Electronic Journal of Statistics
Vol. 14 (2020) 4395–4420
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1775

On Dantzig and Lasso estimators of the drift in a high dimensional Ornstein-Uhlenbeck model[*]

Gabriela Ciołek and Dmytro Marushkevych and Mark Podolskij

Department of Mathematics
Faculty of Science, Technology and Medicine, University of Luxembourg
6 Avenue de la Fonte, 4364 Esch-sur-Alzette, Luxembourg
e-mail: gabriela.ciolek@uni.lu; dmytro.marushkevych@uni.lu; mark.podolskij@uni.lu

Abstract: In this paper we present new theoretical results for the Dantzig and Lasso estimators of the drift in a high dimensional Ornstein-Uhlenbeck model under sparsity constraints. Our focus is on oracle inequalities for both estimators and error bounds with respect to several norms. In the context of the Lasso estimator our paper is strongly related to [11], where the same problem was investigated under row sparsity. We improve their rates and also prove the restricted eigenvalue property solely under ergodicity assumption on the model. Finally, we demonstrate a numerical analysis to uncover the finite sample performance of the Dantzig and Lasso estimators.

MSC2020 subject classifications: Primary 62M05, 60G15; secondary 62H12, 62M99.
Keywords and phrases: Dantzig estimator, high dimensional statistics, Lasso, Ornstein-Uhlenbeck process, parametric estimation.

Received July 2020.

Contents

arXiv: 2008.00847

[*] The authors gratefully acknowledge financial support of ERC Consolidator Grant 815703 “STAMFORD: Statistical Methods for High Dimensional Diffusions”.

1. Introduction

During past decades an immense progress has been achieved in statistics for stochastic processes. Nowadays, comprehensive studies on statistical inference for diffusion processes under low and high frequency observation schemes can be found in monographs [14, 17, 19]. Most of the existing literature is considering a fixed dimensional parameter space, while a high dimensional framework received much less attention in the diffusion setting.

Since the pioneering work of McKean [20, 21], high dimensional diffusions entered the scene in the context of modelling the movement of gas particles. More recently, they found numerous applications in economics and biology, among other disciplines [3, 6, 9]. Typically, high dimensional diffusions are studied in the framework of *mean field theory*, which aims at bridging the interaction of particles at the microscopic scale and the mesoscopic features of the system (see e.g. [28] for a mathematical study). In physics particles are often assumed to be statistically equal, but this homogeneity assumption is not appropriate in other applications. For instance, in [6] high dimensional SDEs are used to model the wealth of trading agents in an economy, who are often far from being equal in their trading behaviour. Another example is the flocking phenomenon of individuals [3], where it seems natural to assume that there are only very few “leaders” who have a distinguished role in the community. These examples motivate to investigate statistical inference for diffusion processes under sparsity constraints.

This paper is focusing on statistical analysis of a d-dimensional Ornstein-Uhlenbeck model of the form

$$dX_t = -A_0 X_t dt + dW_t, \qquad t \geq 0, \tag{1.1}$$

defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, with underlying observation $(X_t)_{t\in[0,T]}$. Here W denotes a standard d-dimensional Brownian motion and $A_0 \in \mathbb{R}^{d\times d}$ represents the unknown interaction matrix. Ornstein-Uhlenbeck processes are one of the most basic parametric diffusion models. When the dimension d is fixed and $T \to \infty$, statistical estimation of the parameter A_0 has been discussed in several papers. Asymptotic analysis of the maximum likelihood estimator in the ergodic case can be found in e.g. [17] while investigations of the non-ergodic setting can be found in [16, 18]. The adaptive Lasso estimation for multivariate diffusion models has been investigated in [12].

Our main goal is to study the estimation of A_0 under sparsity constraints in the large d/large T setting. Such a mathematical problem finds its main motivation in the analysis of bank connectedness whose wealth is modelled by the diffusion process X. This field of economics, which studies linkages between a large number of banks associated with e.g. asset/liability positions and contractual relationships, is key to understanding systemic risk in a global economy [13]. Typically, the connectivity structure, which is represented by the parameter A_0, is quite sparse since only few financial players are significant in an economy, and the main focus is on estimation of non-zero components of A_0.

Theoretical results in the high dimensional diffusion setting are rather scarce. In this context we would like to mention the Dantzig selector which was introduced in [5] and primarily designed for linear regression models. More specifically, [5] established sharp non-asymptotic bounds on the l_2-error in the estimated coefficients and proved that the error is within a factor of $\log(d)$ of the error that would have been reached if the locations of the non-zero coefficients were known. Further extensions of the aforementioned results can be found in [10] and [24], which study the Dantzig selector for discretely observed linear diffusions and support recovery for the drift coefficient, respectively. Our work is closely related to the recent article [11], where estimation of A_0 under row sparsity has been investigated. The authors propose to use the classical Lasso approach and derive upper and lower bounds for the estimation error. We build upon their analysis and provide oracle inequalities and non-asymptotic theory for the the Lasso and Dantzig estimators. In comparison to [11], we obtain an improved upper bound for the Lasso estimator, which essentially matches the theoretical lower bound, and also show that the *restricted eigenvalue property* is automatically satisfied under ergodicity condition on the model (1.1) (in [11] the extra assumption (H4) has been imposed). The latter is proved via Malliavin calculus methods proposed in [22]. Moreover, we show that the Lasso and Dantzig estimators are asymptotically efficient, which is a well known fact in linear regression models (cf. [2]). Finally, we present a simulation study to uncover the finite sample properties of both estimators.

The paper is organised as follows. Section 2 is devoted to the exposition of the classical estimation theory in the fixed dimensional setting and to definition of the Lasso and Dantzig estimators. Concentration inequalities for various stochastic terms are derived in Section 3. In particular, we show the restricted eigenvalue property under the ergodicity assumption via Malliavin calculus methods. In Section 4 we present oracle inequalities and error bounds for both estimators. Numerical simulation results are demonstrated in Section 5. Finally, some proofs are collected in Section 6.

2. The model, notation and main definitions

2.1. Notation

In this subsection we briefly introduce the main notations used throughout the paper. For a vector or a matrix x the transpose of x is denoted by $x^\top$. For $p \geq 1$

and $A \in \mathbb{R}^{d_1 \times d_2}$, we define the l_p-norm as

$$\|A\|_p := \left(\sum_{1 \le i \le d_1, 1 \le j \le d_2} |A_{ij}|^p \right)^{1/p}.$$

