Information Theoretic Limits on Learning Stochastic Differential Equations

José Bento and Morteza Ibrahimi
Department of Electrical Engineering
Stanford University

Andrea Montanari
Department of Electrical Engineering and
Department of Statistics
Stanford University

***Abstract*—Consider the problem of learning the drift coefficient of a stochastic differential equation from a sample path. In this paper, we assume that the drift is parametrized by a high-dimensional vector. We address the question of how long the system needs to be observed in order to learn this vector of parameters. We prove a general lower bound on this time complexity by using a characterization of mutual information as time integral of conditional variance, due to Kadota, Zakai, and Ziv. This general lower bound is applied to specific classes of linear and non-linear stochastic differential equations. In the linear case, the problem under consideration is the one of learning a matrix of interaction coefficients. We evaluate our lower bound for ensembles of sparse and dense random matrices. The resulting estimates match the qualitative behavior of upper bounds achieved by computationally efficient procedures.**

I. Introduction

Consider a continuous-time stochastic process $\{x_t\}_{t\geq 0}$, that is defined by a stochastic differential equation (SDE) of the form

$$\mathrm{d}x_t = F(x_t; A)\,\mathrm{d}t + \mathrm{d}b_t\,, \tag{1}$$

where $x_t \in \mathbb{R}^p$, b_t is a p-dimensional standard Brownian motion and the *drift coefficient* $F(x_t; A) = [F_1(x_t; A), ..., F_p(x_t; A)] \in \mathbb{R}^p$, is a function of x_t parametrized by A, which is an unknown high-dimensional vector.

In this paper we consider the problem of learning information about the vector of parameters A from the observation of a sample trajectory $X^T \equiv \{x_t\}_{t=0}^T$. More precisely, we consider the high dimensional case (where the dimensions of A and x_t are large) and investigate what is the minimum time length T we need to observe the system in order to be able to recover A, with some confidence.

Models based on SDE's play a crucial role in several domains of science and technology, ranging from chemistry to finance. As an example, gene regulatory networks can be modeled by systems of non-linear stochastic differential equations, whose variables encode concentrations of certain gene expression products (e.g. proteins) [1]. Complex chemical networks are also described by SDE's that can involve hundreds of reactants [2], [3]. The problem of learning the parameters (reaction coefficients) of such an SDE or simply reconstructing the underlying network structure (i.e. which parameters are non-vanishing) plays crucial role in this context [4].

An important subclass of models consists in linear SDE's, whereby the drift is a linear function of x_t, namely $F(x_t; A) = Ax_t$ with $A \in \mathbb{R}^{p\times p}$. This can be a good approximation for many systems near a stable equilibrium. Linear SDE's are a special case of a broader class for which the drift is a linear combination of a finite set of basis functions $F(x_t) = [f_1(x_t), f_1(x_t), \ldots, f_m(x_t)]$, with $f_i : \mathbb{R}^p \to \mathbb{R}$. The drift is then given as $F(x_t; A) = AF(x_t)$, with $A \in \mathbb{R}^{p\times m}$. As an example, within models of chemical reactions, the drift 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 noise term affecting the measurement of x_C. In order to learn a model of this type, one can consider a basis of functions that contain all monomials up to a maximum degree.

A. Illustration

As an illustration, consider a system of p masses in $\mathbb{R}^d$ connected by springs. 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 and unit elastic coefficients, the dynamics of this system in the presence of external noisy forces can be modeled by the following damped Newton equations

$$\mathrm{d}v_t = -\gamma v_t \mathrm{d}t - \nabla U(q_t)\,\mathrm{d}t + \sigma\,\mathrm{d}b_t, \tag{2}$$

$$\mathrm{d}q_t = v_t \mathrm{d}t\,, \tag{3}$$

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

where $q_t = (q_t^{(1)}, \ldots, q_t^{(p)})$, $v_t = (v_t^{(1)}, \ldots, v_t^{(p)})$, and $q_t^{(i)}, v_t^{(i)} \in \mathbb{R}^d$ denote the position and velocity of mass i at time t. This system of SDE's can be written in the form (1) by letting $x_t = [q_t, v_t]$ and $A = [C^0, D^0]$. A straightforward calculation shows that the drift $F(x_t; A)$ can be further written as a linear combination of the following basis of non-linear functions

$$F(x_t) = \Big[\{v_t^{(i)}\}_{i\in[p]}, \{\Delta_t^{(ij)}\}_{i,j\in[p]}, \Big\{\frac{\Delta_t^{(ij)}}{\|\Delta_t^{(ij)}\|}\Big\}_{i,j\in[p]}\Big], \tag{4}$$

where $\Delta_t^{(ij)} = q_t^{(i)} - q_t^{(j)}$ and $[p] = \{1, \ldots, p\}$. In many situations only specific properties of the parameters are of

