Learning Networks of Stochastic Differential Equations

José Bento
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
jbento@stanford.edu

Morteza Ibrahimi
Department of Electrical Engineering
Stanford University
Stanford, CA 94305
ibrahimi@stanford.edu

Andrea Montanari
Department of Electrical Engineering and Statistics
Stanford University
Stanford, CA 94305
montanari@stanford.edu

Abstract

We consider linear models for stochastic dynamics. To any such model can be associated a network (namely a directed graph) describing which degrees of freedom interact under the dynamics. We tackle the problem of learning such a network from observation of the system trajectory over a time interval T.

We analyze the ℓ_1-regularized least squares algorithm and, in the setting in which the underlying network is sparse, we prove performance guarantees that are *uniform in the sampling rate* as long as this is sufficiently high. This result substantiates the notion of a well defined 'time complexity' for the network inference problem.

keywords: Gaussian processes, model selection and structure learning, graphical models, sparsity and feature selection.

1 Introduction and main results

Let $G = (V, E)$ be a directed graph with weight $A^0_{ij} \in \mathbb{R}$ associated to the directed edge (j, i) from $j \in V$ to $i \in V$. To each node $i \in V$ in this network is associated an independent standard Brownian motion b_i and a variable x_i taking values in $\mathbb{R}$ and evolving according to

$$\mathrm{d}x_i(t) = \sum_{j \in \partial_+ i} A^0_{ij} x_j(t)\, \mathrm{d}t + \mathrm{d}b_i(t)\,,$$

where $\partial_+ i = \{j \in V : (j, i) \in E\}$ is the set of 'parents' of i. Without loss of generality we shall take $V = [p] \equiv \{1, \dots, p\}$. In words, the rate of change of x_i is given by a weighted sum of the current values of its neighbors, corrupted by white noise. In matrix notation, the same system is then represented by

$$\mathrm{d}x(t) = A^0 x(t)\, \mathrm{d}t + \mathrm{d}b(t)\,, \qquad (1)$$

with $x(t) \in \mathbb{R}^p$, $b(t)$ a p-dimensional standard Brownian motion and $A^0 \in \mathbb{R}^{p\times p}$ a matrix with entries $\{A^0_{ij}\}_{i,j\in[p]}$ whose sparsity pattern is given by the graph G. We assume that the linear system $\dot{x}(t) = A^0 x(t)$ is stable (i.e. that the spectrum of A^0 is contained in $\{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}$). Further, we assume that $x(t = 0)$ is in its stationary state. More precisely, $x(0)$ is a Gaussian random variable independent of $b(t)$, distributed according to the invariant measure. Under the stability assumption, this a mild restriction, since the system converges exponentially to stationarity.

A portion of time length T of the system trajectory $\{x(t)\}_{t\in[0,T]}$ is observed and we ask under which conditions these data are sufficient to reconstruct the graph G (i.e., the sparsity pattern of A^0). We are particularly interested in computationally efficient procedures, and in characterizing the scaling of the learning time for large networks. Can the network structure be learnt in a time scaling linearly with the number of its degrees of freedom?

As an example application, chemical reactions can be conveniently modeled by systems of non-linear stochastic differential equations, whose variables encode the densities of various chemical species [1, 2]. Complex biological networks might involve hundreds of such species [3], and learning stochastic models from data is an important (and challenging) computational task [4]. Considering one such chemical reaction network in proximity of an equilibrium point, the model (1) can be used to trace fluctuations of the species counts with respect to the equilibrium values. The network G would represent in this case the interactions between different chemical factors. Work in this area focused so-far on low-dimensional networks, i.e. on methods that are guaranteed to be correct for fixed p, as $T \to \infty$, while we will tackle here the regime in which both p and T diverge.

Before stating our results, it is useful to stress a few important differences with respect to classical graphical model learning problems:

(i) Samples are not independent. This can (and does) increase the sample complexity.

