# Support Recovery for the Drift Coefficient of High-Dimensional Diffusions

José Bento and Morteza Ibrahimi

***Abstract*—Consider the problem of learning the drift coefficient of a $p$-dimensional stochastic differential equation 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 when both $p$ and $T$ can tend to infinity. In particular, we prove 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 stochastic differential equations, the drift coefficient is parametrized by a $p \times p$ matrix which describes which degrees of freedom interact under the dynamics. In this case, we analyze a $\ell_1$-regularized least squares estimator and prove an upper bound on $T$ that nearly matches the lower bound on specific classes of sparse matrices.**

***Index Terms*—Stochastic differential equation, sparse recovery, dynamical systems, maximum likelihood**

## I. Introduction

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

$$\mathrm{d}x(t) = F(x(t); \Theta^0)\,\mathrm{d}t + \mathrm{d}b(t)\,, \tag{I.1}$$

where $b(t)$ is a $p$-dimensional standard Brownian motion and the *drift coefficient*[1]

$$F(x(t); \Theta^0) = [F_1(x(t); \Theta^0), \dots, F_p(x(t); \Theta^0)] \in \mathbb{R}^p,$$

is a function of $x(t)$ parametrized by $\Theta^0$. This is an unknown vector, with dimensions scaling polynomially with $p$.

In this paper we consider the problem of learning the support of the vector $\Theta^0$ from a sample trajectory $X_0^T \equiv \{x(t) : \ t \in [0,T]\}$. More precisely, we focus on the high-dimensional scenario where $p$ and $T$ are allowed to increase simultaneously. Our goal is to determine necessary and sufficient conditions for recovering the support of $\Theta^0$ and the sign of its entries with high probability. We refer to the smallest $T$ that allows 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 $\Theta^0$ and an algorithm $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ that outputs an estimate $\widehat{\Theta}$, we define the sample-complexity $T_{\mathsf{Alg}}(\Theta^0)$ as

$$\inf\big\{T_0 \in \mathbb{R}^+ : \mathbb{P}_{\Theta^0,T}\{\mathrm{sign}(\widehat{\Theta}) = \mathrm{sign}(\Theta^0)\} \geq 1-\delta, \text{ for all } T \geq T_0\big\}. \tag{I.2}$$

[1]Throughout the paper, vectors are 'column vector' even if they are represented in row form for typographical reasons.

In the expression above, $\mathbb{P}_{\Theta^0,T}$ denotes probability with respect to the trajectory $X_0^T$. The function $\mathrm{sign}(.)$ acts element-wise on its vector-valued argument and to each scalar applies the mapping $\mathrm{sign} : \mathbb{R} \mapsto \{-1, 0, 1\}$ such that

$$\mathrm{sign}(x) = \begin{cases} -1 & \text{if} \quad x < 0, \\ +1 & \text{if} \quad x > 0, \\ 0 & \text{if} \quad x = 0. \end{cases}$$

Obviously, $T_{\mathsf{Alg}}(\Theta^0)$ defined above is an upper bound for sample-complexity of learning the support alone. In addition to this definition, given some class $\mathcal{A}$ of parameters, we define

$$T_{\mathsf{Alg}}(\mathcal{A}) = \sup_{\Theta^0 \in \mathcal{A}} T_{\mathsf{Alg}}(\Theta^0). \tag{I.3}$$

Models based on SDEs play a crucial role in several domains of science and technology, ranging from chemistry to finance. Consequently, estimating their parameters has been a topic of great interest in several fields. We refer to Section III for a brief overview. A complete understanding of support recovery 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. Explicitly, we are given a set of basis functions

$$\mathbf{F}(x) = [f_1(x), f_2(x), \dots, f_m(x)], \tag{I.4}$$

with $f_i : \mathbb{R}^p \to \mathbb{R}$. The drift is then given as $F(x; \Theta^0) = \Theta^0 \mathbf{F}(x)$, with matrix $\Theta^0 \equiv \{\theta^0_{ij}\}_{i\in[p], j\in[m]} \in \mathbb{R}^{p\times m}$, $[p] = \{1, ..., p\}$ and $[m] = \{1, ..., m\}$. We then have, for each $i \in \mathbb{R}^p$,

$$\mathrm{d}x_i(t) = \sum_{j=1}^{m} \theta^0_{ij} f_j(x(t))\,\mathrm{d}t + \mathrm{d}b_i(t)\,. \tag{I.5}$$

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 $\mathrm{d}x_\mathsf{C} = k_{\mathsf{C},\mathsf{AB}}\, x_\mathsf{A} x_\mathsf{B}^2 \mathrm{d}t - k_{\mathsf{AB},\mathsf{C}}\, x_\mathsf{C} + \mathrm{d}b_\mathsf{C}$ 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 $\mathbf{F}(x)$ that comprises all monomials up to a maximum degree. In this case, the support of $\Theta^0$ tells which species react with which species, i.e. a network interactions. The sign of its entries distinguishes "inhibitory" effects from "excitatory" effects. In the end of this section we give a concrete example of using our method to learn chemical reactions.

An important subclass of models of the last type is provided by linear SDEs. In this case, the drift is a linear function of

$x(t)$, namely $F(x;\Theta^0) = \Theta^0 x(t)$ with $\Theta^0 \equiv \{\theta^0_{ij}\}_{i,j\in[p]} \in \mathbb{R}^{p\times p}$. Explicitly, for each $i \in \mathbb{R}^p$,

$$\mathrm{d}x_i(t) = \sum_{j=1}^{p} \theta^0_{ij} x_j(t)\,\mathrm{d}t + \mathrm{d}b_i(t)\,. \tag{I.6}$$

A model of this type is a good approximation for many systems near a stable equilibrium. The model (I.6) can be used to trace fluctuations of the species' concentrations in proximity of an equilibrium point in chemical reactions. In this case, the matrix $\Theta^0$ would represent the linearized interactions between different chemical factors.

More generally, we can associate to the model (I.6) a directed graph $G = (V, E)$ with edge weight $\theta^0_{ij} \in \mathbb{R}$ associated to the directed edge $(j, i)$ from $j \in V$ to $i \in V$. Each component $x_i(t)$ of the vector $x(t)$ 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 SDEs can be seen as graphical models – a probabilistic model parametrized by a graph.

This paper establishes lower bounds on the sample-complexity for estimating the support of $\Theta^0$ in the general model (I.1). These are based on information theoretic techniques and apply irrespective of computational considerations. For linear models of the form (I.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 $\Theta^0$.

Before stating our results more formally, it is useful to stress two key differences with respect to other high-dimensional estimation problems.

($i$) *Samples are not independent.*

($ii$) *Infinitely many samples are given as data* (in fact a collection indexed by $t \in [0, T]$).

A simple approach would be to select a finite subsample set. For instance, one can select a sampling interval $\eta > 0$ and only use samples at regularly spaced times $\{x(\eta), x(2\eta), x(3\eta), \dots\}$. At first sight, this reduces the problem to a more classical one. A closer consideration illustrates instead the new challenges posed by the present model.

- If $\eta$ is small, one obtains a large number of strongly dependent samples and earlier analysis does not apply. In particular, a careful analysis must reveal that there is limited information to be harnessed from a given time interval $T$.
- One might be lead into the conclusion that $\eta$ must be taken sufficiently large as to make samples approximately independent. However, this approach will waste important information contained in the sample path. For example, for a linear SDE, the matrix $\Theta^0$ contains more information than the stationary distribution of the process (I.6)[2].

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

[2]Let $\Theta^0_1 = \{\{-2,-1,-1\},\{1,-2,-1\},\{1,1,-2\}\}$ and $\Theta^0_2 = \{\{-2,1,0\},\{-1,-2,1\},\{0,-1,-2\}\}$. The linear systems defined by these matrices have different support. Yet, their stationary behavior is described by the same covariance matrix $\Sigma = \{\{1/4,0,0\},\{0,1/4,0\},\{0,0,1/4\}\}$.

### A. Regularized least squares

Regularized least squares, Rls, is an efficient and well-studied method for support recovery. We discuss relations with existing literature in Section III. In this paper we study its application to estimating the drift coefficient of a high-dimensional diffusion and show that its sample-complexity compares favorably with our information-theoretic lower bounds.

Its use is better explained for the general linearly parametrized model (I.5). For this model, we estimate independently each row of the matrix $\Theta^0 \in \mathbb{R}^{p\times m}$. The $r^{\text{th}}$ row, denoted by $\Theta^0_r$, is estimated by solving the following convex optimization problem for $\Theta_r \in \mathbb{R}^p$

$$\text{minimize} \quad \mathcal{L}(\Theta_r; X_0^T) + \lambda\|\Theta_r\|_1\,, \tag{I.7}$$

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

$$\begin{aligned}\mathcal{L}(\Theta_r; X_0^T) =& \frac{1}{2T}\int_0^T \langle\Theta_r, \mathbf{F}(x(t))\rangle^2\,\mathrm{d}t \\ &- \frac{1}{T}\int_0^T \langle\Theta_r, \mathbf{F}(x(t))\rangle\,\mathrm{d}x_r(t)\,.\end{aligned} \tag{I.8}$$

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 $\mathsf{Rls}(\lambda)$. The $\ell_1$ regularization term in Eq. (I.7) has the role of shrinking to 0 all the entries $\theta_{rj}$, except the most significant ones, thus effectively selecting the support of $\Theta$.

By minimizing the function $\mathcal{L}$ alone, i.e. setting $\lambda = 0$, one obtains the maximum likelihood estimator for the diffusion process (I.1). Maximum likelihood optimization has been used before in the context of estimating diffusions in the low-dimension setting [3]. See [1] and other references in Section III. In particular, the normalized log-likelihood function (I.8) is the appropriate generalization of the sum of square residuals for a continuous-time process. To see this heuristically, one can *formally* write $\dot{x}_r(t) = \mathrm{d}x_r(t)/\mathrm{d}t$. A careless sum of square residuals would take the form $\int(\langle\Theta_r, \mathbf{F}(x(t))\rangle - \dot{x}_r(t))^2\mathrm{d}t$. Unfortunately, this expression is not defined because $x_r(t)$ is not differentiable. On the other hand, expanding the square, we get $2T\mathcal{L}(\Theta_r; X_0^T) + \int(\dot{x}_r(t))^2\mathrm{d}t$. The first term is well defined, as is clear from Eq. (I.8), and the second is independent of $\Theta$ and hence can be dropped.

Notice that constructing a well-defined cost function as in Eq. (I.8) is not a purely academic problem. Indeed, a cost function that included the time derivative $\dot{x}(t)$ would in practice require to estimate $\dot{x}(t)$ itself. This is all but hopeless because $\dot{x}(t)$ does not exist in the model.

## II. Main Results

Our main contributions are the followings:

**Information-theoretic lower bound:** We establish a general lower bound on the sample-complexity for estimating the drift coefficient of a diffusion of the form (I.1). By specializing this result, we obtain bounds for the linearly parametrized model

[3]Low-dimensional in the sense of keeping the number of degrees of freedom, $p$, fixed and letting $T$ converge to infinity.

(I.5), and the linear model (I.6).

**Upper bound via regularized least squares:** For the linear model (I.6), and suitable class of sparse matrices $\Theta^0$, we prove high-dimensional consistency of the penalized least-squares method introduced in Section I-A. The resulting upper bound on sample-complexity matches the information theoretic lower bound up to constant factors in $p$.

For the sake of simplicity, in this section we focus on the case of sparse linear SDEs, stating upper and lower bounds, cf. Section II-B. We then illustrate the general theory by analyzing a specific but rich problem: learning the Laplacian of a sparse graph, cf. Section II-C. In Section IV we give numerical illustrations of our main results. Extensions, in particular, general lower bounds on the sample complexity, are discussed in Section V. Finally, in Section VI, we present numerical illustrations of these extensions, part of which are motivated by real-world applications.

Proofs for the technical lemmas are provided in the appendix.

### A. Notation

For any $N \in \mathbb{N}$, we let $[N] = \{1, 2, \dots, N\}$.

Given any 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 $Q_{ij} \neq 0$.

For a vector $v \in \mathbb{R}^N$, $\mathrm{supp}(v)$ is defined analogously. 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 $\mathrm{sign}(Q)_{ij} = \mathrm{sign}(Q_{ij})$ where the function $\mathrm{sign}(Q_{ij})$ is defined as

$$\mathrm{sign}(Q)_{ij} = \begin{cases} +1 & \text{if } Q_{ij} > 0 \\ 0 & \text{if } Q_{ij} = 0 \\ -1 & \text{if } Q_{ij} < 0 \end{cases} \qquad \text{(II.1)}$$

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 $\ell_q$ 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 of $v$, namely $\|v\|_0$ is the number of non-zero entries of $v$. The $\ell_q$ operator norm of a matrix $Q \in \mathbb{R}^{M \times N}$ is denoted by $|\!|\!| Q |\!|\!|_q$. In particular the $\ell_\infty$ operator norm 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 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 $0 \leq \sigma_1(Q) \leq \cdots \leq \sigma_{\min\{M,N\}}(Q)$ denote its singular values. Further $\sigma_{\min}(Q) = \sigma_1(Q)$ and $\sigma_{\max}(Q) = \sigma_{\min\{M,N\}}(Q)$ are the minimum and maximum singular values.

Throughout the paper, we denote by $C$, $C_1$, $C_2$, etc, constants that can be adjusted from point to point.

### B. Sample complexity for sparse linear SDEs

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

$$\mathcal{A}^{(S)} = \mathcal{A}^{(S)}(k, p, \theta_{\min}, \rho_{\min}) \subseteq \mathbb{R}^{p \times p} \qquad \text{(II.2)}$$

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

$(i)$ $\|\Theta_r\|_0 \leq k$ 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 ODEs $\dot{x}(t) = \Theta x(t)$ is stable. Equivalently, the spectrum of $\Theta$ 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 (I.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 [2]

$$\Theta^0 Q^0 + Q^0 (\Theta^0)^* + I = 0. \qquad \text{(II.3)}$$

Hence $X_0^T = \{x(t) : \ t \in [0, T]\}$ is a stationary trajectory distributed according to the linear model (I.6) if $x(t = 0) \sim \mathsf{N}(0, Q^0)$ is a Gaussian random variable independent of $b(t)$.

We consider the linear model (I.6) with $\Theta^0 \in \mathcal{A}^{(S)}$. Given a row index $r \in [p]$, let $S^0 = S^0(r)$ be the support of $\Theta^0_r$.

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

$$\lambda_{\min}(Q^0_{S^0, S^0}) \geq C_{\min}\,. \qquad \text{(II.4)}$$

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

$$|\!|\!| Q^0_{(S^0)^C, S^0} \left( Q^0{}_{S^0, S^0} \right)^{-1} |\!|\!|_\infty \leq 1 - \alpha\,. \qquad \text{(II.5)}$$

We refer to [3], [4] for the original development of these conditions in the context of sparse regression.

Our first theorem establishes high-dimensional consistency of $\ell_1$-penalized least squares for estimating $\mathrm{sign}(\Theta^0)$ from a stationary trajectory $X_0^T$ according to the linear model (I.6) when $\Theta^0 \in \mathcal{A}^{(S)}$.

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

$$T_{\mathsf{Rls}(\lambda)}(\Theta^0) \leq \frac{2 \cdot 10^4 k^2 (k\, \rho_{\min}^{-2} + \theta_{\min}^{-2})}{\alpha^2 \rho_{\min} C_{\min}^2} \log\Big(\frac{4pk}{\delta}\Big). \qquad \text{(II.6)}$$

*In particular, one can choose*

$$\lambda = \sqrt{\frac{36}{T \alpha^2 \rho_{\min}} \log\Big(\frac{4p}{\delta}\Big)}\,. \qquad \text{(II.7)}$$

**Remark II.1.** *Note that the notions of sample-complexity introduced in I.2 and I.3 are 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 $\mathsf{Rls}(\lambda)$, since $\lambda$ can depend on $T$.*

**Remark II.2.** *If there exists $C_{\min}, \alpha > 0$ such that assumptions 1 and 2 hold for all $r \in [p]$ and for all $\Theta^0 \in \mathcal{A}^{(S)}(k, p, \theta_{\min}, \rho_{\min})$, then we can replace $T_{\mathsf{Rls}(\lambda)}(\Theta^0)$ by $T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(S)})$ in* (II.6).

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 (I.6).