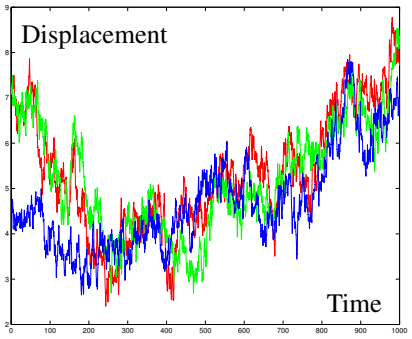

Fig. 1. Evolution of the horizontal component of the position of three masses in a system with $p = 36$ masses interacting via elastic springs (cf. Fig. 2 for the network structure). The time interval is here $T = 1000$. All the springs have rest length $D_{ij} = 1$, the damping coefficient is $\gamma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

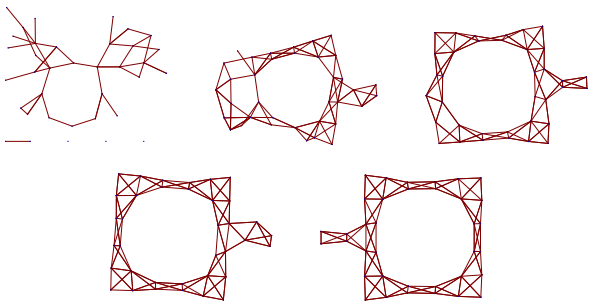

Fig. 2. From left to right and top to bottom: structures reconstructed using the algorithm of [5] with observation time $T = 500$, 1500, 2500, 3500 and 4500. For $T = 4500$ exact reconstruction is achieved.

interest, for instance one might be interested only in the network structure in the present example.

Figure 1 shows the trajectories of three masses in a two-dimensional network of 36 masses and 90 springs evolving according to Eq. (2) and Eq. (3). How long does one need to observe these (and the other masses) trajectories in order to learn the structure of the underlying network? Figure 2 reproduces the network structure reconstructed using the algorithm of [5] for increasing observation intervals T. The inferred structure converges to the actual one only if T is large enough.

B. Related Work

Over the last few years, a significant effort has been devoted to developing methods and sample complexity bounds for learning graphical models from data. Particular effort was devoted to learning sparse graphical models using convex regularizations that promote sparsity. Well known examples in the context of Gaussian graphical models include the *graphical* LASSO [6] and the pseudo-likelihood method of [7]. These papers assume that the data are i.i.d. samples from a high-dimensional Gaussian distribution. However in many cases samples are produced by an underlying dynamical process and the i.i.d. assumption is unrealistic.

In [5], a convex regularization method was developed to learn linear SDE's with a sparse network structure from data. The upper bounds on the sample complexity proved in [5] match in several cases the lower bounds developed here. The related topic of learning graphical models for autoregressive processes was studied recently in [8], [9]. These papers propose a convex relaxation different from the one of [5], without however developing estimates on the sample complexity for model selection.

Finally, a substantial literature addresses various questions related to learning SDE's [3], [10], [11]. However this line of work did not yield quantitative estimates on the scaling of sample complexity with the problem dimensionality.

II. Main Results

Without loss of generality, assume that the parameter A is a random variable chosen with some unknown prior distribution $\mathbb{P}_A$ (subscript will be often omitted). We are interested in a specific property of A that is given by a function $A \mapsto M(A)$. Unless specified otherwise $\mathbb{P}$ and $\mathbb{E}$ denote probability and expectation with respect to the joint law of $\{x_t\}_{t\geq 0}$ and A. As mentioned above $X^T \equiv \{x_t\}_{0\leq t\leq T}$ will denote 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.,

$$\mathrm{Var}_{A|X^t}(F(x_t;A)) = \sum_{i=1}^{p} \mathrm{Var}_{A|X^t}(F_i(x_t;A)). \qquad (5)$$

Our main tool is the following general lower bound, that follows from an identity between mutual information and the integral of conditional variance proved by Kadota, Zakai and Ziv [12].

