LEARNING GRAPHICAL MODELS
FUNDAMENTAL LIMITS AND EFFICIENT ALGORITHMS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL
ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

José Bento
August 2012

Preface

Graphical models provide a flexible and yet powerful language to describe high dimensional probability distributions. Over the last 20 years, graphical models methods have been successfully applied to a broad range of problems, from computer vision to error correcting codes to computational biology, to name a few.

In a graphical model, the graph structure characterizes the conditional independence properties of the underlying probability distributions. Roughly speaking, it encodes key information about which variables influence each other. It allows us to answer questions of the type: are variables X and Y dependent because they 'interact' directly, or because they are both dependent on a third variable Z? In many applications, this information has utmost practical importance, and it is therefore crucial to develop efficient algorithms to learn the graphical structure from data. This problem is largely unsolved and for a long time several heuristics have been used without a solid theoretical foundation in place.

In the first part of this work, we consider the problem of learning the structure of Ising models (pairwise binary Markov random fields) from i.i.d. samples. While several methods have been proposed to accomplish this task, their relative merits and limitations remain somewhat obscure. By analyzing a number of concrete examples, we show that low-complexity algorithms often fail when the Markov random field develops long-range correlations. More precisely, this phenomenon appears to be related to the Ising model phase transition (although it does not coincide with it).

An example of an important algorithm that exhibits this behavior is the ℓ_1-regularized logistic regression estimator introduced by Ravikumar et al. [94]. Ravikumar et al. [94] proved a set sufficient conditions under which this algorithm exactly

learns Ising models, the most interesting being a so-called *incoherence condition.* In this thesis we show that this incoherence condition is also necessary and analytically establish whether it holds for several families of graphs. In particular, denoting by θ the edge strength and by Δ the maximum degree, we prove that regularized logistic regression succeeds on any graph with $\Delta \leq 3/(10\theta)$ and fails on *most* regular graphs with $\Delta \geq 2/\theta$.

In the second part of this work, we address the important scenario in which data is not composed of i.i.d. samples. We focus on the problem of learning the drift coefficient of a p-dimensional stochastic differential equation (SDE) from a sample path of length T. We assume that the drift is parametrized by a high-dimensional vector, and study the support recovery problem in the case where p is allowed to grow with T.

In particular, we describe a general lower bound on the sample-complexity T by using a characterization of mutual information as a time integral of conditional variance, due to Kadota, Zakai, and Ziv. For linear SDEs, the drift coefficient is parametrized by a p-by-p matrix which describes which degrees of freedom interact under the dynamics. In this case, we analyze an ℓ_1-regularized least-squares estimator and describe an upper bound on T that nearly matches the lower bound on specific classes of sparse matrices.

We describe how this same algorithm can be used to learn non-linear SDEs and in addition show by means of a numerical experiment why one should expect the sample-complexity to be of the same order as that for linear SDEs.

Acknowledgements

The work of this thesis was accomplished with support of many ‘hands’.

First, I am happy to thank God for His love. Without Him, my endeavors at Stanford are worthless.

I would also like to thank Professor Andrea Montanari, my advisor, for his support, encouragement, and guidance, and above all, for being an example of excellence in research. The passion he has for his work has taught me first-hand a valuable lesson: only those who love their work can excel.

I have also been privileged to be a part Andrea’s research group. During my stay at Stanford I was able to interact with Sewoong, Satish, Moshen, Morteza, Raghu, Adel, Yash and Yash. In particular, I was able to collaborate with Sewoong and Morteza on two different projects.

In addition, I must thank my friends for their care and presence in my life throughout these years. I would have to double the number of pages in this thesis if I even attempted to summarize in a list the many benefits I have been granted through them.

My family has also been indispensable in my success, for the indescribable joy they bring me which fuels my life and gives me strength. In particular, I must thank my mother, Nina, for always having many interesting stories to tell every Sunday afternoon over Skype. I am also very happy for the support I have received from my Godmother, Teresa. More visibly in the first years of my life but certainly still now, at a distance.

I am also very thankful for all the help Professor Persi Diaconis gave me during my time at Stanford and for the several interesting topics I have learned about from

him. In particular, I am thankful for his availability to meet and help me overcome mathematical hurdles and for kindly and habitually inquiring about the wellbeing of my family.

I would also like to acknowledge Professor Iain Johnstone for his availability to regularly discuss my research with him and also Professor Emmanuel Candes and Professor Ramesh Johari for being part of my thesis defense and reading committee. More over, I must thank Professor Tsachy Weissman for the availability to work with me and guide me in my first year at Stanford.

Furthermore, I would like to thank Hewlett-Packard Labs and Technicolor Labs in Palo Alto for providing me two wonderful summer internships. In particular, I would like to thank Niranjan, Emon and Jerry from HP and Jinyun, Stratis, Nadia and Jean from technicolor.

Finally, this work was partially supported by the Portuguese Doctoral FCT fellowship SFRH / BD / 28924 / 2006.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In a nutshell, in this thesis we look at stochastic models parametrized by an unknown graph - *graphical models* - and address the following question:

Question. *Is it possible to recover the graph from data?*

1.1 Graphical models

Graphical models are a language to compactly describe large joint probability distributions using a set of 'local' relationships among neighboring variables in a graph [33, 74, 77, 71]. The 'local' relationships are described by functions involving neighboring variables in this graph that are related to conditional probability distributions over these variables. The product of these functions equals the joint distribution. Let us be precise. Given a set of variables $\underline{x} = \{x_1, ..., x_p\}$ that take a value in $\{-1, +1\}$, we can start from a local model of how likely it is that any two variables assume the same value, -1 or $+1$, and obtain a joint model for the probability that $\underline{x}$ takes a certain value in $\{-1, 1\}^p$. For example, the local model can describe two scenarios: any two variables x_i, x_j, are either 'connected' with strength $\theta_{ij} \neq 0$ or 'disconnected', $\theta_{ij} = 0$. If two variables x_i and x_j are connected, we include a factor $e^{\theta_{ij} x_i x_j}$ in the joint probability distribution, and if disconnected, we include a factor 1. Given a weighted graph $G = (V, E)$ with vertex set V and edge set E, ($V = [p] \equiv \{1, ..., p\}$, $E = \{(i, j) : \theta_{i,j} \neq 0\}$) and assuming every two variables obey the same local model described above, the joint probability distribution of a configuration of $\underline{x}$ is given by

$$\mathbb{P}_G(\underline{x}) = \frac{1}{Z_G} \prod_{(i,j)\in E} e^{\theta_{ij} x_i x_j}, \tag{1.1.1}$$

where Z_G is a normalization constant. The above model is an example of a graphical model and is known as the Ising model.

Graphical models find many applications. For example, the above model has been used to understand the evolution of opinions in closed communities [105]. There, G describes a set of acquaintanceships among individuals in a community. Individuals 'connected' by a positive bond $\theta_{ij} > 0$ are more likely to assume the same opinion, -1 or -1, on a given matter. Ising models are also used in computer vision, where G describes a set of pixels in an image to be de-noised [112, 79] or pixels in a pair of stereoscopic images to be used for depth perception [106]. Computational biologists use other graphical models where G represents a network of interacting genes that regulate cell activities [48, 47, 73, 58] or the amino-acids in a protein that interact with themselves and other bio-entities to determine their shape and function [59, 111]. In digital communications, codes are constructed by designing graphs that represent parity-check constraints on the bits of codewords and decoding is done by computations on these graphs [50, 28, 80, 95]. In computational neurobiology, to understand the functioning of the brain, scientists use graphical models where G represents a map of neural connections in the brain [34, 64, 22, 113]. Graphical models also find applications in meteorology [23].

These applications pose different challenges related to graphical models. Four of the main problems are: *representation*, *sampling*, *inference* and *learning*. See [69, 18, 75] for a review on the several research questions associated with graphical models. Representation concerns choosing the right model for the application at hand. Different kinds of graphical models include Markov random fields (based on undirected graphs) , Bayesian networks (based on directed graphs) and factor graphs. In this thesis we focus only on pair-wise Markov random fields and on stochastic differential equations. Sampling refers to the problem of generating samples from the model's probability distribution. Inference is about using the model to answer probabilistic queries - for example, computing marginal probabilities or inferring the value of unobserved variables. The problem of learning focuses on recovering the model from data, as when we estimate the values of the parameters $\{\theta_{ij}\}$ in model (1.1.1). This is important because often we start with no details about the model. In many applications the interest is specifically in recovering the *support* of these parameters. For the Ising model, this corresponds to determining which coefficients $\{\theta_{ij}\}$ are non-zero. Herein lies the focus of this thesis, which can be described as an investigation into how the graph G can be learned from data. Concerning this last problem, we note that several variants can be conceived which we do not address. For example, how do we learn the graph from partially hidden data? Regarding the questions addressed in this thesis, these hidden data can come, for example, from unobserved nodes in the graph [97, 46, 7]. This is relevant in evolutionary biology as a way to model the change in gene-data along the evolution-tree of a group of species [84, 99]. Hidden data can also arise from points in the trajectory of a stochastic differential equation which are not sampled [31, 91, 5]. This is the case in many applications where sampled data come at a very low frequency compared with the dominant frequency modes in the system.

1.2 Structural learning in graphical models

One of the most important properties of graphical models is that the underlying graph describes a set of conditional independence relations among variables. This relation is made precise by the Hammersley-Clifford theorem [57, 17],

Theorem 1.2.1. *Let $G = (V = [p], E)$ be an undirected graph and C_G the set of all maximal cliques in G. A probability distribution $\mathbb{P}(x_1, ..., x_p) > 0$ factorizes according to G*

$$\mathbb{P}(\underline{x}) = \frac{1}{Z} \prod_{c \in C_G} \Phi_c(\underline{x}_c) \tag{1.2.1}$$

if and only if,

$$\mathbb{P}(x_i|\underline{x}_{G\setminus i}) = \mathbb{P}(x_i|\underline{x}_{\partial i}), \text{ for all } i \in V. \tag{1.2.2}$$

In the expressions above, Z is a normalization constant, for any set $c \subseteq V$, $\underline{x}_c = \{x_j : j \in c\}$ and $\partial i = \{j \in V : (i,j) \in E\}$.

In fact, when the above equivalence holds, all the following three conditional independence relations hold:

- A variable is conditionally independent of all other variables given its neighbors;
- Any two non-connected variables ($(i,j) \notin E$) are conditionally independent given all other variables;
- Any two subsets of variables, A, B, are conditionally independent given a separating subset C, where every path from A to B passes through C.

For the Ising model (1.1.1) over a graph G, since $\mathbb{P}_G(\underline{x}) > 0$ and $\mathbb{P}_G(\underline{x})$ factorizes according to G, all conditional independence property above hold.

Understanding the dependencies among a set of variables is of great importance for many applications and is almost always a prerequisite for further modeling efforts. In particular, it allows us to solve the confounding variable problem: Given three different variables - X, Y and Z - does X affect Z directly, or only through Y?

For concrete examples, these conditional dependencies can have interesting interpretations. Consider the Ising model where the vertices represent people holding one of two opinions, -1 or $+1$, on a certain matter. The edges, all with equal positive weight, represent influence links. Two people who are connected have a higher chance of holding the same opinion. Now let A, B and C be groups of individuals ($A, B, C \subset V$) such that, if members of group C were to disappear, there would be no connection between group A and B. Then, conditioned on the opinion of all the members of group C being fixed, the members of group A and B cannot influence each other.

The question formulated in the beginning which motivates the work in this thesis amounts to recovering this set of dependencies. For the Ising model (1.1.1), a particular instance of this problem can be written as

Question. *Is it possible to recover G given n* independent identically distributed *(i.i.d.) samples from $\mathbb{P}_G(\underline{x})$?*

This problem appears in the literature under the name of *structural learning* of graphical models (e.g. [25, 42, 47, 18, 3, 29, 2]). Learning G is equivalent to learning which coefficients of $\underline{\theta} \equiv \{\theta_{ij}\}$ are non-zero, i.e. learning the support of $\underline{\theta}$, i.e. $\text{supp}(\underline{\theta})$. The problem of learning G is a special case of the general estimation problem [78, 41]. In the general estimation problem, given samples $\{\underline{x}^{(\ell)}\}$ from a parametrized probability distribution $\mathbb{P}_{\underline{\theta}}(\underline{x})$, the objective is to compute an estimate $\hat{\underline{\theta}}$ for $\underline{\theta}$ under an appropriate loss function. Structural learning is estimation under the loss function

$$C(\underline{\theta}, \hat{\underline{\theta}}) = \mathbb{I}(\text{supp}(\underline{\theta}) \neq \text{supp}(\hat{\underline{\theta}})). \tag{1.2.3}$$

When related to graphical models, the estimation problem often goes under the simpler name of learning graphical models, to distinguish it from structural learning, and often the cost function assumed is the Euclidean norm of the difference between true and estimated parameters.

This difference in norms does change our problem from the usual estimation problem in a fundamental way. In particular, there is a difference between knowing whether a given coefficient is approximately zero (approximating $\underline{\theta}$) and knowing whether a coefficient is exactly zero, i.e., estimating G. Recovering $G = (V, E)$ from data allows us, for example, to find a sets $A, B, C \in V$ such that, conditioned on the variables in C, the variables in A and B are independent. However, from a set real parameters $\{\theta_{ij}\}$, some of smaller magnitude than others, it is unclear how we should select such sets.

In addition, in parameter learning for graphical models we know from the start the structure of the solution, i.e., which coefficients are non-zero [86]; while in this thesis, finding them is our objective.

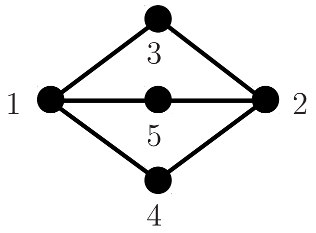

Figure 1.1: Topology of G_5: $\mathbb{P}_{G_5}(x_1, ..., x_5) = \exp(x_1x_3 + x_1x_4 + x_1x_5 + x_2x_3 + x_2x_4 + x_2x_5)$.

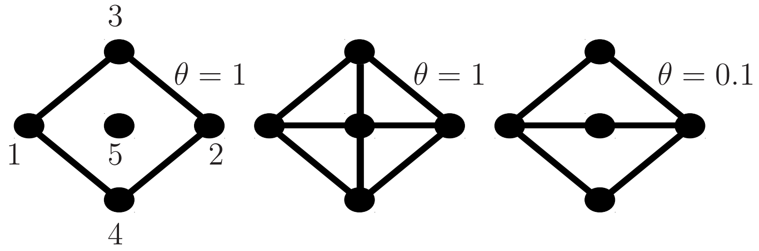

Figure 1.2: From left to right: Graph for variant 1, variant 2 and variant 3.

1.2.1 Sample-complexity and computational-complexity

To illustrate some of the challenges of structural learning we now focus on a particular instance of the model (1.1.1). There is a set of five nodes connected by two kinds of links: any two nodes are either connected, $\theta_{ij} = 1$, or disconnected, $\theta_{ij} = 0$. In this model, the graph, G_5, has the topology shown in Figure 1.1. Since (1.1.1) is an exponential family, given n i.i.d samples from $\mathbb{P}_{G_5}(\underline{x})$, the set of empirical covariances $\hat{C}_{ij} = (1/n)\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}$ is a sufficient statistic to recover G_5 [78]. As a first attempt to recover G_5, we compute all $\hat{C}_{ij}$ and use the following threshold rule: If $\hat{C}_{ij} > \tau$, we conclude that $(i, j) \in E$; otherwise, we do not. However, even in the favorable case where $n = \infty$, one can see that such an attempt would not work. For $n = \infty$, $\hat{C}_{12} \approx 0.963 > \hat{C}_{13} \approx 0.946$ and the threshold rule says either that both edges $(1, 2)$ and $(1, 3)$ are in E or that E excludes both. In either case, we do not recover G_5 correctly. Consider now the following three variants of the network of Figure 1.1. These variants are illustrated in Figure 1.2.

- Variant 1 (topological change): Start from G_5 and remove edges $(1, 5)$ and $(5, 2)$.

- Variant 2 (topological change): Start from G_5 and add edges $(3,4)$ and $(4,5)$.
- Variant 3 (edge-strength change): Start from G_5 and reduce the edge-weights from $\theta_{ij} \in \{0,1\}$ to $\theta_{ij} \in \{0,0.1\}$.

Surprisingly, for these three variants, at least when $n = \infty$, we can recover the underlying graph from this simple threshold rule. When $n = \infty$, successful reconstruction by thresholding for variant 1 and variant 3 is equivalent, by symmetry, to $\hat{C}_{13} > \max\{\hat{C}_{34}, \hat{C}_{12}\}$. Namely, for variant 1, $\hat{C}_{13} \approx 0.100$, $\hat{C}_{34} \approx 0.020$ and $\hat{C}_{12} \approx 0.020$ and for variant 3, $\hat{C}_{13} \approx 0.102$, $\hat{C}_{34} \approx 0.020$ and $\hat{C}_{12} \approx 0.030$. For variant 2, when $n = \infty$, the success is equivalent, by symmetry, to $\min\{\hat{C}_{13}, \hat{C}_{34}\} > \hat{C}_{12}$. Correspondingly, $\hat{C}_{13} \approx 0.989$, $\hat{C}_{34} \approx 0.993$ and $\hat{C}_{12} \approx 0.987$.

The above examples show that while some graphs are recoverable by fairly simple algorithms, other similar graphs might be harder to learn. Because of this, most algorithms proposed in the literature are only guaranteed to work under specific restrictions on the class of graphs [3, 21, 94, 1]. In addition, the class of graphs that in principle it might be possible to learn does not seem to have a simple characterization. Making a graph denser (by adding edges) or sparser (by removing edges) sometimes makes a particular algorithm succeed and other times fail. Furthermore, not only the topology, but also the edges-weights, must be taken into account.

There are other nuances: among graphs that are recoverable with simple algorithms there are also differences. The gap between the correlation of connected and disconnected nodes is greater than 0.07 in variants 1 and 3 but in variant 2 is smaller than 0.002. Using the thresholding algorithm to recover variant 1 and 3 therefore requires using less samples (smaller n) than to recover variant 2. The notions of *sample-complexity* and *computational-complexity* are introduced to quantify these differences. Given an algorithm Alg that receives as input n samples from $\mathbb{P}_G(\underline{x})$ and outputs a graph $\hat{G}$, the sample-complexity is defined by

$$N_{\mathsf{Alg}}(G) \equiv \min\left\{n_0 \in \mathbb{N} : \mathbb{P}_{G,n}\{\hat{G} = G\} \geq 1 - \delta \text{ for all } n \geq n_0\right\} \tag{1.2.4}$$

where $\mathbb{P}_{G,n}$ denotes probability with respect to n i.i.d. samples with distribution $\mathbb{P}_{G,n}$. The computational complexity $\chi_{\mathsf{Alg}}(G)$ is the running time of Alg on an input of size

$n = N_{\mathsf{Alg}}(G)$. The question generally addressed in this thesis can be expressed using these two quantities:

Question. *What are the values of $N_{\mathsf{Alg}}(G)$ and $\chi_{\mathsf{Alg}}(G)$ for different families of graphs and algorithms?*

We are particularly interested in understanding how $N_{\mathsf{Alg}}(G)$ and $\chi_{\mathsf{Alg}}(G)$ scale with the size p of G (recall $|V| = p$). This is different from the classical learning setting where $n \gg p$. The recent explosion of data collection and storage puts many applications in the regime $n \gg p$, but for many problems, the number of variables involved is unavoidably greater than the amount of data. To attack these problems, we wish to find algorithms for which both the sample-complexity and computational-complexity scale slowly with p. How slowly can these quantities scale? As a reference for computational-complexity, notice that the thresholding algorithm used above, one of the simplest algorithms we can think of, has a running time that scales likes $O(np^2)$ [1]. As a reference for sample-complexity, consider the minimum number of i.i.d. samples needed to identify a graph among the class of graphs of degree bounded by Δ. Each sample on a graph with p nodes gives p bits of information and the number of graphs of degree bounded by Δ on p nodes is $O\left(\binom{p}{\Delta}^p\right)$ hence $pn = O\left(p \log \binom{p}{\Delta}\right)$ or $n = O(\Delta \log p)$.

1.2.2 Dependencies in data

In our discussion of learning the Ising model, we have assumed data are composed of i.i.d. samples from the constructed probability distribution. However, in many applications data are gathered in time and samples are not independent but correlated. This reality brings additional questions to the problem of structural learning. This thesis addresses some of them. Going back to the Ising model, a simple dynamical model can be obtained by constructing a Gibbs sampler for (1.1.1) [51, 24]: If at time

[1] There are p^2 empirical correlations to be computed each taking $O(n)$ time steps to be computed.

t the configuration is $\underline{x}(t)$, we form the configuration at time $t+1$, $\underline{x}(t+1)$, by choosing a node i uniformly at random from V and 'flipping' its value with probability $\min\{1, \exp(-2x_i \sum_j \theta_{ij} x_j(t))\}$. This dynamical model is a reversible Markov chain with unique stationary measure equal to the model (1.1.1). If we have an algorithm Alg for learning G that provably works when the input values are i.i.d. samples, one can now think of circumventing the correlation among samples by inputting to Alg the subsequence $\underline{x}(m), \underline{x}(2m), ..., \underline{x}(m\lfloor n/m \rfloor)$. If m is sufficiently large, $\{\underline{x}(im\}$ is close to set of i.i.d. samples and we expect that $\mathsf{Alg}(\{\underline{x}(im\}) = G$ with high probability. But is this the best we can do? Discarding information increases the sample-complexity, hence the general question:

Question. *How does $N_{\mathsf{Alg}}(G)$ change when learning is done from correlated samples?*

The answer to this question depends on the stochastic process generating the correlated samples. In this thesis we study an extreme case of structural learning with correlations in data: dynamical processes in continuous time. In particular, our focus is on *stochastic differential equations* parametrized by graphs. A good introduction to SDEs is found in [88]. At this point it is easier to have in mind a concrete example. One of the simplest SDE models represents the fluctuation of the value of a node i, $x_i(t)$, as a linear combination of the fluctuation of the value of neighboring nodes in a graph G. More precisely, given a weighted graph $G = (V, E)$ with $V = [p]$, $E = \{(i, j) : \theta_{ij} \neq 0\}$, we define

$$\mathrm{d}x_i(t) = \sum_j \theta_{ij} x_j(t) \mathrm{d}t + \mathrm{d}b_i(t), \tag{1.2.5}$$

where $\underline{b}(t)$ is a p-dimensional standard Brownian motion. The above equation is a linear stochastic differential equation and is also a graphical model: the support of the matrix $\Theta = \{\theta_{ij}\}$ defines the adjacency matrix of the graph G. Given the evolution of these values, $\underline{x}(t)$, in a time window of length T, the particular question we are interested in is:

Question. *It is possible to recover G from $\{\underline{x}(t)\}_{t=0}^{T}$?*

Unlike in the previous questions, no matter how small T is, we now have at our disposal an infinite number of samples, since the trajectories are continuous. Perhaps, then, we only need an arbitrarily small time window to recover G? But samples obtained close it time exhibit a strong correlation, and it is reasonable to expect that larger graphs (bigger p) require longer observation time windows to be recovered. Hence, perhaps T scales like $O(p)$? Given an algorithm Alg that receives as input a trajectory $\{\underline{x}(t)\}_{t=0}^{T}$ and outputs a graph $\hat{G}$, we introduce the following modified notion of sample-complexity for SDEs,

$$T_{\mathsf{Alg}}(G) = \inf\big\{T_0 \in \mathbb{R}^+ : \mathbb{P}_{G,T}\{\hat{G} = G\} \geq 1 - \delta \text{ for all } T \geq T_0\big\}, \tag{1.2.6}$$

where $\mathbb{P}_{G,T}$ denotes probability with respect to a trajectory of length T. The problem of structural learning for SDEs that we address in this thesis is now precisely defined:

Question. *What are the values of $T_{\mathsf{Alg}}(G)$ and $\chi_{\mathsf{Alg}}(G)$ for different families of SDEs and algorithms?*

Here also, our focus is on the regime of 'large-systems and few data', or $p \gg T$. The more classical problem of estimating the values of Θ [76, 12], known as *system-identification* or *drift-estimation*, is again related to the problem we address, under a suitable choice of error function. However, past work has not focused on support recovery.

1.3 Contributions

Apart from the Introduction and Conclusion, this thesis is divided into two main chapters. Chapter 2 concerns learning the structure the Ising model. This model is interesting because of its applications (e.g., in computer vision and biology) and its simplicity: it is the simplest model on binary variables for which the set of all pair-wise correlations forms a sufficient statistic. While several methods have been proposed to recover G in the Ising model, their relative merits and limitations remain somewhat obscure. We analyze three different reconstruction algorithms and relate their success or failure to a single criterion [68]:

Contribution. *When the Ising model develops long-range correlations these algorithms fail to reconstruct G in polynomial time ($\chi_G = O(poly(p))$).*

More concretely, for three polynomial time algorithms, and for Ising models of bounded degree Δ with homogeneous edge-weights of strength θ, there are constants C and C' such if $\Delta\theta < C$ then $N_{\mathsf{Alg}}(G) = O(\log p)$ and if if $\Delta\theta > C'$ then $N_{\mathsf{Alg}}(G) = \infty$.

Among the three algorithms studied, most of our focus was on the regularized logistic regression algorithm introduced by by Ravikumar et al. [94]. [94]. [94] proved a set sufficient conditions under which this algorithm exactly learns Ising models, the most interesting being a so-called *incoherence condition*. In Chapter 2, we show that this incoherence condition is also necessary and analytically establish whether it holds or not for several families of graphs. In particular, for this algorithm we obtain a sharp characterization of the two unspecified constants above [16].

Contribution. *Regularized logistic regression succeeds on any graph with $\Delta\theta \leq 3/10$ and fails on* most *regular graphs with $\Delta\theta \geq 2$.*

These results are well illustrated by Figure 1.3. The plot illustrates the probability of successful reconstruction of regular graphs of degree 4 using regularized logistic regression as a function of θ when $\theta_{ij} \in \{0, \theta\}$. When $\theta > \theta_c$, it is no longer possible to learn G. θ_c is the critical temperature of the lattice and, for regular graphs, scales like $1/\Delta$, just as predicted by our sharp bounds.

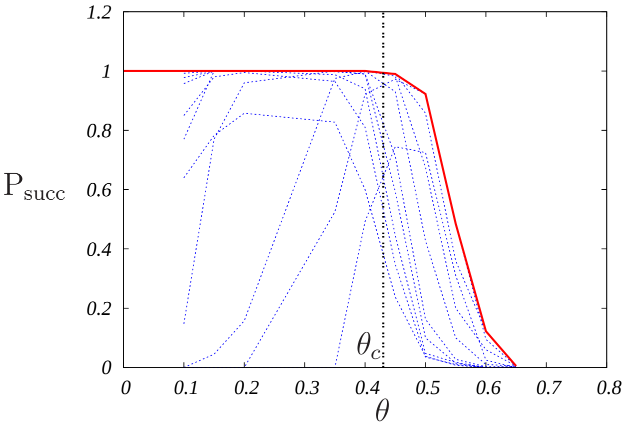

Figure 1.3: Learning uniformly generated random regular graphs of degree $\Delta = 4$ for the Ising model from samples using regularized logistic regression. Red curve: success probability as a function of the edge-strength, i.e. $\theta_{ij} \in \{0, \theta\}$.

Chapter 3 treats of learning systems of SDEs. There we prove a general lower bound on the sample-complexity $T_{\mathsf{Alg}}(G)$ by using a characterization of mutual information as time a integral of conditional variance, due to Kadota, Zakai, and Ziv. For linear SDEs, we analyze an ℓ_1-regularized least-squares algorithm, Rls, and prove an upper bound on $T_{\mathsf{Rls}}(G)$ which nearly matches the general lower bound [14, 15, 13].

Contribution. *For linear and stable SDEs, if G has maximum degree bounded by Δ then $T_{\mathsf{Rls}}(G) = O(\log p)$ and any algorithm* Alg *with probability of success greater than $1/2$ on this class of graphs has $T_{\mathsf{Alg}}(G) = \Omega(\log p)$. If G is a dense graph then $T_{Rls}(G) = O(p)$ and any algorithm* Alg *with probability of success greater than $1/2$ on this class of graphs has $T_{\mathsf{Alg}}(G) = \Omega(p)$. In both cases, the upper bound is achieved by Rls and $\chi_{Rls} = O(poly(p))$.*

Although our theoretical results only apply for linear SDEs, the algorithm proposed has much greater applicability. In particular, it seems to be able to learn even non-linear SDEs. This result is summarized in Figure 1.4 for the case of sparse graphs. As the figure shows, even for non-linear SDEs, the sample-complexity in

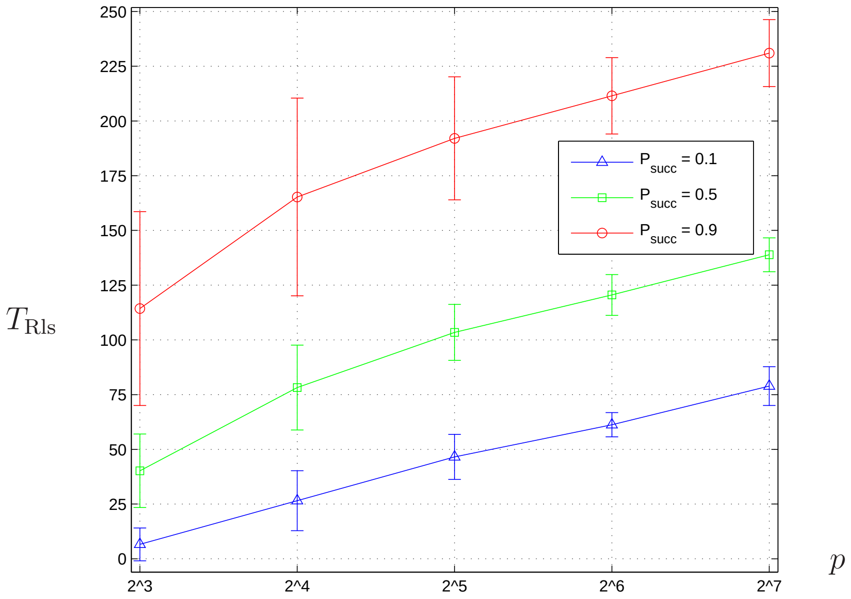

Figure 1.4: Reconstruction of non-linear SDEs. Curves show minimum observation time $T_{\rm Rls}$ required to achieve a probability of reconstruction of $\rm P_{succ} = 0.1, 0.5$ and 0.9 versus the size of the network p. All non-linear SDEs are associated to random regular graphs of degree 4 sampled uniformly at random. The points in the plot are averages over different graphs and different SDEs trajectories.

reconstructing sparse graphs scales like $O(\log p)$.

1.4 Notation

Throughout this thesis, notation is introduced as needed. However, for convenience, we summarize here the main conventions used, unless stated otherwise.

$\mathbb{N}$	Set of natural numbers $\{1, 2, ...\}$;
$\mathbb{Z}$	Set of integer numbers $\{..., -2, -1, 0, 1, 2, ...\}$;
$\mathbb{R}$	Set of real numbers;
$[i]$	Subset of integer numbers $\{1, 2, ..., i\}$;
$\mathrm{supp}(.)$	Support of a vector/matrix, i.e., the set of indices with non-zero values;
$\mathrm{sign}(.)$	Signed support of a vector/matrix, i.e., the set of indices with positive values and the set of indices with negative values;
$A^C \subseteq B$	If A is a subset of B then A^C is the complement of A in B. It will be clear from the context what B is;
$\mathbb{1}$	All-ones vector;
$\mathbb{I}$	Identity matrix;
v^*,M^*	Transpose of vector v or matrix M;
$\lVert.\rVert_0, \lVert.\rVert_1, \lVert.\rVert_2, \lVert.\rVert_F, \lVert.\rVert_\infty$	0-norm, 1-norm, euclidean norm, Frobenius norm, infinity norm;
$\Lambda_{\max}(M), \Lambda_{\min}(M)$	Maximum and minimum eigenvalue of matrix M;
$\sigma_{\max}(M), \sigma_{\min}(M)$	Maximum and minimum singular of matrix M;
$\langle v, w \rangle$	Inner product of vector v and w (dot-product in euclidean space);
$\mathrm{Tr}(A)$	Trace of matrix A;
$\lvert A \rvert$	Determinant of matrix A;
v_A, M_{AB}	If A and B are subset of indices then v_A is the vector formed by the entries of v whose indices are in A and M_{AB} is matrix formed by the entries of M whose

	row indices are in A and column indices are in B;
∇f	Gradient of function f;
$\mathsf{Hess}(f)$	Hessian of function f;
$J(f)$	Jacobian of function f;
$\theta, \underline{\theta}, \Theta$	Parameters describing probability distributions;
$\underline{\theta}^0, \Theta^0$	Unknown value of parameters whose support estimation is this thesis' main focus;
$\hat{\underline{\theta}}, \hat{\Theta}$	Estimate of parameters $\underline{\theta}^0$ and Θ^0;
S, S^0	support of $\underline{\theta}$ and $\underline{\theta}^0$;
$\hat{S}$	Estimate of S^0;
z^0	Sub-gradient of $\|\underline{\theta}\|_1$ evaluated at $\underline{\theta}^0$;
$\hat{z}$	Sub-gradient of $\|\underline{\theta}\|_1$ evaluated at $\hat{\underline{\theta}}$;
$\mathbb{P}_{(.)}$	Probability distribution parameterized by (.);
'i.i.d.'	Independent and identically distributed;
$\mathbb{P}_{(.),n}$	Probability distribution of n i.i.d samples from $\mathbb{P}_{(.)}$;
$\mathbb{E}_{(.)}$	Expected value over the probability distribution $\mathbb{P}_{(.)}$;
$\mathbb{E}_{(.),n}$	Expected value over the probability distribution $\mathbb{P}_{(.),n}$;
$\mathrm{Var}_{(.)}$	Variance with respect to the probability distribution $\mathbb{P}_{(.)}$;
$\mathrm{P}_{\mathrm{succ}}$	Probability of successful reconstruction of a graph $G = (V, E)$ (successful reconstruction means full exact recovery of E);
$\underline{b}(t)$	Standard Brownian motion;
$\underline{x}, \underline{X}$	Sample from $\mathbb{P}_{(.)}$ (deterministic and random variables respectively);
$\underline{x}^{(\ell)}, \underline{X}^{(\ell)}, \underline{x}(t), \underline{X}(t)$	Samples from $\mathbb{P}_{(.)}$ indexed by integer number ℓ and indexed by real number t;
X_0^n, X_0^T	Samples $\{\underline{x}^{(\ell)}\}_{\ell=0}^n$ and $\{\underline{x}(t)\}_{t=0}^T$;
$G = (V, E)$	Graph with edge set E and vertex set V;
Δ^G	Laplacian of graph G;

$\partial r \subseteq V$	Neighborhood of node $r \in V$ of a graph G;
$\deg(r)$	Degree of node r, i.e., number of neighbors of node $r \in V$ of a graph G;
Δ	Maximum degree across all nodes in a graph G;
p	Number of nodes in graph G, the dimension of $\underline{\theta}$ or the 'width' of a matrix $\Theta \in \mathbb{R}^{m \times p}$;
$\mathcal{G}_{\rm one}$	Family of 'one-edge' graphs introduced in Section 2.6.3;
$\mathcal{G}_{\rm diam}$	Family of 'diamond' graphs introduced in Section 2.6.3;
$\mathcal{G}_{\rm rand}$	Family of regular graphs introduced in Section 2.6.3;
n	Number of samples being used for reconstruction;
T	Length of trajectory being used for reconstruction;
Alg	General reconstruction algorithm;
Thr	Thresholding algorithm of Section 2.3.1;
Ind	Conditional independence algorithm of Section 2.3.2;
Rlr	Regularized logistic regression algorithm of Section 2.3.3;
Rls	Regularized least squares algorithm of Section 3.3.2;
λ	Regularization parameter;
$N_{\mathsf{Alg}}(G)$	Minimum number of samples that Alg requires to reconstruct the graph G of a specific graphical model;
$N_{\mathsf{Alg}}(G, \theta)$	Minimum number of samples that Alg requires to reconstruct G for an homogeneous-edge-strength Ising model of edge-strength θ;
$N_{\mathsf{Alg}}(p, \Delta, \theta)$	Minimum number of samples required by Alg to reconstruct any graph G with p nodes and maximum degree Δ for an homogeneous-edge-strength Ising model of edge-strength θ;

$T_{\mathsf{Alg}}(G)$	Minimum observation time that Alg requires to reconstruct the graph G of a SDE parametrized by G;
$T_{\mathsf{Alg}}(\Theta)$	Minimum observation time that Alg requires to reconstruct the signed support of Θ of a SDE parametrized by Θ;
$T_{\mathsf{Alg}}(\mathcal{A})$	Minimum observation time required by Alg to reconstruct a family of SDEs denoted by $\mathcal{A}$;
$\chi_{\mathsf{Alg}}(G)$	Number of computation steps that Alg requires to reconstruct the graph G of a specific graphical model when run on $N_{\mathsf{Alg}}(G)$ samples;
$\chi_{\mathsf{Alg}}(G,\theta)$	Number of computation steps that Alg requires to reconstruct G for an homogeneous-edge-strength Ising model of edge-strength θ when run on $N_{\mathsf{Alg}}(G,\theta)$ samples;
$\chi_{\mathsf{Alg}}(p,\Delta,\theta)$	Number of computation steps required to reconstruct any graph G with p nodes and maximum degree Δ for an homogeneous-edge-strength Ising model of edge-strength θ when run on $N_{\mathsf{Alg}}(p,\Delta,\theta)$ samples;
SDE	Stochastic differential equation;
NRMSE	Normalized root mean squared error $(\|\Theta^0 - \hat{\Theta}\|_2/\|\Theta^0\|_2)$;

Chapter 2

Learning the Ising model

This chapter is devoted to studying the sample-complexity and computational-complexity of learning the Ising model for a number of reconstruction algorithms and graph models. The Ising model has already appeared in Chapter 1 in equation (1.1.1), but is introduced in more detail in Section 2.1. In particular, we consider homogeneous edge-strengths, i.e. $\theta_{ij} \in \{0, \theta\}$, and graphs of maximum degree bounded by Δ.

Results of this analysis are presented in Section 2.3 for three algorithms. A simple thresholding algorithm is discussed in Section 2.3.1. In Section 2.3.2, we look at the conditional independence test method of [21]. Finally, in Section 2.3.3, we study the penalized pseudo-likelihood method of [94]. In Section 2.5, we validate our analysis through numerical simulations, and Section 2.6 contains the proofs of these conclusions, with some technical details deferred to the appendices.

Our analysis unveils a general pattern: *when the model develops strong correlations, several low-complexity algorithms fail, or require a large number of samples.* What does 'strong correlations' mean? Correlations arise from a trade-off between the degree (which we characterize here via the maximum degree Δ), and the interaction strength θ. It can be ascribed to a few strong connections (large θ) or to a large number of weak connections (large Δ). Is there any meaningful way to compare and combine these quantities (θ and Δ)? An answer is suggested by the theory of Gibbs measures which predicts a dramatic change of behavior of the Ising model when θ crosses the so-called 'uniqueness threshold' $\theta_{\text{uniq}}(\Delta) = \text{atanh}(1/(\Delta - 1))$ [52]. For

$\theta < \theta_{\text{uniq}}(\Delta)$, Gibbs sampling mixes rapidly and far-apart variables in G are roughly independent [85]. Conversly, for any $\theta > \theta_{\text{uniq}}(\Delta)$, there exist graph families on which Gibbs sampling is slow, and far-apart variables are strongly dependent [53]. While polynomial sampling algorithms exist for all $\theta > 0$ [70], for $\theta < 0$, in the regime $|\theta| > \theta_{\text{uniq}}(\Delta)$ sampling is #-P hard [101]. Related to the uniqueness threshold is the critical temperature, which is graph-dependent, with typically $\theta_{\text{crit}} \leq \text{const.}/\Delta$.