**Theorem II.2.** *Let* $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ *be an estimator of* $\mathrm{sign}(\Theta^0)$. *There is a constant* $C(k,\delta)$, *such that, for all* $p$ *large enough,*

$$T_{\mathsf{Alg}}(\mathcal{A}^{(S)}) \geq C(k,\delta)\, \max\left\{\frac{\rho_{\min}}{\theta_{\min}^2}, \frac{1}{\theta_{\min}}\right\} \log p\,. \qquad \text{(II.8)}$$

**Remark II.3.** *Theorem II.2 cannot be used to conclude that, if* $T$ *is 'small', then* $\mathsf{Rls}(\lambda)$ *always fails to reconstruct* $\Theta^0$ *from* $X_0^T$ *regardless of the choice of* $\lambda$. *What the lower bound says is that, if* $T$ *is 'small', then, for every choice of* $\lambda = \lambda(X_0^T)$, *there exists a* $\Theta^0 \in \mathcal{A}^{(S)}$ *that cannot be reconstructed. The particular* $\Theta^0$ *that cannot be reconstructed, however, can depend on the choice of* $\lambda$.

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

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. (I.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 (I.6). The upper bound is essentially tight if the matrix $\Theta^0$ is symmetric.

Theorems II.1 and II.2 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. (II.3). In order to clarify this dependency, we apply our general results to the problem of learning the Laplacian of an undirected graph.

### C. 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 [5]. (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 $\Theta^0 = -m\, I + \Delta^{\mathcal{G}}$ fits into the setting of Theorem II.1 for $m > 0$. The corresponding model (I.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 $\mathcal{G}_{\text{bounded}} = \mathcal{G}_{\text{bounded}}(k,p)$ be the class of graphs on $p$ nodes with maximum vertex degree bounded by $k$. Define,

$$\mathcal{A}^{(L)}(m,p,k) = \{\Theta^0 = -m\, I + \Delta^G \mid m > 0, G \in \mathcal{G}_{\text{bounded}}\} \qquad \text{(II.9)}$$

The following theorem holds regarding the sample-complexity of learning the signed support of $\Theta^0$ from a stationary trajectory $X_0^T$ of a linear SDE with $\Theta^0 \in \mathcal{A}^{(L)}$.

**Theorem II.3.** *If* $\Theta^0 \in \mathcal{A}^{(L)}(m,p,k)$ *then there exists* $\lambda = \lambda(T) > 0$ *such that*

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

*In particular one can take,*

$$\lambda = \sqrt{36(k+m)^2 \log(4p/\delta)/(Tm^3)}.$$

In other words, for $m$ bounded away from 0 and $\infty$, 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 II.1 and II.2 above.

**Corollary II.4.** *Assume the same setting as in Theorem II.3. There exist constants* $\lambda = \lambda(T)$, $C_1 = C_1(k,\delta)$ *and* $C_2 = C_2(k,\delta)$ *such that, for all* $p$ *large enough,*

$$m < k \;\;\Rightarrow\;\; C_1 \log p \leq T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(L)}) \leq C_2 m^{-5} \log p,$$
$$m \geq k \;\;\Rightarrow\;\; C_1 m \log p \leq T_{\mathsf{Rls}(\lambda)}(\mathcal{A}^{(L)}) \leq C_2 m^2 \log p\,.$$

*In addition, the lower-bounds hold regardless of the choice of* $\lambda$.

*Proof:* The proof of this corollary follows immediately from Theorem II.3 and Theorem II.2. ∎

Notice that the upper bound on $T_{\mathsf{Rls}}$ 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 $\Theta^0$. Vice-versa for large $m$, $\Theta^0$ 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 $T_{\mathsf{Rls}}$ as $m \to 0$ remains an open question since our lower bound stays bounded in that limit.

## III. Related work

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

In the first context, diffusions are used to model the evolution of price indices [6]. While the most elementary process is the (geometric) Brownian motion [7], [8], 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 [1], [9], [10] 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 [11]–[14]. In particular, asymptotically consistent methods based on approximate transition probabilities exist, see for instance [15], [16]. Nonparametric estimation of the drift coefficient has been studied as well [17]–[19].

However, 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 I. A large variety of chemical reactions are modeled by diffusions with suitably parametrized drift terms [20], [21]. Of particular interest here are special classes of drift coefficients, for instance those exhibiting time-scale separation [22] or gradients of a potential [23]. [24] use regularized least squares to learn SDEs and from them recover both intracellular and intercellular biological networks. In this work, several regularizations are studied, including $\ell$-1 regularization, but no guarantees are proved. In a different work, [25], the same method is applied to study the functional connectivity of the brain. As with the 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 *independent and identically distributed* (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 [26] and the pseudo-likelihood method of [4]. In the context of binary pairwise graphical models, similar methods were developed in [27].

More closely related to our paper is the work reported in [28]. It proposes an algorithm to learn the interference graph in a wireless network from passive measurements of the traffic. The paper is concerned with the number of samples required in order to recover the interference graph correctly. Both information theoretic lower bounds and upper bounds using a practical algorithm are provided. The model used in this work is a time-evolving discrete time model and the algorithm is domain specialized. In contrast, the emphasis of our work is on the continuous time models and indeed a significant portion of our effort is dedicated to obtaining the right scaling in this scenario. Furthermore, the algorithms analyzed in these two works are completely different.

To the best of our knowledge the present work is the first one moving beyond the assumption of independent samples from a continuous time diffusion process when dealing with the sample complexity of learning the structure of the underlying graph. While we extend ideas and methods from this literature, dealing with dependent samples raises new mathematical challenges.

Our methods build on the work on $\ell_1$-regularized least squares, and its variants [29]–[33]. The most closely related results are the one concerning high-dimensional consistency for support recovery [3], [4], [27]. Our proof for our upper bound follows indeed the approach developed in these papers, with two important challenges. First, the design matrix in our case is produced by a stochastic diffusion, and it does not necessarily satisfy 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 [3]. 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 $\eta$ tends to 0 for $n\eta = T$ fixed. In particular, they do not allow to prove an analogous of our continuous time result, Theorem II.1. A large part of our effort is devoted to proving more accurate probability estimates that capture the correct scaling for small $\eta$.

Finally, the related topic of learning graphical models for autoregressive processes was studied recently in [34]–[36]. These papers propose a convex relaxation that is different from the one studied in this paper, without however establishing high-dimensional consistency for model selection.

Preliminary report of our work were presented at NIPS 2010 [37] and ISIT 2011 [38]. Subsequent work by Bolstad, Van Veen and Nowak [39] establishes high-dimensional consistency for estimating autoregressive models through a related approach. These guarantees are non-uniform in the sampling rate $\eta$. The work of [40] provides upper bounds on the error of regularized least square when observations are not independent. Although bounding the error of Rls is related to our problem of support recovery, in the context of learning SDEs, the conditions under which their result holds are never reduced or related to properties of the dynamics of the SDE alone. In addition, it is unclear whether their conditions hold uniformly with the sampling rate $\eta$( the results presented only apply directly to discrete time). The more recent work of [41] relates to ours by showing that, under suitable conditions, sparse linear quadratic systems can be estimated and adaptively controlled with few observations. Finally, [42] provides a framework for filtering $X_0^T$ which could be used to estimate $\Theta^0$. It is an interesting open problem to investigate how an estimator obtained from their framework compares to ours.

## IV. Numerical illustrations of the main theoretical results

In this section we illustrate our main results on synthetic data. These numerical results agree with our observations in Theorems II.1, II.2 and II.3 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 agree with the implication of Theorem V.1 that the time-complexity is roughly independent of the sampling rate, assuming that we are in the regime of small $\eta$. Or, in other words, that our reconstruction guarantees are uniform in the sampling rate for small $\eta$.

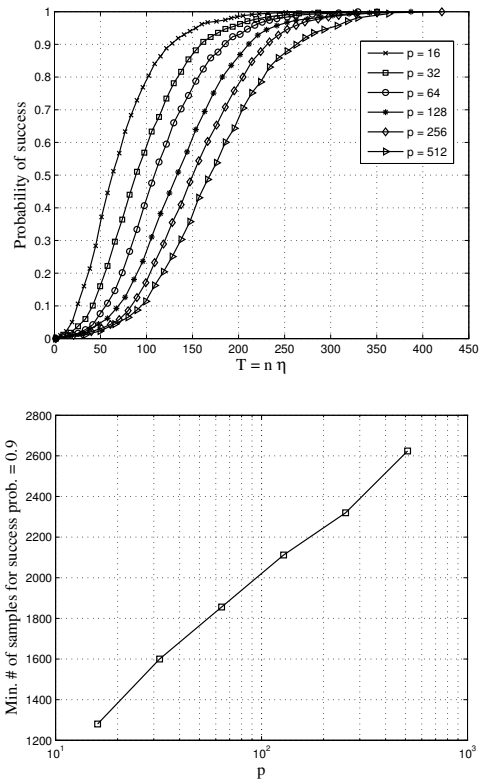


Fig. 1. (top) Probability of success vs. length of the observation interval $n\eta$. (bottom) Sample complexity for 90% probability of success vs. p.

Note that, in order to obtain numerical values for the time-complexity that do not depend on $\lambda$, we use a definition for sample-complexity and time-complexity that is slightly different than the one used when stating our main results.

We start by analyzing the performance of Rls for the discrete analogue of (I.6) (See equation (V.1) in Section V). Our results are summarized in Figures 1 and 2. First, we generate 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 $\mathbb{P}(\theta^0_{ij}=1)=k/p$, $k=5$, and form $\Theta^0=-7\mathbb{I}+\tilde{\Theta}^0$ [4]. Second, a sample path $X_0^n\equiv\{x(t):0\le t\le n\}$ is obtained from Eq. (V.1). Finally, 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 $\lambda$. We record a 1 or a 0 if the correct signed support of $\Theta^0_r$ is recovered or not. For every value of $n$ and $\lambda$, the probability of successful recovery is then estimated by taking the average of these errors over all realizations of $\Theta^0$, $X_0^n$ and $r$. Finally, for each fixed $n$, we take the maximum over $\lambda$ of these probability of success. The top plot in Figure 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 $\Theta^0$. In addition, from this plot of $n$ vs. probability of success, we generate the bottom plot in Figure 1: sample-complexity vs. $p$. To be explicit, the definition of sample-complexity in use is

$$N_{\mathsf{Rls}}(\mathcal{A})=\inf\{n_0\in\mathbb{N}_0:\sup_{\lambda>0}\hat{\mathbb{E}}\{\hat{\mathbb{P}}_{\Theta^0,n}\{\mathsf{Rls}(\lambda)=\mathrm{sign}(\Theta^0)\}\}\ge 1-\delta \text{ for all } n\ge n_0\},\tag{IV.1}$$

where we choose a probability of success of $\delta=0.9$. Above, $\hat{\mathbb{E}}$ represents empirical expectation over $\Theta^0$ and $\hat{\mathbb{P}}$ empirical probability over $X_0^n$, and, $\mathcal{A}$ is the class of all matrices that can be generated by the random procedure described before. In agreement with Theorem II.3, the curve shows the logarithmic scaling of the sample-complexity with $p$.

In Figure 2 we turn to the continuous-time model (I.6). Trajectories are generated by 'discretizing' this stochastic differential equation with step $\eta'$ much smaller than the sampling rate $\eta$. We draw random matrices $\Theta^0$ as above and plot the probability of success for $p=16$, $k=4$ and different values of $\eta$, as a function of $T$. We used $2^{11}$ instances for each curve. The time-complexity in use for these plots is the continuous-time analog of (IV.1). Again in agreement with Theorem V.1, for a fixed observation interval $T$, the probability of success converges to some limiting value as $\eta\to 0$.

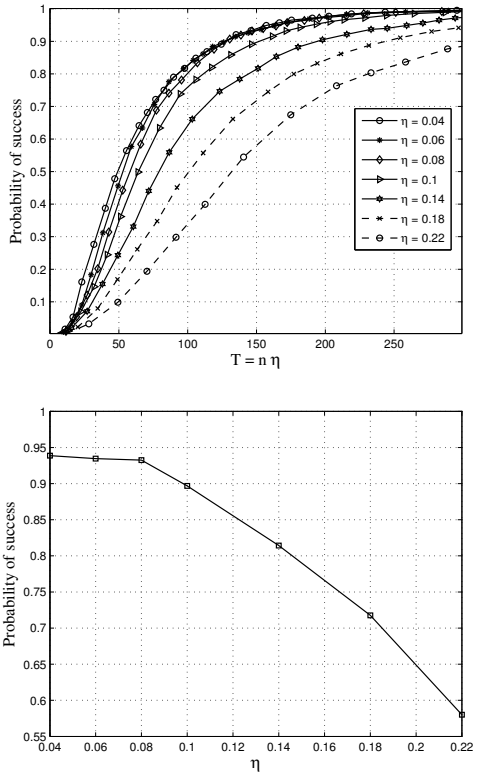


Fig. 2. (top) Probability of success vs. length of the observation interval $n\eta$ for different values of $\eta$. (bottom) Probability of success vs. $\eta$ for a fixed length of the observation interval, ($n\eta=150$) . The process is generated for a small value of $\eta$ and sampled at different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

## V. Extensions

In this section we present some extensions to our previous results. We begin by presenting an analogous theorem of Theorem II.1 for the case of a discrete time system. This is an important result in itself and also constitutes the basis for

[4] For $p$ large, the SDE generated is stable with high-probability.

[5] For discrete-time SDEs, the cost function is given explicitly in Eq. (V.2).

the proof of Theorem II.1. In fact, Theorem II.1 is proved by letting $\eta \to 0$ in the result below. We then present a general lower bound on the time-complexity of learning continuous stochastic differential equations. Using this result, lower bounds for the time-complexity of linear SDEs with dense matrices $\Theta^0$ and non-linear SDEs are derived.

### A. 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 II-B. However, much of the theoretical analysis concerning the regularized least square algorithm is in fact devoted to the analysis of the following discrete-time dynamics, with parameter $\eta > 0$:

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

Here $x(t) \in \mathbb{R}^p$ is the vector collecting the dynamical variables, $\Theta^0 \in \mathbb{R}^{p\times p}$ specifies the dynamics as above, and $\{w(t)\}_{t\geq 0}$ is a sequence of i.i.d. normal vectors with covariance $\eta\, I_{p\times p}$ (i.e. with independent components of variance $\eta$). We assume that $n+1$ consecutive samples are given, $X_0^n \equiv \{x(t) : 0 \leq t \leq t\}$, and ask under which conditions regularized least squares reconstructs the signed support of $\Theta^0$.

The parameter $\eta$ has the meaning of a time-step size. The continuous-time model (I.6) is recovered, in a sense made precise below, by letting $\eta \to 0$. Indeed, for this discrete time model, 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 II-B ) is kept constant. For a formal statement we refer to Theorem V.1. Theorem II.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 (V.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. (I.1). Notice that in the latter case the matrices $\Theta^0$ appearing in Eq. (V.1) and (I.1) coincide only to the zeroth order in $\eta$. 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 (V.1), and let $X_0^n$ be the observed portion of the trajectory. The $r^{\text{th}}$ row of $\Theta^0$, $\Theta_r^0$, is estimated by solving the following convex optimization problem

$$\underset{\Theta_r\in\mathbb{R}^p}{\text{minimize}} \quad \mathcal{L}(\Theta_r; X_0^n) + \lambda\|\Theta_r\|_1\,, \qquad \text{(V.2)}$$

where the log-likelihood function $\mathcal{L}(\Theta_r; X_0^n)$ is defined as

$$\frac{1}{2\eta^2 n}\sum_{t=0}^{n-1}\left\{x_r(t+1) - x_r(t) - \eta\,\langle\Theta_r, x(t)\rangle\right\}^2\,. \qquad \text{(V.3)}$$

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. (I.8). Indeed the proof of Theorem II.1 will amount 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 $\Theta_r$ within model (V.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 II-B but with condition $(iii)$ replaced by

$$\frac{1-\sigma_{\max}(I+\eta\,\Theta^0)}{\eta} \geq D > 0 \qquad \text{(V.4)}$$

If $\Theta^0 \in \mathcal{A}'^{(S)}$ then, under the model (V.1), $x(t)$ has a unique stationary measure which is Gaussian with covariance $Q^0$ determined by the following modified Lyapunov equation

$$\Theta^0 Q^0 + Q^0(\Theta^0)^* + \eta\Theta^0 Q^0(\Theta^0)^* + I = 0\,. \qquad \text{(V.5)}$$

It will be clear from the context whether $\Theta^0$ (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 $\ell_1$-regularized least squares recovers $\mathrm{sign}(\Theta^0)$ with high probability.