Theorem II.1. *Let $\widehat{M}_T(X^T)$ be an estimator of $M(A)$ based on X^T. If $\mathbb{P}(\widehat{M}_T(X^T) \neq M(A)) < \frac{1}{2}$ then*

$$T \geq \frac{H(M(A)) - 2I(A;x_0)}{\frac{1}{T}\int_0^T \mathbb{E}_{X^t}\{\mathrm{Var}_{A|X^t}(F(x_t;A))\}\mathrm{d}t}\,. \qquad (6)$$

Proof: Equation (1) can be regarded as describing a white Gaussian channel with feedback where A denotes the message to be transmitted. For this scenario, Kadota et al. [12] give the following identity for the mutual information between X^T and A when the initial condition is $x_0 = 0$,

$$I(X^T;A) = \frac{1}{2}\int_0^T \mathbb{E}_{X^t}\{\mathrm{Var}_{A|X^t}(F(x_t;A))\}\mathrm{d}t. \qquad (7)$$

For the general case where $x_0 \neq 0$ and might depend on A (if for example x_0 is the stationary state of the system) we can write $I(X^T;A) = I(x_0;A) + I(X^T;A|x_0)$ and apply the previous identity to $I(X^T;A|x_0)$. Taking into account that $I(\widehat{M}_T(X^T)); M(A)) \leq I(X^T;A)$ and making use of Fano's inequality $I(\widehat{M}_T(X^T)); M(A)) \geq \mathbb{P}(\widehat{M}_T(X^T) = M(A))H(\widehat{M}_T(X^T)))$ the results follows. ∎

The bound in Theorem II.1 is often too complex to be evaluated. Instead, the following corollary provides a more easily computable bound.

Corollary II.2. *Assume that the process $\{x_t\}_{t\geq 0}$ is stationary. Let $\widehat{M}_T(X^T)$ be an estimator of $M(A)$ based on X^T. If $\mathbb{P}(\widehat{M}_T(X^T) \neq M(A)) < \frac{1}{2}$ then*

$$T \geq \frac{H(M(A)) - 2I(A; x_0)}{\mathbb{E}_{x_0}\{\mathrm{Var}_{A|x_0}(F(x_0; A))\}}. \tag{8}$$

Proof: Since conditioning reduces variance, we have $\mathbb{E}_{X^t}\{\mathrm{Var}_{A|X^t}(F(x_t; A))\} \leq \mathbb{E}_{x_t}\{\mathrm{Var}_{A|x_t}(F(x_t; A))\}$. Using stationarity, we have $\mathbb{E}_{x_t}\{\mathrm{Var}_{A|x_t}(F(x_t; A))\} = \mathbb{E}_{x_0}\{\mathrm{Var}_{A|x_0}(F(x_t; A))\}$, which simplifies (6) to (8). ∎

In the rest of this section, we apply this lower bound to special classes of SDE's. In all of our applications it is understood that the process $\{x_t\}_{t\geq 0}$ is stationary.

A. Learning Sparse Linear SDE's

Consider the linear SDE,

$$\mathrm{d}x_t = Ax_t\mathrm{d}t + \mathrm{d}b_t. \tag{9}$$

The goal is to learn the interaction matrix $A \in \mathbb{R}^{p\times p}$. The first two theorems stated below provide lower bounds for sample complexity T, for the two regimes of sparse and dense matrices. Throughout this paper Q^* will denote the transpose of matrix Q. Given a matrix Q, its $\mathrm{supp}(Q)$ is the $0-1$ matrix such that $\mathrm{supp}(Q)_{ij} = 1$ if and only if $Q_{ij} \neq 0$. Its 'signed support' $\mathrm{sign}(Q)$ is the matrix such that $\mathrm{sign}(Q)_{ij} = \mathrm{sign}(Q_{ij})$ if $Q_{ij} \neq 0$ and $\mathrm{sign}(Q)_{ij} = 0$ otherwise.

Define the class of matrices $\mathcal{A}^{(S)} \subset \mathbb{R}^{p\times p}$ by letting $A \in \mathcal{A}^{(S)}$ if and only if

(i) A has at most k non-zero elements per row, $k \geq 3$,
(ii) $\min_{ij} |A_{ij}| > a_{\min}$,
(ii) Letting $\lambda_{\min}(Q)$ denote the smallest eigenvalue of matrix Q, $\lambda_{\min}(-(A + A^*)/2) \geq \rho > 0$.

The next theorem provides a lower bound on the time complexity of learning the signed support of models from the class $\mathcal{A}^{(S)}$.