(ii) On the other hand, infinitely many samples are given as data (in fact a collection indexed by the continuous parameter $t \in [0, T]$). Of course one can select a finite subsample, for instance at regularly spaced times $\{x(i\,\eta)\}_{i=0,1,\dots}$. This raises the question as to whether the learning performances depend on the choice of the spacing η.

(iii) In particular, one expects that choosing η sufficiently large as to make the configurations in the subsample approximately independent can be harmful. Indeed, the matrix A^0 contains more information than the stationary distribution of the above process (1), and only the latter can be learned from independent samples.

(iv) On the other hand, letting $\eta \to 0$, one can produce an arbitrarily large number of distinct samples. However, samples become more dependent, and intuitively one expects that there is limited information to be harnessed from a given time interval T.

Our results confirm in a detailed and quantitative way these intuitions.

1.1 Results: Regularized least squares

Regularized least squares is an efficient and well-studied method for support recovery. We will discuss relations with existing literature in Section 1.3.

In the present case, the algorithm reconstructs independently each row of the matrix A^0. The $r^{\rm th}$ row, A^0_r, is estimated by solving the following convex optimization problem for $A_r \in \mathbb{R}^p$

$$\text{minimize} \quad \mathcal{L}(A_r; \{x(t)\}_{t\in[0,T]}) + \lambda \|A_r\|_1 \,, \tag{2}$$

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

$$\mathcal{L}(A_r; \{x(t)\}_{t\in[0,T]}) = \frac{1}{2T}\int_0^T (A_r^* x(t))^2 \, {\rm d}t - \frac{1}{T}\int_0^T (A_r^* x(t)) \, {\rm d}x_r(t) \,. \tag{3}$$

(Here and below M^* denotes the transpose of matrix/vector M.) To see that this likelihood function is indeed related to least squares, one can *formally* write $\dot{x}_r(t) = {\rm d}x_r(t)/{\rm d}t$ and complete the square for the right hand side of Eq. (3), thus getting the integral $\int (A_r^* x(t) - \dot{x}_r(t))^2 {\rm d}t - \int \dot{x}_r(t)^2 \, {\rm d}t$. The first term is a sum of square residuals, and the second is independent of A. Finally the ℓ_1 regularization term in Eq. (2) has the role of shrinking to 0 a subset of the entries A_{ij} thus effectively selecting the structure.

Let S^0 be the support of row A^0_r, and assume $|S^0| \le k$. We will refer to the vector ${\rm sign}(A^0_r)$ as to the *signed support* of A^0_r (where ${\rm sign}(0) = 0$ by convention). Let $\lambda_{\rm max}(M)$ and $\lambda_{\rm min}(M)$ stand for the maximum and minimum eigenvalue of a square matrix M respectively. Further, denote by $A_{\min}$ the smallest absolute value among the non-zero entries of row A^0_r.

When stable, the diffusion process (1) has a unique stationary measure which is Gaussian with covariance $Q^0 \in \mathbb{R}^{p\times p}$ given by the solution of Lyapunov's equation [5]

$$A^0Q^0 + Q^0(A^0)^* + I = 0. \tag{4}$$

Our guarantee for regularized least squares is stated in terms of two properties of the covariance Q^0 and one assumption on $\rho_{\min}(A^0)$ (given a matrix M, we denote by $M_{L,R}$ its submatrix $M_{L,R} \equiv (M_{ij})_{i\in L, j\in R}$):

(a) We denote by $C_{\min} \equiv \lambda_{\min}(Q^0_{S^0,S^0})$ the minimum eigenvalue of the restriction of Q^0 to the support S^0 and assume $C_{\min} > 0$.

(b) We define the incoherence parameter α by letting $\|\!| Q^0{}_{(S^0)^C,S^0}\left(Q^0{}_{S^0,S^0}\right)^{-1} |\!\|_\infty = 1-\alpha$, and assume $\alpha > 0$. (Here $\|\!| \cdot |\!\|_\infty$ is the operator sup norm.)

(c) We define $\rho_{\min}(A^0) = -\lambda_{\max}((A^0 + A^{0^*})/2)$ and assume $\rho_{\min}(A^0) > 0$. Note this is a stronger form of stability assumption.