**Theorem V.1.** *Assume that $\Theta^0 \in \mathcal{A}'^{(S)}(k, p, \theta_{\min}, D)$ and that $\Theta_r^0$ satisfies assumptions 1 and 2 of Section II-B. Let $X_0^n$ be a stationary trajectory distributed according to the linear model* (V.1)*. If*

$$n\eta > \frac{10^4 k^2(kD^{-2} + \theta_{\min}^{-2})}{\alpha^2 D C_{\min}^2}\,\log\Big(\frac{4pk}{\delta}\Big)\,, \qquad \text{(V.6)}$$

*then there exists $\lambda = \lambda(n\eta) > 0$ such that $\ell_1$-regularized least squares recovers the signed support of $\Theta_r^0$ 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 $[0, T]$.

**Remark V.1.** *The form of Theorem V.1 is different than that of Theorem II.1. In Theorem V.1 we do not compute a bound on*

$$N_{\mathsf{Rls}(\lambda)}(\Theta^0) \equiv \min\big\{n_0 > 0 : \mathbb{P}_{\Theta^0,n}\{\mathrm{sign}(\widehat{\Theta}) = \mathrm{sign}(\Theta^0)\} \geq 1-\delta \text{ for all } n \geq n_0\big\},$$

*the sample-complexity of reconstructing* $\mathrm{sign}(\Theta^0)$*, but rather a bound on the sample-complexity of reconstructing the signed support of a* particular *row $r$,* $\mathrm{sign}(\Theta_r^0)$*. Obviously, if assumptions 1 and 2 hold for the same constants $C_{\min}, \alpha > 0$ across*

*$r \in [p]$, then replacing $\delta$ by $\delta/p$ in* (V.6) *allows us to use union bound and conclude that there exists $\lambda$ for which*

$$N_{\mathsf{RIs}(\lambda)}(\Theta^0)\,\eta \leq \frac{2\cdot 10^4 k^2(kD^{-2}+\theta_{\min}^{-2})}{\alpha^2 D C_{\min}^2}\,\log\left(\frac{4pk}{\delta}\right).$$

*(Notice the factor of 2). The reason why we present Theorem V.1 in a different form is to emphasize the fact that the proofs for the upper bounds are based on the success of* RIs *for reconstructing a particular row $r$.*

### *B. 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 $\Theta^0$ from a trajectory $X_0^T$ distributed according to the general model (I.1). For our problem, $M(\Theta^0)$ is the signed-support of $\Theta^0$. However, the bound holds in general. This result is used afterwards to derive lower bounds for the time-complexity of learning linear SDEs with dense matrices $\Theta^0$ (Section V-C) and for the time-complexity of learning non-linear SDEs (Section V-D).

The general form of the results in this section, and in the remainder of Section V, is as follow: If $\widehat{M}_T(X^T)$, an estimator of $M(\Theta^0)$ based on $X^T$, achieves successful recovery with probability greater than $1/2$ for every $\Theta^0$ in a class $\mathcal{A}$, then $T$ must be greater then a certain value that is dependent on properties of $\mathcal{A}$ (cf. Theorems V.4 and V.5). These results however are a corollary of a more relaxed result (Theorem V.2 and Corollary V.3) where we only require that the expected rate of miss-estimation is small when $\Theta^0$ 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 $\mathcal{A}$.

Without loss of generality, in the remainder of Section V-B, the parameter $\Theta^0$ is a random variable chosen with some unknown prior distribution $\mathbb{P}_{\Theta^0}$ (subscript will be often omitted). Also, in the following theorems we assume that $M(\Theta^0)$ can be described by an alphabet $\mathcal{M}$ of finite size $|\mathcal{M}| < \infty$. For example, if $\Theta^0 \in \mathbb{R}^{p\times p}$ and $M(.) = \mathrm{supp}(.)$ then $\mathcal{M}$ can be a set of $2^{p^2}$ symbols, one per possible support of $\Theta^0$. If $M(.) = \mathrm{sign}(.)$ then $|\mathcal{M}| = 3^{p^2}$ symbols suffice to describe all possible signed-supports of $\Theta^0$.

**Remark V.2 (Special notation).** *In this section we make a small change in our notation. Outside Section V-B, where $\Theta^0$ is a matrix of real numbers, $\mathbb{P}_{\Theta^0}$ represents a probability distribution over $X_0^T$ parametrized by $\Theta^0$. 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 $\Theta^0$*

*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 $\Theta^0$. As mentioned above $X_0^T \equiv \{x(t) : t \in [0,T]\}$ 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. In particular,*

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

*where $\mathrm{Var}_{\Theta^0|X_0^t}$ is the variance with respect to $\Theta^0$ conditioned on $X_0^t$.*

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 [43] and a similar result by Duncan [44].

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

$$T \geq \frac{2H(M(\Theta^0)) - \log(|\mathcal{M}|) - 2I(\Theta^0; x(0)) - 2}{\frac{1}{T}\int_0^T \mathbb{E}_{X_0^t}\{\mathrm{Var}_{\Theta^0|X_0^t}(F(x(t);\Theta^0))\}\mathrm{d}t}; \quad \text{(V.7)}$$

*where $|\mathcal{M}|$ is the size of the alphabet of $M(\Theta^0)$.*

*Proof:* Equation (I.1) can be regarded as describing a white Gaussian channel with feedback where $\Theta^0$ denotes the message to be transmitted. For this scenario, Kadota et al. [43] give the following identity for the mutual information between $X_0^T$ and $\Theta^0$ when the initial condition is $x(0) = 0$,

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

For the general case where $x(0)$ might depend on $\Theta^0$ (if, for example, $x(0)$ is the stationary state of the system) we can write $I(X_0^T;\Theta^0) = I(x(0);\Theta^0) + I(X_0^T;\Theta^0|x(0))$ and apply the previous identity to $I(X_0^T;\Theta^0|x(0))$. Taking into account that $I(\widehat{M}_T(X_0^T)); M(\Theta^0)) \leq I(X_0^T;\Theta^0)$ and making use of Fano's inequality $I(\widehat{M}_T(X_0^T)); M(\Theta^0)) \geq H(M(\Theta^0)) - 1 - (\mathbb{P}(\widehat{M}_T(X_0^T) \neq M(\Theta^0)))\log(|\mathcal{M}|)$ the results follows. ∎

The bound in Theorem V.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 V.3.** *Assume that* (I.1) *has a stationary distribution for every realization of $\Theta^0$ and let $X_0^T$ be a trajectory following any such stationary distribution for a specific realization of the random variable $\Theta^0$. 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*

$$T \geq \frac{2H(M(\Theta^0)) - \log(|\mathcal{M}|) - 2I(\Theta^0; x(0)) - 2}{\mathbb{E}_{x(0)}\{\mathrm{Var}_{\Theta^0|x(0)}(F(x(0);\Theta^0))\}}, \quad \text{(V.8)}$$

*where $|\mathcal{M}|$ is the size of the alphabet of $M(\Theta^0)$.*

*Proof:* Since conditioning reduces variance, we have

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

Using stationarity, we have

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

which simplifies (V.7) to (V.8). ∎

In the rest of section V, we apply this lower bound to special classes of SDEs, namely linear SDEs with dense matrices $\Theta^0$

and non-linear SDEs. In all of our applications it is to be understood that the process $\{x_t\}_{t\geq 0}$ is stationary.

### C. Learning dense linear SDEs

A different regime of interest in learning the network of interactions for a linear SDE 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 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 (I.6).

**Theorem V.4.** *Let* $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ *be an estimator of* $\mathrm{sign}(\Theta^0)$. *There is a constant* $C(\delta)$ *such that, for all* $p$ *large enough,*

$$T_{\mathsf{Alg}}(\mathcal{A}^{(D)}) \geq C(\delta)\max\Big\{\frac{\rho_{\min}}{\theta_{\min}^2}, \frac{1}{\theta_{\min}}\Big\}p. \tag{V.9}$$

The sample-complexity bound is similar to the one in Theorem II.2 but the scaling with $p$ has now changed from $\mathsf{O}(\log p)$ to $O(p)$. The lack of structure in $\Theta^0$ requires exponentially more samples for successful reconstruction. The proof is deferred to Section B-C in the appendix.

**Remark V.3.** *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* $\lambda$. *In particular, it is not hard to upper bound* $T_{\mathsf{Rls}(\lambda=0)}(\mathcal{A}^{(D)})$ *by* $\mathsf{O}(p)$. *This can be done in two steps. First, taking* $\lambda = 0$, *one can compute a closed form solution for* Rls. *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 II.1, and compute the trajectory length* $T$ *required to guarantee that*

$$\|\widehat{\Theta} - \Theta^0\|_\infty \leq \theta_{\min}/2 \tag{V.10}$$

*with probability greater than* $1-\delta$. *This value of* $T$ *is an upper bound on* $T_{\mathsf{Rls}(0)}(\mathcal{A}^{(D)})$ *since* (V.10) *plus a simple thresholding decision rule* [6] *is enough to guarantee that*

$$\mathrm{sign}(\widehat{\Theta}) = \mathrm{sign}(\Theta^0). \tag{V.11}$$

*We start Section VI with a numerical illustration of this behaviour.*

### D. Learning (sparse) non-Linear SDEs

We now assume that the observed samples $X_0^T$ come from a stochastic process driven by a general SDE of the form (I.1).

In what follows, $v_i$ denotes the $i^{th}$ component of vector $v$. For example, $x_3(2)$ is the $3^{th}$ component of the vector $x(t)$ at time $t=2$. $JF(\,\cdot\,;\Theta^0) \in \mathbb{R}^{p\times p}$ 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 $F(x;\Theta) \in \mathcal{A}^{(N)}$ if and only if

(i) the support of $JF(x;\Theta)$ has at most $k$ non-zero entries for every $x$,
(ii) the SDE (I.1) admits a stationary solution with covariance matrix, $Q$, satisfying $\lambda_{\min}(Q) \geq L$,
(iii) $\mathrm{Var}_{x(0)|\Theta}(x_i(0)) \leq B\ \forall i$,
(iv) $|\partial F_i(x;\Theta)/\partial x_j| \leq D$ for all $x \in \mathbb{R}^p$ $i,j \in [p]$.

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

Note that 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 $\mathrm{supp}(JF(x;\Theta^0)) \subseteq M(\Theta^0)\ \forall x$. Hence, we consider the following modified definition of sample-complexity that can be applied to learning SDEs of the form (I.1),

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

The following theorem holds for learning $M(\Theta^0)$, $\Theta^0 \in \mathcal{A}^{(N)}$, from a stationary trajectory of (I.1).

**Theorem V.5.** *Let* $\mathsf{Alg} = \mathsf{Alg}(X_0^T)$ *be an estimator of* $M(\Theta^0)$. *Then*

$$T_{\mathsf{Alg}}(\mathcal{A}^{(D)}) \geq \frac{k\log p/k - \log B/L}{C + 2k^2D^2B}, \tag{V.12}$$

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

**Remark V.4.** *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* (I.1) *with finite expected energy, [45].*

## VI. Numerical illustration of some extensions

In Theorem II.1 we describe a set of conditions under which Rls successfully reconstructs the dynamics of a sparse system of linear SDEs. These sufficient conditions naturally raise several questions: do they hold when the entries of $\Theta^0$ are related to some real world problem? Can Rls perform well even when these conditions do not hold? Even more generally, can Rls learn SDEs in a scenario completly different than the one described in Theorem II.1, e.g. in the presense of non-linearities? Answering these questions is non-trivial because it is hard to get a clear intuition of what assumptions like Assumption 1 and Assumption 2 of Section II-B mean in practice. The same difficulty arises with analogous results on the high-dimensional consistency of the LASSO method [3], [27].

In this section we provide concrete illustrations of the performance of Rls when applied to scenarios for which our upper bounds on time-complexity do not hold. We compare its performance to the performance predicted by our lower bounds, that hold in greater generality, and observe that, in these examples, they match. Finally, although not the focus of this paper, our last example also illustrates the effect of $\lambda$ on the performance of $\mathsf{Rls}(\lambda)$.

[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 $+1$.

### A. Time-complexity for dense linear SDEs

First we study the time-complexity for learning dense linear SDEs. We repeat the experiment of Section IV for continuous-time linear SDEs but with a dense matrix $\Theta^0$ that we generate as follows.

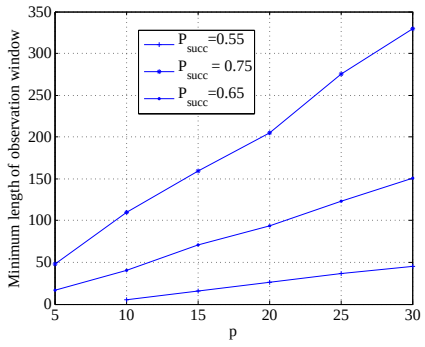


Fig. 3. Time-complexity to learn linear dense SDEs as a function of the dimension of $\Theta^0$ for different probabilities of success.

Generate $\tilde{\Theta}^0 \in \mathbb{R}^{p\times p}$ by sampling each entry from a standard Gaussian distribution; set every entry to zero with probability $1/2$; set $\Theta^0 = -(\rho+\sqrt{2})\mathbb{I} + p^{-1/2}\tilde{\Theta}^0$. For large $p$, almost all such generated matrices lead to a stable SDE.

The time-complexity curves we obtain are depicted in the figure above. Just like pointed out in Remark V.3, we observe that the time-complexity scales linearly with $p$, compared to $\mathsf{O}(\log p)$ for sparse matrices. The slope is larger for larger probabilities of success.

### B. Time-complexity for non-linear SDEs

The example in this subsection illustrates that the time-complexity of RIs scales like $\mathsf{O}(\log p)$ even when learning sparse non-linear systems of SDEs.

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 equations

$$\mathrm{d}v(t) = -\gamma v(t)\mathrm{d}t - \nabla U(q(t))\,\mathrm{d}t + \sigma\,\mathrm{d}b(t), \qquad \text{(VI.1)}$$

$$\mathrm{d}q(t) = v(t)\mathrm{d}t\,, \qquad \text{(VI.2)}$$

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

where $q(t) = (q_1(t),\dots,q_p(t))$, $v(t) = (v_1(t),\dots,v_p(t))$, 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 SDEs can be written in the form (I.1) by letting $x(t) = [q(t), v(t)]$ and $\Theta^0 = [C^0, D^0]$. A straightforward calculation shows that the drift $F(x(t);\Theta^0)$ can be further written as a linear combination of the following basis of non-linear functions

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

where $\Delta_{ij}(t) = q_i(t) - q_j(t)$ and $[p] = \{1,\dots,p\}$. Hence, the system can be modeled according to (I.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.

We consider the trajectories of three masses in a two-dimensional network of 36 masses and 90 springs evolving according to Eq. (VI.1) and Eq. (VI.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 4 reproduces the network structure reconstructed using the RLS algorithm described in Sec. I-A for increasing observation intervals $T$.

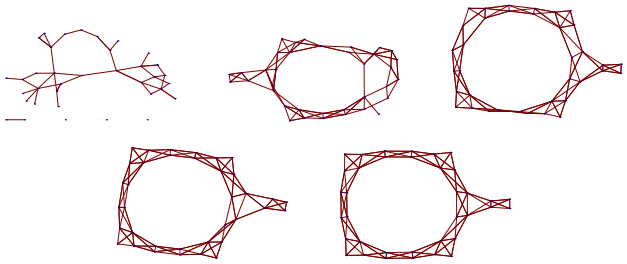

Fig. 4. From left to right, top to bottom: structures reconstructed using RIs with observation time $T = 500$, 1500, 2500, 3500 and 4500. For $T = 4500$ exact reconstruction is achieved.

Despite the non-linearities, the inferred structure converges to the true one when $T$ is large enough [7].

To quantify the efficiency of the regularized least-squares in learning non-linear SDEs, we generate multiple spring-mass networks of sizes $p = 8, 16, 32, 64$ and 128 and study the mean minimum length of the observation window required for successful reconstruction. The spring-mass networks are sampled uniformly from the ensemble of regular graphs of vertex degree 4. Like for the previous system, the data is generated by simulating the dynamics using an Euler approximation with a time step of 0.1s. The noise level, $\sigma$, is set to 0.5 and the damping parameter, $\gamma$, is set to 0.1.

