Acknowledgments

Philippe Rigollet was supported by NSF awards IIS-1838071, DMS-1712596, DMS-TRIPODS- 1740751, and ONR grant N00014-17- 1-2147.

References

Abrahamsen, N. and Rigollet, P. (2018). Sparse Gaussian ICA. *arXiv preprint arXiv:1804.00408*.

Bellec, P. C., Lecué, G., and Tsybakov, A. B. (2018). Slope meets lasso: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642.

Bello, K. and Honorio, J. (2018). Computationally and statistically efficient learning of causal Bayes nets using path queries. In *Advances in Neural Information Processing Systems*, pages 10931–10941.

Bielby, W. T. and Hauser, R. M. (1977). Structural equation models. *Annual review of sociology*, 3(1):137–161.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, Ltd.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Cai, M. (1984). On a problem of Katona on minimal completely separating systems with restrictions. *Discrete Mathematics*, 48(1):121–123.

Cai, X., Bazerque, J. A., and Giannakis, G. B. (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology*, 9(5):e1003068.

Chickering, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498.

Dickson, T. J. (1969). On a problem concerning separating systems of a finite set. *Journal of Combinatorial Theory*, 7(3):191–196.

Duncan, O. D. (1966). Path analysis: Sociological examples. *American journal of Sociology*, 72(1):1–16.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40.

Ghoshal, A. and Honorio, J. (2017). Information-theoretic limits of Bayesian network structure learning. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 767–775, Fort Lauderdale, FL, USA. PMLR.

Ghoshal, A. and Honorio, J. (2018). Learning linear structural equation models in polynomial time and sample complexity. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76.

Hauser, A. and Bühlmann, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2012). Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(Nov):3387–3439.

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2013). Experiment selection for causal discovery. *The Journal of Machine Learning Research*, 14(1):3041–3071.

Itani, S., Ohannessian, M., Sachs, K., Nolan, G. P., and Dahleh, M. A. (2010). Structure learning in causal cyclic networks. In *Causality: Objectives and Assessment*, pages 165–176.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636.

Keats, B. W. and Hitt, M. A. (1988). A causal model of linkages among environmental dimensions, macro organizational characteristics, and performance. *Academy of management journal*, 31(3):570–598.

Kocaoglu, M., Dimakis, A., and Vishwanath, S. (2017). Cost-optimal learning of causal graphs. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1875–1884.

Lacerda, G., Spirtes, P., Ramsey, J., and Hoyer, P. O. (2008). Discovering cyclic causal models by Independent Components Analysis. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pages 366–374.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105.

Loh, P.-L. and Wainwright, M. J. (2011). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734.

Loh, P.-L. and Wainwright, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484.

Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, 37(6A):3133–3164.

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

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, second edition.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, second edition.

Peters, J. and Bühlmann, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228.

Richardson, T. (1996). *Feedback Models: Interpretation and Discovery*. PhD thesis, Ph. D. thesis, Carnegie Mellon.

Richardson, T. and Spirtes, P. (1996). Automated discovery of linear feedback models. manuscript.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Schmidt, M. and Murphy, K. (2009). Modeling discrete interventional data using directed cyclic graphical models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 487–495. AUAI Press.

Schmidt, M., Niculescu-Mizil, A., and Murphy, K. (2007). Learning graphical model structure using L1-regularization paths. In *AAAI*, volume 7, pages 1278–1283.

Shanmugam, K., Kocaoglu, M., Dimakis, A. G., and Vishwanath, S. (2015). Learning Causal Graphs with Small Interventions. In *Advances in Neural Information Processing Systems*, pages 3195–3203.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

van de Geer, S. and Bühlmann, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.*, 41(2):536–567.

Wang, Y., Segarra, S., and Uhler, C. (2018). High-Dimensional Joint Estimation of Multiple Directed Gaussian Graphical Models. *arXiv preprint arXiv:1804.00778*.

Supplement to: Estimation Rates for Sparse Linear Cyclic Causal Models

APPENDIX A: NUMERICAL EXPERIMENTS, FULL VERSION

Recall that $\ell(\Theta, \hat{\Sigma}) = \mathsf{Tr}(\hat{\Sigma}\Theta) - \log\det(\Theta)$ and that we want to find solutions to the two regularized maximum likelihood problems,

$$\hat{B}_{\text{init}} \in \underset{B \in \mathcal{B}_0}{\operatorname{argmin}} \left\{ \sum_{e \in \mathcal{E}} \ell(\Theta^e(B), \hat{\Sigma}^e) + \lambda_{\text{init}} \sum_{e \in \mathcal{E}} \|\Theta^e(B)\|_1 \right\}, \tag{A.1}$$

$$\hat{B}_{\text{loc}} \in \underset{\substack{B \in \mathcal{B}_0 \\ \|B - \hat{B}_{\text{init}}\|_F \le R_{\text{loc}}}}{\operatorname{argmin}} \left\{ \sum_{e \in \mathcal{E}} \ell(\Theta^e(B), \hat{\Sigma}^e) + \lambda_{\text{loc}} \|B\|_1 \right\}. \tag{A.2}$$

Both problems are non-convex and there is no obvious strategy for how to find global minima. However, since they are continuous, we can empirically study the performance of optimization algorithms designed for convex problems, hoping to obtain at least local minima. In the following, we describe how candidate solutions for both (A.1) and (A.2) can be found efficiently and demonstrate their performance based on experiments with synthetic data. Additionally, we give a low-rank update approach in Appendix H that can be used to speed up calculations when the number of experiments E is large, but for each experiment, the number of controlled variables $|\mathcal{J}_e|$ is small.

A.1 Solving the initialization problem by non-convex ADMM

The difficulty in solving problem (A.1) is to handle the non-smooth penalty terms of non-linear transformations of B, $\|\Theta^e(B)\|_1$. We use a non-linear version of the Alternating Direction Method of Multipliers (ADMM) algorithm, which allows us to introduce additional variables Θ^e, constrain them to fulfill $\Theta^e = \Theta^e(B)$, and keep the resulting dimensionality blowup manageable.

The ADMM algorithm [GM76, GM75, BPC$^+$11] is a splitting algorithm intended to solve convex optimization problems of the form

$$\begin{gathered} \min f(x) + g(y) \\ \text{s. t. } Fx + Gy = b, \end{gathered}$$

where $x \in \mathbb{R}^m, y \in \mathbb{R}^\ell$, f and g are convex functions on $\mathbb{R}^m$ and $\mathbb{R}^\ell$, respectively, and $F \in \mathbb{R}^{m \times k}$, $G \in \mathbb{R}^{\ell \times k}$, $b \in \mathbb{R}^k$. Introducing the dual variable u, a step size $\rho > 0$, and starting with an initialization x^0, y^0, u^0, its iterations are given by

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}} f(x) + \frac{\rho}{2}\|Fx + Gy^k - b + u^k\|_2^2 \\ y^{k+1} &= \underset{y}{\operatorname{argmin}} g(y) + \frac{\rho}{2}\|Fx^{k+1} + Gy - b + u^k\|_2^2 \\ u^{k+1} &= u^k + Fx^{k+1} + Gy^{k+1} - b, \end{aligned}$$

which is the so-called *scaled form* of ADMM.

Note that while in the case of convex objective functions and linear constraints, there are well-established convergence results for ADMM, [Gab83, EB92], results about convergence to a stationary point for non-convex variants are scarce, requiring either linear constraints [WYZ19] or further modifications and additional assumptions [BKSV15].

In order to apply a non-convex ADMM variant, we rewrite problem (A.1) as

$$\begin{aligned}\min_{B\in\mathcal{B}_0}\sum_{e\in\mathcal{E}}&\left(\ell(\Theta^e,\hat{\Sigma}^e)+\lambda_{\text{init}}\|\Theta^e\|_1\right)\\ \text{s. t. }&\Theta^e=(I-U_eB)^\top(I-U_eB)\quad\text{for }e\in\mathcal{E}.\end{aligned}$$

Then, introducing dual variables $\Lambda^e\in\mathbb{R}^{p\times p}$, $e=1,\ldots,E$, the outer iteration of our algorithm is given by

$$\begin{aligned}\Theta^{e,k+1}&=\underset{\Theta^e}{\operatorname{argmin}}\,\mathsf{Tr}(\hat{\Sigma}^e\Theta^e)-\log\det\Theta^e+\lambda_{\text{init}}\|\Theta^e\|_1\\&\qquad+\frac{\rho}{2}\|\Theta^e-(I-U_eB^k)^\top(I-U_eB^k)+\Lambda^{e,k}\|_F^2,\quad(e=1,\ldots,E)\qquad\text{(A.3)}\\ B^{k+1}&=\underset{B}{\operatorname{argmin}}\sum_e\|\Theta^{e,k+1}-(I-U_eB)^\top(I-U_eB)+\Lambda^{e,k}\|_F^2\qquad\text{(A.4)}\\ \Lambda^{e,k+1}&=\Lambda^{e,k}+\Theta^{e,k+1}-(I-U_eB^{k+1})^\top(I-U_eB^{k+1}),\quad(e=1,\ldots,E).\end{aligned}$$

Note that (A.3) is a convex problem, resembling the graphical LASSO [FHT08] or SPICE [RBLZ08] but with an additional quadratic penalty term. We can solve these subproblems with an extension of the QUIC algorithm [HDRS11] that employs coordinate descent to iteratively find Newton directions.

Problem (A.4) on the other hand is a non-convex problem, albeit without constraints. Hence, we can use any local optimization algorithm. For our experiments, we choose L-BFGS [LN89] to perform this approximate minimization, yielding a stationary point of the objective function.

In order to find a suitable step size parameter ρ, we allow varying ρ^k and employ the dual-balancing strategy from [HYW00, WL01].

A.2 Solving local problem by Augmented Lagrangian Method

In order to find a local minimum of (A.2), we employ the Augmented Lagrangian Method [NW06] that transforms the inequality constraint into a box constraint and iteratively solves for the associated dual variable. It leads to the following iteration, where u is a slack variable for the ℓ_2 constraint and λ is the associated dual variable.

$$\begin{aligned}B^{k+1}&=\underset{B,\,u\le R_{\text{loc}}^2}{\operatorname{argmin}}\sum_{e\in\mathcal{E}}\ell(\Theta^e(B),\hat{\Sigma}^e)+\lambda_{\text{loc}}\|B\|_1+\frac{\rho}{2}\left(u^k-\|B-\hat{B}_1\|_2^2+\frac{\lambda^k}{\rho}\right)^2\qquad\text{(A.5)}\\ u^{k+1}&=u^k+\rho(u-\|B^{k+1}-\hat{B}_1\|_2^2)\end{aligned}$$

To solve (A.5), we use L-BFGS-B [ZBLN97], transforming the ℓ^1-regularization into a linear term with additional non-negativity constraints,

$$B=B_+-B_-,\quad B_+\ge0,\quad B_-\ge0,\quad\|B\|_1=\sum_{i,j}((B_+)_{ij}+(B_-)_{ij}).$$

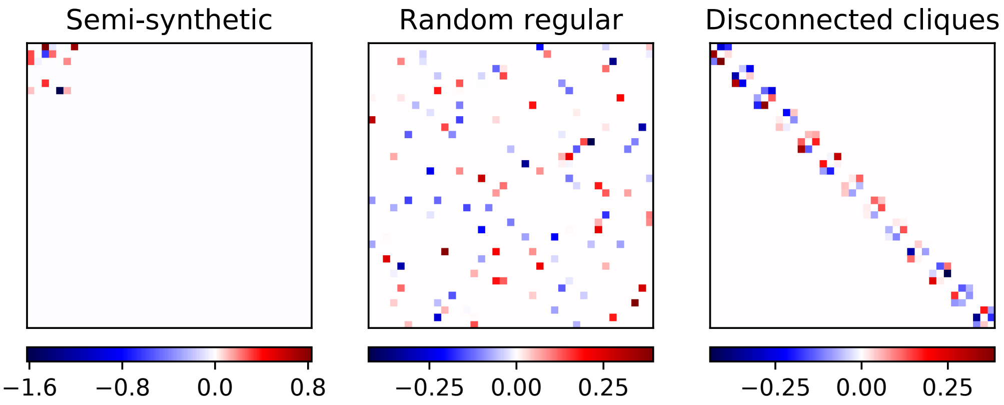


Figure 1: Heatmaps visualizing example matrices B^* for the three studied models. In all examples, $p = 39$ and $d = 3$. From left to right: Semi-synthetic data from [CBG13], Random regular graphs, and disconnected cliques.

A.3 Experimental setup

We perform experiments with synthetic and semi-synthetic data to gauge the performance of the maximum likelihood procedure, comparing it to the LLC algorithm [HEH12].

For the synthetic benchmarks, we study two types of graph structures: (directed) random regular graphs and graphs composed of disconnected cliques. For the semi-synthetic benchmarks, we use a gene-regulatory network from [CBG13] consisting of 39 genes. Note that in [CBG13], the authors employ a model very similar to ours, but instead of allowing controlled experiments on certain nodes, they consider so-called expression quantitative trait loci (eQTL) as proxies for interventions, which changes their model compared to the one considered here. Nonetheless, part of the output of their estimator is a linear causal network, which is what we consider as ground truth to simulate data following the Gaussian model introduced in Section 2. Example ground truth matrices for the two random models and the semi-synthetic matrix from [CBG13] are given in Figure 1. There, we set $p = 39$ and $d = 3$ for the random models to coincide with p and d in the semi-synthetic case.

In the following two sections, we give more details about data generation and parameter tuning.

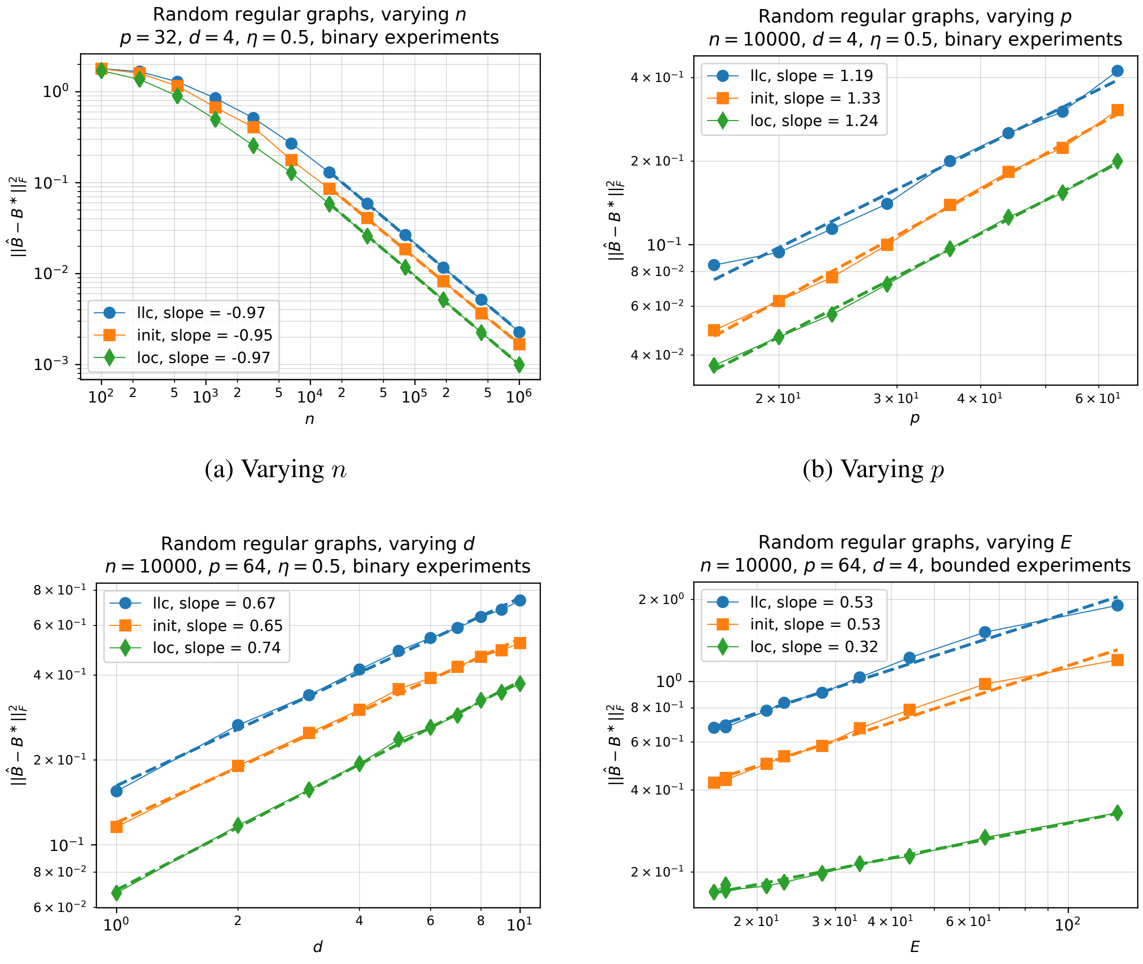
A.3.1 Models

Synthetic graphs: The ground truth graphs are generated by first obtaining the (directed) adjacency matrix $B_{\mathrm{adj}} \in \{0,1\}^{p\times p}$, a matrix $B_{\mathrm{val}} \in \mathbb{R}^{p\times p}$ containing edge values, and finally setting B^* to be the Hadamard product of the two, normalized to have operator norm $1 - \eta = 0.5$,

$$\tilde{B} = B_{\mathrm{adj}} \odot B_{\mathrm{val}}, \quad B^* = \frac{(1-\eta)}{\|\tilde{B}\|_{\mathrm{op}}} \tilde{B}.$$

Here, B_{val} consists of independent standard Gaussian entries, and B_{adj} is the adjacency matrix of either a regular random graph or one composed of disconnected cliques.