We denote by $\|A\|_\infty = \lim_{p \to \infty} \|A\|_p$ the maximum norm and set $\|A\|_0 := \sum_{1 \le i \le d_1, 1 \le j \le d_2} 1_{\{A_{ij} \neq 0\}}$. We associate to the Frobenius norm $\| \cdot \|_2$ the scalar product

$$\langle A_1, A_2 \rangle_{\mathrm{F}} := \mathrm{tr}(A_1^\top A_2), \qquad A_1, A_2 \in \mathbb{R}^{d_1 \times d_2},$$

where tr denotes the trace. For a symmetric matrix $A \in \mathbb{R}^{d \times d}$ we write $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ for the largest and the smallest eigenvalue of A, respectively. We denote by $\|A\|_{\mathrm{op}} := \sqrt{\lambda_{\max}(A^\top A)}$ the operator norm of $A \in \mathbb{R}^{d \times d}$. For any $J \subset \{1, \dots, d_1\} \times \{1, \dots, d_2\}$ and $A \in \mathbb{R}^{d_1 \times d_2}$, the matrix $A_{|J}$ is defined via

$$(A_{|J})_{ij} := A_{ij} 1_{(i,j) \in J}.$$

For a quadratic matrix $A \in \mathbb{R}^{d \times d}$, diag$(A)$ stands for the diagonal matrix satisfying diag$(A)_{ii} = A_{ii}$. We also introduce the notation

$$\mathcal{C}(s, c_0) := \left\{ A \in \mathbb{R}^{d \times d} \setminus \{0\} : \ \|A\|_1 \le (1 + c_0) \|A_{|\mathcal{I}_s(A)}\|_1 \right\}, \qquad (2.1)$$

where $c_0 > 0$ and $\mathcal{I}_s(A)$ is a set of coordinates of s largest elements of A. Furthermore, vec denotes the vectorisation operator and $\otimes$ stands for the Kronecker product. For $z \in \mathbb{C}$ we denote by $\mathfrak{Re}(z)$ (resp. $\mathfrak{Im}(z)$) the real (resp. imaginary) part of z. Finally, for stochastic processes $(X_t)_{t \in [0,T]}, (Y_t)_{t \in [0,T]} \in L^2([0,T], dt)$ we introduce the scalar product

$$\langle X, Y \rangle_{L^2} := \frac{1}{T} \int_0^T X_t Y_t dt.$$

2.2. The setting and fixed dimensional theory

We consider a d-dimensional Ornstein-Uhlenbeck process introduced in (1.1). Throughout this paper the matrix A_0 is assumed to satisfy the following condition:

(H) Matrix A_0 is diagonalisable with eigenvalues $\theta_1, \dots, \theta_d \in \mathbb{C}$, i.e.

$$A_0 = P_0 \mathrm{diag}(\theta_1, \dots, \theta_d) P_0^{-1},$$

where the column vectors of P_0 are eigenvectors of A_0. Furthermore, the eigenvalues $\theta_1, \dots, \theta_d \in \mathbb{C}$ have strictly positive real parts:

$$\mathfrak{r}_0 := \min_{1 \le j \le d} (\mathfrak{Re}(\theta_i)) > 0. \qquad (2.2)$$

It is well known that under condition (H) the stochastic differential equation (1.1) exhibits a unique stationary solution, which can be written explicitly as

$$X_t = \int_{-\infty}^{t} \exp\left(-(t-s)A_0\right) dW_t.$$

In this case we have that

$$X_t \sim \mathcal{N}(0, C_\infty) \qquad \text{with} \qquad C_\infty := \int_0^\infty \exp(-sA_0)\exp(-sA_0^\top)ds.$$

We assume that the complete path $(X_t)_{t\in[0,T]}$ is observed and we are interested in estimating the unknown parameter A_0. Let us briefly recall the classical maximum likelihood theory when d is fixed and $T \to \infty$. When $\mathbb{P}_A^T$ denotes the law of the process (1.1) with transition matrix A restricted to $\mathcal{F}_T$, the log-likelihood function is explicitly computed via Girsanov's theorem as

$$\log(\mathbb{P}_A^T/\mathbb{P}_0^T) = -\int_0^T (AX_t)^\top dX_t - \frac{1}{2}\int_0^T (AX_t)^\top (AX_t)dt. \tag{2.3}$$

Consequently, the maximum likelihood estimator $\widehat{A}_{\mathrm{ML}}$ is given by

$$\widehat{A}_{\mathrm{ML}} = -\Big(\int_0^T dX_t X_t^\top\Big)\Big(\int_0^T X_t X_t^\top dt\Big)^{-1}.$$

Under condition (H) the estimator $\widehat{A}_{\mathrm{ML}}$ is asymptotically normal, i.e.

$$\sqrt{T}\left(\mathrm{vec}(\widehat{A}_{\mathrm{ML}}) - \mathrm{vec}(A_0)\right) \xrightarrow{\mathrm{d}} \mathcal{N}_{d^2}\left(0, C_\infty^{-1}\otimes \mathrm{id}\right) \tag{2.4}$$

with id denoting the d-dimensional identity matrix. Indeed, we have the identity $\widehat{A}_{\mathrm{ML}} - A_0 = -\varepsilon_T \widehat{C}_T^{-1}$ with

$$\varepsilon_T := \frac{1}{T}\int_0^T dW_t X_t^\top \qquad \text{and} \qquad \widehat{C}_T := \frac{1}{T}\int_0^T X_t X_t^\top dt \xrightarrow{\text{a.s.}} C_\infty, \tag{2.5}$$

and the result (2.4) follows from the standard martingale central limit theorem. We refer to [17, p. 120–124] for a more detailed exposition.

When assumption (H) is violated the asymptotic theory for the maximum likelihood estimator $\widehat{A}_{\mathrm{ML}}$ is more complex. If some eigenvalues θ_j satisfy $\mathfrak{Re}(\theta_i) < 0$ exponential rates appear as it has been shown in [18]. A further application of Ornstein-Uhlenbeck processes to co-integration is discussed in [16], where the condition $\mathfrak{Re}(\theta_i) = 0$ appears for some i's.

2.3. The Lasso and Dantzig estimators

Now we turn our attention to large d/large T setting. We consider the Ornstein-Uhlenbeck model (1.1) satisfying the assumption (H) and assume that the unknown transition matrix A_0 satisfies the constraint

$$\|A_0\|_0 \le s_0. \tag{2.6}$$

We remark that due to condition (2.2) it must necessarily hold that $s_0 \geq d$. A standard approach to estimate A_0 under the sparsity constraint (2.6) is the Lasso method, which has been investigated in [11] in the framework of an Ornstein-Uhlenbeck model. The Lasso estimator is defined as

$$\widehat{A}_{\mathrm{L}} := \operatorname*{argmin}_{A \in \mathbb{R}^{d\times d}} \left(\mathcal{L}_T(A) + \lambda \|A\|_1 \right) \quad \textit{with} \qquad \mathcal{L}_T(A) := -\frac{1}{T} \log(\mathbb{P}_A^T / \mathbb{P}_0^T), \tag{2.7}$$

where $\lambda > 0$ is a tuning parameter. We remark that $\widehat{A}_{\mathrm{L}}$ can be computed efficiently, since it is a solution of a convex optimisation problem.

Next, we are going to introduce the Dantzig estimator of the parameter A_0. According to (2.3) the quantity $\mathcal{L}_T(A)$ can be written as

$$\mathcal{L}_T(A) = \operatorname{tr}\left((\varepsilon_T - A_0 \widehat{C}_T) A^\top + \frac{1}{2} A \widehat{C}_T A^\top \right) \text{ and } \nabla \mathcal{L}_T(A) = \varepsilon_T - A_0 \widehat{C}_T + A \widehat{C}_T. \tag{2.8}$$

We recall that B belongs to a subdifferential of a convex function $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ at point B_0, $B \in \partial f(B_0)$, if $\langle B, A - B_0 \rangle_{\mathrm{F}} \leq f(A) - f(B_0)$ for all $A \in \mathbb{R}^{d\times d}$. In particular, $B \in \partial \|B_0\|_1$ satisfies the constraint $\|B\|_\infty \leq 1$. A necessary and sufficient condition for the minimiser at (2.7) is the fact that 0 belongs to the subdifferential of the function $A \mapsto \mathcal{L}_T(A) + \lambda \|A\|_1$. This implies that the Lasso estimator $\widehat{A}_{\mathrm{L}}$ satisfies the constraint

$$\|\widehat{A}_{\mathrm{L}} \widehat{C}_T + \varepsilon_T - A_0 \widehat{C}_T\|_\infty \leq \lambda. \tag{2.9}$$