In this chapter we see that the theory of Gibbs measure is indeed a relevant way of comparing interaction strength and graph degree for the problem of structural learning. All the algorithms we analyzed provably fail for $\theta \gg \text{const.}/\Delta$, for a number of 'natural' graph families. This chapter raises several fascinating questions, the most important being the construction of structural learning algorithms with provable performance guarantees in the strongly dependent regime $\theta_{\text{crit}} \gg \text{const.}/\Delta$. The question as to whether such an algorithm exists is left open by the present thesis (but see Section 2.2 for an overview of earlier work).

Let us finally emphasize that we do not think that any of the specific families of graphs studied in the present thesis is intrinsically 'hard' to learn. For instance, we show in Section 2.3.3 that the regularized logistic regression method of [94] fails on random regular graphs, while it is easy to learn such graphs using the simple thresholding algorithm of Section 2.3.1. The specific families were indeed chosen mostly because they are analytically tractable.

The work in this chapter is based on joint work with Montanari [68, 16].

2.1 Introduction

Given an undirected graph $G = (V = [p], E)$, and a positive parameter $\theta > 0$, the *ferromagnetic Ising model on G* is the pair-wise Markov random field

$$\mathbb{P}_{G,\theta}(\underline{x}) = \frac{1}{Z_{G,\theta}} \prod_{(i,j)\in E} e^{\theta x_i x_j} \tag{2.1.1}$$

over binary variables $\underline{x} = (x_1, x_2, \ldots, x_p)$, $x_i \in \{+1, -1\}$. Apart from being one of the best-studied models in statistical mechanics [66, 56], the Ising model is a prototypical undirected graphical model. Since the seminal work of Hopfield [65] and Hinton and Sejnowski [62], it has found application in numerous areas of machine learning, computer vision, clustering and spatial statistics. The obvious generalization of the distribution (2.1.1) to edge-dependent parameters θ_{ij}, $(i,j) \in E$ is of central interest in such applications

$$\mathbb{P}_{\underline{\theta}}(\underline{x}) = \frac{1}{Z_{\underline{\theta}}} \prod_{(i,j)\in E(K_p)} e^{\theta_{ij}x_i x_j}, \qquad (2.1.2)$$

where $E(K_p) \equiv \{(i,j) : \ i,j \in V\}$ is the edge set of the complete graph and $\underline{\theta} = \{\theta_{ij}\}_{(i,j)\in E(K_p)}$ is a vector of real parameters. The support of the parameter $\underline{\theta}$ specifies a graph. In fact, model (2.1.1) corresponds to $\theta_{ij} = 0$, $\forall(i,j) \notin E$ and $\theta_{ij} = \theta$, $\forall(i,j) \in E$. Let us stress that we follow the statistical mechanics convention of calling (2.1.1) an Ising model even if the graph G is not a grid.

In this section we focus on the following structural learning problem:

> *Given n i.i.d. samples $\underline{x}^{(1)}$, $\underline{x}^{(2)}$,..., $\underline{x}^{(n)} \in \{+1,-1\}^p$ with distribution $\mathbb{P}_{G,\theta}(\,\cdot\,)$, reconstruct the graph G.*

For the sake of simplicity, we assume the parameter θ is known, and that G has no double edges (it is a 'simple' graph). It follows from the general theory of exponential families that, for any $\theta \in (0,\infty)$, the model (2.1.1) is identifiable [78]. In particular, the structural learning problem is solvable with unbounded sample complexity and computational resources. The question we address is: for which classes of graphs and values of the parameter θ is the problem solvable under realistic complexity constraints? More precisely, given a graph G, an algorithm Alg that outputs an estimate $\widehat{G} = \mathsf{Alg}(\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(n)})$, a value θ of the model parameter, and a small $\delta > 0$, the sample complexity is defined as

$$N_{\mathsf{Alg}}(G,\theta) \equiv \min\left\{n_0 \in \mathbb{N} : \mathbb{P}_{G,\theta,n}\{\widehat{G} = G\} \geq 1-\delta \text{ for all } n \geq n_0\right\}, \qquad (2.1.3)$$

where $\mathbb{P}_{G,\theta,n}$ denotes probability with respect to n i.i.d. samples with distribution $\mathbb{P}_{G,\theta}$. Further, we let $\chi_{\mathsf{Alg}}(G,\theta)$ denote the number of operations of the algorithm Alg, when applied to $N_{\mathsf{Alg}}(G,\theta)$ samples. The general problem is therefore to characterize the functions $N_{\mathsf{Alg}}(G, \theta)$ and $\chi_{\mathsf{Alg}}(G, \theta)$, and to design algorithms that minimize the complexity.

Let us emphasize that these are not the only possible definitions of sample and computational complexity. Alternative definitions are obtained by requiring that the reconstructed structure $\mathsf{Alg}(\underline{x}^{(1)}, \dots, \underline{x}^{(n)})$ is only partially correct. However, for the algorithms considered in this paper, such definitions should not result in qualitatively different behavior[1]

General upper and lower bounds on the sample complexity $N_{\mathsf{Alg}}(G, \theta)$ were proved by Santhanam and Wainwright [108, 98], without however taking into account computational complexity. At the other end of the spectrum, several low complexity algorithms have been developed in the last few years (see Section 2.2 for a brief overview). Yet the resulting sample complexity bounds only hold under specific assumptions on the underlying model (i.e., on the pair (G, θ)). A general understanding of the trade-offs between sample complexity and computational complexity is largely lacking.

This paper is devoted to the study of the tradeoff between sample complexity and computational complexity for some specific structural learning algorithms, when applied to the Ising model. An important challenge consists in the fact that the model (2.1.1) induces subtle correlations between the binary variables $(x_1, \dots, x_p)$. The objective of a structural learning algorithm is to disentangle pairs x_i, x_j that are conditionally independent given the other variables (and hence are not connected by an edge) from those that are instead conditionally dependent (and hence connected by an edge in G). This becomes particularly difficult when θ becomes large and hence pairs x_i, x_j that are not connected by an edge in G become strongly dependent. The next section sets the stage for our work by discussing a simple and concrete illustration of this phenomenon.

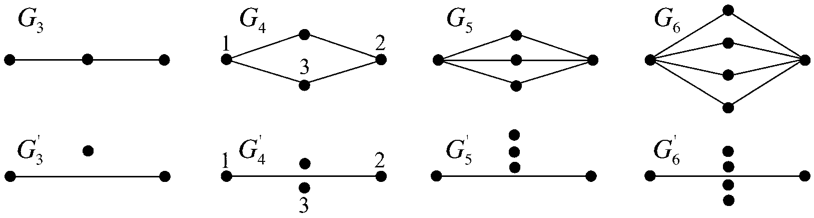

Figure 2.1: Two families of graphs G_p and G_p' whose distributions $\mathbb{P}_{G_p,\theta}$ and $\mathbb{P}_{G_p',\theta'}$ merge as p gets large.

2.1.1 A toy example

As a toy illustration[2] of the challenges of structural learning, we will study the two families of graphs in Figure 2.4. The two families will be denoted by $\{G_p\}_{p\geq 3}$ and $\{G_p'\}_{p\geq 3}$ and are indexed by the number of vertices p. Later on, in Section 2.6.3, the family $\{G_p\}$ will again be studied under the name of $\mathcal{G}_{\rm diam}(p)$.

Graph G_p has p vertices and $2(p-2)$ edges. Two of the vertices (vertex 1 and vertex 2) have degree $(p-2)$, and $(p-2)$ have degree 2. Graph G_p' has also p vertices, but only one edge between vertices 1 and 2. In other words, graph G_p' corresponds to variables x_1 and x_2 interacting 'directly' (and hence not in a conditionally independent way), while graph G_p describes a situation in which the two variables interact 'indirectly' through numerous weak intermediaries (but they are still conditionally independent since they are not connected). Fix p, and assume that one of G_p or G_p' is chosen randomly and i.i.d. samples $\underline{x}^{(1)},\dots,\underline{x}^{(n)}$ from the corresponding Ising distribution are given to us.

Can we efficiently distinguish the two graphs, i.e., infer whether the samples were generated using G_p or G_p'? As mentioned above, since the model is identifiable, this task can be achieved with unbounded sample and computational complexity. Further, since model (2.1.1) is an exponential family, the $p\times p$ matrix of empirical covariances $(1/n)\sum_{\ell=1}^{n}\underline{x}^{(\ell)}(\underline{x}^{(\ell)})^T$ provides a sufficient statistic for inferring the graph structure.

In this specific example, we assume that different edge strengths are used in the

[1] Indeed the algorithms considered in this paper reconstruct G by separately estimating the neighborhood of each node i. This implies that any significant probability of error results in a substantially different graph.

[2] A similar example was considered in [87].

two graphs: θ for graph G_p and θ' for graph G'_p (i.e. we have to distinguish between $\mathbb{P}_{G_p,\theta}$ and $\mathbb{P}_{G'_p,\theta'}$). We claim that, by properly choosing the parameters θ and θ', we can ensure that the covariances approximately match $|\mathbb{E}_{G_p,\theta}\{x_i x_j\} - \mathbb{E}_{G'_p,\theta'}\{x_i x_j\}| = O(1/\sqrt{p})$. Indeed the same remains true for all marginals involving a bounded number of variables. Namely, for all subsets of vertices $U \subseteq [p]$ of bounded size $|\mathbb{P}_{G_p,\theta}(\underline{x}_U) - \mathbb{P}_{G'_p,\theta'}(\underline{x}_U)| = O(1/\sqrt{p})$. Low-complexity algorithms typically estimate each edge using only a small subset low–dimensional marginal. Hence, they are bound to fail unless the number of samples n diverges with the graph size p. On the other hand, a naive information-theoretic lower bound (in the spirit of [108, 98]) only yields $N_{\mathsf{Alg}}(G,\theta) = \Omega(1)$. This sample complexity is achievable by using global statistics to distinguish the two graphs.

In other words, even for this simple example, a dichotomy emerges: either the number of samples has to grow with the number of parameters, or the algorithms have to exploit a large number of marginals of $\mathbb{P}_{G,\theta}$.

To confirm our claim, we need to compute the covariances of the Ising measures distributions $\mathbb{P}_{G_p,\theta}$, $\mathbb{P}_{G'_p,\theta'}$. We easily obtain, for the latter graph

$$\mathbb{E}_{G'_p,\theta'}\{x_1 x_2\} = \tanh\theta' \,, \tag{2.1.4}$$

$$\mathbb{E}_{G'_p,\theta'}\{x_i x_j\} = 0\,. \qquad (i,j) \neq (1,2)\,. \tag{2.1.5}$$

The calculation is somewhat more intricate for graph G_p. The details can be found in [67]. Here we report only the result for $p \gg 1$, $\theta \ll 1$:

$$\mathbb{E}_{G_p,\theta}\{x_1 x_2\} = \tanh\left\{p\theta^2 - O(p\theta^4)\right\}\,, \tag{2.1.6}$$

$$\mathbb{E}_{G_p,\theta}\{x_i x_j\} = O(\theta + p\theta^3)\,, \qquad i \in \{1,2\}, j \in \{3,\dots,p\}\,, \tag{2.1.7}$$

$$\mathbb{E}_{G_p,\theta}\{x_i x_j\} = O(\theta^2 + p\theta^4)\,, \qquad i,j \in \{3,\dots,p\}\,. \tag{2.1.8}$$

In other words, variables x_1 and x_2 are strongly correlated (although not connected), while all the other variables are weakly correlated. By letting $\theta = \sqrt{\theta'/p}$ this covariance structure matches Eqs. (2.1.4), (2.1.5) up to corrections of order $1/\sqrt{p}$.

Notice that the ambiguity between the two models G_p and G'_p arises because several weak, indirect paths between x_1 and x_2 in graph G_p add up to the same effect as a strong direct connection. This toy example is hence suggestive of the general phenomenon that strong long-range correlations can 'fake' a direct connection. However, the example is not completely convincing for several reasons:

i. Most algorithms of interest estimate each edge on the basis of a large number of low-dimensional marginals (for instance *all* pairwise correlations).

ii. Reconstruction guarantees have been proved for graphs with bounded degree [3, 21, 108, 98, 94], while here we are letting the maximum degree be as large as the system size. The graph is sparse but only on 'average'.

iii. It may appear that the difficulty in distinguishing graph G_p from G'_p is related to the fact that in the former we take $\theta = O(1/\sqrt{p})$. This is however the natural scaling when the degree of a vertex is large, in order to obtain a non-trivial distribution. If the graph G_p had θ bounded away from 0, this would result in a distribution $\mu_{G_p,\theta}(\underline{x})$ concentrated on the two antipodal configurations: all-$(+1)$ and all-(-1). Structural learning would be equally difficult in this case.

Despite these shortcommings, this model provides already a useful counter-example. In Appendix A.4.1 show why, even for bounded p (and hence θ bounded away from 0) the model G_p in Figure 2.1 'fools' the regularized logistic regression algorithm of Ravikumar, Wainwright and Lafferty [94]. Regularized logistic regression reconstructs G'_p instead of G_p.

2.2 Related work

Traditional algorithms for learning Ising models were developed in the context of Boltzmann machines [62, 4, 61]. These algorithms try to solve the maximum likelihood problem by gradient ascent. Estimating the gradient of the log-likelihood function requires to compute expectations with respect to the Ising distribution. In these works, expectations were computed using the Markov Chain Monte Carlo (MCMC) method, and more specifically Gibbs sampling.

This approach presents two type of limitations. First of all, it does not output a 'structure' (i.e. a sparse subset of the $\binom{p}{2}$ potential edges): because of approximation errors, it yields non-zero values for all the edges. This problem can in principle be overcome by using suitably regularized objective functions, but such a modified algorithm was never studied.

Second, the need to compute expectation values with respect to the Ising distribution, and the use of MCMC to achieve this goal, poses some fundamental limitations. As mentioned above, the Markov chain commonly used by these methods is simple Gibbs sampling. This is known to have mixing time that grows exponentially in the number of variables for $\theta > \theta_{\text{uniq}}(\Delta)$, and hence does not yield good estimates of the expectation values in practice. While polynomial sampling schemes exist for models with $\theta > 0$ [70], they do not apply to $\theta < 0$ or to general models with edge-dependent parameters θ_{ij}. Already in the case $\theta < 0$, estimating expectation values of the Ising distribution is likely to be #-P hard [100, 101].

Abbeel, Koller and Ng [3] first developed a method with computational complexity provably polynomial in the number of variables, for bounded maximum degree, and logarithmic sample complexity. Their approach is based on ingenious use of the Hammersley-Clifford representation of Markov random fields. Unfortunately, the computational complexity of this approach is of order $p^{\Delta+2}$ which becomes impractical for reasonable values of the degree and network size (and superpolynomial for Δ diverging with p). The algorithm by Bresler, Mossel and Sly [21] mentioned in Section 2.3.2 presents similar limitations, that the authors overcome (in the small θ regime) by exploiting the correlation decay phenomenon.

An alternative point of view consists in using standard regression methods. This approach was pionereed by Meinshausen and Bühlmann [82] in the context of Gaussian graphical models. More precisely, [82] proposes to reconstruct the graph G by sequentially reconstructing the neighborhood of each vertex $i \in V$. In order to achieve the latter, the observed values of variable x_i are regressed against the observed value of all the other variables, using ℓ_1-penalized least squares (a.k.a. the Lasso [107]). The neighborhood of i is hence identified with the subset of variables x_j, $j \in V \setminus i$ whose regression coefficients are non-vanishing. The regularized logistic regressionn method of [94] studied in the present paper extends the work of Meinshausen and Bühlmann [82] to non-Gaussian graphical models. Let us notice in passing that maximum likelihood or ℓ_1-regularized maximum likelihood are computationally tractable in the case of Gaussian graphical models [43].

More recently, several interesting results were obtained in research directions to the ones addressed in this thesis.

Anandkumar, Tan and Willsky [1, 9] considered Gaussian graphical models under a 'local- separation property', and proposed a conditional independence test that is effective under the so-called walk-summability condition. The latter can be thought of as a sufficient condition for correlation decay, and is hence related to the general theme of the present Chapter.

The same authors considered Ising models in [1, 8], and prove structural consistency of a conditional independence test under a condition $\theta_{\max} \leq \theta_0$. Here θ_0 depends on the graph family but is related once more to the correlation decay property. For instance, in the case of random regular graphs, they prove $\Delta \tanh \theta_0 = 1$ (while, as already stated, the correlation decay threshold is $(\Delta - 1) \tanh \theta = 1$). In the case of random irregular graphs, the average degree is showed to play a more important role (again, in correspondence with correlation decay).

The conditional independence tests of [1, 9, 8] have complexity $O(p^{\eta+2})$ with η depending on the graph family. For general graphs of maximum degree Δ, we have $\eta = \Delta$, but η can be significantly smaller for locally tree-like graphs.

In a recent paper, Jalali, Johnson and Ravikumar [2] study a reconstruction algorithm that optimizes the likelihood function (2.3.14) over sparse neighborhoods through a greedy procedure. They prove that this procedure is structurally consistent under weaker conditions than the one of [94], and has lower sample complexity, namely $n = O(\Delta^2 \log p)$. It would be interesting to investigate whether an analogous of Theorem 2.3.6 hold for this algorithm as well.

Finally, Cocco and Monasson [29] propose an 'adaptive cluster' heuristics and demonstrated empirically good performances for specific graph families, also in the highly correlated regime i.e. for $\theta\Delta$ large. A mathematical analysis of their method is lacking.

2.3 Main results

Our main results mostly concern learning Ising models of maximum degree bounded by Δ. As such, and before we proceed, it is convenient to introduce special notions of sample-complexity and computational-complexity.

First, consider an algorithm Alg whose full specification requires choosing a value for the set of parameters s in some domain $\mathcal{D}$. Strictly speaking, a priori, $\mathsf{Alg}(s)$ and $\mathsf{Alg}('s)$ can be different algorithms. In particular, $N_{\mathsf{Alg}(s)}(G,\theta)$ and $N_{\mathsf{Alg}(s')}(G,\theta)$might have different values. When dealing with algorithms whose output $\widehat{G}(s)$ depends on a free parameter $s \in \mathcal{D}$ that must be chosen, we use the following definition for sample-complexity

$$N_{\mathsf{Alg}}(G,\theta) \equiv \min\left\{n_0 \in \mathbb{N} : \max_{s\in\mathcal{D}} \mathbb{P}_{G,\theta,n}\{\widehat{G}(s) = G\} \geq 1-\delta \text{ for all } n \geq n_0\right\}. \tag{2.3.1}$$

This is to be distinguished from $N_{\mathsf{Alg}(s)}(G,\theta)$ for any particular $s \in \mathcal{D}$. Similarly, $\chi_{\mathsf{Alg}}(G,\theta)$ is defined as the running time of $\mathsf{Alg}(s^0)$ when running on $N_{\mathsf{Alg}}(G,\theta)$ samples and where $s^0 = \arg\max_{s\in\mathcal{D}} \mathbb{P}_{G,\theta,N_{\mathsf{Alg}}(G,\theta)}\{\widehat{G}(s) = G\}$. This is to be distinguished from $\chi_{\mathsf{Alg}(s)}(G,\theta)$ for any particular $s \in \mathcal{D}$.

Second, consider the family $\mathcal{G}(p,\Delta)$ of graphs on p nodes with maximum degree Δ and an algorithm Alg (dependent or not on free parameters) that attempts to reconstruct G from n i.i.d. samples from (2.1.1). We define (with a slight abuse of notation)

$$N_{\mathsf{Alg}}(p,\Delta,\theta) \equiv \max_{G\in\mathcal{G}(p,\Delta)} N_{\mathsf{Alg}}(G,\theta)\,. \tag{2.3.2}$$

In words, $N_{\mathsf{Alg}}(p,\Delta,\theta)$ is the minimax sample complexity for learning graphs with p vertices, maximum degree Δ and edge strength θ, using Alg [3]. Similarly, we define,

$$\chi_{\mathsf{Alg}}(p,\Delta,\theta) \equiv \max_{G\in\mathcal{G}(p,\Delta)} \chi_{\mathsf{Alg}}(G,\theta)\,. \tag{2.3.3}$$

[3] In fact, using Alg in the best possible way if there is a set of parameters s for tunning.

2.3.1 Simple thresholding algorithm

In order to illustrate the interplay between graph structure, sample-complexity and interaction strength θ, it is instructive to consider a simple example. The thresholding algorithm reconstructs G by thresholding the empirical correlations

$$\widehat{C}_{ij} \equiv \frac{1}{n} \sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)} , \tag{2.3.4}$$

for $i, j \in V$.

```
THRESHOLDING( samples $\{\underline{x}^{(\ell)}\}$, threshold $\tau$ )
1:  Compute the empirical correlations $\{\widehat{C}_{ij}\}_{(i,j)\in V\times V}$;
2:  For each $(i,j) \in V \times V$
3:      If $\widehat{C}_{ij} \geq \tau$, set $(i,j) \in E$;
```

We denote this algorithm by $\mathsf{Thr}(\tau)$. Notice that its complexity is dominated by the computation of the empirical correlations, i.e. $\chi_{\mathsf{Thr}}(G, \theta) = O(p^2 n)$. The sample complexity $N_{\mathsf{Thr}}(G, \theta)$ is bounded for specific classes of graphs as follows (for proofs see Section A.1).

Theorem 2.3.1. *If G is a tree, then*

$$N_{\mathsf{Thr}}(G, \theta) \leq \frac{32}{(\tanh\theta - \tanh^2\theta)^2} \, \log \frac{2p}{\delta} \, . \tag{2.3.5}$$

In particular $\mathsf{Thr}(\tau)$ with $\tau(\theta) = (\tanh\theta + \tanh^2\theta)/2$, achieves this bound.

Theorem 2.3.2. *If G has maximum degree $\Delta > 1$ and if $\theta < \operatorname{atanh}(1/(2\Delta))$ then*

$$N_{\mathsf{Thr}}(G, \theta) \leq \frac{32}{(\tanh\theta - \frac{1}{2\Delta})^2} \, \log \frac{2p}{\delta} \, . \tag{2.3.6}$$

Further, $\mathsf{Thr}(\tau)$ with the choice $\tau(\theta) = (\tanh\theta + (1/2\Delta))/2$ achieves this bound.

Theorem 2.3.3. *There exists a numerical constant K such that the following is true. If $\Delta > 3$ and $\theta > K/\Delta$, there are graphs of bounded degree Δ such that,*

$N_{\mathsf{Thr}}(G, \theta) = \infty$, *i.e. for these graphs the thresholding algorithm always fails with high probability regardless of the value of* τ.

These results confirm the idea that the failure of low-complexity algorithms is related to long-range correlations in the underlying graphical model. If the graph G is a tree, then correlations between far apart variables x_i, x_j decay exponentially with the distance between vertices i, j. Hence trees can be learnt from $O(\log p)$ samples irrespectively of their topology and maximum degree (assuming $\theta \neq \infty$). The same happens on bounded-degree graphs if $\theta \leq \text{const.}/\Delta$. However, for $\theta > \text{const.}/\Delta$, there exists families of bounded degree graphs with long-range correlations.

2.3.2 Conditional independence test

A recurring approach to structural learning consists in exploiting the conditional independence structure encoded by the graph [3, 21, 30, 49].

Let us consider, to be definite, the approach of [21], specializing it to the model (2.1.1). Fix a vertex r, whose neighborhood ∂r we want to reconstruct, and consider the conditional distribution of x_r given its neighbors[4]: $\mathbb{P}_{G,\theta}(x_r|\underline{x}_{\partial r})$. Any change of x_i, $i \in \partial r$, produces a change in this distribution which is bounded away from 0. Let U be a candidate neighborhood, and assume $U \subseteq \partial r$. Then changing the value of x_j, $j \in U$ will produce a noticeable change in the marginal of X_r, even if we condition on the remaining values in U and in any W, $|W| \leq \Delta$. On the other hand, if $U \nsubseteq \partial r$, then it is possible to find W (with $|W| \leq \Delta$) and a node $i \in U$ such that, changing its value after fixing all other values in $U \cup W$ will produce no noticeable change in the conditional marginal. (Just choose $i \in U \backslash \partial r$ and $W = \partial r \backslash U$). This procedure allows us to distinguish subsets of ∂r from other sets of vertices, thus motivating the following algorithm.

[4] If $\underline{a}$ is a vector and R is a set of indices then we denote by $\underline{a}_R$ the vector formed by the components of $\underline{a}$ with index in R.

```
LOCAL INDEPENDENCE TEST( samples $\{\underline{x}^{(\ell)}\}$, thresholds $(\epsilon, \gamma)$ )
Select a node $r \in V$;
Set as its neighborhood the largest candidate neighbor $U$ of
size at most $\Delta$ for which the score function SCORE($U$) $> \epsilon/2$;
Repeat for all nodes $r \in V$;
```

The score function SCORE($\cdot$) depends on $(\{\underline{x}^{(\ell)}\}, \Delta, \gamma)$ and is defined as follows,

$$\min_{W,j} \max_{x_i, \underline{x}_W, \underline{x}_U, x_j} |\widehat{\mathbb{P}}_{G,\theta,n}\{X_i = x_i | \underline{X}_W = \underline{x}_W, \underline{X}_U = \underline{x}_U\} - \\ \widehat{\mathbb{P}}_{G,\theta,n}\{X_i = x_i | \underline{X}_W = \underline{x}_W, \underline{X}_{U\setminus j} = \underline{x}_{U\setminus j}, X_j = x_j\}| \,. \tag{2.3.7}$$

In the minimum, $|W| \leq \Delta$ and $j \in U$. In the maximum, the values must be such that

$$\begin{aligned} &\widehat{\mathbb{P}}_{G,\theta,n}\{\underline{X}_W = \underline{x}_W, \underline{X}_U = \underline{x}_U\} > \gamma/2 \\ &\widehat{\mathbb{P}}_{G,\theta,n}\{\underline{X}_W = \underline{x}_W, \underline{X}_{U\setminus j} = \underline{x}_{U\setminus j}, X_j = x_j\} > \gamma/2 \end{aligned} \tag{2.3.8}$$

$\widehat{\mathbb{P}}_{G,\theta,n}$ is the empirical distribution calculated from the samples $\{\underline{x}^{(\ell)}\}_{\ell=1}^{n}$. We denote this algorithm by $\mathsf{Ind}(\epsilon, \gamma)$. The search over candidate neighbors U, the search for minima and maxima in the computation of the SCORE(U) and the computation of $\widehat{\mathbb{P}}_{G,\theta,n}$ all contribute for $\chi_{\mathsf{Ind}}(G, \theta)$.

Both theorems that follow are consequences of the analysis of [21], hence proofs are omitted.

Theorem 2.3.4. *Let G be a graph of bounded degree $\Delta \geq 1$. For every θ there exists (ϵ^0, γ^0), and a numerical constant K, such that*

$$N_{\mathsf{Ind}}(G, \theta) \leq \frac{100\Delta}{(\epsilon^0)^2 (\gamma^0)^4} \log \frac{2p}{\delta} \,, \tag{2.3.9}$$

$$\chi_{\mathsf{Ind}}(G, \theta) \leq K \, (2p)^{2\Delta+1} \log p \,. \tag{2.3.10}$$

More specifically, one can take $\epsilon^0 = \frac{1}{4}\sinh(2\theta)$, $\gamma^0 = e^{-4\Delta\theta}\, 2^{-2\Delta}$.

This first result implies in particular that G can be reconstructed with polynomial computational-complexity for any bounded Δ. However, the degree of such polynomial is pretty high and non-uniform in Δ. This makes the above approach impractical.

A way out was proposed in [21]. The idea is to identify a set of 'potential neighbors' of vertex r via thresholding:

$$B(r) = \{i \in V : \widehat{C}_{ri} > \kappa/2\}\,. \tag{2.3.11}$$

For each node $r \in V$, we evaluate $\textsc{Score}(U)$ by restricting the minimum in Eq. (2.3.7) over $W \subseteq B(r)$, and search only over $U \subseteq B(r)$. We call this algorithm $\mathsf{IndD}(\epsilon, \gamma, \kappa)$. The basic intuition here is that C_{ri} decreases rapidly with the graph distance between vertices r and i. As mentioned above, this is true at low temperature.

Theorem 2.3.5. *Let G be a graph of bounded degree $\Delta \geq 1$. Assume that $\theta < K'/\Delta$ for some small enough constant K'. Then there exists $\epsilon^0, \gamma^0, \kappa^0$ such that*

$$N_{\mathsf{IndD}}(G, \theta) \leq 16 \times 8^{\Delta} \log \frac{4p}{\delta}\,, \tag{2.3.12}$$

$$\chi_{\mathsf{IndD}}(G, \theta) \leq K' p \Delta^{\Delta^{\frac{\log(4/(\Delta\kappa^0))}{\log(1/K')}}} + K' \Delta p^2 \log p\,. \tag{2.3.13}$$

More specifically, we can take $\kappa^0 = \tanh\theta$, $\epsilon^0 = \frac{1}{4}\sinh(2\theta)$ and $\gamma^0 = e^{-4\Delta\theta}\, 2^{-2\Delta}$.

2.3.3 Regularized logistic regression

A common approach to learning the Ising model consists in maximizing an appropriate empirical likelihood function [94, 63, 11, 114, 82, 107]. In order to control statistical fluctuations, and select sparse graphs, a regularization term is often added to the cost function. In this section we focus on a specific implementation of this idea, the ℓ_1-regularized logistic regression method of [94]. This algorithm is interesting because of its low computational complexity and good empirical performance.

For each node r, the following likelihood function is considered

$$\mathcal{L}^n(\underline{\theta}; \{\underline{x}^{(\ell)}\}_{\ell=1}^n) = -\frac{1}{n}\sum_{\ell=1}^{n} \log \mathbb{P}_{\underline{\theta}}(x_r^{(\ell)}|\underline{x}_{\backslash r}^{(\ell)}), \tag{2.3.14}$$

where $\underline{x}_{\backslash r} = \{x_i : i \in V \setminus r\}$ is the vector of all variables except x_r. Henceforth, to simplify notation, we denote the function $\mathcal{L}^n(\underline{\theta}; \{\underline{x}^{(\ell)}\}_{\ell=1}^n)$ by $\mathcal{L}^n(\underline{\theta})$. From the definition of $\mathbb{P}_{\underline{\theta}}$ in (2.1.2) and Bayes rule we have

$$\log \mathbb{P}_{\underline{\theta}}(x_r|\underline{x}_{\backslash r}) = -\log\left(e^{\sum_{j\in V\backslash\{r\}}\theta_{rj}x_j} + e^{-\sum_{j\in V\backslash\{r\}}\theta_{rj}x_j}\right) + \sum_{j\in V\backslash\{r\}} \theta_{rj}x_r x_j\,. \tag{2.3.15}$$

In particular, the function $\mathcal{L}^n(\underline{\theta})$ depends only on the parameters $\underline{\theta}_{r,\cdot} = \{\theta_{rj} : j \in V\backslash\{r\}\}$. This is used to estimate the neighborhood of each node by the following algorithm, denoted by $\mathsf{Rlr}(\lambda)$.

```
REGULARIZED LOGISTIC REGRESSION( samples $\{\underline{x}^{(\ell)}\}_{\ell=1}^n$, regularization $(\lambda)$)
Select a node $r \in V$;
Calculate $\hat{\underline{\theta}}_{r,\cdot} = \arg\min_{\underline{\theta}_{r,\cdot}\in\mathbb{R}^{p-1}} \{\mathcal{L}^n(\underline{\theta}_{r,\cdot}) + \lambda\|\underline{\theta}_{r,\cdot}\|_1\}$;
    If $\hat{\theta}_{rj} \neq 0$, set $(r,j) \in E$;
```

For each node $r \in V$, $\mathsf{Rlr}(\lambda)$ solves a convex optimization problem in p variables whose overall computational-complexity can be bounded by $O(\max\{p,n\}p^3)$ [94]. In this section we focus on the algorithm's sample-complexity, i.e. in the smallest number of samples that are required to reconstruct the graph G. In particular, we are interested in computing bounds for the sample-complexity when the regularization parameter λ is tuned *optimally*, as a function of the graph G, and when the graph is of bounded degree Δ. (see (2.3.1)).

This is a somewhat optimistic assumption, that makes our negative results stronger, and is further discussed below. Our main result establishes an approximate dichotomy for this sample complexity. It might be usefull to recall the definition of $N_{\mathsf{Alg}}(p,\Delta,\theta)$ introduced in (2.3.2).

Theorem 2.3.6. *There exists universal constants C (with $C \leq 10^6$) and Δ_0 (with*

$\Delta_0 \leq 50$), such that

$$\theta\Delta \leq \frac{3}{10} \implies N_{\mathsf{Rlr}}(p,\Delta,\theta) \leq C\,\frac{\Delta}{\theta^2}\log\left(\frac{p}{\delta}\right)\,, \tag{2.3.16}$$

$$2 \leq \theta\Delta \leq 3 \implies N_{\mathsf{Rlr}}(p,\Delta,\theta) = \infty,\;\; \textit{for } \Delta \geq \Delta_0 \textit{ and all } p \textit{ large enough.} \tag{2.3.17}$$

In particular, for $\theta\Delta \leq (3/10)$, the above sample complexity is achieved by $\lambda = \theta/(50\sqrt{\Delta})$.

Further, for all $\theta\Delta \geq 2$, $\Delta \geq 3$ and $\epsilon > 0$, for any $\lambda_1(p) \to \infty$ as $p \to \infty$ and $\lambda_2(n) \to 0$ as $n \to \infty$,

$$\max_{\lambda\in[\lambda_1(p)/\sqrt{n},\lambda_2(n)]} \mathbb{P}\{\widehat{G}(\lambda) = G\} \leq \epsilon \qquad \textit{for all } n \in \mathbb{N}, \tag{2.3.18}$$

for all but a vanishing fraction of regular graphs G with p vertices and degree Δ.

Notice that the requirement $\lambda \in [\lambda_1(p)/\sqrt{n}, \lambda_2(n)]$ in the last part of this statement is very natural. Indeed, $\lambda \geq \lambda_1(p)/\sqrt{n}$ is needed for the regularizer to overcome statistical fluctuations, and $\lambda \leq \lambda_2(n)$ is reuired for the estimator $\hat{\theta}(\lambda)$ to be asymptotically consistent as $n \to \infty$. Typical prescriptions for λ are of the form $\lambda \sim \sqrt{(\log p)/n}$. We also note that the universal constants C, Δ_0 are in practice significantly smaller than what stated formally.

Remark 2.3.1. *The smallest value of the maximum degree Δ_0 for which the negative result (2.3.17) holds can be determined by optimizing a two-variable function (see Appendix A.4.1). Numerical optimization implies that we can take $\Delta_0 = 3$, and that the condition $\theta\Delta \leq 3$ is not required.*

Let us briefly outline the main technical developments in the proof. Ravikumar, Wainwright and Lafferty [94] introduced a set sufficient conditions under which regularized logistic regression reconstructs graphs of maximum degree Δ. Under these conditions the sample complexity was bounded[5] in [94] as $N_{\mathsf{Rlr}}(\lambda, G, \theta) = O(\Delta^3 \log p)$.

[5] Notice that the upper bound in Eq. (2.3.16) is consistent with the one of [94] since, for $\theta = \Theta(1/\Delta)$ it yields $N_{\mathsf{Rlr}}(p,\Delta,\theta) = O(\Delta^3 \log p)$.

A crucial role was played in particular by the so-called *incoherence condition* that depends on the Hessian of the expected likelihood function

$$\mathcal{L}(\underline{\theta}) \equiv \mathbb{E}_{G,\theta^*}\{\log \mathbb{P}_{\underline{\theta}}(X_r|\underline{X}_{\backslash r})\}. \tag{2.3.19}$$

Bellow we use θ^0 or $\underline{\theta}^*$ to denote the true values of the parameters whenever useful for clarity. Further, $\underline{X} \in \mathbb{R}^p$ is a random variable with law $\mathbb{P}_{G,\theta^0}$ and $\mathbb{E}_{G,\theta^0}$ the expectation with respect to this law.

Definition 2.3.2. *Define $Q^0 \equiv \mathsf{Hess}(\mathcal{L})(\underline{\theta}^0)$, where $\mathsf{Hess}(\mathcal{L})$ denotes the Hessian of $\mathcal{L}(\,\cdot\,)$. Let $S = \{j : (r,j) \in E\}$ and $S^C = V \setminus (\{r\} \cup S)$. Define the matrices $Q^0_{SS} = \{Q^0_{ij} : i,j \in S\}$ and $Q^0_{S^CS} = \{Q^0_{ij} : i \in S^C, j \in S\}$. Then (G,θ) satisfies the* incoherence condition with parameter α *if*

$$\|Q^0_{S^CS}(Q^0_{SS})^{-1}\|_\infty \le 1 - \alpha, \tag{2.3.20}$$

where the matrix sup-norm is defined by $\|M\|_\infty = \max_i \sum_j |M_{ij}|$.

The proof of Theorem 2.3.6 involves the following novel technical developments.

i. We prove that that incoherence is necessary, and when incoherence does not hold, under some reasonable assumptions on λ, regularized logistic regression fails with high probability.

ii. We prove that incoherence holds on bounded-degree graphs under the condition $\Delta \le 3/(10\theta)$, cf. Eq. (2.3.16). This requires bounding the entries of $\mathsf{Hess}(\mathcal{L})$. To estabilish such bounds, we use a technique based on a self avoiding walks representation, due to Fisher [40].

iii. We prove that, if G is a uniformly random regular graph of degree Δ, then incoherence fails to hold, with high probability for large p, provided $\Delta \ge 2/\theta$, cf. Eq. (2.3.17). In other words, regularized logistic regression fails on most Δ-regular graphs, if $\Delta \ge 2/\theta$, under some reasonable assuptions on λ.

iv. We construct a family of simple graphs with degree Δ and $p = \Delta + 2$ vertices on which regularized logistic regression fails assymptotically (i.e. as $n \to \infty$) if $\Delta \geq 2/\theta$ and λ is bounded away from 0.

Theorem 2.3.6 follows by merging these developments.

Discussion and further results

Let us discuss a few extensions of Theorem 2.3.6, as well as some outstanding challenges.

Heterogeneous edge strengths. In the statement of Theorem 2.3.6 we assume that all the edge strengths are equal, i.e. $\theta_{ij} = \theta > 0$. In applications, it is more realistic to assume unequal θ_{ij}'s. A natural class would be given by all models of the form (2.1.2) with p nodes, degree bounded by Δ and $0 < \theta_{\min} \leq |\theta_{ij}| \leq \theta_{\max}$. We denote this class of models (weighted graphs) by $\mathcal{G}(p, \Delta, \theta_{\min}, \theta_{\max})$. The algorithm Rlr remains unchanged in this case [94].

Consider first the negative result in Theorem 2.3.6, namely that for $2 \leq \theta\Delta \leq 3$ regularized logistic regression fails irrespective of the number of samples. Of course this conclusion does not change if we consider a more general class. In particular Rlr fails on $\mathcal{G}(p, \Delta, \theta_{\min}, \theta_{\max})$ if $\theta_{\max}\Delta \geq 2$.

Next consider the positive part in Theorem 2.3.6, namely that for $\theta\Delta \leq 3/10$ regularized logistic regression reconstructs G from a number samples that is logarithmic in p. It is not hard to check that the proof applies to the more general model $\mathcal{G}(p, \Delta, \theta_{\min}, \theta_{\max})$, essentially unchanged. The only part that need to be modified is the estimate in Lemma A.3.1, that can be shown to hold with θ replaced by $\theta_{\max}$. Summarizing, we have the following.

Remark 2.3.3. *Denoting by $N_{\mathsf{Rlr}}(p, \Delta, \theta_{\min}, \theta_{\max})$ the sample complexity of regularized logistic regression for the class $\mathcal{G}(p, \Delta, \theta_{\min}, \theta_{\max})$, we have*

$$\theta_{\max}\Delta \leq \frac{3}{10} \qquad \Longrightarrow \qquad N_{\mathsf{Rlr}}(p, \Delta, \theta_{\min}, \theta_{\max}) \leq C\,\frac{\Delta}{\theta_{\min}^2} \log\left(\frac{p}{\delta}\right). \tag{2.3.21}$$

Small graphs. The negative part of Theorem 2.3.6, cf. Eq. (2.3.17) is stated for sufficiently large graph size p. A natural question is whether regularized logistic regression also fails for moderate p. In Section **??** we will construct a class of graphs with $p = \Delta + 2$ for which Rlr fails at $n = \infty$, for any $\lambda > 0$ and $\theta\Delta \geq 2$. These graphs are used as an intermediate step in the proof of Theorem 2.3.6 and indeed suggests that (2.3.17) already holds at small p.