Our main result is to show that there exists a well defined *time complexity*, i.e. a minimum time interval T such that, observing the system for time T enables us to reconstruct the network with high probability. This result is stated in the following theorem.

Theorem 1.1. *Consider the problem of learning the support S^0 of row A^0_r of the matrix A^0 from a sample trajectory $\{x(t)\}_{t\in[0,T]}$ distributed according to the model (1). If*

$$T > \frac{10^4k^2(k\,\rho_{\min}(A^0)^{-2} + A_{\min}^{-2})}{\alpha^2\rho_{\min}(A^0)C_{\min}^2}\,\log\left(\frac{4pk}{\delta}\right), \tag{5}$$

then there exists λ such that ℓ_1-regularized least squares recovers the signed support of A^0_r with probability larger than $1-\delta$. This is achieved by taking $\lambda = \sqrt{36\,\log(4p/\delta)/(T\alpha^2\rho_{\min}(A^0))}$.

The time complexity is logarithmic in the number of variables and polynomial in the support size. Further, it is roughly inversely proportional to $\rho_{\min}(A^0)$, which is quite satisfying conceptually, since $\rho_{\min}(A^0)^{-1}$ controls the relaxation time of the mixes.

1.2 Overview of other results

So far we focused on continuous-time dynamics. While, this is useful in order to obtain elegant statements, much of the paper is in fact devoted to the analysis of the following discrete-time dynamics, with parameter $\eta > 0$:

$$x(t) = x(t-1) + \eta A^0x(t-1) + w(t), \qquad t \in \mathbb{N}_0\,. \tag{6}$$

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

The parameter η has the meaning of a time-step size. The continuous-time model (1) is recovered, in a sense made precise below, by letting $\eta \to 0$. Indeed we will 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 previous section) is kept constant. For a formal statement we refer to Theorem 3.1. Theorem 1.1 is indeed proved by carefully controlling this limit. The mathematical challenge in this problem is related to the fundamental fact that the samples $\{x(t)\}_{0\leq t\leq n}$ are dependent (and strongly dependent as $\eta \to 0$).

Discrete time models of the form (6) can arise either because the system under study evolves by discrete steps, or because we are subsampling a continuous time system modeled as in Eq. (1). Notice that in the latter case the matrices A^0 appearing in Eq. (6) and (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.

1.3 Related work

A substantial amount of work has been devoted to the analysis of ℓ_1 regularized least squares, and its variants [6, 7, 8, 9, 10]. The most closely related results are the one concerning high-dimensional consistency for support recovery [11, 12]. Our proof follows indeed the line of work 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 corrupted by i.i.d. noise (since successive configurations are correlated) and therefore elementary concentration inequalities are not sufficient.

Learning sparse graphical models via ℓ_1 regularization is also a topic with significant literature. In the Gaussian case, the *graphical* LASSO was proposed to reconstruct the model from i.i.d. samples [13]. In the context of binary pairwise graphical models, Ref. [11] proves high-dimensional consistency of regularized logistic regression for structural learning, under a suitable irrepresentability conditions on a modified covariance. Also this paper focuses on i.i.d. samples.

Most of these proofs builds on the technique of [12]. 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 $\eta \to 0$ for $n\eta = T$ fixed. In particular, they do not allow to prove an analogous of our continuous time result, Theorem 1.1. A large part of our effort is devoted to producing more accurate probability estimates that capture the correct scaling for small η.

Similar issues were explored in the study of stochastic differential equations, whereby one is often interested in tracking some slow degrees of freedom while ‘averaging out’ the fast ones [14]. The relevance of this time-scale separation for learning was addressed in [15]. Let us however emphasize that these works focus once more on system with a fixed (small) number of dimensions p.

Finally, the related topic of learning graphical models for autoregressive processes was studied recently in [16, 17]. The convex relaxation proposed in these papers is different from the one developed here. Further, no model selection guarantee was proved in [16, 17].