Now, the Dantzig estimator $\widehat{A}_{\mathrm{D}}$ of the parameter A_0 is defined as a matrix with the smallest l_1-norm that satisfies the inequality (2.9), i.e.

$$\widehat{A}_{\mathrm{D}} := \operatorname*{argmin}_{A \in \mathbb{R}^{d\times d}} \left\{ \|A\|_1 : \ \|A \widehat{C}_T + \varepsilon_T - A_0 \widehat{C}_T\|_\infty \leq \lambda \right\}. \tag{2.10}$$

By definition of the Dantzig estimator we have that $\|\widehat{A}_{\mathrm{D}}\|_1 \leq \|\widehat{A}_{\mathrm{L}}\|_1$. In particular, when the tuning parameters λ for Lasso and Dantzig estimators are preset to be the same, then the Lasso estimate is always a feasible solution to the Dantizg selector minimization problem although it may not necessarily be the optimal solution. This implies, that when respective solutions are not identical, the Dantizg selector solution is sparser (in l_1-norm) than the Lasso solution (see [15], Appendix A for details). From the computational point view, the Dantzig estimator can be found numerically via linear programming for convex optimisation with constraints.

The following basic inequality, which is a direct consequence of the fact that $\mathcal{L}_T(\widehat{A}_{\mathrm{L}}) + \lambda \|\widehat{A}_{\mathrm{L}}\|_1 \leq \mathcal{L}_T(A) + \lambda \|A\|_1$ for all $A \in \mathbb{R}^{d\times d}$, provides the necessary basis for the analysis of the error $\widehat{A}_{\mathrm{L}} - A_0$.

Lemma 2.1 ([11, Lemma 3]). *For any $A \in \mathbb{R}^{d\times d}$ and $\lambda > 0$ it holds that*

$$\begin{aligned} \|(\widehat{A}_L - A_0)X\|_{L^2}^2 - \|(A - A_0)X\|_{L^2}^2 \leq 2\langle \varepsilon_T, A - \widehat{A}_L \rangle_F - \|(A - \widehat{A}_L)X\|_{L^2}^2 \\ + 2\lambda(\|A\|_1 - \|\widehat{A}_L\|_1), \end{aligned}$$

where the quantity ε_T is defined in (2.5).

From Lemma 2.1 it is obvious that we require a good control over martingale term $\langle \varepsilon_T, V\rangle_{\mathrm{F}}$ for certain matrices $V \in \mathbb{R}^{d\times d}$ to get an upper bound on the prediction error $\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}$. Another important ingredient is the *restricted eigenvalue property*, which is a standard requirement in the analysis of Lasso estimators (see e.g. [2, 4]). In our setting the restricted eigenvalue property amounts in showing that

$$\inf_{V\in\mathcal{C}(s,c_0)} \frac{\|VX\|_{L^2}^2}{\|V\|_2^2} \text{ is bounded away from 0 with high probability.}$$

Interestingly, the latter is a consequence of the model assumption (H) and not an extra condition as in the framework of linear regression. This has been noticed in [11], but an additional condition (H4) was required which is in fact not needed as we will show in the next section.

In order to establish the connection between the Dantzig and the Lasso estimators we will show the inequality

$$\Big| \|(\widehat{A}_{\mathrm{D}} - A_0)X\|_{L^2} - \|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2} \Big| \le c\|\widehat{A}_{\mathrm{L}}\|_0\lambda^2$$

for a certain constant $c > 0$, which holds with high probability. Once the term $\|\widehat{A}_{\mathrm{L}}\|_0$ is controlled, we deduce statements about the error term $\widehat{A}_{\mathrm{D}} - A_0$ via the corresponding analysis of $\widehat{A}_{\mathrm{L}} - A_0$.

3. Concentration bounds for the stochastic terms

In this section we derive various concentration inequalities, which play a central role in the analysis of the estimators $\widehat{A}_{\mathrm{L}}$ and $\widehat{A}_{\mathrm{D}}$.

3.1. The restricted eigenvalue property

This subsection is devoted to the proof of the restricted eigenvalue property. The main result of this subsection relies heavily on some theoretical techniques presented in [22], where Malliavin calculus is applied in order to obtain tail bounds for certain functionals of Gaussian processes. In the following, we introduce some basic notions of Malliavin calculus; we refer to the monograph [23] for a more detailed exposition.

Let $\mathbb{H}$ be a real separable Hilbert space. We denote by $B = \{B(h) :\ h \in \mathbb{H}\}$ an *isonormal Gaussian process* over $\mathbb{H}$. That is, B is a centred Gaussian family with covariance kernel given by

$$\mathbb{E}\big[B(h_1)B(h_2)\big] = \langle h_1, h_2\rangle_{\mathbb{H}}.$$

We shall use the notation $L^2(B) = L^2(\Omega, \sigma(B), \mathbb{P})$. For every $q \ge 1$, we write $\mathbb{H}^{\otimes q}$ to indicate the qth tensor product of $\mathbb{H}$; $\mathbb{H}^{\odot q}$ stands for the symmetric qth tensor. We denote by I_q the isometry between $\mathbb{H}^{\odot q}$ and the qth Wiener chaos of X. It is well-known (see e.g. [23, Chapter 1]) that any random variable $F \in L^2(B)$ admits the *chaotic expansion*

$$F = \sum_{q=0}^{\infty} I_q(f_q), \qquad I_0(f_0) := \mathbb{E}[F],$$

where the series converges in L^2 and the kernels $f_q \in \mathbb{H}^{\odot q}$ are uniquely determined by F. The operator L, called the *generator of the Ornstein-Uhlenbeck semigroup*, is defined as

$$LF := -\sum_{q=1}^{\infty} q I_q(f_q)$$

whenever the latter series converges in L^2. The pseudo inverse L^{-1} of L is defined by $L^{-1}F = -\sum_{q=1}^{\infty} q^{-1} I_q(f_q)$.

Next, let us denote by $\mathcal{S}$ the set of all smooth cylindrical random variables of the form $F = f\big(B(h_1), \ldots, B(h_n)\big)$, where $n \geq 1$, $f : \mathbb{R}^n \to \mathbb{R}$ is a C^∞-function with compact support and $h_i \in \mathbb{H}$. The Malliavin derivative DF of F is defined as

$$DF := \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\big(B(h_1), \ldots, B(h_n)\big) h_i.$$

The space $\mathbb{D}^{1,2}$ denotes the closure of $\mathcal{S}$ with respect to norm $\|F\|_{1,2}^2 := \mathbb{E}[F^2] + \mathbb{E}[\|DF\|_{\mathbb{H}}^2]$. The Malliavin derivative D verifies the following *chain rule*: when $\varphi : \mathbb{R}^n \to \mathbb{R}$ is in C_b^1 (the set of continuously differentiable functions with bounded partial derivatives) and if $(F_i)_{i=1,\ldots,n}$ is a vector of elements in $\mathbb{D}^{1,2}$, then $\varphi(F_1, \ldots, F_n) \in \mathbb{D}^{1,2}$ and

$$D\varphi(F_1, \ldots, F_n) = \sum_{i=1}^{n} \frac{\partial \varphi}{\partial x_i}(F_1, \ldots, F_n) DF_i.$$

The next theorem establishes left and right tail bounds for certain elements $Z \in \mathbb{D}^{1,2}$.