Strucured graphs. Finally, Theorem 2.3.6 states that, for $\theta\Delta > 2$, regularized logistic regression fails on most Δ-regular graphs. However, graphs encountered in applications are often highly structured (although this structure is *a priori* unknown). One might wonder what happens if we consider a structured subclass of the graph family $\mathcal{G}(p, \Delta)$. We consider two such examples:

i. If G is a tree, then $\mathsf{Rlr}(\lambda)$ recovers G with high probability for any θ (for a suitable λ);

ii. if G is a large two dimensional grid, $\mathsf{Rlr}(\lambda)$ fails with high probability for θ large enough and any λ that vanishes with n and satisfies $n\lambda^2 \to \infty$ with p.

Incoherence is necessary. An important technical step in proving the negative part of Theorem 2.3.6, consists in proving that the incoherence condition is necessary for Rlr to successfully reconstruct G. This is stated formally as Lemma 2.6.1. Although a similar result was proven in [116] for model selection using the Lasso, the present paper (and its conference version [68]) is the first to prove that a similar incoherence condition is also necessary when the underlying model is the Ising model.

The intuition behind this condition is quite simple. When $n \to \infty$, and under the restriction that $\lambda \to 0$, solutions given by Rlr converge to the ground truth $\underline{\theta}^0$ as $n \to \infty$ [94]. Hence, for large n, we can expand $\mathcal{L}^n$ in a quadratic function centered around $\underline{\theta}^*$ plus a small stochastic error term. Consequently, when adding the regularization term to $\mathcal{L}^n$, we obtain a cost function analogous to the Lasso plus an error term that needs to be controlled. The study of the dominating contribution leads to the incoherence condition.

Correlation decay and computational phase transitions. The effectiveness of regularized logistic regression changes dramatically as $\theta\Delta$ increases from 3/10 to 2. Similar *computational phase transitions* have been the object of significant attention in the context of approximate counting. This is the problem of computing marginals of (2.1.2), or the partition function $Z_{\underline{\theta}}$, given the parameters $\underline{\theta}$. We refer to [100, 101] and references therein for this line of work.

The proof of Theorem 2.3.6 indicates that the underlying mechanism is the same in these two cases, namely the break down of 'correlation decay' as $\theta\Delta$ increases. More precisely for $\theta\Delta$ small, the correlation between x_i and x_j under the measure (2.1.1) decreases exponentially with the distance between i and j. On the other hand, for $\theta\Delta$ large it does not decrease anymore with the distance. The threshold separating these two behaviors is located at $(\Delta - 1)\tanh\theta = 1$, i.e. for $\Delta\theta = \Theta(1)$.

Choice of the regularization parameter. The choice of λ is crucial for the accuracy of regularized logistic regression. Our definition of sample complexity, cf. Eq. (2.3.1) assumes that the same value of λ is used for all vertices, and that this is chosen as to maximize the probability of correct reconstruction. In practice, the optimal value of λ might be difficult to find and hence this assumption makes our negative result (for $\theta\Delta \geq 2$) stronger. As for our positive result (for $\theta\Delta \leq 3/10$), an explicit value of λ is given that works uniformly over all graphs with the prescribed maximum degree.

An interesting open question is whether these are the only two possible behaviors for the graph ensemble $\mathcal{G}(p, \Delta, \theta_{\min}, \theta_{\max})$: either a universal value of λ exists that performs uniformly well over the ensemble, or even tuning λ 'graph by graph' is unsuccessful.

Finally, there is an even more 'optimistic' definition of sample complexity. In principle, one might assume that a different value of λ is used for each vertex $r \in V$ and each of them is tuned optimally. This is however unrealistic as it would require tuning p regularization parameters, and is unlikely to be ineffective. Indeed, the graphs constructed for the negative part of Theorem 2.3.6 are regular and highly homogeneous. Local graph properties do not appear to be sufficient to determine λ.

2.4 Important remark

If $\theta\Delta \geq C$, for C large enough, all the above described algorithms fail in the regime where they have polynomial computational-complexity. It is natural to ask whether this difficulty is related to the specific algorithms, or is rather an 'intrinsic' computational barrier. While our results do not provide a definite answer to this question, they point at an interesting possibility. Two simple but important observations are the following. First of all, if such a barrier exists, it is computational and not statistical. In fact, the conditional independence test of [21] analysed in Section 2.3.2 reconstructs G with high probability for all θ, Δ, although with complexity $p^{O(\Delta)}$ (see Theorem 2.3.5). Second, it is likely that this barrier does not exist for attractive interactions, i.e. if $\theta_{ij} > 0$ for all $(i, j) \in E$. Indeed counting is polynomial in this case [70], and hence the maximum likelihood estimator can be approximately evaluated.

With these caveats, we notice that, for general Δ, $\theta_{\max} \geq \theta_{\min} > 0$, no algorithm is known that provably reconstructs $G \in \mathcal{G}(p, \Delta, \theta_{\min}, \theta_{\max})$ with computational complexity bounded by p^C, and sample complexity bounded by $(\Delta \log p)^C$ for C a universal constant.

2.5 Numerical results

In order to explore the practical relevance of the above results, we carried out extensive numerical simulations using the regularized logistic regression algorithm $\mathsf{Rlr}(\lambda)$. For a given graph construction with maximum degree Δ, it is convenient to compare the edge strength θ with two distinct thresholds. The first one, defined as $\theta_{\rm uniq}(\Delta) = {\rm atanh}(1/(\Delta - 1))$, is the threshold for correlation decay (and uniqueness of Gibbs measures) on graphs of maximum degree Δ. The second, denoted by $\theta_{\rm crit}$, is the phase transition threshold. The latter is defined only for specific graph sequences, but provides a finer (not worst case) control of correlation decay in many cases.

Samples from the Ising model (2.1.1) were generated using Gibbs sampling (a.k.a. Glauber dynamics). Mixing time can be very large for $\theta \geq \theta_{\rm uniq}$, and was estimated using the time required for the overall bias to change sign (this is a quite conservative

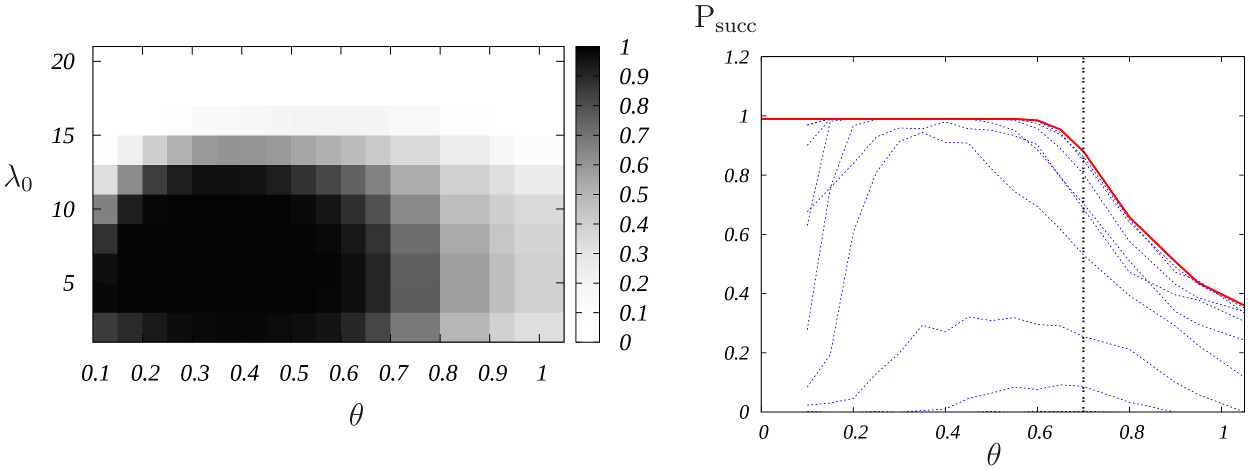

Figure 2.2: Learning random subgraphs of a 7×7 ($p = 49$) two-dimensional grid from $n = 4500$ Ising models samples, using regularized logistic regression. Left: success probability as a function of the model parameter θ and of the regularization parameter λ_0 (darker corresponds to highest probability). Right: the same data plotted for several choices of λ versus θ. The vertical line corresponds to the model critical temperature. The thick line is an envelope of the curves obtained for different λ, and should correspond to optimal regularization.

estimate at low temperature). Generating the samples $\{\underline{x}^{(\ell)}\}$ was indeed the bulk of our computational effort and took about 50 days CPU time on Pentium Dual Core processors. Notice that $\mathsf{Rlr}(\lambda)$ had been tested in [94] only on tree graphs G, or in the weakly coupled regime $\theta < \theta_{\text{uniq}}$. In these cases sampling from the Ising model is easy, but structural learning is also intrinsically easier.

Figure 2.2 reports the success probability of $\mathsf{Rlr}(\lambda)$ when applied to random subgraphs of a 7×7 two-dimensional grid. Each such graphs was obtained by removing each edge independently with probability $\rho = 0.3$. Success probability was estimated by applying $\mathsf{Rlr}(\lambda)$ to each vertex of 8 graphs (thus averaging over 392 runs of $\mathsf{Rlr}(\lambda)$), using $n = 4500$ samples. We scaled the regularization parameter as $\lambda = 2\lambda_0\theta(\log p/n)^{1/2}$ (this choice is motivated by the algorithm analysis [94] and is empirically the most satisfactory), and searched over λ_0.

The data clearly illustrate the phenomenon discussed in the previous pages. Despite the large number of samples $n \gg \log p$, when θ crosses a threshold, the algorithm

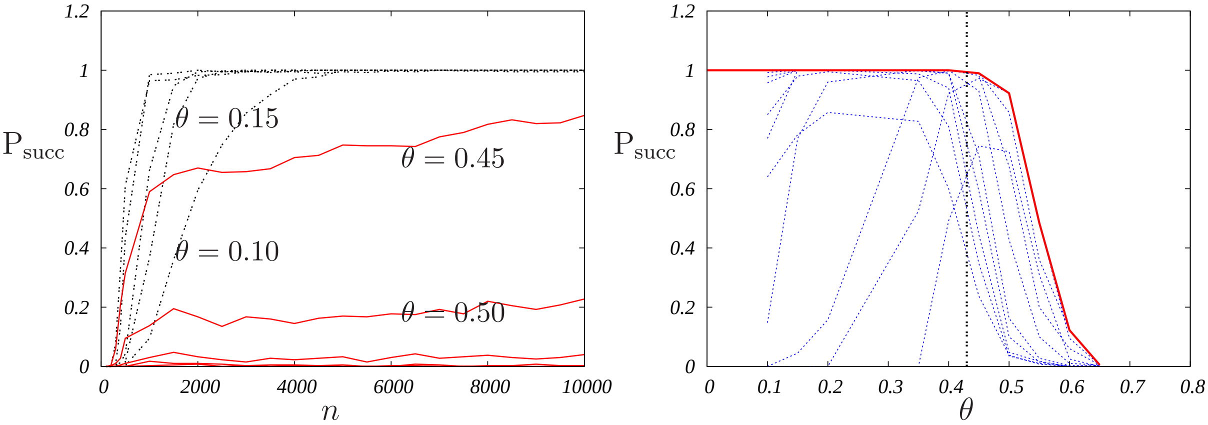

Figure 2.3: Learning uniformly random graphs of degree $\Delta = 4$ from Ising models samples, using regularized logistic regression. Left: success probability as a function of the number of samples n for several values of θ. Dotted: $\theta = 0.10, 0.15, 0.20, 0.35, 0.40$ (in all these cases $\theta < \theta_{\mathrm{thr}}(\Delta = 4)$). Dashed: $\theta = 0.45, 0.50, 0.55, 0.60, 0.65$ ($\theta > \theta_{\mathrm{thr}}(4)$, some of these are indistinguishable from the axis). Right: the same data plotted for several choices of λ versus θ as in Fig. 2.2, right panel.

starts performing poorly irrespective of λ. Intriguingly, this threshold is not far from the critical point of the Ising model on a randomly diluted grid $\theta_{\mathrm{crit}}(\rho = 0.3) \approx 0.7$ [118, 40].

Figure 2.3 presents similar data when G is a uniformly random graph of degree $\Delta = 4$, over $p = 50$ vertices. The evolution of the success probability with n clearly shows a dichotomy. When θ is below a threshold, a small number of samples is sufficient to reconstruct G with high probability. Above the threshold even $n = 10^4$ samples are too few. In this case we can predict the threshold analytically, cf. Lemma ?? below, and get $\theta_{\mathrm{thr}}(\Delta = 4) \approx 0.4203$, which compares favorably with the data.

2.6 Proofs for regularized logistic regression

In this section we present the proof of our results for the success and failure of Rlr. Regarding the chapter on the Ising model, these are the most interesting and novel results. The proofs regarding the simple thresholding algorithm are put in the appendix (cf. A.1) and the proofs regarding the conditional independence can be found in [21]. The proofs for Rlr were first introduced in [68, 16].

The proof regarding the result on Rlr builds on four lemmas that will be proved in the appendices. Conceptually, the proof relies on two types of technical results. First, we estabilish that the incoherence condition introduced in [94] is roughly necessary for regularized logistic regession to succeed, cf. Lemma 2.6.1. Second, we use this fact, together with the main result of [94] to characterize the success probability on three graph families, cf. Lemmas 2.6.2, 2.6.3, 2.6.4. Finally, in order to prove the negative part in Theorem 2.3.6 we simply construct a graph by taking the disjoint union of elements from each of these families.

2.6.1 Notation and preliminary remarks

Before proceeding it is convenient to recall some notation and make some preliminary remarks.

We denote by $[m] = \{1, 2, \dots, m\}$ the set of first m integers. If $v \in \mathbb{R}^m$ is a vector and $R \subseteq [m]$ is an index set then $v_R \equiv (v_i)_{i \in R}$ denotes the vector formed by the entries with index in R. Similarly, if $M \in \mathbb{R}^{m \times n}$ is a matrix and $R \subseteq [m]$, $P \subseteq [n]$ are index sets, then $M_{R,P} \equiv (M_{ij})_{i \in R, j \in P}$ denotes the submatrix indexed by rows in R and columns in P. We denote the maximum and minimum eigenvalue of a symmetric matrix M by $\sigma_{\max}(M)$ and $\sigma_{\min}(M)$ respectively. Recall that $\|M\|_\infty = \max_i \sum_j |M_{ij}|$. We denote by $\mathbb{1}$ the all-ones vector

As before, we let r be the vertex whose neighborhood we are trying to reconstruct and define $S = \partial r$ and $S^C = V \setminus (\partial r \cup \{r\})$. Since the function $\mathcal{L}^n(\underline{\theta}; \{\underline{x}^{(\ell)}\}_{\ell=1}^n) + \lambda \|\underline{\theta}\|_1$ only depends on $\underline{\theta}$ through the components $\underline{\theta}_{r,\cdot} = \{\theta_{rj} : j \in V \setminus \{r\}\}$ (see equation (2.3.15)), we hereafter neglect all the other parameters and write $\underline{\theta}$ as a shorthand of $\underline{\theta}_{r,\cdot}$. As mentioned above, whenever necessary to avoid confusion, we will write θ^0 or $\underline{\theta}^0$ (if viewed as a vector) for the true parameters values. Namely $\underline{\theta}^0 = \{\theta_{rj} : \theta_{ij} = 0,\ \forall j \notin \partial r, \theta_{rj} = \theta^0,\ \forall j \in \partial r\}$.

We denote by z^0 a sub-gradient of $\|\underline{\theta}\|_1$ evaluated at the true parameters values θ^0. Note that, since we are working in the ferromagnetic domain, i.e. $(\underline{\theta}^0_S)_i > 0$ for all $i \in S$, we have $z^0_S = \mathbb{1}$. We define $W^n(\underline{\theta})$ to be minus the gradient of $\mathcal{L}^n$,

$W(\underline{\theta})$ to be the minus the gradient of $\mathcal{L}$, $Q^n(\underline{\theta})$ to be the Hessian of $\mathcal{L}^n(\underline{\theta})$ and $Q(\underline{\theta})$ to be the Hessian of $\mathcal{L}(\underline{\theta})$. Notice that, by the law of large numbers, for every $\underline{\theta}$, $\lim_{n\to\infty} \mathcal{L}^n(\underline{\theta}) = \mathcal{L}(\underline{\theta})$, $\lim_{n\to\infty} W^n(\underline{\theta}) = W(\underline{\theta})$, and $\lim_{n\to\infty} Q^n(\underline{\theta}) = Q(\underline{\theta})$.

The gradient and Hessian of $\mathcal{L}$ admit indeed fairly explicit expressions. For all $i, j \in V\backslash\{r\}$ we have,

$$Q^n_{ij}(\underline{\theta}) = \frac{1}{n}\sum_{\ell=1}^{n} \frac{x_i^{(\ell)}x_j^{(\ell)}}{\cosh^2\left(\sum_{t\in V\backslash\{r\}}\theta_{rt}x_t^{(\ell)}\right)}, \tag{2.6.1}$$

$$Q_{ij}(\underline{\theta}) = \mathbb{E}_{G,\theta}\left(\frac{X_iX_j}{\cosh^2(\sum_{t\in V\backslash\{r\}}\theta_{rt}X_t)}\right), \tag{2.6.2}$$

$$-W^n_i(\underline{\theta}) = [\nabla\mathcal{L}^n(\underline{\theta})]_i = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)}\Big(\tanh\Big(\sum_{t\in V\backslash\{r\}}\theta_{rt}x_t^{(\ell)}\Big) - x_r^{(\ell)}\Big), \tag{2.6.3}$$

$$-W_i(\underline{\theta}) = [\nabla\mathcal{L}(\underline{\theta})]_i = \mathbb{E}_{G,\theta}\Big\{X_i\tanh\Big(\sum_{t\in V\backslash\{r\}}\theta_{rt}X_t\Big)\Big\} - \mathbb{E}_{G,\theta}\{X_iX_r\}. \tag{2.6.4}$$

Note that, from the last expression, it follows that $\nabla\mathcal{L}(\underline{\theta}^0) = 0$. This in turn is related to asymptotic consistency: if $\lambda \to 0$ as $n \to \infty$ then $\hat{\underline{\theta}} \to \underline{\theta}^0$.

We let $\hat{\underline{\theta}}$ denote the parameter estimate computed by $\mathsf{Rlr}(\lambda)$ when applied to samples $\{\underline{x}^{(\ell)}\}_{\ell=1}^n$.

We will omit arguments whenever clear from the context. Any quantity evaluated at the true parameter values will be represented with a ^{0}, e.g. $Q^0 = Q(\underline{\theta}^0)$. We use instead a hat (as in $\hat{\underline{\theta}}$) to denote estimates based on $\{\underline{x}^{(\ell)}\}_{\ell=1}^n$. When clear from the context we might write $\mathbb{E}_{G,\theta}$ as simply $\mathbb{E}$. Similarly, $\mathbb{P}_{G,\theta}$ will be sometimes written as simply $\mathbb{P}$.

Throughout this paper, $\mathrm{P}_{\mathrm{succ}}$ will denote the probability of success of Rlr, that is, the probability that the algorithm is able to recover the underlying G exactly. Also, G will be a graph of maximum degree Δ.

2.6.2 Necessary conditions for the success of Rlr

Our first technical result establishes that, if λ is small and $\|Q^0_{S^CS}(Q^0_{SS})^{-1}z^0_S\|_\infty > 1$, then $\mathsf{Rlr}(\lambda)$ fails to reconstruct the neighborhood S correctly. (Recall that z^0_S is the

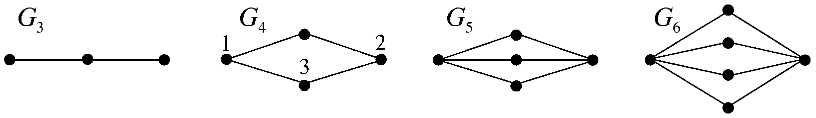

Figure 2.4: Diamond graphs $\mathcal{G}_{\rm diam}(p)$.

subgradient of $\|\underline{\theta}\|_1$ evaluated at $\underline{\theta}^0$.) Notice that, under the incoherence condition $\|Q^0_{S^CS}(Q^0_{SS})^{-1}z^0_S\|_\infty \le (1-\alpha)\|z^0_S\|_\infty \le (1-\alpha)$. Hence this lemma suggests that incoherence is roughly necessary for regularized logistic regression to succeed. Its proof can be found in Appendix A.2. (It is not quite necessary because it could in principle be the case that $\|Q^0_{S^CS}(Q^0_{SS})^{-1}\|_\infty > 1$ but $\|Q^0_{S^CS}(Q^0_{SS})^{-1}z^0_S\|_\infty < 1$.)

Lemma 2.6.1. *Assume* $[Q^0_{S^CS}(Q^0_{SS})^{-1}z^0_S]_i \ge 1+\epsilon$ *for some* $\epsilon > 0$ *and some row* $i \in V$, $\sigma_{\min}(Q^0_{SS}) \ge C_{\min} > 0$, *and* $\lambda < C^3_{\min}\epsilon/(2^7(1+\epsilon^2)\Delta^3)$. *Then the success probability of* $\mathsf{Rlr}(\lambda)$ *is upper bounded as*

$$\mathbb{P}\{\widehat{G}(\lambda) = \widehat{G}\} \le 4\Delta^2 e^{-n\delta_A^2} + 4\Delta\, e^{-n\lambda^2\delta_B^2} \tag{2.6.5}$$

where $\delta_A = (C^2_{\min}/32\Delta)\epsilon$ *and* $\delta_B = (C_{\min}/64\sqrt{\Delta})\epsilon$.

2.6.3 Specific graph ensembles

In this section we consider the performances of $\mathsf{Rlr}(\lambda)$ on three graph ensembles. In the proof of Theorem 2.3.6, we will take the disjoint union of one graph from each ensemble. This trick allows us to rule out a subset of values of λ for each graph ensemble. Notice that, in each case, we single out a specific vertex r whose neighborhood is to be reconstructed. Equivalently, we define here ensembles of *rooted* graphs. The three graph ensembles are defined as follows:

One-edge graphs $\mathcal{G}_{\rm one}(p)$. This is the set of graphs with vertex set $V = [p]$ and only one edge involving the distinguished vertex r, $E = \{(r,i)\}$ (e.g. $r = 1$, $i = 2$).

Diamond graph $\mathcal{G}_{\rm diam}(p)$. This is the set of graphs with p vertices and $2(p-2)$ edges (see Figure 2.4). Two of the vertices to be denoted as vertex 1 and vertex

2 have degree $(p-2)$. They are connected to each of vertices $\{3,\dots,p\}$ which have in turn degree 2. The maximum degree is $\Delta = p-2$. We will identify the root vertex with $r=1$.

These graphs capture a situation in which the two variables interact 'indirectly' through numerous weak intermediaries.

Random regular graph $\mathcal{G}_{\rm rand}(p,\Delta)$. Finally we denote by $\mathcal{G}_{\rm rand}(p,\Delta)$ the set of regular graphs with p nodes and degree Δ. This is naturally endowed with the uniform probability distribution. With a slight abuse of notation, we shall use $\mathcal{G}_{\rm rand}(p,\Delta)$ to denote the resulting probability law over graphs, and write $G \sim \mathcal{G}_{\rm rand}(p,\Delta)$ for a uniformly random regular graph. The root r is also chosen uniformly at random in V.

The proof of the first lemma can be found in [16] and the proofs of the two other are included in Appendices A.4 and A.5.

The first lemma of this section considers one-edge graphs in $\mathcal{G}_{\rm one}(p)$. It implies that, unless $n\lambda^2$ is increasing with p, $\mathsf{Rlr}(\lambda)$ fails with significant probability.

Lemma 2.6.2 (One-edge graphs $\mathcal{G}_{\rm one}(p)$)**.** *There exist $M = M(K,\theta) > 0$ decreasing with K for $\theta > 0$ such that the following is true. If $G \in \mathcal{G}_{\rm one}(p)$ and $n\lambda^2 \le K$, then*

$$\mathbb{P}\{\widehat{G}(\lambda) = \widehat{G}\} \le e^{-M(K,\theta)p} + e^{-n(1-\tanh\theta)^2/32}\,. \tag{2.6.6}$$

The same upper bounds holds for the probability of reconstructing the neighborhood of the root r.

The second lemma deals with diamond graphs $\mathcal{G}_{\rm diam}(p)$. It shows that $\mathsf{Rlr}(\lambda)$ fails if λ is bounded away from 0 provided $\theta\Delta \ge 2$.

Lemma 2.6.3 (Diamond graphs $\mathcal{G}_{\rm diam}(p)$)**.** *There exists $\Delta_0 \ge 3$ (with $\Delta_0 \le 50$) such that the following happens for all $\Delta \ge \Delta_0$ and $2 \le \theta\Delta \le 3$. For any $G \in \mathcal{G}_{\rm diam}(p = \Delta+2)$ and any fixed $\lambda_{\rm min} > 0$,*

$$\sup_{\lambda\ge\lambda_{\rm min}} \mathbb{P}\{\widehat{G}(\lambda) = G\} \le \epsilon \quad \textit{for all } n > n_0(\Delta,\epsilon,\lambda_{\rm min})\,. \tag{2.6.7}$$

The same upper bounds holds for the probability of reconstructing the neighborhood of the root r.

The proof of this lemma is based on the analysis of the expected likelihood function $\mathcal{L}(\underline{\theta})$, cf. Eq. (2.3.19). More precisely we consider the regularized cost function $\mathcal{L}(\underline{\theta}) + \lambda\|\underline{\theta}\|_1$ and show that it is minimized at a parameter vector $\underline{\theta}$ with wrong support. Using the graph symmetries the proof is reduced to minimizing a function in two dimensions.

Remark 2.6.1. *Numerical solution of the mentioned two-dimensional minimization problem indicates that Lemma 2.6.3 holds for $\theta\Delta \geq 2$, $\Delta \geq 3$. (see Appendix A.4.1)*

Finally, we consider random regular graphs.

Lemma 2.6.4 (Random graphs $\mathcal{G}_{\rm rand}(p, \Delta)$**.** *Let $\Delta \geq 3$ and $G \sim \mathcal{G}_{\rm rand}(p, \Delta)$. Then there exists $\theta_{\rm thr}(\Delta)$ and, for any $\theta > \theta_{\rm thr}(\Delta)$, there exists $\delta = \delta(\theta, \Delta) > 0$, $\lambda_{\rm thr} = \lambda_{\rm thr}(\theta, \Delta)$ such that the success probability of* $\mathsf{Rlr}(\lambda)$ *is upper bounded as*

$$\max_{\lambda \leq \lambda_{\rm thr}} \mathbb{P}\{\widehat{G}(\lambda) = G\} \leq 4\Delta^2 p^{-\delta} + 2\Delta\, e^{-\delta n \lambda^2}\,, \tag{2.6.8}$$

with high probability with respect to choice of G. The same upper bounds holds for the probability of reconstructing the neighborhood of the root r.

For large Δ, $\theta_{\rm thr}(\Delta) = h_\infty^2\,\Delta^{-1}(1 + o_\Delta(1))$. The constant h_∞^2 (≈ 1.439) is the unique positive solution of

$$h_\infty \tanh h_\infty = 1\,. \tag{2.6.9}$$

In addition, for any $\Delta \geq 3$, $\theta_{\rm thr}(\Delta) \leq 2\Delta^{-1}$.

The proof of this result relies on a local weak convergence result for ferromagnetic Ising models on random graphs proved in [35]. This allows us to prove that, under the lemma's assumptions $\|Q^0_{S^CS}(Q^0_{SS})^{-1}z^0_S\|_\infty \geq 1 + \epsilon(\theta, \Delta)$ with high probability.

2.6.4 Proof of Theorem 2.3.6

We distinguish two parts: (i) Proving that Rlr succeds with high probability if $\theta\Delta < 3/10$; (ii) Construct a family of graphs that Rlr fails to reconstruct for $\theta\Delta \geq 2$ (plus the mentioned conditions).

The second part is the most challenging technically, and builds on Lemmas 2.6.2, 2.6.3, 2.6.4.

2.6.5 Proof of Theorem 2.3.6: $\theta\Delta \leq 3/10$

The proof consists in checking the sufficient conditions for the success of Rlr described in [94, Theorem 1] hold true. Namely we prove the following Lemma.

Lemma 2.6.5. *For $\theta\Delta \leq 3/10$, there exists constants $C_{\min} > 0$, $D_{\max} < \infty$ and $\alpha < 1$ such that*

$$\sigma_{\min}(Q^0_{SS}) > C_{\min}, \tag{2.6.10}$$

$$\sigma_{\max}(\mathbb{E}_{G,\theta}(X_S X_S^*)) < D_{\max}, \tag{2.6.11}$$

$$\|Q^0_{S^C S}(Q^0_{SS})^{-1}\|_\infty \leq 1 - \alpha. \tag{2.6.12}$$

This is proved in Appendix A.3 by estabilishing estimates on the entries Q^0 using a technique of Fisher [40]. Once this lemma is established, the upper-bound on the sample-complexity follows by carefully carrying all the proofs in [94, Theorem 1] without dropping any constants.

2.6.6 Proof of Theorem 2.3.6: $\theta\Delta \geq 2$

First consider the last part of the Theorem, cf. Eq. (2.3.18). By Lemma 2.6.4, we have

$$\max_{\lambda\in[\lambda_1(p)/\sqrt{n},\lambda_2(n)]} \mathbb{P}\{\widehat{G}(\lambda) = G\} \leq 4\Delta^2\, p^{-\delta} + 2\Delta\, e^{-\delta\lambda_1(p)^2)}\,,$$

for all but a vanishing fraction of graphs $G \in \mathcal{G}_{\rm rand}(p,\Delta)$. Since $\delta > 0$ and $\lambda_1(p) \to \infty$ as $p \to \infty$, the right hand side will be smaller than ϵ for all p large enough. Therefore,

$$\max_{\lambda\in[\lambda_1(p)/\sqrt{n},\lambda_2(n)]} \mathbb{P}\{\widehat{G}(\lambda) = G\} \leq \epsilon\,, \qquad \text{for all } n \in \mathbb{N},$$

for all but a vanishing fraction of graphs $G \in \mathcal{G}_{\rm rand}(p,\Delta)$.

Next consider the claim (2.3.17). Fix $\Delta \geq \Delta_0$, whereby Δ_0 is defined in such a way that Lemma 2.6.3 holds, and θ so that $2 \leq \theta\Delta \leq 3$. We construct a graph G with $p_{\rm tot} = 2p + (\Delta + 2)$ vertices as a disjoint union of $G = G_1 \cup G_2 \cup G_3$. Here $G_1 = (V_1, E_1) \in \mathcal{G}_{\rm one}(p)$ is a one-edge graph with p vertices and root r_1; $G_2 = (V_2, E_2) \in \mathcal{G}_{\rm diam}(\Delta + 2)$ is a diamond graph with $\Delta + 2$ vertices and root r_2; $G_3 = (V_3, E_3) \sim \mathcal{G}_{\rm rand}(p, \Delta)$ is a uniformly random regular graph with degree Δ, rooted at r_3. It is sufficient to prove that, for all n large enough

$$\sup_{\lambda \in \mathbb{R}_+} \mathbb{P}_{G,\theta,n}(\widehat{G}(\lambda) = G) \leq \frac{1}{2}.$$

For $\ell \in \{1, 2, 3\}$, let $\mathcal{E}_\ell(\lambda)$ be the event that $\mathsf{Rlr}(\lambda)$ reconstructs the neighborhood of r_ℓ correctly. Further let $\mathcal{E}_0(\lambda)$ be the event that $\mathsf{Rlr}(\lambda)$ does not return any edge across vertex sets V_1, V_2, V_3 (such edges are absent from the ground truth). We then have

$$\sup_{\lambda \in \mathbb{R}_+} \mathbb{P}(\widehat{G}(\lambda) = G) \leq \sup_{\lambda \in \mathbb{R}_+} \min \Big\{ \mathbb{P}(\mathcal{E}_1(\lambda) \cap \mathcal{E}_0(\lambda));\ \mathbb{P}(\mathcal{E}_2(\lambda) \cap \mathcal{E}_0(\lambda));\ \mathbb{P}(\mathcal{E}_3(\lambda) \cap \mathcal{E}_0(\lambda)) \Big\}$$

It is therefore sufficient to show that there exists a function $p \mapsto \lambda_1(p)$ and a constant $\lambda_2 > 0$ such that, with positive probability with respect to the choice of the random graph G, we have

$$\begin{aligned} \sup_{\lambda \in [0, \lambda_1(p)/\sqrt{n}]} \mathbb{P}(\mathcal{E}_1(\lambda) \cap \mathcal{E}_0(\lambda)) &\leq \frac{1}{2}, \\ \sup_{\lambda \in [\lambda_1(p)/\sqrt{n}, \lambda_2]} \mathbb{P}(\mathcal{E}_3(\lambda) \cap \mathcal{E}_0(\lambda)) &\leq \frac{1}{2}, \\ \sup_{\lambda \in [\lambda_2, \infty)} \mathbb{P}(\mathcal{E}_2(\lambda) \cap \mathcal{E}_0(\lambda)) &\leq \frac{1}{2}. \end{aligned}$$

We fix the function $p \mapsto \lambda_1(p)$ such that $\lambda_1(p) \to \infty$ as $p \to \infty$ and $M(\lambda_1(p)^2, \theta)p \geq \log p$, where $M(\cdots)$ is defined as per Lemma 2.6.2. This is possible since $M(K, \theta)$ is decreasing and strictly positive at any $K < \infty$.

Notice that, under $\mathcal{E}_0(\lambda)$, $\mathsf{Rlr}(\lambda)$ outputs $\theta_{r_1,j} = 0$ for any $j \in V_2 \cup V_3$. We can therefore bound $\mathbb{P}(\mathcal{E}_1(\lambda) \cap \mathcal{E}_0(\lambda))$ as if G_2 and G_3 were absent. Using Lemma 2.6.2

we get

$$\begin{aligned}\sup_{\lambda\in[0,\lambda_1(p)/\sqrt{n}]}\mathbb{P}(\mathcal{E}_1(\lambda)\cap\mathcal{E}_0(\lambda)) &\le e^{-M(\lambda_1(p)^2,\theta)p}+e^{-n(1-\tanh\theta)^2/32}\\ &\le \frac{1}{p}+e^{-n(1-\tanh\theta)^2/32}\le\frac{1}{2}\,,\end{aligned}$$

where the last inequality follows for $p\ge 4$ and $n\ge 64/(1-\tanh\theta)^2$.

Analogously, we can use Lemma 2.6.4 to upper bound the probability of $\mathcal{E}_3(\lambda)$. We get, for $\delta=\delta(\theta,\Delta)>0$,

$$\begin{aligned}\sup_{\lambda\in[\lambda_1(p)/\sqrt{n},\lambda_2]}\mathbb{P}(\mathcal{E}_3(\lambda)\cap\mathcal{E}_0(\lambda)) &\le 4\Delta^2\,p^{-\delta}+2\Delta\,e^{-\delta n\lambda^2}\\ &\le 4\Delta^2\,p^{-\delta}+2\Delta\,e^{-\delta\lambda_1(p)^2}\le\frac{1}{2}\,,\end{aligned}$$

where the last inequality follows for $p\ge p_0(\Delta,\theta)$.

Finally, using Lemma 2.6.3 we obtain the desired bound on $\mathbb{P}(\mathcal{E}_2(\lambda)\cap\mathcal{E}_0(\lambda))$, for $n>n_0(\Delta,1/2,\lambda_2)$.

2.6.7 Discussion

The construction used in the proof of Theorem 2.3.6 for $\theta\Delta>2$ might appear somewhat contrived. In reality it exposes quite clearly the basic mechanisms at work.

First of all, it is necessary to scale $\lambda\ge\sqrt{\lambda_1(p)/n}$, with $\lambda_1(p)\to\infty$. Indeed, if this is not the case, $\mathsf{Rlr}(\lambda)$ will fail reconstructing the neighborhood of any vertex that is connected only to a few neighbors, and essentially independent from most other vertices. This phenomenon is independent of the tradeoff between Δ and θ, and of the correlation decay properties. The graph $G_1\in\mathcal{G}_{\rm one}(p)$ is only a particularly simple example that demonstrates it.

On the other hand we expect that, to demonstrat failure of $\mathsf{Rlr}(\lambda)$ for $\lambda\ge\sqrt{\lambda_1(p)/n}$, either of the two graphs $G_2\in\mathcal{G}_{\rm diam}(\Delta+2)$ or $G_3\sim\mathcal{G}_{\rm rand}(p,\Delta)$ should be sufficient. The main reason to use the disjoint union of G_2 and G_3 is that each of the two cases is easier to treat in a distinct regime. Further, each of the two types of graphs is interesting in its own. While $G_2 \in \mathcal{G}_{\text{diam}}(\Delta+2)$ is extremely simple, $G_3 \sim \mathcal{G}_{\text{rand}}(p,\Delta)$ is 'typical' (as it reproduces the behavior of most graphs with degree Δ).

The insight associated to failure on $\mathcal{G}_{\text{diam}}(\Delta+2)$ or $\mathcal{G}_{\text{rand}}(p,\Delta)$ is similar. Strong correlations between vertices that are not directly connected fool the algorithm.

2.7 Regularized logistic regression and graph families with additional structure

Our main result concerning Rlr, Theorem 2.3.6, implies that regularized logistic regression generally fails when $\Delta \max_{(ij)\in E} |\theta_{ij}| \geq 2$. It is natural to wonder whether this conclusion changes under additional assumptions on the graph G.

The next result is a striking example of this type.

Proposition 2.7.1. *If G is a tree with maximum degree Δ then, for any $\theta > 0$,*

$$N_{\mathsf{Rlr}}(G,\theta) \leq C\frac{\Delta}{\theta^2}\log\left(\frac{p}{\delta}\right). \tag{2.7.1}$$

The proof of this proposition is obtained by showing that, for any θ, Δ, the Ising measure on a tree G with maximum degree Δ satisfies the incoherence condition. The proof of this proposition can be found in the appendix of [67].

The last statement might appear to contradict Lemma 2.6.4. The proof of the latter is based on the local weak convergence of the distribution (2.1.1) for random regular graph to an Ising measure on a regular tree. For the latter measure, we prove that the incoherence condition is violated for $\Delta\theta$ large enough.

Remark 2.7.1. *There is no contradiction between Proposition 2.7.1 and Lemma 2.6.4. The Ising measure on a random regular graph converges to the measure on a tree with* plus boundary conditions. *The relevant measure for a finite tree, used in Proposition 2.7.1 has* free boundary conditions*.*

Our last example concerns loopy graphs, namely two dimensional grids. It implies that $\mathsf{Rlr}(\lambda)$ fails above a certain threshold in θ: in this case, additional structure does not change the qualitative conclusion of Theorem 2.3.6.

Proposition 2.7.2. *There exists $\theta_{\rm 2d}$ such that the following happens for all $\theta > \theta_{\rm 2d}$. Let G be a two-dimensional grid of size $\sqrt{p} \times \sqrt{p}$ with periodic boudary condition, and, $p \mapsto \lambda_1(p)$ and $n \mapsto \lambda_2(n)$ be such that $\lambda_1(p) \to \infty$ as $p \to \infty$ and $\lambda_2(n) \to 0$ as $n \to \infty$. Then for any $\epsilon > 0$ there esist p_0, n_0 such that, if $p \geq p_0$, $n \geq n_0$, then*

$$\sup_{\lambda \in [\lambda_1(p)/\sqrt{n}, \lambda_2(d)]} \mathbb{P}_{G,\theta,n}(\widehat{G}(\lambda) = G) \leq \epsilon\, .$$

The proof uses once more Lemma 2.6.1 and shows that the incoherence condition is violated at large enough θ. The proof of this fact is very technical and can be found in the appendix of [67].