2 Illustration of the main results

It might be difficult to get a clear intuition of Theorem 1.1, mainly because of conditions (a) and (b), which introduce parameters $C_{\min}$ and α. The same difficulty arises with analogous results on the high-dimensional consistency of the LASSO [11, 12]. In this section we provide concrete illustration both via numerical simulations, and by checking the condition on specific classes of graphs.

2.1 Learning the laplacian of graphs with bounded degree

Given a simple graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ on vertex set $\mathcal{V} = [p]$, its laplacian $\Delta^{\mathcal{G}}$ is the symmetric $p \times p$ matrix which is equal to the adjacency matrix of $\mathcal{G}$ outside the diagonal, and with entries $\Delta^{\mathcal{G}}_{ii} = -\deg(i)$ on the diagonal [18]. (Here $\deg(i)$ denotes the degree of vertex i.)

It is well known that $\Delta^{\mathcal{G}}$ is negative semidefinite, with one eigenvalue equal to 0, whose multiplicity is equal to the number of connected components of $\mathcal{G}$. The matrix $A^0 = -m\,I + \Delta^{\mathcal{G}}$ fits into the setting of Theorem 1.1 for $m > 0$. The corresponding model (1.1) 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. We obtain the following result.

Theorem 2.1. *Let $\mathcal{G}$ be a simple connected graph of maximum vertex degree k and consider the model (1.1) with $A^0 = -m\,I + \Delta^{\mathcal{G}}$ where $\Delta^{\mathcal{G}}$ is the laplacian of $\mathcal{G}$ and $m > 0$. If*

$$T \geq 2 \cdot 10^5 k^2 \left(\frac{k+m}{m}\right)^5 (k+m^2)\,\log\left(\frac{4pk}{\delta}\right), \tag{7}$$

then there exists λ such that ℓ_1-regularized least squares recovers the signed support of A^0_r with probability larger than $1-\delta$. This is achieved by taking $\lambda = \sqrt{36(k+m)^2\log(4p/\delta)/(Tm^3)}$.

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 system size. Notice that once the graph is known, the laplacian $\Delta^{\mathcal{G}}$ is uniquely determined. Also, the proof technique used for this example is generalizable to other graphs as well.

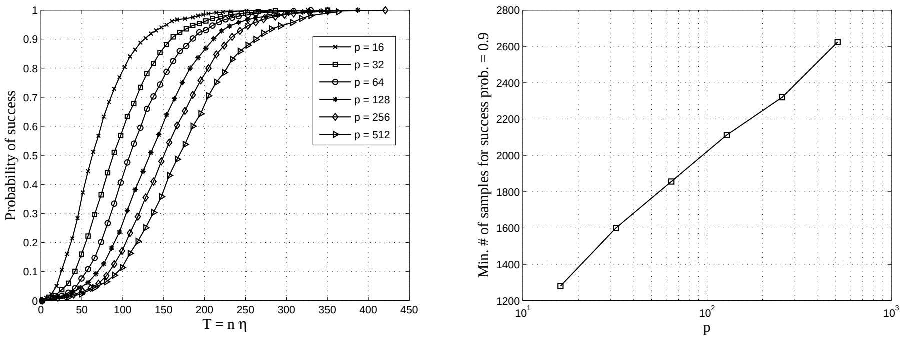

Figure 1: (left) Probability of success vs. length of the observation interval $n\eta$. (right) Sample complexity for 90% probability of success vs. p.

2.2 Numerical illustrations

In this section we present numerical validation of the proposed method on synthetic data. The results confirm our observations in Theorems 1.1 and 3.1, below, namely that the time complexity scales logarithmically with the number of nodes in the network p, given a constant maximum degree. Also, the time complexity is roughly independent of the sampling rate. In Fig. 1 and 2 we consider the discrete-time setting, generating data as follows. We draw A^0 as a random sparse matrix in $\{0,1\}^{p\times p}$ with elements chosen independently at random with $\mathbb{P}(A^0_{ij} = 1) = k/p$, $k = 5$. The process $x_0^n \equiv \{x(t)\}_{0\le t\le n}$ is then generated according to Eq. (6). We solve the regularized least square problem (the cost function is given explicitly in Eq. (8) for the discrete-time case) for different values of n, the number of observations, and record if the correct support is recovered for a random row r using the optimum value of the parameter λ. An estimate of the probability of successful recovery is obtained by repeating this experiment. Note that we are estimating here an average probability of success over randomly generated matrices.