Theorem 3.1 ([22, Theorem 4.1])**.** *Assume that $Z \in \mathbb{D}^{1,2}$ and define the function*

$$g_Z(z) := \mathbb{E}[\langle DZ, -DL^{-1}Z\rangle_{\mathbb{H}} |\ Z = z].$$

Suppose that the following conditions hold for some $\alpha \geq 0$ and $\beta > 0$:

(i) $g_Z(Z) \leq \alpha Z + \beta$ holds $\mathbb{P}$-almost surely,
(ii) The law of Z has a Lebesgue density.

Then, for any $z > 0$, it holds that

$$\mathbb{P}(Z \geq z) \leq \exp\left(-\frac{z^2}{2\alpha z + 2\beta}\right) \qquad \text{and} \qquad \mathbb{P}(Z \leq -z) \leq \exp\left(-\frac{z^2}{2\beta}\right).$$

Now, we apply Theorem 3.1 to certain quadratic forms of the Ornstein-Uhlenbeck process X. The following result is crucial for proving the restricted eigenvalue property.

Proposition 3.2. *Suppose that assumption (H) is satisfied and let $\widehat{C}_T$ be defined as in (2.5). Then it holds for all $x > 0$:*

$$\sup_{v\in\mathbb{R}^d:\ \|v\|_2=1} \mathbb{P}\left(|v^\top(\widehat{C}_T - C_\infty)v| \geq x\right) \leq 2\exp\left(-TH_0(x)\right), \tag{3.1}$$

where the function H_0 is defined as

$$H_0(x) = \frac{\mathfrak{r}_0}{8\mathfrak{p}_0\mathfrak{K}_\infty}\frac{x^2}{x+\mathfrak{K}_\infty}$$

with $\mathfrak{K}_\infty = \lambda_{\max}(C_\infty)$ and $\mathfrak{p}_0 = \|P_0\|_{\mathrm{op}}\|P_0^{-1}\|_{\mathrm{op}}$, and the quantities P_0 and $\mathfrak{r}_0$ are introduced in assumption (H).

Proof. We define the centred stationary Gaussian process $Y_t^v = v^\top X_t$ and note that its covariance kernel is given by $\mathbb{E}[Y_t^vY_s^v] = \rho_v(|t-s|)$ with $\rho_v(r) := v^\top\exp(-rA_0)C_\infty v$. By submultiplicativity of the operator norm we conclude that

$$|\rho_v(r)| \leq \|\exp(-rA_0)\|_{\mathrm{op}}\|C_\infty\|_{\mathrm{op}} \leq \exp(-\mathfrak{r}_0 r)\mathfrak{p}_0\mathfrak{K}_\infty.$$

We observe that $(Y_t^v)_{t\in[0,T]}$ can be considered as an isonormal Gaussian process indexed by a separable Hilbert space $\mathbb{H}$ whose scalar product is induced by the covariance kernel of $(Y_t^v)_{t\in[0,T]}$. In particular, we can write $Y_t^v = B(h_t)$ and $\langle h_t, h_s\rangle_{\mathbb{H}} = \rho_v(|t-s|)$. We introduce the quantity

$$Z_T^v := v^\top(\widehat{C}_T - C_\infty)v = \frac{1}{T}\int_0^T (Y_t^v)^2 - \mathbb{E}[(Y_t^v)^2]dt$$

and notice that Z_T^v is an element of the second order Wiener chaos. Hence, Z_T^v has a Lebesgue density and we have $L^{-1}Z_T^v = -Z_T^v/2$, and we conclude by the chain rule that

$$\begin{aligned}
\langle DZ_T^v, -DL^{-1}Z_T^v\rangle_{\mathbb{H}} &= \frac{1}{2}\|DZ_T^v\|_{\mathbb{H}}^2 \leq \frac{2}{T^2}\int_0^T\int_0^T |Y_t^vY_s^v||\rho_v(t-s)|dtds \\
&\leq \frac{2}{T^2}\int_0^T\int_0^T (Y_t^v)^2|\rho_v(t-s)|dtds \leq \frac{4}{T}\int_0^\infty |\rho_v(r)|dr\big(Z_T^v + \rho_v(0)\big) \\
&\leq \frac{4}{T}\mathfrak{p}_0\mathfrak{K}_\infty\int_0^\infty \exp(-\mathfrak{r}_0 r)dr\big(Z_T^v + \mathfrak{K}_\infty\big) = \frac{4}{T}\frac{\mathfrak{p}_0\mathfrak{K}_\infty}{\mathfrak{r}_0}(Z_T^v + \mathfrak{K}_\infty).
\end{aligned}$$

Consequently, the conditions of Theorem 3.1 are satisfied with $\alpha = \frac{4}{T}\frac{\mathfrak{p}_0\mathfrak{K}_\infty}{\mathfrak{r}_0}$ and $\beta = \frac{4}{T}\frac{\mathfrak{p}_0\mathfrak{K}_\infty^2}{\mathfrak{r}_0}$, which completes the proof of Proposition 3.2 since $\mathbb{P}(|Z_T^v| \geq x) = \mathbb{P}(Z_T^v \geq x) + \mathbb{P}(Z_T^v \leq -x)$. □

The statement of Proposition 3.2 corresponds to assumption (H4) in [11], which has been shown to be valid via a log-Sobolev inequality only when A_0 is *symmetric* (cf. [11, Theorem]). In other words, the extra assumption (H4) is not required as it directly follows from the modelling setup.

The next theorem proves the restricted eigenvalue property.

Theorem 3.3. *Suppose that assumption (H) is satisfied and define $\mathfrak{k}_\infty := \lambda_{\min}(C_\infty) > 0$. Then for any $\epsilon_0 \in (0,1)$ it holds that*

$$\mathbb{P}\Big(\inf_{V \in \mathcal{C}(s,c_0)} \frac{\|VX\|_{L^2}^2}{\|V\|_2^2} \geq \frac{\mathfrak{k}_\infty}{2} \Big) \geq 1 - \epsilon_0,$$

for all

$$T \geq T_0(\epsilon_0, s, c_0) := \mathfrak{T}_0(\epsilon_0, s, c_0)\Big((4s+1)\log d - 2s\big(\log \frac{2s}{21} - 1 \big) + \log \frac{2}{\epsilon_0} \Big),$$

where the constant $\mathfrak{T}_0(\epsilon_0, s, c_0)$ is defined as

$$\mathfrak{T}_0(\epsilon_0, s, c_0) = \frac{144\mathfrak{p}_0\mathfrak{K}_\infty(c_0+2)^2(\mathfrak{k}_\infty + 18(c_0+2)^2\mathfrak{K}_\infty)}{\mathfrak{r}_0\mathfrak{k}_\infty^2}.$$

Proof. See Section 6.1. □

The next corollary presents a deviation bound for the quantity $\widehat{C}_T$.

Corollary 3.4. *For any $\epsilon_0 > 0$ and $T \geq T_0(\epsilon_0, s, c_0)$ it holds that*

$$\mathbb{P}\left(\inf_{V \in \mathcal{C}(s,c_0)} \frac{\|VX\|_{L^2}^2}{\|V\|_2^2} \geq \frac{\mathfrak{k}_\infty}{2},\ \|\text{diag}\ \widehat{C}_T\|_\infty \leq \mathfrak{m}_\infty + \frac{\mathfrak{k}_\infty}{2},\ \|\widehat{C}_T\|_\infty \leq \mathfrak{M}_\infty + \frac{3\mathfrak{k}_\infty}{2} \right) \geq 1 - \epsilon_0,$$

where $\mathfrak{m}_\infty := \|\text{diag}\ C_\infty\|_\infty$ and $\mathfrak{M}_\infty := \|C_\infty\|_\infty$.