Assuming $\lambda \in [\lambda_1(p)/\sqrt{n}, \lambda_2(d)]$ is natural for the same reasons as explained after the statement of Theorem 2.3.6.

Chapter 3

Learning stochastic differential equations

In this chapter we study the sample-complexity and computational-complexity in learning graphical models from data with correlations. The focus is on stochastic differential equations (SDEs) and learning from continuous trajectories. A specific type of SDE already appeared in the Introduction (Chapter 1 equation (1.2.5)): in linear SDEs the rate of change of each variable is a linear combination of neighboring variables in a graph G. However, our interest is broader, and in Section 3.1, a very general family of non-linear SDEs parametrized by graphs is introduced and is then analyzed throughout the chapter.

An ℓ_1-regularized least-squares algorithm to learn these non-linear SDEs is discussed in Section 3.3.2. We denote it by Rls. Our main results (Section 3.3) characterize the sample-complexity of Rls in learning linear SDEs with sparse representation from continuous trajectories. A general result is presented in Section 3.3.3, and, to facilitate its interpretation, in Section 3.3.4 we look at the simpler case of linear SDEs parametrized by the Laplacian of a graph. In Section 3.5, we study two essential results that are the basis of our main results. Namely, in Section 3.5.1, we focus on learning stochastic linear difference equations [1] – of which learning SDEs from continuous trajectories is a limiting case – and in Section 3.5.2, we describe a general lower bound on the sample-complexity of learning SDEs – from which we derive a lower bound matching the upper bound on the sample-complexity of Rls. In Section 3.7.1, we look at a lower bound on the sample-complexity of learning dense linear SDEs and in Section 3.7.2 at a lower bound when learning non-linear SDEs.

Our results show that the presence of correlations in the samples is balanced by the infinite number of data points from the continuous trajectories and results in a change of the sample-complexity (regarding the dependency in p) only by a multiplicative factor in comparison with when samples are independent. Why is this so and what is this multiplicative factor?

Let us look at the autocorrelation function of the sampled stationary trajectory of the following 1-D linear SDE

$$\mathrm{d}x(t) = -\rho x(t)\mathrm{d}t + \mathrm{d}b(t). \tag{3.0.1}$$

For $\rho > 0$ the SDE admits a unique stationary solution [88] and we obtain

$$\mathbb{E}\{x(0)x(\ell\eta)\} = \frac{1}{2\rho}e^{-\rho\ell\eta}. \tag{3.0.2}$$

If $\eta \gg 1/\rho$, consecutive samples are approximately independent. From a different perspective, in n samples spaced in time by η, the number of approximately independent samples is $O(n\min\{1,\eta\rho\})$. Hence, for example, if learning sparse graphs of bounded degree Δ from n independent samples can be, in principle, done for $n < O(\Delta\log p)$ [21], then learning sparse graphs for SDEs should be possible for $n < O(\Delta\max\{1,(\eta\rho)^{-1}\}\log p)$. Since $n\eta$ is the length T of the observation window of the trajectory, we can write $T < O(\Delta\max\{\eta,\rho^{-1}\}\log p)$. As $\eta \to 0$, the bound converges to $T < O(\Delta\rho^{-1}\log p)$ and it is tempting to regard this metric as a bound on the sample complexity when learning from continuous time data. In particular,

[1] Throughout this thesis we often refer to stochastic difference equations as linear SDEs in discrete time. A simple example of a linear SDE in discrete time is $x(n+1) = 0.5x(n) + w(n)$, $n = 0, 1, ...$ where $x(0) = 0$ and $\{w(n)\}_{n=0}^{\infty}$ are i.i.d. $\mathsf{N}(0,1)$ random variables.

this bound predicts that the smaller ρ is, the more information we need to learn G.

In this section we show the above relation between η, n and ρ is partially correct. More precisely, we prove an upper bound that is similar to the bounds for independent samples and behaves like ρ^{-1} for ρ small. In our result, ρ is the decay rate of the slowest mode of the SDE. In other words, Rls performs worse as the SDE approaches instability. However, the same upper-bound also predicts that, as ρ increases, the sample-complexity degrades. This is at odds with the bound we obtained here but is the correct behavior. In fact, this is the best possible behavior for ρ large since, for this regime, we can prove a matching lower bound for any algorithm with success probability greater than 1/2. To understand this, look at the SDE as a dynamical system driven by white-noise. For ρ large, the system quickly filters any inputs and the driving white-noise cannot excite the system enough for us to learn it from data.

The question of characterizing the sample-complexity of learning from a sampled continuous trajectory remains open (but see Section 3.2 for an overview on related work).

The work in this chapter is based on joint work with Ibrahimi and Montanari [14, 15, 13].

3.1 Introduction

Consider a continuous-time stochastic process $\{\underline{x}(t)\}_{t\geq 0}$, $\underline{x}(t) = [x_1(t), \dots, x_p(t)] \in \mathbb{R}^p$, which is defined by a stochastic differential equation (SDE) of diffusion type

$$\mathrm{d}\underline{x}(t) = F(\underline{x}(t); \underline{\theta}^0)\,\mathrm{d}t + \mathrm{d}\underline{b}(t)\,, \tag{3.1.1}$$

where $\underline{b}(t)$ is a p-dimensional standard Brownian motion and the *drift coefficient* [2], $F(\underline{x}(t); \underline{\theta}^0) = [F_1(\underline{x}(t); \underline{\theta}^0), \dots, F_p(\underline{x}(t); \underline{\theta}^0)] \in \mathbb{R}^p$, is a function of $\underline{x}(t)$ parametrized by $\underline{\theta}^0$. This is an unknown vector, with dimensions scaling polynomially with p.

In this chapter we consider the problem of learning the support of the vector $\underline{\theta}^0$

[2] Throughout this chapter, vectors are 'column vector' even if they are represented in row form for typographical reasons.

from a sample trajectory $X_0^T \equiv \{\underline{x}(t) : \ t \in [0,T]\}$. More precisely, we focus on the high-dimensional limit, where p can grow with T, and determine necessary and sufficient conditions for recovering the signed support of $\underline{\theta}^0$ with high probability [3].

As stated in the introductory Chapter 1, we refer to the smallest T that allows us to achieve a prescribed success probability as the 'sample-complexity' of the problem (although the number of samples is, strictly speaking, infinite). We are particularly interested in achieving the optimal scaling of sample-complexity with the problem dimensions through computationally efficient procedures.

Concretely, given a SDE parametrized by $\underline{\theta}^0$ and an algorithm $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ that outputs an estimate $\hat{\underline{\theta}}$, we define the sample-complexity $T_{\mathsf{Alg}}(\underline{\theta}^0)$ as

$$T_{\mathsf{Alg}}(\underline{\theta}^0) = \inf\{T_0 \in \mathbb{R}^+ : \mathbb{P}_{\underline{\theta}^0,T}\{\text{sign}(\hat{\underline{\theta}}) = \text{sign}(\underline{\theta}^0)\} \geq 1-\delta \text{ for all } T \geq T_0\}. \quad (3.1.2)$$

In the expression above, $\mathbb{P}_{\underline{\theta}^0,T}$ denotes probability with respect to the trajectory X_0^T. Obviously, $T_{\mathsf{Alg}}(\underline{\theta}^0)$ defined above is an upper bound for sample-complexity of learning the support alone.

In addition to this definition, given a class of $\mathcal{A}$ of parameters we define,

$$T_{\mathsf{Alg}}(\mathcal{A}) = \max_{\Theta^0 \in \mathcal{A}} T_{\mathsf{Alg}}(\underline{\theta}^0). \quad (3.1.3)$$

Models based on SDEs play a crucial role in several domains of science and technology, ranging from chemistry to finance. Correspondingly, parameter estimation has been intensely studied in this context. We refer to Section 3.2 for a brief overview. A complete understanding of parameter estimation in a high-dimensional setting is nevertheless missing.

Our results address these challenges for special classes of SDEs of immediate relevance. A first class is constituted by drift coefficients that are parametrized linearly.

[3] Recall that the signed support of $\underline{\theta}^0$ is represented by sign($\underline{\theta}^0$) and corresponds to the partition of the set of indices of $\underline{\theta}^0$ into three sets: indices with positive value, indices with negative value and indices with zero value.

Explicitly, we are given a set of basis functions

$$F(\underline{x}) = [f_1(\underline{x}), f_2(\underline{x}), \dots, f_m(\underline{x})], \tag{3.1.4}$$

with $f_i : \mathbb{R}^p \to \mathbb{R}$. The drift is then given as $F(\underline{x}; \Theta^0) = \Theta^0 F(x)$, with $\Theta^0 \equiv (\theta^0_{ij})_{i\in[p],j\in[m]} \in \mathbb{R}^{p\times m}$. In this chapter we often use the notation Θ^0 instead of $\underline{\theta}^0$ to stress that the unknown parameter has a natural matrix presentation. We then have, for each $i \in \mathbb{R}^p$,

$$\mathrm{d}x_i(t) = \sum_{j=1}^{m} \theta^0_{ij} f_j(\underline{x}(t))\, \mathrm{d}t + \mathrm{d}b_i(t)\,. \tag{3.1.5}$$

Suitable sets of basis functions can be provided by domain-specific knowledge. As an example, within stochastic models of chemical reactions, the drift coefficient is a low-degree polynomial. For instance, the reaction $\mathsf{A} + 2\mathsf{B} \to \mathsf{C}$ is modeled as $\mathrm{d}x_{\mathsf{C}} = k_{\mathsf{C},\mathsf{AB}} x_{\mathsf{A}} x_{\mathsf{B}}^2 \mathrm{d}t + \mathrm{d}b_{\mathsf{C}}$, where x_A, x_B and x_C denote the concentration of the species A, B and C respectively, and $\mathrm{d}b_C$ is a chemical noise term. In order to learn a model of this type, one can consider a basis of functions $F(x)$ that comprises all monomials up to a maximum degree.

An important subclass of models of the last type is provided by linear SDEs. In this case, the drift is a linear function of $\underline{x}(t)$, namely $F(\underline{x}; \Theta^0) = \Theta^0 \underline{x}(t)$ with $\Theta^0 \equiv (\theta^0_{ij})_{i,j\in[p]} \in \mathbb{R}^{p\times p}$. Explicitly, for each $i \in \mathbb{R}^p$,

$$\mathrm{d}x_i(t) = \sum_{j=1}^{p} \theta^0_{ij} x_j(t)\, \mathrm{d}t + \mathrm{d}b_i(t)\,. \tag{3.1.6}$$

A model of this type is a good approximation for many systems near a stable equilibrium. The model (3.1.6) can be used to trace fluctuations of the species' concentrations in proximity to an equilibrium point. The matrix Θ^0 would represent in this case the linearized interactions between different chemical factors.

More generally, we can associate to the model (3.1.6) a directed graph $G = (V, E)$ with edge weight $\theta^0_{ij} \in \mathbb{R}$ associated with the directed edge (j, i) from $j \in V$ to $i \in V$. Each component $x_i(t)$ of the vector $\underline{x}(t)$ describes the state of a node $i \in V$. The graph G describes which nodes interact: the rate of change of $x_i(t)$ is given by a weighted sum of the current values of its neighbors, corrupted by white noise. In other words linear SDE's can be seen as graphical models – a probabilistic model parametrized by a graph.

This thesis establishes lower bounds on the sample-complexity for estimating the general model (3.1.1). These bounds are based on information theoretic techniques and apply irrespective of computational considerations. For linear models of the form (3.1.6), we put forward a low-complexity estimator and derive upper bounds on its sample-complexity. Upper and lower bounds are shown to be within a constant factor for special classes of sparse networks Θ^0.

3.2 Related work

The problem of estimating the parameters of a diffusion plays a central role in several applied domains, the most notable being econometrics, chemistry and system biology.

In the first context, diffusions are used to model the evolution of price indices [92]. While the most elementary process is the (geometric) Brownian motion [10, 19], a number of parametric families have been introduced to account for nonlinearities. The number of parameters is usually small and parameter estimation is addressed via maximum likelihood (ML). We refer to [12, 76] for proofs of consistency and asymptotic normality of the ML estimator. Much of the recent research has focused on dealing with the challenges posed by the fact that the diffusion is sampled at discrete intervals, and the transition probabilities cannot be computed in closed form. A short list of contributions on this problem includes [31, 91, 5]. In particular, asymptotically consistent methods based on approximate transition probabilities exist, see for instance [90, 26]. Nonparametric estimation of the drift coefficient has been studied as well [39, 104, 32].

Let us emphasize that all of these works focus on the low-dimensional setting: the vector of parameters to be estimated is p-dimensional, and the diffusion is observed for a time $T \to \infty$. Hence there is little overlap with the present work. In particular, simple ML estimators are not viable in the high-dimensional setting. At the same time, it would be interesting to address the problems posed by discrete sampling and non-parametric estimation in the high-dimensional setting as well.

Applications to chemistry and system biology have been mentioned in Section 3.1. A large variety of chemical reaction is modeled by diffusions with suitably parametrized drift terms [54, 60]. Of particular interest here are special classes of drift coefficients, for instance those exhibiting time-scale separation [89] or gradients of a potential [93]. As for econometrics applications, these works have focused on low-dimensional diffusions.

Technically, our work fits on recent developments in learning high-dimensional graphical models. The typical setting assumes that the data are i.i.d. samples from a high-dimensional Gaussian distribution with sparse inverse covariance. The underlying graph structure (the support of the inverse covariance) is estimated using convex regularizations that promote sparsity. Well known examples include the *graphical* LASSO [44] and the pseudo-likelihood method of [81]. In the context of binary pairwise graphical models, similar methods were developed in Ref. [110]. To the best of our knowledge the work described by this thesis is the first one moving beyond the assumption of i.i.d. samples. While we extend ideas and methods from this literature, dealing with dependent samples raises new mathematical challenges.

Our methods build on the work on ℓ_1-regularized least squares, and its variants [107, 36, 37, 115, 109]. The most closely related results are the one concerning high-dimensional consistency for support recovery [81, 110, 116]. Our proof for our upper bound follows indeed the approach developed in these papers, with two important challenges. First, the design matrix is in our case produced by a stochastic diffusion, and it does not necessarily satisfies the irrepresentability conditions used by these works. Second, the observations are not independent and therefore elementary concentration inequalities are not sufficient.

Most of these proofs build on the technique of [116]. A naive adaptation to the present case allows to prove some performance guarantee for the discrete-time setting. However the resulting bounds are not uniform as the sampling interval η tends to 0 for $n\eta = T$ fixed. In particular, they do not allow to prove an analogous of our continuous time result, Theorem 3.3.1. A large part of our effort is devoted to proving more accurate probability estimates that capture the correct scaling for small η.

Finally, the related topic of learning graphical models for autoregressive processes was studied recently in [102, 103]. These papers propose a convex relaxation that is different from the one studied in this paper, without however estabilishing high-dimensional consistency for model selection.

Preliminary report of our work were presented at NIPS 2010 [14] and ISIT 2011 [15]. Subsequent work by Bolstad, Van Veen and Nowak [20] establishes high-dimensional consistency for estimating autoregressive models through a related approach. These guarantees are non-uniform in the sampling rate η.

3.3 Main results

Our main contributions are the following:

Information-theoretic lower bound. We establish a general lower bound on the sample complexity for estimating the signed support of the drift coefficient of a diffusion of the form (3.1.1). By specializing this result, we obtain bounds for the linearly parametrized model (3.1.5), and the linear model (3.1.6).

Upper bound via penalized least squares. For the linear model (3.1.6), and suitable classes of sparse matrices Θ^0, we prove high-dimensional consistency of the penalized least-squares method introduced in Section 3.3.2. The resulting upper bound on sample complexity matches the information theoretic lower bound up to constant factors.

In this section we focus on the case of sparse linear SDE's, stating upper and lower bounds in this case, cf. Section 3.3.3. We then illustrate the general theory by analyzing a specific but rich problem: learning the Laplacian of a sparse graph, cf. Section 3.3.4.

Related results extending the ones presented here (in particular, general lower bounds on the sample complexity) are discussed in Section 3.5 and in Section 3.7.

3.3.1 Notation

Let us recall some important notation is used in this chapter.

For $N \in \mathbb{N}$, we let $[N] = \{1, 2, \dots, q\}$. Given a matrix Q, its transpose is denoted by Q^* and its support $\mathrm{supp}(Q)$ is the $0-1$ matrix such that $\mathrm{supp}(Q)_{ij} = 1$ if and only if $Q_{ij} \neq 0$. The support $\mathrm{supp}(v)$ is defined analogously for a vector $v \in \mathbb{R}^N$. With a slight abuse of notation, we occasionally write $\mathrm{supp}(v)$ for the subset of indices $i \in [N]$ such that $v_i \neq 0$. The *signed support* of a matrix (or vector) Q, denoted by $\mathrm{sign}(Q)$, is the matrix defined by $\mathrm{sign}(Q)_{ij} = \mathrm{sign}(Q_{ij})$ if $Q_{ij} \neq 0$ and $\mathrm{sign}(Q)_{ij} = 0$ otherwise. The r-th row of a matrix Q is denoted by Q_r.

Given a matrix $Q \in \mathbb{R}^{M \times N}$, and sets $L \subseteq [M]$, $R \subseteq [N]$, we denote by $Q_{L,R}$ the sub-matrix $Q_{L,R} \equiv (Q_{ij})_{i \in L, j \in R}$.

For $q \geq 1$, the ℓ_q norm of a vector $v \in \mathbb{R}^N$ is given by $\|v\|_q \equiv (\sum_{i \in [N]} |v_i|^q)^{1/q}$. This is extended in the usual way to $q = \infty$. As usual, the misnomer '0-norm' is used for the size of the support ov v, namely $\|v\|_0$ is the number of non-zero entries of v. The ℓ_q operator norm of a matrix $Q \in \mathbb{R}^{M \times N}$ is denoted by $|||Q|||_q$. In fact, we only use the ℓ_∞ operator norm, which is given by $|||Q|||_\infty \equiv \max_{r \in [M]} \|Q_r\|_1$.

If $Q \in \mathbb{R}^{N \times N}$ is symmetric, then its eigenvalues are denoted by $\lambda_1(Q) \leq \lambda_2(Q) \leq \cdots \leq \lambda_N(Q)$. The minimum and maximum eigenvalues are also denoted as $\lambda_{\min}(Q) \equiv \lambda_1(Q)$ and $\lambda_{\max}(Q) \equiv \lambda_N(Q)$. For a general (non-symmetric) matrix $Q \in \mathbb{R}^{M \times N}$ we let $0 \leq \sigma_1(Q) \leq \cdots \leq \sigma_{M \wedge N}(Q)$ denote its singular values. Further $\sigma_{\min}(Q) = \sigma_1(Q)$ and $\sigma_{\max}(Q) = \sigma_{M \wedge N}(Q)$ are the minimum and maximum singular values.

Throughout this chapter, we denote by C, C_1, C_2, etc, constants that can be adjusted from point to point.

3.3.2 Regularized least squares

Before introducing the upper bounds for the sample-complexity of learning SDEs we must introduce the algorithm with which we achieve them.

Regularized least squares is an efficient and well-studied method for support recovery. In order to describe its application to estimating the signed support of the drift coefficient of a high-dimensional diffusion, we consider the general linearly parametrized model (3.1.5).

We estimate independently each row of the matrix $\Theta^0 \in \mathbb{R}^{p\times m}$. The $r^{\rm th}$ row, denoted by Θ^0_r, is estimated by solving the following convex optimization problem

$$\min_{\Theta_r \in \mathbb{R}^p} \mathcal{L}(\Theta_r; \{\underline{x}(t)\}_{t\in[0,T]}) + \lambda \|\Theta_r\|_1 \,, \tag{3.3.1}$$

where the log-likelihood function $\mathcal{L}$ is defined by

$$\mathcal{L}(\Theta_r; \{\underline{x}(t)\}_{t\in[0,T]}) = \frac{1}{2T}\int_0^T \langle \Theta_r, F(\underline{x}(t))\rangle^2 \,{\rm d}t - \frac{1}{T}\int_0^T \langle \Theta_r, F(\underline{x}(t))\rangle \,{\rm d}x_r(t)\,. \tag{3.3.2}$$

(Here and below $\langle u, v\rangle$ denotes the standard scalar product of vectors $u, v \in \mathbb{R}^N$.). We denote this algorithm by $\mathsf{Rls}(\lambda)$.

Notice that in Chapter 2 we used two different notations ($\mathcal{L}^n$ and $\mathcal{L}$) to distinguish the likelihood function when $n < \infty$ from when $n = \infty$. In this chapter, $\mathcal{L}$ always represents the case $T < \infty$.

The ℓ_1 regularization term in Eq. (3.3.1) has the role of shrinking to 0 all the entries θ_{rj}, except the most significant ones, thus effectively selecting the support of Θ.

The normalized log-likelihood function (3.3.2) is the appropriate generalization of the sum of square residuals for a continuous-time process. To see this heuristically, one can *formally* write $\dot{x}_r(t) = {\rm d}x_r(t)/{\rm d}t$. A naive sum of square residuals would take the form $\int (\langle \Theta_r, F(\underline{x}(t))\rangle - \dot{x}_r(t))^2 {\rm d}t$. Unfortunately, this expression is not defined because $x_r(t)$ is not differentiable. On the other hand, expanding the square, we get $2T\mathcal{L}(\Theta_r; \{\underline{x}(t)\}_{[0,T]}) + \int (\dot{x}_r(t))^2 {\rm d}t$. The first term is well defined, as is clear from Eq. (3.3.2), and the second is independent of Θ and hence can be dropped.

Notice that constructing a well-defined cost function as in Eq. (3.3.2) is not a purely academic problem. Indeed, a cost function that included the time derivative $\dot{\underline{x}}(t)$ would in practice require to estimate $\dot{\underline{x}}(t)$ itself. This is all but hopeless because $\dot{\underline{x}}(t)$ does not exist in the model.

3.3.3 Sample complexity for sparse linear SDE's

In order to state our results, it is convenient to define the class of sparse matrices $\mathcal{A}^{(S)}$, depending on parameters $\Delta, p \in \mathbb{N}$, $\Delta \geq 3$, $\theta_{\min}, \rho_{\min} > 0$

$$\mathcal{A}^{(S)} = \mathcal{A}^{(S)}(\Delta, p, \theta_{\min}, \rho_{\min}) \subseteq \mathbb{R}^{p\times p} \tag{3.3.3}$$

by letting $\Theta \in \mathcal{A}^{(S)}$ if and only if

(*i*) $\|\Theta_r\|_0 \leq \Delta$ for all $r \in [p]$.

(*ii*) $|\theta_{ij}| \geq \theta_{\min}$ for all $i, j \in [p]$ such that $\theta_{ij} \neq 0$.

(*iii*) $\lambda_{\min}(-(\Theta + \Theta^*)/2) \geq \rho_{\min} > 0$.

Notice in particular that condition (*iii*) implies that the system of linear ordinary differential equations $\dot{\underline{x}}(t) = \Theta \underline{x}(t)$ is stable. Equivalently, the spectrum of Θ is contained in the half plane $\{z \in \mathbb{C} \, : \, \mathrm{Re}(z) < 0\}$. As a consequence, if $\Theta^0 \in \mathcal{A}^{(S)}$, then the diffusion process (3.1.6) has a unique stationary measure which is Gaussian with covariance $Q^0 \in \mathbb{R}^{p\times p}$ and is given by the unique solution of Lyapunov's equation [117]

$$\Theta^0 Q^0 + Q^0 (\Theta^0)^* + I = 0. \tag{3.3.4}$$

Hence $X_0^T = \{\underline{x}(t) : \; t \in [0, T]\}$ is stationary trajectory distributed according to the linear model (3.1.6) if $\underline{x}(t = 0) \sim \mathsf{N}(0, Q^0)$ is a Gaussian random variable independent of $\underline{b}(t)$.

We consider the linear model (3.1.6) with $\Theta^0 \in \mathcal{A}^{(S)}$. Considering a row index $r \in [p]$, let $S^0 = S^0(r)$ be the support of Θ^0_r.

Assumption 1 (Restricted convexity). For $C_{\min} > 0$, we have

$$\lambda_{\min}(Q^0_{S^0,S^0}) \geq C_{\min} \, . \tag{3.3.5}$$

Assumption 2 (Irrepresentability). For some $\alpha > 0$, we have

$$\|{Q^0}_{(S^0)^C,S^0} \left({Q^0}_{S^0,S^0}\right)^{-1} \|_\infty \leq 1 - \alpha \, . \tag{3.3.6}$$

We refer to [116, 81] for the original development of these conditions in the context of sparse regression.

Our first theorem establishes high-dimensional consistency of ℓ_1-penalized least squares for estimating $\mathrm{sign}(\Theta^0)$ from a stationary trajectory X_0^T according to the linear model (3.1.6) when $\Theta^0 \in \mathcal{A}^{(S)}$. The details of its proof can be found in Section 3.5 and Appendix B.5. In particular in Appendix B.5.

Theorem 3.3.1. *If $\Theta^0 \in \mathcal{A}^{(S)}(\Delta, p, \theta_{\min}, \rho_{\min})$ satisfies assumptions* 1 *and* 2 *above for all $r \in [p]$ and some $C_{\min}, \alpha > 0$, then there exists $\lambda = \lambda(T)$ such that*

$$T_{\mathsf{Rls}(\lambda)}(\Theta^0) \leq \frac{2 \cdot 10^4 \Delta^2 (\Delta\, \rho_{\min}^{-2} + \theta_{\min}^{-2})}{\alpha^2 \rho_{\min} C_{\min}^2} \log\Big(\frac{4p\Delta}{\delta}\Big). \tag{3.3.7}$$

In particular, one can choose

$$\lambda = \sqrt{\frac{36}{T\alpha^2 \rho_{\min}} \log\Big(\frac{4p}{\delta}\Big)}\,. \tag{3.3.8}$$

Remark 3.3.1. *Note that our notion of sample-complexity is well-defined for reconstruction algorithms that depend on T, the length of the stationary trajectory X_0^T. This is the case with the regularized least squares algorithm $\mathsf{Rlr}(\lambda)$, since λ can depend on T.*

Corollary 3.3.2. *If there exists $C_{\min}, \alpha > 0$ such that assumptions 1 and 2 hold for all $r \in [p]$ and for all $\Theta^0 \in \mathcal{A}^{(S)}(\Delta, p, \theta_{\min}, \rho_{\min})$, then we can replace $T_{\mathsf{Rls}(\lambda)}(\Theta^0)$ by $T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(S)})$ in* (3.3.7).

The next theorem establishes a lower bound on the sample complexity of learning the signed support of $\Theta^0 \in \mathcal{A}^{(S)}$ from a stationary trajectory, X_0^T, distributed according to the linear model (3.1.6). The details of the proof of this theorem can be found in appendix B.6. In particular, in appendix B.6.2.

Theorem 3.3.3. *Let $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ be an estimator of* $\mathrm{sign}(\Theta^0)$. *There is a constant $C(\Delta, \delta)$, such that, for all p large enough,*

$$T_{\mathsf{Alg}}(\mathcal{A}^{(S)}) \geq C(\Delta, \delta) \max\Big\{\frac{\rho_{\min}}{\theta_{\min}^2}, \frac{1}{\theta_{\min}}\Big\} \log p\,. \tag{3.3.9}$$

The last two theorems establish that, under assumptions 1 and 2 above, the time complexity of learning the support of the diffusion coefficient for sparse linear SDEs in the class $\mathcal{A}^{(S)}$ is of order $\log p$.

Notice that both upper and lower bounds depend in a non-trivial way on the parameter $\rho_{\min}$. In order to gain intuition on this quantity, consider Eq. (3.1.6) in absence of the driving term $\mathrm{d}b_i(t)$. By using the Lyapunov function $\|x(t)\|_2^2$, it is easy to verify that $\|x(t)\|_2 \leq \|x(0)\|_2 \, e^{-\rho_{\min} t/2}$. Hence $\rho_{\min}^{-1}$ provides a general upper bound on the mixing time of the diffusion (3.1.6). The upper bound is essentially tight if the matrix Θ^0 is symmetric.

Theorems 3.3.1 and 3.3.3 can therefore be used to characterize the dependence of the sample complexity on the mixing time. One subtle aspect is that $C_{\min}$ and $\rho_{\min}$ cannot be varied independently because of the Lyapunov equation, Eq. (3.3.4). In order to clarify this dependency, we apply our general results to the problem of learning the Laplacian of an undirected graph.

3.3.4 Learning the laplacian of graphs with bounded degree

Given a simple graph $G = (V, E)$ on vertex set $V = [p]$, its Laplacian Δ^G is the symmetric $p \times p$ matrix which is equal to the adjacency matrix of G outside the diagonal, and with entries $\Delta^G_{ii} = -\mathrm{deg}(i)$ on the diagonal [27]. (Here $\mathrm{deg}(i)$ denotes the degree of vertex i.)

It is well known that Δ^G is negative semidefinite, with one eigenvalue equal to 0, whose multiplicity is equal to the number of connected components of G. The matrix $\Theta^0 = -m\, I + \Delta^G$ fits into the setting of Theorem 3.3.1 for $m > 0$. The corresponding model (3.1.6) describes the over-damped dynamics of a network of masses connected by springs of unit strength, and connected by a spring of strength m to the origin.

Let $\mathcal{G}_{\text{bounded}} = \mathcal{G}_{\text{bounded}}(\Delta, p)$ be the class of graphs on p nodes with maximum vertex degree bounded by Δ. Define,

$$\mathcal{A}^{(L)}(m, p, \Delta) = \{\Theta^0 : \Theta^0 = -m\, I + \Delta^G, \text{ st. } m > 0, G \in \mathcal{G}_{\text{bounded}}\} \tag{3.3.10}$$

The following result holds. Its proof can be found in the appendix of [14, 13].

Theorem 3.3.4. *Consider the model* (3.1.6) *with* $\Theta^0 \in \mathcal{A}^{(L)}(m, p, \Delta)$*. Then there exists* λ *such that*

$$T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(L)}) \leq 4 \cdot 10^5 \Delta^2 \left(\frac{\Delta + m}{m}\right)^5 (\Delta + m^2) \log\left(\frac{4p\Delta}{\delta}\right), \tag{3.3.11}$$

In particular one can take, $\lambda = \sqrt{36(\Delta + m)^2 \log(4p/\delta)/(Tm^3)}$*.*

In other words, for m bounded away from 0 and ∞, regularized least squares regression correctly reconstructs the graph $\mathcal{G}$ from a trajectory of time length which is polynomial in the degree and logarithmic in the graph size.

Using this theorem we can write the following corollary that helps compare the bounds obtained in Theorems 3.3.1 and 3.3.3 above.

Corollary 3.3.5. *Assume the same setting as in Theorem 3.3.4. There exist constants* $\lambda = \lambda(T)$*,* $C_1 = C_1(\Delta, \delta)$ *and* $C_2 = C_2(\Delta, \delta)$ *such that, for all* p *large enough,*

$$m < \Delta \;\Rightarrow\; C_1 \log p \leq T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(L)}) \leq C_2 m^{-5} \log p, \tag{3.3.12}$$

$$m \geq \Delta \;\Rightarrow\; C_1 m \log p \leq T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(L)}) \leq C_2 m^2 \log p\,. \tag{3.3.13}$$

In addition, the lower-bounds hold regardless of the choice of λ*.*

Proof. The proof of this corollary follows immediately from Theorem 3.3.4 and Theorem 3.3.3. □

Notice that the upper bound on T_{Rlr} presents a non-trivial behavior in m. It diverges both at large m, and at small m. The reasons of these behaviors are different. For small m, the mixing time of the diffusion (which is proportional to $1/m$) gets large, and hence a large time is necessary to accumulate information about Θ^0. Vice-versa for large m, Θ^0 gets close to $-m\,I$ and hence it depends weakly on the graph structure.

Notice that the lower bound also diverges as $m \to \infty$, hence confirming the above picture. On the other hand, the behavior of T_{Rls} as $m \to 0$ remains an open question since our lower bound stays bounded in that limit.

3.4 Important remark

Given that $\mathsf{Rls}(\lambda)$ can be tuned using λ, it is natural to asks whether we can write the above results in a from resembling Theorem 2.3.6 of Chapter 2 for $\mathsf{Rlr}(\lambda)$. In particular, if we define

$$T_{\mathsf{Rls}}(\Theta^0) = \inf\{T_0 \in \mathbb{R}^+ : \sup_{\lambda>0} \mathbb{P}_{\Theta^0,T}\{\mathsf{Rls}(\lambda) = \text{sign}(\Theta^0)\} \geq 1-\delta \text{ for all } T \geq T_0\},$$

can we say something about $T_{\mathsf{Rls}}(\mathcal{A}) = \max_{\Theta^0 \in \mathcal{A}} T_{\mathsf{Rls}}(\Theta^0)$?

Since $T_{\mathsf{Rls}}(\Theta^0) \leq T_{\mathsf{Rls}(\lambda)}(\Theta^0)$ for all λ and for all Θ^0, the upper bounds in this chapter still hold if $T_{\mathsf{Rls}(\lambda)}$ is replaced by T_{Rls}. However, the same is not true for the lower bounds.

Here is the difference. The lower bound of Theorem 2.3.6 tell us that, for n smaller than a certain value (∞ in this case), there exists a graph G such that, for all $\lambda > 0$, $\mathsf{Rlr}(\lambda)$ fails with high probability. The lower bound of Theorem 3.3.3 says that, for T smaller than a certain value, for every $\lambda > 0$ there exists a graph G [4] such that $\mathsf{Rls}(\lambda)$ fails with high probability.

In other words, in the first case we prove there is a graph so 'pathological' that $\mathsf{Rlr}(\lambda)$ fail on this graph no matter what λ is, while in the second case, we did not find such a graph. Rather, for each λ we might have to find a different graph to make $\mathsf{Rls}(\lambda)$ fail.

There reason for this difference can be traced back to the kind of argument used to prove the two bounds. The lower bound of Chapter 2 is proved by showing that a certain necessary condition for success is violated uniformly over λ for certain graphs. The lower bound of this chapter however, is proved by a 'counting' argument. Concretely, unless T is large enough, the amount of information available is not enough for any algorithm to distinguish among the graphs in a certain class of graphs. Just like with one bit we cannot distinguish among more than 2 objects/graphs. Hence, for a particular algorithm (i.e. a particular λ), there are graphs in this class that cannot be distinguished from each other. The particular graphs that cannot be distinguished

[4] G is the support of Θ^0.

might dependent on the algorithm (i.e. on λ).

Does this mean that, for every graph G we can find a λ for which $\mathsf{Rls}(\lambda)$ succeeds? In principle yes. However, we do not know G a priori, λ cannot be a function of G. What is true is that, $\lambda = \lambda(X_0^T)$ and hence the right way to interpret our lower bound is: for any function $\lambda = \lambda(X_0^T)$, if T is smaller than what the lower bound prescribes, then there exists a graph that with high probability cannot be correctly recovered by Rls.

A more conceptual way to distinguish the results from Chapter 2 and the results from this chapter is as follows. Regarding the Ising model, the lower bounds were derived to prove that the algorithms fail on certain graphs. Here the lower bounds are derived to inform us about how tight the upper bounds are.

Finally, it is important to point out that, although the results in this chapter cannot be encapsulated in a form similar to the one of Chapter 2, the form in which we presented them in Sections 3.3.3 and 3.3.4 is certainty not the only one.

For example, the following alternative theorem also follows from the proofs of our main results. Given a class $\mathcal{A}$ of parameters Θ^0 and a probability distribution over this class, define

$$T_{\mathsf{Rls}}(\mathcal{A}) = \inf\{T_0 \in \mathbb{R}^+ : \sup_{\lambda > 0} \mathbb{E}\{\mathbb{P}_{\Theta^0,T}\{\mathsf{Rls}(\lambda) = \text{sign}(\Theta^0)\}\} \geq 1 - \delta \text{ for all } T \geq T_0\}.$$

Above, $\mathbb{E}$ represents expectation over the random variable $\Theta^0 \in \mathcal{A}$.

Theorem 3.4.1. *Consider the model* (3.1.6)*. There exists a constant $C(\Delta, \delta)$ such that, for all p large enough,*

$$T_{\mathsf{Rls}}(\mathcal{A}^{(L)}) \leq 4 \cdot 10^5 \Delta^2 \left(\frac{\Delta + m}{m}\right)^5 (\Delta + m^2) \log\left(\frac{4p\Delta}{\delta}\right)$$

and

$$T_{\mathsf{Rls}}(\mathcal{A}^{(L)}) \geq C(\Delta, \delta) \max\{m, 1\} \log p \,.$$

Note that in the above theorem there is no longer an explicit reference to λ.

3.5 Important steps towards the proof of main results

In this section we begin by presenting an analogous of Theorem 3.3.1 for the case of a discrete time system. This is an important result in itself and also constitutes the basis for the proof of Theorem 3.3.1. In fact, Theorem 3.3.1 is proved by letting $\eta \to 0$ in the result bellow. We then present a general lower bound on the time complexity of learning continuous stochastic differential equations. Theorem 3.3.3 follows as a consequence of this general bound. Later, in Section 3.5.2, using this result, lower bounds for the time complexity of linear SDE's with dense matrices Θ^0 and non-linear SDE's are derived.

3.5.1 Discrete-time model

The problem of learning stochastic differential equations in discrete time is important in itself and also because it relates to the problem of learning a continuous-time stochastic differential equation from discretely sampling its continuous trajectory. Focusing on continuous-time dynamics allowed us to obtain the elegant statements of Section 3.3.3. However, much of the theoretical analysis concerning the regularized least squares algorithm is in fact devoted to the analysis of the following discrete-time dynamics, with parameter $\eta > 0$:

$$\underline{x}(t) = \underline{x}(t-1) + \eta\Theta^0\underline{x}(t-1) + \underline{w}(t), \qquad t \in \mathbb{N}_0\,. \tag{3.5.1}$$

Here $\underline{x}(t) \in \mathbb{R}^p$ is the vector collecting the dynamical variables, $\Theta^0 \in \mathbb{R}^{p\times p}$ specifies the dynamics as above, and $\{\underline{w}(t)\}_{t\geq 0}$ is a sequence of i.i.d. normal vectors with covariance $\eta\,\mathbb{I}_{p\times p}$ (i.e. with independent components of variance η). We assume that consecutive samples $X_0^n \equiv \{\underline{x}(t) : 0 \leq t \leq n\}$ are given and ask under which conditions regularized least squares reconstructs the signed support of Θ^0.

The parameter η has the meaning of a time-step size. The continuous-time model (3.1.6) is recovered, in a sense made precise below, by letting $\eta \to 0$. Indeed we prove reconstruction guarantees that are uniform in this limit as long as the product $n\eta$

(which corresponds to the time interval T in the Section 3.3.3) is kept constant. For a formal statement we refer to Theorem 3.5.1. Theorem 3.3.1 is indeed proved by carefully controlling this limit. The mathematical challenge in this problem is related to the fundamental fact that the samples $\{\underline{x}(t)\}_{0\le t\le n}$ are dependent (and strongly dependent as $\eta \to 0$).

Discrete time models of the form (3.5.1) can arise either because the system under study evolves by discrete steps, or because we are sub-sampling a continuous time system modeled as in Eq. (3.1.1). Notice that in the latter case the matrices Θ^0 appearing in Eq. (3.5.1) and (3.1.1) coincide only to the zeroth order in η. Neglecting this technical complication, the uniformity of our reconstruction guarantees as $\eta \to 0$ has an appealing interpretation already mentioned above. Whenever the samples spacing is not too large, the time complexity (i.e. the product $n\eta$) is roughly independent of the spacing itself.