The left plot in Fig.1 depicts the probability of success vs. $n\eta$ 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 A^0. The right plot in Fig.1 is the corresponding curve of the sample complexity vs. p where sample complexity is defined as the minimum value of $n\eta$ with probability of success of 90%. As predicted by Theorem 2.1 the curve shows the logarithmic scaling of the sample complexity with p.

In Fig. 2 we turn to the continuous-time model (1). Trajectories are generated by discretizing this stochastic differential equation with step δ much smaller than the sampling rate η. We draw random matrices A^0 as above and plot the probability of success for $p = 16$, $k = 4$ and different values of η, as a function of T. We used 2^{11} instances for each curve. As predicted by Theorem 1.1, for a fixed observation interval T, the probability of success converges to some limiting value as $\eta \to 0$.

3 Discrete-time model: Statement of the results

Consider a system evolving in discrete time according to the model (6), and let $x_0^n \equiv \{x(t)\}_{0\le t\le n}$ be the observed portion of the trajectory. The r^{th} row A^0_r is estimated by solving the following convex optimization problem for $A_r \in \mathbb{R}^p$

$$\text{minimize} \quad L(A_r; x_0^n) + \lambda \|A_r\|_1 \,, \tag{8}$$

where

$$L(A_r; x_0^n) \equiv \frac{1}{2\eta^2 n} \sum_{t=0}^{n-1} \{x_r(t+1) - x_r(t) - \eta\, A_r^* x(t)\}^2 \ . \tag{9}$$

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). Indeed the proof of Theorem 1.1 will amount to a more precise version of this statement. Furthermore, $L(A_r; x_0^n)$ is easily seen to be the log-likelihood of A_r within model (6).

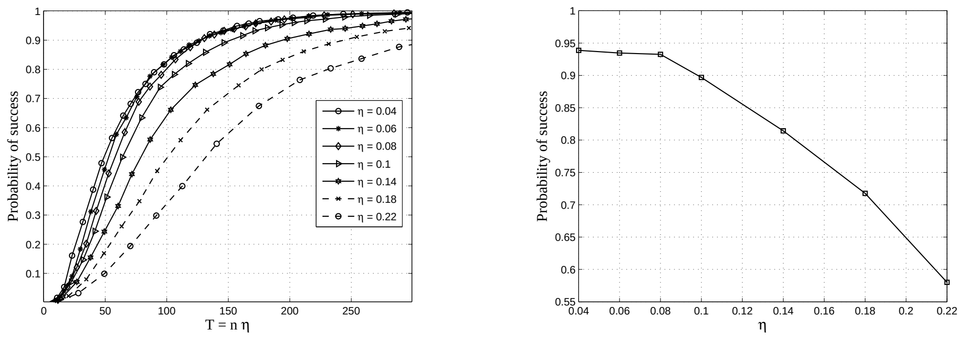

Figure 2: (right)Probability of success vs. length of the observation interval $n\eta$ for different values of η. (left) Probability of success vs. η for a fixed length of the observation interval, $(n\eta = 150)$. The process is generated for a small value of η and sampled at different rates.