Proof. See Section 6.2. □

3.2. Deviation bounds for the martingale term ε_T and final estimates

As mentioned earlier controlling the stochastic term $\langle \varepsilon_T, V \rangle_{\text{F}}$ for matrices $V \in \mathbb{R}^{d \times d}$ is crucial for the analysis of the estimators $\widehat{A}_{\text{L}}$ and $\widehat{A}_{\text{D}}$. The martingale property of ε_T turns out to be the key in the next proposition. We remark that the following result is an improvement of [11, Theorem 8].

Proposition 3.5. *For any $\epsilon_0 \in (0,1)$ the following inequality holds:*

$$\mathbb{P}\left(\sup_{V \in \mathbb{R}^{d\times d}, V \neq 0} \frac{\langle \varepsilon_T, V \rangle_{\text{F}}}{\|V\|_1} \geq \mu \right) \leq \epsilon_0$$

for any

$$T \geq \frac{48\mathfrak{p}_0\mathfrak{K}_\infty}{\mathfrak{r}_0} \frac{\mathfrak{k}_\infty + 6\mathfrak{K}_\infty}{\mathfrak{k}_\infty^2} \big((2s+1)\ln d - s(\ln s - 1) + \ln(4/\epsilon_0) \big)$$

and

$$\mu \geq \sqrt{\big(2\mathfrak{m}_\infty + \mathfrak{k}_\infty\big) \frac{\ln(2d^2/\epsilon_0)}{T}}.$$

Proof. We first recall Bernstein's inequality for continuous local martingales. Let $(M_t)_{t\geq 0}$ be a real-valued continuous local martingale with quadratic variation $(\langle M\rangle_t)_{t\geq 0}$. Then for any $a, b > 0$ it holds that

$$\mathbb{P}(M_t \geq a, \langle M\rangle_t \leq b) \leq \exp(-a^2/(2b)). \tag{3.2}$$

This result is a straightforward consequence of exponential martingale technique (cf. Chapter 4, Exercise 3.16 in [25]).

By definition $\varepsilon_T^{ij} = \frac{1}{T}\int_0^T dW_t^i X_t^j$ is a continuous martingale with $\langle \varepsilon^{ij}\rangle_T = \frac{1}{T}\widehat{C}_T^{ii}$. Therefore, we obtain by Corollary 3.4 and (3.2)

$$\begin{aligned}
\mathbb{P}\left(\sup_{V\in\mathbb{R}^{d\times d}, V\neq 0} \frac{\langle \varepsilon_T, V\rangle_{\mathrm{F}}}{\|V\|_1} \geq \mu\right) &\leq \mathbb{P}\Big(\|\mathrm{diag}\ \widehat{C}_T\|_\infty > \mathfrak{m}_\infty + \frac{\mathfrak{k}_\infty}{2}\Big) \\
+\mathbb{P}\left(\sup_{V\in\mathbb{R}^{d\times d}, V\neq 0} \frac{\langle \varepsilon_T, V\rangle_{\mathrm{F}}}{\|V\|_1} \geq \mu,\ \|\mathrm{diag}\ \widehat{C}_T\|_\infty \leq \mathfrak{m}_\infty + \frac{\mathfrak{k}_\infty}{2}\right) \\
\leq \sum_{i,j=1}^d \mathbb{P}\Big(\varepsilon_T^{ij} \geq \mu,\ \langle \varepsilon^{ij}\rangle_T \leq \frac{1}{T}\big(\mathfrak{m}_\infty + \frac{\mathfrak{k}_\infty}{2}\big)\Big) + \frac{\epsilon_0}{2} \\
\leq d^2 \exp\Big(-T\frac{\mu^2}{2\mathfrak{m}_\infty + \mathfrak{k}_\infty}\Big) + \frac{\epsilon_0}{2} \leq \epsilon_0,
\end{aligned}$$

which completes the proof. □

Summarising all previous deviation bounds we obtain the following result.

Corollary 3.6. *For $s \geq s_0$ and $c_0 > 0$ define the event*

$$\mathcal{E}(s, c_0) := \Big\{\inf_{V\in\mathcal{C}(s,c_0)} \frac{\|VX\|_{L^2}^2}{\|V\|_2^2} \geq \frac{\mathfrak{k}_\infty}{2}\Big\} \bigcap \Big\{\sup_{V\neq 0} \frac{\langle \varepsilon_T, V\rangle_F}{\|V\|_1} \leq \frac{\lambda}{2}\Big\}$$
$$\bigcap \Big\{\|\varepsilon_T\|_\infty \leq \frac{\lambda}{2}\Big\} \bigcap \Big\{\|\widehat{C}_T\|_\infty \leq \mathfrak{M}_\infty + \frac{3\mathfrak{k}_\infty}{2}\Big\}.$$

Then, for any $\epsilon_0 \in (0,1)$, it holds that $\mathbb{P}(\mathcal{E}(s, c_0)) \geq 1 - \epsilon_0$ for any $T \geq T_0\big(\epsilon_0/2, s, c_0\big)$ and

$$\lambda \geq 2\sqrt{\big(2\mathfrak{m}_\infty + \mathfrak{k}_\infty\big)\frac{\ln(2d^2/\epsilon_0)}{T}}.$$

4. Oracle inequalities and error bounds for the Lasso and Dantzig estimators

In this section we present the main theoretical results for the Lasso and Dantzig estimators. More specifically, we derive oracle inequalities for $\widehat{A}_{\mathrm{L}}$ and $\widehat{A}_{\mathrm{D}}$, and show the error bounds for the norms $\|\cdot\|_{L^2}$, $\|\cdot\|_1$ and $\|\cdot\|_2$. In particular, we establish the asymptotic equivalence between the Lasso and Dantzig estimators.

4.1. Properties of the Lasso estimator

We start this subsection with proving a statement, which is important for obtaining oracle inequality for the Lasso estimator $\widehat{A}_{\mathrm{L}}$.

Lemma 4.1. *Suppose that condition (2.6) holds. For any matrix $A \in \mathbb{R}^{d\times d}\backslash\{0\}$ denote $\mathcal{A} := \operatorname{supp}(A)$. Then for any $s \geq s_0$ and $c_0 > 0$ on $\mathcal{E}(s, c_0)$ the following inequality holds:*

$$\|(\widehat{A}_L - A_0)X\|_{L^2}^2 + \lambda\|\widehat{A}_L - A\|_1 \leq \|(A - A_0)X\|_{L^2}^2 + 4\lambda\|\widehat{A}_{L|\mathcal{A}} - A\|_1. \quad (4.1)$$

In particular, it implies that $\widehat{A}_L - A_0 \in \mathcal{C}(s_0, 3)$ on $\mathcal{E}(s, c_0)$.