Random regular graphs: $\mathrm{supp}((B_{\mathrm{adj}})_{i,:})$ is constructed by sampling d times uniformly at random without replacement from $\{1,\ldots,p\} \setminus \{i\}$ and all elements in the support are assigned 1.



(a) Varying n

(b) Varying p

(c) Varying d

(d) Varying E

Figure 2: Experiments for random regular graphs, varying one parameter while keeping the other ones fixed. "llc" refers to $\hat{B}_{\text{llc}}$, "init" to $\hat{B}_{\text{init}}$, "loc" to $\hat{B}_{\text{loc}}$.

Disconnected cliques: B_{adj} is the adjacency matrix of a graph consisting of $\lfloor p/d \rfloor$ disconnected d-cliques and an additional disconnected $p - d\lfloor p/d \rfloor$ clique if d does not divide p. This model is meant to illustrated clustered variables that operate in modules, akin to the ones arising in gene regulatory networks.

A.3.2 Tuning parameters

Choice of λ: To keep the comparison simple, we use an oracle choice of λ_{init}, λ_{loc} and R_{loc}. For the first two, this means choosing them such that $\|\hat{B}_{\text{init}} - B^*\|_F$ and $\|\hat{B}_{\text{loc}} - B^*\|_F$ is minimal. For R_{loc}, we choose $R_{\text{loc}} = 2\|\hat{B}_{\text{init}} - B^*\|_F$. In practice, both parameters could be chosen by cross-validation.

Initialization of optimization algorithm: We initialize the calculation of $\hat{B}_{\text{init}}$ for the largest value of λ_{init} with the all zeros matrix and then warm-start the calculation with the output of the calculation for the next larger value of λ_{init}. The calculation of $\hat{B}_{\text{loc}}$ is initialized with the output of $\hat{B}_{\text{init}}$. To investigate the dependence on the initialization, we also try initializing the calculation of $\hat{B}_{\text{init}}$ with a strict triangular matrix whole upper elements consist of independent $\mathcal{N}(0, 10)$ random variables, as well as running the likelihood optimization without constraints as described previously with the same initialization.

Systems of interventions: We consider three choices for the experiments $\mathcal{E}$. The first one, which we call *binary*, consists of separating the nodes with a bisection approach similar to the construction given in [Dic69] that leads to $E = O(\log p)$. The second one, which we call *bounded*, is given by [Cai84] and produces experiments whose sizes $|\mathcal{J}_e|$ are bounded by k. In this case, $E = O(n/k)$. The third kind corresponds to $k = 1$, taking $\mathcal{J}_i = \{i\}$ for $i \in [p]$, which we call *single-node experiments*.

Repetitions: All errors are averaged over 32 random repetitions of sampling B^* and the observations X^e_k.

A.4 Results

In all cases, we performed linear regression on (a subset of) the log-transformed values to arrive at an estimate of the polynomial dependence of the error rate on the parameters, indicated by a dashed line.

A.4.1 Performance

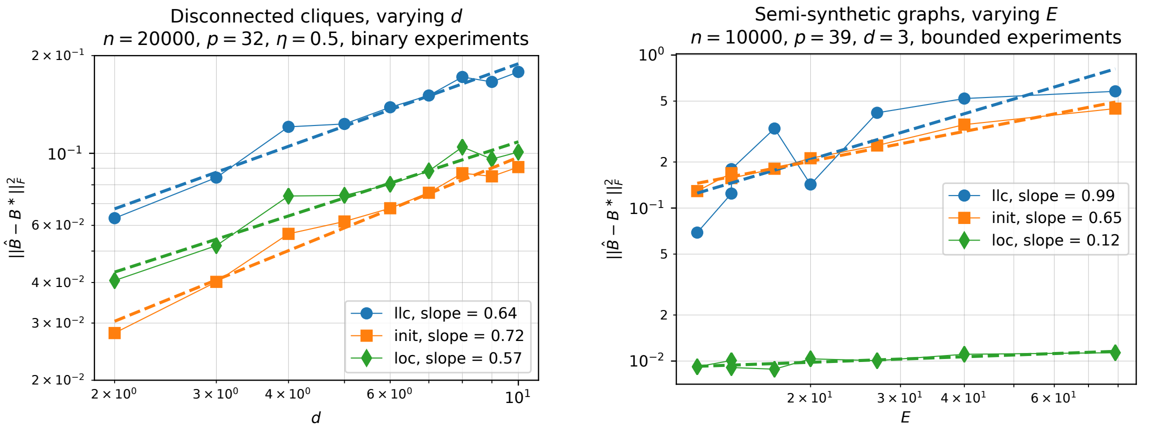
Random regular graphs with oracle choice: In Figure 2, we collect comparisons for the estimation rates of $\hat{B}_{\text{llc}}$, $\hat{B}_{\text{init}}$, and $\hat{B}_{\text{loc}}$, varying n, p, d, and E, respectively, where the varying E case is given by bounded experiments with a varying bound on the size k of the experiments which, of course, governs the total number E of experiments needed for separation. In all other cases, we consider binary experiments.

Figure 2(a) indicates that all three estimators exhibit a risk that scales as $1/n$ and displays a clear ordering in the performance of the three candidates where $\hat{B}_{\text{llc}}$ performs worse than $\hat{B}_{\text{init}}$, which in turn is worse than $\hat{B}_{\text{loc}}$. This corroborates the potentially sub-optimal dependence of LLC on the conditioning of the problem, see Remark 5, and agrees with what we expect given the upper bounds in (3.8) and (E.8) for $\hat{B}_{\text{loc}}$ and $\hat{B}_{\text{init}}$, respectively.

In Figure 2(b), we observe a scaling with respect to p that is slightly worse than guaranteed by our theorems and could be due to the presence of log factors. In Figure 2(c), we in turn see that the scaling with respect to d is slightly better than expected, hinting at good adaptation to the sparsity parameter d. Most interestingly, in Figure 2(d), we observe that the scaling with respect to E when increasing the number of experiments appears to be better than predicted by our theory: about $E^{1/2}$ for $\hat{B}_{\text{llc}}$ and $\hat{B}_{\text{init}}$, about $E^{1/3}$ for $\hat{B}_{\text{loc}}$. This different behavior is even more striking in Figure 3(b) where the performance of $\hat{B}_{\text{loc}}$ appears to decay at most logarithmically in E.

Disconnected clique graphs: In Figure 3(a), we plot the same experiment as in Figure 2(a), only this time with disconnected clusters instead of random regular graphs. We notice a similar behavior, with the key difference of the performance of $\hat{B}_{\text{init}}$ surpassing that of $\hat{B}_{\text{loc}}$. This could be explained by the fact that the penalization in the objective $\hat{B}_{\text{init}}$ is particularly suited for the estimation of this kind of graphs since the sparsity of $(I - B^*)^\top (I - B^*)$ in this case almost coincides with the one of B^*, which can be seen from the argument that led to (E.6) in the proof of Theorem 7.

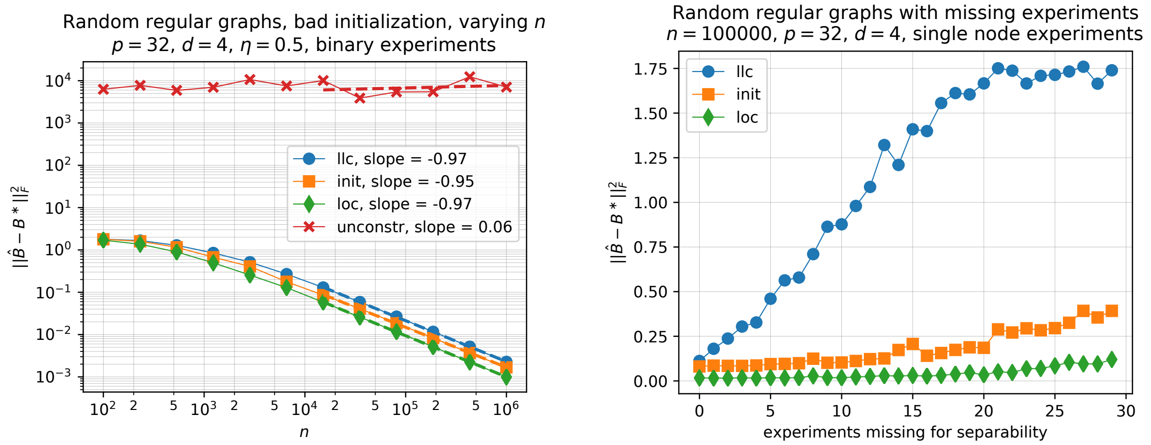
Semi-synthetic graph: The performance of the three estimators on the semi-synthetic data built from the graph taken in [CBG13] appears in Figure 3(b). The LLC estimator $\hat{B}_{\text{llc}}$ performs similarly to $\hat{B}_{\text{init}}$ and both suffer in comparison to $\hat{B}_{\text{loc}}$ either in terms of absolute performance and in terms of scaling with E.



(a) Varying d for disconnected cliques graph.

(b) Varying E for semi-synthetic graph based on [CBG13].

Figure 3: Experiments for other types of graphs



(a) Same case as Figure 2(a), but with bad initialization. "unconstr" refers to the case where $R_{\text{loc}} = \infty$ and illustrates the need for localization.

(b) Random regular graphs case where Definiton 1 is violated.

Figure 4: Additional computational experiments, running the algorithm with bad initializations and with experiments not satisfying complete separability.

A.4.2 Stability

Role of initialization: In Figure 4(a), we show the same setup as in Figure 2(a), only this time, the calculation for $\hat{B}_{\text{init}}$ is initialized with a random matrix as outlined in Section A.3. Additionally, we plot the result of optimizing an unconstrained version of $\hat{B}_{\text{loc}}$ with the same bad initialization, denoted by $\hat{B}_{\text{unconstr}}$. We observe that the performance of the latter is very bad due to the non-convex nature of the objective together with the fact that a bad initialization point is chosen. However, even though $\hat{B}_{\text{init}}$ is found through solving a non-convex objective as well, it seems to be robust enough to yield comparable performance and hence serve as a good initialization for calculating $\hat{B}_{\text{loc}}$ even with a poor initial choice of B.

Missing experiments: In Figure 4(b) we investigate the robustness to systems of interventions that do not fulfill the separability condition in Definition 1. For this, we consider single-node experiments and plot the number of experiments that are missing from a completely separating set of such experiments (in which case we would have $E = p$). The likelihood-based approaches are much more robust in this case, and to a larger degree than the degree-of-freedom calculations as in Appendix G would suggest.

APPENDIX B: EXTENDED NOTATION

Here, we recapitulate and extend the notation used in the main paper.

We write $a \lesssim b$ for two quantities a and b if there exists an absolute constant $C > 0$ such that $a \leq Cb$, and similarly for $a \gtrsim b$. For two real numbers $a, b \in \mathbb{R}$, we write $a \wedge b$ for their minimum and $a \vee b$ their maximum, respectively. For a natural number p, we denote by $[p] = \{1, \dots, p\}$. Given a set S, we write $|S|$ for its cardinality.

Let $x, y \in \mathbb{R}^p$. We write $\operatorname{supp} x$ for the indices of non-zero elements of x,

$$d_H(x, y) = |\{i \in [p] : x_i \neq y_i\}|$$

for the Hamming distance between x and y, and $\|x\|_p$ for the ℓ^p norm of x.

For two matrices $A, B \in \mathbb{R}^{p_1 \times p_2}$, we abbreviate the ith row by $B_{i,:}$ and the ith column by $B_{:,i}$. Similarly, $B_{i,-j}$ denotes the ith row of B where the jth element is omitted. Further, $\|B\|_F$ denotes the Frobenius norm, $\|B\|_{\mathrm{op}}$ the operator norm,

$$\|B\|_\infty = \max_{i,j} |B_{i,j}|, \quad \|B\|_1 = \sum_{i,j} |B_{i,j}|,$$

and $\|B\|_{\infty,\infty}$ the operator norm of B with respect to the ℓ^∞ norm, which is

$$\|B\|_{\infty,\infty} = \max_{i \in [p_1]} \|B_{i,:}\|_1.$$

If A is a square invertible matrix, we denote by A^{-1} its inverse and by $A^{-\top}$ the transpose of A^{-1}. We denote the smallest and largest singular value of A by $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$, respectively. If A and B are symmetric, we write $A \prec B$ if $B - A$ is positive definite, and similarly for $A \succ B$. By $I \in \mathbb{R}^{p \times p}$, we denote the identity matrix.

For a function $f : \mathbb{R}^{p_1} \to \mathbb{R}^{p_2}$, we denote its derivative at a point $x \in \mathbb{R}^{p_1}$ applied to a vector $h \in \mathbb{R}^{p_1}$ by $Df(x)[h]$. We write subG and subE to denote sub-Gaussian and sub-Exponential distributions as defined in Definition 16.

APPENDIX C: PROOF OF LOWER BOUNDS

C.1 Proof of Theorem 3

To begin, recall the definition of the redundancy factor

$$\kappa = \kappa(\mathcal{E}) = \max_{(i,j)} |\{e \in \mathcal{E} : (i,j) \text{ separated in } e\}|.$$

The proof of Theorem 3 is based on standard techniques for minimax lower bounds [Tsy09].

THEOREM 1 ([Tsy09, Theorem 2.5]). *Denote by $\mathcal{G} \subseteq \mathbb{R}^{p \times p}$ a set of possible hypotheses with associated probablity measures P_B for $B \in \mathcal{G}$.*

Fix $M \geq 2, s > 0, \alpha \in (0, 1/8)$, and assume that there exists $B_0, \dots, B_M \in \mathcal{G}$, such that *(i)* $\|B_j - B_k\|_F \geq 2s > 0$ *for all* $0 \leq j < k \leq M$*;*

(ii) $\mathsf{KL}(P_j|P_0) \leq \alpha \log M$ *for all* $j = 1, \ldots, M$*, where* $P_j = P_{B_j}$*.*

Then,

$$\inf_{\hat{B}} \sup_{B^* \in \mathcal{G}} P_B(\|\hat{B} - B^*\|_F \geq s) \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right), \tag{C.1}$$

where the infimum in (C.1) *is taken over all measurable functions* $\hat{B}$ *on the observations.*

Before proceeding with the proof, we first present two lemmas. Lemma 2 gives a way to upper bound the Kullback-Leibler divergence between two Gaussian distributions in terms of their concentration matrices, while Lemma 3 contains a version of the Varshamov-Gilbert lemma adopted to produce candidate matrices with the same row sparsity. Their proofs can be found in Section C.2 and Section C.3, respectively.

In the following, we denote by $d_H(A, B)$ the Hamming distance between two matrices $A, B \in \mathbb{R}^{p \times p}$. It is defined by $d_H(A, B) = |\{(i, j) \in [p]^2 : A_{i,j} \neq B_{i,j}\}|$.

LEMMA 2. *Let* $\Theta_1, \Theta_2 \in \mathbb{R}^{p \times p}$ *be two positive definite concentration matrices and* $P_1 = \mathcal{N}(0, \Theta_1^{-1})$ *and* $P_2 = \mathcal{N}(0, \Theta_2^{-1})$ *the associated Gaussian distributions. If*

$$\|\Theta_1 - \Theta_2\|_{\mathrm{op}} \leq \frac{\lambda_{\min}(\Theta_2)}{2},$$

then,

$$\mathsf{KL}(P_1|P_2) \leq \frac{1}{\lambda_{\min}(\Theta_2)^2} \|\Theta_1 - \Theta_2\|_F^2. \tag{C.2}$$

LEMMA 3. *Given* $m \geq 1$ *and* $1 \leq d \leq m/2$*, there is a family* $H_1, \ldots, H_M$ *of matrices in* $\{0, 1\}^{m \times m}$ *such that*

(i) *Every row of* H_i *is* d*-sparse for* $i = 1, \ldots, M$*;*

(ii) $d_H(H_i, H_j) \geq \dfrac{md}{2}, \quad i \neq j$*;*

(iii) $\log M \geq \dfrac{md}{16} \log\left(1 + \dfrac{m}{2d}\right)$.

Taking the above lemmas as given, we proceed to prove Theorem 3.

We apply Theorem 1 by constructing an appropriate set of hypotheses $B_0, \ldots, B_M$. Without loss of generality, assume that p is even. Set $B_0 = 0$ to be the all zeros matrix, and apply Lemma 3 with $m = p/2$ to obtain M matrices $H_1, \ldots, H_M \in \{0, 1\}^{p/2 \times p/2}$ with pairwise Hamming distance at least $pd/4$ and with

$$\log M \geq \frac{pd}{32} \log\left(1 + \frac{p}{4d}\right).$$

We define B_i, $i = 1, \ldots, M$ as block matrices by setting

$$\gamma = \sqrt{\frac{\alpha}{64\kappa}}, \quad \beta = \gamma \sqrt{\frac{E}{n} \log\left(1 + \frac{p}{4d}\right)}, \quad \text{and} \quad B_i = \begin{bmatrix} 0 & \beta H_i \\ -\beta H_i^\top & 0 \end{bmatrix}.$$

By construction, for every $i \in [M]$, every row of B_i is d-sparse, and B_i has zero-diagonal. Moreover, by $\kappa \geq 1$, $\alpha < 1/8$, and by assumption

$$n \geq pdE^2 \log\left(1 + \frac{p}{4d}\right) \geq d^2 E \log\left(1 + \frac{p}{4d}\right),$$

so we get from the Gershgorin circle theorem that

$$\|B_i\|_{\mathrm{op}} \le \gamma d \sqrt{\frac{E}{n} \log\left(1+\frac{p}{4d}\right)} \le \gamma \le \frac{1}{5} \le 1-\eta$$

in light of $\eta \le 1/2$. Hence, $B_i \in \mathcal{B}(p,d,\eta)$ for all $i \in [M]$.