Figure 5–top 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 5–bottom shows the minimum length of observation window for successful reconstruction of the networks versus their size for different

[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

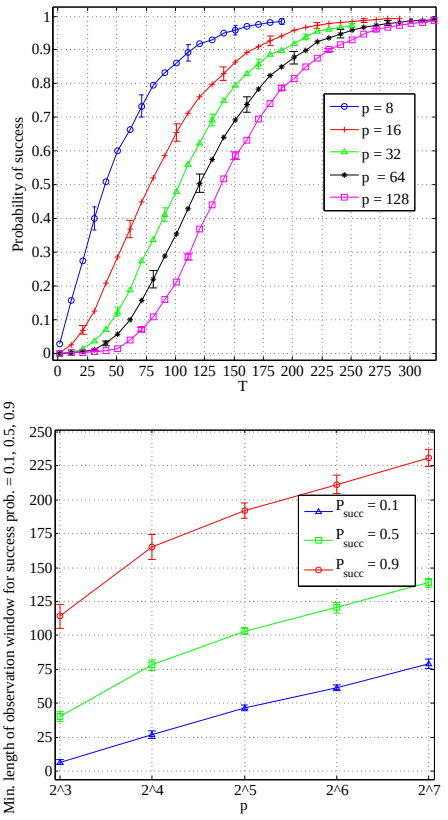


Fig. 5. (top) Probability of success versus length of observation time window, $T$, for different network sizes ($p = 8, 16, 32, 64$ and 128). (bottom) Minimum number of samples required to achieve a probability of reconstruction of $\mathrm{P_{succ}} = 0.1, 0.5$ and 0.9 versus the size of the network $p$. 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 $\gamma = 0.1$

probabilities of success ($\mathrm{P_{succ}} = 0.1, 0.5$ and 0.9). In both pictures, error bars represent $\pm$ two standard errors. We define a successful reconstruction by an exact recovery of the whole network. Since networks are sampled uniformly over regular graphs, 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 is used to minimize the number of simulations required to achieve a small fluctuation in our numerical results.

In agreement with the lower bound of Theorem V.5 for non-linear SDEs, the time-complexity of Rls in learning these sparse non-linear system of SDEs also scales logarithmically with $p$. The behavior of the plot also agrees with the $\mathsf{O}(\log p)$ time-complexity for sparse linear SDEs, even though the mass-spring system is non-linear (compare Figure 5 with Figure 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 should follow despite the non-linearities. The difficulty in proving a generalization of Theorem II.1 to general non-linear SDEs of the from (I.5) stems from the fact that it is hard in 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.

### *C. Learning biochemical pathways and the effect of the regularization parameter*

We now look at a biochemical pathway describing a general response of a cell to a change in its environment. We model the pathway behavior using non-linear SDEs, produce synthetic data by simulation and then try to recover it from the data using Rls. In this example, we also analyze how the regularization parameter, $\lambda$, affects the support recovery and the (normalized) error in estimating the values of $\Theta^0$.

The pathway in consideration is described in [46] and reproduced below.

$$R + L \underset{k_{r1}}{\overset{k_{f1}}{\rightleftharpoons}} (LR^*), \tag{VI.3}$$

$$(LR^*) + K \underset{k_{r2}}{\overset{k_{f2}}{\rightleftharpoons}} (LR^*K), \tag{VI.4}$$

$$(LR^*K) \xrightarrow{k_{f3}} (LR^*) + K^*, \tag{VI.5}$$

$$K^* + S \underset{k_{r4}}{\overset{k_{f4}}{\rightleftharpoons}} (K^*S), \tag{VI.6}$$

$$(K^*S) \xrightarrow{k_{f5}} K^* + S^*. \tag{VI.7}$$

This pathway can describe, for example, the response of cells to 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 a 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 symbols, $k_f$ and $k_r$ are the forward and backward rates of reaction. Expressions inside parenthesis, e.g. $(LR^*)$, represent specific intermediary stages or compounds along the pathway.

We assume the following correspondence between the concentration of each species and the variables $x_i(t), i \in [9]$: $x_1 \leftrightarrow R$, $x_2 \leftrightarrow L$, $x_3 \leftrightarrow (LR^*)$, $x_4 \leftrightarrow (LR^*K)$, $x_5 \leftrightarrow K$, $x_6 \leftrightarrow K^*$, $x_7 \leftrightarrow S$, $x_8 \leftrightarrow (K^*S)$, $x_9 \leftrightarrow (S^*)$. With this notation, the model proposed in [46] takes the form of a system of non-linear SDEs. Bellow are a few of the equations in the model.

$$\begin{aligned}
\mathrm{d}x_1(t) &= (k_{r1}x_3(t) - k_{f1}x_1(t)x_2(t))\mathrm{d}t + \mathrm{d}b_1(t)\\
\mathrm{d}x_2(t) &= (k_{r1}x_3(t) - k_{f1}x_1(t)x_2(t))\mathrm{d}t + \mathrm{d}b_2(t)\\
&\ldots\\
\mathrm{d}x_8(t) &= (k_{f4}x_6(t)x_7(t) - (k_{r4} + k_{f5})x_8(t))\mathrm{d}t + \mathrm{d}b_8(t)\\
\mathrm{d}x_9(t) &= (k_{f5}x_8(t))\mathrm{d}t + \mathrm{d}b_9(t)
\end{aligned}$$

The data we use for learning are synthetic sample-trajectories for the concentrations, $\{x_i(t)\}_{i=1,t\geq 0}^{9}$, obtained from these equations using the Euler-Maruyama method.

We learn the network of interaction as the support of a non-linear SDE of the form (I.5) with a basis of functions consisting of monomials up to order two, i.e., all the functions of the form $x_i^{\alpha_1} x_j^{\alpha_2}$ with $\alpha_1, \alpha_2 \in \{0, 1\}$. Although there are only 9 species in the model, the adjacency matrix whose support we want to learn is $\Theta^0 \in \mathbb{R}^{9\times 46}$ which translates into $414$ parameters to be estimated.

Figure 6 summarizes the performance of Rls in recovering the support of $\Theta^0$. Figure 6-top shows that, for a fixed value of $\lambda$, as the length of the observation increases from $T = 150$ to $T = 3000$, in steps of 285, the number of species that do not interact that are estimated as interacting (false positives) decreases and the number of species that do interact that are estimated as interacting (true positives) increases. It also shows that one can go from a high true positive rate and a high false positive rate to a low true positive rate and a low false positive rate by increasing $\lambda$. Figure 6-bottom shows the area under the previous curves as a function of $T$. In this case, the area under the curve does not have the usual probabilistic interpretation, but it does provide a metric of performance for Rls that is independent of $\lambda$. The area increases with $T$ and approaches 1, i.e. Rls can recover the exact structure of the biochemical pathway if enough data is available.

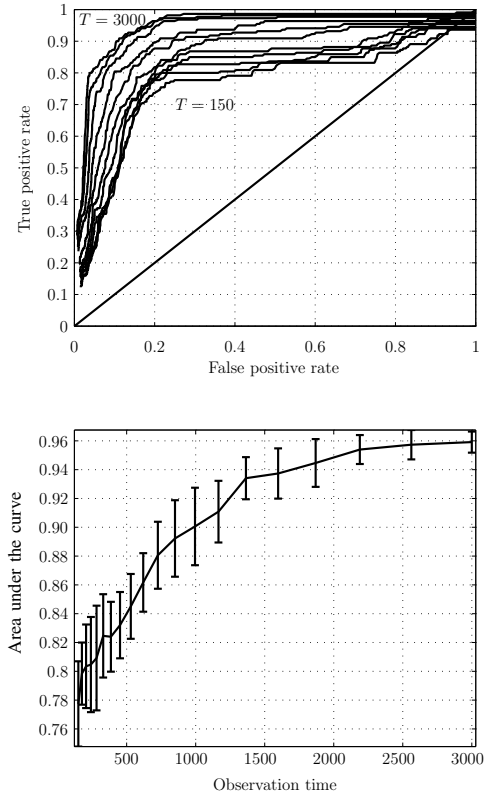


Fig. 6. (top) True positive rate versus false positive rate for the recover of the entries in the support of $\Theta^0$ using Rls. The regularization parameter $\lambda$ changes along each curve. As $\lambda$ decreases (from $\infty$ to 0), the true positive rate increases but the false positive rate also increases. (bottom) Area under the curves above as a function of $T$.

Although the focus of this paper is on support recovery, Rls also outputs real-value estimates for the entries of $\Theta^0$. Hence one can look at the normalized RMSE $\|\hat{\Theta} - \Theta^0\|_{\text{fro}} / \|\Theta^0\|_{\text{fro}}$ and its relation with $\lambda$. Figure 7-top shows this relation when running Rls on $T = 1200$ seconds of data. The curve follows the typical behavior described in [47]. In particular, there is a value of $\lambda$ that gives best parameter estimation. In Figure 7-bottom we show the evolution of the value of the minimum normalized RMSE as a function of $T$ up to the maximum duration we simulated.

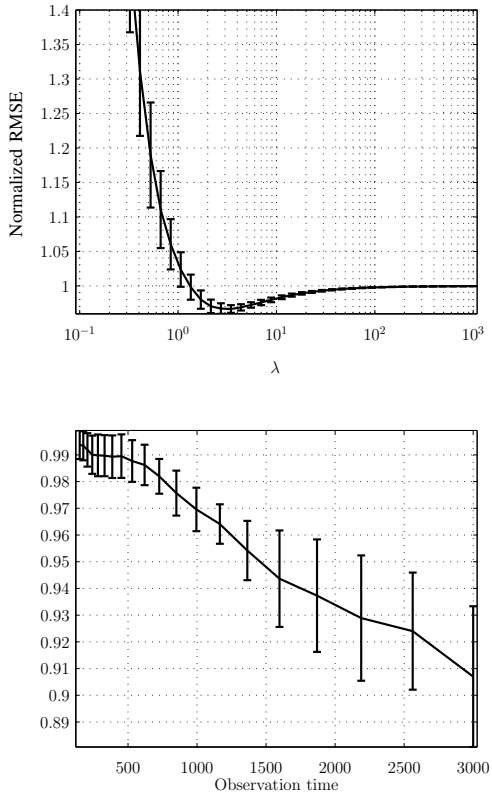


Fig. 7. (top) Normalized RMSE versus $\lambda$ for $T = 1200$. (bottom) Normalized RMSE versus $T$ for best value of $\lambda$.

*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. A great part of this work was done under the suppervision of Professor Andrea Montanari. We are very thankful for his help and contribution. Finally, we thank Jiantao Jiao and Peter Trocha for their feedback on this document.

## REFERENCES

[1] BM Brown and JI Hewitt, "Asymptotic likelihood theory for diffusion processes," *Journal of Applied Probability*, pp. 228–238, 1975.

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

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

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

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

[6] Peter C. B. Phillips and Jun Yu, "Maximum likelihood and gaussian estimation of continuous time models in finance," in *Handbook of Financial Time Series*, Thomas Mikosch, Jens-Peter Krei, Richard A. Davis, and Torben Gustav Andersen, Eds., pp. 497–530. Springer, 2009.

[7] L. Bachelier, "Théorie de la spculation," *Annales Scientifiques de lEcole Normale Supŕieure*, vol. 3, pp. 2186, 1900.

[8] F. Black and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy*, vol. 81, pp. 637654, 1973.

[9] I.V. Basawa and B.L.S. Prakasa Rao, *Statistical inference for stochastic processes*, Academic Press, London, 1980.

[10] Yu.A. Kutoyants, *Statistical Inference for Ergodic Diffusion Processes*, Springer, New York, 2004.

[11] Andrew W Lo, "Maximum likelihood estimation of generalized itô processes with discretely sampled data," 1986.

[12] D. Dacunha-Castelle and D. Florens-Zmirou, "Estimation of the coefficients of a diffusion from discrete observations," *Stochastics*, vol. 19, pp. 263–284, 1986.

[13] Asger Roer Pedersen, "A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations," *Scandinavian Journal of Statistics*, vol. 22, no. 1, pp. pp. 55–71, 1995.

[14] Y. AïtSahalia, "Maximum likelihood estimation of discretely sampled diffusions: a closed-form approach," *Econometrica*, vol. 70, pp. 223–262, 2002.

[15] Asger Roer Pedersen, "Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes," *Bernoulli*, vol. 1, pp. pp. 257–279, 1995.

[16] J. Chang and Chen S. X, "On the approximate maximum likelihood estimation for diffusion processes," *The Annals of Statistics*, vol. 39, pp. 2820–2851, 2011.

[17] J. Fan, "A selective overview of nonparametric methods in financial econometrics," *Statist. Sci.*, vol. 20, pp. 317357, 2005.

[18] V.G. Spokoiny, "Adaptive drift estimation for nonparametric diffusion model," *The Annals of Statistics*, vol. 28, pp. 815–836, 2000.

[19] A. Dalalyan, "Sharp adaptive estimation of the drift function for ergodic diffusions," *The Annals of Statistics*, vol. 33, pp. 2507–2528, 2005.

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

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

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

[23] Y. Pokern, A.M. Stuart, and E. Vanden-Eijnden, "Remarks on drift estimation for diffusion processes," *Multiscale Modeling & Simulation*, vol. 8, pp. 69–95, 2009.

[24] Chris J Oates and Sach Mukherjee, "Network inference and biological dynamics," *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 1209–1235, 2012.

[25] Pedro A Valdés-Sosa, Jose M Sánchez-Bornot, Agustín Lage-Castellanos, Mayrim Vega-Hernández, Jorge Bosch-Bayard, Lester Melie-García, and Erick Canales-Rodríguez, "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 969–981, 2005.

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

[27] 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*, vol. 19, pp. 1465, 2007.

[28] Jing Yang, Stark Draper, and Robert Nowak, "Learning the interference graph of a wireless network," *arXiv preprint arXiv:1208.0562*, 2012.

[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[30] 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*, vol. 59, no. 7, pp. 907–934, 2006.

[31] 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*, vol. 59, no. 6, pp. 797–829, 2006.

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

[33] M.J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso)," *IEEE Trans. Information Theory*, vol. 55, pp. 2183–2202, 2009.

[34] Stefan Haufe, Guido Nolte, Klaus-Robert Mueller, and Nicole Krämer, "Sparse causal discovery in multivariate time series," *arXiv preprint arXiv:0901.2234*, 2009.

[35] J. Songsiri, J. Dahl, and L. Vandenberghe, *Graphical models of autoregressive processes*, pp. 89–116, Cambridge University Press, 2010.

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

[37] José Bento, Morteza Ibrahimi, and Andrea Montanari, "Learning networks of stochastic differential equations," *Advances in Neural Information Processing Systems 23*, pp. 172–180, 2010.

[38] José Bento, Morteza Ibrahimi, and Andrea Montanari, "Information theoretic limits on learning stochastic differential equations," in *IEEE Intl. Symp. on Inform. Theory*, St. Perersbourg, Aug. 2011.

[39] A. Bolstad, B. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE transactions on signal processing*, vol. 59, pp. 2628–2641, 2011.

[40] Po-Ling Loh and Martin J Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity," *arXiv preprint arXiv:1109.3714*, 2011.

[41] Morteza Ibrahimi, Adel Javanmard, and Benjamin Van Roy, "Efficient reinforcement learning for high dimensional linear quadratic systems," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 2645–2653.

[42] Albert No and Tsachy Weissman, "Minimax filtering regret via relations between information and estimation," *arXiv preprint arXiv:1301.5096*, 2013.

[43] 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.

[44] T.E. Duncan, "On the calculation of mutual information," *SIAM Journal on Applied Mathematics*, vol. 19, no. 1, pp. 215–220, 1970.

[45] B.K. Øksendal, *Stochastic differential equations: an introduction with applications*, Springer Verlag, 2003.

[46] B.B. Aldridge, J.M. Burke, D.A. Lauffenburger, and P.K. Sorger, "Physicochemical modelling of cell signalling pathways," *Nature cell biology*, vol. 8, no. 11, pp. 1195–1203, 2006.

[47] Mohsen Bayati, José Pereira, and Andrea Montanari, "The lasso risk: asymptotic results and real world examples," in *Advances in Neural Information Processing Systems*, 2010, pp. 145–153.

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

[49] Joel 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.

[50] Guy Bresler, Elchanan Mossel, and Allan Sly, "Reconstruction of markov random fields from samples: Some observations and algorithms," *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pp. 343–356, 2008.

[51] Greg W Anderson, Alice Guionnet, and Ofer Zeitouni, *An introduction to random matrices*, Cambridge University Press, 2009.