Proof. Let us set $\delta_L(A) := A - \widehat{A}_{\mathrm{L}}$. Applying Lemma 2.1 we obtain the following inequality

$$\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 + \lambda\|\delta_L(A)\|_1$$
$$\leq \|(A - A_0)X\|_{L^2}^2 + 2\langle\varepsilon_T, \delta_L(A)\rangle_{\mathrm{F}} + \lambda\|\delta_L(A)\|_1 + 2\lambda\big(\|A\|_1 - \|\widehat{A}_{\mathrm{L}}\|_1\big).$$

Hence, on $\mathcal{E}(s, c_0)$ it holds that

$$\begin{aligned}&\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 + \lambda\|\delta_L(A)\|_1\\ &\quad\leq \|(A - A_0)X\|_{L^2}^2 + 2\lambda\big(\|\delta_L(A)\|_1 + \|A\|_1 - \|\widehat{A}_{\mathrm{L}}\|_1\big).\end{aligned}$$

We observe next that $\|\delta_L(A)\|_1 + \|A\|_1 - \|\widehat{A}_{\mathrm{L}}\|_1 \leq 2\|\delta_L(A)_{|\mathcal{A}}\|_1$, which immediately implies (4.1). Applying (4.1) to $A = A_0$ we deduce that

$$\|\delta_L(A_0)\|_1 \leq 4\|\delta_L(A_0)_{|\mathcal{A}}\|_1 \leq 4\|\delta_L(A_0)_{|\mathcal{I}_{s_0}(\delta_L(A_0))}\|_1,$$

where the last inequality holds due to the sparsity assumption $\|A_0\|_0 \leq s_0$. Consequently, $\widehat{A}_{\mathrm{L}} - A_0 \in \mathcal{C}(s_0, 3)$ and the proof is complete. □

We are now in the position to present an oracle inequality for the Lasso estimator $\widehat{A}_{\mathrm{L}}$, which is one of the main results of our paper.

Theorem 4.2. *Fix $\gamma > 0$ and $\epsilon_0 \in (0, 1)$. Consider the Lasso estimator $\widehat{A}_L$ defined at (2.7) and assume that condition (H) holds. Then for*

$$\lambda \geq 2\sqrt{\big(2\mathfrak{m}_\infty + \mathfrak{k}_\infty\big)\frac{\ln(2d^2/\epsilon_0)}{T}}$$

and $T \geq T_0\big(\epsilon_0/2, s_0, 3 + 4/\gamma\big)$, with probability at least $1 - \epsilon_0$ it holds that

$$\|(\widehat{A}_L - A_0)X\|_{L^2}^2 \leq (1+\gamma) \inf_{A:\ \|A\|_0\leq s_0}\Big\{\|(A - A_0)X\|_{L^2}^2 + \frac{9(2+\gamma)^2}{2\mathfrak{k}_\infty\gamma(1+\gamma)}\|A\|_0\lambda^2\Big\}.$$

Proof. Consider an arbitrary matrix $A \in \mathbb{R}^{d\times d}$ with $\|A\|_0 \le s_0$. Then, on $\mathcal{E}(s_0, 3+4/\gamma)$, according to Lemma 4.1 and Cauchy-Schwarz inequality:

$$\begin{aligned}&\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 + \lambda\|\widehat{A}_{\mathrm{L}} - A\|_1 \\ &\le \|(A - A_0)X\|_{L^2}^2 + 4\lambda\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_1 \\ &\le \|(A - A_0)X\|_{L^2}^2 + 4\lambda\sqrt{\|A\|_0}\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_2. \end{aligned} \tag{4.2}$$

Now, if $4\lambda\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_1 \le \gamma\|(A - A_0)X\|_{L^2}^2$ the result immediately follows from Lemma 4.1. Hence, we only need to treat the case $4\lambda\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_1 > \gamma\|(A - A_0)X\|_{L^2}^2$. The latter implies that $\widehat{A}_{\mathrm{L}} - A_0 \in \mathcal{C}(s_0, 3+4/\gamma)$ due to (4.2). Then, on the event $\mathcal{E}(s_0, 3+4/\gamma)$, we have

$$\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_2^2 \le \|\widehat{A}_{\mathrm{L}} - A\|_2^2 \le \frac{2}{\mathfrak{k}_\infty}\|(\widehat{A}_{\mathrm{L}} - A)X\|_{L^2}^2$$

and consequently we obtain from (4.2) that

$$\begin{aligned}\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 &\le \|(A - A_0)X\|_{L^2}^2 + 3\lambda\sqrt{\frac{2\|A\|_0}{\mathfrak{k}_\infty}}\|(\widehat{A}_{\mathrm{L}} - A)X\|_{L^2}^2 \\ &\le \|(A - A_0)X\|_{L^2}^2 + 3\lambda\sqrt{\frac{2\|A\|_0}{\mathfrak{k}_\infty}}\big(\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 + \|(A - A_0)X\|_{L^2}^2\big).\end{aligned}$$

Using the inequality $2xy \le ax^2 + y^2/a$ for $a > 0$, we then conclude that

$$\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 \le (1+\gamma)\|(A - A_0)X\|_{L^2}^2 + \frac{9(2+\gamma)^2}{2\mathfrak{k}_\infty\gamma(1+\gamma)}\|A\|_0\lambda^2,$$

which completes the proof. □

Theorem 4.2 enables us to find upper bounds on the various norms of $\widehat{A}_{\mathrm{L}} - A_0$ as well as on the sparsity of $\widehat{A}_{\mathrm{L}}$. We remark that the bound in (4.6) will be useful to provide the connection between the Lasso and Dantzig estimators in the next subsection.

Corollary 4.3. *Fix $\epsilon_0 \in (0,1)$. Consider the Lasso estimator $\widehat{A}_L$ defined in (2.7) and assume that conditions (2.6) and (H) hold. Then for*

$$\lambda \ge 2\sqrt{\big(2\mathfrak{m}_\infty + \mathfrak{k}_\infty\big)\frac{\ln(2d^2/\epsilon_0)}{T}}$$

and $T \ge T_0\big(\epsilon_0/2, s_0, 3\big)$, with probability at least $1-\epsilon_0$, it holds that

$$\|(\widehat{A}_L - A_0)X\|_{L^2}^2 \le \frac{18}{\mathfrak{k}_\infty}s_0\lambda^2 \tag{4.3}$$

$$\|\widehat{A}_L - A_0\|_2^2 \le \frac{36}{\mathfrak{k}_\infty^2}s_0\lambda^2 \tag{4.4}$$

$$\|\widehat{A}_L - A_0\|_1 \leq \frac{24}{\mathfrak{k}_\infty} s_0 \lambda \tag{4.5}$$

$$\|\widehat{A}_L\|_0 \leq \Big(48\frac{\mathfrak{M}_\infty}{\mathfrak{k}_\infty} + 72\Big) s_0. \tag{4.6}$$

Proof. On the event $\mathcal{E}(s_0, 3)$, taking $A = A_0$ and $\mathcal{A}_0 = \operatorname{supp}(A_0)$, we obtain the inequality

$$\|(\widehat{A}_\mathrm{L} - A_0)X\|_{L^2}^2 + \lambda\|\widehat{A}_\mathrm{L} - A\|_1 \leq 4\lambda\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_1$$

due to Lemma 4.1. Since on $\mathcal{E}(s_0, 3)$ we have $\widehat{A}_\mathrm{L} - A_0 \in \mathcal{C}(s_0, 3)$, we conclude that

$$\|(\widehat{A}_\mathrm{L} - A_0)X\|_{L^2}^2 \leq 3\lambda\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_1$$
$$\leq 3\lambda\sqrt{s_0}\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_2 \leq 3\lambda\sqrt{\frac{2s_0}{\mathfrak{k}_\infty}}\|(\widehat{A}_\mathrm{L} - A_0)X\|_{L^2}^2.$$