Next, we can lower bound the pairwise distances by

$$\begin{aligned}\|B_i - B_j\|_F^2 &\ge 2\beta^2 d_H(H_i,H_j) \ge \gamma^2 \frac{pdE}{4n} \log\left(1+\frac{p}{4d}\right),\\ \|B_i - B_0\|_F^2 &\ge \gamma^2 \frac{pd}{2n} \log\left(1+\frac{p}{4d}\right) \ge \gamma^2 \frac{pdE}{4n} \log\left(1+\frac{p}{4d}\right),\end{aligned}$$

which yields the needed separation in Theorem 1(i).

We proceed to estimate the KL divergence between two distributions corresponding to matrices B_0 and any B_i, $i \ge 1$. Decompose the difference between the concentration matrices as

$$\begin{aligned}\Theta^e(B_i) - \Theta^e(B_0) &= (I - U_e B_i)^\top (I - U_e B_i) - I &\text{(C.3)}\\ &= (U_e B_i)^\top + U_e B_i + B_i U_e B_i. &\text{(C.4)}\end{aligned}$$

Because $\|B_i\|_{\mathrm{op}} \le 1/5$, (C.3) together with $\|U_e\|_{\mathrm{op}} \le 1$ and the sub-multiplicativity of the operator norm implies

$$\|\Theta^e(B_2) - \Theta^e(B_1)\|_{\mathrm{op}} \le \frac{1}{5} + \frac{1}{5} + \frac{1}{25} \le \frac{1}{2} = \frac{\lambda_{\min}(I)}{2} = \frac{\lambda_{\min}(\Theta^e(B_0))}{2},$$

so the hypothesis of Lemma 2 is satisfied for all pairs $(\Theta^e(B_i), \Theta^e(B_0))$. By Lemma 2, (C.4), and the tensorization property of the KL divergence, we obtain

$$\begin{aligned}\mathsf{KL}(P_{B_i}|P_{B_0}) &\le \sum_e n_e \|\Theta^e(B_i) - \Theta^e(B_0)\|_F^2 \le \frac{2n}{E} \sum_e \|U_e B_i + (U_e B_i)^\top\|_F^2 + 2n\|B_i\|_F^4\\ &\le \frac{2n}{E} \sum_e \left[2 \sum_{\substack{k\in\mathcal{U}_e\\ \ell\in\mathcal{J}_e}} (B_i)_{k,\ell}^2 + \sum_{k,\ell\in\mathcal{U}_e} ((B_i)_{k,\ell} + (B_i)_{k,\ell})^2\right] + 2n\|B_i\|_F^4. &\text{(C.5)}\end{aligned}$$

Since B_i is defined to be anti-symmetric, we have

$$\sum_{k,\ell\in\mathcal{U}_e} ((B_i)_{k,\ell} + (B_i)_{k,\ell})^2 = 0, \quad i = 1,\dots,M,\ e = 1,\dots,\mathcal{E}. \tag{C.6}$$

Moreover,

$$\sum_e \sum_{\substack{k\in\mathcal{U}_e\\ l\in\mathcal{J}_e}} (B_i)_{k,\ell}^2 \le \kappa(\mathcal{E}) \|B_i\|_F^2. \tag{C.7}$$

Combining (C.5), (C.6) and (C.7), we arrive at

$$\mathsf{KL}(P_{B_i}|P_{B_0}) \le \frac{4n\kappa}{E} \|B_i\|_F^2 + 2n\|B_i\|_F^4.$$

It remains to compute the Frobenius norm of each B_i,

$$\|B_i\|_F^2 = \gamma^2 \frac{pdE}{4n} \log\left(1+\frac{p}{4d}\right).$$

Hence, because

$$n \geq pdE^2 \log\left(1 + \frac{p}{2d}\right), \quad \gamma < 1, \quad \text{and} \quad \gamma^2 = \frac{\alpha}{32\kappa},$$

we obtain

$$\begin{aligned}\mathsf{KL}(P_j|P_0) &\leq \frac{4n\kappa}{E}\gamma^2\frac{pdE}{4n}\log\left(1+\frac{p}{4d}\right) + 2n\gamma^4\frac{p^2d^2E^2}{16n^2}\left(\log\left(1+\frac{p}{4d}\right)\right)^2\\ &= \kappa pd\gamma^2\log\left(1+\frac{p}{4d}\right) + \gamma^4\frac{p^2d^2E^2}{8n}\left(\log\left(1+\frac{p}{4d}\right)\right)^2\\ &\leq 2\kappa pd\gamma^2\log\left(1+\frac{p}{4d}\right)\\ &\leq \alpha\frac{pd}{32}\log\left(1+\frac{p}{4d}\right) = \alpha\log M.\end{aligned}$$

Finally, we can pick $\alpha = \frac{1}{16}$ in Theorem 1 to conclude that

$$\inf_{\hat{B}} \sup_{B^* \in \mathcal{G}} P_B\left(\|\hat{B} - B^*\|_F \geq \frac{1}{2^{14}\kappa}\frac{pdE}{n}\log\left(1+\frac{p}{4d}\right)\right) \geq c,$$

for some constant $c > 0$.

C.2 Proof of Lemma 2

The Kullback-Leibler divergence between two Gaussians $P_1 = \mathcal{N}(0, \Theta_1^{-1})$ and $P_2 = \mathcal{N}(0, \Theta_2^{-1})$ is given by

$$\mathsf{KL}(P_1|P_2) = \frac{1}{2}\left(\mathsf{Tr}(\Theta_1^{-1}(\Theta_2 - \Theta_1) - \log\det\Theta_2 + \log\det\Theta_1\right).$$

Using the fact that the first derivative of $\Theta \mapsto -\log\det\Theta$ is $-\Theta^{-1}$ and the second derivative is $\Theta^{-1}\otimes\Theta^{-1}$, we employ a Taylor expansion about Θ_1 (compare (E.10) in proof of Lemma 9) to obtain

$$\mathsf{KL}(P_1|P_2) = \frac{1}{4}\mathsf{Tr}(\tilde{\Theta}^{-1}(\Theta_2 - \Theta_1)\tilde{\Theta}^{-1}(\Theta_2 - \Theta_1)),$$

for some $\tilde{\Theta} = t\Theta_1 + (1-t)\Theta_2$, $t \in [0,1]$. By considering the square root of $\tilde{\Theta}^{-1}$, this can be expressed in terms of the Frobenius norm of the difference $\Theta_2 - \Theta_1$,

$$\frac{1}{4}\mathsf{Tr}(\tilde{\Theta}^{-1}(\Theta_2 - \Theta_1)\tilde{\Theta}^{-1}(\Theta_2 - \Theta_1)) = \frac{1}{4}\|\tilde{\Theta}^{-1/2}(\Theta_2 - \Theta_1)\tilde{\Theta}^{-1/2}\|_F^2 \leq \frac{1}{4\lambda_{\min}(\tilde{\Theta})^2}\|\Theta_2 - \Theta_1\|_F^2. \quad \text{(C.8)}$$

By Weyl's inequality, [Fra12, Section 6.7, Theorem 2],

$$|\lambda_{\min}(\Theta_2) - \lambda_{\min}(\tilde{\Theta})| \leq \|\Theta_2 - \tilde{\Theta}\|_{\mathrm{op}} \leq \|\Theta_1 - \Theta_2\|_{\mathrm{op}}.$$

Hence, if $\|\Theta_2 - \Theta_1\|_{\mathrm{op}} \leq \lambda_{\min}(\Theta_2)/2$, then

$$\frac{1}{\lambda_{\min}(\tilde{\Theta})^2} \leq \frac{4}{\lambda_{\min}(\Theta_2)^2},$$

which together with (C.8) implies (C.2).

C.3 Proof of Lemma 3

We use the probabilistic method to show the existence of the family $H_1, \dots, H_M$, modifying a standard argument that can be found in [Tsy09, Lemma 2.9].

Let $H_1, \dots, H_M$ be M independent random matrices H_k, where each row of H_k is a zero-one-vector corresponding to a subset of $[p]$ with cardinality d drawn uniformly at random. More precisely, for the ith row of the matrix H_k, draw U_1^i uniformly from $\{1, \dots, m\}$, and U_j^i conditioned on $U_1^i, \dots, U_{j-1}^i$ uniformly from the set $\{1, \dots, m\} \setminus \{U_1^i, \dots, U_{i-1}^i\}$, $j = 2, \dots, d$. Then, set

$$(H_k)_{i,j} = \begin{cases} 1, & j \in \{U_1^i, \dots, U_d^i\} \\ 0, & \text{otherwise.} \end{cases}$$

By a union bound, the probability that there exists a pair k, ℓ for which $d_H(H_k, H_\ell) \leq md/2$ can be bounded by the probability of this occurring for one draw of H_1, comparing to a fixed H_0 with d-sparse rows, say $(H_0)_{kl} = \mathbb{1}_{\{k \leq d\}}$,

$$P\left(\exists \ell \neq k : d_H(H_\ell, H_k) < \frac{md}{2}\right) \leq \binom{M}{2} P\left(d_H(H_1, H_0) < \frac{md}{2}\right), \tag{C.9}$$

because for each row, every d sparse pattern is equally likely.

We can lower bound the Hamming distance by the number of elements in $\operatorname{supp}(H_0)$ on which H_1 is one, *i.e.*,

$$d_H(H_1, H_0) \geq md - \sum_{i=1}^{m} \sum_{j=1}^{d} Z_{i,j}$$

with $Z_{i,j} = \mathbb{1}(U_j^i \leq d)$. Then, $Z_{i,j} \sim \mathsf{Bern}(q_{i,j})$ with $q_{i,1} = \frac{d}{m}$ and, noting that $d \leq m/2$,

$$q_{i,j} = \frac{d - \sum_{\ell=1}^{j-1} Z_{i,j}}{m - (j-1)} \leq \frac{d}{m-d} \leq \frac{2d}{m}, \quad j \geq 2.$$

From there, apply a Chernoff bound, that is, pick $\lambda > 0$ and estimate

$$\begin{aligned} P\left(d_H(H_1, H_0) < \frac{md}{2}\right) &\leq P\left(\sum_{i=1}^{m} \sum_{j=1}^{d} Z_{i,j} \geq \frac{md}{2}\right) \\ &\leq \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{m} \sum_{j=1}^{d} Z_{i,j}\right)\right] \exp\left(-\lambda \frac{md}{2}\right) \end{aligned} \tag{C.10}$$

For a Bernoulli distribution $Z \sim \mathsf{Bern}(q)$, the moment generating function is given by

$$\mathbb{E}\left[\exp(\lambda Z)\right] = q(\exp(\lambda) - 1) + 1.$$

Together with the observation that $Z_{i,j}$ is stochastically dominated by a $\mathsf{Bern}(2d/m)$ distribution, we can estimate

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{m} \sum_{j=1}^{d} Z_{i,j}\right)\right] \leq \left(\frac{2d}{m}(\exp(\lambda) - 1) + 1\right)^{md}.$$

Setting $\lambda = \log\left(1 + \frac{m}{2d}\right)$, we get that

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{m}\sum_{j=1}^{d} Z_{i,j}\right)\right] \leq 2^{md}. \tag{C.11}$$

Combining (C.11), (C.10), (C.9), and the estimate $\binom{M}{2} \leq M^2$, we obtain

$$\begin{aligned} P\left(\exists \ell \neq k : d_H(H_\ell, H_k) \leq \frac{md}{2}\right) &\leq \exp\left(2\log M + md\log 2 - \frac{md}{2}\log\left(1 + \frac{m}{2d}\right)\right) \\ &\leq \exp\left(2\log M - \frac{md}{4}\log\left(1 + \frac{m}{2d}\right)\right) < 1, \end{aligned}$$

since $d \leq m/2$, provided we choose M such that

$$\log M \leq \frac{md}{8}\log\left(1 + \frac{m}{2d}\right).$$

Setting,

$$\log M = \frac{md}{16}\log\left(1 + \frac{m}{2d}\right).$$

we have thus shown that there exists a family fulfilling the conditions of Lemma 3.

APPENDIX D: PROOF OF LLC UPPER BOUNDS

D.1 Notation and lemmas

We start by recalling the notation from Section 3.2. We denote the ith row of B^* by $b_i^* \in \mathbb{R}^{p-1}$, omitting the diagonal element which is assumed to be zero. From empirical covariances of the performed experiments, for $i \in [p]$, we obtain estimators $\hat{T}_i$ for T_i^* and $\hat{t}_i$ for t_i^*, where $T_i^* b_i^* = t_i^*$. Then, we solve the associated ℓ^1-regularized least squares problem,

$$\hat{b}_i = \underset{b}{\operatorname{argmin}}\, \|\hat{T}_i b - \hat{t}_i\|_2^2 + \lambda\|b\|_1, \quad i \in [p],$$

and assemble its solutions into $\hat{B}_{\text{llc}}$ as

$$(\hat{B}_{\text{llc}})_{i,:} = (P_i^\top \hat{b}_i)^\top, \quad i \in [p],$$

where $P_i \in \mathbb{R}^{(p-1)\times p}$ denotes the projection matrix that omits the ith coordinate.

In particular, the T_i^* are defined row-wise, adding a row $\mathfrak{e}_j^\top \Sigma^{*,e} P_i^\top$. for each experiment e such that $i \in \mathcal{U}_e$ and each entry $j \in \mathcal{J}_e$. Similarly, the vector t_i^* is defined by appending the corresponding entries $\Sigma_{j,i}^{*,e}$. Estimators for T^ast_i and t_i^* in turn are given by row-wise assembling empirical counterparts of the above quantities, so that a generic ℓth row of T_i^* and ℓth entry of t_i^* are given by

$$\hat{T}_{\ell,:} = e_j^\top (J_e + \hat{\Sigma}^e U_e) P_i^\top \quad \text{and} \quad (\hat{t}_i)_\ell = \hat{\Sigma}_{j,i}^e,$$

respectively.

Next, recall the following quantities that enter the rate:

$$\rho(d) = \min_{i\in[p]} \inf_{v\in\mathcal{C}(d), v\neq 0} \frac{\|T_i^* v\|_2}{\|v\|_2}, \tag{D.1}$$

$$R(d) = \max_{i\in[p]} \sup_{\substack{v\in\mathbb{R}^p, v\neq 0,\\ |\operatorname{supp}(v)|\leq d}} \frac{\|T_i^* v\|_2}{\|v\|_2}, \tag{D.2}$$

$$\tilde{R} = \max_{i\in[p]} \|(T_i^*)^\top\|_{\infty,\infty} = \max_{i\in[p]} \max_{j\in[p]} \sum_{k\in[p]} |(T_i^*)_{k,j}|,$$

where

$$\mathcal{C}(d) = \{v \in \mathbb{R}^p : \text{for all } S \subseteq [p] \text{ with } |S| \leq d, \|v_{S^c}\|_1 \leq 3\|v_S\|_1\}. \tag{D.3}$$

We restate Theorem 4 for convenience.

THEOREM 4. *Let assumptions A1 – A3 hold and fix $\delta \in (0,1)$. Assume further that*

$$n \gtrsim \left(1 \vee \frac{p^2}{\tilde{R}^2\eta^4} \vee \frac{pd}{(R(d)+1)^2\eta^4\rho(d)^4}\right) E\log(e\kappa p/\delta), \tag{D.4}$$

Then, the LLC estimator $\hat{B}_{\text{llc}}$ defined in (3.2) with λ chosen such that

$$\lambda \asymp \tilde{R}\sqrt{\frac{E\log(e\kappa p/\delta)}{n}},$$

satisfies

$$\|\hat{B}_{\text{llc}} - B^*\|_F^2 \lesssim \frac{\tilde{R}^2}{\rho(d)^4\eta^4} \frac{pdE\log(e\kappa p/\delta)}{n},$$

with probability at least $1-\delta$.

The proof relies on the following key lemmas. Lemma 4 yields control on the stochastic error, while Lemma 5 ensures that the linear system we solve via ℓ^1-regularization is well-conditioned for that purpose.

LEMMA 4. *Under the assumptions of Theorem 4, writing*

$$\phi_n = C\eta^{-2}\sqrt{\frac{E\log(e\kappa p/\delta)}{n}},$$

for a fixed $C > 0$, there is an event $\mathcal{A}$ such that $\mathbb{P}(\mathcal{A}) \geq 1-\delta$, and on $\mathcal{A}$,

$$\|\hat{T}_i^\top(\hat{T}_i b_i^* - \hat{t}_i)\|_\infty \leq 4\tilde{R}\phi_n, \quad \textit{for all } i \in [p].$$

LEMMA 5. *Assume that the same hypotheses as in Theorem 4 hold. On the same event $\mathcal{A}$ as in Lemma 4, we have*

$$\|\hat{T}_i h\|_2^2 \geq \frac{1}{2}\rho(d)^2\|h\|_2^2, \quad \textit{for all } h \in \mathcal{C}(d),\ i \in [p],$$

where $\mathcal{C}(d)$ is the set of vectors fulfilling the cone condition in (D.3).