Consider a system evolving in discrete time according to the model (3.5.1), and let X_0^n be the observed portion of the trajectory. The r^{th} row Θ_r^0 is estimated by solving the following convex optimization problem

$$\underset{\Theta_r\in\mathbb{R}^p}{\text{minimize}} \quad \mathcal{L}(\Theta_r; X_0^n) + \lambda\|\Theta_r\|_1\,, \tag{3.5.2}$$

where

$$\mathcal{L}(\Theta_r; X_0^n) \equiv \frac{1}{2\eta^2 n}\sum_{t=0}^{n-1}\left\{x_r(t+1) - x_r(t) - \eta\,\langle\Theta_r, \underline{x}(t)\rangle\right\}^2\,. \tag{3.5.3}$$

Apart from an additive constant, the $\eta \to 0$ limit of this cost function can be shown to coincide with the cost function in the continuous time case, cf. Eq. (3.3.2). Indeed the proof of Theorem 3.3.1 amounts to a more precise version of this statement. Furthermore, $\mathcal{L}(\Theta_r; X_0^n)$ is easily seen to be the log-likelihood of Θ_r within model (3.5.1).

Let us introduce the class of sparse matrices $\mathcal{A}'^{(S)}$ as being exactly equal to the class $\mathcal{A}^{(S)}$ introduced in Section 3.3.3 but with condition (iii) replaced by

$$\frac{1-\sigma_{\max}(I+\eta\,\Theta^0)}{\eta} \ge D > 0 \tag{3.5.4}$$

If $\Theta^0 \in \mathcal{A}'^{(S)}$ then, under the model (3.5.1), $\underline{x}(t)$ has a unique stationary measure which is Gaussian with covariance Q^0 determined by the following modified Lyapunov equation

$$\Theta^0 Q^0 + Q^0(\Theta^0)^* + \eta\Theta^0 Q^0(\Theta^0)^* + I = 0\,. \tag{3.5.5}$$

It will be clear from the context whether Θ^0 (or Q^0) refers to the dynamics matrix (or covariance of the stationary distribution) from the continuous or discrete time system.

The following theorem establishes the conditions under which ℓ_1-penalized least squares recovers $\text{sign}(\Theta^0)$ with high probability. Its proof can be found in Appendix B.4. The adaptation of the proof of this theorem to the proof of the main Theorem 3.3.1 can be found in Appendix B.5.

Theorem 3.5.1. *Assume that $\Theta^0 \in \mathcal{A}'^{(S)}(\Delta, p, \theta_{\min}, D)$ and that Θ^0_r satisfies assumptions 1 and 2 of Section 3.3.3 for constants $C_{\min}, \alpha > 0$. Let X_0^n be a stationary trajectory distributed according to the linear model* (3.5.1). *There exists $\lambda = \lambda(n\eta) > 0$ such that if*

$$n\,\eta \leq \frac{10^4\Delta^2(\Delta D^{-2} + \theta_{\min}^{-2})}{\alpha^2 D C_{\min}^2}\log\Big(\frac{4p\Delta}{\delta}\Big)\,., \tag{3.5.6}$$

the ℓ_1 regularized least-squares recovers the signed support of Θ^0 with probability larger than $1-\delta$. In particular one can take $\lambda = \sqrt{(36\ \log(4p/\delta))/(D\alpha^2 n\eta)}$.

In other words the discrete-time sample complexity, n, is logarithmic in the model dimension, polynomial in the maximum network degree and inversely proportional to the time spacing between samples. The last point is particularly important. It enables us to derive the bound on the continuous-time sample complexity as the limit $\eta \to 0$ of the discrete-time sample complexity. It also confirms our intuition mentioned in the Introduction: although one can produce an arbitrary large number of samples by sampling the continuous process with finer resolutions, there is limited amount of information that can be harnessed from a given time interval $[0, T]$.

Remark 3.5.1. *The form of Theorem 3.5.1 is different than that of Theorem 3.3.1. In Theorem 3.5.1 we do not compute a bound on $N_{\mathsf{Rls}(\lambda)}(\Theta^0)$, the sample-complexity* *of reconstructing* $\mathrm{sign}(\Theta^0)$*, but rather a bound on the sample-complexity,* n*, of reconstructing the signed support of a* particular *row* r*,* $\mathrm{sign}(\Theta^0_r)$*. Obviously, if assumptions 1 and 2 hold for the same constants* $C_{\min}, \alpha > 0$ *across* $r \in [p]$*, then replacing* δ *by* δ/p *in 3.5.6 allows us to use union bound an conclude that there exists* λ *for which*

$$N_{\mathsf{Rls}(\lambda)}(\Theta^0)\,\eta \leq \frac{2 \cdot 10^4 \Delta^2(\Delta D^{-2} + \theta_{\min}^{-2})}{\alpha^2 D C_{\min}^2} \log\left(\frac{4p\Delta}{\delta}\right). \tag{3.5.7}$$

(Notice the factor of 2*). The reason why we present Theorem 3.5.1 in a different form is to emphasize the fact that the proofs for the upper bounds focus only on the success of* Rls *for reconstructing a particular row* r*. This was the case in Chapter 2 is* Rlr *and is also the case with* Rls *here.*

3.5.2 General lower bound on time complexity

In this section we derive a general lower bound on the minimum time T required to learn a property $M(\Theta^0)$ associated to Θ^0 from a trajectory X_0^T distributed according to the general model (3.1.1). This result is used in Section 3.7 to derive lower bounds for the time complexity of linear SDE's with dense matrices Θ^0 (Section 3.7.1) and non-linear SDE's (Section 3.7.2).

The general form of the results in this section, and in the remainder of Section 3.7, is as follow: If $\widehat{M}_T(X_0^T)$, an estimator of $M(\Theta^0)$ based on X_0^T, achieves successful recover with probability greater than 1/2 for every Θ^0 in a class $\mathcal{A}$, then T must be greater then a certain value that is dependent on properties of $\mathcal{A}$ (cf. Theorems 3.7.1 and 3.7.2). These results however are a corollary of a more relaxed version (Theorem 3.5.2 and Corollary 3.5.3) where we only require that the expected rate of miss-estimation is small when Θ^0 is drawn *at random* from the ensemble $\mathcal{A}$. Clearly, if an estimator performs well over all $\Theta^0 \in \mathcal{A}$ then it must also perform well in expectation regardless of the distribution assumed over $\mathcal{A}$.

Special notation

Without loss of generality, in the remainder of this section, the parameter Θ^0 is a random variable chosen with some unknown prior distribution $\mathbb{P}_{\Theta^0}$ (subscript is be often omitted).

Regarding this section we have to point out a small change in our notation. Outside Section 3.5.2, where Θ^0 is a matrix of real numbers, $\mathbb{P}_{\Theta^0}$ represents a probability distribution over X_0^T parametrized by Θ^0. In this section however, subscripts indicate that probabilities and expectations are to be taken with respect to the random variable in the subscript. Hence, $\mathbb{P}_{\Theta^0}$ is a probability distribution *for* the random variable $\mathbb{P}_{\Theta^0}$

Unless specified otherwise, $\mathbb{P}$ and $\mathbb{E}$ denote probability and expectation with respect to the joint law of $\{\underline{x}(t)\}_{t\geq 0}$ and Θ^0. As mentioned before $X_0^T \equiv \{\underline{x}(t) : t \in [0,T]\}$ denotes the trajectory up to time T. Also, we define the variance of a vector-valued random variable as the sum of the variances over all components, i.e.,

$$\mathrm{Var}_{\Theta^0|X_0^t}(F(\underline{x}(t);\Theta^0)) = \sum_{i=1}^{p} \mathrm{Var}_{\Theta^0|X_0^t}(F_i(\underline{x}(t);\Theta^0)). \tag{3.5.8}$$

Results

The following general lower bound is a consequence of an identity between mutual information and the integral of conditional variance proved by Kadota, Zakai and Ziv [72] and a similar result by Duncan [38].

Theorem 3.5.2. *Let X_0^T be a trajectory of system* (3.1.1) *with initial state $\underline{x}(0)$ for a specific realization of the random variables $\underline{x}(0)$ and Θ^0. Let $\widehat{M}_T(X_0^T)$ be an estimator of $M(\Theta^0)$ based on X_0^T. If $\mathbb{P}_{\underline{x}(0),\Theta^0,X_0^T}(\widehat{M}_T(X_0^T) \neq M(\Theta^0)) < \frac{1}{2}$ then*

$$T \geq \frac{H(M(\Theta^0)) - 2I(\Theta^0;\underline{x}(0))}{\frac{1}{T}\int_0^T \mathbb{E}_{X_0^t}\{\mathrm{Var}_{\Theta^0|X_0^t}(F(\underline{x}(t);\Theta^0))\}\mathrm{d}t}. \tag{3.5.9}$$

Proof. Equation (3.1.1) can be regarded as describing a white Gaussian channel with feedback where Θ^0 denotes the message to be transmitted. For this scenario, Kadota et al. [72] give the following identity for the mutual information between X_0^T and Θ^0 when the initial condition is $\underline{x}(0) = 0$,

$$I(X_0^T; \Theta^0) = \frac{1}{2} \int_0^T \mathbb{E}_{X_0^t}\{\mathrm{Var}_{\Theta^0|X_0^t}(F(\underline{x}(t); \Theta^0))\}\mathrm{d}t. \tag{3.5.10}$$

For the general case where $\underline{x}(0)$ might depend on Θ^0 (if, for example, $\underline{x}(0)$ is the stationary state of the system) we can write $I(X_0^T; \Theta^0) = I(\underline{x}(0); \Theta^0) + I(X_0^T; \Theta^0|\underline{x}(0))$ and apply the previous identity to $I(X_0^T; \Theta^0|\underline{x}(0))$. Taking into account that

$$I(\widehat{M}_T(X_0^T)); M(\Theta^0)) \leq I(X_0^T; \Theta^0) \tag{3.5.11}$$

and making use of Fano's inequality

$$I(\widehat{M}_T(X_0^T)); M(\Theta^0)) \geq \mathbb{P}(\widehat{M}_T(X_0^T) = M(\Theta^0))H(\widehat{M}_T(X_0^T))) \tag{3.5.12}$$

the results follows. □

The bound in Theorem 3.5.2 is often too complex to be evaluated. Instead, the following corollary provides a more easily computable bound for the case when X_0^T is a stationary process.

Corollary 3.5.3. *Assume that* (3.1.1) *has a stationary distribution for every realization of* Θ^0 *and let* X_0^T *be a trajectory following any such stationary distribution for a specific realization of the random variable* Θ^0. *Let* $\widehat{M}_T(X_0^T)$ *be an estimator of* $M(\Theta^0)$ *based on* X_0^T. *If* $\mathbb{P}_{\Theta^0, X_0^T}(\widehat{M}_T(X_0^T) \neq M(\Theta^0)) < \frac{1}{2}$ *then*

$$T \geq \frac{H(M(\Theta^0)) - 2I(\Theta^0; \underline{x}(0))}{\mathbb{E}_{\underline{x}(0)}\{\mathrm{Var}_{\Theta^0|\underline{x}(0)}(F(\underline{x}(0); \Theta^0))\}}. \tag{3.5.13}$$

Proof. Since conditioning reduces variance, we have

$$\mathbb{E}_{X_0^t}\{\mathrm{Var}_{\Theta^0|X_0^t}(F(\underline{x}(t); \Theta^0))\} \leq \mathbb{E}_{\underline{x}(t)}\{\mathrm{Var}_{\Theta^0|\underline{x}(t)}(F(\underline{x}(t); \Theta^0))\}. \tag{3.5.14}$$

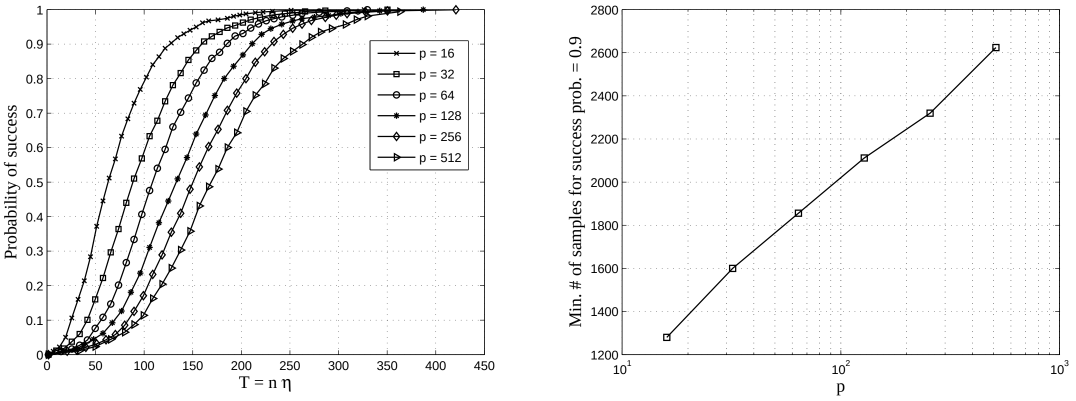

Figure 3.1: (left) Probability of success vs. length of the observation interval $n\eta$. (right) Sample complexity for 90% probability of success vs. p.

Using stationarity, we have

$$\mathbb{E}_{\underline{x}(t)}\{\mathrm{Var}_{\Theta^0|\underline{x}(t)}(F(\underline{x}(t);\Theta^0))\} = \mathbb{E}_{\underline{x}(0)}\{\mathrm{Var}_{\Theta^0|\underline{x}(0)}(F(\underline{x}(0);\Theta^0))\}, \tag{3.5.15}$$

which simplifies (3.5.9) to (3.5.13). □

In Section 3.7, we apply this lower bound to special classes of SDE's, namely linear SDE's with dense matrices Θ^0 and non-linear SDE's.

3.6 Numerical illustrations of the main theoretical results

In this section we present numerical validation of our main results on synthetic data. They confirm our observations in Theorems 3.3.1, 3.3.3 and 3.3.4 that the time complexity for learning linear sparse SDEs scales logarithmically with the number of nodes in the network p, given a constant maximum degree. They also confirm the implication of Theorem 3.5.1 that the time complexity is roughly independent of the sampling rate, assuming that we are in the regime of small η. Or, in other words, that our reconstruction guarantees are uniform in the sampling rate for small η. In Figures 3.1 and 3.2 we consider the discrete-time setting, generating data as follows. We draw $\tilde{\Theta}^0$ as a random sparse matrix in $\{0,1\}^{p\times p}$ with elements chosen independently at random with $\mathbb{P}(\theta^0_{ij} = 1) = \Delta/p$, $\Delta = 5$ and form $\Theta^0 = -7\mathbb{I} + \tilde{\Theta}^0$. The process $X_0^n \equiv \{\underline{x}(t) : 0 \le t \le n\}$ is then generated according to Eq. (3.5.1). Then we choose an $r \in [p]$ uniformly at random and solve the regularized least squares problem [5] for a different number of observations n and different values of λ. We record a 1 or a 0 if the correct signed support of Θ^0_r is recovered or not. For every value of n and λ, the probability of successful recovery is then estimated by taking the average of these errors over all realizations of Θ^0, X_0^n and r. Finally, for each fixed n, we take the maximum over λ of these probability of success. In other words, the numerical definition of sample-complexity we are using is

$$T_{\mathsf{Rls}}(\mathcal{A}) = \inf\{T_0 \in \mathbb{R}^+ : \sup_{\lambda>0} \hat{\mathbb{E}}\{\hat{\mathbb{P}}_{\Theta^0,T}\{\mathsf{Rls}(\lambda) = \text{sign}(\Theta^0)\}\} \ge 1-\delta \text{ for all } T \ge T_0\}.$$

Above, $\hat{\mathbb{P}}$ and $\hat{\mathbb{E}}$ represent empirical expectation and empirical probability and $\mathcal{A}$ is the class of all matrices that can be generated by the random procedure described before. Hence, the definition is a numerical approximation of the sample-complexity introduced in Section 3.4.

The plots we present would look similar similar if instead we used the notion of sample complexity adopted for our main results.

The left plot in Fig. 3.1 depicts the probability of success versus $n\eta$ for $\eta = 0.1$ and different values of p. Each curve is obtained using 2^{11} instances, and each instance is generated using a new random matrix Θ^0. The right plot in Fig.3.1 is the corresponding curve of the sample complexity versus p. The sample-complexity is computed for as the minimum value of $n\eta$ with probability of success of $1-\delta$ =90%. As predicted by Theorem 3.3.4, the curve shows a logarithmic scaling of the sample complexity with p.

In Fig. 3.2 we turn to the continuous-time model (3.1.6). Trajectories are generated by 'discretizing' this stochastic differential equation with a time step much smaller than the sampling rate η. We draw random matrices Θ^0 as before and plot the probability of success for $p = 16$, $\Delta = 4$ and different values of η, as a function of

[5] For discrete-time SDEs, the cost function is given explicitly in Eq. (3.5.2).

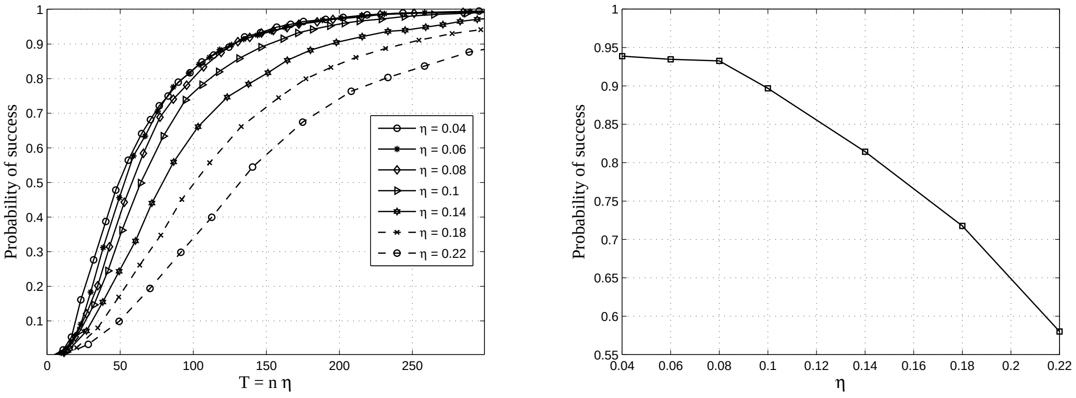

Figure 3.2: (right)Probability of success vs. length of the observation interval $n\eta$ for different values of η. (left) Probability of success vs. η for a fixed length of the observation interval, $(n\eta = 150)$. The process is generated for a small value of η and sampled at different rates.

T. We used 2^{11} instances for each curve. As predicted by Theorem 3.5.1, for a fixed observation interval T, the probability of success converges to some limiting value as $\eta \to 0$.

3.7 Extensions

In this section we present some extensions to our previous result.

3.7.1 Learning Dense Linear SDE's

A different regime of interest in learning the network of interactions for a linear SDE is the case of dense matrices. This regime exhibits fundamentally different behavior in terms of sample complexity compared to the regime of sparse matrices.

Let $\mathcal{A}^{(D)} \subset \mathbb{R}^{p\times p}$ be the set of dense matrices defined as $\Theta \in \mathcal{A}^{(D)}$ if and only if,

(i) $\theta_{\min} \leq |\theta_{ij}|p^{1/2} \leq \theta_{\max} \forall i, j : \theta_{ij} \neq 0$,

(ii) $\lambda_{\min}(-(\Theta + \Theta^*)/2) \geq \rho_{\min} > 0$.

The following theorem provides a lower bound for learning the signed support of models from the class $\mathcal{A}^{(D)}$ from stationary trajectories X_0^T of (3.1.6).

Theorem 3.7.1. *Let* $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ *be an estimator of* $\text{sign}(\Theta^0)$. *There is a constant* $C(\Delta, \delta)$ *such that, for all* p *large enough,*

$$T_{\mathsf{Alg}}(\mathcal{A}^{(D)}) \geq C(\Delta, \delta) \max\left\{\frac{\rho_{\min}}{\theta_{\min}^2}, \frac{1}{\theta_{\min}}\right\} p. \tag{3.7.1}$$

The sample complexity bound is similar to the one in Theorem 3.3.3 but the scaling with p has now increased from $O(\log p)$ to $O(p)$. The lack of structure in Θ^0 requires exponentially more samples for successful reconstruction. The proof of this theorem can be bound in appendix of [13].

Remark 3.7.1. *Although the above theorem only gives a lower bound on* $T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(D)})$, *it is not hard to upper bound* $T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(D)})$ *for linear dense systems of SDEs and certain values of* λ. *In particular, it is not hard to upper bound* $T_{\mathsf{Rls}(\lambda=0)}(\mathcal{A}^{(D)})$ *by* $O(p)$. *This can be done in two steps. First, taking* $\lambda = 0$, *one can compute a closed form solution for* Rlr. *This solution is an unbiased estimator involving sums of dependent Gaussian random variables. Second, one can prove concentrations bounds similar to the ones proved for Theorem 3.3.1, and compute the trajectory length* T *required to guarantee that*

$$\|\hat{\Theta} - \Theta^0\|_\infty \leq \theta_{\min}/2 \tag{3.7.2}$$

with probability greater than $1 - \delta$. *This value of* T *is an upper bound on* $T_{\mathsf{Rlr}(0)}(\mathcal{A}^{(D)})$ *since* (3.7.2) *is enough to guarantee that, using a simple thresholding* [6]

$$\text{sign}(\hat{\Theta}) = \text{sign}(\Theta^0). \tag{3.7.3}$$

3.7.2 Learning Non-Linear SDE's

In this section we assume that the observed samples X_0^T come from a stochastic process driven by a general SDE of the form (3.1.1).

[6] If $|\hat{\theta}_{ij}| < \theta_{\min}/2$ declare 0, if $\hat{\theta}_{ij} < -\theta_{\min}/2$ declare -1 and if $\hat{\theta}_{ij} > \theta_{\min}/2$ declare $+1$.

We recall that, v_i denotes the i^{th} component of vector v. For example, $x_3(2)$ is the 3^{th} component of the vector $\underline{x}(t)$ at time $t = 2$. In this section, $JF(\,\cdot\,; \Theta^0) \in \mathbb{R}^{p\times p}$ denotes the Jacobian of the function $F(\,\cdot\,; \Theta^0)$.

For fixed L, B and $D \geq 0$, define the class of functions $\mathcal{A}^{(N)} = \mathcal{A}^{(N)}(L, B, D)$ by letting $F(\underline{x}; \Theta) \in \mathcal{A}^{(N)}$ if and only if

(i) the support of $JF(\underline{x}; \Theta)$ has at most Δ non-zero entries for every $\underline{x}$,

(ii) the SDE (3.1.1) admits a stationary solution with covariance matrix, Q^0, satisfying $\lambda_{\min}(Q^0) \geq L$,

(iii) $\text{Var}_{\underline{x}(0)|\Theta}(x_i(0) \leq B \;\forall i$,

(iv) $|\partial F_i(\underline{x}; \Theta)/\partial x_j| \leq D$ for all $\underline{x} \in \mathbb{R}^p$ $i, j \in [p]$.

For simplicity we write $F(\underline{x}; \Theta^0) \in \mathcal{A}^{(N)}$ as $\Theta^0 \in \mathcal{A}^{(N)}$.

Our objective is different than before. Given $\Theta^0 \in \mathcal{A}^{(N)}$, we are interested in recovering the smallest support $M(\Theta^0)$ for which $\text{supp}(JF(\underline{x}; \Theta^0)) \subseteq M(\Theta^0)$ $\forall \underline{x}$. Hence, we consider the following modified definition of sample-complexity that can be applied to learning SDEs of the form (3.1.1),

$$T_{\mathsf{Alg}}(\mathcal{A}^{(N)}) = \sup_{\Theta^0 \in \mathcal{A}^{(N)}} \inf\{T_0 \in \mathbb{R}^+ : \mathbb{P}_{\Theta^0,T}\{\mathsf{Alg}(X_0^T) = M(\Theta^0)\} \geq 1-\delta \text{ for all } T \geq T_0\}.$$

The following theorem holds for this class of functions and stationary trajectories of (3.1.1). Its proof can be found the appendix of [13, 15].

Theorem 3.7.2. *Let* $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ *be an estimator of* $M(\Theta^0)$. *Then*

$$T_{\mathsf{Alg}}(\mathcal{A}^{(D)}) \geq \frac{\Delta \log p/\Delta - \log B/L}{C + 2\Delta^2 D^2 B}, \tag{3.7.4}$$

where $C = \max_{i\in[p]} \mathbb{E}\{F_i(\mathbb{E}_{\underline{x}(0)|\Theta^0}(\underline{x}(0)); \Theta^0)\}$.

Remark 3.7.2. *The assumption that* F *is Lipschitz is not very restrictive as it is a sufficient condition commonly used to guarantee existence and uniqueness of a solution of the SDE* (3.1.1) *with finite expected energy, [88].*

3.8 Numerical illustration of some extensions

In Theorem 3.3.1 we described a set of conditions under which it is possible to successfully reconstruct the dynamics of a sparse system of linear SDEs. A natural question that arises is: Are these conditions natural enough to hold when Θ^0 describes a system of SDEs related to some real world problem? Or even more generally, given that all models described in this thesis constitute most often nothing but approximations to a real phenomenon we are trying to understand, how good are the models and learning algorithms developed when applied to these? Answering this question is non-trivial. In part because it is also non-trivial to get a clear intuition of what assumptions like Assumption 1 and Assumption 2 of Section 3.3.3 translate to in practice. The same difficulty arises with analogous results on the high-dimensional consistency of the LASSO [110, 116].

In this section we study the performance of Rls when applied to more realistic scenarios. We compare its performance to the performance bounds predicted by our theorems and observe that, despite the potentially assumptions made in them, the bounds capture the right behavior of Rls.

3.8.1 Mass-spring system

Consider a system of p masses in $\mathbb{R}^d$ connected by damped springs that is vibrating under the influence of white-noise. These can be thought of, for example, as points on a vibrating object whose physical structure we are trying to reconstruct from the measured amplitude of vibrations over time on a grid of points at its surface.

Let C^0 be the corresponding adjacency matrix, i.e. $C^0_{ij} = 1$ if and only if masses i and j are connected, and D^0_{ij} be the rest length of the spring (i, j). Assuming unit masses, unit rest lengths and unit elastic coefficients, the dynamics of this system in the presence of external noisy forces can be modeled by the following damped Newton

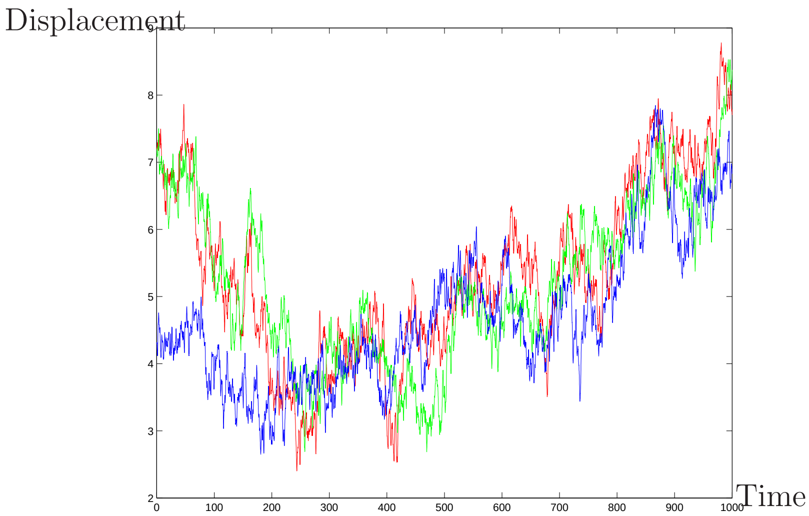

Figure 3.3: Evolution of the horizontal component of the position of three masses in a system with $p = 36$ masses interacting via elastic springs (cf. Fig. 3.4 for the network structure). The time interval is $T = 1000$. All the springs have rest length $D_{ij} = 1$, the damping coefficient is $\gamma = 2$, cf. Eq. (3.8.1), and the noise variance is $\sigma^2 = 0.25$.

equations

$$\mathrm{d}\underline{v}(t) = -\gamma \underline{v}(t)\mathrm{d}t - \nabla U(\underline{q}(t))\,\mathrm{d}t + \sigma\,\mathrm{d}\underline{b}(t), \tag{3.8.1}$$

$$\mathrm{d}\underline{q}(t) = \underline{v}(t)\mathrm{d}t\,, \tag{3.8.2}$$

$$U(\underline{q}) \equiv \frac{1}{2}\sum_{(i,j)} C^0_{ij}(\|q_i - q_j\| - D^0_{ij})^2\,,$$

where $\underline{q}(t) = (q_1(t), \dots, q_p(t))$, $\underline{v}(t) = (v_1(t), \dots, v_p(t))$, and $q_i(t), v_i(t) \in \mathbb{R}^d$ denote the position and velocity of mass i at time t. This system of SDE's can be written in the form (3.1.1) by letting $\underline{x}(t) = [\underline{q}(t), \underline{v}(t)]$ and $\Theta^0 = [C^0, D^0]$. A straightforward calculation shows that the drift $F(\underline{x}(t); \Theta^0)$ can be further written as a linear combination of the following basis of non-linear functions

$$F(\underline{x}(t)) = \Big[\{v_i(t)\}_{i\in[p]}, \{\Delta_{ij}(t)\}_{i,j\in[p]}, \Big\{\frac{\Delta_{ij}(t)}{\|\Delta_{ij}(t)\|}\Big\}_{i,j\in[p]}\Big], \tag{3.8.3}$$

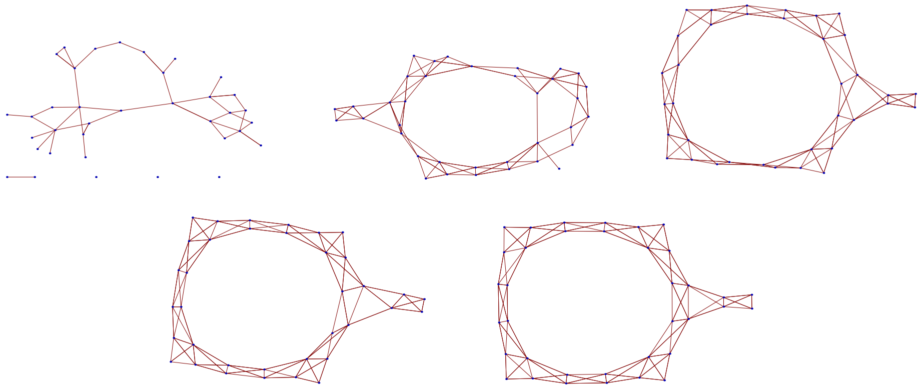

Figure 3.4: From left to right and top to bottom: structures reconstructed using Rlr with observation time $T = 500$, 1500, 2500, 3500 and 4500. For $T = 4500$ exact reconstruction is achieved.

where $\Delta_{ij}(t) = q_i(t) - q_j(t)$ and $[p] = \{1, \dots, p\}$. Hence, the system can be modeled according to (3.1.5). In many situations, only specific properties of the parameters are of interest, for instance one might be interested only in the network structure of the springs.

Figure 3.3 shows the trajectories of three masses in a two-dimensional network of 36 masses and 90 springs evolving according to Eq. (3.8.1) and Eq. (3.8.2). How long does one need to observe these (and the other masses) trajectories in order to learn the structure of the underlying network? Notice that the system being considered is non-linear and hence, a priori, we cannot apply any of our theorems to guarantee that correct reconstruction will be achieved for any T. Figure 3.4 shows the network structure reconstructed using the least-squares algorithm described in Sec. 3.3.2 for increasing observation intervals T. Despite the non-linearities, the inferred structure converges to the true one when T is large enough [7].

To understand the efficiency of the regularized least-squares in learning non-linear SDEs we generated multiple spring-mass networks of sizes $p = 8, 16, 32, 64$

[7] The data was generated by a simulation of Newton's equations of motion using an Euler approximation with discrete time step of size 0.1s

and 128 and studied the minimum number of samples required for successful reconstruction.The spring-mass networks were generated from random regular graphs of vertex degree 4. The data was generated by simulating the dynamics using an Euler approximation with a time step of 0.1s. The noise level, σ, was set to 0.5 and the damping parameter, γ, was set to 0.1.

Figure 3.5–Left shows the probability of success versus the length of the observation time window for systems of different sizes ($p = 8, 16, 32, 64$ and 128) and Figure 3.5–Right shows the minimum number of samples for successful reconstruction of the networks versus their size for different probabilities of success ($\mathrm{P}_{\mathrm{succ}} = 0.1, 0.5$ and 0.9). In both pictures, error bars represent ± one standard deviation. We define a successful reconstruction by an exact recovery of the whole network. Since networks were generated by sampling regular graphs uniformly at random, the probability of full exact reconstruction of the network equals the probability of full exact reconstruction of any node's neighborhood in the network. This fact was used to minimize the number of simulations required to achieve a small fluctuation in our numerical results.

It is interesting to observe that the sample complexity in learning these non-linear system of SDEs also scales logarithmically with p (compare Figure 3.5 with Figure 3.1). A careful look into the proof of our main theorem suggests that as long as the correlation between consecutive samples decays exponentially with time, the same proof would follow. The difficulty in proving a generalization of Theorem 3.3.1 to general non-linear SDEs of the from (3.1.5) stems from the fact that it is hard to know what kind of correlations a general SDE will induce on its trajectory. However, given a sufficiently 'nice' trajectory, the success of the least-square method should not be affected by the fact that we are considering a non-linear basis of functions. In fact, even in this case, the method still consists of minimizing a quadratic function under a norm-1 constrain.

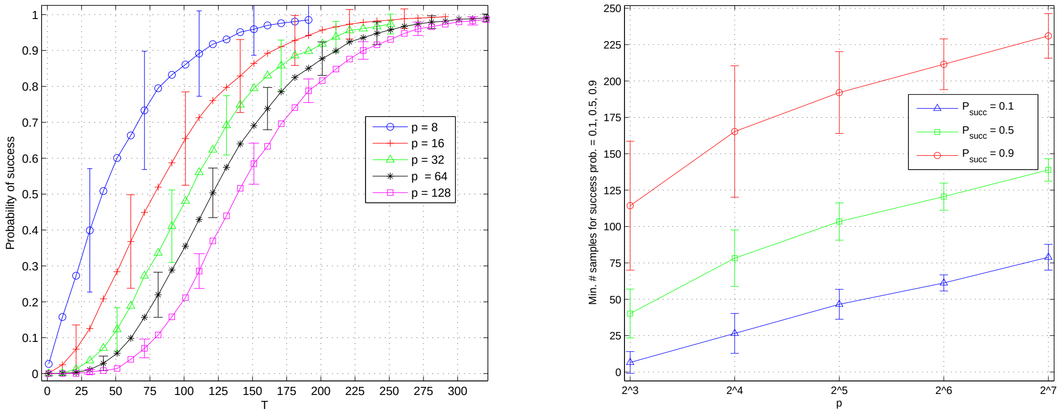

Figure 3.5: (left) Probability of success versus length of observation time window, T, for different network sizes ($p = 8, 16, 32, 64$ and 128). (right) Minimum number of samples required to achieve a probability of reconstruction of $\mathrm{P}_{\mathrm{succ}} = 0.1, 0.5$ and 0.9 versus the size of the network p. All networks where generated from random regular graphs of degree 4 sampled uniformly at random. The dynamics' parameters were set to $\sigma = 0.5$ and $\gamma = 0.1$

3.8.2 Biochemical pathway

As another example, we look at a biochemical pathway describing a general response of a cell to a change in its environment. This change can be, for instance, a lesion on the skin. The lesion causes some cells to generate diffusible ligands (L). These ligands come upon receptors (R) on the cell membrane, which act like antennas. Receptors that have caught a ligand can then be modified (phosphorylated $*$) by enzymes called kinases (K). These modifications enable interactions with other substrates (S) which eventually turn on the genetic program of platelets to move towards the source of the injury. This sequence of events is what is called the chemical pathway and can be thought of as a sequence of chemical reactions describing the interaction between difference species inside and outside the cell. The general pathway in consideration is described in [6] and reproduced bellow for completeness. Below, k_f and k_r describe forward and backward rates of reaction. Expressions inside parenthesis, e.g. (LR^*), represent specific intermediary stages or compounds along the pathway.

$$R + L \underset{k_{r1}}{\overset{k_{f1}}{\rightleftharpoons}} (LR^*), \tag{3.8.4}$$

$$(LR^*) + K \underset{k_{r2}}{\overset{k_{f2}}{\rightleftharpoons}} (LR^*K), \tag{3.8.5}$$

$$(LR^*K) \xrightarrow{k_{f3}} (LR^*) + K^*, \tag{3.8.6}$$

$$K^* + S \underset{k_{r4}}{\overset{k_{f4}}{\rightleftharpoons}} (K^*S), \tag{3.8.7}$$

$$(K^*S) \xrightarrow{k_{f5}} K^* + S^*. \tag{3.8.8}$$

The network of interactions comprises nine biochemical species and can be described by a set of SDE's copied below from [6].

$$\begin{aligned}
\mathrm{d}x_1(t) &= (-k_{f1}x_1(t)x_2(t) + k_{r1}x_3(t))\mathrm{d}t + \mathrm{d}b_1(t)\\
\mathrm{d}x_2(t) &= (-k_{f1}x_1(t)x_2(t) + k_{r1}x_3(t))\mathrm{d}t + \mathrm{d}b_2(t)\\
\mathrm{d}x_3(t) &= (+k_{f1}x_1(t)x_2(t) - k_{f2}x_3(t) - k_{f2}x_3(t)x_5(t) + (k_{r2} + k_{f3})x_4(t))\mathrm{d}t + \mathrm{d}b_3(t)\\
\mathrm{d}x_4(t) &= (+k_{f2}x_3(t)x_5(t) - (k_{r2} + k_{f3})x_4(t))\mathrm{d}t + \mathrm{d}b_4(t)\\
\mathrm{d}x_5(t) &= (-k_{f2}x_3(t)x_5(t) + k_{r2}x_4(t))\mathrm{d}t + \mathrm{d}b_5(t)\\
\mathrm{d}x_6(t) &= (-k_sx_6(t) + k_{f3}x_4(t) - k_{f4}x_6(t)x_7(t) + (k_{r4} + k_{f5})x_8(t))\mathrm{d}t + \mathrm{d}b_6(t)\\
\mathrm{d}x_7(t) &= (-k_{f4}x_6(t)x_7(t) + k_{r4}x_8(t))\mathrm{d}t + \mathrm{d}b_7(t)\\
\mathrm{d}x_8(t) &= (+k_{f4}x_6(t)x_7(t) - (k_{r4} + k_{f5})x_8(t))\mathrm{d}t + \mathrm{d}b_8(t)\\
\mathrm{d}x_9(t) &= (-k_sx_9(t) + k_{f5}x_8(t))\mathrm{d}t + \mathrm{d}b_9(t)
\end{aligned} \tag{3.8.9}$$

In (3.8.9) we assume the following correspondence between the concentration of each compound and the variables $x_i(t), i \in [9]$: $x_1 \leftrightarrow R$, $x_2 \leftrightarrow L$, $x_3 \leftrightarrow (LR^*)$, $x_4 \leftrightarrow K$, $x_5 \leftrightarrow (LR^*K)$, $x_6 \leftrightarrow K^*$, $x_7 \leftrightarrow S$, $x_8 \leftrightarrow (K^8 * S)$, $x_9 \leftrightarrow (S^*)$.

Our objective is to attempt to learn this network as a nonlinear SDE of the form (3.1.5) with a basis of functions consisting of polynomials of up to order two, i.e., all the functions of the form $x_i^{\alpha_1}x_j^{\alpha_2}$ with $\alpha_1, \alpha_2 \in \{0, 1\}$. The adjacency matrix to be learned is $\Theta^0 \in \mathbb{R}^{9\times 46}$ which translates into approximately 400 parameters to be estimated. We simulate the system in (3.8.9) using the Euler-Maruyama method to obtain sample trajectories. We repeat the experiment 10 times, sampling independent traces of the SDE's trajectory. We obtain the following curves. Error bars represent plus-minus one (empirical) standard deviation.

Figure 3.6 shows the true and false positive rates in reconstructing the support of matrix Θ^0.