The appendix is divided into two parts. The first half contains the proofs of the upper bounds on the sample-complexity and the second half the proofs of the lower bounds.

## APPENDIX A
## PROOFS OF THE UPPER BOUNDS ON THE SAMPLE-COMPLEXITY OF THE REGULARIZED LEAST SQUARE ALGORITHM

Our bounds for the continuous model follow from an analysis of the problem for discrete case (introduced in Section V-A) when taking the limit when $\eta \to 0$. Hence, we first prove Theorem V.1. We begin by giving an outline of the proof in Section A-A based on three propositions. The three propositions are proved in Section A-D and, in particular, the details of how to combine them to complete the proof of Theorem V.1 are in Section A-D3. Afterwards, in Section A-B, we prove Theorem II.1. Finally, in Section A-C, we specialize this bound to the case of the Laplacian of a graph and prove Theorem II.3 .

### A. Proof of Theorem V.1

In this Section we detail the proof of our main result for discrete-time dynamics, i.e., Theorem V.1. We start by stating a set of sufficient conditions for regularized least squares to recover the correct support and sign of the entries of $\Theta^0$. 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 in the main text, the proof strategy, and in particular the following proposition, Proposition A.1, which provides a compact set of sufficient conditions for the sign-support to be recovered correctly, is analogous to the one in [3]. A proof of this proposition can be found in in Section A-D1.

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 $x(t)$, i.e. $X = [x(0), x(1), \dots, x(n-1)]$. Furthermore, $\Delta X \in \mathbb{R}^{p\times n}$ is the matrix containing consecutive state changes, namely $\Delta X = [x(1)-x(0), \dots, x(n)-x(n-1)]$. It is important not to confuse $X_0^n \equiv \{x(t) : t-1 \in [n-1]\}$ with $X$ defined here. These are not the same, although both are related. In addition, $X_0^n$ should not be confused with the $n^{th}$ power of $X$ (which is never mentioned in this paper). Finally we write $W = [w(1), \dots, w(n-1)] \in \mathbb{R}^{p\times n}$ for the matrix containing the Gaussian noise realization and observe that

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

The $r^{\text{th}}$ 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 (V.3) and simply write $\mathcal{L}(\Theta_r)$. We define its normalized gradient and Hessian by

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

**Proposition A.1.** *Let $\alpha, C_{\min} > 0$ be defined by*

$$\lambda_{\min}(Q^0_{S^0,S^0}) \equiv C_{\min}$$
$$\|\!| Q^0_{(S^0)^C,S^0} \left(Q^0_{S^0,S^0}\right)^{-1} |\!\|_\infty \equiv 1-\alpha\,. \tag{A.2}$$

*If the following conditions hold then the regularized least square solution* (V.2) *correctly recovers the signed-support of $\Theta^0$, i.e. $\operatorname{sign}(\Theta^0_r)$:*

$$\|\widehat{G}\|_\infty \le \frac{\lambda\alpha}{3}\,, \tag{A.3}$$
$$\|\widehat{G}_{S^0}\|_\infty \le \frac{\Theta_{\min} C_{\min}}{4k} - \lambda, \tag{A.4}$$
$$|\!|\!| \widehat{Q}_{(S^0)^C,S^0} - Q^0_{(S^0)^C,S^0} |\!|\!|_\infty \le \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}\,, \tag{A.5}$$
$$|\!|\!| \widehat{Q}_{S^0,S^0} - Q^0_{S^0,S^0} |\!|\!|_\infty \le \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}\,. \tag{A.6}$$

*Further the same statement holds for the continuous model I.8, provided $\widehat{G}$ and $\widehat{Q}$ are the gradient and the Hessian of the likelihood (I.8).*

The proof of Theorem V.1 consists in checking that, under the hypothesis (V.6) on the number of consecutive configurations, conditions (A.4) to (A.6) hold with high probability. Checking these conditions can be regarded in turn as concentration-of-measure statement. Indeed, if expectation is taken with respect to a stationary trajectory, we have $\mathbb{E}\{\widehat{G}\} = 0$, $\mathbb{E}\{\widehat{Q}\} = Q^0$.

*1) Technical lemmas:* In this section we state the necessary concentration lemmas for proving Theorem V.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 A.2 and Proposition A.3 can be found in Section A-D2.

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

**Proposition A.2.** *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*

$$\mathbb{P}\big\{\|\widehat{G}_S\|_\infty > \epsilon\big\} \le 2|S|\,\exp\left(-n(1-\sigma_{\max})\,\epsilon^2/4\right). \tag{A.7}$$

Furthermore, we need to bound the matrix norms as per (A.6) in proposition $A.1$. 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 $\{1, ..., p\}$. Namely, we have

$$\mathbb{P}(|\!|\!|\widehat{Q}_{JS} - Q^0_{JS})|\!|\!|_\infty > \epsilon)$$
$$\le |J||S| \max_{i\in J, j\in S} \mathbb{P}(|\widehat{Q}_{ij} - Q^0_{ij}| > \epsilon/|S|). \tag{A.8}$$

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

**Proposition A.3.** *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,*

$$\mathbb{P}(|\widehat{Q}_{ij} - Q^0_{ij})| > \epsilon) \le 2\exp\left(-\frac{n}{32\eta^2}(1-\sigma_{\max})^3\epsilon^2\right). \tag{A.9}$$

Finally, the next corollary follows from Proposition A.3 and Eq. (A.8).

**Corollary A.4.** *Let $J, S$ ($|S| \leq k$) be any two subsets of $\{1, ..., p\}$ and $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\Theta^0) < 1$, $\epsilon < 2k/D$ and $n\eta > 3/D$ (where $D = (1 - \sigma_{\max})/\eta$) then,*

$$\mathbb{P}(\|\widehat{Q}_{JS} - Q^0_{JS}\|_\infty > \epsilon) \leq 2|J|k \exp\left(-\frac{n}{32k^2\eta^2}(1-\sigma_{\max})^3\epsilon^2\right). \quad \text{(A.10)}$$

*2) Outline of the proof of Theorem V.1:* With these concentration bounds we can now easily prove Theorem V.1. All we need to do is to compute the probability that the conditions given by Proposition A.1 hold. From the statement of the theorem we have that the first two conditions ($\alpha, C_{\min} > 0$) of Proposition A.1 hold. In order to make the first condition on $\widehat{G}$ imply the second condition on $\widehat{G}$, we assume that $\lambda\alpha/3 \leq (\theta_{\min}C_{\min})/(4k) - \lambda$ which is guaranteed to hold if

$$\lambda \leq \theta_{\min}C_{\min}/8k. \quad \text{(A.11)}$$

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

$$\|\widehat{Q}_{[p],S^0} - Q^0_{[p],S^0}\|_\infty \leq \frac{\alpha}{12}\frac{C_{\min}}{\sqrt{k}}, \quad \text{(A.12)}$$

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 A.2 and Corollary A.4. It is shown in the end of this section, Section A-D3, that this is satisfied if condition (V.6) holds.

### B. Proof of Theorem II.1

To prove Theorem II.1 we recall that Proposition A.1 holds provided the appropriate continuous time expressions are used for $\widehat{G}$ and $\widehat{Q}$, namely

$$\widehat{G} = -\nabla\mathcal{L}(\Theta^0_r) = \frac{1}{T}\int_0^T x(t)\,\mathrm{d}b_r(t)\,,$$
$$\widehat{Q} = \nabla^2\mathcal{L}(\Theta^0_r) = \frac{1}{T}\int_0^T x(t)x(t)^*\,\mathrm{d}t\,. \quad \text{(A.13)}$$

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 (II.6), and making use of Lemma A.5 it is easy to show that (V.6) holds for all $n > n_0$ with $n_0$ a sufficiently large constant.

Therefore, by the proof of Theorem V.1, the conditions in Proposition A.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 (V.1) will be denoted by $x(i)$ where $i \in \mathbb{N}$ and the state of continuous time system (I.1) by $x(t)$ where $t \in \mathbb{R}$. We denote by $Q^0$ the solution of (II.3) and by $Q^0(\eta)$ the solution of (V.5). It is easy to check that $Q^0(\eta) \to Q^0$ as $\eta \to 0$ by the uniqueness of stationary state distribution. We couple the process as follows.

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 to be generated by the respective dynamical systems using the same matrix $\Theta^0$ 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.

**Lemma A.5.** *Let $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\Theta^0)$ and $\rho_{\min} = -\lambda_{\max}((\Theta^0 + (\Theta^0)^*)/2) > 0$ then,*

$$-\lambda_{\min}\left(\frac{\Theta^0 + (\Theta^0)^*}{2}\right) \geq \limsup_{\eta\to 0}\frac{1-\sigma_{\max}}{\eta}, \quad \text{(A.14)}$$
$$\liminf_{\eta\to 0}\frac{1-\sigma_{\max}}{\eta} \geq -\lambda_{\max}\left(\frac{\Theta^0 + (\Theta^0)^*}{2}\right). \quad \text{(A.15)}$$

*Proof:*

$$\frac{1-\sigma_{\max}}{\eta} = \frac{1-\lambda^{1/2}_{\max}((I+\eta\Theta^0)^*(I+\eta\Theta^0))}{\eta} \quad \text{(A.16)}$$
$$= \frac{1-\lambda^{1/2}_{max}(I+\eta(\Theta^0+(\Theta^0)^*)+\eta^2(\Theta^0)^*\Theta^0)}{\eta} \quad \text{(A.17)}$$
$$= \frac{1-(1+\eta u^*(\Theta^0+(\Theta^0)^*+\eta(\Theta^0)^*\Theta^0)u)^{1/2}}{\eta}, \quad \text{(A.18)}$$

where $u$ is some unit vector that depends on $\eta$. Thus, since $\sqrt{1+x} = 1 + x/2 + \mathsf{O}(x^2)$,

$$\liminf_{\eta\to 0}\frac{1-\sigma_{\max}}{\eta} = -\limsup_{\eta\to 0} u^*\left(\frac{\Theta^0+(\Theta^0)^*}{2}\right)u \geq -\lambda_{\max}\left(\frac{\Theta^0+(\Theta^0)^*}{2}\right). \quad \text{(A.19)}$$

The other inequality is proved in a similar way. ∎

### C. Proof of Theorem II.3

In order to prove Theorem II.3 we first state and prove the following lemma,

**Lemma A.6.** *Let $G$ be a simple connected graph of vertex degree bounded above by $k$. Let $\tilde{\Theta}$ be its adjacency matrix and $\Theta^0 = -hI + \tilde{\Theta}$ with $h > k$. Then, for this $\Theta^0$, the system in* (I.1) *has $Q^0 = -(1/2)(\Theta^0)^{-1}$ and,*

$$\|Q^0_{(S^0)^C,S^0}(Q^0_{S^0,S^0})^{-1}\|_\infty = \|(\Theta^0_{(S^0)^C,(S^0)^C})^{-1}\Theta^0_{(S^0)^C,S^0}\|_\infty \leq k/h. \quad \text{(A.20)}$$

*Proof:* $\tilde{\Theta}$ is symmetric so $\Theta^0$ is symmetric. Since $\tilde{\Theta}$ is irreducible and non-negative, Perron-Frobenious theorem tells that $\lambda_{\max}(\tilde{\Theta}) \leq k$ and consequently $\lambda_{\max}(\Theta^0) \leq -h + \lambda_{\max}(\tilde{\Theta}) \leq -h + k$. Thus $h > k$ implies that $\Theta^0$ is negative definite and using equation (II.3) we can compute $Q^0 = -(1/2)(\Theta^0)^{-1}$.

Now notice that, by the block matrix inverse formula, we have

$$(Q^0_{S^0,S^0})^{-1} = -2C^{-1}, \tag{A.21}$$

$$Q^0_{(S^0)^C,S^0} = \frac{1}{2}((\Theta^0_{(S^0)^C,(S^0)^C})^{-1}\Theta^0_{(S^0)^C,S^0}C), \tag{A.22}$$

where $C = \Theta^0_{S^0,S^0} - \Theta^0_{S^0,(S^0)^C}(\Theta^0_{(S^0)^C,(S^0)^C})^{-1}\Theta^0_{(S^0)^C,S^0}$ and thus

$$\|\!| Q^0_{(S^0)^C,S^0}(Q^0_{S^0,S^0})^{-1} |\!\|_\infty = \|\!| (\Theta^0_{(S^0)^C,(S^0)^C})^{-1}\Theta^0_{(S^0)^C,S^0} |\!\|_\infty. \tag{A.23}$$

Recall the definition of the infinity norm of a matrix $B$, $\|\!| B |\!\|_\infty$,

$$\|\!| B |\!\|_\infty = \max_i \sum_j |B_{ij}|. \tag{A.24}$$

Let $z = h^{-1}$ and write,

$$\begin{aligned}(\Theta^0_{(S^0)^C,(S^0)^C})^{-1} &= -z(I - z\tilde{\Theta}_{(S^0)^C,(S^0)^C})^{-1} \\ &= -z\sum_{n=0}^{\infty}(z\tilde{\Theta}_{(S^0)^C,(S^0)^C})^n, \qquad \text{(A.25)}\\ \Theta^0_{(S^0)^C,S^0} &= z^{-1}z\tilde{\Theta}_{(S^0)^C,S^0}. \qquad \text{(A.26)}\end{aligned}$$

This allows us to conclude that

$$\|\!| (\Theta^0_{(S^0)^C,(S^0)^C})^{-1}\Theta^0_{(S^0)^C,S^0} |\!\|_\infty$$

is in fact the maximum over all path generating functions of paths starting from a node $i \in (S^0)^C$ and hitting $S^0$ for a first time. Let $\Omega_i$ denote this set of paths, $\omega$ a general path in $G$ and $|\omega|$ its length. Let $k_1, ..., k_{|\omega|}$ denote the degree of each vertex visited by $\omega$ and note that $k_m \leq k, \forall m$. Then each of these path generating functions can be written in the following form,

$$\sum_{\omega\in\Omega_i} z^{|\omega|} \leq \sum_{\omega\in\Omega_i} \frac{1}{k_1...k_{|\omega|}}(kz)^{|\omega|} = \mathbb{E}_G((kz)^{T_{i,S^0}}), \tag{A.27}$$

where $T_{i,S^0}$ is the first hitting time of the set $S^0$ by a random walk that starts at node $i \in {S^0}^C$ and moves with equal probability to each neighboring node. But $T_{i,S^0} \geq 1$ and $kz < 1$ so the previous expression is upper bounded by $kz$. ■

Now what remains to complete the proof of Theorem II.3 is to compute values for the upper bound constants $\alpha$, $\theta_{\min}$, $\rho_{\min}$ and $C_{\min}$ in Theorem II.1 . From Lemma A.6 we can set $\alpha = 1 - k/(k+m)$. In addition, clearly, we can choose $\theta_{\min} = 1$. We also have that $\sigma_{\min}(\Theta^0) \geq k + m - \sigma_{\max}(\tilde{\Theta}) \geq m + k - k = m$ so we set $\rho_{\min} = m$. Finally, notice that

$$\begin{aligned}\lambda_{\min}(Q^0_{S^0,S^0}) &= \frac{1}{2}\lambda_{\min}(-(\Theta^0)^{-1}) \\ &= \frac{1}{2}\frac{1}{\lambda_{\max}(-\Theta^0)} \\ &\geq \frac{1}{2}\frac{1}{m+k+k} \\ &\geq \frac{1}{4(m+k)}, \qquad \text{(A.28)}\end{aligned}$$

where in the last step we made use of the fact that $m+k > k$. Hence, we choose $C_{\min} = 1/(4(m+k))$. Substituting these values in the inequality from Theorem II.1 gives the desired result.

### *D. Proofs of auxiliary results for the discrete-time model*

*1) Proof of Proposition A.1:* In order to prove Proposition A.1 we first introduce two technical lemmas.

The following Lemma is taken from the proof of Lemma 6 in [48].