Theorem II.3. *Let $M(A) = \mathrm{sign}(A)$ be the signed support of A and $\widehat{M}_T(X^T)$ an estimator of $M(A)$ based on X^T. There is a constant $C(k)$ such that, for all p large enough, if $\sup_{A\in\mathcal{A}^{(S)}} \mathbb{P}_{X^T|A}(M(A) \neq \widehat{M}_T(X^T)) < \frac{1}{2}$ then*

$$T > \frac{C(k)}{a_{\min}} \max\{\rho/a_{\min}, 1\} \log(p). \tag{10}$$

B. Learning Dense Linear SDE's

A different regime of interest in learning the network of interactions for a linear SDE's is the case of dense matrices. As we shall see shortly, 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 matrices with the following properties: $A \in \mathcal{A}^{(D)}$ if and only if,

(i) $a_{\min} \leq |A_{ij}|p^{1/2} \leq a_{\max}$.
(ii) $\lambda_{\min}(-(A + A^*)/2) \geq \rho > 0$.

The second theorem provides a lower bound for learning the signed support of models from class $\mathcal{A}^{(D)}$.

Theorem II.4. *Let $M(A) = \mathrm{sign}(A)$ be the signed support of A and $\widehat{M}_T(X^T)$ an estimator of $M(A)$ based on X^T. There exists a constant C such that, for all p large enough, if $\sup_{A\in\mathcal{A}^{(D)}} \mathbb{P}_{X^T|A}(M(A) \neq \widehat{M}_T(X^T)) < \frac{1}{2}$ then*

$$T > \frac{C}{a_{\min}} \max\{\rho/a_{\min}, 1\} p. \tag{11}$$

Together with the upper bounds from [5], Theorem II.3 establishes that the time complexity of learning sparse linear SDE's is $T = \Theta(\log(p))$. Further, this task can be performed efficiently using ℓ_1 penalized least squares [5]. On the other hand, Theorem II.4 implies a dramatic dichotomy. The time complexity of learning dense linear SDE's is at least linear in p (and indeed matching upper bounds can be proved in this case as well [13]).

C. Learning Non-Linear SDE's

In this section we assume that the observed samples X^T come from a stochastic process driven by a general SDE of the form (1).

In what follows, $v^{(i)}$ denotes the i^{th} component of vector v. For example, $x_2^{(3)}$ is the 3^{th} component of the vector x_t at time $t = 2$. $JF(\,\cdot\,; A) \in \mathbb{R}^{p\times p}$ will denote the Jacobian of the function $F(\,\cdot\,; A)$.

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

(i) the support of $JF(x; A)$ has at most k non-zero entries for every x,
(ii) the covariance matrix for the stationary process, Σ_∞, satisfies $\lambda_{\min}(\Sigma_\infty) \geq L$,
(iii) $\mathrm{Var}_{x_0|A}(x_0^{(i)}) \leq B\ \forall i$,
(iv) $|\partial F_i(x; A)/\partial x^{(j)}| \leq D$ for all $x \in \mathbb{R}^p\ i, j \in [p]$.

For simplicity we write $F(x; A) \in \mathcal{A}^{(N)}$ by $A \in \mathcal{A}^{(N)}$.

Theorem II.5. *Let $M(A)$ be the smallest support for which supp$(JF(x; A)) \subseteq M(A)\ \forall x$. If $\widehat{M}_T(X^T)$ is an estimator of $M(A)$ based on X^T and $\sup_{A\in\mathcal{A}^{(N)}} \mathbb{P}_{X^T|A}(\widehat{M}_T(X^T) \neq M(A)) < 1/2$ then*

$$T > \frac{k \log p/k - \log B/L}{C + 2k^2D^2B}. \tag{12}$$

In the above expression $C = \max_{i\in[p]} \mathbb{E}\{F_i(\mathbb{E}_{x_0|A}(x_0); A)\}$.

Remark II.1. *Note that the assumption that F is Lipschitz is not very strong in the sense that it is usually required for existence and uniqueness of a solution of the SDE* (1) *with finite expected energy, [14].*

III. Proofs and technical lemmas

In this section we prove Theorems II.3 to II.5. Throughout, $\{x_t\}_{t\geq 0}$ is assumed to be a stationary process. It is immediate to check that under the assumptions of the Theorems II.3 and II.4, the SDE admit a unique stationary measure, with bounded covariance. We let $\Sigma_\infty = \mathbb{E}\{x_0 x_0^*\} - \mathbb{E}\{x_0\}(\mathbb{E}\{x_0\})^* = \mathbb{E}\{x_t x_t^*\} - \mathbb{E}\{x_t\}(\mathbb{E}\{x_t\})^*$ denote this covariance.