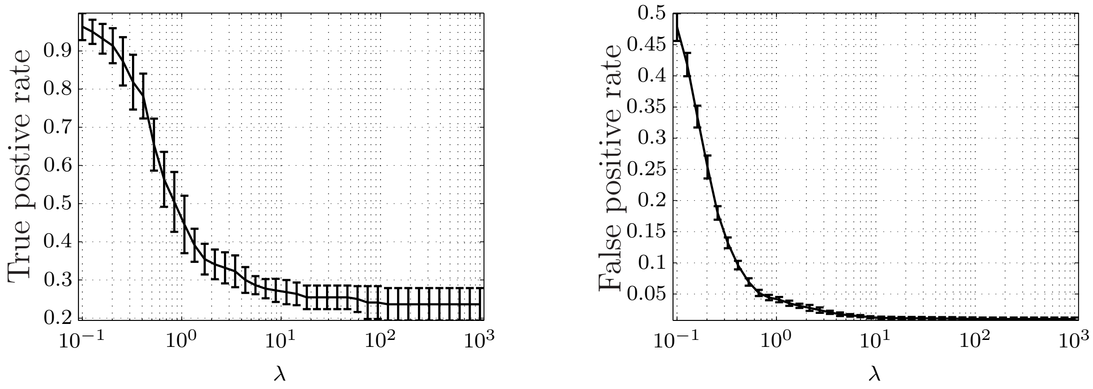

Figure 3.6: (left) True positive rate (right) false positive rate vs. λ for the duration of observation $T = 1200$.

Figure 3.7 shows the corresponding true positive rate versus false positive rate curve (ROC curve) for different values of observation length T. The area under the ROC curve is shown on the right plot in this figure. This value is an approximation of the probability of choosing a random existing edge over a random non-existing negative.

Finally, Fig. 3.8 shows the normalized root mean squared error (RMSE) defined as

$$NRMSE \equiv \frac{\|\Theta^0 - \hat{\Theta}\|_2}{\|\Theta^0\|_2}. \tag{3.8.10}$$

The left plot shows the NRMSE versus λ. There exist an optimal value for the parameter λ that minimized NRMSE. The right plot depicts NRMSE at the optimum value of the parameter λ versus the length of the observation interval. NRMSE

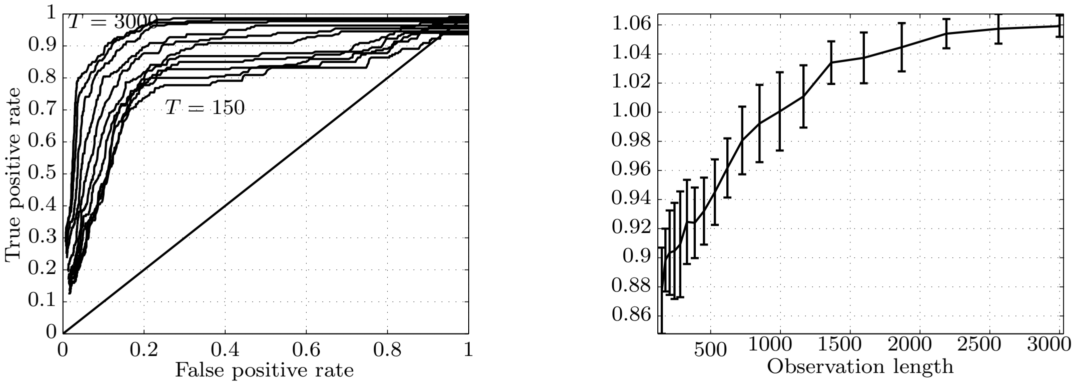

Figure 3.7: (left) ROC curves for different values of T. (right) Area under the ROC curve vs. T.

drops below one as the observation length increases suggesting that the algorithm is succeeding in reconstructing the signed support of the adjacency matrix Θ^0.

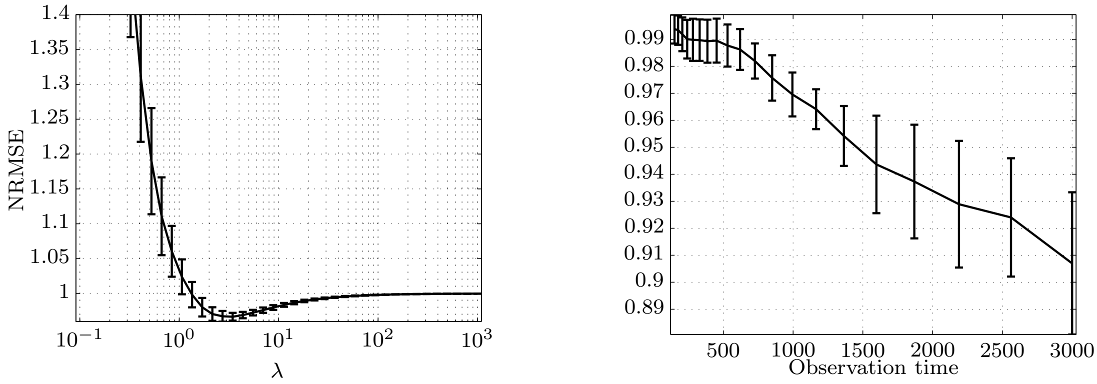

Figure 3.8: (left) NRMSE vs. λ; $T = 1200$. (right) NRMSE vs. T for the optimum value of λ.

Appendix A

Learning the Ising model

A.1 Simple Thresholding

In the following we let $C_{ij} \equiv \mathbb{E}_{G,\theta}\{X_i X_j\}$ where expectation is taken with respect to the Ising model (2.1.1).

Proof. (Theorem 2.3.1) If G is a tree then $C_{ij} = \tanh\theta$ for all $(ij) \in E$ and $C_{ij} \leq \tanh^2\theta$ for all $(ij) \notin E$. To see this notice that only paths that connect i to j contribute to C_{ij} and given that G is a tree there is only one such path and its length is exactly 1 if $(i,j) \in E$ and at least 2 when $(i,j) \notin E$. The probability that $\mathsf{Thr}(\tau)$ fails is

$$1 - \mathrm{P}_{\mathrm{succ}} = \mathbb{P}_{n,G,\theta}\{\widehat{C}_{ij} < \tau \text{ for some } (i,j) \in E \text{ or } \widehat{C}_{ij} \geq \tau \text{ for some } (i,j) \notin E\} \tag{A.1.1}$$

Let $\tau = (\tanh\theta + \tanh^2\theta)/2$. Applying Azuma-Hoeffding inequality to $\widehat{C}_{ij} = \frac{1}{n}\sum_{\ell=1}^{n} x_i^{(\ell)} x_j^{(\ell)}$ we have that if $(i,j) \in E$ then,

$$\mathbb{P}_{n,G,\theta}(\widehat{C}_{ij} < \tau) = \mathbb{P}_{n,G,\theta}\left(\sum_{\ell=1}^{n}(x_i^{(\ell)} x_j^{(\ell)} - C_{ij}) < n(\tau - \tanh\theta)\right) \leq e^{-\frac{1}{32}n(\tanh\theta - \tanh^2\theta)^2} \tag{A.1.2}$$

and if $(i,j) \notin E$ then similarly,

$$\mathbb{P}_{n,G,\theta}(\widehat{C}_{ij} \geq \tau) = \mathbb{P}_{n,G,\theta}\left(\sum_{\ell=1}^{n}(x_i^{(\ell)}x_j^{(\ell)} - C_{ij}) \geq n(\tau - \tanh^2\theta)\right) \leq e^{-\frac{1}{32}n(\tanh\theta - \tanh^2\theta)^2}. \tag{A.1.3}$$

Applying union bound over the two possibilities, $(i,j) \in E$ or $(i,j) \notin E$, and over the edges ($|E| < p^2/2$), we can bound $\mathrm{P}_{\mathrm{succ}}$ by

$$\mathrm{P}_{\mathrm{succ}} \geq 1 - p^2\, e^{-\frac{1}{32}n(\tanh\theta - \tanh^2\theta)^2}\,. \tag{A.1.4}$$

Imposing the right hand side to be larger than δ proves our result. □

Proof. (Theorem 2.3.2) We will prove that, for $\theta < \operatorname{arctanh}(1/(2\Delta))$, $C_{ij} \geq \tanh\theta$ for all $(i,j) \in E$ and $C_{ij} \leq 1/(2\Delta)$ for all $(ij) \notin E$. In particular $C_{ij} < C_{kl}$ for all $(i,j) \notin E$ and all $(k,l) \in E$. The theorem follows from this fact via union bound and Azuma-Hoeffding inequality as in the proof of Theorem 2.3.1.

The bound $C_{ij} \geq \tanh\theta$ for $(ij) \in E$ is a direct consequence of Griffiths inequality [55] : compare the expectation of x_ix_j in G with the same expectation in the graph that only includes edge (i,j).

The second bound is derived using the technique of [40], i.e., bound C_{ij} by the generating function for self-avoiding walks on the graphs from i to j. More precisely, assume $l = \operatorname{dist}(i,j)$ and denote by $N_{ij}(k)$ the number of self avoiding walks of length k between i and j on G. Then [40] proves that

$$C_{ij} \leq \sum_{k=l}^{\infty}(\tanh\theta)^k N_{ij}(k) \leq \sum_{n=l}^{\infty}\Delta^{k-1}(\tanh\theta)^k \leq \frac{\Delta^{l-1}(\tanh\theta)^l}{1-\Delta\tanh\theta} \leq \frac{\Delta(\tanh\theta)^2}{1-\Delta\tanh\theta} \tag{A.1.5}$$

If $\theta < \operatorname{arctanh}(1/(2\Delta))$ the above implies $C_{ij} \leq 1/(2\Delta)$ which is our claim. □

Proof. (Theorem 2.3.3) The theorem is proved by constructing G as follows: sample a uniformly random regular graph of degree Δ over the $p-2$ vertices $\{1,2,\dots,p-2\} \equiv [p-2]$. Add an extra edge between nodes $p-1$ and p. The resulting graph is not connected. We claim that for $\theta > K/\Delta$ and with probability converging to 1 as $p \to \infty$, there exist $i, j \in [p-2]$ such that $(i,j) \notin E$ and $C_{ij} > C_{p-1,p}$. As a consequence, thresholding fails.

Obviously $C_{p-1,p} = \tanh\theta$. Choose $i \in [p-2]$ uniformly at random, and j a node at a fixed distance t from i. We can compute C_{ij} as $p \to \infty$ using the same local weak convergence result as in the proof of Lemma **??**. Namely, C_{ij} converges to the correlation between the root and a leaf node in the tree Ising model (A.5.1). In particular one can show, [83], that

$$\lim_{p\to\infty} C_{ij} \geq m(\theta)^2 \, , \tag{A.1.6}$$

where $m(\theta) = \tanh(\Delta h^0/(\Delta - 1))$ and h^0 is the unique positive solution of Eq. (A.5.2).

The proof is completed by showing that $\tanh\theta < m(\theta)^2$ for all $\theta > K/\Delta$. □

A.2 Incoherence is a necessary condition: Proof of Lemma 2.6.1

This proof follows closely the proof of Proposition 1 in [94]. For a matter of clarity of exposition we include all the steps, even if these do not differ from the exposition done in [94].

Here we show that, under the assumptions of the Lemma on the incoherence condition, on $\sigma_{\min}(Q^0_{SS})$ and on λ, the probability that $\mathsf{Rlr}(\lambda)$ returns a $\hat{\theta}$ satisfying $\underline{\hat{\theta}} = (\underline{\hat{\theta}}_S, \underline{\hat{\theta}}_{S^C}) = (\underline{\hat{\theta}}_S, 0)$ with $\underline{\hat{\theta}}_S > 0$ is upper bounded as in Eq. (2.6.5). More specifically, we will show that this $\underline{\hat{\theta}}$ will not satisfy the stationarity condition $\nabla \mathcal{L}^n(\underline{\hat{\theta}}) + \lambda \hat{z} = 0$ with high probability for any sub-gradient $\hat{z}$ of the function $\|\underline{\theta}\|_1$ at $\underline{\hat{\theta}}$.

To simplify notation we omit $\{\underline{x}^{(\ell)}\}$ in all the expressions involving and derived from $\mathcal{L}^n$.

Assume the event $\nabla \mathcal{L}^n(\underline{\hat{\theta}}) + \lambda \hat{z} = 0$ holds for some $\underline{\hat{\theta}}$ as specified above. An application of the mean value theorem yields

$$\nabla^2 \mathcal{L}^n(\underline{\theta}^0)[\underline{\hat{\theta}} - \underline{\theta}^0] = W^n - \lambda \hat{z} - R^n \, , \tag{A.2.1}$$

where we recall $W^n = -\nabla\mathcal{L}^n(\underline{\theta}^0)$ and $[R^n]_j = [\nabla^2\mathcal{L}^n(\bar{\underline{\theta}}^{(j)}) - \nabla^2\mathcal{L}^n(\underline{\theta}^0)]^*_j(\hat{\underline{\theta}} - \underline{\theta}^0)$ with $\bar{\underline{\theta}}^{(j)}$ a point in the line from $\hat{\underline{\theta}}$ to $\underline{\theta}^0$. Notice that by definition $\nabla^2\mathcal{L}^n(\underline{\theta}^0) = Q^{n0} = Q^n(\underline{\theta}^0)$. To simplify notation we omit the ^{0} in all Q^{n0}. All Q^n in this proof are thus evaluated at $\underline{\theta}^0$.

Breaking this expression into its S and S^C components and since $\hat{\underline{\theta}}_{S^C} = \underline{\theta}^0_{S^C} = 0$ we can write

$$Q^n_{S^CS}(\hat{\underline{\theta}}_S - \underline{\theta}^0_S) = W^n_{S^C} - \lambda\hat{z}_{S^C} + R^n_{S^C}, \tag{A.2.2}$$

$$Q^n_{SS}(\hat{\underline{\theta}}_S - \underline{\theta}^0_S) = W^n_S - \lambda\hat{z}_S + R^n_S. \tag{A.2.3}$$

Eliminating $\hat{\underline{\theta}}_S - \underline{\theta}^0_S$ from the two expressions we obtain

$$[W^n_{S^C} - R^n_{S^C}] - Q^n_{S^CS}(Q^n_{SS})^{-1}[W^n_S - R^n_S] + \lambda Q^n_{S^CS}(Q^n_{SS})^{-1}\hat{z}_S = \lambda\hat{z}_{S^C}\,. \tag{A.2.4}$$

Now notice that $Q^n_{S^CS}(Q^n_{SS})^{-1} = T_1 + T_2 + T_3 + T_4$ where

$$T_1 = Q^0_{S^CS}[(Q^n_{SS})^{-1} - (Q^0_{SS})^{-1}]\,, \qquad T_2 = [Q^n_{S^CS} - Q^0_{S^CS}]{Q^0_{SS}}^{-1}\,,$$
$$T_3 = [Q^n_{S^CS} - Q^0_{S^CS}][(Q^n_{SS})^{-1} - (Q^0_{SS})^{-1}]\,, \qquad T_4 = Q^0_{S^CS}{Q^0_{SS}}^{-1}\,.$$

Recalling that $\hat{z}_S = \mathbb{1}$ and using the above decomposition we can lower bound the absolute value of the indexed-i component of $\hat{z}_{S^C}$ by

$$\begin{aligned} |\hat{z}_i| \geq \|[Q^0_{S^CS}{Q^0_{SS}}^{-1}\hat{z}_S]_i\|_\infty - \|T_{1,i}\|_1 - \|T_{2,i}\|_1 - \|T_{3,i}\|_1 \\ -\Big|\frac{W^n_i}{\lambda}\Big| - \Big|\frac{R^n_i}{\lambda}\Big| - \|[Q^n_{S^CS}(Q^n_{SS})^{-1}]_i\|\left(\Big\|\frac{W^n_S}{\lambda}\Big\|_\infty + \Big\|\frac{R^n_S}{\lambda}\Big\|_\infty\right). \end{aligned} \tag{A.2.5}$$

We now assume that the samples $\{\underline{x}^{(\ell)}\}$ are such that the following event holds (notice that $i \in S^C$),

$$\mathcal{E}_i \equiv \left\{\|Q^n_{S\cup\{i\}\;S} - Q^0_{S\cup\{i\}\;S}\|_\infty < \xi_A, \Big\|\frac{W^n_{S\cup\{i\}}}{\lambda}\Big\|_\infty < \xi_B\right\}, \tag{A.2.6}$$

where $\xi_A \equiv C^2_{\min}\epsilon/(8\Delta)$ and $\xi_B \equiv C_{\min}\epsilon/(16\sqrt{\Delta})$.

From relations (2.6.1) to (2.6.4) in Section 2.6.1 we know that $\mathbb{E}_{G,\theta}(Q^n) = Q^0$, $\mathbb{E}_{G,\theta}(W^n) = 0$ and that both $Q^n - Q^0$ and W^n are sums i.i.d. random variables bounded by 2. From this, a simple application of Azuma-Hoeffding inequality yields [1].

$$\mathbb{P}_{G,\theta,n}(|Q^n_{ij} - Q^0_{ij}| > \delta) \leq 2e^{-\frac{\delta^2 n}{8}}, \tag{A.2.7}$$

$$\mathbb{P}_{G,\theta,n}(|W^n_{ij}| > \delta) \leq 2e^{-\frac{\delta^2 n}{8}}, \tag{A.2.8}$$

for all i and j. Applying union bound we conclude that the event $\mathcal{E}_i$ holds with probability at least

$$1 - 2\Delta(\Delta+1)e^{-\frac{n\xi_A^2}{8}} - 2(\Delta+1)e^{-\frac{n\lambda^2\xi_B^2}{8}} \geq 1 - 4\Delta^2 e^{-n\delta_A^2} - 4\Delta e^{-n\lambda^2\delta_B}, \tag{A.2.9}$$

where $\delta_A = C_{\min}^2\epsilon/(32\Delta)$ and $\delta_B = C_{\min}\epsilon/(64\sqrt{\Delta})$.

If the event $\mathcal{E}_i$ holds then $\sigma_{\min}(Q^n_{SS}) > \sigma_{\min}(Q^0_{SS}) - C_{\min}/2 > C_{\min}/2$. Since $\|[Q^n_{S^CS}(Q^n_{SS})^{-1}]_i\|_\infty \leq \|{Q^n_{SS}}^{-1}\|_2\|Q^n_{Si}\|_2$ and $|Q^n_{ji}| \leq 1 \forall i,j$ we can write $\|[Q^n_{S^CS}(Q^n_{SS})^{-1}]_i\|_\infty \leq 2\sqrt{\Delta}/C_{\min}$ and simplify our lower bound to

$$\begin{aligned} |\hat{z}_i| \geq \|[Q^0_{S^CS}{Q^0_{SS}}^{-1}\hat{z}_S]_i\|_\infty - \|T_{1,i}\|_1 - \|T_{2,i}\|_1 - \|T_{3,i}\|_1 \\ -\Big|\frac{W^n_i}{\lambda}\Big| - \Big|\frac{R^n_i}{\lambda}\Big| - \frac{2\sqrt{\Delta}}{C_{\min}}\left(\Big\|\frac{W^n_S}{\lambda}\Big\|_\infty + \Big\|\frac{R^n_S}{\lambda}\Big\|_\infty\right). \end{aligned} \tag{A.2.10}$$

The proof is completed by showing that the event $\mathcal{E}_i$ and the assumptions of the theorem imply that each of last 7 terms in this expression is smaller than $\epsilon/8$. Since $|[Q^0_{S^CS}(Q^0_{SS})^{-1}]^*_i\hat{z}_S| \geq 1+\epsilon$ by assumption, this implies $|\hat{z}_i| \geq 1+\epsilon/8 > 1$ which cannot be true since any sub-gradient of the 1-norm has components of magnitude at most 1.

Taking into account that $\sigma_{\min}(Q^0_{SS}) \leq \max_{ij} Q^0_{ij} \leq 1$ and that $\Delta > 1$, the last condition on $\mathcal{E}_i$ immediately bounds all terms involving W^n by $\epsilon/8$. Some straightforward

[1] For full details see the proof of Lemma 2 and the discussion following Lemma 6 in [94]

manipulations imply (see Lemma 7 from [94] for a similar computation)

$$\|T_{1,i}\|_1 \leq \frac{\Delta}{C_{\min}^2}\|Q_{SS}^n - Q_{SS}^0\|_\infty\,, \qquad \|T_{2,i}\|_1 \leq \frac{\sqrt{\Delta}}{C_{\min}}\|[Q_{S^CS}^n - Q_{S^CS}^0]_i\|_\infty\,,$$
$$\|T_{3,i}\|_1 \leq \frac{2\Delta}{C_{\min}^2}\|Q_{SS}^n - Q_{SS}^0\|_\infty\|[Q_{S^CS}^n - Q_{S^CS}^0]_i\|_\infty\,,$$

and thus, again making use of the fact that $\sigma_{\min}(Q_{SS}^0) \leq 1$, all will be bounded by $\epsilon/8$ when $\mathcal{E}_i$ holds. The final step of the proof consists in showing that if $\mathcal{E}_i$ holds and λ satisfies the condition given in the Lemma enunciation then the terms involving R^n will also be bounded above by $\epsilon/8$. The details of this calculation can be found in [16].

A.3 Proof of Theorem 2.3.6: $\theta\Delta \leq 3/10$

Before we prove the first part of Theorem 2.3.6 we need an auxiliary step to bound the covariance $C_{ij} \equiv \mathbb{E}_{G,\theta}(X_iX_j)$ between any two variables $i, j \in V$ in our model (2.1.1). This bound is derived using the technique of [40], i.e., bound C_{ij} by the generating function for self-avoiding walks on the graphs from i to j.

Lemma A.3.1. *Assume $l = \mathrm{dist}(i,j)$ is the distance between node i and j in G and denote by $N_{ij}(k)$ the number of self avoiding walks of length k between i and j on G. Then [40] proves that*

$$C_{ij} \leq \sum_{k=l}^{\infty}(\tanh\theta)^k N_{ij}(k) \leq \sum_{n=l}^{\infty}\Delta^{k-1}(\tanh\theta)^k \tag{A.3.1}$$
$$\leq \frac{\Delta^{l-1}(\tanh\theta)^l}{1-\Delta\tanh\theta} \leq \frac{\Delta(\tanh\theta)^2}{1-\Delta\tanh\theta}. \tag{A.3.2}$$

The proof of the first part of our main theorem consists in showing that, in the regime when $\theta\Delta \leq 3/10$, all conditions of Theorem 1 in [94] (denoted there by A1 and A2) hold. In the process of verifying these conditions, we get explicit bounds for several constants that remain unspecified in [94]. With this, we are able obtain our bound on the sample complexity.

In this proof all the functions are assumed to be evaluated at the true value of the parameters $\underline{\theta}^0$. Hence, throughout this proof, and to simplify notation, we denote $Q^{n0} = Q^n(\underline{\theta}^0)$ simply by Q^n and $W^{n0} = W^n(\underline{\theta}^0)$ by W^n. In what follows, $C_{\min}$ is a lower bound for $\sigma_{\min}(Q^0_{SS})$ and $D_{\max}$ [2] is an upper bound for $\sigma_{\max}(\mathbb{E}_\theta(\underline{X}_S \underline{X}^*_S))$. We define $1-\alpha \equiv \|Q^0_{S^CS}(Q^0_{SS})^{-1}\|_\infty$ and let $\theta_{\min}$ denote the minimum absolute value of the components of $\underline{\theta}^0$. In our case we have $\theta_{\min} = \theta$. Throughout this proof we also have $\hat{C}_{\min} \equiv \sigma_{\min}(Q^n_{SS})$, $\hat{D}_{\max} \equiv \sigma_{\max}\left(\frac{1}{n}\sum_{l=1}^n x^{(l)}_S {x^{(l)}}^*_S\right)$ and $1-\hat{\alpha} \equiv \|Q^n_{S^CS}(Q^n_{SS})^{-1}\|_\infty$.

We begin by noting that Theorem 1 in [94] can be rewritten in the following form,

Theorem A.3.1. *If* $0 < \alpha < 1$ *and the events* $\mathcal{E}$ *and* $\mathcal{A}$ *hold true, then* Rlr *will not fail.*

The event $\mathcal{E}$ consists of the following conditions [3],

$$\begin{array}{lll} \text{In Lemma 5,[94]:} & & \|Q^n_{SS} - Q^0_{SS}\|_2 < \dfrac{C_{\min}}{2}\,, \qquad \text{(A.3.3)} \\ \text{In Lemma 6,[94]:} & \text{for T1} & \sigma_{\min}({Q^n}_{SS}) \geq \dfrac{C_{\min}}{2}\,, \qquad \text{(A.3.4)} \\ & \text{for T1} & \|Q^n_{SS} - Q^0_{SS}\|_\infty \leq \dfrac{1}{12}\dfrac{\alpha}{1-\alpha}\dfrac{C_{\min}}{\sqrt{\Delta}}\,, \qquad \text{(A.3.5)} \\ & \text{for T2} & \|Q^n_{S^CS} - Q^0_{S^CS}\|_\infty \leq \dfrac{\alpha}{6}\dfrac{C_{\min}}{\sqrt{\Delta}}\,, \qquad \text{(A.3.6)} \\ & \text{for T3} & \|Q^n_{S^CS} - Q^0_{S^CS}\|_\infty \leq \sqrt{\dfrac{\alpha}{6}}\,, \qquad \text{(A.3.7)} \\ \text{In Lemma 7,[94]:} & & \sigma_{\min}({Q^n}_{SS}) \geq \dfrac{C_{\min}}{2} \text{ and} \qquad \text{(A.3.8)} \\ & & \|Q^n_{SS} - Q^0_{SS}\|_2 \leq \sqrt{\dfrac{\alpha}{6}}\dfrac{C^2_{\min}}{2\sqrt{\Delta}}\,, \qquad \text{(A.3.9)} \\ \text{In Proposition 1,[94]:} & & \dfrac{\|W^n\|_\infty}{\lambda} < \dfrac{\hat{\alpha}}{4(2-\hat{\alpha})}. \qquad \text{(A.3.10)} \end{array}$$

[2] It is easy to prove that $C_{\min} \leq D_{\max}$

[3] These conditions are the conditions required for Theorem 1 in [94] to be applicable and are labeled by the names of the intermediary results in [94] required to prove Theorem 1 in [94]

The event $\mathcal{A}$ consists of the following bounds on the value of λ in $\mathsf{Rlr}(\lambda)$,

$$\text{In Lemma 3,[94]:} \qquad \lambda\Delta \le \frac{\hat{C}^2_{\min}}{10\hat{D}_{\max}}\,, \tag{A.3.11}$$

$$\text{In Proposition 3,[94]:} \qquad \frac{5}{\hat{C}_{\min}}\lambda\sqrt{\Delta} \le \frac{\theta_{\min}}{2}\,. \tag{A.3.12}$$

Ravikumar et al. [94] show that the condition from Lemma 5 together with the definition of $C_{\min}$ and $D_{\max}$ implies that $\hat{C}_{\min} \ge C_{\min}/2$ and $\hat{D}_{\max} \le 2D_{\max}$. In addition, the proof of Lemma 6 in [94] shows that, if (A.3.4) to (A.3.7) hold, then, without loss of generality, we can assume $\hat{\alpha} = \alpha/2$. This allows us to rewrite the right hand side of all the above inequalities with constants only and no random variables (remember that $\hat{\alpha}$, $\hat{D}_{\max}$ and $\hat{C}_{\min}$ are random variables). We call the new events involving only constants on the right hand side by $\mathcal{E}' \subseteq \mathcal{E}$ and $\mathcal{A}' \subseteq \mathcal{A}$ respectively.

Having the definition of $\mathcal{E}'$ and $\mathcal{A}'$ in mind, Theorem 1 in [94] can be rewritten in the following form,

Theorem A.3.2. *If* $1-\alpha < 1$ *and the events* $\mathcal{A}'$ *and* $\mathcal{E}'$ *hold true, then* Rlr *will not fail.*

The event $\mathcal{E}'$ consists of deviations of random vectors and matrices, under different norms, to their corresponding expected values. A straightforward application of Azuma's inequality yields the following upper bound on the probability of these assumptions not occurring together [4],

$$\begin{aligned}\mathbb{P}_{n,G,\theta}(\mathcal{E}'^c) \le\ & 2e^{-n\frac{1}{32\Delta^2}(d^{(2)}_{SS})^2+2\log\Delta} + 2e^{-n\frac{1}{32\Delta^2}(d^{(\infty)}_{SS})^2+2\log\Delta} \\ & + 2e^{-n\frac{1}{32\Delta^2}(d^{(\infty)}_{S^CS})^2+\log\Delta+\log p-\Delta} + 2e^{-n\frac{\lambda^2}{2^7}(\frac{\alpha}{4-\alpha})^2+\log p}\,,\end{aligned} \tag{A.3.13}$$

[4] The first three terms are for the conditions involving matrix Q^n and the fourth with the event dealing with matrix W^n

where

$$d_{SS}^{(2)} = \min\Big\{\frac{C_{\min}}{2}, \sqrt{\frac{\alpha}{6}}\frac{C_{\min}^2}{2\sqrt{\Delta}}\Big\}, \quad \text{(A.3.14)}$$

$$d_{SS}^{(\infty)} = \frac{1}{12}\frac{\alpha}{1-\alpha}\frac{C_{\min}}{\sqrt{\Delta}}, \quad \text{(A.3.15)}$$

$$d_{S^CS}^{(\infty)} = \min\Big\{\frac{\alpha}{6}\frac{C_{\min}}{\sqrt{\Delta}}, \sqrt{\frac{\alpha}{6}}\Big\}. \quad \text{(A.3.16)}$$

We want this probability to be upper bounded by δ/p. If this is the case, an application of union bound allows to conclude that Rlr correctly recovers the neighborhood of every node, and hence the whole graph G, with probability greater then $1-\delta$.

In the regime where $\theta\Delta \leq 3/10$, we now compute explicit bounds for $C_{\min}$, $D_{\max}$ and α. Replacing them in (A.3.13) we simplify expression (A.3.13). Finally, from this expression, we prove that the λ chosen in our main theorem suffices to obtain the sample complexity stated.

In what follows we let $K_1 = 3/10$. Also, recall that $\tanh x \leq x$ for all $x \geq 0$. First notice that by (2.6.2) we have $C_{\min} = \sigma_{\min}\{\mathbb{E}_{G,\theta}((1-\tanh^2\theta M)\underline{X}_S\underline{X}_S^*)\}$ where $M = \sum_{t\in\partial r} X_t$. Since $\theta M \leq \theta\Delta \leq K_1$ we have, $C_{\min} \geq (1-K_1^2)\sigma_{\min}(\mathbb{E}_{G,\theta}\{\underline{X}_S\underline{X}_S^*\})$. Now write $\mathbb{E}_{G,\theta}\{\underline{X}_S\underline{X}_S^*\} \equiv I + T$ and notice that by (A.3.2) in Lemma A.3.1, T is a symmetric matrix whose entries are non-negative and smaller than $\tanh\theta/(1-\Delta\tanh\theta)$. Since $\sigma_{\min}(\mathbb{E}_{G,\theta}\{\underline{X}_S\underline{X}_S^*\}) = 1 - v^*(-T)v$ for some unit norm vector v and since, by Cauchy–Schwarz inequality, we have $v^*(-T)v \leq \|v\|_1^2 \max_{ij}|Q_{ij}| \leq \Delta\tanh\theta/(1-\Delta\tanh\theta) \leq K_1/(1-K_1)$, it follows that $\sigma_{\min}(\mathbb{E}_{G,\theta}\{\underline{X}_S\underline{X}_S^*\}) \geq (1-2K_1)/(1-K_1)$. Consequently, $C_{\min} \geq (1+K_1)(1-2K_1) = 13/25$. Again using the bound (A.3.2), we can write $D_{\max} \leq 1+\Delta\tanh\theta/(1-\Delta\tanh\theta) \leq (1-K_1)^{-1} = 10/7$. Finally, we bound $\|Q_{S^CS}^0(Q_{SS}^0)^{-1}\|_\infty$.

Before proceeding however, we need the follow technical Lemma.

Lemma A.3.2. *If $\theta > 0$, $i \in S$ and $j \in S^C$ then,*

$$Q_{ij}^0 = \mathbb{E}_{G,\theta}\{(1-\tanh^2(\theta M))X_iX_j\} \leq \mathbb{E}_{G,\theta}\{1-\tanh^2(\theta M)\}\mathbb{E}_{G,\theta}\{X_iX_j\}. \quad \text{(A.3.17)}$$

Proof. Start by writing,

$$\mathbb{E}\{\tanh^2(\theta M)X_iX_j\} = \mathbb{E}\{\tanh^2(\theta M)|X_iX_j = 1\}\mathbb{P}\{X_iX_j = 1\} \tag{A.3.18}$$

$$- \mathbb{E}\{\tanh^2(\theta M)|X_iX_j = -1\}\mathbb{P}\{X_iX_j = -1\}. \tag{A.3.19}$$

Given that $\theta > 0$, two applications of FKG inequality allow us to conclude that,

$$\mathbb{E}\{\tanh^2(\theta M)|X_iX_j = 1\} = \mathbb{E}\{\tanh^2(\theta M)|X_i = 1, X_j = 1\} \geq \mathbb{E}\{\tanh^2(\theta M)\} \tag{A.3.20}$$

$$\mathbb{E}\{\tanh^2(\theta M)|X_iX_j = -1\} = \mathbb{E}\{\tanh^2(\theta M)|X_i = 1, X_j = -1\} \leq \mathbb{E}\{\tanh^2(\theta M)\}. \tag{A.3.21}$$

Making use of these two inequalities we obtain,

$$\mathbb{E}_{G,\theta}\{\tanh^2(\theta M)X_iX_j\} \geq \mathbb{E}_{G,\theta}\{\tanh^2(\theta M)\}\mathbb{E}_{G,\theta}\{X_iX_j\}, \tag{A.3.22}$$

and the Lemma follows. □

From Lemma A.3.2 and using the bound (A.3.2) we have that, for $i \in S$ and $j \in S^C$, $Q^0_{ij} \leq \tanh\theta/(1 - \Delta\tanh\theta)$. We can now write,

$$\|Q^0_{S^CS}(Q^0_{SS})^{-1}\|_\infty = \max_{j \in S^C} \|(Q^0_{SS})^{-1}Q^0_{S,j}\|_1 \leq \|(Q^0_{SS})^{-1}\|_1\|Q^0_{S,j}\|_1 \tag{A.3.23}$$

$$\leq \Delta\|(Q^0_{SS})^{-1}\|_2\|Q^0_{S,j}\|_\infty \leq \Delta C^{-1}_{\min}\|Q^0_{S,j}\|_\infty \leq \frac{K_1}{(1 - K_1^2)(1 - 2K_1)} = \frac{75}{91} < 1. \tag{A.3.24}$$

Having a lower bound for $C_{\min}$, $D_{\max}$ and $1 - \alpha$ we now proceed to simplify (A.3.13). First notice that, since the lower bound on $C_{\min}$ is smaller than 1 and since $0 < \alpha < 1$, we can write,

$$\mathbb{P}_{n,G,\theta}(\mathcal{E}'^c) \leq 6e^{-n\frac{\alpha^2C^2_{\min}}{32\times 24^2\Delta^3}+2\log p} + 2e^{-n\frac{\alpha^2\lambda^2}{2^1 1}+2\log p}, \tag{A.3.25}$$

where we also made use of $\log\Delta \leq \log p$.

Bounding the first term by $(3\delta)/(4p)$ and the second term by $\delta/(4p)$ is enough to guarantee that the probability of full reconstruction of G is greater then $1-\delta$. Taking into account the values of the bounds for α and $C_{\min}$, we conclude that

$$n > 32 \times 24^2 \Delta^3 \alpha^{-2} C_{\min}^{-2} \log(8p^3/\delta), \tag{A.3.26}$$

suffices for the bound on the first term to hold. For the bound on the second term to hold, it suffices that

$$n\lambda^2 > 2^1 1 \alpha^{-2} \log(8p^3/\delta). \tag{A.3.27}$$

Now notice that event $\mathcal{A}'$ imposes the following two conditions on λ,

$$\lambda \leq \frac{C_{\min}^2}{80\Delta D_{\max}}, \tag{A.3.28}$$

$$\lambda \leq \frac{\theta C_{\min}}{20\sqrt{\Delta}}. \tag{A.3.29}$$

It is not hard to see that, for $\Delta \geq 3$ and if $\theta\Delta \leq K_1$, the second condition on λ implies the first one. Having this in mind one easily sees that, when $\lambda \leq \frac{13\theta}{500\sqrt{\Delta}}$, the second restriction on n (A.3.27) implies the first one (A.3.26). This proves the first part of Theorem 2.3.6.

A.4 Graphs in $\mathcal{G}_{\rm diam}(p, \Delta)$: Proof of Lemma 2.6.3

Let us focus on reconstructing the neighborhood of node $r = 1$ in $G \in \mathcal{G}_{\rm diam}(p, \Delta)$. If we cannot reconstruct this neighborhood then we also cannot reconstruct the graph G correctly.

First notice that, by strong duality, the convex problem $\min_{\underline{\theta} \in \mathbb{R}^{p-1}} \mathcal{L}^n(\underline{\theta}) + \lambda \|\underline{\theta}\|_1$ can be equivalently written as $\min_{\underline{\theta} \in \mathbb{R}^{p-1} : \|\underline{\theta}\|_1 \leq C} \mathcal{L}^n(\underline{\theta})$ for some $C = C(\lambda) > 0$. We denote any solution of these problem by $\hat{\underline{\theta}}(\lambda)$ and $\hat{\underline{\theta}}(C)$ respectively. Notice also that $\lambda \geq \lambda_{\min}$ is equivalent to $C \in [0, C_{\max}]$ for some $C_{\max} = C_{\max}(\Delta, \lambda_{\min})$. Therefore, without loss of generality, we can assume that $\underline{\theta}$ belongs to the compact set $\|\underline{\theta}\|_1 \leq C_{\max}$. Furthermore, notice that $\mathcal{L}^n(\underline{\theta}) = (1/n)\sum_{\ell=1}^{n} f(\underline{\theta}, \underline{X}^{(\ell)})$ where

$f(x, y) \leq 2C_{\max} + \log 2$ and is Lipschitz in x for every y. Consequently, since $\{\underline{X}^{(\ell)}\}$ are i.i.d. random variables, this representation of $\mathcal{L}^n(.)$ guarantees that with probability one, $\mathcal{L}^n(\underline{\theta})$ converges to $\mathcal{L}(\underline{\theta})$ uniformly over $\|\underline{\theta}\|_1 \leq C_{\max}$ as $n \to \infty$, [96].

Secondly, it is not hard to see that $\mathcal{L}(\underline{\theta})$ is strictly convex. Hence the problem $\min_{\underline{\theta} \in \mathbb{R}^{p-1}: \|\underline{\theta}\|_1 \leq C} \mathcal{L}(\underline{\theta})$ has a unique solution. Call it $\underline{\theta}^\infty(C) \in \mathbb{R}^{p-1}$. Because with probability one, $\mathcal{L}^n(\underline{\theta})$ converges to $\mathcal{L}(\underline{\theta})$ uniformly over $\|\underline{\theta}\| \leq C_{\max}$, and because both $\mathcal{L}$ and $\mathcal{L}^n$ are continuous, we have that any $\hat{\underline{\theta}}(C)$ converges to $\underline{\theta}^\infty(C)$ uniformly over $C \in [0, C_{\max}]$ as $n \to \infty$.

Finally, as a consequence the equivalence between the two optimization problems mentioned in the first paragraph, we have

$$\sup_{\lambda \geq \lambda_{\min}} \mathbb{P}\{\widehat{G}(\lambda) = G\} \leq \sup_{C \in [0, C_{\max}]} \mathbb{P}\{\text{supp}(\underline{\theta}^\infty(C)) = \partial r\}. \tag{A.4.1}$$

Consequently, if we prove that $\text{supp}(\underline{\theta}^\infty(C)) \neq \partial r$ for all $C \in [0, C_{\max}]$ then

$$\sup_{\lambda \geq \lambda_{\min}} \mathbb{P}\{\widehat{G}(\lambda) = G\} \leq \sup_{C \in [0, C_{\max}]} \mathbb{P}\{\text{supp}(\underline{\theta}^\infty(C)) = \partial r\} \to 0, \text{ as } n \to \infty \tag{A.4.2}$$

and therefore,

$$\sup_{\lambda \geq \lambda_{\min}} \mathbb{P}\{\widehat{G}(\lambda) = G\} \leq \epsilon, \text{ for all } n \geq n_0(\Delta, \epsilon, \lambda_{\min}), \tag{A.4.3}$$

which would finish the proof.