**Lemma A.7.** *For any subset $S \subseteq [p]$ the following decomposition holds,*

$$\widehat{Q}_{S^C,S}\left(\widehat{Q}_{S,S}\right)^{-1} = T_1 + T_2 + T_3 + Q^0_{S^C,S}\left(Q^0_{S,S}\right)^{-1}, \tag{A.30}$$

*where,*

$$T_1 = Q^0_{S^C,S}\left(\left(\widehat{Q}_{S,S}\right)^{-1} - \left(Q^0_{S,S}\right)^{-1}\right), \tag{A.31}$$

$$T_2 = (\widehat{Q}_{S^C,S} - Q^0_{S^C,S})\left(Q^0_{S,S}\right)^{-1} \quad \textit{and} \tag{A.32}$$

$$T_3 = (\widehat{Q}_{S^C,S} - Q^0_{S^C,S})\left(\left(\widehat{Q}_{S,S}\right)^{-1} - \left(Q^0_{S,S}\right)^{-1}\right). \tag{A.33}$$

(A.34)

*In addition, if $|S| \leq k$, if $\|\!| Q^0_{S^C,S}\left(Q^0_{S,S}\right)^{-1} |\!\|_\infty < 1$ and $\lambda_{\min}(\widehat{Q}_{S,S}) \geq C_{\min}/2 > 0$ the following relations hold,*

$$\|\!| T_1 |\!\|_\infty \leq \frac{2\sqrt{k}}{C_{\min}}\|\!| \widehat{Q}_{S,S} - Q^0_{S,S} |\!\|_\infty, \tag{A.35}$$

$$\|\!| T_2 |\!\|_\infty \leq \frac{\sqrt{k}}{C_{\min}}\|\!| \widehat{Q}_{S^C,S} - Q^0_{S^C,S} |\!\|_\infty \quad \textit{and} \tag{A.36}$$

$$\|\!| T_3 |\!\|_\infty \leq \frac{2\sqrt{k}}{C^2_{\min}}\|\!| \widehat{Q}_{S^C,S} - Q^0_{S^C,S} |\!\|_\infty \|\!| \widehat{Q}_{S,S} - Q^0_{S,S} |\!\|_\infty. \tag{A.37}$$

The following lemma, directly obtained from the proofs of Proposition 1 in [48] and Proposition 1 in [3] respectively, resumes the conditions that guarantee correct signed-support reconstruction of $\Theta^0_r$.

**Lemma A.8.** *If $\widehat{Q}_{S^0,S^0} > 0$, then the dual vector $\hat{z}$ from the KKT conditions of the optimization problem* (V.2) *satisfies the following inequality,*

$$\begin{aligned}\|\hat{z}_{(S^0)^C}\|_\infty \leq{}& \|\!| \widehat{Q}_{(S^0)^C,S^0}\left(\widehat{Q}_{S^0,S^0}\right)^{-1} |\!\|_\infty \\ &\left(1 + \frac{\|\widehat{G}_{S^0}\|_\infty}{\lambda}\right) + \frac{\|\widehat{G}_{(S^0)^C}\|_\infty}{\lambda}. \qquad \text{(A.38)}\end{aligned}$$

$$\tilde{R}(j) = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 \\ \rho(m,j) & \rho(m-1,j) & \dots & \rho(1,j) & \rho(0,j) & 0 & \dots & 0 & 0 \\ \rho(m+1,j) & \rho(m,j) & \dots & \rho(2,j) & \rho(1,j) & \rho(0,j) & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ \rho(m+n-1,j) & \rho(m+n-2,j) & \dots & \rho(n,j) & \rho(n-1,j) & \rho(n-2,j) & \dots & \rho(0,j) & 0 \end{pmatrix}. \quad \text{(A.29)}$$

*In addition, if*

$$\|\widehat{G}_{S^0}\|_\infty \leq \frac{\theta_{\min}\lambda_{\min}(\widehat{Q}_{S^0,S^0})}{2k} - \lambda \quad \text{(A.39)}$$

*then* $\|\Theta^0_r - \hat{\Theta}_r\|_\infty \leq \theta_{\min}/2$. *The same result holds for problem* (I.7).

**Proof of Proposition A.1:** To guarantee that our estimated support is at least contained in the true support we need to impose that $\|\hat{z}_{S^C}\|_\infty < 1$. To guarantee that we do not introduce extra elements in estimating the support and also to determine the correct sign of the solution we need to impose that $\|\Theta^0_r - \hat{\Theta}_r\|_\infty \leq \theta_{\min}/2$.

Now notice that since $\lambda_{\min}(Q^0_{S^0,S^0}) = C_{\min}$ the relation $\lambda_{\min}(\widehat{Q}_{S^0,S^0}) \geq C_{\min}/2$ is guaranteed as long as $\|\widehat{Q}_{S^0,S^0} - Q^0_{S^0,S^0}\|_\infty \leq C_{\min}/2$. Using the norm triangle inequality together with Lemma A.7 to bound $\|...\|_\infty$ in (A.38), it is easy to see that the bounds of Proposition A.1 lead to $|\hat{z}_{(S^0)^C}|_\infty < 1$ and to (A.39) being verified. In turn, these lead to a correct recovery of the sign and support of $\Theta^0_r$. □

*2) Proof of Propositions A.2 and A.3:* To prove the concentration bounds of Propositions A.2 and A.3 we need the following lemmas.

**Lemma A.9.** *Let $r, j \in [p]$ and let $\rho(\tau, j)$ represent a $p \times p$ matrix with all rows equal to zero except the $r^{th}$ row which equals the $j^{th}$ row of $(I + \eta\Theta^0)^\tau$ (the $\tau^{th}$ power of $I+\eta\Theta^0$ ). Let $\tilde{R}(j) \in \mathbb{R}^{(n+m+1)p \times (n+m+1)p}$ be defined as in Eq. A.29.*

*Define $R(j) = 1/2(\tilde{R} + \tilde{R}^*)$ and let $\nu_i$ denote its $i^{th}$ eigenvalue and assume $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\Theta^0) < 1$. Then,*

$$\sum_{i=1}^{p(n+m+1)} \nu_i = 0, \quad \text{(A.41)}$$

$$\max_i |\nu_i| \leq \frac{1}{1-\sigma_{\max}} \quad \textit{and} \quad \text{(A.42)}$$

$$\sum_{i=1}^{p(n+m+1)} \nu_i^2 \leq \frac{1}{2}\frac{n}{1-\sigma_{\max}}. \quad \text{(A.43)}$$

*Proof:* First it is immediate to see that $\sum_{i=1}^{p(n+m+1)} \nu_i = Tr(R) = 0$. Let $I_{1\tau}$ represent a $(n+m+1) \times (n+m+1)$ matrix with zeros everywhere and ones in the block-positions of $R(j)$ where $\rho(\tau, j)$ appears and $I_{2\tau}$ represent a similar matrix but with ones in the block-position of $R(j)$ where $\rho(\tau, j)^*$ appears. Then $R$ can be written as,

$$R = \frac{1}{2}\left(\sum_{\tau=0}^{m+n-1} I_{1\tau} \otimes \rho(\tau, j) + I_{2\tau} \otimes \rho(\tau, j)^*\right), \quad \text{(A.44)}$$

where $\otimes$ denotes the Kronecker product of matrices. This expression can be used to compute an upper bound on $|\nu_i|$. Namely,

$$\begin{aligned} \max_i |\nu_i| &= \sigma_{\max}(R) \\ &\leq \sum_{\tau=0}^{\infty} \sigma_{\max}(I_{1\tau} \otimes \rho(\tau, j)) \\ &\leq \sum_{\tau=0}^{\infty} \sigma_{\max}(I_{1\tau})\sigma_{\max}(\rho(\tau, j)) \\ &\leq \sum_{\tau=0}^{\infty} \sigma_{\max}(\rho(\tau, j)) \\ &\leq \sum_{\tau=0}^{\infty} \sigma^\tau_{\max} = \frac{1}{1-\sigma_{\max}}. \end{aligned} \quad \text{(A.45)}$$

For the other bound we do,

$$\begin{aligned} \sum_{i=1}^{(n+m+1)p} \nu_i^2 &= Tr(R^2) \\ &\leq \frac{2}{4}\, n \sum_{\tau=0}^{\infty} Tr(\rho(\tau, j)\rho(\tau, j)^*) \\ &= \frac{1}{2} n \sum_{\tau=0}^{\infty} \|\rho(\tau, j)\|_2^2 \\ &\leq \frac{1}{2} n \sum_{\tau=0}^{\infty} \sigma^{2\tau}_{\max} \\ &\leq \frac{1}{2}\frac{n}{1-\sigma_{\max}}, \end{aligned} \quad \text{(A.46)}$$

where in the last step we used the fact that $0 \leq \sigma_{\max} < 1$. ■

**Lemma A.10.** *Let $j \in [p]$. Define* [8] *$\rho(\tau, j) \in \mathbb{R}^{1\times p}$ to be the $j^{th}$ row of $(I + \eta\Theta^0)^\tau$. Let $\Phi_j \in \mathbb{R}^{n\times(n+m)p}$ be defined as in Eq.* (A.40)

*Let $\nu_l$ denote the $l^{th}$ eigenvalue of the matrix $R(i, j) = 1/2(\Phi_j^*\Phi_i + \Phi_i^*\Phi_j) \in \mathbb{R}^{(n+m)p\times(n+m)p}$ (where $i \in [p]$) and assume $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\Theta^0) < 1$ then,*

$$|\nu_l| \leq \frac{1}{(1-\sigma_{\max})^2} \quad \textit{and} \quad \text{(A.47)}$$

$$\frac{1}{n}\sum_{l=1}^{(n+m)p} \nu_l^2 \leq \frac{2}{(1-\sigma_{\max})^3}\left(1 + \frac{3}{2n}\frac{1}{1-\sigma_{\max}}\right). \quad \text{(A.48)}$$

[8] Note that, with regards to Lemma A.9, we are redefining the meaning of $\rho(\tau, j)$

$$\Phi_j = \begin{pmatrix} \rho(m,j) & \rho(m-1,j) & \dots & \rho(1,j) & \rho(0,j) & 0 & \dots & 0 \\ \rho(m+1,j) & \rho(m,j) & \dots & \rho(2,j) & \rho(1,j) & \rho(0,j) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & 0 \\ \rho(m+n-1,j) & \rho(m+n-2,j) & \dots & \rho(n,j) & \rho(n-1,j) & \rho(n-2,j) & \dots & \rho(0,j) \end{pmatrix}. \tag{A.40}$$

*Proof:* The first bound can be proved in a trivial manner. In fact, since for any matrix $A$ and $B$ we have $\sigma_{\max}(A+B) \leq \sigma_{\max}(A) + \sigma_{\max}(B)$ and $\sigma_{\max}(AB) \leq \sigma_{\max}(A)\sigma_{\max}(B)$ we can write

$$\begin{aligned} \max_l |\nu_l| &= \sigma_{\max}(1/2(\Phi_j^*\Phi_i + \Phi_i^*\Phi_j)) \\ &\leq 1/2(\sigma_{\max}(\Phi_j^*\Phi_i) + \sigma_{\max}(\Phi_i^*\Phi_j)) \qquad \text{(A.49)} \\ &\leq \sigma_{\max}(\Phi_i^*\Phi_j) \\ &\leq \frac{1}{(1-\sigma_{\max})^2}, \qquad \text{(A.50)} \end{aligned}$$

where in the last inequality we used the fact $\sigma_{\max}(\Phi_j) \leq 1/(1-\sigma_{\max})$. The proof of this last fact is just a copy of the proof of the bound (A.42) in Lemma A.9.

Now notice that $\Phi_i^*\Phi_j$ can be written as a block matrix

$$\begin{pmatrix} \tilde{A} & \tilde{D} \\ \tilde{C} & \tilde{B} \end{pmatrix} \tag{A.51}$$

where $\tilde{A}, \tilde{B}, \tilde{C}$ and $\tilde{D}$ are block-matrices. Each block is a $p$ by $p$ matrix. $\tilde{A}$ has $p \times p$ blocks, $\tilde{B}$ has $n \times n$ blocks, $\tilde{C}$ has $n \times m$ blocks and $\tilde{D}$ has $m \times n$ blocks. If we index the blocks of each matrix with the indices $x, y$ these can be described in the following way

$$\tilde{A}_{xy} = \sum_{s=1}^{m} \rho(m-x+s,i)^*\rho(m-y+s,j) \tag{A.52}$$

$$\tilde{B}_{xy} = \sum_{s=0}^{n-x} \rho(s,i)^*\rho(s+x-y,j), \quad x \geq y \tag{A.53}$$

$$\tilde{B}_{xy} = \sum_{s=0}^{n-y} \rho(s+y-x,i)^*\rho(s,j), \quad x \leq y \tag{A.54}$$

$$\tilde{C}_{xy} = \sum_{s=0}^{n-x} \rho(s,i)^*\rho(m-y+x+s,j) \tag{A.55}$$

$$\tilde{D}_{xy} = \sum_{s=0}^{n-y} \rho(m-x+y+s,i)^*\rho(s,j). \tag{A.56}$$

With this in mind and denoting by $A, B, C$ and $D$ the 'symmetrized' versions of these same matrices (e.g.: $A = 1/2(\tilde{A} + \tilde{A}^*)$) we can write,

$$\sum_{l=1}^{(n+m)p} \nu_l^2 = Tr(R(i,j)^2) = Tr(A^2) + Tr(B^2) + 2Tr(CD). \tag{A.57}$$

We now compute a bound for each one of the terms. We exemplify in detail the calculation of the first bound only. First write,

$$Tr(A^2) = \sum_{x=1}^{m}\sum_{y=1}^{m} Tr(A_{xy}A_{xy}^*). \tag{A.58}$$

Now notice that each $Tr(A_{xy}A_{xy}^*)$ is a sum over $\tau_1, \tau_2 \in [m]$ of terms of the type,

$$\begin{aligned} &(\rho(m-x+\tau_1,i)^*\rho(m-y+\tau_1,j) \\ &\quad + \rho(m-x+\tau_1,j)^*\rho(m-y+\tau_1,i)) \\ &\times (\rho(m-y+\tau_2,j)^*\rho(m-x+\tau_2,i) \\ &\quad + \rho(m-y+\tau_2,i)^*\rho(m-x+\tau_2,j)). \end{aligned} \tag{A.59}$$

The trace of a matrix of this type can be easily upper bounded by

$$\begin{aligned} &(\sigma_{\max})^{m-x+\tau_1+m-y+\tau_1+m-y+\tau_2+m-x+\tau_2} \\ &\quad = (\sigma_{\max})^{2(m-x)+2(m-y)+2\tau_1+2\tau_2} \end{aligned} \tag{A.60}$$

which finally leads to

$$Tr(A^2) \leq \frac{1}{(1-\sigma_{\max})^4}. \tag{A.61}$$

Similarly for the other terms

$$Tr(B^2) \leq \sum_{x,y}^{n,n}\sum_{\tau_1,\tau_2} \sigma_{\max}^{2\tau_1+2\tau_2+2|x-y|} \leq \frac{2n}{(1-\sigma_{\max})^3} \tag{A.62}$$

$$\begin{aligned} Tr(DC) &= \sum_{x=1}^{m}\sum_{y=1}^{n} Tr(C_{xy}D_{yx}) \\ &\leq \sum_{x,y,\tau_1,\tau_2}^{m,n,n-y,n-y} \sigma_{\max}^{2(m-x)+2y+2\tau_1+2\tau_2} \\ &\leq \frac{1}{(1-\sigma_{\max})^4}. \end{aligned} \tag{A.63}$$

Putting all these together leads to the desired bound. ■

**Proof of Proposition A.2:** We will start by proving that this exact same bound holds when the probability of the event $\{\|\widehat{G}_S\|_\infty > \epsilon\}$ is computed with respect to a trajectory $\{x(t)\}_{t=-m}^{n}$ that is initiated at instant $t = -m$ with the value $x(-m) = w(-m)$. Assume we have done so. Now notice that as $m \to \infty$, $X$ converges in distribution to $n$ consecutive samples from the model (V.1) when this is initiated from stationary state. Since $\|\widehat{G}_S\|_\infty$ is a continuous function of $X = [x(0), ..., x(n-1)]$, by the Continuous Mapping Theorem, $\|\widehat{G}_S\|_\infty$ converges in distribution to the corresponding random variable in the case when the trajectory $\{x(i)\}_{i=0}^{n}$ is initiated from stationary state. Since the probability bound does not depend on $m$ we have that this same bound holds for stationary trajectories too.

We now prove our initial claim. Recall that $\widehat{G}_j = (X_j W_r^*)/(n\eta)$. Since $X$ is a linear function of the independent Gaussian random variables $W$ we can write $X_j W_r^* = \eta z^* R(j) z$, where $z \in \mathbb{R}^{p(n+m+1)}$ is a vector of i.i.d. $\mathsf{N}(0,1)$

random variables and $R(j) \in \mathbb{R}^{p(n+m+1)\times p(n+m+1)}$ is the symmetric matrix defined in Lemma A.9.