A. 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 (9) with symmetric interaction matrix A.

Lemma III.1. *Assume that $\{x_t\}_{t\geq 0}$ is a stationary process generated by the linear SDE (9), with A symmetric. Let $\widehat{M}_T(X^T)$ be an estimator of $M(A)$ based on X^T. If $\mathbb{P}(\widehat{M}_T(X^T) \neq M(A)) < \frac{1}{2}$ then*

$$T \geq \frac{H(M(A)) - 2I(A;x_0)}{\frac{1}{2}\mathrm{Tr}\{\mathbb{E}\{-A\} - (\mathbb{E}\{-A^{-1}\})^{-1}\}\}}. \tag{13}$$

Proof: The bound follows from Corollary II.2 after showing that $\mathbb{E}_{x_0}\{\mathrm{Var}_{A|x_0}(Ax_0)) \leq (1/2)\mathrm{Tr}\{\mathbb{E}\{-A\} - (\mathbb{E}\{-A^{-1}\})^{-1}\}$. First note that

$$\mathbb{E}_{x_0}\{\mathrm{Var}_{A|x_0}(Ax_0)\} = \mathbb{E}_{x_0}||Ax_0 - \mathbb{E}_{A|x_0}(Ax_0|x_0)||_2^2. \tag{14}$$

The quantity in (14) can be thought of as the ℓ_2-norm error of estimating Ax_0 based on x_0, using $\mathbb{E}_{A|x_0}(Ax_0|x_0)$. Since conditional expectation is the minimal mean square error estimator, replacing $\mathbb{E}_{A|x_0}(Ax_0|x_0)$ by any estimator of Ax_0 based on x_0 gives an upper bound for the expression in (14). We choose as an estimator a linear estimator , i.e., an estimator in the form Bx_0 where $B = (\mathbb{E}_A A\Sigma_\infty)(\mathbb{E}_A \Sigma_\infty)^{-1}$,

$$\begin{aligned}&\mathbb{E}_{x_0}||Ax_0 - \mathbb{E}_{A|x_0}(Ax_0|x_0)||_2^2 \leq \mathbb{E}_{x_0}||Ax_0 - Bx_0||_2^2\\ &= \mathrm{Tr}\{\mathbb{E}\{Ax_0(x_0)^*A^*\}\} - 2\mathrm{Tr}\{B\mathbb{E}\{x_0(x_0)^*A^*\}\}\\ &+ \mathrm{Tr}\{B\mathbb{E}\{x_0(x_0)^*\}B^*\}.\end{aligned} \tag{15}$$

Furthermore, for a linear system, Σ_∞ satisfies the Lyapunov equation $A\Sigma_\infty + \Sigma_\infty A^* + I = 0$. For A symmetric, this implies $\Sigma_\infty = -(1/2)A^{-1}$. Substituting this expression in (14) and (15) finishes the proof. ∎

B. Proof of Theorem II.3

We prove the theorem by showing that the same complexity bound holds in the case when we are trying to estimate the signed support of A for an A 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 will exist $A \in \mathcal{A}^{(S)}$ for which the probability of error is biger than $1/2$. The complexity bound for random matrices A is proved using Lemma III.1.

In order to generate A at random we proceed as follows. Let G be the a random matrix constructed from the adjacency matrix of a uniformly random k-regular graph. Generate $\tilde{A}$ by flipping the sign of each non-zero entry in G with probability $1/2$ independently. We define A to be the random matrix $A = -(\gamma + 2a_{\min}\sqrt{k-1})I + a_{\min}\tilde{A}$ where $\gamma = \gamma(\tilde{A}) > 0$ is the smallest value such that the maximum eigenvalue of A is smaller than $-\rho$. This guarantees that all these A satisfy the four properties of the class $\mathcal{A}^{(S)}$.

The following lemma encapsulates the necessary random matrix calculations.

Lemma III.2. *Let A be a random matrix defined as above and*

$$Q(a_{\min}, k, \rho) \equiv \lim_{p\to\infty} \frac{1}{p}\{\mathrm{Tr}\{\mathbb{E}(-A)\} - \mathrm{Tr}\{(\mathbb{E}(-A^{-1}))^{-1}\}\}. \tag{16}$$

Then, there exists a constant C' only dependent on k such that

$$Q(a_{\min}, k, \rho) \leq \min\{\frac{C' k a_{\min}^2}{\rho}, \frac{k a_{\min}}{\sqrt{k-1}}\}. \tag{17}$$