As before, we let S^0 be the support of row A^0_r, and assume $|S^0| \leq k$. Under the model (6) $x(t)$ has a Gaussian stationary state distribution with covariance Q^0 determined by the following modified Lyapunov equation

$$A^0Q^0 + Q^0(A^0)^* + \eta A^0Q^0(A^0)^* + I = 0\,. \tag{10}$$

It will be clear from the context whether A^0/Q^0 refers to the dynamics/stationary matrix from the continuous or discrete time system. We assume conditions (a) and (b) introduced in Section 1.1, and adopt the notations already introduced there. We use as a shorthand notation $\sigma_{\max} \equiv \sigma_{\max}(I+\eta\, A^0)$ where $\sigma_{\max}(.)$ is the maximum singular value. Also define $D \equiv \big(1 - \sigma_{\max}\big)/\eta$. We will assume $D > 0$. As in the previous section, we assume the model (6) is initiated in the stationary state.

Theorem 3.1. *Consider the problem of learning the support S^0 of row A^0_r from the discrete-time trajectory $\{x(t)\}_{0\leq t\leq n}$. If*

$$n\eta > \frac{10^4 k^2 (kD^{-2} + A_{\min}^{-2})}{\alpha^2 D C_{\min}^2} \log\left(\frac{4pk}{\delta}\right), \tag{11}$$

then there exists λ such that ℓ_1-regularized least squares recovers the signed support of A^0_r with probability larger than $1-\delta$. This is achieved by taking $\lambda = \sqrt{(36\ \log(4p/\delta))/(D\alpha^2 n\eta)}$.

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 $[0, T]$.

4 Proofs

In the following we denote by $X \in \mathbb{R}^{n\times p}$ the matrix whose $(t+1)^{\rm th}$ column corresponds to the configuration $x(t)$, i.e. $X = [x(0), x(1), \dots, x(n-1)]$. Further $\Delta X \in \mathbb{R}^{n\times p}$ is the matrix containing configuration changes, namely $\Delta X = [x(1) - x(0), \dots, x(n) - x(n-1)]$. Finally we write $W = [w(1), \dots, w(n-1)]$ for the matrix containing the Gaussian noise realization. Equivalently,

$$W = \Delta X - \eta A\, X\,.$$

The $r^{\rm th}$ row of W is denoted by W_r.

In order to lighten the notation, we will omit the reference to x_0^n in the likelihood function (9) and simply write $L(A_r)$. We define its normalized gradient and Hessian by

$$\widehat{G} = -\nabla L(A^0_r) = \frac{1}{n\eta} X W_r^*\,, \qquad \widehat{Q} = \nabla^2 L(A^0_r) = \frac{1}{n} X X^*\,. \tag{12}$$

4.1 Discrete time

In this Section we outline our prove for our main result for discrete-time dynamics, i.e., Theorem 3.1. 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 finally we sketch the outline of the proof.

As mentioned, the proof strategy, and in particular the following proposition which provides a compact set of sufficient conditions for the support to be recovered correctly is analogous to the one in [12]. A proof of this proposition can be found in the supplementary material.

Proposition 4.1. *Let $\alpha, C_{\min} > 0$ be be defined by*

$$\lambda_{\min}(Q^0_{S^0,S^0}) \equiv C_{\min}\,, \qquad |||Q^0_{(S^0)^C,S^0}\left(Q^0_{S^0,S^0}\right)^{-1}|||_\infty \equiv 1-\alpha\,. \tag{13}$$

If the following conditions hold then the regularized least square solution (8) *correctly recover the signed support* $\mathrm{sign}(A^0_r)$*:*

$$\|\widehat{G}\|_\infty \le \frac{\lambda\alpha}{3}\,, \qquad \|\widehat{G}_{S^0}\|_\infty \le \frac{A_{\min}C_{\min}}{4k} - \lambda, \tag{14}$$

$$|||\widehat{Q}_{(S^0)^C,S^0} - Q^0_{(S^0)^C,S^0}|||_\infty \le \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}\,, \qquad |||\widehat{Q}_{S^0,S^0} - Q^0_{S^0,S^0}|||_\infty \le \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}\,. \tag{15}$$

Further the same statement holds for the continuous model 3, provided $\widehat{G}$ and $\widehat{Q}$ are the gradient and the hessian of the likelihood (3).