D.2 Proof of Theorem 4

Since the experiments are completely separating, it follows from [HEH12] that

$$\rho(d) \geq \min_{i\in[p]} \sigma_{\min}(T_i^*) > 0.$$

Fix $i \in [p]$ and abbreviate $T^* = T_i^*$, $\hat{T} = \hat{T}_i$, $t^* = t_i^*$, $\hat{t} = \hat{t}_i$, $b^* = b_i^*$, and $\hat{b} = \hat{b}_i$. On the event $\mathcal{A}$ from Lemma 4 the following holds. By definition of $\hat{b}$, we have

$$\|\hat{T}\hat{b} - \hat{t}\|_2^2 + \lambda\|\hat{b}\|_1 \leq \|\hat{T}b^* - \hat{t}\|_2^2 + \lambda\|b^*\|_1.$$

Set $h = \hat{b} - b^*$ and rearrange to obtain

$$\|\hat{T}h\|_2^2 \leq -2h^\top \hat{T}^\top(\hat{T}b^* - \hat{t}) + \lambda(\|b^*\|_1 - \|\hat{b}\|_1). \tag{D.5}$$

By Hölder's inequality

$$|h^\top \hat{T}^\top(\hat{T}b^* - \hat{t})| \leq \|h\|_1\|\hat{T}^\top(\hat{T}b^* - \hat{t})\|_\infty.$$

By the assumptions on n and Lemma 4, $\|\hat{T}^\top(\hat{T}b^* - \hat{t})\|_\infty \leq 4\tilde{R}\phi_n$. Denote by S the support of b^*. By triangle inequality and splitting between S and S^c, we can bound the regularization term by

$$\|b^*\|_1 - \|\hat{b}\|_1 \leq \|h_S\|_1 + \|\hat{b}_S\|_1 - \|\hat{b}_S\|_1 - \|\hat{b}_{S^c}\|_1 \leq \|h_S\|_1 - \|h_{S^c}\|_1.$$

Add $\lambda\|h\|_1/2$ on both sides of (D.5) to obtain

$$\|\hat{T}h\|_2^2 + \frac{\lambda}{2}\|h\|_1 \leq \left(4\tilde{R}\phi_n + \frac{\lambda}{2}\right)\|h\|_1 + \lambda\|h_S\|_1 - \lambda\|h_{S^c}\|_1 \leq 2\lambda\|h_S\|_1. \tag{D.6}$$

Now, assume $\lambda \geq 8\tilde{R}\phi_n$, which by Lemma 4 matches the assumed scaling of

$$\lambda \asymp \tilde{R}\sqrt{\frac{E\log(e\kappa p/\delta)}{n}}. \tag{D.7}$$

Together with (D.6), we get that h fulfills the cone condition $\|h_{S^c}\|_1 \leq 3\|h_S\|_1$. In turn, by Lemma 5, taking into account that the assumptions on n and ϕ_n are fulfilled by assumption, we obtain

$$\|\hat{T}h\|_2^2 \geq \frac{1}{2}\rho(d)^2\|h\|_2^2.$$

Moreover, by the Cauchy-Schwarz inequality $\|h_S\|_1 \leq \sqrt{d}\|h\|_2$, so combined with (D.6) and (D.7), we have

$$\|h\|_2^2 \leq \frac{16d}{\rho(d)^4}\lambda^2 \lesssim \frac{\tilde{R}^2}{\rho(d)^4\eta^4}\frac{dE\log(e\kappa p/\delta)}{n}.$$

Re-introducing the index i and summing the above over all $i \in [p]$, we get

$$\|\hat{B}_{\text{llc}} - B^*\|_F^2 \leq \frac{\tilde{R}^2}{\rho(d)^4\eta^4}\frac{pdE\log(e\kappa p/\delta)}{n}.$$

D.3 Proof of Lemma 4

The proof of Lemma 4 consists of two parts that correspond to Lemma 6 and Lemma 7 below. Let $\phi_n > 0$ and define the events $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ as follows:

$$\mathcal{A}_1 = \left\{\max_i \|(\hat{T}_i - T_i^*)b_i^*\|_\infty \le \phi_n\right\}$$
$$\mathcal{A}_2 = \left\{\max_i \|\hat{T}_i - T_i^*\|_\infty \le \phi_n\right\}$$
$$\mathcal{A}_3 = \left\{\max_i \|\hat{t}_i - t_i^*\|_\infty \le \phi_n\right\}$$

Lemma 6 gives an upper bound on $\|\hat{T}^\top(\hat{T}b^* - \hat{t})\|_\infty$ in terms of ϕ_n, while Lemma 7 gives a high-probability bound on ϕ_n. We give the proofs of both of these lemmas after finishing the proof of Lemma 4.

LEMMA 6 (Trace term estimate). *If*

$$\phi_n \le \tilde{R}/p, \tag{D.8}$$

then on the event $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$,

$$\|\hat{T}_i^\top(\hat{T}_i b_i^* - \hat{t}_i)\|_\infty \le 4\tilde{R}\phi_n, \quad \textit{for all } i \in [p].$$

LEMMA 7 (Control on stochastic error). *Let $\delta \in (0,1)$. If $n \gtrsim E\log(e\kappa p/\delta)$,*

$$n \gtrsim E\log(e\kappa p/\delta), \tag{D.9}$$

and we set

$$\phi_n = C\eta^{-2}\sqrt{\frac{E\log(e\kappa p/\delta)}{n}},$$

for a fixed constant $C > 0$, we have that

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) \ge 1 - \delta.$$

Adjusting the constants in the requirement on n (D.4) in Theorem 4, we can ensure the requirements (D.8) and (D.9) and thus Lemma 4 follows by setting $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$ and combining Lemma 6 and Lemma 7.

PROOF OF LEMMA 6. We fix $i \in [p]$ and, as before, omit it for notational convenience It holds

$$\begin{aligned}\|\hat{T}^\top(\hat{T}b^* - \hat{t})\|_\infty &= \|(\hat{T} - T^* + T^*)^\top((\hat{T} - T^* + T^*)b^* - (\hat{t} - t^* + t^*))\|_\infty \\ &\le \|(T^*)^\top(\hat{T} - T^*)b^*\|_\infty + \|(T^*)^\top(\hat{t} - t^*)\|_\infty \\ &\quad + \|(\hat{T} - T^*)^\top(\hat{T} - T^*)b^*\|_\infty + \|(\hat{T} - T^*)^\top(\hat{t} - t^*)\|_\infty \\ &\le \left(\|(T^*)^\top\|_{\infty,\infty} + \|(\hat{T} - T^*)^\top\|_{\infty,\infty}\right)\left(\|(\hat{T} - T^*)b^*\|_\infty + \|t^* - \hat{t}\|_\infty\right),\end{aligned}$$

where we used the fact that $T^*b^* = t^*$ and that for an arbitrary matrix $A \in \mathbb{R}^{p\times p}$ and vector $x \in \mathbb{R}^p$, $\|Ax\|_\infty \le \|A\|_{\infty,\infty}\|x\|_\infty$. Since by the definition of $\mathcal{A}_2$,

$$\|(\hat{T} - T^*)^\top\|_{\infty,\infty} = \max_{j\in[p]}\sum_{i\in[p]}|(\hat{T} - T^*)_{i,j}| \le p\max_{i,j}|(\hat{T} - T^*)_{i,j}| \le p\phi_n,$$

we have that combined with the definitons of $\mathcal{A}_1$, $\mathcal{A}_3$, and $\tilde{R}$,

$$\|\hat{T}^\top(\hat{T}b^* - \hat{t})\|_\infty \leq \left(\|(T^*)^\top\|_{\infty,\infty} + p\phi_n\right) 2\phi_n \leq 4\tilde{R}\phi_n,$$

if $\phi_n \leq \tilde{R}/p$. □

PROOF OF LEMMA 7. For all three events, we write each element of the associated matrices or vectors as a sum over independent sub-exponential random variables and apply Bernstein's inequality, Lemma 19.

We start by controlling $\max_i \|(\hat{T}_i - T_i^*)b_i^*\|_\infty$. Let $i \in \{1,\dots,p\}$ and $\ell \in \{1,\dots,m_i\}$. The ℓth row of $\hat{T}_i - T_i^*$ corresponds to an experiment $e = e(i,\ell)$ such that $i \in \mathcal{U}_e$, and an index $j = j(\ell) \in \mathcal{J}_e$, which means we can write

$$\mathfrak{e}_\ell^\top \hat{T}_i b_i^* = \mathfrak{e}_j^\top \left(J_e + \hat{\Sigma}^e U_e\right) P_i^\top b_i^* \mathfrak{e}_j^\top \left(J_e + \hat{\Sigma}^e U_e\right) (B^*_{i,:})^\top$$

where we used that $B^*_{i,i} = 0$, so that $P_i^\top b_i^* = (B^*_{i,:})^\top$. Moreover, with independent normal random vectors Z_k^e for $e = 1,\dots,E$ and $k = 1,\dots,n/E$, $\hat{\Sigma}^e$ is of the form

$$\hat{\Sigma}^e = \frac{E}{n}(I - U_eB)^{-1}\sum_{k=1}^{n/E} Z_k^e (Z_k^e)^\top (I - U_eB)^{-\top},$$

so that

$$\mathfrak{e}_\ell^\top \hat{T}_i b_i^* = \mathfrak{e}_j^\top \left[J_e + \frac{E}{n}(I - U_eB)^{-1}\sum_{k=1}^{n/E} Z_k^e (Z_k^e)^\top (I - U_eB)^{-\top} U_e\right] (B^*_{i,:})^\top.$$

We proceed to control the ℓ^2 norm of the vectors that are being multiplied with Z_k^e. Lemma 15 yields that

$$\|(I - U_eB^*)^{-1}U_eB^*_{:,i}\|_2 \leq \|(I - U_eB^*)^{-1}\|_{\text{op}}\, \|U_e\|_{\text{op}}\, \|B^*_{:,i}\|_2 \leq \eta^{-1},$$

and

$$\|(I - U_eB^*)^{-\top}\mathfrak{e}_j\|_2 \leq \|(I - U_eB^*)^{-\top}\|_{\text{op}}\, \|\mathfrak{e}_j\|_2 \leq \eta^{-1}.$$

Hence, by Lemma 18,

$$B^*_{i,:}U_e(I - U_eB^*)^{-1}Z_k^e \sim \mathsf{subG}(\eta^{-2}), \text{ and}$$
$$\mathfrak{e}_j^\top (I - U_eB)^{-1}Z_k^e \sim \mathsf{subG}(\eta^{-2}),$$

and by Lemma 17, $\mathfrak{e}_\ell^\top \hat{T}_i b_i^* \sim \mathsf{subE}(\eta^{-2})$. Now, Bernstein's inequality in Lemma 19 and $\mathbb{E}[\hat{T}_i] = T_i^*$ allows us conclude that for $t_1 > 0$,

$$\mathbb{P}\left(|\mathfrak{e}_\ell^\top (\hat{T}_i - T_i^*)(B^*_{i,:})^\top| > t_1\right) \leq 2\exp\left[-c_B\left(\left(\frac{\eta^4 n t_1^2}{E}\right) \wedge \left(\frac{\eta^2 n t_1}{E}\right)\right)\right].$$

A union bound over all indices $i \in [p]$ and all ℓ, taking into account that there are at most κp rows in every T_i^*, then yields

$$\begin{aligned}\mathbb{P}\left(\max_{i\in[p],\ell} |\mathfrak{e}_\ell^\top (\hat{T}_i - T_i^*)(B^*_{i,:})^\top| > t_1\right) &\leq 2\kappa p^2 \exp\left[-c_B\left(\left(\frac{\eta^4 n t_1^2}{E}\right) \wedge \left(\frac{\eta^2 n t_1}{E}\right)\right)\right] \qquad \text{(D.10)}\\ &\leq \exp\left[-c_B\left(\left(\frac{\eta^4 n t_1^2}{E}\right) \wedge \left(\frac{\eta^2 n t_1}{E}\right)\right) + 4\log(e\kappa p)\right].\end{aligned}$$

Similarly, for $j \in \{1, \ldots, p-1\}$ and any row index ℓ,

$$\mathfrak{e}_\ell^\top \hat{T}_i \mathfrak{e}_j = \mathfrak{e}_\ell^\top \left[J_e + \frac{E}{n}(I - U_e B)^{-1} \sum_{k=1}^{n/E} Z_k^e (Z_k^e)^\top (I - U_e B)^{-\top} U_e \right] P_i^\top \mathfrak{e}_j, \tag{D.11}$$

and, as before,

$$\|(I - U_e B^*)^{-\top} U_e P_i^\top \mathfrak{e}_j\|_2 \leq \|(I - U_e B^*)^{-\top}\|_{\mathrm{op}} \, \|U_e\|_{\mathrm{op}} \|P_i^\top\|_{\mathrm{op}} \|\mathfrak{e}_j\|_2 \leq \eta^{-1},$$

so that $\mathfrak{e}_\ell^\top \hat{T}_i \mathfrak{e}_j \sim \mathsf{subE}(\tilde{\eta}^{-2})$. Hence, by Bernstein's inequality, for $t_2 > 0$,

$$\mathbb{P}\left(|\mathfrak{e}_\ell^\top (\hat{T}_i - T_i^*) \mathfrak{e}_j| > t_2 \right) \leq 2 \exp\left[-c_B \left(\left(\frac{\eta^4 n t_2^2}{E} \right) \wedge \left(\frac{\eta^2 n t_2}{E} \right) \right) \right].$$

A union bound over all $i \in [p]$, $j \in [p-1]$, and row indices ℓ yields

$$\begin{aligned} \mathbb{P}\left(\max_{i,j,\ell} |\mathfrak{e}_\ell^\top (\hat{T}_i - T_i^*) \mathfrak{e}_j| > t_2 \right) &\leq 2\kappa p^3 \exp\left[-c_B \left(\left(\frac{\eta^4 n t^2}{E} \right) \wedge \left(\frac{\eta^2 n t_2}{E} \right) \right) \right] \\ &\leq \exp\left[-c_B \left(\left(\frac{\eta^4 n t_2^2}{E} \right) \wedge \left(\frac{\eta^2 n t_2}{E} \right) \right) + 6 \log(e\kappa p) \right]. \end{aligned} \tag{D.12}$$

In particular, the union of the two events in (D.10) and (D.12) occurs with probability at most δ if

$$t_1 \wedge t_2 \gtrsim \eta^{-2} \left[\sqrt{\frac{E \log(e\kappa p/\delta)}{n}} \vee \frac{E \log(e\kappa p/\delta)}{n} \right].$$

Taking into account that all $\hat{T}_i$ and $\hat{t}_i$ are of the form we investigated in (D.11), we get the claim of the lemma if we choose

$$\phi_n = C \eta^{-2} \sqrt{\frac{E \log(e\kappa p/\delta)}{n}},$$

for a suitable constant C and assume $n \gtrsim E \log(e\kappa p/\delta)$. □

D.4 Proof of Lemma 5

To obtain the result, we employ the following lemma.

LEMMA 8 ([LW11, Lemma 12]). *If for a matrix $\Gamma \in \mathbb{R}^{k \times k}$, $k \in \mathbb{N}$ and an integer $s \geq 1$, it holds that*

$$|v^\top \Gamma v| \leq \delta \quad \textit{for all } v \in \mathbb{R}^k, \ \|v\|_0 \leq 2s, \ \|v\|_2 = 1,$$

then

$$|v^\top \Gamma v| \leq 27\delta(\|v\|_2^2 + \frac{1}{s}\|v\|_1^2) \quad \textit{for all } v \in \mathbb{R}^k.$$

To this end, let $v \in \mathbb{R}^{p-1}$ be a d sparse vector with $\|v\|_2 = 1$, as well as $i \in [p]$, and denote by $\hat{T} = \hat{T}_i$, $T^* = T_i^*$. Then,

$$\begin{aligned} |\|\hat{T} v\|_2^2 - \|T^* v\|_2^2| &= |\|(\hat{T} - T^* + T^*) v\|_2^2 - \|T^* v\|_2^2| \\ &= |\|(\hat{T} - T^*) v\|_2^2 + 2(T^* v)^\top (\hat{T} - T^*) v + \|T^* v\|_2^2 - \|T^* v\|_2^2| \\ &\leq \|(\hat{T} - T^*) v\|_2^2 + 2\|T^* v\|_2 \|(\hat{T} - T^*) v\|_2, \end{aligned}$$

On the one hand, by the definition of $R(d)$, (D.2),

$$\|T^*v\|_2 \leq R(d)\|v\|_2 = R(d). \tag{D.13}$$

On the other hand, if the event $\mathcal{A}_2$ occurred, then by definition

$$\|\hat{T} - T^*\|_\infty \leq \phi_n.$$

Thus, denoting by S the support of v, we can further estimate

$$\begin{aligned}\|(\hat{T} - T^*)v\|_2^2 &= \sum_{i=1}^{p}\left(\sum_{j\in S}(\hat{T}_{ij} - T^*_{ij})v_j\right)^2 \leq \sum_{i=1}^{p}\|(\hat{T} - T^*)_{i,S}\|_2^2\|v\|_2^2 \\ &\leq \sum_{i=1}^{p} d\|(\hat{T} - T^*)_{i,S}\|_\infty^2 \leq pd\|(\hat{T} - T^*)\|_\infty^2 \leq pd\phi_n^2, \end{aligned} \tag{D.14}$$

Combined, (D.13) and (D.14) yield

$$|\|\hat{T}v\|_2^2 - \|Tv\|_2^2| \leq (R(d) + 2\sqrt{pd}\phi_n)\sqrt{pd}\phi_n.$$

Now, let $h \in \mathbb{R}^{p-1}$ be a vector that fulfills the cone condition of order d. That is, there is a set of indices $S \subseteq [p-1]$ with $|S| \leq d$ such that $\|h_{S^c}\|_1 \leq 3\|h_S\|_1$. This in turn implies that

$$\|h\|_1 = \|h_S\|_1 + \|h_{S^c}\|_1 \leq 4\|h_S\|_1 \leq 4\sqrt{d}\|h\|_2$$

by the Cauchy-Schwarz inequality. By Lemma 8, (D.14), and the definition of $R(d)$ in (D.1), we have

$$\begin{aligned}\|\hat{T}h\|_2^2 &\geq \|T^*h\|_2^2 - |h^\top(\hat{T} - T^*)^\top(\hat{T} - T^*)h| \\ &\geq \|T^*h\|_2^2 - 27\left((R(d) + 2\sqrt{pd}\phi_n)\sqrt{pd}\phi_n\right)(\|h\|_2^2 + \frac{2}{d}\|h\|_1^2) \\ &\geq \left(\rho(d)^2 - 432\left((R(d) + 2\sqrt{pd}\phi_n)\sqrt{pd}\phi_n\right)\right)\|h\|_2^2.\end{aligned}$$

Combined, if

$$\phi_n \lesssim \frac{\rho(d)^2}{\sqrt{pd}}(R(d) + 1),$$

which is guaranteed from the assumptions of Theorem 4, we get the claim,

$$\|\hat{T}h\|_2^2 \geq \frac{1}{2}\rho(d)^2\|h\|_2^2.$$

APPENDIX E: PROOF OF UPPER BOUNDS FOR PENALIZED MAXIMUM LIKELIHOOD ESTIMATOR

E.1 Notation and lemmas

In the following section, we present the proof of Theorem 7, whereas the proofs of several key lemmas are deferred to later sections.