Now apply the standard Bernstein method. First by union bound we have

$$\mathbb{P}\big\{\|\widehat{G}_S\|_\infty > \epsilon\big\} \leq 2|S| \max_{j\in S} \mathbb{P}\big\{z^* R(j) z > n\epsilon\big\}.$$

Next denoting by $\{\nu_i\}_{1\leq i\leq p(n+m+1)}$ the eigenvalues of $R(j)$, we have, for any $\gamma > 0$,

$$\begin{aligned}
&\mathbb{P}\big\{z^* R(j) z > n\epsilon\big\} \\
&\quad = \mathbb{P}\Big\{ \sum_{i=1}^{p(n+m+1)} \nu_i z_i^2 > n\epsilon \Big\} \\
&\quad \leq e^{-n\gamma\epsilon} \prod_{i=1}^{p(n+m+1)} \mathbb{E}\big\{e^{\gamma\nu_i z_i^2}\big\} \\
&\quad = \exp\left( -n\Big(\gamma\epsilon + \frac{1}{2n} \sum_{i=1}^{(n+m+1)p} \log(1-2\nu_i\gamma)\Big)\right).
\end{aligned}$$

Let $\gamma = \frac{1}{2}(1-\sigma_{\max})\epsilon$. Using the bound obtained for $\max_i |\nu_i|$ in Eq. (A.42) (Lemma A.9) we have $|2\nu_i\gamma| \leq \epsilon$. Now notice that if $|x| < 1/2$ then $\log(1-x) > -x - x^2$. Thus, if we assume $\epsilon < 1/2$ and given that $\sum_{i=1}^{(n+m+1)p} \nu_i = 0$ (see Eq. (A.41) in Lemma A.9) we can continue the chain of inequalities,

$$\begin{aligned}
&\mathbb{P}(\|\widehat{G}_S\|_\infty > \epsilon) \\
&\quad \leq 2|S| \max_j \exp\left( -n\Big(\gamma\epsilon - 2\gamma^2 \frac{1}{n} \sum_{i=1}^{(n+m+1)p} \nu_i^2\Big)\right) \\
&\quad \leq 2|S| \exp\left( -\frac{n}{2}(1-\sigma_{\max})\epsilon^2 \Big(1 + \frac{1}{2}\frac{1-\sigma_{\max}}{1-\sigma_{\max}}\Big)\right) \\
&\quad \leq 2|S| \exp\left( -\frac{n}{4}(1-\sigma_{\max})\epsilon^2 \right). \qquad \text{(A.64)}
\end{aligned}$$

where the second inequality is obtained using the bound in Eq. (A.43) from Lemma A.9. □

**Proof of Proposition A.3:** The proof is very similar to that of proposition A.2. We will first show that the bound

$$\mathbb{P}(|\widehat{Q}_{ij} - \mathbb{E}(\widehat{Q}_{ij})| > \epsilon) \leq 2e^{-\frac{n}{32\eta^2}(1-\sigma_{max})^3\epsilon^2}, \qquad \text{(A.65)}$$

holds in the case where the probability measure and expectation are taken with respect to trajectories $\{x(i)\}_{i=0}^n$ that started at time instant $t = -m$ with $x(-m) = w(-m)$. Assume we have done so. Now notice that as $m \to \infty$, $X$ converges in distribution to $n$ consecutive samples from the model V.1 when this is initiated from stationary state. In addition, as $m \to \infty$, we have from lemma A.11 that $\mathbb{E}(\widehat{Q}_{ij}) \to Q^0_{ij}$. Since $\widehat{Q}_{ij}$ is a continuous function of $X = [x(0), ..., x(n-1)]$, a simple application of the Continuous Mapping Theorem plus the fact that the upper bound is continuous in $\epsilon$ leads us to conclude that the bound also holds when the system is initiated from stationary state.

To prove our initial statement first recall the definition of $\widehat{Q}$ and notice that we can write,

$$\widehat{Q}_{ij} = \frac{\eta}{n} z^* R(i,j) z, \qquad \text{(A.66)}$$

where $z \in \mathbb{R}^{m+n}$ is a vector of i.i.d. $\mathsf{N}(0,1)$ and $R(i,j) \in \mathbb{R}^{(n+m)\times(n+m)}$ is defined as in lemma A.10. Letting $\nu_l$ denote the $l^{th}$ eigenvalue of the symmetric matrix $R(i,j)$ we can further write,

$$\widehat{Q}_{ij} - \mathbb{E}(\widehat{Q}_{ij}) = \frac{\eta}{n} \sum_{l=1}^{(n+m)p} \nu_l(z_l^2 - 1). \qquad \text{(A.67)}$$

By Lemma A.10 we know that,

$$|\nu_l| \leq \frac{1}{(1-\sigma_{\max})^2} \quad \text{and} \qquad \text{(A.68)}$$

$$\begin{aligned}
\frac{1}{n} \sum_{l=1}^{(n+m)p} \nu_l^2 &\leq \frac{2}{(1-\sigma_{\max})^3} \left(1 + \frac{3}{2n}\frac{1}{1-\sigma_{\max}}\right) \\
&\leq \frac{3}{(1-\sigma_{\max})^3}, \qquad \text{(A.69)}
\end{aligned}$$

where we applied $T > 3/D$ in the last line.

Now we are done since applying Bernstein method, this time with $\gamma = 1/8\,(1-\sigma_{\max})^3\epsilon/\eta$, and making again use of the fact that $\log(1-x) > -x - x^2$ for $|x| < 1/2$ we get,

$$\begin{aligned}
&\mathbb{P}(\widehat{Q}_{ij} - \mathbb{E}(\widehat{Q}_{ij}) > \epsilon) \\
&\quad = \mathbb{P}\Big( \sum_{l=1}^{(n+m)p} \nu_l(z_l^2 - 1) > \epsilon n/\eta \Big) \\
&\quad \leq e^{-\frac{\gamma\epsilon n}{\eta}} e^{-\gamma\sum_{l=1}^{(n+m)p}\nu_l} e^{-1/2\sum_{l=1}^{(m+n)p}\log(1-2\gamma\nu_l)} \\
&\quad \leq e^{-\frac{\gamma\epsilon n}{\eta} - \gamma\sum_{l=1}^{(n+m)p}\nu_l + \gamma\sum_{l=1}^{(n+m)p}\nu_l + 2\gamma^2\sum_{l=1}^{(n+m)p}\nu_l^2} \\
&\quad \leq e^{-\frac{n}{32\eta^2}(1-\sigma_{\max})^3\epsilon^2}. \qquad \text{(A.70)}
\end{aligned}$$

Above, in order to apply the bound on $\log(1-x)$, we require that $\epsilon < 2/D$.

An analogous reasoning leads us to,

$$\mathbb{P}(\widehat{Q}_{ij} - \mathbb{E}(\widehat{Q}_{ij}) < -\epsilon) \leq e^{-\frac{n}{32\eta^2}(1-\sigma_{\max})^3\epsilon^2} \qquad \text{(A.71)}$$

and the results follows.

□

**Lemma A.11.** *As before, assume* $\sigma_{\max} \equiv \sigma_{\max}(I + \eta\Theta^0) < 1$ *and consider that model* (V.1) *was initiated at time* $-m$ *with* $x(-m) = w(-m)$ *then*

$$|\mathbb{E}(\widehat{Q}_{ij}) - Q^0_{ij}| \leq \frac{1}{n+m}\frac{\eta}{(1-\sigma_{\max})^2}. \qquad \text{(A.72)}$$

*Proof:* Let $\rho = I + \eta\Theta^0$. Taking the expectation of $\hat{Q}_{ij}$ in (A.66), and recalling that $z$ is a vector of i.i.d. standard Gaussian variables, we can write,

$$\mathbb{E}(\widehat{Q}_{ij}) = \eta \sum_{l=0}^{n+m-1} \frac{m+n-l}{n+m} (\rho^l \rho^{*l})_{ij}. \qquad \text{(A.73)}$$

We also have that

$$Q^0_{ij} = \eta \sum_{l=0}^{\infty} (\rho^l \rho^{*l})_{ij}. \qquad \text{(A.74)}$$

This last expression can be proved, for example, by taking $n \to \infty$ in (A.73). Putting these two expressions together we obtain

$$Q^0_{ij} - \mathbb{E}(\widehat{Q}_{ij}) = \eta \left( \sum_{l=m+n}^{\infty} (\rho^l \rho^{*l})_{ij} + \sum_{l=1}^{n+m-1} \frac{l}{m+n} (\rho^l \rho^{*l})_{ij} \right). \tag{A.75}$$

Using the fact that for any matrix $A$ and $B$ $\max_{ij}(A_{ij}) \leq \sigma_{\max}(A)$, $\sigma_{\max}(AB) \leq \sigma_{\max}(A)\sigma_{\max}(B)$ and $\sigma_{\max}(A+B) \leq \sigma_{\max}(A) + \sigma_{\max}(B)$, and introducing the notation $\zeta = \rho^2$, we can write,

$$\begin{aligned} |\mathbb{E}(\widehat{Q}_{ij}) - Q^0_{ij}| &\leq \eta \left( \frac{\zeta^{n+m}}{1-\zeta} + \frac{\zeta}{n+m} \sum_{l=0}^{m+n-2} \zeta^l \right) \\ &= \frac{\eta(\zeta^2 + \zeta^{n+m} - 2\zeta^{m+n+1})}{(m+n)(1-\zeta)^2} \\ &\leq \frac{\eta}{(m+n)(1-\sigma_{max})^2}. \end{aligned} \tag{A.76}$$

Above, we used the fact that for $\zeta \in [0,1]$ and $n \in \mathbb{N}$ we have $1 - \zeta \geq 1 - \sqrt{\zeta}$ and $\zeta^2 + \zeta^n - 2\zeta^{1+n} \leq 1$. ∎

*3) Proof of Theorem V.1 for discrete case system:* In order to prove Theorem V.1 we need to compute the probability that the conditions given by Proposition A.1 hold.

From the statement of the theorem we have that the two conditions, $\alpha, C_{\min} > 0$, of Proposition A.1 hold.

In order to make the first condition on $\widehat{G}$ imply the second condition on $\widehat{G}$ we assume that

$$\frac{\lambda\alpha}{3} \leq \frac{\theta_{\min} C_{\min}}{4k} - \lambda \tag{A.77}$$

which is guaranteed to hold if

$$\lambda \leq \theta_{\min} C_{\min}/8k. \tag{A.78}$$

We also combine the two last conditions on $\widehat{Q}$ and obtain the sufficient condition

$$\|\!| \widehat{Q}_{[p],S^0} - Q^0_{[p],S^0} |\!\|_\infty \leq \frac{\alpha}{12} \frac{C_{\min}}{\sqrt{k}}. \tag{A.79}$$

Note that $[p] = S^0 \cup (S^0)^c$.

We then impose that both the probability that condition (A.79) on $\widehat{Q}$ fails and the probability that condition (A.3) on $\widehat{G}$ fails are upper bounded by $\delta/2$. Using Proposition A.2, we can guarantee that the condition on $\widehat{G}$ fails with probability smaller than $\delta/2$ if we set

$$\lambda^2 = 36\alpha^{-2}(n\eta D)^{-1} \log(4p/\delta). \tag{A.80}$$

Since we also want (A.78) to be satisfied, we substitute $\lambda$ from the previous expression in (A.78) and we conclude that $n$ must satisfy

$$n \geq 2304 k^2 {C_{\min}}^{-2} {\theta_{\min}}^{-2} \alpha^{-2} (D\eta)^{-1} \log(4p/\delta). \tag{A.81}$$

Since in addition, the application of the probability bound in Proposition A.2 requires that

$$\frac{\lambda^2 \alpha^2}{9} < 1/4, \tag{A.82}$$

we need to impose further that,

$$n \geq 16 (D\eta)^{-1} \log(4p/\delta). \tag{A.83}$$

To use Corollary A.4 for computing the probability that the condition on $\widehat{Q}$ holds we need that,

$$n\eta > 3/D, \tag{A.84}$$

and

$$\frac{\alpha C_{\min}}{12\sqrt{k}} < 2kD^{-1}. \tag{A.85}$$

The last expression imposes the following conditions on $k$,

$$k^{3/2} > 24^{-1} \alpha C_{\min} D. \tag{A.86}$$

We then have that the condition on $\widehat{Q}$ holds with probability smaller than $1/2$ if

$$n > 4608 \eta^{-1} k^3 \alpha^{-2} {C_{\min}}^{-2} D^{-3} \log 4pk/\delta. \tag{A.87}$$

Note that the restriction (A.86) on $k$ looks unfortunate but, since $k \geq 1$, we can actually show it always holds. Just notice $\alpha < 1$ and that

$$\begin{aligned} \sigma_{\max}(Q^0_{S^0,S^0}) \leq \sigma_{\max}(Q^0) \leq \frac{\eta}{1-\sigma_{\max}} \\ \Leftrightarrow D \leq \sigma^{-1}_{\max}(Q^0_{S^0,S^0}) \end{aligned} \tag{A.88}$$

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 $n$ (namely (A.81), (A.83), (A.84) and (A.87)) 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 (A.81), and (A.87). A little more of algebra shows that these two inequalities are satisfied if

$$n\eta > \frac{10^4 k^2 (kD^{-2} + \theta^{-2}_{\min})}{\alpha^2 D C^2_{\min}} \log(4pk/\delta). \tag{A.89}$$

This concludes the proof of Theorem V.1.

## Appendix B
## Proofs of the Lower Bounds on the Sample-Complexity of General Reconstruction Algorithms

In this section we prove Theorem II.2 and Theorem V.2 to Theorem V.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.2 and V.4, the SDE admit a unique stationary measure, with bounded covariance $Q^0$. Recall that

$$Q^0 = \mathbb{E}\{x(0)x(0)^*\} - \mathbb{E}\{x(0)\}(\mathbb{E}\{x(0)\})^* \tag{B.1}$$

$$= \mathbb{E}\{x(t)x(t)^*\} - \mathbb{E}\{x(t)\}(\mathbb{E}\{x(t)\})^*. \tag{B.2}$$

### A. A general bound for linear SDEs

Before passing to the actual proofs, it is useful to establish a general bound for linear SDEs (I.6) with symmetric interaction matrix $\Theta^0$.

**Lemma B.1.** *Assume that $\{x(t)\}_{t\geq 0}$ is a stationary process generated by the linear SDE (I.6), with $\Theta^0$ symmetric. Let $\widehat{M}_T(X^T)$ be an estimator of $M(\Theta^0)$ based on $X^T$. If $\mathbb{P}(\widehat{M}_T(X^T) \neq M(\Theta^0)) < \frac{1}{2}$ then*

$$T \geq \frac{H(M(\Theta^0)) - \log(|\mathcal{M}|) - 2I(\Theta^0; x(0)) - 2}{\frac{1}{2}\mathrm{Tr}\{\mathbb{E}\{-\Theta^0\} - (\mathbb{E}\ \{-(\Theta^0)^{-1}\})^{-1}\}\}}, \quad \text{(B.3)}$$

*where $|\mathcal{M}|$ is the size of the alphabet of $M(\Theta^0)$.*

*Proof:* The bound follows from Corollary V.3 after showing that

$$\mathbb{E}_{x(0)}\{\mathrm{Var}_{\Theta^0|x(0)}(\Theta^0 x(0))) \leq (1/2)\mathrm{Tr}\{\mathbb{E}\{-\Theta^0\} - (\mathbb{E}\ \{-(\Theta^0)^{-1}\})^{-1}\}. \quad \text{(B.4)}$$

First note that

$$\mathbb{E}_{x(0)}\{\mathrm{Var}_{\Theta^0|x(0)}(\Theta^0 x(0))\} = \mathbb{E}_{x(0)}\|\Theta^0 x(0) - \mathbb{E}_{\Theta^0|x(0)}(\Theta^0 x(0)|x(0))\|_2^2. \quad \text{(B.5)}$$

The quantity in (B.5) can be thought of as the $\ell_2$-norm error of estimating $\Theta^0 x(0)$ based on $x(0)$ using $\mathbb{E}_{\Theta^0|x(0)}(\Theta^0 x(0)|x(0))$. Since conditional expectation is the minimal mean square error estimator, replacing $\mathbb{E}_{\Theta^0|x(0)}(\Theta^0 x(0)|x(0))$ by any estimator of $\Theta^0 x(0)$ based on $x(0)$ gives an upper bound for the expression in (B.5). We choose as an estimator a linear estimator, i.e., an estimator of the form $Bx(0)$ where $B = (\mathbb{E}_{\Theta^0}\Theta^0 Q^0)(\mathbb{E}_{\Theta^0} Q^0)^{-1}$. We then have

$$\begin{aligned}\mathbb{E}_{x(0)}||\Theta^0 x(0) - \mathbb{E}_{\Theta^0|x(0)}(\Theta^0 x(0)|x(0))||_2^2 &\leq \mathbb{E}_{x(0)}||\Theta^0 x(0) - Bx(0)||_2^2 \\ &= \mathrm{Tr}\{\mathbb{E}\{\Theta^0 x(0)(x(0))^*\Theta^{0^*}\}\} \\ &\quad - 2\mathrm{Tr}\{B\mathbb{E}\{x(0)(x(0))^*\Theta^{0^*}\}\} \\ &\quad + \mathrm{Tr}\{B\mathbb{E}\{x(0)(x(0))^*\}B^*\}. \end{aligned} \quad \text{(B.6)}$$

Furthermore, for a linear system, $Q^0$ satisfies the Lyapunov equation $\Theta^0 Q^0 + Q^0(\Theta^0)^* + I = 0$. For $\Theta^0$ symmetric, this implies $Q^0 = -(1/2)(\Theta^0)^{-1}$. Substituting this expression in (B.5) and (B.6) finishes the proof. ∎

### B. Proof of Theorem II.2

We prove Theorem II.2 by showing that the same complexity bound holds in the case when we are trying to estimate the signed support of $\Theta^0$ for an $\Theta^0$ 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 $\Theta^0$ is proved using Lemma B.1 together with Lemma B.2 about random matrices.