The proof of Theorem 3.1 consists in checking that, under the hypothesis (11) on the number of consecutive configurations, conditions (14) to (15) will 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 $\mathbb{E}\{\widehat{G}\} = 0$, $\mathbb{E}\{\widehat{Q}\} = Q^0$.

4.1.1 Technical lemmas

In this section we will state the necessary concentration lemmas for proving Theorem 3.1. These are non-trivial because $\widehat{G}$, $\widehat{Q}$ are quadratic functions of *dependent* random variables $\big($the samples $\{x(t)\}_{0\le t\le n}\big)$. The proofs of Proposition 4.2, of Proposition 4.3, and Corollary 4.4 can be found in the supplementary material provided.

Our first Proposition implies concentration of $\widehat{G}$ around 0.

Proposition 4.2. *Let $S \subseteq [p]$ be any set of vertices and $\epsilon < 1/2$. If $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\, A^0) < 1$, then*

$$\mathbb{P}\big\{\|\widehat{G}_S\|_\infty > \epsilon\big\} \le 2|S|\, e^{-n(1-\sigma_{\max})\,\epsilon^2/4}. \tag{16}$$

We furthermore need to bound the matrix norms as per (15) in proposition 4.1. First we relate bounds on $|||\widehat{Q}_{JS} - Q^0{}_{JS}|||_\infty$ with bounds on $|\widehat{Q}_{ij} - Q^0_{ij}|$, $(i \in J, i \in S)$ where J and S are any subsets of $\{1, ..., p\}$. We have,

$$\mathbb{P}(|||\widehat{Q}_{JS} - Q^0_{JS})|||_\infty > \epsilon) \le |J||S| \max_{i,j\in J} \mathbb{P}(|\widehat{Q}_{ij} - Q^0_{ij}| > \epsilon/|S|). \tag{17}$$

Then, we bound $|\widehat{Q}_{ij} - Q^0_{ij}|$ using the following proposition

Proposition 4.3. *Let $i, j \in \{1, ..., p\}$, $\sigma_{\max} \equiv \sigma_{max}(I + \eta A^0) < 1$, $T = \eta n > 3/D$ and $0 < \epsilon < 2/D$ where $D = (1 - \sigma_{\max})/\eta$ then,*

$$\mathbb{P}(|\widehat{Q}_{ij} - Q^0_{ij})| > \epsilon) \le 2e^{-\frac{n}{32\eta^2}(1-\sigma_{\max})^3\epsilon^2}. \tag{18}$$

Finally, the next corollary follows from Proposition 4.3 and Eq. (17).

Corollary 4.4. *Let J, S ($|S| \le k$) be any two subsets of $\{1, ..., p\}$ and $\sigma_{\max} \equiv \sigma_{\max}(I + \eta A^0) < 1$, $\epsilon < 2k/D$ and $n\eta > 3/D$ (where $D = (1 - \sigma_{\max})/\eta$) then,*

$$\mathbb{P}(|||\widehat{Q}_{JS} - Q^0_{JS}|||_\infty > \epsilon) \le 2|J|ke^{-\frac{n}{32k^2\eta^2}(1-\sigma_{\max})^3\epsilon^2}. \tag{19}$$

4.1.2 Outline of the proof of Theorem 3.1

With these concentration bounds we can now easily prove Theorem 3.1. All we need to do is to compute the probability that the conditions given by Proposition 4.1 hold. From the statement of the theorem we have that the first two conditions ($\alpha, C_{\min} > 0$) of Proposition 4.1 hold. In order to make the first condition on $\widehat{G}$ imply the second condition on $\widehat{G}$ we assume that $\lambda\alpha/3 \leq (A_{\min}C_{\min})/(4k) - \lambda$ which is guaranteed to hold if

$$\lambda \leq A_{\min}C_{\min}/8k. \tag{20}$$

We also combine the two last conditions on $\widehat{Q}$, thus obtaining the following

$$|\!|\!|\widehat{Q}_{[p],S^0} - Q^0_{[p],S^0}|\!|\!|_\infty \leq \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}\,, \tag{21}$$

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 4.2 and Corollary 4.4. It is shown in the supplementary material that this is satisfied if condition (11) holds.