We begin by recalling the estimators and restating Theorem 7. The loss functions are given by

$$\ell(\Theta, \hat{\Sigma}) = \mathsf{Tr}(\hat{\Sigma}\Theta) - \log\det(\Theta), \quad \mathcal{L}(B) = \mathcal{L}(B, \hat{\Sigma}^1, \ldots, \hat{\Sigma}^E) = \sum_{e\in\mathcal{E}}\ell(\Theta^e(B), \hat{\Sigma}^e),$$

where

$$\Theta^e(B) = (I - U_e B)^\top (I - U_e B).$$

We consider the penalty terms

$$\mathrm{pen}_{\mathrm{init}}(B) = \mathrm{pen}_{\mathrm{init},\lambda_{\mathrm{init}}}(B) = \lambda_{\mathrm{init}} \sum_{e \in \mathcal{E}} \|\Theta^e(B)\|_1, \quad \mathrm{pen}_{\mathrm{loc}}(B) = \mathrm{pen}_{\mathrm{loc},\lambda_{\mathrm{loc}}}(B) = \lambda_{\mathrm{loc}} \|B\|_1,$$

leading to the objective functions

$$\mathcal{T}_{\mathrm{init}}(B) = \mathcal{L}(B, \hat{\Sigma}^1, \ldots, \hat{\Sigma}^E) + \mathrm{pen}_{\mathrm{init},\lambda_{\mathrm{init}}}(B),$$

and

$$\mathcal{T}_{\mathrm{loc}}(B) = \mathcal{L}(B, \hat{\Sigma}^1, \ldots, \hat{\Sigma}^E) + \mathrm{pen}_{\mathrm{loc},\lambda_{\mathrm{loc}}}(B).$$

Finally, the estimators are defined as

$$\hat{B}_{\mathrm{init}} \in \operatorname*{argmin}_{B \in \mathcal{B}_0} \mathcal{T}_{\mathrm{init}}(B), \quad \hat{B}_{\mathrm{loc}} \in \operatorname*{argmin}_{\substack{B \in \mathcal{B}_0 \\ \|B - \hat{B}_{\mathrm{init}}\|_F \le R_{\mathrm{loc}}}} \mathcal{T}_{\mathrm{loc}}(B),$$

where $\lambda_{\mathrm{init}}, \lambda_{\mathrm{loc}}$ and R_{loc} are tuning parameters that are to be determined.

THEOREM 7. *Under assumptions A1 – A3, if*

$$n \gtrsim \left(E^2 \vee \frac{1}{\eta^4} \vee p^2 \right) \frac{p^2 (d+1)^2 E^3}{\eta^4} \log(epE/\delta)$$

and the parameters for the estimators B_{init} and B_{loc} are chosen such that

$$R_{\mathrm{loc}} \asymp \frac{1}{\sqrt{E}} \wedge \eta \wedge \frac{1}{\sqrt{p}}, \quad \lambda_{\mathrm{init}} \asymp \sqrt{\frac{E \log(epE/\delta)}{n}}, \quad \textit{and} \quad \lambda_{\mathrm{loc}} \asymp \sqrt{\frac{E^2 \log(epE/\delta)}{n}}$$

then

$$\|\hat{B}_{\mathrm{loc}} - B^*\|_F^2 \lesssim \frac{p(d+1)E^2}{\eta^8\, n} \log(pE/\delta),$$

with probability at least $1 - \delta$.

First, we present three key lemmas used in the proof of Theorem 7. Lemma 9 yields curvature estimates of the likelihood function in terms of the difference of the concentration matrices associated with a candidate matrix B while Lemma 10 allows us to relate the difference of the concentration matrices to the difference in the underlying matrices, $B - B^*$. Finally, Lemma 11 gives bounds on a stochastic error term.

To facilitate the presentation, we present the lemmas with the following set of notations and assumptions. Let $B \in \mathbb{R}^{p \times p}$ be an arbitrary matrix and $\mathcal{E}$ a set of completely separating experiments as in assumption A2 with associated matrices $\{J_e, U_e\}_{e \in \mathcal{E}}$. Moreover, assume that $B^* \in \mathcal{B}(p, d, \eta)$. Then, we denote by

$$\Theta^e = \Theta^e(B) = (I - U_e B)^\top (I - U_e B), \quad \Theta^{*,e} = \Theta^e(B^*),$$

the concentration matrices associated with B and B^*, respectively, as well as the associated differences between the structure matrices and the concentration matrices by

$$H = B - B^*, \quad \Delta^e = \Theta^e - \Theta^{*,e},$$

respectively. We also abbreviate

$$\|\Delta\|_F^2 = \sum_{e \in \mathcal{E}} \|\Delta^e\|_F^2.$$

The first lemma follows from convexity arguments that also appear in [RBLZ08, LW13].

LEMMA 9 (Lower bounds on Gaussian log-likelihood function, [RBLZ08, LW13]). *With $\mathcal{L}$ defined as in (3.3.1), it holds for any $B \in \mathbb{R}^{p \times p}$ that*

$$\mathcal{L}(B) - \mathcal{L}(B^*) \geq \sum_{e \in \mathcal{E}} \mathsf{Tr}((\hat{\Sigma}^e - \Sigma^{*,e})(\Theta^e(B) - \Theta^{*,e})) + (c_1 \|\Delta\|_F \wedge c_1 \|\Delta\|_F^2), \tag{E.1}$$

where $c_1 = 18^{-1}$.

LEMMA 10 (Upper and lower bounds on $\|\Delta\|_F$ in terms of $\|H\|_F$). *If $B \in \mathcal{B}$, that is, B has zero diagonal, we have*

$$\|\Delta\|_F^2 \gtrsim \frac{\eta^4}{pE} \|H\|_F^4, \tag{E.2}$$

$$\|\Delta\|_F^2 \gtrsim \eta^4 \|H\|_F^2 (1 - 2\eta^{-4} \|H\|_F^4), \tag{E.3}$$

$$\|\Delta\|_F^2 \lesssim E(\|H\|_F^2 + \|H\|_F^4). \tag{E.4}$$

LEMMA 11 (Trace term estimates). *Let $\delta \in (0,1)$. Denote by ϕ_n and ψ_n the rates*

$$\psi_n = C\sqrt{\frac{E \log(epE/\delta)}{n}}, \qquad \phi_n = C\sqrt{\frac{E^2 \log(ep/\delta)}{n}},$$

for an appropriately chosen constant $C > 0$. If $n \gtrsim E \log(epE/\delta)$, then with probability at least $1 - \delta$, it holds for any $B \in \mathbb{R}^{p \times p}$ that

$$\sum_e \mathsf{Tr}\left((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta^e - \Theta^{*,e})\right) \leq \psi_n \sum_e \|\Delta^e\|_1$$

and

$$\sum_e \mathsf{Tr}\left((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta^e - \Theta^{*,e})\right) \leq (2 + \|H\|_{\infty,\infty}) \phi_n \|H\|_1.$$

For the proof of Theorem 7, we additionally introduce the following abbreviations. For $\diamond \in \{\mathrm{init}, \mathrm{loc}\}$, let

$$\Theta_\diamond^e = \Theta^e(\hat{B}_\diamond), \qquad H_\diamond = \hat{B}_\diamond - B^*,$$
$$\Delta_\diamond^e = \Theta_\diamond^e - \Theta^{*,e}, \quad \|\Delta_\diamond\|_F^2 = \sum_{e \in \mathcal{E}} \|\Delta_\diamond^e\|_F^2.$$

With this, we are ready to give the proof of 7.

E.2 Proof of Theorem 7

Proof sketch: The proof of Theorem 7 is split into two parts. First, we show that the initialization estimator $\hat{B}_{\text{init}}$ performs well enough to allow us to choose R_{loc} sufficiently small, so that the log-likelihood in an R_{loc}-neighborhood of B_{init} has large enough curvature. Second, we show that locally, $\hat{B}_{\text{loc}}$ achieves the desired rate.

Both proofs are based on re-arranging the optimality condition for the penalized log-likelihood, bounding the occurring trace term with high-probability, and exploiting the curvature of the log-likelihood function.

Step 1, basic inequality: By definition of the estimator $\hat{B}_{\text{init}}$,

$$\hat{B}_{\text{init}} \in \operatorname*{argmin}_{B} \mathcal{T}_{\text{init}}(B).$$

Comparing to the ground truth B^* yields the basic inequality

$$\mathcal{T}_{\text{init}}(\hat{B}_{\text{init}}) \leq \mathcal{T}_{\text{init}}(B^*),$$

which implies

$$\mathcal{L}(\hat{B}_{\text{init}}) - \mathcal{L}(B^*) \leq \text{pen}_{\text{init}}(B^*) - \text{pen}_{\text{init}}(\hat{B}).$$

Applying the lower bound on the negative log-likelihood (E.1) in Lemma 9 then yields

$$c_1\|\Delta_{\text{init}}\|_F \wedge c_1\|\Delta_{\text{init}}\|_F^2 \leq \sum_{e\in\mathcal{E}} \mathsf{Tr}((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta^e_{\text{init}} - \Theta^{*,e})) + \text{pen}_{\text{init}}(B^*) - \text{pen}_{\text{init}}(\hat{B}_{\text{init}}). \tag{E.5}$$

Step 1, estimate error term: Next, we bound the trace term

$$\sum_{e\in\mathcal{E}} \mathsf{Tr}((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta^e_{\text{init}} - \Theta^{*,e})),$$

with high probability using Lemma 11. For the remainder of the proof, we place ourselves on the event of probability at least $1-\delta$ on which the statement of Lemma 11 holds. Thus, we can estimate the trace term in (E.5) by

$$c_1\|\Delta_{\text{init}}\|_F \wedge c_1\|\Delta_{\text{init}}\|_F^2 \leq \psi_n \sum_{e} \|\Delta^e_{\text{init}}\|_1 + \text{pen}_{\text{init}}(B^*) - \text{pen}_{\text{init}}(\hat{B}_{\text{init}}).$$

Denoting the support of $\Theta^{*,e} = (I - U_e B^*)^\top (I - U_e B^*)$ by S^e_{init}, we have

$$\sum_{e} \|\Delta^e_{\text{init}}\|_1 = \sum_{e,i,j} |(\Delta^e_{\text{init}})_{i,j}| = \sum_{e} (\|(\Delta^e_{\text{init}})_{S^e_{\text{init}}}\|_1 + \|(\Delta^e_{\text{init}})_{(S^e_2)^c}\|_1).$$

Moreover, by triangle inequality,

$$\|\Theta^{*,e}\|_1 - \|\Theta^e_{\text{init}}\|_1 \leq \|(\Delta^e_{\text{init}})_{S^e_{\text{init}}}\|_1 - \|(\Delta^e_{\text{init}})_{(S^e_2)^c}\|_1.$$

Combined with the definition of the penalization term,

$$\text{pen}_{\text{init}}(B) = \lambda_{\text{init}} \sum_{e\in\mathcal{E}} \|\Theta^e(B)\|_1.$$

Now, assume $\lambda_{\text{init}} \geq \psi_n$, which matches the assumed scaling of λ_{init} to obtain

$$c_1 \|\Delta_{\text{init}}\|_F \wedge c_1 \|\Delta_{\text{init}}\|_F^2 \leq 2\lambda_{\text{init}} \sum_e \|(\Delta^e_{\text{init}})_{S^e_{\text{init}}}\|_1.$$

Note that we can control the size of the support $|S^e_{\text{init}}|$ by the in-degree of B^*. Namely, if we decompose

$$\Theta^{*,e} = (I - U_e B^*)^\top (I - U_e B^*) = \sum_{k=1}^p (I - U_e B^*)^\top_{k,:} (I - U_e B^*)_{k,:},$$

which is a sum over the outer product of $d+1$ sparse vectors by the assumption that the in-degree of the underlying graph is bounded by d, and hence

$$|S^e_{\text{init}}| \leq p(d+1)^2.$$

In turn, Hölder's inequality yields

$$2\lambda_{\text{init}} \sum_e \|\Delta^e_{S^e_{\text{init}}}\|_1 \leq 2\lambda_{\text{init}} \sqrt{p(d+1)^2 E} \|\Delta\|_F. \tag{E.6}$$

Bounds on $\|\Delta_{\text{init}}\|_F^2$: If $\|\Delta_{\text{init}}\|_F \geq 1$, by (E.5) and (E.6), we have

$$\|\Delta\|_F \leq 2 \frac{\lambda_{\text{init}}}{c_1} \sqrt{p(d+1)^2 E} \|\Delta\|_F,$$

which yields a contradiction if

$$\lambda_{\text{init}} \leq \frac{c_1}{4\sqrt{p(d+1)^2 E}}.$$

By the assumption that $\lambda_{\text{init}} \asymp \psi_n$ and the value of ψ_n in 11, this holds if

$$n \gtrsim p(d+1)^2 E^2 \log(epE/\delta).$$

If $\|\Delta_{\text{init}}\|_F \leq 1$, again by combining (E.5) and (E.6), we have

$$\|\Delta_{\text{init}}\|_F^2 \leq 2 \frac{\lambda_{\text{init}}}{c_1} \sqrt{p(d+1)^2 E} \|\Delta_{\text{init}}\|_F.$$

Dividing by $\|\Delta_{\text{init}}\|_F$ and squaring then implies

$$\|\Delta_{\text{init}}\|_F^2 \leq 4 \frac{\lambda_{\text{init}}^2}{c_1^2} \sqrt{p(d+1)^2 E}.$$

By Lemma 11 and the choice of $\lambda_{\text{init}} \asymp \psi_n$, this leads to

$$\|\Delta_{\text{init}}\|_F^2 \lesssim \frac{p(d+1)^2 E^2 \log(epE/\delta)}{n}. \tag{E.7}$$

Bounds on $\|H_{\text{init}}\|_F$: In order to relate $\|\Delta_{\text{init}}\|_F$ to $\|H_{\text{init}}\|_F$ we appeal to Lemma 10. If n is large enough for (E.7) to hold, then by the lower bound (E.2) in Lemma 10,

$$\frac{\eta^4}{pE} \|H_{\text{init}}\|_F^4 \lesssim \|\Delta_{\text{init}}\|_F^2 \lesssim \frac{p(d+1)^2 E^2 \log(epE/\delta)}{n},$$

and hence

$$\|H_{\text{init}}\|_F^4 \lesssim \frac{p^2(d+1)^2E^3\log(epE/\delta)}{\eta^4 n}, \tag{E.8}$$

which concludes the analysis for the initialization estimator.