In order to prove that $\text{supp}(\underline{\theta}^\infty(C)) \neq \partial r$ first observe that by symmetry, the unique solution $\underline{\theta}^\infty = \{\theta^\infty_{12}, \theta^\infty_{13}, ..., \theta^\infty_{1p}\} \in \mathbb{R}^{p-1}$ satisfies, $\theta^\infty_{13} = \theta^\infty_{14}, ..., \theta^\infty_{1p}$. Hence, all we need to prove is that we cannot have $\theta^\infty_{12} = 0$ and $\theta^\infty_{13} \neq 0$. To study these two components of $\underline{\theta}^\infty$, we define $\tilde{\mathcal{L}}(\theta_{13}, \theta_{12}) = \mathcal{L}(\theta_{12}, \theta_{13}, \theta_{13}, ..., \theta_{13})$ and solve the optimization problem

$$\min_{\theta_{13}, \theta_{12} \in \mathbb{R}} \tilde{\mathcal{L}}(\theta_{13}, \theta_{12}) + \lambda \Delta |\theta_{13}| + \lambda |\theta_{13}|. \tag{A.4.4}$$

The following lemma now completes the proof.

Lemma A.4.1. *For $\Delta \geq 50$, if $2 \leq \Delta\theta \leq 3$ and $\lambda > 0$ then no solution of*

$$\min_{\theta_{13},\theta_{12}} \tilde{\mathcal{L}}(\theta_{13},\theta_{12}) + \lambda\Delta|\theta_{13}| + \lambda|\theta_{13}| \tag{A.4.5}$$

simultaneously satisfies $\theta_{12} = 0$ and $\theta_{13} \neq 0$.

Proof. Let $\mathcal{L}(\theta_{13},\theta_{12}) = \tilde{\mathcal{L}}(\theta_{13},\theta_{12}) + \lambda\Delta|\theta_{13}| + \lambda|\theta_{13}|$. A solution of the convex optimization problem (A.4.5) satisfies $\theta_{12} = 0$ and $\theta_{13} \neq 0$ for some $\lambda > 0$ if and only if, for some $\lambda > 0, \theta_{12} = 0$ and $\theta_{13} \neq 0$, we have

$$\frac{\partial \mathcal{L}}{\partial \theta_{12}} \ni 0, \tag{A.4.6}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{13}} \ni 0. \tag{A.4.7}$$

Above we use '$\ni 0$' to signify that the sub-gradient of $\mathcal{L}$ must contain 0.

Let $Z = X_3 + ... + X_p$ and notice that

$$\tilde{\mathcal{L}}(\theta_{13},\theta_{12}) = \log 2 + \mathbb{E}\{\log\cosh(Z\theta_{13} + X_2\theta_{12})\} - \Delta\theta_{13}\mathbb{E}\{X_1X_3\} - \theta_{12}\mathbb{E}\{X_1X_2\}. \tag{A.4.8}$$

Since $(\log\cosh x)' = \tanh x$ the above optimality conditions can be written as,

$$\left.\frac{\partial \mathcal{L}}{\partial \theta_{12}}\right|_{\theta_{12}=0,\theta_{13}>0} = \mathbb{E}\{X_2\tanh(Z\theta_{13})\} - \mathbb{E}\{X_1X_2\} + \lambda\frac{\partial|\theta_{12}|}{\partial\theta_{12}} \ni 0, \tag{A.4.9}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{13}} = \mathbb{E}\{Z\tanh(Z\theta_{13})\} - \Delta\mathbb{E}\{X_1X_3\} + \lambda\Delta\frac{\partial|\theta_{13}|}{\partial\theta_{13}} \tag{A.4.10}$$

$$= \Delta\mathbb{E}\{X_3\tanh(Z\theta_{13})\} - \Delta\mathbb{E}\{X_1X_3\} + \lambda\Delta\frac{\partial|\theta_{13}|}{\partial\theta_{13}} \tag{A.4.11}$$

$$= \Delta\mathbb{E}\{X_3\tanh(Z\theta_{13})\} - \Delta\mathbb{E}\{X_1X_3\} \pm \lambda\Delta = 0, \tag{A.4.12}$$

where in the last line we used the fact that, for $\theta_{13} \neq 0$ we have $\frac{\partial|\theta_{13}|}{\partial\theta_{13}} \in \{-1, 1\}$.

From the last condition we conclude that λ must satisfy $\lambda = \mathbb{E}\{X_1X_3\} - \mathbb{E}\{X_3\tanh(Z\theta_{13})\}$ if $\theta_{13} > 0$ or $\lambda = -\mathbb{E}\{X_1X_3\} + \mathbb{E}\{X_3\tanh(Z\theta_{13})\}$ if $\theta_{13} < 0$. But by symmetry $\mathbb{E}\{X_3\tanh(Z\theta_{13})\} = \Delta^{-1}\mathbb{E}\{Z\tanh(Z\theta_{13})\}$ which is strictly negative if $\theta_{13} < 0$. Hence, not to contradict the fact that $\lambda > 0$, it must be that $\theta_{13} > 0$ and $\lambda =$

$\mathbb{E}\{X_1X_3\} - \mathbb{E}\{X_3 \tanh(Z\theta_{13})\}$.

From now on we assume that $\mathbb{E}\{X_1X_3\} - \mathbb{E}\{X_3 \tanh(Z\theta_{13})\} > 0$, otherwise $\lambda \leq 0$ and the proof again follows by contradiction. Replacing this value in the first condition we obtain,

$$\frac{\partial|\theta_{12}|}{\partial\theta_{12}} \ni \frac{\mathbb{E}\{X_1X_2\} - \mathbb{E}\{X_2 \tanh(Z\theta_{13})\}}{\mathbb{E}\{X_1X_3\} - \mathbb{E}\{X_3 \tanh(Z\theta_{13})\}}. \tag{A.4.13}$$

The sub-gradient of the modulus function can only take values in $[-1, 1]$. To finish the proof we now show that, under the conditions of Lemma A.4.1 and for any $\theta_{13} > 0$,

$$\mathbb{E}\{X_1X_2\} - \mathbb{E}\{X_2 \tanh(Z\theta_{13})\} - \mathbb{E}\{X_1X_3\} + \mathbb{E}\{X_3 \tanh(Z\theta_{13})\} > 0. \tag{A.4.14}$$

This implies that the optimality conditions cannot be satisfied for any $\lambda > 0$ when $\theta_{13} > 0$ and $\theta_{12} = 0$.

Let us define, $\beta \equiv \theta_{13}$ and

$$F(\beta) \equiv \mathbb{E}\{X_1X_2\} - \mathbb{E}\{X_2 \tanh(Z\beta)\} - \mathbb{E}\{X_1X_3\} + \mathbb{E}\{X_3 \tanh(Z\beta)\}. \tag{A.4.15}$$

Since $\mathbb{P}_{G_p,\theta}(X_1 = 1|X_2, ..., X_p) = e^{\theta Z}/(e^{\theta Z}+e^{-\theta Z})$ we have that $\mathbb{E}\{X_1|Z\} = \mathbb{E}\{X_2|Z\} = \tanh(Z\theta)$. This allows us to make the following substitutions in (A.4.15),

$$\mathbb{E}\{X_1X_2\} = \mathbb{E}\{X_2 \tanh(\theta Z)\}, \tag{A.4.16}$$

$$\mathbb{E}\{X_1X_3\} = \mathbb{E}\{X_3 \tanh(\theta Z)\}, \tag{A.4.17}$$

$$\mathbb{E}\{X_2 \tanh(Z\beta)\} = \mathbb{E}\{\tanh(\theta Z) \tanh(Z\beta)\}, \tag{A.4.18}$$

and by symmetry, $\mathbb{E}\{X_3 \tanh(\theta Z)\} = \Delta^{-1}\mathbb{E}\{Z \tanh(\theta Z)\}$. With these substitutions, and introducing $\tilde{Z} = \Delta^{-1}Z$, we obtain

$$F(\beta) = \mathbb{E}\{(\tanh(\theta\Delta\tilde{Z}) - \tilde{Z})(\tanh(\theta\Delta\tilde{Z}) - \tanh(\beta\Delta\tilde{Z}))\}. \tag{A.4.19}$$

Let us denote the expression inside the expectation by $H(\tilde{Z})$, that is, $F(\beta) = \mathbb{E}\{H(\tilde{Z})\}$.

Since $H(\tilde{Z})$ is an even function of $\tilde{Z}$ we can write,

$$F(\beta) = 2\mathbb{E}\{H(\tilde{Z})\mathbb{I}_{\tilde{Z}>0}\} = 2\mathbb{E}\{H(\tilde{Z})\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}>\gamma}\} + 2\mathbb{E}\{H(\tilde{Z})\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}\leq\gamma}\}. \quad \text{(A.4.20)}$$

We now make use of the following inequalities

$$(1-\tanh^2 a)(a-b) \leq \tanh(a) - \tanh(b) \leq a-b,\ 0 \leq a \leq b, \quad \text{(A.4.21)}$$

$$\tanh(a) \geq a\gamma^{-1}\tanh\gamma,\ 0 \leq a \leq \gamma, \quad \text{(A.4.22)}$$

to obtain a lower bound on $F(\beta)$. First notice that,

$$2\mathbb{E}\{H(\tilde{Z})\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}\leq\gamma}\} \geq 2\mathbb{E}\Big\{\left(\frac{\theta\Delta\tilde{Z}\tanh\gamma}{\gamma} - \tilde{Z}\right)(1-\tanh^2\gamma)(\theta-\beta)\Delta\tilde{Z}\,\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}\leq\gamma}\Big\} \quad \text{(A.4.23)}$$

$$= 2\left(\frac{\theta\Delta\tanh\gamma}{\gamma} - 1\right)(1-\tanh^2\gamma)(\theta-\beta)\Delta\mathbb{E}\Big\{\tilde{Z}^2\,\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}\leq\gamma}\Big\} \quad \text{(A.4.24)}$$

$$= \left(\frac{\theta\Delta\tanh\gamma}{\gamma} - 1\right)(1-\tanh^2\gamma)(\theta-\beta)\Delta\mathbb{E}\Big\{\tilde{Z}^2\,\mathbb{I}_{|\tilde{Z}|\leq\frac{\gamma}{\Delta\theta}}\Big\}, \quad \text{(A.4.25)}$$

and also that,

$$2\mathbb{E}\{H(\tilde{Z})\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}>\gamma}\} \geq -2\mathbb{E}\{(1-\tanh(\theta\Delta))(\theta-\beta)\Delta\tilde{Z}\,\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}>\gamma}\} \quad \text{(A.4.26)}$$

$$\geq -2(1-\tanh(\theta\Delta))(\theta-\beta)\Delta\mathbb{E}\{\mathbb{I}_{\tilde{Z}>0,\theta\Delta\tilde{Z}>\gamma}\} \quad \text{(A.4.27)}$$

$$= -2(1-\tanh(\theta\Delta))(\theta-\beta)\Delta\mathbb{P}\{\tilde{Z}>0,\theta\Delta\tilde{Z}>\gamma\} \quad \text{(A.4.28)}$$

$$= -(1-\tanh(\theta\Delta))(\theta-\beta)\Delta\mathbb{P}\Big\{|\tilde{Z}| > \frac{\gamma}{\theta\Delta}\Big\}. \quad \text{(A.4.29)}$$

Since $|\tilde{Z}| \leq 1$,

$$\mathbb{E}\Big\{\tilde{Z}^2\,\mathbb{I}_{|\tilde{Z}|\leq\frac{\gamma}{\Delta\theta}}\Big\} \geq \mathbb{E}\{\tilde{Z}^2\} - \mathbb{P}\Big\{|\tilde{Z}| \leq \frac{\gamma}{\Delta\theta}\Big\}, \quad \text{(A.4.30)}$$

summing both expressions above and rearranging terms we obtain

$$\frac{F(\beta)}{\Delta(\theta-\beta)} \geq \left(\frac{\theta\Delta\tanh\gamma}{\gamma}-1\right)(1-\tanh^2\gamma)\mathbb{E}\{\tilde{Z}^2\} \tag{A.4.31}$$

$$-\left(\left(\frac{\theta\Delta\tanh\gamma}{\gamma}-1\right)(1-\tanh^2\gamma)+(1-\tanh(\theta\Delta))\right)\mathbb{P}\Big\{|\tilde{Z}|>\frac{\gamma}{\theta\Delta}\Big\}. \tag{A.4.32}$$

We now set $\gamma = 1.64$. Since $\Delta\theta \geq 2$, a quick numerical calculation shows that $\left(\frac{\theta\Delta\tanh\gamma}{\gamma}-1\right)(1-\tanh^2\gamma) > 3.3224(1-\tanh(\theta\Delta)) > 0$ and hence,

$$\frac{F(\beta)}{\Delta(\theta-\beta)} \geq \left(\frac{\theta\Delta\tanh\gamma}{\gamma}-1\right)(1-\tanh^2\gamma)\left(\mathbb{E}\{\tilde{Z}^2\}-1.3010\mathbb{P}\Big\{|\tilde{Z}|>\frac{\gamma}{\theta\Delta}\Big\}\right). \tag{A.4.33}$$

We compute a value for $\mathbb{E}\{\tilde{Z}^2\}$ and then an upper bound for $\mathbb{P}\Big\{|\tilde{Z}|>\frac{\gamma}{\theta\Delta}\Big\}$.

Using the fact that $\mathbb{E}\{X_iX_j\} = \mathbb{E}\{X_3X_4\}$ for all $i \neq j, i, j \notin \{1,2\}$ we have

$$\mathbb{E}\{\tilde{Z}^2\} = \Delta^{-2}\mathbb{E}\{Z^2\} = \Delta^{-2}(\Delta+\Delta(\Delta-1)\mathbb{E}\{X_3X_4\}) = \Delta^{-1}+(\Delta-1)\Delta^{-1}\mathbb{E}\{X_3X_4\}. \tag{A.4.34}$$

For the probability bound, start by noticing that $\mathbb{P}\{X_1 = X_2 = 1\} \geq \mathbb{P}\{X_1 = 1, X_2 = -1\}$. Then we can write,

$$\mathbb{P}\Big\{|\tilde{Z}|>\frac{\gamma}{\theta\Delta}\Big\} \leq 2\left(\mathbb{P}\Big\{|\tilde{Z}|>\frac{\gamma}{\theta\Delta}\Big|X_1=X_2=1\Big\}+\mathbb{P}\Big\{|\tilde{Z}|>\frac{\gamma}{\theta\Delta}\Big|X_1=1,X_2=-1\Big\}\right) \tag{A.4.35}$$

$$\times\,\mathbb{P}\{X_1=X_2=1\}. \tag{A.4.36}$$

Conditioned on $X_1 = X_2 = 1$, the random variable $\tilde{Z}$ is the average of i.i.d. random variables that take values in $\{-1,1\}$ and whose mean is $\tanh 2\theta$. Now we apply the Chernoff-Hoeffding bound. Since $\Delta \geq 50$ and $\theta\Delta \leq 3$ we have $\frac{\gamma}{\theta\Delta} - \tanh(2\theta) \geq$

$1.64/3 - \tanh(2 \times 3/50) \geq 0.4272$ and we can write,

$$\mathbb{P}\Big\{|\tilde{Z}| > \frac{\gamma}{\theta\Delta}\Big| X_1 = X_2 = 1\Big\} \leq 2\mathbb{P}\Big\{\tilde{Z} > \frac{\gamma}{\theta\Delta} - \tanh 2\theta \Big| X_1 = X_2 = 1\Big\} \tag{A.4.37}$$

$$\leq 2e^{-\frac{1}{2}\Delta\left(\frac{\gamma}{\theta\Delta} - \tanh 2\theta\right)^2} \leq 2e^{-0.09126\Delta}. \tag{A.4.38}$$

Conditioned on $X_1 = 1, X_2 = -1$, the random variable $\tilde{Z}$ is the average of i.i.d. random variables that take values in $\{-1, 1\}$ and whose mean is zero. Again using the Chernoff-Hoeffding bound, and since $\frac{\gamma}{\theta\Delta} \geq 1.64/3$ we obtain,

$$\mathbb{P}\Big\{|\tilde{Z}| > \frac{\gamma}{\theta\Delta}\Big| X_1 = 1, X_2 = -1\Big\} \leq 2e^{-\frac{1}{2}\Delta\left(\frac{\gamma}{\theta\Delta}\right)^2} \leq 2e^{-0.1494\Delta}. \tag{A.4.39}$$

Putting these two last bounds together, we obtain

$$\mathbb{P}\Big\{|\tilde{Z}| > \frac{\gamma}{\theta\Delta}\Big\} \leq 4e^{-0.09126\Delta}\mathbb{P}\{X_1 = X_2 = 1\} = 4e^{-0.09126\Delta}\left(\frac{1 + \mathbb{E}\{X_1X_2\}}{4}\right). \tag{A.4.40}$$

In order to guarantee that $F(\beta) > 0$ it suffices that,

$$\Delta^{-1} > 1.3010e^{-\frac{1}{2}\Delta(\tanh 2\theta)^2}, \tag{A.4.41}$$

$$\frac{\Delta - 1}{\Delta}\mathbb{E}\{X_3X_4\} > 1.3010e^{-\frac{1}{2}\Delta(\tanh 2\theta)^2}\mathbb{E}\{X_1X_2\}. \tag{A.4.42}$$

Since

$$\mathbb{E}\{X_1X_2\} = \tanh(\Delta \text{atanh}((\tanh\theta)^2)), \tag{A.4.43}$$

$$\mathbb{E}\{X_3X_4\} = 2(\tanh(2\theta))^2\mathbb{P}\{X_1 = X_2 = 1\} = 1/2(\tanh(2\theta))^2(1 + \mathbb{E}\{X_1X_2\}), \tag{A.4.44}$$

we have $\mathbb{E}\{X_3X_4\}/\mathbb{E}\{X_1X_2\} \geq 2/\Delta$. In addition, since $\Delta \geq 50$ we have $(\Delta - 1)/\Delta \geq 49/50$ and hence it suffices that

$$\Delta^{-1} - 1.3010e^{-0.09126\Delta} > 0. \tag{A.4.45}$$

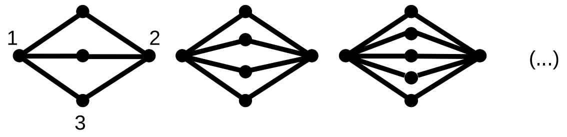

Figure A.1: For this family of graphs of increasing maximum degree Δ $\mathsf{Rlr}(\lambda)$ will fail for any $\lambda > 0$ if $\theta \geq 2/\Delta$.

Since $1/50 - 1.3010e^{-0.09126\times 50} = 0.006429 > 0$ the proof now follows. □

A.4.1 Graphs $\mathcal{G}_{\mathbf{diam}}(p)$ from the Section 2.6.3: Remark 2.3.1 and Remark 2.6.1

In Section A.4 we proved Lemma 2.6.3: $\mathsf{Rlr}(\lambda)$, $\lambda > 0$, fails to reconstruct graphs from $\mathcal{G}_{\text{diam}}(p)$ assymptotically (as $n \to \infty$ with λ fixed) when $2 \leq \theta\Delta \leq 3$. In this section we numerically investigate the behavior of Rlr in more detail and numerically demonstrate that in fact we do not need to impose that $\Delta\theta \leq 3$ and that we can take $\Delta_0 = 3$.

As we saw in Section A.4, for large n, the success of Rlr is dictated by

$$\min_{\theta_{13},\theta_{12}\in\mathbb{R}} \tilde{\mathcal{L}}(\theta_{13},\theta_{12}) + \lambda\Delta|\theta_{13}| + \lambda|\theta_{12}| \tag{A.4.46}$$

having an optimal solution with $\theta_{13} \neq 0$ and $\theta_{12} = 0$.

Since it only involves two variables, we can easily analyze this optimization problem by solving it numerically. Figure A.2 shows the solution path of this problem as a function of λ for $p = 5$ and for different values of θ.

From the plots we see that for high values of θ, Rlr will never yield a correct reconstruction (unless we assume $\lambda = 0$) since for these θs the curves are strictly above the horizontal ($\hat{\theta}_{12} > 0$) for all $\lambda > 0$. However, if θ is bellow a certain value, call it θ_T ($\theta_T \approx 0.61$ for $p = 5$), then there are $\lambda > 0$ for which the solution yields a correct reconstruction. In fact, for $\theta < \theta_T$ all curves exhibit a portion (above a certain λ) that have $\hat{\theta}_{12} = 0$ and $\hat{\theta}_{13} > 0$. In addition, we observe that there is a value θ_L such that if $\theta < \theta_L$ the curves identify themselves with the horizontal axis for all λ.

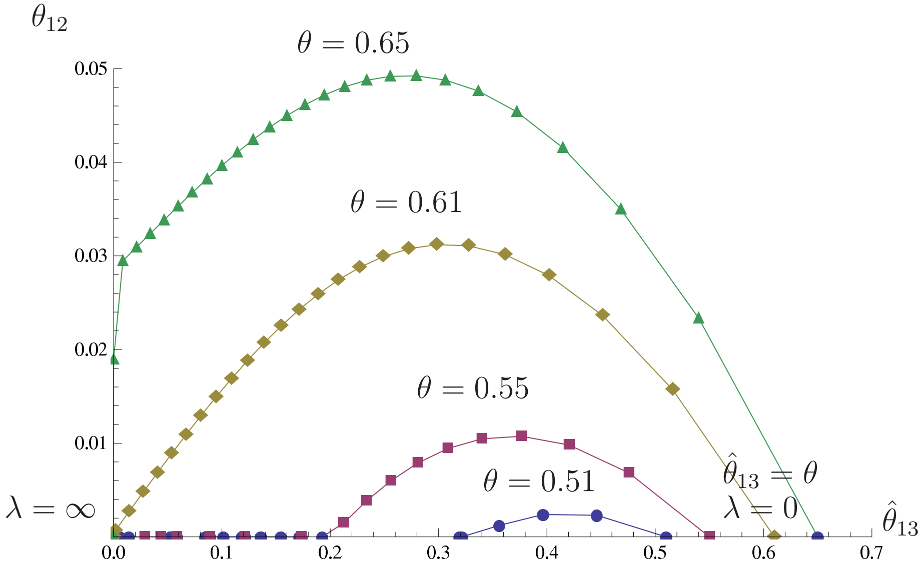

Figure A.2: Solution curves of $\mathsf{Rlr}(\lambda)$ as a function of λ for different values of θ and $p = 5$. Along each curve, λ increases from right to left. Plot points separated by $\delta\lambda = 0.05$ are included to show the speed of the parameterization with λ. For $\lambda \to \infty$ all curves tend to the point $(0, 0)$. For $\lambda = 0$, $\hat{\theta}_{13} = \theta$. Remark: Curves like the one for $\theta = 0.55$ are identically zero above a certain value of λ.

All the above observations hold in the limit when $n \to \infty$. For high finite n, with probability close to one (and uniformly over θ_{13}, θ_{13} and λ – cf. Section A.4) the solution curves will not be the ones plotted but rather be random fluctuations around these. For $\lambda = 0$, finite n and $\theta > \theta_L$, the solution curves will no longer start from $\underline{\hat{\theta}} = (\theta_{13}, \theta_{13}) = \underline{\theta}^0 = (\theta, 0)$ but will have a positive non vanishing probability of having $\hat{\theta}_{12} > 0$. This reflects the fact that for finite n the success of $\mathsf{Rlr}(\lambda)$ requires λ to be positive. However, for $\theta < \theta_L$ and $\lambda > 0$ such that we are in the region where the curves for $n = \infty$ are identically zero, the curves for finite n will have an increasing probability of being identically zero too. Thus, for these values of λ and θ, the probability of successful reconstruction will tend to 1 as $n \to \infty$. From the plots we also conclude that, unless the whole curve (for $n = \infty$) is identified with zero, $\mathsf{Rlr}(\lambda)$ restricted to the assumption $\lambda \to 0$ will fail with positive non vanishing probability for finite n. For $\theta < \theta_L$, when the curves (for $n = \infty$) become identically zero, there will be a scaling of λ with n to zero that will allow for a probability of success converging to 1 as $n \to \infty$.

When $\lambda \to 0$ with n, θ_L is the critical value above which reconstruction with Rlr fails. This is the related to the requirement that λ is small in Lemma 2.6.3. In fact, θ_L coincides with the value above which $\|Q^0_{S^C S}(Q^0_{SS})^{-1} z^0_S\|_\infty > 1$. However, we do not have to choose $\lambda \to 0$. We thus conclude that, for graphs in $\mathcal{G}_{\text{diam}}(p)$, the true condition required for successful reconstruction is not $\|Q^0_{S^C S}(Q^0_{SS})^{-1}\|_\infty < 1$ but rather that $\theta < \theta_T$. Surprisingly, for graphs in $\mathcal{G}_{\text{diam}}(p)$, this condition coincides with $\mathbb{E}_{G,\theta}(X_1 X_3) > \mathbb{E}_{G,\theta}(X_1 X_2)$, i.e. the correlation between neighboring nodes must be bigger than that between non-neighboring nodes.

This can be see in the following way. In the proof of Lemma 2.6.3 we proved that the failure of $\mathsf{Rlr}(\lambda)$ for $\lambda > 0$ is equivalent to $F(\beta)$ being positive for all $0 \leq \beta < \theta$. From equation (A.4.19) in the proof, we also know that $F(\theta) = 0$. We now argue that F also has the following property: θ is the smallest positive solution of $F(\beta) = 0$. A

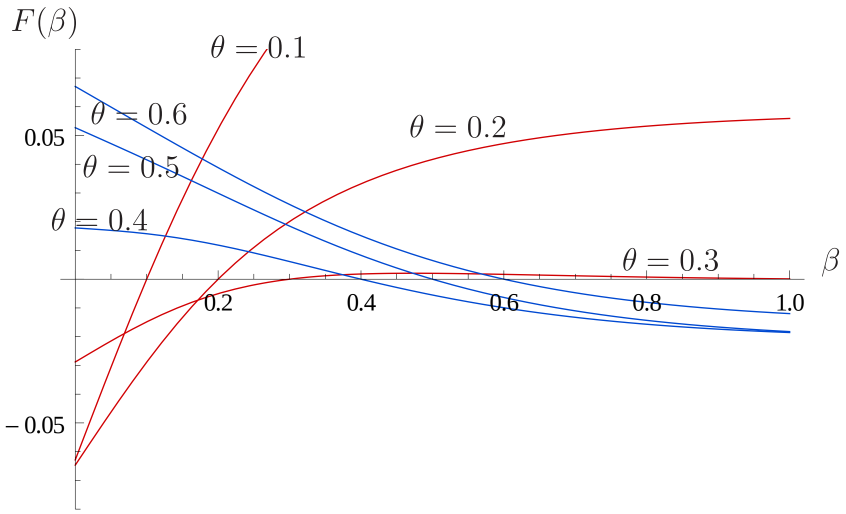

Figure A.3: $F(\beta)$ for $p = 6$. Red $\theta = 0.1, 0.2, 0.3$. Blue $\theta = 0.4, 0.5, 0.6$.

brute force computation shows that $F(\beta)$ can be written as,

$$\begin{aligned} F(\beta) = &\left(\frac{1}{2}\left(\tanh\left((p-2)\tanh^{-1}\left(\tanh^2(\theta)\right)\right)+1\right)-1\right) \\ &\times\left(\sum_{r=0}^{p-3} 2^{3-p}\binom{p-3}{r}\tanh(\beta(-p+2r+2))+1\right) \\ &-\left(1-\frac{e^{2\theta}}{e^{-2\theta}+e^{2\theta}}\right)\left(\tanh\left((p-2)\tanh^{-1}\left(\tanh^2(\theta)\right)\right)+1\right) \\ &\times\left(\sum_{r=0}^{p-3}\left(\frac{e^{2\theta}}{e^{-2\theta}+e^{2\theta}}\right)^r\binom{p-3}{r}\left(1-\frac{e^{2\theta}}{e^{-2\theta}+e^{2\theta}}\right)^{p-r-3}\tanh(\beta(-p+2r+2))-1\right). \end{aligned}$$

With this function we produced Figure A.3, where we plot $F(\beta)$ for several values of θ for $p = 6$ (similar plots are observed for other values of $p \geq 5$). From the plots, one observes that in fact, the smallest value of $\beta > 0$ for which $F(\beta) = 0$ corresponds to $\beta = \theta$. Because of this, $F(0) > 0$ if and only if $F(\beta) > 0$, $\forall \beta \in [0, \theta)$. Hence, the asymptotic failure of $\mathsf{Rlr}(\lambda)$ for $\lambda > 0$ is equivalent to $F(0) = \mathbb{E}\{X_1X_2\} - \mathbb{E}\{X_1X_3\} > 0$. Or in other words, the success is equivalent to $\mathbb{E}\{X_1X_3\} > \mathbb{E}\{X_1X_2\}$.

Remark A.4.1. *Notice that this condition is also the condition required for the simple thresholding algorithm* $\mathsf{Thr}(\tau)$ *to reconstruct a graph: the correlation between connected nodes needs to be greater than the correlation between non-connected nodes.*

We will now prove that if $\Delta \geq 3$ and $\Delta\theta \geq 2$ then $\mathbb{E}\{X_1 X_2\} > \mathbb{E}\{X_1 X_3\}$.

Let 1 and 2 be the two nodes with degree greater than 2 and let 3 be any other node (of degree 2), see Figure A.1. Define $x_\Delta = \mathbb{E}_{G,\theta}(X_1 X_2)$ and $y_\Delta = \mathbb{E}_{G,\theta}(X_1 X_3)$. It is not hard to see that,

$$x_{\Delta+1} = \frac{x_\Delta + \tanh^2\theta}{1 + \tanh^2\theta\, x_\Delta} \qquad y_{\Delta+1} = \frac{\tanh\theta\, x_\Delta + \tanh\theta}{1 + \tanh^2\theta\, x_\Delta}. \tag{A.4.47}$$

From these expression we see that the condition $x_\Delta(\theta) > y_\Delta(\theta)$ is equivalent to $x_{\Delta-1}(\theta) > \tanh\theta$.

Since $\tanh(a+b) = \frac{\tanh a + \tanh b}{1+\tanh a \tanh b}$, from the recursion for x_Δ we can obtain

$$x_{\Delta-1} = \tanh\left((\Delta-1)\text{arctanh}(\tanh^2(\theta))\right). \tag{A.4.48}$$

Therefore, $x_\Delta(\theta) > y_\Delta(\theta)$ is equivalent to $(\Delta-1)\text{arctanh}(\tanh^2(\theta)) > \theta$. But for $\theta > 0$, $\theta > \theta^{-1}\text{arctanh}(\tanh^2(\theta))$. Hence, if $(\Delta-1)\theta \geq 1$ we have $\mathbb{E}\{X_1X_2\} > \mathbb{E}\{X_1X_3\}$. But $(\Delta-1)\theta \geq 1$ is equivalent to $\Delta\theta \geq \Delta/(\Delta-1)$ and for $\Delta \geq 3$, $\Delta/(\Delta-1) \leq 3/2 \leq 2$ so our claim follows.

A.5 Graphs in $\mathcal{G}_{\rm rand}(p,\Delta)$: Proof of Lemma 2.6.4

To prove this theorem we will show that there exists $\theta_{\mathsf{Thr}}(\Delta)$ such that, when $\theta > \theta_{\mathsf{Thr}}(\Delta)$, we can compute constants $\epsilon = \epsilon(\Delta,\theta)$ and $C_{\min} = C_{\min}(\Delta,\theta)$ such that all the conditions of Lemma 2.6.1 hold when we define $\lambda_{\mathsf{Thr}}(\Delta,\theta) = C_{\min}^3\epsilon/(2^7(1+\epsilon^2)\Delta^3)$.

First, let us state explicitly the local weak convergence result mentioned in Sec. **??** right after our statement of Lemma 2.6.4. For $t \in \mathbb{N}$, let $\mathsf{T}(t) = (V_\mathsf{T}, E_\mathsf{T})$ be the regular rooted tree of degree Δ of t generations and define the associated Ising measure as

$$\mu^+_{\mathsf{T},\theta}(\underline{x}) = \frac{1}{Z_{\mathsf{T},\theta}} \prod_{(i,j)\in E_\mathsf{T}} e^{\theta x_i x_j} \prod_{i\in\partial\mathsf{T}(t)} e^{h^0 x_i}. \tag{A.5.1}$$

Here $\partial\mathsf{T}(t)$ is the set of leaves of $\mathsf{T}(t)$ and h^0 is the unique positive solution of

$$h = (\Delta - 1)\,\mathrm{atanh}\,\{\tanh\theta\,\tanh h\}\,. \tag{A.5.2}$$

It was proved in [83] that non-trivial local expectations with respect to $\mathbb{P}_{G,\theta}(\underline{x})$ converge to local expectations with respect to $\mu^{+}_{\mathsf{T},\theta}(\underline{x})$, as $p \to \infty$.

More precisely, let $\mathsf{B}_r(t)$ denote a ball of radius t around node $r \in G$ (the node whose neighborhood we are trying to reconstruct). For any fixed t, the probability that $\mathsf{B}_r(t)$ is not isomorphic to $\mathsf{T}(t)$ goes to 0 as $p \to \infty$.

Let $g(\underline{x}_{\mathsf{B}_r(t)})$ be any function of the variables in $\mathsf{B}_r(t)$ such that $g(\underline{x}_{\mathsf{B}_r(t)}) = g(-\underline{x}_{\mathsf{B}_r(t)})$. Then almost surely over graph sequences G_p of uniformly random regular graphs with p nodes (expectations here are taken with respect to the measures (2.1.1) and (A.5.1))

$$\lim_{p\to\infty} \mathbb{E}_{G,\theta}\{g(\underline{X}_{\mathsf{B}_r(t)})\} = \mathbb{E}_{\mathsf{T}(t),\theta,+}\{g(\underline{X}_{\mathsf{T}(t)})\}\,. \tag{A.5.3}$$

Notice that this characterizes expectations completely since if $g(\underline{x}_{\mathsf{B}_r(t)}) = -g(-\underline{x}_{\mathsf{B}_r(t)})$ then,

$$\mathbb{E}_{G,\theta}\{g(\underline{X}_{\mathsf{B}_r(t)})\} = 0. \tag{A.5.4}$$

The main part of the proof consists in considering $[Q^0_{S^CS}{Q^0_{SS}}^{-1}z^0_S]_i$ for $t = \mathrm{dist}(r,i)$ bounded. We then write $(Q^0_{SS})_{lk} = \mathbb{E}_{G,\theta}\{g_{l,k}(\underline{X}_{\mathsf{B}_r(t)})\}$ and $(Q^0_{S^CS})_{il} = \mathbb{E}_{G,\theta}\{g_{i,l}(\underline{X}_{\mathsf{B}_r(t)})\}$ for some functions $g_{\cdot,\cdot}(\underline{X}_{\mathsf{B}_r(t)})$ and apply the weak convergence result (A.5.3) to these expectations. We thus reduced the calculation of $[Q^0_{S^CS}(Q^0_{SS})^{-1}z^0_S]_i$ to the calculation of expectations with respect to the tree measure (A.5.1). The latter can be implemented explicitly through a recursive procedure, with simplifications arising thanks to the tree symmetry and by taking $t \gg 1$. A similar computation is done with regards to obtaining $C_{\min}$, the lower bound on $\sigma_{\min}(Q^0_{SS})$. The actual calculations consist in a (very) long exercise in calculus and can be found in the appendix of [16].

Appendix B

Learning stochastic differential equations

In this appendix we include all details of the proofs of Chapter 3. We start by proving the upper bounds on the sample complexity and then proceed to prove the lower bounds. Each of these proofs are included in a separate section. Auxiliary lemmas are be introduced as needed and are proven in a separate subsection inside sections.

B.1 Upper bounds on sample complexity of the regularized least squares algorithm

Our result for the continuous time model follows from an analysis of the problem for discrete time, introduced in Section 3.5.1, and taking the limit when $\eta \to 0$. Hence, we first prove Theorem 3.5.1, then we prove Theorem 3.3.1 and finally we prove the specialization this bound to the case of the Laplacian of a graph, Theorem 3.3.4.

B.2 Necessary condition for successful reconstruction of SDEs

In this Section we prove our main result for discrete-time dynamics, i.e., Theorem 3.5.1. We start by stating a set of sufficient conditions for regularized least squares to work. Then we present a series of concentration lemmas to be used to prove the validity of these conditions, and then finalize the proof.

As mentioned, the proof strategy, and in particular the following proposition, Proposition B.2.1, which provides a compact set of sufficient conditions for the support to be recovered correctly, is analogous to the one in [116]. A proof of this proposition can be found in the end of this subsection.

In the following we denote by $X \in \mathbb{R}^{p\times n}$ the matrix whose $(t+1)^{\text{th}}$ column corresponds to the configuration $\underline{x}(t)$, i.e. $X = [\underline{x}(0), \underline{x}(1), \dots, \underline{x}(n-1)]$. Further $\Delta X \in \mathbb{R}^{p\times n}$ is the matrix containing consecutive state changes, namely $\Delta X = [\underline{x}(1) - \underline{x}(0), \dots, \underline{x}(n) - \underline{x}(n-1)]$. It is important not to confuse $X_0^n \equiv \{\underline{x}(t) : 0 \le t \le n\}$ with X defined here. These are not the same. In addition, although both are related, X_0^n should not be confused with the n^{th} power of X (which is never mentioned in this thesis). Finally we write $W = [\underline{w}(1), \dots, \underline{w}(n-1)] \in \mathbb{R}^{p\times n}$ for the matrix containing the Gaussian noise realization. Equivalently,

$$W = \Delta X - \eta \Theta \, X \, .$$

The r^{th} row of W is denoted by W_r.

In order to lighten the notation, we omit the reference to X_0^n in the likelihood function (3.5.3) and simply write $\mathcal{L}(\Theta_r)$. We define its normalized gradient and Hessian by

$$\widehat{G} = -\nabla \mathcal{L}(\Theta_r^0) = \frac{1}{n\eta} X W_r^* \, , \qquad \widehat{Q} = \nabla^2 \mathcal{L}(\Theta_r^0) = \frac{1}{n} X X^* \, . \tag{B.2.1}$$

Proposition B.2.1. *Let $\alpha, C_{\min} > 0$ be be defined by*

$$\lambda_{\min}(Q^0_{S^0,S^0}) \equiv C_{\min} \, , \qquad \| Q^0_{(S^0)^C,S^0} \left(Q^0_{S^0,S^0} \right)^{-1} \|_\infty \equiv 1 - \alpha \, . \tag{B.2.2}$$

If the following conditions hold,

$$\|\widehat{G}\|_\infty \le \frac{\lambda\alpha}{3}\,, \qquad \|\widehat{G}_{S^0}\|_\infty \le \frac{\Theta_{\min}C_{\min}}{4\Delta} - \lambda, \tag{B.2.3}$$

$$\|\!|\widehat{Q}_{(S^0)^C,S^0} - Q^0_{(S^0)^C,S^0}|\!\|_\infty \le \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{\Delta}}\,, \qquad \|\!|\widehat{Q}_{S^0,S^0} - Q^0_{S^0,S^0}|\!\|_\infty \le \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{\Delta}}\,, \tag{B.2.4}$$

then the regularized least squares solution (3.5.2) *correctly recovers* $\mathrm{sign}(\Theta^0_r)$. *Further the same statement holds for the continuous model 3.3.2, provided $\widehat{G}$ and $\widehat{Q}$ are the gradient and the hessian of the likelihood (3.3.2).*

The proof of Theorem 3.5.1 consists in checking that, under the hypothesis (3.5.6) on the number of consecutive configurations, conditions (B.2.3) to (B.2.4) hold with high probability. Checking these conditions can be regarded in turn as concentration-of-measure statements. Indeed, if expectation is taken with respect to a stationary trajectory, we have $\mathbb{E}\{\widehat{G}\} = 0$, $\mathbb{E}\{\widehat{Q}\} = Q^0$.