This gives (4.3) and (4.4). Moreover, on the same event it holds

$$\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_1 \leq 4\sqrt{s_0}\|\widehat{A}_{\mathrm{L}|\mathcal{A}} - A\|_2$$

and hence (4.5) follows. Now, it remains to prove (4.6). Note that necessary and sufficient condition for $\widehat{A}_\mathrm{L}$ to be the solution of the optimisation problem (2.7) is the existence of a matrix $B \in \partial\|\widehat{A}_\mathrm{L}\|_1$ such that

$$\varepsilon_T + \big(\widehat{A}_\mathrm{L} - A_0\big)\widehat{C}_T + \lambda B = 0.$$

Furthermore, $\widehat{A}_\mathrm{L}^{ij} \neq 0$ implies that $B^{ij} = \operatorname{sign}(\widehat{A}_\mathrm{L}^{ij})$. Thus, we conclude that

$$\|(\widehat{A}_\mathrm{L} - A_0)\widehat{C}_T\|_1 = \|\lambda B + \varepsilon_T\|_1 = \sum_{i,j=1}^{d}\Big|\lambda B^{ij} + \varepsilon_T^{ij}\Big|$$
$$\geq \sum_{i,j:\widehat{A}_\mathrm{L}^{ij}\neq 0}\Big|\lambda B^{ij} + \varepsilon_T^{ij}\Big| \geq \sum_{i,j:\widehat{A}_\mathrm{L}^{ij}\neq 0}\Big|\lambda - |\varepsilon_T^{ij}|\Big| \geq \|\widehat{A}_\mathrm{L}\|_0\frac{\lambda}{2},$$

where the last inequality holds on $\mathcal{E}(s_0, 3)$. On the other hand, on the same event we obtain

$$\|(\widehat{A}_\mathrm{L} - A_0)\widehat{C}_T\|_1 \leq \|\widehat{C}_T\|_\infty\|\widehat{A}_\mathrm{L} - A_0\|_1 \leq \left(\mathfrak{M}_\infty + \frac{3\mathfrak{k}_\infty}{2}\right)\frac{24}{\mathfrak{k}_\infty}s_0\lambda,$$

which implies (4.6). □

The upper bounds in (4.3)-(4.5) improve the bounds obtained in [11, Corollary 1] and they are in line with the classical results for linear regression models. We recall that the paper [11] considers row sparsity of the unknown parameter A_0, i.e.

$$\|A_0^i\|_0 \leq \mathfrak{s} \qquad \text{for all } 1 \leq i \leq d,$$

where A_0^i denotes the ith row of A_0. Obviously, this constraint corresponds to $s_0 = d\mathfrak{s}$ in our setting. The authors of [11] obtained the upper bound for $\|\widehat{A}_{\mathrm{L}} - A_0\|_2^2$ of order

$$\frac{d\mathfrak{s}(\log d + \log\log T)}{T}$$

in contrast to our improved bound $T^{-1} d\mathfrak{s} \log d$. Thus, we essentially match the lower bound

$$\inf_{\widehat{A}} \sup_{A:\ \max_i \|A_0^i\|_0 \leq \mathfrak{s}} \mathbb{E}[\|\widehat{A} - A\|_2^2] \geq \frac{c_1 d\mathfrak{s} \log(c_2 d/\mathfrak{s})}{T} \qquad \text{for some } c_1, c_2 > 0,$$

which has been derived in [11, Theorem 2].

Remark 4.4. Unfortunately, an extension of the analysis to more general diffusion models does not seem to be straightforward. The linearity of the drift function in the parameter A is absolutely crucial for the proofs. First of all, it allows for an explicit computation of the log-likelihood function, whereas in the more general setting we would require an approximative analysis of the likelihood, which is expected to be much more involved (see e.g. [26, 27] for an example of such analysis for general parametric models). Secondly, the linear form of the drift function leads to the quadratic form of the term $\widehat{C}_T$. We remark however that the methodology of [22], which is the basis of Proposition 3.2, only applies to quadratic functionals of X and thus different mathematical techniques are needed to show this type of concentration phenomena in the general framework. Hence, we leave this investigation for future research. □

Remark 4.5. In this section we showed that the Lasso estimator has asymptotically optimal estimation rate. Another preferable property for an estimator in sparse model is consistency in selection of variables. For Ornstein-Uhlenbeck model we say that estimator $\hat{A}$ is consistent for selection of variables if

$$\mathbb{P}\big(\mathrm{supp}(\hat{A}) = \mathrm{supp}(A_0)\big) \to 1, \quad \text{when} \quad T \to \infty.$$

These two properties are referred to as oracle properties (see [8]), and it is well known that the Lasso estimator for linear Gaussian models cannot satisfy both of them with the same tuning parameter λ (see [30]), while the adaptive Lasso estimator can. The authors of [11] have introduced the adaptive Lasso estimator for the Ornstein-Uhlenbeck model, which is defined as

$$\widehat{A}_{\mathrm{ad}} := \underset{A \in \mathbb{R}^{d \times d}}{\mathrm{argmin}} \left(\mathcal{L}_T(A) + \lambda \|A \circ |\widehat{A}_{\mathrm{ML}}|^{-\gamma}\|_1 \right),$$

where $\circ$ denotes the Hadamard product and $(|\widehat{A}_{\mathrm{ML}}|^{-\gamma})_{ij} := |\widehat{A}_{\mathrm{ML}}^{ij}|^{-\gamma}$ for a $\gamma > 0$. They have proved that the adaptive estimator $\widehat{A}_{\mathrm{ad}}$ is consistent for support selection and showed the asymptotic normality of $\widehat{A}_{\mathrm{ad}}$ when restricted to the elements in $\mathrm{supp}(A_0)$; see [11, Theorem 4] for more details. □

4.2. Properties of the Dantzig estimator

In this subsection we will establish a connection between the prediction errors associated with the Lasso and Dantzig estimators. This step is essential for the derivation of error bounds for $\widehat{A}_{\mathrm{D}}$. Our results are an extension of the study in [2], where it was shown that under sparsity conditions, the Lasso and the Dantizg estimators show similar behaviour for linear regression and for nonparametric regression models, for l_2 prediction loss and for l_p loss in the coefficients for $1 \leq p \leq 2$.

In what follows, we will derive analogous bounds for the Ornstein-Uhlenbeck process.

Proposition 4.6. *Consider the Dantzig estimator $\widehat{A}_D$ defined in (2.10) and assume that condition (H) holds.*

(i) Define $\delta_D(A) := A - \widehat{A}_D$ and $\mathcal{A} := supp(A)$, and assume that A satisfies the Dantzig constraint (2.9). Then it holds that

$$\|\delta_D(A)_{|\mathcal{A}^c}\|_1 \leq \|\delta_D(A)_{|\mathcal{A}}\|_1.$$

(ii) On the event $\big\{\|\widehat{A}_L\|_0 \leq s\big\} \cap \mathcal{E}(s,1)$ the following inequality holds:

$$\Big| \|(\widehat{A}_L - A_0)X\|^2_{L^2} - \|(\widehat{A}_D - A_0)X\|^2_{L^2} \Big| \leq \frac{18}{\mathfrak{k}_\infty} \|\widehat{A}_L\|_0 \lambda^2.$$

Proof. See Section 6.3. □

Proposition 4.6 implies an oracle inequality for the Dantzig estimator, which is formulated in the next theorem.