Step 2, basic inequality: We have

$$\hat{B}_{\text{loc}} \in \underset{\|B-\hat{B}_{\text{init}}\|_F \le R_{\text{loc}}}{\operatorname{argmin}} \mathcal{T}_{\text{loc}}(B).$$

Suppose $R_{\text{loc}} \ge \|H_{\text{init}}\|_F$, which we achieve by (E.8) and choosing n large enough later, once R_{loc} has been chosen. Then, comparing to the ground truth B^* yields the basic inequality

$$\mathcal{T}_{\text{loc}}(\hat{B}_{\text{loc}}) \le \mathcal{T}_{\text{loc}}(B^*),$$

which implies

$$\mathcal{L}(\hat{B}_{\text{loc}}) - \mathcal{L}(B^*) \le \text{pen}_{\text{loc}}(B^*) - \text{pen}_{\text{loc}}(\hat{B}_{\text{loc}}).$$

Applying the lower bound on the negative log-likelihood (E.1) in Lemma 9 yields

$$c_1\|\Delta_{\text{loc}}\|_F \wedge c_1\|\Delta_{\text{loc}}\|_F^2 \le \sum_{e\in\mathcal{E}} \mathsf{Tr}((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta_{\text{loc}}^e - \Theta^{*,e})) + \text{pen}_{\text{loc}}(B^*) - \text{pen}_{\text{loc}}(\hat{B}_{\text{loc}}).$$

Step 2, estimate error term: We resort to Lemma 11, this time in the form of (E.19), which yields

$$c_1\|\Delta_{\text{loc}}\|_F \wedge c_1\|\Delta_{\text{loc}}\|_F^2 \le (2 + \|H_{\text{loc}}\|_{\infty,\infty})\phi_n\|H_{\text{loc}}\|_1 + \text{pen}_{\text{loc}}(B^*) - \text{pen}_{\text{loc}}(\hat{B}_{\text{loc}}). \tag{E.9}$$

First, we want to ensure $\|\Delta_{\text{loc}}\|_F \le 1$. The upper bound on Δ in Lemma 10, (E.4), achieves this if

$$\|H_{\text{loc}}\|_F \le \frac{c_2}{\sqrt{E}},$$

for a small enough constant $c_2 \le 1$. By the triangle inequality and (E.8), this is true if

$$R_{\text{loc}} \le \frac{1}{2}\frac{c_2}{\sqrt{E}}, \quad \text{and} \quad n \gtrsim \frac{p^2(d+1)^2E^5}{\eta^4}\log(epE/\delta).$$

Second, since we want the bound (E.3) to be effective within the ball $\|B - \hat{B}_{\text{init}}\|_F \le R_{\text{loc}}$ over which the optimization in step 2 is constrained, we choose n large enough to guarantee

$$\|H_{\text{loc}}\|_F^4 \le \frac{\eta^4}{4}.$$

This again follows from triangle inequality and (E.8) if

$$R_{\text{loc}} \le \frac{\eta}{2\sqrt{2}}, \quad \text{and} \quad n \gtrsim \frac{p^2(d+1)^2E^3}{\eta^8}\log(epE/\delta).$$

Third, to control the $\|H_{\text{loc}}\|_{\infty,\infty}$ term in (E.9), observe that

$$\|H_{\text{loc}}\|_{\infty,\infty} \le \sqrt{p}\|H_{\text{loc}}\|_F$$

by Hölder inequality. To guarantee $\|H_{\text{loc}}\|_{\infty,\infty} \le 2$, it is enough to ask for $\|H_{\text{init}}\|_F^4 \le 1/p^2$ and $R_{\text{loc}} \le 1/\sqrt{p}$ by triangle inequality. By (E.8), the former is be satisfied if

$$n \gtrsim \frac{p^4(d+1)^2E^3}{\eta^4}\log(epE/\delta).$$

Combined, in addition to the assumptions made in step 1, if

$$R_{\text{loc}} \le c_3\left[\frac{1}{\sqrt{E}} \wedge \eta \wedge \frac{1}{\sqrt{p}}\right] \quad \text{and} \quad n \gtrsim \left[E^2 \vee \frac{1}{\eta^4} \vee p^2\right]\frac{p^2(d+1)^2E^3}{\eta^4}\log(epE/\delta),$$

then

$$\|\Delta_{\text{loc}}\|_F \le 1, \quad \|H_{\text{loc}}\|_{\infty,\infty} \le 2 \quad \text{and} \quad \|H_{\text{loc}}\|_F^4 \le \frac{\eta^4}{4}.$$

In turn, from (E.3), we obtain

$$\|H_{\text{loc}}\|_F^2 \le 2\eta^4\|\Delta_{\text{loc}}\|_F^2.$$

Writing $S := \text{supp}(I - B^*)$, we then see that

$$\|H_{\text{loc}}\|_1 = \|(H_{\text{loc}})_S\|_1 + \|(H_{\text{loc}})_{S^c}\|_1,$$

and by triangle inequality,

$$\|B^*\|_1 - \|\hat{B}_{\text{loc}}\|_1 \le \|(H_{\text{loc}})_S\|_1 - \|(H_{\text{loc}})_{S^c}\|_1.$$

Together with (E.9) and observing that we can assume $\lambda_{\text{loc}} \ge 4\phi_n$, it follows that

$$\|H_{\text{loc}}\|_F^2 \lesssim \frac{\lambda_{\text{loc}}}{\eta^4}\|(H_{\text{loc}})_S\|_1.$$

Applying the Cauchy-Schwarz inequality gives

$$\|H_{\text{loc}}\|_F^2 \lesssim \frac{\lambda_{\text{loc}}}{\eta^4}\sqrt{|S|}\|H_{\text{loc}}\|_F.$$

Finally, we divide by $\|H_{\text{loc}}\|_F$, take squares, observe that $|S| \le p(d+1)$ use $\lambda_{\text{loc}} \asymp \phi_n$, and plug in the value of ϕ_n in Lemma 11 to obtain

$$\|H_{\text{loc}}\|_F^2 \lesssim \frac{p(d+1)E^2}{\eta^8\, n}\log(pE/\delta),$$

which concludes the proof.

E.3 Proof of Lemma 9

Let $R_1 > 0$ and recall the notation

$$\ell(\Theta, \Sigma) = \mathsf{Tr}(\Sigma\Theta) - \log\det(\Theta)$$

for the negative log-likelihood of a centered multivariate Gaussian distribution. Let $\Theta^*, \hat{\Sigma}$ be a positive definite matrix and a positive semi-definite matrix, respectively, and set $\Sigma^* = (\Theta^*)^{-1}$. Noting that the first derivative of $\Theta \mapsto -\log\det\Theta$ is $-\Theta^{-1}$ and the second derivative is $\Theta^{-1} \otimes \Theta^{-1}$, by computing a Taylor expansion of ℓ with differential remainder term about Θ^*, we have that

$$\ell(\Theta, \hat{\Sigma}) - \ell(\Theta^*, \hat{\Sigma}) = \mathsf{Tr}(\hat{\Sigma}(\Theta - \Theta^*)) - \mathsf{Tr}(\Sigma^*(\Theta - \Theta^*)) + \frac{1}{2}\mathsf{Tr}(\tilde{\Theta}^{-1}(\Theta - \Theta^*)\tilde{\Theta}^{-1}(\Theta - \Theta^*)) \tag{E.10}$$

for some $t \in [0, 1]$ and $\tilde{\Theta} = \Theta^* + t(\Theta - \Theta^*)$.

Denote the matrix square root of $\tilde{\Theta}^{-1}$ by $\tilde{\Theta}^{-1/2}$. Then, we can further lower bound the quadratic term by

$$\begin{aligned}\mathsf{Tr}(\tilde{\Theta}^{-1}(\Theta - \Theta^*)\tilde{\Theta}^{-1}(\Theta - \Theta^*)) &= \mathsf{Tr}(\tilde{\Theta}^{-1/2}(\Theta - \Theta^*)\tilde{\Theta}^{-1/2}\tilde{\Theta}^{-1/2}(\Theta - \Theta^*)\tilde{\Theta}^{-1/2}) \\ &= \|\tilde{\Theta}^{-1/2}(\Theta - \Theta^*)\tilde{\Theta}^{-1/2}\|_F^2 \\ &\geq \lambda_{\min}(\tilde{\Theta}^{-1/2})^4 \|\Theta - \Theta^*\|_F^2. \end{aligned} \tag{E.11}$$

By the spectral theorem, we can express the smallest eigenvalue of $\tilde{\Theta}^{-1/2}$ in terms of the largest eigenvalue of $\tilde{\Theta}$,

$$\lambda_{\min}(\tilde{\Theta}^{-1/2}) = (\lambda_{\max}(\tilde{\Theta}))^{-1/2}.$$

Now, recall

$$\mathcal{L}(B) = \sum_{e \in \mathcal{E}} \ell(\Theta^e(B), \hat{\Sigma}^e),$$

where

$$\Theta^e = \Theta^e(B) = (I - U_e B)^\top (I - U_e B),$$

and introduce

$$\tilde{\Delta}^e := \tilde{\Theta}^e - \Theta^{*,e}$$

and denote by

$$\|\tilde{\Delta}\|_F = \sqrt{\sum_{e \in \mathcal{E}} \|\tilde{\Delta}^e\|_F^2}$$

the Frobenius norm of the collection of $\tilde{\Delta}^e$ when viewed as a tensor. We now apply the expansion (E.10) and the estimate (E.11) to each of the summands, distinguishing two cases.

First, if $\|\Delta\|_F \leq R_1$, then also $\|\Theta^e - \Theta^{*,e}\|_F \leq R_1$ for all $e \in \mathcal{E}$ and we get

$$\begin{aligned}\lambda_{\max}(\tilde{\Theta}^e) &= \|\tilde{\Theta}^e\|_{\mathrm{op}} = \|\Theta^{*,e} + \tilde{\Delta}^e\|_{\mathrm{op}} \leq \|\Theta^{*,e}\|_{\mathrm{op}} + \|\tilde{\Delta}^e\|_{\mathrm{op}} \\ &\leq \|\Theta^{*,e}\|_{\mathrm{op}} + \|\tilde{\Delta}^e\|_F \leq \|\Theta^{*,e}\|_{\mathrm{op}} + \|\Delta^e\|_F \leq \|\Theta^{*,e}\|_{\mathrm{op}} + R_1.\end{aligned}$$

Therefore, from (E.11) we get a lower bound of the form

$$\mathcal{L}(B) - \mathcal{L}(B^*) \geq \sum_{e \in \mathcal{E}} \mathsf{Tr}((\hat{\Sigma}^e - \Sigma^{*,e})(\Theta^e - \Theta^{*,e})) + c_1 \|\Delta\|_F^2, \tag{E.12}$$

with $c_1 = (\max_{e \in \mathcal{E}} \|\Theta^{*,e}\|_{\mathrm{op}} + R_1)^{-2}/2$.

Second, if $\|\Delta\|_F > R_1$, we can leverage the convexity of $\Theta \mapsto -\log\det\Theta$ to again obtain lower bounds. Define $g(s)$ for $s \in [0, 1]$ by

$$g(s) = \sum_{e \in \mathcal{E}} \left[\ell(\Theta^{*,e} + s\Delta^e, \hat{\Sigma}^e) - \ell(\Theta^{*,e}, \hat{\Sigma}^e)\right].$$

Since ℓ is convex in Θ, g is convex in s, and we obtain

$$\frac{g(1)-g(0)}{1} \geq \frac{g(s)-g(0)}{s}, \quad \text{for all } s \in (0,1].$$

Plugging in $t = R_1/\|\Delta\|_F$, we are in the first case that was discussed and can appeal to (E.12), which yields

$$\begin{aligned}\sum_{e\in\mathcal{E}} \left[\ell(\Theta^e(B),\hat{\Sigma}^e) - \ell(\Theta^{*,e},\hat{\Sigma}^e)\right] &\geq \frac{\|\Delta\|_F}{R_1}\sum_{e\in\mathcal{E}}\left(\ell(\Theta^{*,e} + \frac{R_1}{\|\Delta\|_F}\Delta^e,\hat{\Sigma}^e) - \ell(\Theta^{*,e},\hat{\Sigma}^e)\right)\\ &\geq \frac{\|\Delta\|_F}{R_1}\sum_{e\in\mathcal{E}}\left(\mathsf{Tr}((\hat{\Sigma}^e - \Sigma^{*,e})\frac{R_1}{\|\Delta\|_F}\Delta^e) + R_1^2 c_1\right)\\ &= \sum_{e\in\mathcal{E}}\mathsf{Tr}((\hat{\Sigma}^e - \Sigma^{*,e})\Delta^e) + R_1 c_1\|\Delta\|_F.\end{aligned}$$

Combined, we get

$$\mathcal{L}(B) - \mathcal{L}(B^*) \geq \sum_{e\in\mathcal{E}}\mathsf{Tr}((\hat{\Sigma}^e - \Sigma^{*,e})(\Theta^e - \Theta^{*,e})) + (R_1 c_1\|\Delta\|_F \wedge c_1\|\Delta\|_F^2)$$

Finally, setting $R_1 = 1$ and observing that $\max_e \|\Theta^{*,e}\|_{\mathrm{op}} \leq 2$ by Lemma 15 yields the claim.

E.4 Proof of Lemma 10

In this section, we abbreviate

$$H = B - B^*, \quad A = (I - B^*)^{-1}, \quad A_e = (I - U_e B^*)^{-1}. \tag{E.13}$$

We also need the following linear transformation of H, which we denote by G,

$$G = HA = H(I - B^*)^{-1}. \tag{E.14}$$

First, we give a lemma that allows us to estimate the Frobenius norm of G by its off-diagonal elements.

LEMMA 12. *Let $B \in \mathbb{R}^{p\times p}$ and denote by H and G the matrices in* (E.13) *and* (E.14)*, respectively. Moreover, write G_D and G_{D^c} for the restriction of the matrix G to its diagonal indices and off-diagonal elements, respectively. If $\|B^*\|_{\mathrm{op}} < 1$ and $H_D = 0$, then*

$$\|G\|_F^2 \leq 2\|G_{D^c}\|_F^2.$$

PROOF. By the definition of G, we know that $H = GA^{-1} = G(I - B^*)$. The restriction $H_D = 0$ implies

$$\sum_{k=1}^p G_{ik}(I - B^*)_{ki} = 0, \quad \text{for all } i \in [p].$$

Since B^* has zero diagonal, for each $i \in [p]$, we can solve for G_{ii} and obtain

$$G_{ii} = \sum_{k\neq i} G_{ik}B^*_{ki}.$$

By the Cauchy-Schwarz inequality,

$$G_{ii}^2 \leq \left(\sum_{k\neq i} G_{ik}^2\right)\left(\sum_{k\neq i}(B^*_{ki})^2\right).$$

Finally, summing over all i gives

$$\|G_D\|_F^2 = \sum_i G_{ii}^2 \leq \left[\max_i \sum_k (B_{ki}^*)^2\right] \sum_i \sum_{k\neq i} G_{ik}^2.$$

Since $\|B^*\|_{\text{op}} < 1$ and by Lemma 15,

$$\max_i \sum_k (B_{ki}^*)^2 \leq 1,$$

we have the claim,

$$\|G\|_F^2 = \|G_{D^c}\|_F^2 + \|G_D\|_F^2 \leq 2\|G_{D^c}\|_F^2. \qquad \square$$

With this, we proceed to prove Lemma 10.

To start, let $e \in \mathcal{E}$. We have

$$\begin{aligned}
\Delta^e &= \Theta^e - \Theta^{*,e} \\
&= (I - U_e\hat{B})^\top (I - U_e\hat{B}) - (I - U_eB^*)^\top (I - U_eB^*) \\
&= (I - U_e(B^* + H))^\top (I - U_e(B^* + H)) - (I - U_eB^*)^\top (I - U_eB^*) \\
&= \; - (U_eH)^\top A_e^{-1} - A_e^{-\top}(U_eH) + (U_eH)^\top (U_eH).
\end{aligned} \tag{E.15}$$

Since $U_e^\top U_e = U_e^2 = U_e$, we can simplify the terms in the above expression as

$$(I - U_eB^*)^\top (U_eH) = U_eH - (B^*)^\top U_e^\top U_eH = (I - B^*)^\top (U_eH),$$

which leads to

$$\begin{aligned}
\Delta^e &= \; - (U_eH)^\top A^{-1} - A^{-\top}(U_eH) + A^{-\top}A^\top (U_eH)^\top (U_eH)AA^{-1} \\
&= A^{-\top}\left(-A^\top (U_eH)^\top - (U_eH)A + A^\top (U_eH)^\top (U_eH)A\right) A^{-1} \\
&= A^{-\top}\left(-(U_eHA)^\top - (U_eHA) + (U_eHA)^\top (U_eHA)\right) A^{-1}.
\end{aligned}$$

Hence, by Lemma 15,

$$\begin{aligned}
\|\Delta^e\|_F &\geq \sigma_{\min}^2(A^{-1})\| - (U_eHA)^\top - (U_eHA) + (U_eHA)^\top (U_eHA)\|_F \\
&\geq \eta^2 \| - (U_eHA)^\top - (U_eHA) + (U_eHA)^\top (U_eHA)\|_F.
\end{aligned} \tag{E.16}$$

Write $G = HA$.