Proof: First notice that

$$\lim_{p\to\infty} \frac{1}{p}\mathbb{E}\mathrm{Tr}\{-A\} = \lim_{p\to\infty} \mathbb{E}(\gamma) + 2a_{\min}\sqrt{k-1} \tag{18}$$

$$= \rho + 2a_{\min}\sqrt{k-1} \tag{19}$$

since by Kesten-McKay law [15], for large p, the spectrum of $\tilde{A}$ has support in $(-\epsilon - 2a_{\min}\sqrt{k-1}, 2a_{\min}\sqrt{k-1} + \epsilon)$ with high probability. Notice that unless we randomize each entry of $\tilde{A}$ with $\{-1,+1\}$ values, every $\tilde{A}$ will have k 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 $\mathbb{E}(-A^{-1})$. We can write,

$$\frac{1}{p}\mathrm{Tr}\{(\mathbb{E}(-A^{-1}))^{-1}\} = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i} \tag{20}$$

$$\geq \frac{1}{\frac{1}{p}\sum_{i=1}^{p}\lambda_i} = \frac{1}{\mathbb{E}\{\frac{1}{p}\mathrm{Tr}\{(-A)^{-1}\}\}} \tag{21}$$

where we applied Jensen's inequality in the last step. By Kesten-McKay law we now have that,

$$\lim_{p\to\infty}\mathbb{E}\{\frac{1}{p}\mathrm{Tr}\{(-A)^{-1}\}\} = \mathbb{E}\{\lim_{p\to\infty}\frac{1}{p}\mathrm{Tr}\{(-A)^{-1}\}\} \tag{22}$$

$$= \frac{1}{a_{\min}}G(k, \rho/a_{\min} + 2\sqrt{k-1}) \tag{23}$$

where

$$G(k,z) = \int \frac{-1}{\nu - z}\mathrm{d}\mu(\nu) \tag{24}$$

and

$$\mathrm{d}\mu(\nu) = \frac{k}{2\pi}\frac{\sqrt{4(k-1)-\nu^2}}{k^2-\nu^2}\mathrm{d}\nu \tag{25}$$

for $\nu \in [-2\sqrt{k-1}, -2\sqrt{k-1}]$ and zero otherwise. Expression (25) defines the Kesten-McKay distribution. Computing the above integral we obtain

$$G(k,z) = -\frac{(k-2)z - k\sqrt{-4k+z^2+4}}{2\,(z^2-k^2)} \tag{26}$$

whence

$$\lim_{\rho\to 0} Q(a_{\min},k,\rho) = \frac{a_{\min}k}{\sqrt{k-1}}, \quad (27)$$

$$\lim_{\rho\to\infty} \rho\, Q(a_{\min},k,\rho) = k(a_{\min})^2. \quad (28)$$

Since $Q(a_{\min},k,\rho)/a_{\min}$ is a function of k and $\rho/a_{\min}$ that is strictly decreasing with $\rho/a_{\min}$, the claimed bound follows. ∎

Proof (Theorem II.3): Starting from the bound of Lemma III.1, we divide both terms in the numerator and the denominator by p. The term $H(M(A))/p$ can be lower bounded by $p^{-1}\log\left(\binom{p}{k}2^k\right)^p \geq k\log(2p/k)$ and Lemma III.2 gives an upper bound on the denominator when $p\to\infty$. We now prove that $\lim_{p\to\infty} I(x_0;A)/p \leq 1$. This finishes the proof of Theorem II.3 since after multiplying by a small enough constant (only dependent on k) the bound obtained by replacing the numerator and denominator with these limits will be valid for all p large enough.

We start by writing,

$$I(x_0;A) = h(x_0) - h(x_0|A) \quad (29)$$

$$\leq \frac{1}{2}\log(2\pi e)^p|\mathbb{E}(\Sigma_\infty)| - \mathbb{E}\frac{1}{2}\log(2\pi e)^p|\Sigma_\infty|, \quad (30)$$

where $\Sigma_\infty = -(1/2)A^{-1}$ is the covariance matrix of the stationary process x_t and $|.|$ denotes the determinant of a matrix. Then we write,

$$I(x_0;A) \leq \frac{1}{2}\log|\mathbb{E}(-(\beta A)^{-1})| + \frac{1}{2}\mathbb{E}\log(|-\beta A|) \quad (31)$$

$$\leq \frac{1}{2}\mathrm{Tr}\mathbb{E}(-I-(\beta A)^{-1}) + \frac{1}{2}\mathbb{E}\mathrm{Tr}\{-I-\beta A\} \quad (32)$$