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

The following lemma encapsulates the necessary random matrix calculations to prove the complexity bound for random matrices.

**Lemma B.2.** *Let $\Theta$ be a random matrix defined as above and*

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

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

$$Q(\theta_{\min}, k, \rho) \leq \min\Big\{\frac{C' k\theta_{\min}^2}{\rho}, \frac{k\theta_{\min}}{\sqrt{k-1}}\Big\}. \quad \text{(B.7)}$$

*Proof:* First notice that

$$\lim_{p\to\infty}\frac{1}{p}\mathbb{E}\mathrm{Tr}\{-\Theta\} = \lim_{p\to\infty}\mathbb{E}(\gamma) + 2\theta_{\min}\sqrt{k-1} = \rho + 2\theta_{\min}\sqrt{k-1} \quad \text{(B.8)}$$

since by Kesten-McKay law [49], for large $p$, the spectrum of $\tilde{\Theta}$ has support in $(-\epsilon - 2\theta_{\min}\sqrt{k-1}, 2\theta_{\min}\sqrt{k-1} + \epsilon)$ with high probability. Notice that unless we randomize each entry of $\tilde{\Theta}$ with $\{-1, +1\}$ values, every $\tilde{\Theta}$ will have $k$ as its largest eigenvalue and the above limit will not hold.

For the second term, $\mathrm{Tr}\{(\mathbb{E}(-\Theta^{-1}))^{-1}\}$, 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,

$$\begin{aligned}\frac{1}{p}\mathrm{Tr}\{(\mathbb{E}(-\Theta^{-1}))^{-1}\} &= \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_i} \\ &\geq \frac{1}{\frac{1}{p}\sum_{i=1}^{p}\lambda_i} \\ &= \frac{1}{\mathbb{E}\{\frac{1}{p}\mathrm{Tr}\{(-\Theta)^{-1}\}\}} \end{aligned} \quad \text{(B.9)}$$

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

$$\lim_{p\to\infty}\mathbb{E}\Big\{\frac{1}{p}\mathrm{Tr}\{(-\Theta)^{-1}\}\Big\} = \mathbb{E}\Big\{\lim_{p\to\infty}\frac{1}{p}\mathrm{Tr}\{(-\Theta)^{-1}\}\Big\} = \frac{1}{\theta_{\min}}G(k, \rho/\theta_{\min} + 2\sqrt{k-1}) \quad \text{(B.10)}$$

where

$$G(k, z) = \int\frac{-1}{\nu - z}\mathrm{d}\mu(\nu). \quad \text{(B.11)}$$

Above, $\mu(\nu)$ is the Kesten-McKay distribution and, inside its support, $\nu \in [-2\sqrt{k-1}, -2\sqrt{k-1}]$, it is defined by

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

The integral (B.11) can be computed exactly

$$G(k, z) = -\frac{(k-2)z - k\sqrt{-4k + z^2 + 4}}{2\,(z^2 - k^2)}. \quad \text{(B.12)}$$

From the closed form expression for $G(k,z)$ one can see that

$$\lim_{\rho\to 0} Q(\theta_{\min},k,\rho) = \frac{\theta_{\min}k}{\sqrt{k-1}} \quad \text{and} \tag{B.13}$$

$$\lim_{\rho\to\infty} \rho\, Q(\theta_{\min},k,\rho) = k(\theta_{\min})^2. \tag{B.14}$$

Finally, notice that $Q(\theta_{\min},k,\rho)/\theta_{\min}$ can be see as function of $k$ and $\rho/\theta_{\min}$ alone. In addition, because it is strictly decreasing with $\rho/\theta_{\min}$, the limits above imply that $Q(\theta_{\min},k,\rho)/\theta_{\min} \le k/\sqrt{k-1}$ and that there is a large $C'$ such that $Q(\theta_{\min},k,\rho)/\theta_{\min} \le C'k\theta_{\min}/\rho$ for $\rho$ sufficiently high. From these two bounds, the proof follows. ∎

*Proof (Theorem II.2):* We now show that when $\Theta^0$ is chosen at random from $\mathcal{A}^{(S)}$, the right hand side of (B.3) reduces to the right hand side of (II.8) in Theorem II.2.

Starting from the bound of Lemma B.1, we divide both terms in the numerator and the denominator by $p$. Notice that we can ignore the term $2/p$ in the numerator when $p\to\infty$.

Recall that $\Theta^0$ is built from the adjacency matrix of a regular graph chosen uniformly at random and whose entries have had their sign flipped with probability $1/2$. Therefore, since $M(\Theta^0)$ is the sign-support of $\Theta^0$, we have $H(M(\Theta^0)) = \log(|\mathcal{M}|)$. Hence, we can write $p^{-1}(2H(M(\Theta^0)) - \log(|\mathcal{M}|)) = p^{-1}\log(|\mathcal{M}|)$. In addition, $|\mathcal{M}| = 2^{pk/2}|\mathcal{R}|$, where $|\mathcal{R}|$ is the number of regular graphs of degree $k$ on $p$ nodes and $2^{pk/2}$ accounts for the sign flips in the non-zero non-diagonal entries [9]. From [50], we know that $\log(|\mathcal{R}|) \ge Cpk\log(2p/k)$ for small enough constant $C$. And therefore, $\log(|\mathcal{M}|)/p \ge (k/2)\log(2) + Ck\log(2p/k) \ge C'k\log(2p/k)$ for all $p$ large enough and small enough $C'$.

Lemma B.2 gives an upper bound on the denominator when $p\to\infty$.

To finish the proof of Theorem II.2, we show that $\lim_{p\to\infty} I(x(0);\Theta^0)/p \le 1$. This finishes the proof since, after multiplying by a small enough constant (only dependent on $k$), the bound obtained by replacing the numerator and denominator with the above limiting lower bounds will be valid for all $p$ large enough.

First notice that $h(x(0)) \le (1/2)\log(2\pi e)^p|\mathbb{E}(Q^0)|$ and hence,

$$I(x(0);\Theta^0) = h(x(0)) - h(x(0)|\Theta^0) \tag{B.15}$$

$$\le \frac{1}{2}\log(2\pi e)^p|\mathbb{E}(Q^0)| - \mathbb{E}\frac{1}{2}\log(2\pi e)^p|Q^0|, \tag{B.16}$$

where $Q^0 = -(1/2)(\Theta^0)^{-1}$ is the covariance matrix of the stationary process $x(t)$ and $|.|$ denotes the determinant of a matrix. Then we write, $I(x(0);\Theta^0) \le (1/2)\log|\mathbb{E}(-(\beta\Theta^0)^{-1})| + (1/2)\mathbb{E}\log(|-\beta\Theta^0|) \le \frac{1}{2}\mathrm{Tr}\mathbb{E}(-I-(\beta\Theta^0)^{-1}) + \frac{1}{2}\mathbb{E}\mathrm{Tr}\{-I-\beta\Theta^0\}$ where $\beta>0$ is an arbitrary rescaling factor and the last inequality follows from the matrix inequality $\log(I+(.)) \le \mathrm{Tr}(.)$. From this and equations (B.8) and (B.10) it follows that,

$$\lim_{p\to\infty}\frac{1}{p}I(x(0);\Theta^0) \le -1 + (1/2)(\beta' z + \beta'^{-1}G(k,z)) \tag{B.17}$$

where $z = \rho/\theta_{\min} + 2\sqrt{k-1}$ and $\beta' = \beta\theta_{\min}$. To finish, note that optimizing over $\beta'$ and then over $z$ gives,

$$\beta' z + \beta'^{-1}G(k,z) \le 2\sqrt{zG(k,z)} \le \sqrt{\frac{8(k-1)}{k-2}} \le 4. \tag{B.18}$$

∎

[9] Notice that diagonal entries are constant and equal to $\gamma + 2\theta_{\min}\sqrt{k-1}$.

### C. Proof of Theorem V.4

The proof of this theorem follows closely the proof of Theorem II.2. Basically, the claim follows by proving that the bound (V.9) holds for an $\Theta^0$ chosen at random with a distribution supported on $\mathcal{A}^{(D)}$.

Again, in order to lower bound the sample-complexity for random matrices, we make use of Lemma B.1.

Now, however, we construct the random matrix $\Theta^0$ as follows. Let $\tilde{\Theta}^0$ be a random symmetric matrix with zero-diagonal and with $\{\theta_{ij}\}_{i<j}$ i.i.d. random variables where $\mathbb{P}(\theta_{ij}=\theta_{\min}) = \mathbb{P}(\theta_{ij}=-\theta_{\min}) = 1/4$, and $\mathbb{P}(\theta_{ij}=0) = 1/2$. Notice that the second moment of each entry $i\neq j$ is $\mathbb{E}(\Theta^2_{ij}) = \theta^2_{\min}/2 \equiv \alpha$. We then define $\Theta^0 = -(\gamma+2\sqrt{\alpha})I + \tilde{\Theta}^0/\sqrt{p}$ where $\gamma = \gamma(\tilde{\Theta}^0)$ is the smallest value that guarantees that $\lambda_{\min}(-\Theta) \ge \rho$.

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

**Lemma B.3.** *Let $\Theta$ be a random matrix defined as above and*

$$Q(\theta_{\min},\rho) \equiv \lim_{p\to\infty}\frac{1}{p}\{\mathrm{Tr}\{\mathbb{E}(-\Theta)\} - \mathrm{Tr}\{(\mathbb{E}(-\Theta^{-1}))^{-1}\}\}. \tag{B.19}$$

*Then, there exists a constant $C'$ such that*

$$Q(A_{\min},\rho) \le \min\{\frac{C'\theta^2_{\min}}{2\rho}, \frac{\theta_{\min}}{\sqrt{2}}\}. \tag{B.20}$$

*Proof:* Using Wigner's Semicircle law for random symmetric matrices [51] and the bound described in (B.9) it follows that,

$$\lim_{p\to\infty}\frac{1}{p}\{\mathrm{Tr}\{\mathbb{E}(-\Theta)\} = \rho + 2\sqrt{\alpha} \quad \text{and} \tag{B.21}$$

$$C(\alpha,\rho) \equiv \lim_{p\to\infty}\mathbb{E}\{\frac{1}{p}\mathrm{Tr}\{(-\Theta)^{-1}\}\} \tag{B.22}$$

$$= \frac{-\sqrt{\rho\,(4\sqrt{\alpha}+\rho)} + 2\sqrt{\alpha} + \rho}{2\alpha}. \tag{B.23}$$

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 $\alpha$ or $\rho$ such that $\sqrt{\alpha}G(\rho/\alpha) \le \sqrt{\alpha}\min\{1, C'\sqrt{\alpha}/\rho\}$. The result now follows by replacing $\alpha = \Theta^2_{\min}/2$. ∎

*Proof (Theorem V.4):* Like in the proof of Theorem II.2 we start by dividing both numerator and denominator of (B.3) in Lemma B.1 by $p$. By multiplying the resulting expression by an appropriately small constant we can replace

the denominator and $\lim_{p\to\infty} I(x(0);\Theta^0)/p$ by their limits when $p \to \infty$ and get an expression that is still valid for all $p$ large enough.

Let us produce a lower bound for $2H(M(\Theta^0)) - \log(|\mathcal{M}|)$. First notice that we again have, $H(M(\Theta^0)) = \log(|\mathcal{M}|)$ since every $M(\Theta^0)$ is equally likely. Therefore, $2H(M(\Theta^0)) - \log(|\mathcal{M}|) = H(M(\Theta^0))$.

Since $H(M(\Theta^0))/p = \frac{p-1}{2} H(\{1/2, 1/4, 1/4\}) \geq \frac{(p-1)}{4} \log 2$ [10], and since by Lemma B.3 we already know the limiting expression of the denominator, all we have to do is find $\lim_{p\to\infty} I(x(0);\Theta^0)/p$. By an analysis very similar to that in the proof of Theorem II.2 one can show that

$$\lim_{p\to\infty} \frac{1}{p} I(x(0);\Theta^0) \leq -1 + \sqrt{(z+2)C(1,z)} \leq 1. \quad \text{(B.24)}$$

where $C(\alpha, \rho)$ was defined in (B.22), which finishes the proof. ■

### *D. Proof of Theorem V.5*

The proof consists in evaluating the lower bound in Corollary V.3. We prove the theorem by showing the bound holds for functions uniformly chosen over a specific subset of $\mathcal{A}^{(N)}$. Consider the set of functions such that for each possible support of a $p \times p$ matrix with at most $k$ non-zero entries per row there is one and only one function in the family with $JF$ having that support for all $x$. Note that this implies that, when evaluating V.3, here with $\mathcal{M} = \mathcal{A}^{(N)}$, we have $\log(|\mathcal{M}|) = H(M(\Theta^0))$. Hence $2H(M(\Theta^0)) - \log(|\mathcal{M}|) = H(M(\Theta^0))$.

Now notice that $\mathbb{E}_{x(0)} \mathrm{Var}_{x(0)|\Theta^0} F(x(0);\Theta^0) \leq \mathbb{E}(||F(x(0);\Theta^0)||^2)$. Secondly notice that, if $x$ and $x'$ only differ on the $j^{th}$ component and $(JF)_{ij} \neq 0$ then $|F_i(x;\Theta^0)| \leq |F_i(x';\Theta^0)| + D||x' - x||$. Since $JF$ has at most $k$ non-zero entries per row, we get that for any $x$ and $x'$, $|F_i(x;\Theta^0)| \leq |F_i(x';\Theta^0)| + kD||x' - x||$. If $x = x(0)$ and $x' = \mathbb{E}_{x(0)|\Theta^0}(x(0)|\Theta^0)$ then squaring the previous expression and taking expectations gives us $\mathbb{E}_{x(0)|\Theta^0}(F_i(x;\Theta^0)^2|\Theta^0) \leq 2F_i(x';\Theta^0)^2 + 2k^2D^2B$. From this we get that $\mathbb{E}(||F(x(0);\Theta^0)||^2)/p \leq C + 2k^2D^2B$ where $C$ is a constant independent of $\Theta^0$. For this sub family of functions we have $H(M(\Theta^0)) \geq pk \log(p/k)$ (see [50]). By (B.16), we know that $I(x(0);\Theta^0) \leq (1/2)\log((2\pi e)^p |\mathbb{E}Q^0|) - (1/2)\mathbb{E}\log((2\pi e)^p |Q^0|)$. The first term, which is the entropy of a $p$-dimensional Gaussian with covariance matrix $\mathbb{E}\{Q^0\}$, can be upper bounded by the sum of the entropy of its individual components, which have variance upper bounded by $B$. Finally, since $\lambda_{\min}(Q^0) \geq L$, we have $\log|Q^0| \geq p \log L$ and therefore $I(x(0);\Theta^0) \leq p/2 \log B/L$, which completes the proof. ■

[10] $H(\{1/2, 1/4, 1/4\})$ is the entropy of the distribution $\{1/2, 1/4, 1/4\}$.