4.2 Outline of the proof of Theorem 1.1

To prove Theorem 1.1 we recall that Proposition 4.1 holds provided the appropriate continuous time expressions are used for $\widehat{G}$ and $\widehat{Q}$, namely

$$\widehat{G} = -\nabla\mathcal{L}(A^0_r) = \frac{1}{T}\int_0^T x(t)\,\mathrm{d}b_r(t)\,, \qquad \widehat{Q} = \nabla^2\mathcal{L}(A^0_r) = \frac{1}{T}\int_0^T x(t)x(t)^*\,\mathrm{d}t\,. \tag{22}$$

These are of course random variables. In order to distinguish these from the discrete time version, we will 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 (5), it is easy to show that (11) holds for all $n > n_0$ with n_0 a sufficiently large constant (for a proof see the provided supplementary material). Therefore, by the proof of Theorem 3.1, the conditions in Proposition 4.1 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 (6) will be denoted by $x(i)$ where $i \in \mathbb{N}$ and the state of continuous time system (1) by $x(t)$ where $t \in \mathbb{R}$. We denote by Q^0 the solution of (4) and by $Q^0(\eta)$ the solution of (10). It is easy to check that $Q^0(\eta) \to Q^0$ as $\eta \to 0$ by the uniqueness of stationary state distribution.

The initial state of the continuous time system $x(t=0)$ is a $\mathsf{N}(0, Q^0)$ random variable independent of $b(t)$ and the initial state of the discrete time system is defined to be $x(i=0) = (Q^0(\eta))^{1/2}(Q^0)^{-1/2}x(t=0)$. At subsequent times, $x(i)$ and $x(t)$ are assumed are generated by the respective dynamical systems using the same matrix A^0 using common randomness provided by the standard Brownian motion $\{b(t)\}_{0\leq t\leq T}$ in $\mathbb{R}^p$. In order to couple $x(t)$ and $x(i)$, we construct $w(i)$, the noise driving the discrete time system, by letting $w(i) \equiv (b(Ti/n) - b(T(i-1)/n))$.

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.

Acknowledgments

This work was partially supported by a Terman fellowship, the NSF CAREER award CCF-0743978 and the NSF grant DMS-0806211 and by a Portuguese Doctoral FCT fellowship.

References

[1] D.T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[2] D. Higham. Modeling and Simulating Chemical Reactions. *SIAM Review*, 50:347–368, 2008.

[3] N.D.Lawrence et al., editor. *Learning and Inference in Computational Systems Biology*. MIT Press, 2010.

[4] T. Toni, D. Welch, N. Strelkova, A. Ipsen, and M.P.H. Stumpf. Modeling and Simulating Chemical Reactions. *J. R. Soc. Interface*, 6:187–202, 2009.

[5] K. Zhou, J.C. Doyle, and K. Glover. *Robust and optimal control*. Prentice Hall, 1996.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[7] 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.

[8] 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.

[9] T. Zhang. Some sharp performance bounds for least squares regression with L1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.

[10] 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.

[11] 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.

[12] P. Zhao and B. Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

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

[14] K. Ball, T.G. Kurtz, L. Popovic, and G. Rempala. Modeling and Simulating Chemical Reactions. *Ann. Appl. Prob.*, 16:1925–1961, 2006.

[15] G.A. Pavliotis and A.M. Stuart. Parameter estimation for multiscale diffusions. *J. Stat. Phys.*, 127:741–781, 2007.

[16] J. Songsiri, J. Dahl, and L. Vandenberghe. Graphical models of autoregressive processes. pages 89–116, 2010.

[17] J. Songsiri and L. Vandenberghe. Topology selection in graphical models of autoregressive processes. *Journal of Machine Learning Research*, 2010. submitted.

[18] F.R.K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics, 1997.

[19] P. Ravikumar, M.J. Wainwright, and J. Lafferty. High-dimensional Ising model selection using l1-regularized logistic regression. *Annals of Statistics*, 2008.