The proof of B.2.1 can be bound in the appendix of [13].

In Section B.3 we state the concentration bounds that are used with Proposition B.2.1 to prove Theorem 3.5.1.

B.3 Concentration bounds

In this section we state the necessary concentration lemmas for proving Theorem 3.5.1. These are non-trivial because $\widehat{G}$, $\widehat{Q}$ are quadratic functions of *dependent* random variables $\big($the samples $\{\underline{x}(t)\}_{0\le t\le n}\big)$. The proofs of Proposition B.3.1, of Proposition B.3.2, and Corollary B.3.3 can be found in the appendix of [13].

Our first proposition implies concentration of $\widehat{G}$ around 0.

Proposition B.3.1. *Let $S \subseteq [p]$ be any set of vertices and $\epsilon < 1/2$. If $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\,\Theta^0) < 1$, then*

$$\mathbb{P}\big\{\|\widehat{G}_S\|_\infty > \epsilon\big\} \le 2|S|\, e^{-n(1-\sigma_{\max})\,\epsilon^2/4}. \tag{B.3.1}$$

We furthermore need to bound the matrix norms as per (B.2.4) in proposition

B.2.1. First we relate bounds on $\|\widehat{Q}_{JS} - Q^0{}_{JS}\|_\infty$ with bounds on $|\widehat{Q}_{ij} - Q^0_{ij}|$, $(i \in J, j \in S)$ where J and S are any subsets of $\{1, ..., p\}$. We have,

$$\mathbb{P}(\|\widehat{Q}_{JS} - Q^0_{JS})\|_\infty > \epsilon) \leq |J||S| \max_{i \in J, j \in S} \mathbb{P}(|\widehat{Q}_{ij} - Q^0_{ij}| > \epsilon/|S|). \tag{B.3.2}$$

Then, we bound $|\widehat{Q}_{ij} - Q^0_{ij}|$ using the following proposition

Proposition B.3.2. *Let $i, j \in \{1, ..., p\}$, $\sigma_{\max} \equiv \sigma_{max}(I + \eta\Theta^0) < 1$, $T = \eta n > 3/D$ and $0 < \epsilon < 2/D$ where $D = (1 - \sigma_{\max})/\eta$ then,*

$$\mathbb{P}(|\widehat{Q}_{ij} - Q^0_{ij})| > \epsilon) \leq 2e^{-\frac{n}{32\eta^2}(1-\sigma_{\max})^3\epsilon^2}. \tag{B.3.3}$$

Finally, the next corollary follows from Proposition B.3.2 and Eq. (B.3.2).

Corollary B.3.3. *Let J, S $(|S| \leq \Delta)$ be any two subsets of $\{1, ..., p\}$ and $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\Theta^0) < 1$, $\epsilon < 2\Delta/D$ and $n\eta > 3/D$ (where $D = (1 - \sigma_{\max})/\eta$) then,*

$$\mathbb{P}(\|\widehat{Q}_{JS} - Q^0_{JS}\|_\infty > \epsilon) \leq 2|J|\Delta e^{-\frac{n}{32\Delta^2\eta^2}(1-\sigma_{\max})^3\epsilon^2}. \tag{B.3.4}$$

B.4 Proof of Theorem 3.5.1

With the above concentration bounds, we now prove Theorem 3.5.1. All we need to do is to compute the probability that the conditions given by Proposition B.2.1 hold. From the statement of the theorem we have that the first two conditions $(\alpha, C_{\min} > 0)$ of Proposition B.2.1 hold. In order to make the first condition on $\widehat{G}$ imply the second condition on $\widehat{G}$ we assume that $\lambda\alpha/3 \leq (\theta_{\min}C_{\min})/(4\Delta) - \lambda$ which is guaranteed to hold if

$$\lambda \leq \theta_{\min}C_{\min}/8\Delta. \tag{B.4.1}$$

We also combine the two last conditions on $\widehat{Q}$, thus obtaining the following

$$\|\widehat{Q}_{[p],S^0} - Q^0_{[p],S^0}\|_\infty \leq \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{\Delta}}\,, \tag{B.4.2}$$

since $[p] = S^0 \cup (S^0)^C$. We then impose that both the probability of the condition on $\widehat{Q}$ failing and the probability of the condition on $\widehat{G}$ failing are upper bounded by $\delta/2$ using Proposition B.3.1 and Corollary B.3.3. As explained next, this leads to the bound on the sample-complexity (3.5.6) stated in the theorem.

B.4.1 Details of proof of Theorem 3.5.1

Using Proposition B.3.1 we see that the condition on $\widehat{G}$ fails with probability smaller than $\delta/2$ given that the following is satisfied

$$\lambda^2 = 36\alpha^{-2}(n\eta D)^{-1}\log(4p/\delta). \tag{B.4.3}$$

But we also want (B.4.1) to be satisfied and so substituting λ from the previous expression in (B.4.1) we conclude that n must satisfy

$$n \geq 2304\Delta^2 {C_{\min}}^{-2} {\theta_{\min}}^{-2} \alpha^{-2}(D\eta)^{-1}\log(4p/\delta). \tag{B.4.4}$$

In addition, the application of the probability bound in Proposition B.3.1 requires that

$$\frac{\lambda^2\alpha^2}{9} < 1/4 \tag{B.4.5}$$

so we need to impose further that,

$$n \geq 16(D\eta)^{-1}\log(4p/\delta). \tag{B.4.6}$$

To use Corollary B.3.3 for computing the probability that the condition on $\widehat{Q}$ holds we need,

$$n\eta > 3/D, \tag{B.4.7}$$

and

$$\frac{\alpha C_{\min}}{12\sqrt{\Delta}} < 2\Delta D^{-1}. \tag{B.4.8}$$

The last expression imposes the following conditions on Δ,

$$\Delta^{3/2} > 24^{-1}\alpha C_{\min} D. \tag{B.4.9}$$

The probability of the condition on $\widehat{Q}$ is upper bounded by $\delta/2$ if

$$n > 4608\eta^{-1}\Delta^3\alpha^{-2}{C_{\min}}^{-2}D^{-3}\log 4p\Delta/\delta. \tag{B.4.10}$$

The restriction (B.4.9) on Δ looks unfortunate but since $\Delta \geq 1$ we can actually show it always holds. Just notice $\alpha < 1$ and that

$$\sigma_{\max}(Q^0_{S^0,S^0}) \leq \sigma_{\max}(Q^0) \leq \frac{\eta}{1-\sigma_{\max}} \Leftrightarrow D \leq \sigma^{-1}_{\max}(Q^0_{S^0,S^0}) \tag{B.4.11}$$

therefore $C_{\min}D \leq \sigma_{\min}(Q^0_{S^0,S^0})/\sigma_{\max}(Q^0_{S^0,S^0}) \leq 1$. This last expression also allows us to simplify the four restrictions on n into a single one that dominates them. In fact, since $C_{\min}D \leq 1$ we also have $C^{-2}_{\min}D^{-2} \geq C^{-1}_{\min}D^{-1} \geq 1$ and this allows us to conclude that the only two conditions on n that we actually need to impose are the one at Equations (B.4.4), and (B.4.10). A little more of algebra shows that these two inequalities are satisfied if

$$n\eta > \frac{10^4\Delta^2(\Delta D^{-2} + \theta^{-2}_{\min})}{\alpha^2 D C^2_{\min}} \log(4p\Delta/\delta). \tag{B.4.12}$$

This concludes the proof of Theorem 3.5.1.

B.5 Proof of Theorem 3.3.1

To prove Theorem 3.3.1 we recall that Proposition B.2.1 holds provided the appropriate continuous time expressions are used for $\widehat{G}$ and $\widehat{Q}$, namely

$$\widehat{G} = -\nabla\mathcal{L}(\Theta^0_r) = \frac{1}{T}\int_0^T \underline{x}(t)\,\mathrm{d}b_r(t)\,, \qquad \widehat{Q} = \nabla^2\mathcal{L}(\Theta^0_r) = \frac{1}{T}\int_0^T \underline{x}(t)\underline{x}(t)^*\,\mathrm{d}t\,. \tag{B.5.1}$$

These are of course random variables. In order to distinguish these from the discrete time version, we adopt the notation $\widehat{G}^n$, $\widehat{Q}^n$ for the latter. We claim that these random variables can be coupled (i.e. defined on the same probability space) in such a way that $\widehat{G}^n \to \widehat{G}$ and $\widehat{Q}^n \to \widehat{Q}$ almost surely as $n \to \infty$ for fixed T. Under assumption (3.3.7), and making use of Lemma B.5.1 it is easy to show that (3.5.6) holds for all $n > n_0$ with n_0 a sufficiently large constant.

Therefore, by the proof of Theorem 3.5.1, the conditions in Proposition B.2.1 hold for gradient $\widehat{G}^n$ and Hessian $\widehat{Q}^n$ for any $n \geq n_0$, with probability larger than $1-\delta$. But by the claimed convergence $\widehat{G}^n \to \widehat{G}$ and $\widehat{Q}^n \to \widehat{Q}$, they hold also for $\widehat{G}$ and $\widehat{Q}$ with probability at least $1-\delta$ which proves the theorem.

We are left with the task of showing that the discrete and continuous time processes can be coupled in such a way that $\widehat{G}^n \to \widehat{G}$ and $\widehat{Q}^n \to \widehat{Q}$. With slight abuse of notation, the state of the discrete time system (3.5.1) is denoted by $\underline{x}(i)$ where $i \in \mathbb{N}$ and the state of continuous time system (3.1.1) by $\underline{x}(t)$ where $t \in \mathbb{R}$. We denote by Q^0 the solution of the Lyapunov equation (3.3.4) and by $Q^0(\eta)$ the solution of the modified Lyapunov equation (3.5.5). It is easy to check that $Q^0(\eta) \to Q^0$ as $\eta \to 0$ by the uniqueness of stationary state distribution (recall that the uniqueness follows because we have stability).

The initial state of the continuous time system $\underline{x}(t=0)$ is a $\mathsf{N}(0, Q^0)$ random variable independent of $\underline{b}(t)$ and the initial state of the discrete time system is defined to be $\underline{x}(i=0) = (Q^0(\eta))^{1/2}(Q^0)^{-1/2}\underline{x}(t=0)$. At subsequent times, $\underline{x}(i)$ and $\underline{x}(t)$ are generated by the respective dynamical systems using the same matrix Θ^0 using common randomness provided by the standard Brownian motion $\{\underline{b}(t)\}_{0\leq t\leq T}$ in $\mathbb{R}^p$. In order to couple $\underline{x}(t)$ and $\underline{x}(i)$, we construct $\underline{w}(i)$, the noise driving the discrete time system, by letting $\underline{w}(i) \equiv (\underline{b}(Ti/n) - \underline{b}(T(i-1)/n))$.

The almost sure convergence $\widehat{G}^n \to \widehat{G}$ and $\widehat{Q}^n \to \widehat{Q}$ follows then from standard convergence of random walk to Brownian motion.

B.5.1 Auxiliary lemma for proof of Theorem 3.3.1

Lemma B.5.1. *Let* $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\Theta^0)$ *and* $\rho_{\min} = -\lambda_{\max}((\Theta^0 + (\Theta^0)^*)/2) > 0$ *then,*

$$-\lambda_{\min}\left(\frac{\Theta^0 + (\Theta^0)^*}{2}\right) \geq \limsup_{\eta\to 0} \frac{1-\sigma_{\max}}{\eta}, \tag{B.5.2}$$

$$\liminf_{\eta\to 0} \frac{1-\sigma_{\max}}{\eta} \geq -\lambda_{\max}\left(\frac{\Theta^0 + (\Theta^0)^*}{2}\right). \tag{B.5.3}$$

Proof.

$$\frac{1-\sigma_{\max}}{\eta} = \frac{1-\lambda_{\max}^{1/2}((I+\eta\Theta^0)^*(I+\eta\Theta^0))}{\eta} \tag{B.5.4}$$

$$= \frac{1-\lambda_{max}^{1/2}(I+\eta(\Theta^0+(\Theta^0)^*)+\eta^2(\Theta^0)^*\Theta^0)}{\eta} \tag{B.5.5}$$

$$= \frac{1-(1+\eta u^*(\Theta^0+(\Theta^0)^*+\eta(\Theta^0)^*\Theta^0)u)^{1/2}}{\eta}, \tag{B.5.6}$$

where u is some unit vector that depends on η. Thus, since $\sqrt{1+x} = 1+x/2+O(x^2)$,

$$\liminf_{\eta\to 0}\frac{1-\sigma_{\max}}{\eta} = -\limsup_{\eta\to 0} u^*\left(\frac{\Theta^0+(\Theta^0)^*}{2}\right)u \geq -\lambda_{\max}\left(\frac{\Theta^0+(\Theta^0)^*}{2}\right). \tag{B.5.7}$$

The other inequality is proved in a similar way. □

B.6 Proofs for the lower bounds

In this section we prove Theorem 3.3.3 and Theorem 3.5.2 to Theorem 3.7.2.

Throughout, $\{\underline{x}(t)\}_{t\geq 0}$ is assumed to be a stationary process. It is immediate to check that under the assumptions of the Theorems 3.3.3 and 3.7.1, the SDE admit a unique stationary measure, with bounded covariance. This covariance is denoted by

$$Q^0 = \mathbb{E}\{\underline{x}(0)\underline{x}(0)^*\} - \mathbb{E}\{\underline{x}(0)\}(\mathbb{E}\{\underline{x}(0)\})^* = \mathbb{E}\{\underline{x}(t)\underline{x}(t)^*\} - \mathbb{E}\{\underline{x}(t)\}(\mathbb{E}\{\underline{x}(t)\})^*.$$

Special notation

Recall that regarding the proofs of the lower bounds the notation is a bit different. This is also true in Appendix B.6. In particular, $\mathbb{P}_{\Theta^0}$ is *not* a probability distribution parametrized by Θ^0 but rather a probability distribution for the random variable Θ^0. Similarly, for example, $\mathbb{E}_{\underline{x}(0)}$ denotes an expectation *only* with regards to the random variable $\underline{x}(0)$, keeping everything else fixed. When we write simply $\mathbb{P}$ or $\mathbb{E}$ we mean that the expectation is taken with regards to all random variables.

B.6.1 A general bound for linear SDE's

Before passing to the actual proofs, it is useful to establish a general bound for linear SDE's (3.1.6) with symmetric interaction matrix Θ^0. In what follows Θ^0 is a random variable whose outcome always leads to a stable system of SDEs and $M(\Theta^0)$ is a certain property/function of the matrix Θ^0, e.g. its signed support.

Lemma B.6.1. *Let X_0^T be the unique stationary process generated by the linear SDE (3.1.6) for a certain realization of the symmetric random matrix Θ^0. Let $\widehat{M}_T(X_0^T)$ be an estimator of $M(\Theta^0)$ based on X_0^T. If $\mathbb{P}(\widehat{M}_T(X_0^T) \neq M(\Theta^0)) < \frac{1}{2}$ then*

$$T \geq \frac{H(M(\Theta^0)) - 2I(\Theta^0; x(0))}{\frac{1}{2}\text{Tr}\{\mathbb{E}\{-\Theta^0\} - (\mathbb{E}\ \{-(\Theta^0)^{-1}\})^{-1}\}}. \tag{B.6.1}$$

Proof. The bound follows from Corollary 3.5.3 after showing that

$$\mathbb{E}_{\underline{x}(0)}\{\text{Var}_{\Theta^0|\underline{x}(0)}(\Theta^0\underline{x}(0))\} \leq (1/2)\text{Tr}\{\mathbb{E}\{-\Theta^0\} - (\mathbb{E}\ \{-(\Theta^0)^{-1}\})^{-1}\} \tag{B.6.2}$$

First note that

$$\mathbb{E}_{\underline{x}(0)}\{\mathrm{Var}_{\Theta^0|\underline{x}(0)}(\Theta^0\underline{x}(0))\} = \mathbb{E}_{\underline{x}(0)}||\Theta^0\underline{x}(0) - \mathbb{E}_{\Theta^0|\underline{x}(0)}(\Theta^0\underline{x}(0)|\underline{x}(0))||_2^2. \tag{B.6.3}$$

The quantity in (B.6.3) can be thought of as the ℓ_2-norm error of estimating $\Theta^0\underline{x}(0)$ based on $\underline{x}(0)$ using $\mathbb{E}_{\Theta^0|\underline{x}(0)}(\Theta^0\underline{x}(0)|\underline{x}(0))$. Since conditional expectation is the minimal mean square error estimator, replacing $\mathbb{E}_{\Theta^0|\underline{x}(0)}(\Theta^0\underline{x}(0)|\underline{x}(0))$ by any estimator of $\Theta^0\underline{x}(0)$ based on $\underline{x}(0)$ gives an upper bound for the expression in (B.6.3). We choose as an estimator a linear estimator, i.e., an estimator of the form $B\underline{x}(0)$ where $B = (\mathbb{E}_{\Theta^0}\Theta^0 Q^0)(\mathbb{E}_{\Theta^0}Q^0)^{-1}$. We get,

$$\begin{aligned}&\mathbb{E}_{\underline{x}(0)}||\Theta^0\underline{x}(0) - \mathbb{E}_{\Theta^0|\underline{x}(0)}(\Theta^0\underline{x}(0)|\underline{x}(0))||_2^2 \leq \mathbb{E}_{\underline{x}(0)}||\Theta^0\underline{x}(0) - B\underline{x}(0)||_2^2\\ &= \mathrm{Tr}\{\mathbb{E}\{\Theta^0\underline{x}(0)(\underline{x}(0))^*\Theta^{0*}\}\} - 2\mathrm{Tr}\{B\mathbb{E}\{\underline{x}(0)(\underline{x}(0))^*\Theta^{0*}\}\}\\ &+ \mathrm{Tr}\{B\mathbb{E}\{\underline{x}(0)(\underline{x}(0))^*\}B^*\}.\end{aligned} \tag{B.6.4}$$

Furthermore, for a linear system, Q^0 satisfies the Lyapunov equation

$$\Theta^0 Q^0 + Q^0(\Theta^0)^* + I = 0. \tag{B.6.5}$$

For Θ^0 symmetric, this implies $Q^0 = -(1/2)(\Theta^0)^{-1}$. Substituting this expression in (B.6.3) and (B.6.4) finishes the proof. □

B.6.2 Proof of Theorem 3.3.3

We prove Theorem 3.3.3 by showing that the same complexity bound holds in the case when we are trying to estimate the signed support of Θ^0 for an Θ^0 that is uniformly randomly chosen with a distribution supported on $\mathcal{A}^{(S)}$ and we simultaneously require that the average probability of error is smaller than 1/2. This guarantees that unless the bound holds, there exists $A \in \mathcal{A}^{(S)}$ for which the probability of error is bigger than 1/2. The complexity bound for random matrices Θ^0 is proved using Lemma B.6.1.

In order to generate Θ^0 at random we proceed as follows. Let G be the a random matrix constructed from the adjacency matrix of a uniformly random Δ-regular graph. Generate $\tilde{\Theta}^0$ by flipping the sign of each non-zero entry in G with probability 1/2 independently. We define Θ^0 to be the random matrix

$$\Theta^0 = -(\gamma + 2\theta_{\min}\sqrt{\Delta - 1})I + \theta_{\min}\tilde{\Theta}^0 \tag{B.6.6}$$

where $\gamma = \gamma(\tilde{\Theta}^0) > 0$ is the smallest value such that the maximum eigenvalue of Θ^0 is smaller than $-\rho$. This guarantees that Θ^0 satisfies the four properties of the class $\mathcal{A}^{(S)}$.

The following lemma encapsulates the necessary random matrix calculations.

Lemma B.6.2. *Let Θ be a random matrix defined as above and*

$$Q(\theta_{\min}, \Delta, \rho) \equiv \lim_{p\to\infty} \frac{1}{p}\{\mathrm{Tr}\{\mathbb{E}(-\Theta)\} - \mathrm{Tr}\{(\mathbb{E}(-\Theta^{-1}))^{-1}\}\}.$$

Then, there exists a constant C' only dependent on Δ such that

$$Q(\theta_{\min}, \Delta, \rho) \leq \min\Big\{\frac{C'\Delta\theta_{\min}^2}{\rho}, \frac{\Delta\theta_{\min}}{\sqrt{\Delta - 1}}\Big\}. \tag{B.6.7}$$

Proof. First notice that

$$\lim_{p\to\infty} \frac{1}{p}\mathbb{E}\mathrm{Tr}\{-\Theta\} = \lim_{p\to\infty} \mathbb{E}(\gamma) + 2\theta_{\min}\sqrt{\Delta - 1} \tag{B.6.8}$$

$$= \rho + 2\theta_{\min}\sqrt{\Delta - 1}. \tag{B.6.9}$$

This holds since by Kesten-McKay law [45], for large p, the spectrum of $\tilde{\Theta}$ has support in $(-\epsilon - 2\theta_{\min}\sqrt{\Delta - 1}, 2\theta_{\min}\sqrt{\Delta - 1} + \epsilon)$ with high probability. Notice that unless we randomize each entry of $\tilde{\Theta}$ with $\{-1, +1\}$ values, every $\tilde{\Theta}$ will have Δ as its largest eigenvalue and the above limit will not hold.

For the second term we will compute a lower bound. For that purpose let $\lambda_i > 0$

be the i^{th} eigenvalue of the matrix $\mathbb{E}(-\Theta^{-1})$. We can write,

$$\frac{1}{p}\mathrm{Tr}\{(\mathbb{E}(-\Theta^{-1}))^{-1}\} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i} \tag{B.6.10}$$

$$\geq \frac{1}{\frac{1}{p}\sum_{i=1}^{p}\lambda_i} = \frac{1}{\mathbb{E}\{\frac{1}{p}\mathrm{Tr}\{(-\Theta)^{-1}\}\}} \tag{B.6.11}$$

where we applied Jensen's inequality in the last step. By Kesten-McKay law we now have that,

$$\lim_{p\to\infty}\mathbb{E}\{\frac{1}{p}\mathrm{Tr}\{(-\Theta)^{-1}\}\} = \mathbb{E}\{\lim_{p\to\infty}\frac{1}{p}\mathrm{Tr}\{(-\Theta)^{-1}\}\} \tag{B.6.12}$$

$$= \frac{1}{\theta_{\min}}G(\Delta, \rho/\theta_{\min} + 2\sqrt{\Delta-1}) \tag{B.6.13}$$

where

$$G(\Delta, z) = \int \frac{-1}{\nu - z}\mathrm{d}\mu(\nu). \tag{B.6.14}$$

$\mu(\nu)$ is the Kesten-McKay distribution and inside its support, $\nu \in [-2\sqrt{\Delta-1}, -2\sqrt{\Delta-1}]$, it is defined by

$$\mathrm{d}\mu(\nu) = \frac{\Delta}{2\pi}\frac{\sqrt{4(\Delta-1)-\nu^2}}{\Delta^2-\nu^2}\mathrm{d}\nu.$$

Computing the above integral we obtain

$$G(\Delta, z) = -\frac{(\Delta-2)z - \Delta\sqrt{-4\Delta+z^2+4}}{2\left(z^2-\Delta^2\right)} \tag{B.6.15}$$

whence

$$\lim_{\rho\to 0} Q(\theta_{\min}, \Delta, \rho) = \frac{\theta_{\min}\Delta}{\sqrt{\Delta-1}}, \tag{B.6.16}$$

$$\lim_{\rho\to\infty} \rho\, Q(\theta_{\min}, \Delta, \rho) = \Delta(\theta_{\min})^2. \tag{B.6.17}$$

Since $Q(\theta_{\min}, \Delta, \rho)/\theta_{\min}$ is a function of Δ and $\rho/\theta_{\min}$ that is strictly decreasing with

$\rho/\theta_{\min}$, the claimed bound follows. □

Proof of Theorem 3.3.3

Starting from the bound of Lemma B.6.1, we divide both terms in the numerator and the denominator by p. The term $H(M(\Theta^0))/p$ can be lower bounded by

$$p^{-1}\log\left(\binom{p}{\Delta}2^{\Delta}\right)^{p} \geq \Delta\log(2p/\Delta) \tag{B.6.18}$$

and Lemma B.6.2 gives an upper bound on the denominator when $p \to \infty$.

We now prove that $\lim_{p\to\infty} I(\underline{x}(0);\Theta^0)/p \leq 1$. This finishes the proof of Theorem 3.3.3 since after multiplying by a small enough constant (only dependent on Δ) the bound obtained by replacing the numerator and denominator with these limits is valid for all p large enough.

First notice that $h(\underline{x}(0)) \leq (1/2)\log(2\pi e)^p|\mathbb{E}\{Q^0\}|$ and hence,

$$I(\underline{x}(0);\Theta^0) = h(\underline{x}(0)) - h(\underline{x}(0)|\Theta^0) \tag{B.6.19}$$

$$\leq \frac{1}{2}\log(2\pi e)^p|\mathbb{E}\{Q^0\}| - \mathbb{E}\Big\{\frac{1}{2}\log(2\pi e)^p|Q^0|\Big\}, \tag{B.6.20}$$

where $Q^0 = -(1/2)(\Theta^0)^{-1}$ is the covariance matrix of the stationary process $\underline{x}(t)$ and $|.|$ denotes the determinant of a matrix. Then we write,

$$I(\underline{x}(0);\Theta^0) \leq (1/2)\log|\mathbb{E}\{-(\beta\Theta^0)^{-1}\}| + (1/2)\mathbb{E}\{\log(|-\beta\Theta^0|)\} \tag{B.6.21}$$

$$\leq \frac{1}{2}\mathrm{Tr}\{\mathbb{E}\{(-I-(\beta\Theta^0)^{-1}\}\} + \frac{1}{2}\mathbb{E}\{\mathrm{Tr}\{-I-\beta\Theta^0\}\} \tag{B.6.22}$$

where $\beta > 0$ is an arbitrary rescaling factor and the last inequality follows from $\log(I+M) \leq \mathrm{Tr}(M)$. From this and equations (B.6.8) and (B.6.12) it follows that,

$$\lim_{p\to\infty}\frac{1}{p}I(\underline{x}(0);\Theta^0) \leq -1 + (1/2)(\beta' z + \beta'^{-1}G(\Delta, z)) \tag{B.6.23}$$

where $z = \rho/\theta_{\min} + 2\sqrt{\Delta-1}$ and $\beta' = \beta\theta_{\min}$. To finish, note that optimizing over β'

and then over z gives,

$$\beta' z + \beta'^{-1} G(\Delta, z) \leq 2\sqrt{z G(\Delta, z)} \leq \sqrt{\frac{8(\Delta - 1)}{\Delta - 2}} \leq 4. \tag{B.6.24}$$

Bibliography

[1] V. Tan A. Anandkumar and A. Willsky. High-dimensional graphical model selection: Tractable graph families and necessary conditions. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2011.

[2] C. C. Johnson A. Jalali and P. Ravikumar. On learning discrete graphical models using greedy methods. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2011.

[3] P. Abbeel, D. Koller, and A.Y. Ng. Learning factor graphs in polynomial time and sample complexity. *The Journal of Machine Learning Research*, 7:1743–1788, 2006.

[4] D.H. Ackley, G.E. Hinton, and T.J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

[5] Y. AïtâĂŘSahalia. Maximum likelihood estimation of discretely sampled diffusions: a closed-form approach. *Econometrica*, 70:223–262, 2002.

[6] B.B. Aldridge, J.M. Burke, D.A. Lauffenburger, and P.K. Sorger. Physicochemical modelling of cell signalling pathways. *Nature cell biology*, 8(11):1195–1203, 2006.

[7] A. Anandkumar and E. Mossel. Learning high-dimensional latent graphical models: Girth-constrained graph families. 2011.

[8] A. Anandkumar, V. Tan, and A. Willsky. High dimensional structure learning of ising models on sparse random graphs: Local separation criterion. *Arxiv preprint arXiv:1107.1736*, 2011.

[9] A. Anandkumar, V.Y.F. Tan, and A. Willsky. High-dimensional gaussian graphical model selection: Walk summability and local separation criterion. *Arxiv preprint arXiv:1107.1270*, 2011.

[10] L. Bachelier. Théorie de la spÃl’culation. *Annales Scientifiques de lâĂŹEcole Normale Supŕieure*, 3:21âĂŞ86, 1900.

[11] O. Banerjee, L. El Ghaoui, and A. d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

[12] I.V. Basawa and B.L.S. Prakasa Rao. *Statistical inference for stochastic processes*. Academic Press, London, 1980.

[13] J. Bento, M. Ibrahimi, and A. Montanari. Estimating the drift coefficient of high-dimensional diffusions. In preparation, 2012.

[14] J. Bento, Morteza Ibrahimi, and Andrea Montanari. Learning networks of stochastic differential equations. *Advances in Neural Information Processing Systems 23*, pages 172–180, 2010.

[15] J. Bento, Morteza Ibrahimi, and Andrea Montanari. Information theoretic limits on learning stochastic differential equations. In *IEEE Intl. Symp. on Inform. Theory*, St. Perersbourg, August 2011.

[16] J. Bento and A. Montanari. Thresholds in maximum degree for learning ising models via l1-regularized logistic regression. In preparation, 2012.

[17] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236, 1974.

[18] C.M. Bishop and SpringerLink (Service en ligne). *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[19] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637âĂŞ654, 1973.

[20] A. Bolstad, B. Van Veen, and R. Nowak. Causal network inference via group sparse regularization. *IEEE transactions on signal processing*, 59:2628–2641, 2011.

[21] G. Bresler, E. Mossel, and A. Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 343–356, 2008.

[22] E.T. Bullmore and D.S. Bassett. Brain graphs: graphical models of the human brain connectome. *Annual review of clinical psychology*, 7:113–140, 2011.

[23] R. Cano, C. Sordo, and J.M. Gutiérrez. Applications of bayesian networks in meteorology. *Studies in Fuzziness And Soft Computing*, 146:309–328, 2004.

[24] G. Casella and E.I. George. Explaining the gibbs sampler. *American Statistician*, pages 167–174, 1992.

[25] E. Castillo, J.M. Gutiérrez, and A.S. Hadi. *Expert systems and probabilistic network models*. Springer Verlag, 1997.

[26] J. Chang and Chen S. X. On the approximate maximum likelihood estimation for diffusion processes. *The Annals of Statistics*, 39:2820–2851, 2011.

[27] F.R.K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, 1997.

[28] S.Y. Chung, T.J. Richardson, and R.L. Urbanke. Analysis of sum-product decoding of low-density parity-check codes using a gaussian approximation. *Information Theory, IEEE Transactions on*, 47(2):657–670, 2001.

[29] S. Cocco and R. Monasson. Adaptive cluster expansion for inferring boltzmann machines with noisy data. *Physical Review Letters*, 106, 2011.

[30] I. Csiszar and Z. Talata. Consistent estimation of the basic neighborhood of markov random fields. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, page 170. IEEE, 2004.

[31] D. Dacunha-Castelle and D. Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19:263–284, 1986.

[32] A. Dalalyan. Sharp adaptive estimation of the drift function for ergodic diffusions. *The Annals of Statistics*, 33:2507–2528, 2005.

[33] J.N. Darroch, S.L. Lauritzen, and TP Speed. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, pages 522–539, 1980.

[34] P. Dayan, L.F. Abbott, and L. Abbott. Theoretical neuroscience: Computational and mathematical modeling of neural systems. 2001.

[35] A. Dembo and A. Montanari. Ising models on locally tree-like graphs. *The Annals of Applied Probability*, 20(2):565–592, 2010.

[36] D.L. Donoho. For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006.

[37] D.L. Donoho. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[38] T.E. Duncan. On the calculation of mutual information. *SIAM Journal on Applied Mathematics*, 19(1):215–220, 1970.

[39] J. Fan. A selective overview of nonparametric methods in financial econometrics. *Statist. Sci.*, 20:317âĂŞ357, 2005.

[40] M.E. Fisher. Critical temperatures of anisotropic ising lattices. ii. general upper bounds. *Physical Review*, 162(2):480, 1967.

[41] R.A. Fisher. Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge Univ Press, 1925.

[42] B.J. Frey. *Graphical models for machine learning and digital communication*. MIT press, 1998.

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

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

[45] Joel Friedman. A proof of Alon's second eigenvalue conjecture. *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 720–724, 2003.

[46] N. Friedman. Learning belief networks in the presence of missing values and hidden variables. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 125–133. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

[47] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science's STKE*, 303(5659):799, 2004.

[48] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[49] N. Friedman, I. Nachman, and D. Peér. Learning bayesian network structure from massive datasets: the «sparse candidate «algorithm. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 206–215. Morgan Kaufmann Publishers Inc., 1999.

[50] R. Gallager. Low-density parity-check codes. *Information Theory, IRE Transactions on*, 8(1):21–28, 1962.

[51] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[52] H.O. Georgii. *Gibbs measures and phase transitions*. Walter de Gruyter, 1988.

[53] A. Gerschcnfeld and A. Monianari. Reconstruction for models on random graphs. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 194–204. IEEE, 2007.

[54] D.T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[55] R.B. Griffiths. Correlations in ising ferromagnets. i. *Journal of Mathematical Physics*, 8:478, 1967.

[56] G. Grimmett. *The random-cluster model*, volume 333. Springer Verlag, 2006.

[57] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971.

[58] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke. Gene regulatory network inference: Data integration in dynamic models–a review. *Biosystems*, 96(1):86–103, 2009.

[59] T. Hertz and C. Yanover. Pepdist: a new framework for protein-peptide binding prediction based on learning peptide distance functions. *BMC bioinformatics*, 7(Suppl 1):S3, 2006.

[60] D. Higham. Modeling and Simulating Chemical Reactions. *SIAM Review*, 50:347–368, 2008.

[61] G.E. Hinton, S. Osindero, and Y.W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[62] G.E. Hinton and T.J. Sejnowski. Analyzing cooperative computation. In *Proceedings of the Fifth Annual Conference of the Cognitive Science Society*, pages 2554–2558, 1983.

[63] H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *The Journal of Machine Learning Research*, 10:883–906, 2009.

[64] J. Honorio and D. Samaras. Multi-task learning of gaussian graphical models. In *Proceedings of the 27th Conference on Machine Learning*, 2010.

[65] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[66] K. Huang. Statistical mechanics, 18.3, 1987.

[67] A. Montanar J. Bento. On the trade-off between complexity and correlation decay in structural learning algorithms. *Arxiv preprint arXiv:1110.1769*, 2011.

[68] A. Montanari J. Bento. Which graphical models are difficult to learn? In *Proceedings of Neural Information Processing Systems (NIPS)*, 2009.

[69] F.V. Jensen. *An introduction to Bayesian networks*, volume 36. UCL press London, 1996.

[70] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on Computing*, 22:1087–1116, 1993.

[71] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[72] T. Kadota, M. Zakai, and J. Ziv. Mutual information of the white gaussian channel with and without feedback. *IEEE Trans. Inf. Theory*, IT-17(4):368–371, July 1971.

[73] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780, 2008.

[74] R. Kindermann, J.L. Snell, and American Mathematical Society. *Markov random fields and their applications*. American Mathematical Society Providence, RI, 1980.

[75] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

[76] Yu.A. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer, New York, 2004.

[77] S.L. Lauritzen. *Graphical models*, volume 17. Oxford University Press, USA, 1996.

[78] E.L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer Verlag, 1998.

[79] S.Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag New York Inc, 2009.

[80] M.G. Luby, M. Mitzenmacher, M.A. Shokrollahi, and D.A. Spielman. Improved low-density parity-check codes using irregular graphs. *Information Theory, IEEE Transactions on*, 47(2):585–598, 2001.

[81] N. Meinhshausen and P. Bühlmann. High-Dimensional Graphs and Variable Selection with the LASSO. *Annals of Statistics*, 34:1436–1462, 2006.

[82] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[83] A. Montanari, E. Mossel, and A. Sly. The weak limit of ising models on locally tree-like graphs. *Probability Theory and Related Fields*, pages 1–21, 2012.

[84] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 366–375. ACM, 2005.

[85] E. Mossel and A. Sly. Exact thresholds for ising-gibbs samplers on general graphs. *Arxiv preprint arXiv:0903.2906*, 2009.

[86] K. Murphy. An introduction to graphical models. *Rap. tech*, 2001.

[87] P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai. Greedy learning of markov network structure. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1295–1302. IEEE, 2010.

[88] B.K. Øksendal. *Stochastic differential equations: an introduction with applications*. Springer Verlag, 2003.

[89] G.A. Pavliotis and A.M. Stuart. Parameter estimation for multiscale diffusions. *J. Stat. Phys.*, 127:741–781, 2007.

[90] Asger Roer Pedersen. Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, 1:pp. 257–279, 1995.

[91] Asger Roer Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, 22(1):pp. 55–71, 1995.

[92] Peter C. B. Phillips and Jun Yu. Maximum likelihood and gaussian estimation of continuous time models in finance. In Thomas Mikosch, Jens-Peter Krei, Richard A. Davis, and Torben Gustav Andersen, editors, *Handbook of Financial Time Series*, pages 497–530. Springer, 2009.

[93] Y. Pokern, A.M. Stuart, and E. Vanden-Eijnden. Remarks on drift estimation for diffusion processes. *Multiscale Modeling & Simulation*, 8:69–95, 2009.

[94] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty. High-dimensional ising model selection using 2̆1131-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[95] T.J. Richardson, M.A. Shokrollahi, and R.L. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *Information Theory, IEEE Transactions on*, 47(2):619–637, 2001.

[96] H. Rubin. Uniform convergence of random functions with applications to statistics. *The Annals of Mathematical Statistics*, pages 200–203, 1956.

[97] S. Russell, J. Binder, D. Koller, and K. Kanazawa. Local learning in probabilistic networks with hidden variables. In *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, volume 14, pages 1146–1152. Citeseer, 1995.

[98] N. Santhanam and M.J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *Arxiv preprint arXiv:0905.2639*, 2009.

[99] A. Siepel and D. Haussler. Phylogenetic hidden markov models. *Statistical methods in molecular evolution*, pages 325–351, 2005.

[100] A. Sly. Computational transition at the uniqueness threshold. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 287–296. IEEE, 2010.

[101] A. Sly and N. Sun. The computational hardness of counting in two-spin models on d-regular graphs. *Arxiv preprint arXiv:1203.2602*, 2012.

[102] J. Songsiri, J. Dahl, and L. Vandenberghe. *Graphical models of autoregressive processes*, pages 89–116. Cambridge University Press, 2010.

[103] J. Songsiri and L. Vandenberghe. Topology selection in graphical models of autoregressive processes. *Journal of Machine Learning Research*, 2010. submitted.

[104] V.G. Spokoiny. Adaptive drift estimation for nonparametric diffusion model. *The Annals of Statistics*, 28:815–836, 2000.

[105] K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community. *Arxiv preprint cond-mat/0101130*, 2001.

[106] M.F. Tappen and W.T. Freeman. Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 900–906. Ieee, 2003.

[107] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[108] M.J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *Information Theory, IEEE Transactions on*, 55(12):5728–5741, 2009.

[109] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, 2009.

[110] M.J. Wainwright, P. Ravikumar, and J.D. Lafferty. High-Dimensional Graphical Model Selection Using l˜ 1-Regularized Logistic Regression. *Advances in Neural Information Processing Systems*, 19:1465, 2007.

[111] M. Weigt, R.A. White, H. Szurmant, J.A. Hoch, and T. Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72, 2009.

[112] J. Woods. Markov image modeling. *Automatic Control, IEEE Transactions on*, 23(5):846–850, 1978.

[113] X. Wu, R. Li, A.S. Fleisher, E.M. Reiman, X. Guan, Y. Zhang, K. Chen, and L. Yao. Altered default mode network connectivity in alzheimer's diseaseâĂŤa resting functional mri and bayesian network study. *Human brain mapping*, 32(11):1868–1881, 2011.

[114] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[115] T. Zhang. Some sharp performance bounds for least squares regression with L1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.

[116] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(2):2541, 2007.

[117] K. Zhou, J.C. Doyle, and K. Glover. *Robust and optimal control*. Prentice Hall, 1996.

[118] D. Zobin. Critical behavior of the bond-dilute two-dimensional ising model. *Physical Review B*, 18(5):2387, 1978.