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).

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

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

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, $\|B\|_{\mathrm{op}}$ 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 $\|B^*\|_{\mathrm{op}} < 1$ 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*

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

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 $\|B^*\|_{\mathrm{op}} \leq 1 - \eta$ 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 $i = 1, \ldots, p$, 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 $\mathcal{J}_e = \{1, \ldots, |\mathcal{J}_e|\}$, in which case we have

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

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:

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

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 $\hat{B}_{\mathrm{llc}}$ defined in (3.2) with λ chosen such that

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

satisfies

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

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 $\Theta^e(B) = (I - U_e B)^\top (I - U_e B)$, 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 $\hat{B}_{\mathrm{init}}$ 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 $\hat{B}_{\mathrm{init}}$.

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

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

leading to the penalized log-likelihood

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

The initialization estimator is then given by

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

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,

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

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

and the estimator is given by

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

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*

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

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

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

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 $n \gtrsim (p^2 \vee E^2)p^2d^2E^3\log(Ep)$ 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.

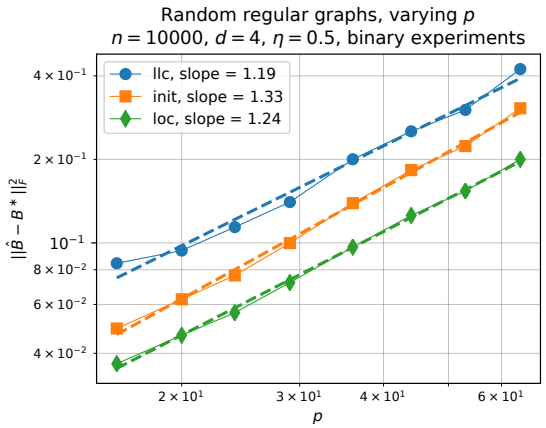
Here, we report results from experiments with synthetic data generated using (directed) random regular graphs to gauge the performance of the maximum likelihood procedure, comparing it to the LLC algorithm (Hyttinen et al., 2012). Further details on the experiments and more experiments on synthetic and 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

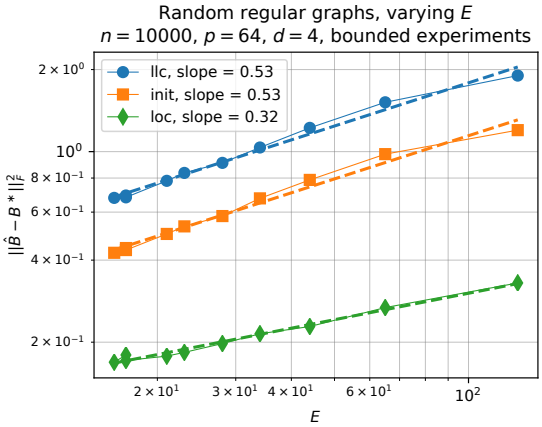
Data generation The ground truth graphs are generated by first obtaining the (directed) adjacency matrix $B_{\mathrm{adj}} \in \{0,1\}^{p\times p}$, a matrix $B_{\mathrm{val}} \in \mathbb{R}^{p\times p}$ containing edge values, and finally setting B^* to be the Hadamard product of the two, normalized 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 $\hat{B}_{\mathrm{llc}}$, "init" to $\hat{B}_{\mathrm{init}}$, "loc" to $\hat{B}_{\mathrm{loc}}$.

random graph, where $\mathrm{supp}((B_{\mathrm{adj}})_{i,:})$ is constructed by sampling d times uniformly at random without replacement from $\{1, \dots, p\} \backslash \{i\}$ 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 $\|\hat{B}_{\mathrm{init}} - B^*\|_F$ and $\|\hat{B}_{\mathrm{loc}} - B^*\|_F$ is minimal. For R_{loc}, we choose $R_{\mathrm{loc}} = 2\|\hat{B}_{\mathrm{init}} - B^*\|_F$. In practice, both parameters could be chosen by cross-validation.

Initialization of optimization algorithm: We initialize the calculation of $\hat{B}_{\mathrm{init}}$ 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 $\hat{B}_{\mathrm{loc}}$ is initialized with the output of $\hat{B}_{\mathrm{init}}$. 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 $\hat{B}_{\mathrm{llc}}$, $\hat{B}_{\mathrm{init}}$, and $\hat{B}_{\mathrm{loc}}$, 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 $\hat{B}_{\mathrm{llc}}$ and $\hat{B}_{\mathrm{init}}$, about $E^{1/3}$ for $\hat{B}_{\mathrm{loc}}$.

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 covari-

ance 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*.