First, to further lower bound the above expression, consider the diagonal of the $\mathcal{J}_e \times \mathcal{J}_e$ block of the matrix. There, we have $(U_eG)_{\mathcal{J}_e,\mathcal{J}_e} = 0$ and $(U_eG)_{\mathcal{J}_e,\mathcal{J}_e}^\top = 0$, and thus

$$\| - (U_eG)^\top - (U_eG) + (U_eG)^\top (U_eG)\|_F^2 \geq \sum_{i\in\mathcal{J}_e}\left(\sum_{u\in\mathcal{U}_e} G_{u,i}^2\right)^2 \geq \frac{1}{p}\left(\sum_{i\in\mathcal{J}_e}\sum_{u\in\mathcal{U}_e} G_{u,i}^2\right)^2,$$

where we used $\|h\|_1 \leq \sqrt{p}\|h\|_2$ for a vector $h \in \mathbb{R}^p$, which follows from Hölder's inquality. Summing over the experiments $\mathcal{E}$, together with the assumption of $\mathcal{E}$ being completely separating, Hölder's inequality, Lemma 12, and Lemma 15, we get

$$\begin{aligned}\|\Delta\|_F^2 &\geq \eta^4 \sum_{e\in\mathcal{E}} \|(U_e G)^\top + (U_e G) + (U_e G)^\top (U_e G)\|_F^2 \\ &\geq \frac{\eta^4}{pE}\|G_{D^c}\|_F^4 \gtrsim \frac{\eta^4}{4pE}\|G\|_F^4 \gtrsim \frac{\eta^4}{pE}\|H\|_F^4.\end{aligned}$$

Second, focusing on the $\mathcal{U}_e \times \mathcal{J}_e$ block of the matrix

$$-(U_e G)^\top - (U_e G) + (U_e G)^\top (U_e G),$$

we note that by the Cauchy-Schwarz inequality and the elementary inequality $(a+b)^2 \geq \frac{1}{2}a^2 - b^2$ for $a, b \in \mathbb{R}$,

$$\begin{aligned}&\| - (U_e G)^\top - (U_e G) + (U_e G)^\top (U_e G)\|_F^2 \\ &\geq \sum_{i\in\mathcal{U}_e} \sum_{j\in\mathcal{J}_e} \left(-G_{i,j} + \sum_{k\in\mathcal{U}_e} G_{k,i} G_{k,j} \right)^2 \\ &\geq \sum_{i\in\mathcal{U}_e} \sum_{j\in\mathcal{J}_e} \left[\frac{1}{2} G_{i,j}^2 - \left(\sum_{k\in\mathcal{U}_e} G_{k,i} G_{k,j} \right)^2 \right] \\ &\geq \sum_{i\in\mathcal{U}_e} \sum_{j\in\mathcal{J}_e} \left[\frac{1}{2} G_{i,j}^2 - \left(\sum_{k\in\mathcal{U}_e} G_{k,i}^2 \right) \left(\sum_{k\in\mathcal{U}_e} G_{k,j}^2 \right) \right] \\ &\geq \left(\sum_{j\in\mathcal{J}_e} \sum_{i\in\mathcal{U}_e} G_{i,j}^2 \right) \left(\frac{1}{2} - \|G\|_F^4 \right).\end{aligned}$$

Summing over the experiments, taking into account that by symmetry the same estimate holds for the $\mathcal{J}_e \times \mathcal{U}_e$ block, and bounding maximum and minimum singular values by Lemma 15, we obtain a lower bound of

$$\begin{aligned}\|\Delta\|_F^2 \geq \eta^4 \|G_{D^c}\|_F^2 (1 - 2\|G\|_F^4) &\gtrsim \eta^4 \|G\|_F^2 (1 - 2\|G\|_F^4) \\ &\gtrsim \eta^4 \|H\|_F^2 (1 - 2\|G\|_F^4) \\ &\geq \eta^4 \|H\|_F^2 (1 - 2\eta^{-4}\|H\|_F^4).\end{aligned}$$

Finally, we can upper bound $\|\Delta\|_F$ in terms of $\|H\|_F$, starting from (E.15), by

$$\begin{aligned}\|\Delta\|_F^2 &= \sum_{e\in\mathcal{E}} \| - (U_e H)^\top A^{-1} - A^{-\top}(U_e H) + (U_e H)^\top (U_e H)\|_F^2 \\ &\lesssim \sum_{e\in\mathcal{E}} (\|H\|_F^2 + \|H\|_F^4) \leq E(\|H\|_F^2 + \|H\|_F^4).\end{aligned}$$

E.5 Proof of Lemma 11

In this section, we abbreviate

$$\begin{aligned} T_1 &:= \max_{e \in \mathcal{E}} \left\| (\Sigma^{*,e} - \hat{\Sigma}^e) \right\|_\infty, \\ T_2 &:= \left\| \sum_{e \in \mathcal{E}} U_e A^{-1} (\Sigma^{*,e} - \hat{\Sigma}^e) \right\|_\infty, \\ T_3 &:= \max_k \left\| \sum_{e \in \mathcal{E}} \mathbb{I}_{k \in U_e} (\Sigma^{*,e} - \hat{\Sigma}^e) \right\|_\infty \end{aligned}$$

and introduce the events

$$\mathcal{A}_1 = \{T_1 \le \psi_n\}, \quad \mathcal{A}_2 = \{T_2 \le \phi_n\}, \quad \mathcal{A}_3 = \{T_3 \le \phi_n\}, \quad \mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3, \tag{E.17}$$

where the terms T_1, T_2, T_3 are upper bounded by rates ϕ_n and ψ_n to be made precise in Lemma 14, while Lemma 13 shows how ϕ_n and ψ_n can be used to estimate the trace term.

LEMMA 13 (Trace term estimates). *On the event $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3$, it holds for any $B \in \mathbb{R}^{p \times p}$ that*

$$\sum_e \mathsf{Tr}\left((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta^e - \Theta^{*,e}) \right) \le \psi_n \sum_e \|\Delta^e\|_1 \tag{E.18}$$

and

$$\sum_e \mathsf{Tr}\left((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta^e - \Theta^{*,e}) \right) \le (2 + \|H\|_{\infty,\infty}) \phi_n \|H\|_1. \tag{E.19}$$

LEMMA 14 (Control on stochastic error). *Let $\delta \in (0,1)$. There exists an absolute constant C such that if*

$$\psi_n = C \sqrt{\frac{E \log(epE/\delta)}{n}}, \qquad \phi_n = C \sqrt{\frac{E^2 \log(ep/\delta)}{n}},$$

and

$$n \ge CE \log(epE/\delta),$$

then with probability at least $1 - \delta$, it holds that

$$\mathbb{P}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \mathcal{A}_3) \ge 1 - \delta,$$

where $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ are defined as in (E.17).

Combined, Lemmas 13 and 14 yield Lemma 11.

PROOF OF LEMMA 13. First, by Hölder's inequality,

$$\sum_{e \in \mathcal{E}} \mathsf{Tr}\left((\Sigma^{*,e} - \hat{\Sigma}^e)(\Theta^e - \Theta^{*,e}) \right) \le \max_{e,i,j} |(\Sigma^{*,e} - \Sigma^e)_{i,j}| \sum_{e,i,j} |\Delta^e_{i,j}|.$$

Identifying the first term as T_1 and using the estimate $T_1 \le \psi_n$ yields (E.18).

Second, by the same calculation that led to (E.16), we decompose the trace term as

$$\begin{aligned}
&\sum_{e\in\mathcal{E}} \mathsf{Tr}\left((\Sigma^{*,e}-\hat{\Sigma}^{e})(\Theta^{e}-\Theta^{*,e})\right)\\
&\qquad = \sum_{e\in\mathcal{E}} \mathsf{Tr}\left((\Sigma^{*,e}-\hat{\Sigma}^{e})\left[-(U_eH)^\top A^{-1}-A^{-\top}(U_eH)+(U_eH)^\top(U_eH)\right]\right)\\
&\qquad = -2\sum_{e\in\mathcal{E}} \mathsf{Tr}\left(H^\top U_e A^{-1}(\Sigma^{*,e}-\hat{\Sigma}^{e})\right)+\sum_{e\in\mathcal{E}} \mathsf{Tr}\left((\Sigma^{*,e}-\hat{\Sigma}^{e})H^\top U_e H\right). \qquad \text{(E.20)}
\end{aligned}$$

The first term in (E.20) can be bounded by

$$\left|-2\sum_{e\in\mathcal{E}} \mathsf{Tr}\left(H^\top U_e A^{-1}(\Sigma^{*,e}-\hat{\Sigma}^{e})\right)\right| \le 2\|H\|_1\|\sum_{e\in\mathcal{E}} U_e A^{-1}(\Sigma^{*,e}-\hat{\Sigma}^{e})\|_\infty \le 2\phi_n\|H\|_1,$$

while the second term can be controlled by

$$\sum_{e\in\mathcal{E}} \mathsf{Tr}\left((\Sigma^{*,e}-\hat{\Sigma}^{e})H^\top U_e H\right) \le \|H\|_1\left\|\sum_{e\in\mathcal{E}} U_e H(\Sigma^{*,e}-\hat{\Sigma}^{e})\right\|_\infty. \qquad \text{(E.21)}$$

For each entry of the matrix on the right of (E.21), indexed by $i,j\in[p]$, we have

$$\begin{aligned}
\left|\left[\sum_{e\in\mathcal{E}} U_e H(\Sigma^{*,e}-\hat{\Sigma}^{e})\right]_{i,j}\right| &\le \left|\sum_{e\in\mathcal{E}}\sum_{k\in[p]} \mathbb{1}_{i\in U_e}(H)_{ik}(\hat{\Sigma}^{e}_{kj}-\Sigma^{*,e}_{kj})\right|\\
&= \left|\sum_{k\in[p]}(H)_{ik}\sum_{e\in\mathcal{E}} \mathbb{1}_{i\in U_e}(\hat{\Sigma}^{e}_{kj}-\Sigma^{*,e}_{kj})\right|\\
&\le \left(\sum_{k\in[p]}|(H)_{ik}|\right)\left(\max_{k\in[p]}\left|\sum_{e\in\mathcal{E}} \mathbb{1}_{i\in U_e}(\hat{\Sigma}^{e}_{kj}-\Sigma^{*,e}_{kj})\right|\right),
\end{aligned}$$

so that

$$\begin{aligned}
\left\|\sum_{e\in\mathcal{E}} U_e H(\Sigma^{*,e}-\hat{\Sigma}^{e})\right\|_\infty &\le \left(\max_{i\in[p]}\sum_{k\in[p]}|(H)_{ik}|\right)\left(\max_{i,j,k\in[p]}\left|\sum_{e} \mathbb{1}_{i\in U_e}(\hat{\Sigma}^{e}_{kj}-\Sigma^{*,e}_{kj})\right|\right)\\
&\le \phi_n \max_{i\in[p]}\sum_{k\in[p]}|(H)_{ik}| = \phi_n\|H\|_{\infty,\infty}.
\end{aligned}$$

Combined with the estimate (E.20), this yields the second claim, (E.19). □

PROOF OF LEMMA 14. To begin, recall the definition of $\hat{\Sigma}^e$, as a sum of n/E *i.i.d.* samples, that is, for $i,j\in[p]$,

$$(\hat{\Sigma}^e)_{i,j} = \frac{E}{n}\sum_{\ell=1}^{n}(X^e_\ell)_i(X^e_\ell)_j. \qquad \text{(E.22)}$$

By the definition of the sample distribution, we can write

$$(X^e_\ell)_i(X^e_\ell)_j = \underbrace{\mathfrak{e}_i^\top(I-U_eB^*)^{-1}Z^e_\ell}_{=:Y_1}\underbrace{(Z^e_\ell)^\top(I-U_eB^*)^{-\top}\mathfrak{e}_j}_{=:Y_2},$$

where the Z_ℓ^e follow a $\mathcal{N}(0,1)$ distribution and are *i.i.d.*, and Lemma 18 ensures that both Y_1 and Y_2 are $\mathsf{subG}(\sigma_{\max}(I - U_eB^*))$ random variables. By Lemma 17, we obtain that $(X_\ell^e)_i(X_\ell^e)_j \sim \mathsf{subE}(\sigma_{\max}(I - U_eB^*)^2)$.

Similarly,

$$\begin{aligned}(A_e^{-1}X_\ell^e)_i X_{\ell,j}^e &= \mathfrak{e}_i^\top (I - U_eB^*)(I - U_eB^*)^{-1} Z_\ell^e (Z_\ell^e)^\top (I - U_eB^*)^{-\top}\mathfrak{e}_j \\ &= \underbrace{\mathfrak{e}_i^\top Z_\ell^e}_{=:\tilde{Y}_1} \underbrace{(Z_\ell^e)^\top (I - U_eB^*)^\top \mathfrak{e}_j}_{=:\tilde{Y}_2}.\end{aligned}$$

Here, we have $\tilde{Y_1} \sim \mathsf{subG}(1)$ and $\tilde{Y_2} = \mathfrak{e}_j^\top (I - U_eB^*)Z_\ell^e \sim \mathsf{subG}(\sigma_{\max}(I - U_eB^*))$. By again applying Lemma 17, this means that $(A_e^{-1}X_\ell^e)_i X_{\ell,j}^e \sim \mathsf{subE}(\sigma_{\max}(I - U_eB^*))$.

Having established this, to obtain an estimate for T_1, we employ Bernstein's inequality, Lemma 19 to the sum in (E.22) for each $e \in \mathcal{E}, i, j \in [p]$ to see

$$\mathbb{P}\left(\left|(\Sigma^{*,e} - \hat{\Sigma}^e)_{i,j}\right| \geq t_1\right) \leq 2\exp\left[-c_B\left(\left(\frac{nt_1^2}{EK_1^2}\right) \wedge \left(\frac{nt_1}{EK_1}\right)\right)\right],$$

for $t_1 > 0$, with an absolute constant c_B and $K_1 = \max_e \sigma_{\max}(I - U_eB)^2$. Here, we made use of the fact that subtracting $\Sigma^{*,e}$ centers the variables in the sum and that there are n/E independent summands in (E.16). By a union bound,

$$\begin{aligned}&\mathbb{P}\left(\max_{e,i,j}\left|(\Sigma^{*,e} - \hat{\Sigma}^e)_{i,j}\right| \geq t_1\right) \\ &\quad\leq 2p^2E\exp\left[-c_B\left(\frac{nt_1^2}{EK_1^2}\right) \wedge \left(\frac{nt_1}{EK_1}\right)\right] \\ &\quad\leq \ \exp\left[-c_B\left(\frac{nt_1^2}{EK_1^2}\right) \wedge \left(\frac{nt_1}{EK_1}\right) + 2\log(2pE)\right].\end{aligned} \tag{E.23}$$

To bound T_2, for $i, j \in [p]$, we write

$$\begin{aligned}\left[\sum_{e\in\mathcal{E}} U_eA^{-1}(\Sigma^{*,e} - \hat{\Sigma}^e)\right]_{i,j} &= \sum_e \mathbb{I}_{i\in\mathcal{U}_e}(A^{-1}(\Sigma^{*,e} - \hat{\Sigma}^e))_{i,j} \\ &= \sum_e \sum_{\ell=1}^{n/E} a_{e,\ell}((A_e^{-1}X_\ell^e)_i(X_\ell^e)_j - \mathbb{E}[(A_e^{-1}X_\ell^e)_i(X_\ell^e)_j])\end{aligned}$$

with

$$a_{e,\ell} = \mathbb{I}_{i\in\mathcal{U}_e}\frac{E}{n}.$$

By Bernstein's inequality, Lemma 19, for $t_2 > 0$,

$$\mathbb{P}\left(\left|\left[\sum_{e\in\mathcal{E}} U_eA^{-1}(\Sigma^{*,e} - \hat{\Sigma}^e)\right]_{i,j}\right| \geq t_2\right) \leq 2\exp\left[-c_B\left(\frac{t_2^2}{K_2^2\|a\|_2^2}\right) \wedge \left(\frac{t_2}{K_2\|a\|_\infty}\right)\right],$$

where c_B is an absolute constant and

$$K_2 = \max_e \sigma_{\max}(I - U_eB), \quad \|a\|_2^2 = \sum_e \frac{E}{n} = \frac{E^2}{n}, \quad \|a\|_\infty = \max_e\left\{\frac{E}{n}\right\} = \frac{E}{n}.$$

A union bound then yields

$$\mathbb{P}\left(\max_{i,j}\left|\left[\sum_{e\in\mathcal{E}} U_e A^{-1}(\Sigma^{*,e}-\hat{\Sigma}^e)\right]_{i,j}\right| \geq t_2\right) \leq 2p^2\exp\left[-c_B\left(\frac{nt_2^2}{K_2^2E^2}\right)\wedge\left(\frac{nt_2}{EK_2}\right)\right] \tag{E.24}$$

$$\leq \exp\left[-c_B\left(\frac{nt_2^2}{K_2^2E^2}\right)\wedge\left(\frac{nt_2}{EK_2}\right)+2\log(2p)\right].$$

To bound T_3, we proceed similarly. Using the fact that $(X_\ell^e)_i(X_\ell^e)_j \sim \mathsf{subE}(\sigma_{\max}(I-U_eB)^2)$ instead of $(A_e^{-1}X_\ell^e)_iX_{\ell,j}^e \sim \mathsf{subE}(\sigma_{\max}(I-U_eB))$, for $t_3>0$, we have

$$\mathbb{P}\left(\max_{i,j,k}\left|\sum_e \mathbb{I}_{k\in U_e}(\hat{\Sigma}_{ij}^e-\Sigma_{ij}^{*,e})\right| \geq t_3\right) \leq 2p^3\exp\left[-c_B\left(\frac{nt_3^2}{K_3^2E^2}\right)\wedge\left(\frac{nt_3}{EK_3}\right)\right] \tag{E.25}$$

$$\leq \exp\left[-c_B\left(\frac{nt^2}{K_3^2E^2}\right)\wedge\left(\frac{nt_3}{EK_3}\right)+3\log(2p)\right],$$

where $K_3=\max_e \sigma_{\max}(I-U_eB^*)^2$.