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 (18) and (22) it follows that,

$$\lim_{p\to\infty}\frac{1}{p}I(x_0;A) \leq -1 + (1/2)(\beta' z + \beta'^{-1}G(k,z)) \quad (33)$$

where $z = \rho/a_{\min} + 2\sqrt{k-1}$ and $\beta' = \beta a_{\min}$. Optimizing over β' and then over z gives,

$$\beta' z + \beta'^{-1}G(k,z) \leq 2\sqrt{zG(k,z)} \leq \sqrt{8}\sqrt{\frac{k-1}{k-2}} \leq 4, \quad (34)$$

which implies $\lim_{p\to\infty} I(x_0;A)/p \leq 1$. ∎

C. Proof of Theorem II.4: Outline

The proof of this theorem follows closely the proof of Theorem II.3. We will prove that same bound (11) holds for an A chosen at random with a distribution supported on $\mathcal{A}^{(D)}$, whence the claim follows. In order to lower bound the error probability for random matrices, we make use of Lemma III.1.

We construct the random matrix A as follows. Let $\tilde{A}$ be a random symmetric matrix with $\{A_{ij}\}_{i\leq j}$ i.i.d. random variables where $\mathbb{P}(A_{ij}=a_{\min}) = \mathbb{P}(A_{ij}=-a_{\min}) = 1/4$, and $\mathbb{P}(A_{ij}=0) = 1/2$. Notice that the second moment of each entry is $\mathbb{E}(A_{ij}^2) = a_{\min}^2/2 \equiv \alpha$. We then define $A = -(\gamma+2\sqrt{\alpha})I + \tilde{A}/\sqrt{p}$ where $\gamma = \gamma(\tilde{A})$ is the smallest value that guarantees that $\lambda_{\min}(-A)\geq\rho$.

D. Proof of Theorem II.5

The proof consists in evaluating the lower bound in Corollary II.2. We again prove the theorem by showing for a random class of functions contained in $\mathcal{A}^{(N)}$.

We consider a the set of functions such that for each possible support of a p by p matrix with at most k non-zero entries per row. Assume there is one and only one function in the family with JF having that support for all x.

Now notice that $\mathbb{E}_{x_0}\mathrm{Var}_{x_0|A}F(x_0;A) \leq \mathbb{E}(||F(x_0;A)||^2)$. Secondly notice that, if x and x' only differ on the j^{th} component and $(JF)_{ij}\neq 0$ then $|F_i(x;A)| \leq |F_i(x';A)| + D||x'-x||$. Since JF has at most k non-zero entries per row, we get that for any x and x', $|F_i(x;A)| \leq |F_i(x';A)| + kD||x'-x||$. If $x = x_0$ and $x' = \mathbb{E}_{x_0|A}(x_0|A)$ then squaring the previous expression and taking expectations gives us $\mathbb{E}_{x_0|A}(F_i(x;A)^2|A) \leq 2F_i(x';A)^2 + 2k^2D^2B$. From this we get that $\mathbb{E}(||F(x_0;A)||^2)/p \leq C + 2k^2D^2B$ where C is a constant independent of A. For this sub family of functions we have $H(M(A)) \geq pk\log(p/k)$. By (29) and (30) we know that $I(x_0;A) \leq (1/2)\log((2\pi e)^p|\mathbb{E}\Sigma_\infty|) - (1/2)\mathbb{E}\log((2\pi e)^p|\Sigma_\infty|)$. The first term, which is the entropy of a p-dimensional Gaussian with covariance matrix $\mathbb{E}\Sigma_\infty$, can be upper bounded by the sum of the entropy of its individual components, which have variance upper bounded by B. Finally, since $\Lambda_{\min}(\Sigma_\infty)\geq L$, we have $\log|\Sigma_\infty| \geq p\log L$ and therefore $I(x_0;A) \leq p/2\log B/L$, which completes the proof. ∎

Acknowledgments: This work was partially supported by the NSF CAREER award CCF-0743978, the NSF grant DMS-0806211, the AFOSR grant FA9550-10-1-0360 and by a Portuguese Doctoral FCT fellowship.

REFERENCES

[1] N. D. Lawrence, Ed., *Learning and Inference in Computational Systems Biology*. MIT Press, 2010.

[2] D. Gillespie, "Stochastic simulation of chemical kinetics," *Annual Review of Physical Chemistry*, vol. 58, pp. 35–55, 2007.

[3] D. Higham, "Modeling and Simulating Chemical Reactions," *SIAM Review*, vol. 50, pp. 347–368, 2008.

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

[5] J. Bento, M. Ibrahimi, and A. Montanari, "Learning networks of stochastic differential equations," *Advances in Neural Information Processing Systems 23*, pp. 172–180, 2010.

[6] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, p. 432, 2008.