Combined, recalling that by Lemma 15, $\sigma_{\max}(I-U_eB^*)\leq 2$ and applying a union bound, we see that the union of the events in (E.23), (E.24), and (E.25) occurs at most with probability δ if

$$t_1 \geq C\left[\sqrt{\frac{E\log(epE/\delta)}{n}}\vee\frac{E\log(epE/\delta)}{n}\right],$$

$$t_2\wedge t_3 \geq C\left[\sqrt{\frac{E^2\log(ep/\delta)}{n}}\vee\frac{E\log(ep/\delta)}{n}\right].$$

Restricting n to be large enough so that the effective part of the bound is the square root term in both cases then yields the claim. □

APPENDIX F: TECHNICAL LEMMAS

LEMMA 15. *If $B\in\mathbb{R}^{p\times p}$ is such that $\|B\|_{\mathrm{op}}\leq 1-\eta$ for some $\eta>0$, then we have*

$$\max_{i\in[p]}\sum_{k=1}^p B_{ki}^2 \vee \max_{i\in[p]}\sum_{k=1}^p B_{ik}^2 \leq 1.$$

Moreover, for any diagonal matrix $U=\mathsf{diag}\,u$ with $u\in\{0,1\}^p$,

$$\sigma_{\max}(I-UB)\leq 2, \qquad \textit{and} \qquad (\sigma_{\min}(I-UB))^{-1}\leq\frac{1}{\eta}.$$

PROOF. Let B and U as in the assumptions above. First, note that by its diagonal structure,

$$\|U\|_{\mathrm{op}}=\max_{i\in[p]}|u_i|\leq 1.$$

Next, We can relate the maximum and minimum singular values to the operator norm and employ sub-additivity and sub-multiplicativity as follows:

$$\sigma_{\max}(I-UB)=\|I-UB\|_{\mathrm{op}}\leq 1+\|U\|_{\mathrm{op}}\|B\|_{\mathrm{op}}\leq 2,$$

$$(\sigma_{\min}(I-UB))^{-1}=\|(I-UB)^{-1}\|_{\mathrm{op}}=\left\|\sum_{k\geq 0}(UB)^k\right\|_{\mathrm{op}}\leq\sum_{k\geq 0}\|U\|_{\mathrm{op}}^k\|B\|_{\mathrm{op}}^k\leq\frac{1}{1-(1-\eta)}=\frac{1}{\eta}.$$

Moreover, for $i \in [p]$, denoting the standard unit vector with 1 in the ith coordinate by $\mathfrak{e}_i$, we have

$$\sum_{k=1}^{p} B_{k,i}^2 = \|B_{:,i}\|_2^2 = \|B\mathfrak{e}_i\|_2^2 \leq \|B\|_{\mathrm{op}}^2 \|\mathfrak{e}_i\|_2^2 \leq (1-\eta)^2 \leq 1$$

and the same argument yields the bound for $\sum_{k=1}^{p} B_{i,k}^2$ by transposing the matrix and $\|B\|_{\mathrm{op}} = \|B^\top\|_{\mathrm{op}}$. □

DEFINITION 16 (Sub-Gaussian and sub-Exponential random variables). *We call a random variable X sub-Gaussian with variance proxy σ^2, written $X \sim \mathsf{subG}(\sigma^2)$, if*

$$\mathbb{E}[\exp(X^2/\sigma^2)] \leq 2.$$

We call a random variable sub-exponential with parameter λ, written $X \sim \mathsf{subE}(\lambda)$, if

$$\mathbb{E}[\exp(|X|/\lambda)] \leq 2.$$

LEMMA 17 (Product of subG random variables is subE, [Ver18, Lemma 2.7.7]).
If $X \sim \mathsf{subG}(\sigma_X^2)$ and $Y \sim \mathsf{subG}(\sigma_Y^2)$, then

$$XY \sim \mathsf{subE}(\sigma_X \sigma_Y).$$

LEMMA 18 (Sum of independent sub-Gaussian variables, [Ver18, Proposition 2.6.1]).
If $X_1, \ldots, X_n$ are n independent mean-zero random variables such that $X_i \sim \mathsf{subG}(\sigma_i^2)$, then

$$\sum_{i=1}^{n} X_i \sim \mathsf{subG}(\sigma^2), \quad \textit{with } \sigma^2 = \sum_{i=1}^{n} \sigma_i^2.$$

LEMMA 19 (Bernstein's inequality, [Ver18, Theorem 2.8.1]). *Let $X_1, \ldots, X_n$ be n independent mean-zero random variables such that $X_i \sim \mathsf{subE}(\lambda_i)$. Then, there is an absolute constant c_B such that for $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-c_B \min\left(\frac{t^2}{\sum_{i=1} \lambda_i^2}, \frac{t}{\max_{i\in[n]} \lambda_i}\right)\right).$$

APPENDIX G: NON-IDENTIFIABILITY IN THE CYCLIC CASE FOR EQUAL VARIANCES

In this section, we give a brief argument to show that for generic matrices B^*, unlike the acyclic case considered in [LB14, PB14], having equal noise variance as required in Assumption A3 does not lead to identifiability from observational data.

The argument is based on counting dimensions of the null space of the non-linear maps

$$\Theta^e(B) = (I - U_e B)^\top (I - U_e B).$$

One limitation of our argument is that it does not cover the potential identifiability of B^* from observational data under the additional assumption of bounded in-degree $d(B^*) \leq d$.

PROPOSITION 20. *Define the integer*

$$m = |\{(i,j)|\exists e : i \in \mathcal{U}_e,\ j \in \mathcal{J}_e,\ i \neq j\}| + |\{(i,j)|i < j, \exists e : i,j \in \mathcal{U}_e\}| + p\,. \tag{G.1}$$

Then the matrix $B^* \in \mathcal{B}$ is not uniquely determined by $\Theta^{*,e} = \Theta^e(B^*), e \in \mathcal{E}$ whenever $m < p^2 - p$. In particular, without interventions, this condition holds as soon as $p \geq 4$.

PROOF. Consider the maps

$$\Theta\colon \mathcal{B}_0 \cong \mathbb{R}^{p^2-p} \to \mathbb{R}^{E\times p^2} \qquad \text{and} \qquad \bar{\Theta}\colon \mathbb{R}^{p^2} \to \mathbb{R}^{E\times p^2},$$

defined by stacking all Θ^e into one vector and accepting respectively matrices with zero diagonal and arbitrary diagonal. Similarly, denote by $\bar{\Theta}^e$ the map Θ^e when not restricted to matrices with zero-diagonal.

We show that the derivative of Θ has constant rank bounded above by $p^2-p-(m-p)$ at a point $B^* \in \mathcal{B}_0$. In turn, whenever $m < p$, this implies the existence of $\tilde{B} \neq B^*$ such that $\Theta(B^*) = \Theta(\tilde{B})$ by the constant rank theorem

First, let $B \in \mathcal{B}$ be arbitrary and compute the derivative of $\bar{\Theta}$ at a point B by computing the derivative $D\bar{\Theta}^e(B)$ of the individual maps $\bar{\Theta}^e : \mathbb{R}^{p\times p} \to \mathbb{R}$. For any $\bar{H} \in \mathbb{R}^{p\times p}$, it holds

$$\begin{aligned} D\bar{\Theta}^e(B)[\bar{H}] &= (I - U_eB)^\top(-U_e\bar{H}) + (-U_e\bar{H})^\top(I - U_eB) \\ &= -(I-B)^\top(U_e\bar{H}) - (U_e\bar{H})^\top(I-B) \\ &= -A^{-\top}\left[(U_e\bar{H}A) + (U_e\bar{H}A)^\top\right]A^{-1}, \qquad (A = (I-B)^{-1}) \end{aligned} \tag{G.2}$$

where we used the fact that $U_e^2 = U_e$.

Next, we compute the dimension of the null space of $D\Theta(B)$. To that end, observe that for any $H \in \mathbb{R}^{p\times p}$ such that $D\Theta^e(B)[H] = 0$, it holds $D\bar{\Theta}^e(B)[\bar{H}] = 0$ and $(\bar{H})_{ii} = 0$ for all $i \in [p]$. To characterize the dimensionality of the subspace of such matrices H, we first consider the null space of $D\bar{\Theta}^e(B)[\bar{H}] = 0$ and then intersect it with the subspace given by $(\bar{H})_{ii} = 0$.

Abbreviate $\bar{G} = \bar{H}A$. By (G.2), $D\bar{\Theta}^e(B)[\bar{H}] = 0$ for all $B \in \mathcal{B}_0$ whenever $(U_e\bar{G}) + (U_e\bar{G})^\top = 0$ for all e. We permute the indices such that $\mathcal{J}_e = \{1, \ldots, |\mathcal{J}_e|\}$ to write this equality in block form:

$$U_e\bar{G} + (U_e\bar{G})^\top = \begin{bmatrix} 0 & J_e\bar{G}^\top U_e \\ U_e\bar{G}J_e & U_e(\bar{G}+\bar{G}^\top)U_e\,. \end{bmatrix}$$

For each $e \in \mathcal{E}$ the three nonzero blocks above translate into the following conditions:

$$\begin{aligned} \bar{G}_{i,j} &= 0 && \text{if } \exists e : i \in \mathcal{U}_e,\ j \in \mathcal{J}_e,\ i \neq j \\ \bar{G}_{i,j} &= -\bar{G}_{j,i} && \text{if } \exists e : i,j \in \mathcal{U}_e\,. \end{aligned}$$

As a result $\bar{H} = \bar{G}(I-B)$ is the image of $(I-B)$ through the linear operator $\bar{G}$ that lives in the intersections of the orthogonal subspaces defined by the above constraints. Thus, each constraint contributes 1 to the codimension of the null space of $D\bar{\Theta}(B)$. Equivalently, each constraint contributes 1 to the rank $D\bar{\Theta}(B)$.

Next, we discuss how to deal with the fact that we need to compute $\mathrm{rank}(D\Theta(B))$ instead of $\mathrm{rank}(D\bar{\Theta}(B))$, where B is restricted to lie in the subspace of matrices with zero-diagonal, thus we need to restrict $\bar{H}$ above accordingly. Intuitively, we want to say that the rank can increase by at most p, the number of additional linear constraints on the null space, but we need to further establish that there is a $B^* \in \mathcal{B}$ such that the rank of $D\Theta(B)$ is constant in a neighborhood of B^*. Adding to that the constraint that $\bar{H}$ has null diagonal, we get $\mathrm{rank}(D\Theta(B)) \leq m$, where m is defined in (G.1).

Next, we show that $\mathrm{rank}(D\Theta(B))$ is, in fact, constant and equal to some r^* in a neighborhood of B^* to apply the constant rank theorem.

To that end, let B^* be such that $r^* := \mathrm{rank}(D\Theta(B^*)) \geq \mathrm{rank}(D\Theta(B))$ for all $B \in \mathcal{B}$. Considering $D\Theta(B^*)$ a matrix let S^* be a maximal principal minor and denote the restriction of $D\Theta(B^*)$ to S^* by $[D\Theta(B^*)]_{S^*}$. By definition, we have

$$\mathrm{rank}(D\Theta(B^*)) = \mathrm{rank}([D\Theta(B^*)]_{S^*}) = r^*.$$

Moreover, the map $B \mapsto f(B) := \det[D\Theta(B^*)]_{S^*}$ is a polynomial in the elements of B such that $f(B^*) \neq 0$. By continuity, it also holds that $f(B) \neq 0$ in an open neighborhood of B^* as well, and thus $\mathrm{rank}(D\Theta(B)) \geq r^*$ in that neighborhood. But since r^* is maximal, $\mathrm{rank}\, D\Theta(B) = r^*$ for B in an open neighborhood of B^*.

The above means that we can apply the constant rank theorem [Boo86, Theorem II.7.1] to obtain diffeomorphisms $\varphi\colon \mathbb{R}^{p^2-p} \supset V_1 \to U_1 \subseteq \mathbb{R}^{p^2-p}$, $\psi\colon \mathbb{R}^{E\times p^2} \supset U_2 \to V_2 \subseteq \mathbb{R}^{E\times p^2}$, with U_j, V_j open sets for $j \in \{1, 2\}$ such that

$$\psi \circ \Theta \circ \varphi^{-1}(x) = (x_1, \dots, x_{r^*}, 0, \dots, 0), \quad \text{and} \quad \varphi^{-1}(0) = B^*.$$

If $r^* < p^2 - p$, we obtain a continuum of pre-images of $\Theta(B^*)$ as

$$\tilde{B}(x) = \varphi^{-1}(0, \dots, 0, x_{p^2-p-r^*+1}, \dots, x_{p^2-p}).$$

for all $(0, \dots, 0, x_{p^2-p-r^*+1}, \dots, x_{p^2-p}) \in V_1$, which includes points other than 0 because V_1 is an open set.

To conclude, recall that $r^* \leq m$ so that $m < p^2 - p$ is a sufficient condition for the failure of injectivity of Θ. This completes the first part of the proof.

To obtain the conclusion without interventions, note that in this case

$$m = |\{(i, j) : i < j, \exists e : i, j \in \mathcal{U}_e\}| + p = \binom{p}{2} + p$$

so that $m < p^2 - p$ whenever $p > 3$. □

APPENDIX H: NUMERICAL SPEED-UP

When many experiments are performed with a small number of nodes that are intervened on, say $|\mathcal{J}_e| \leq k$, calculating the log-likelihood term in the algorithms considered in Section 4 in a naive way takes $O(Ep^3)$ operations: both calculating $\mathsf{Tr}(\Theta^e(B)\hat{\Sigma}^e)$ and performing a Cholesky decomposition for each of the E matrices $\Theta^e(B) = L^e(L^e)^\top$ with L^e lower triangular takes $O(p^3)$ time. The Cholesky decomposition in turn is used to compute

$$\log\det\Theta^e(B) = \sum_{i=1}^{p} 2\log(L_{ii}).$$

The computational complexity can be improved by using a low rank decomposition of $\Theta^e(B)$, both for computing the trace term and the Cholesky decomposition $(I - B)^\top(I - B) = LL^\top$. To see this, write

$J_e = I - U_e$ and decompose

$$\begin{aligned}
&(I - U_eB)^\top(I - U_eB)\\
&\qquad = (I - B + J_eB)^\top(I - B + J_eB)\\
&\qquad = (I - B)^\top(I - B) + (J_eB)^\top(I - B) + (I - B)^\top(J_eB) + (J_eB)^\top(J_eB)\\
&\qquad = (I - B)^\top(I - B) + (J_eB)^\top - (J_eB)^\top B + (J_eB) - (J_eB)^\top(J_eB) + (J_eB)^\top(J_eB)\\
&\qquad = (I - B)^\top(I - B) + (J_eB)^\top + (J_eB) - (J_eB)^\top(J_eB)\\
&\qquad = (I - B)^\top(I - B) - (J_e - J_eB)^\top(J_e - J_eB) + J_e^\top J_e.
\end{aligned}$$

Hence, the Cholesky decomposition L^e can be computed by a rank k update followed by a rank k downdate of L, which takes $O(kp^2)$ [See04]. Computation of the trace terms $\mathsf{Tr}(\hat{\Sigma}^e\Theta^e(B))$ can be sped up analogously, also taking $O(kp^2)$ time.

Hence, the total time to compute the log-likelihood is $O(p^3 + Ekp^2)$. In a similar manner, computing the objective function for step (A.4) in the non-convex ADMM procedure can be done in $O(p^3 + Ekp^2)$ time, although one iteration takes $O(Ep^3)$ time due to the complexity of performing step (A.3).

REFERENCES

[BKSV15] M. Benning, F. Knoll, C.-B. Schönlieb, and T. Valkonen. Preconditioned ADMM with nonlinear operator constraint. In *IFIP Conference on System Modeling and Optimization*, pages 117–126. Springer, 2015.

[Boo86] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*, volume 120. Academic press, 1986.

[BPC+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[Cai84] M. Cai. On a problem of Katona on minimal completely separating systems with restrictions. *Discrete Mathematics*, 48(1):121–123, January 1984.

[CBG13] X. Cai, J. A. Bazerque, and G. B. Giannakis. Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS Computational Biology*, 9(5):e1003068, 2013.

[Dic69] T. J. Dickson. On a problem concerning separating systems of a finite set. *Journal of Combinatorial Theory*, 7(3):191–196, November 1969.

[EB92] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992.

[FHT08] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, July 2008.

[Fra12] J. N. Franklin. *Matrix Theory*. Courier Corporation, 2012.

[Gab83] D. Gabay. Chapter ix applications of the method of multipliers to variational inequalities. *Studies in mathematics and its applications*, 15:299–331, 1983.

[GM75] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975.

[GM76] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[HDRS11] C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, pages 2330–2338, 2011.

[HEH12] A. Hyttinen, F. Eberhardt, and P. O. Hoyer. Learning linear cyclic causal models with latent variables. *Journal of Machine Learning Research*, 13(Nov):3387–3439, 2012.

[HYW00] B. S. He, H. Yang, and S. L. Wang. Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities. *Journal of Optimization Theory and applications*, 106(2):337–356, 2000.

[LB14] P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[LN89] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[LW11] P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

[LW13] P.-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

[NW06] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, second edition, 2006.

[PB14] J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, January 2014.

[RBLZ08] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[See04] M. Seeger. Low rank updates for the Cholesky decomposition. *Infoscience, EPFL Scientific Publications*, 2004.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation. Revised and Extended from the 2004 French Original. Translated by Vladimir Zaiats*. Springer Series in Statistics. Springer, New York, 2009.

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[WL01] S. L. Wang and L. Z. Liao. Decomposition method with a variable parameter for a class of monotone variational inequality problems. *Journal of optimization theory and applications*, 109(2):415–429, 2001.

[WYZ19] Y. Wang, W. Yin, and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019.

[ZBLN97] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.