[7] N. Meinhshausen and P. Bühlmann, "High-Dimensional Graphs and Variable Selection with the LASSO," *Annals of Statistics*, vol. 34, pp. 1436–1462, 2006.

[8] J. Songsiri, J. Dahl, and L. Vandenberghe, *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2010, pp. 89–116.

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

[10] I. Basawa and B. Prakasa Rao, *Statistical inference for stochastic processes*. London: Academic Press, 1980.

[11] G. Pavliotis and A. Stuart, "Parameter estimation for multiscale diffusions," *J. Stat. Phys.*, vol. 127, pp. 741–781, 2007.

[12] T. Kadota, M. Zakai, and J. Ziv, "Mutual information of the white gaussian channel with and without feedback," *IEEE Trans. Inf. Theory*, vol. IT-17, no. 4, pp. 368–371, July 1971.
[13] J. Bento, M. Ibrahimi, and A. Montanari, "Efficient methods for learning high-dimensional stochastic differential equations," 2011, in preparation.
[14] B. Øksendal, *Stochastic differential equations: an introduction with applications*. Springer Verlag, 2003.
[15] J. Friedman, "A proof of Alon's second eigenvalue conjecture," *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pp. 720–724, 2003.
[16] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*. Cambridge University Press, 2009.

APPENDIX

A. Proof of Theorem II.4

The following Lemma contains a matrix theory calculation that will be later used in this proof when applying Lemma III.1. Recall that we defined $\alpha = a_{\min}^2/2$.

Lemma A.1. *Let A be a random matrix defined as above and*

$$Q(a_{\min}, \rho) \equiv \lim_{p\to\infty} \frac{1}{p}\{\mathrm{Tr}\{\mathbb{E}(-A)\} - \mathrm{Tr}\{(\mathbb{E}(-A^{-1}))^{-1}\}\}. \tag{35}$$

Then, there exists a constant C' such that

$$Q(a_{\min}, \rho) \leq \min\{\frac{C' a_{\min}^2}{2\rho}, \frac{a_{\min}}{\sqrt{2}}\}. \tag{36}$$

Proof: Using Wigner's Semicircle law for random symmetric matrices [16] and the bound described in (20) it follows that,

$$\lim_{p\to\infty} \frac{1}{p}\{\mathrm{Tr}\{\mathbb{E}(-A)\} = \rho + 2\sqrt{\alpha}, \tag{37}$$

$$C(\alpha, \rho) \equiv \lim_{p\to\infty} \mathbb{E}\{\frac{1}{p}\mathrm{Tr}\{(-A)^{-1}\}\} \tag{38}$$

$$= \frac{-\sqrt{\rho\,(4\sqrt{\alpha} + \rho)} + 2\sqrt{\alpha} + \rho}{2\alpha}. \tag{39}$$

Since $C(\alpha, \rho) = \alpha^{-1/2} C(1, \rho/\sqrt{\alpha})$ we can write $\rho + 2\sqrt{\alpha} - (C(\alpha, \rho))^{-1} = \sqrt{\alpha} G(\rho/\sqrt{\alpha})$ where $G(x)$ is a strictly decreasing function. Since $\lim_{\rho\to 0} = \sqrt{\alpha} G(\rho/\sqrt{\alpha}) = \sqrt{\alpha}$ and $\lim_{\rho\to\infty} \rho\sqrt{\alpha} G(\rho/\sqrt{\alpha}) = \alpha$ it follows that there is a constant C' independent of α or ρ such that $\sqrt{\alpha} G(\rho/\alpha) \leq \sqrt{\alpha}\min\{1, C'\sqrt{\alpha}/\rho\}$. The result now follows by replacing $\alpha = a_{\min}^2/2$. ∎

Proof (Theorem II.4): Like in the proof of Theorem II.3 we start by dividing both numerator and denominator of (13) in Lemma III.1 by p. By multiplying the resulting expression by an appropriately small constant we can replace the denominator and $\lim_{p\to\infty} I(x_0; A)/p$ by their limits when $p \to \infty$ and get an expression that is still valid for all p large enough. Since $H(M(A))/p = \frac{(1+p))}{4}\log 4$, and since by Lemma A.1 we already know the limiting expression of the denominator, all we have to do is find $\lim_{p\to\infty} I(x_0; A)/p$. By an analysis very similar to that in the proof of Theorem II.3 one can show that

$$\lim_{p\to\infty} \frac{1}{p} I(x_0; A) \leq -1 + \sqrt{(z+2)C(1, z)} \leq 1. \tag{40}$$

where $C(\alpha, \rho)$ was defined in (38), which finishes the proof. ∎