Theorem 4.7. *Fix $\gamma > 0$ and $\epsilon_0 \in (0,1)$. Consider the Dantzig estimator $\widehat{A}_D$ defined in (2.10) and assume that conditions (2.6) and (H) hold. Then for*

$$\lambda \geq 2\sqrt{\left(2\mathfrak{m}_\infty + \mathfrak{k}_\infty\right)\frac{\ln\left(2d^2/\epsilon_0\right)}{T}}$$

and $T \geq T_0\big(\epsilon_0/2, (48\frac{\mathfrak{M}_\infty}{\mathfrak{k}_\infty} + 72)s_0, 3 + 4/\gamma\big)$, with probability at least $1 - \epsilon_0$, it holds that

$$\|(\widehat{A}_D - A_0)X\|^2_{L^2} \leq (1+\gamma) \inf_{A:\|A\|_0 = s_0} \left\{ \|(A - A_0)X\|^2_{L^2} + C_D(\gamma)s_0\lambda^2 \right\}, \quad (4.7)$$

where

$$C_D(\gamma) = \frac{18}{\mathfrak{k}_\infty}\Big(\frac{(\gamma+2)^2}{4\gamma} + 48\frac{\mathfrak{M}_\infty}{\mathfrak{k}_\infty} + 72\Big).$$

Proof. Consider matrix $A \in \mathbb{R}^{d\times d}$ such that $\|A\|_0 = s_0$. Then, on the event $\mathcal{E}(s_1, 3 + 4/\gamma)$, according to Proposition (4.6)

$$\|(\widehat{A}_{\mathrm{D}} - A_0)X\|^2_{L^2} \leq \|(\widehat{A}_{\mathrm{L}} - A_0)X\|^2_{L^2} + \frac{18}{\mathfrak{k}_\infty}\Big(48\frac{\mathfrak{M}_\infty}{\mathfrak{k}_\infty} + 72\Big)s_0\lambda^2.$$

On the other hand, due to Theorem 4.2, we deduce that

$$\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 \le (1+\gamma)\|(A - A_0)X\|_{L^2}^2 + \frac{9(\gamma+2)^2}{2\mathfrak{k}_\infty \gamma} s_0 \lambda^2.$$

Combining both inequalities yields (4.7). □

The statements of Theorems 4.2 and 4.7 suggest that the Lasso and Dantzig estimators are asymptotically equivalent. This is in line with the theoretical findings in linear regression models as it has been shown in [2]. More specifically, we obtain the following result, which is a direct analogue of Corollary 4.3.

Corollary 4.8. *Fix $\epsilon_0 \in (0,1)$. Consider the Dantzig estimator $\widehat{A}_D$ defined in (2.10) and assume that conditions (2.6) and (H) hold. Then for*

$$\lambda \ge 2\sqrt{\left(2\mathfrak{m}_\infty + \mathfrak{k}_\infty\right)\frac{\ln\left(2d^2/\epsilon_0\right)}{T}}$$

and $T \ge T_0\left(\epsilon_0/2, s_0, 1\right)$, with probability at least $1-\epsilon_0$, it holds that

$$\|(\widehat{A}_D - A_0)X\|_{L^2}^2 \le \frac{18}{\mathfrak{k}_\infty} s_0 \lambda^2 \tag{4.8}$$

$$\|\widehat{A}_D - A_0\|_2^2 \le \frac{36}{\mathfrak{k}_\infty^2} s_0 \lambda^2 \tag{4.9}$$

$$\|\widehat{A}_D - A_0\|_1 \le \frac{24}{\mathfrak{k}_\infty} s_0 \lambda.$$

Proof. Denote $\mathcal{A}_0 = \operatorname{supp}(A_0)$. On the event $\mathcal{E}(s_0,1)$ the matrix A_0 satisfies the Dantzig constraint (2.9), $\widehat{A}_{\mathrm{D}} - A_0 \in \mathcal{C}(s_0,1)$ and

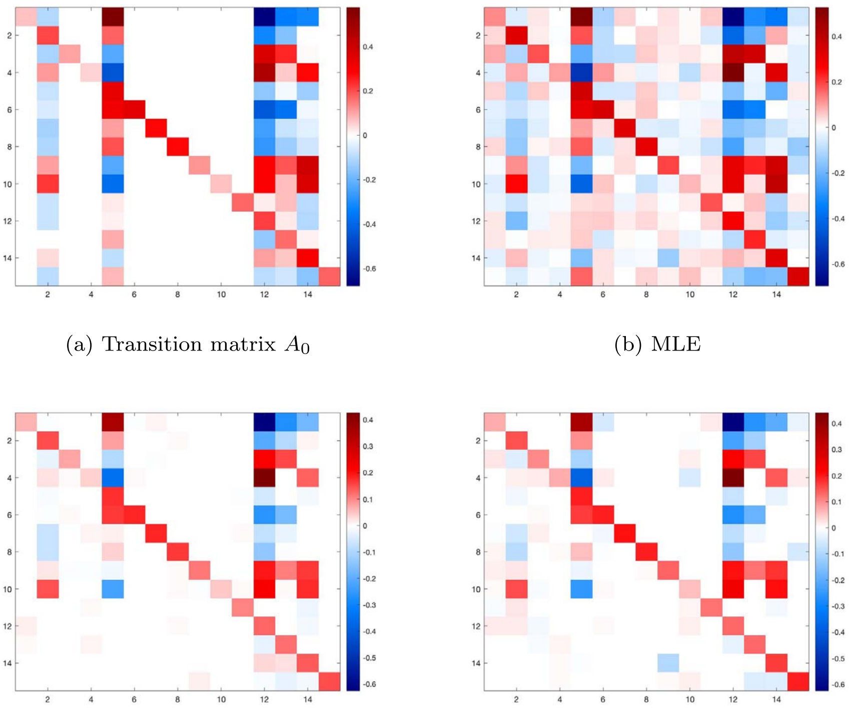
$$\begin{aligned}
&\|(\widehat{A}_{\mathrm{D}} - A_0)X\|_{L^2}^2 \le \|(\widehat{A}_{\mathrm{D}} - A_0)\widehat{C}_T\|_\infty \|\widehat{A}_{\mathrm{D}} - A_0\|_1 \\
&\le 2\big(\|(\widehat{A}_{\mathrm{D}} - A_0)\widehat{C}_T + \varepsilon_T\|_\infty + \|\varepsilon_T\|_\infty\big)\|\widehat{A}_{\mathrm{D}|\mathcal{A}_0} - A_0\|_1 \\
&\le 3\lambda\sqrt{s_0}\|\widehat{A}_{\mathrm{D}} - A_0\|_2 \le 3\lambda\sqrt{\frac{2s_0}{\mathfrak{k}_\infty}\|(\widehat{A}_{\mathrm{D}} - A_0)X\|_{L^2}^2},
\end{aligned}$$

which gives (4.8) and (4.9). Moreover, on the same event it holds that

$$\|\widehat{A}_{\mathrm{D}} - A_0\|_1 \le 2\sqrt{s_0}\|\widehat{A}_{\mathrm{D}} - A_0\|_2,$$

which completes the proof. □

In this work we have shown that the Lasso and Dantzig selector performances are equivalent. It is worth to mention that although we study penalized likelihood methods, it may be of separate interest (in both computational and theoretical context) that the Dantzig estimator can also be applied to settings in which no explicit likelihoods or loss functions are available (see [7] for more details).



(a) Transition matrix A_0

(b) MLE

(c) Lasso

(d) Dantzig

Fig 1. *Comparison of the true matrix with maximum likelihood, Lasso and Dantzig estimators.*

5. Numerical simulations

This sections presents some numerical experiments on simulated data that illustrate our theoretical results.

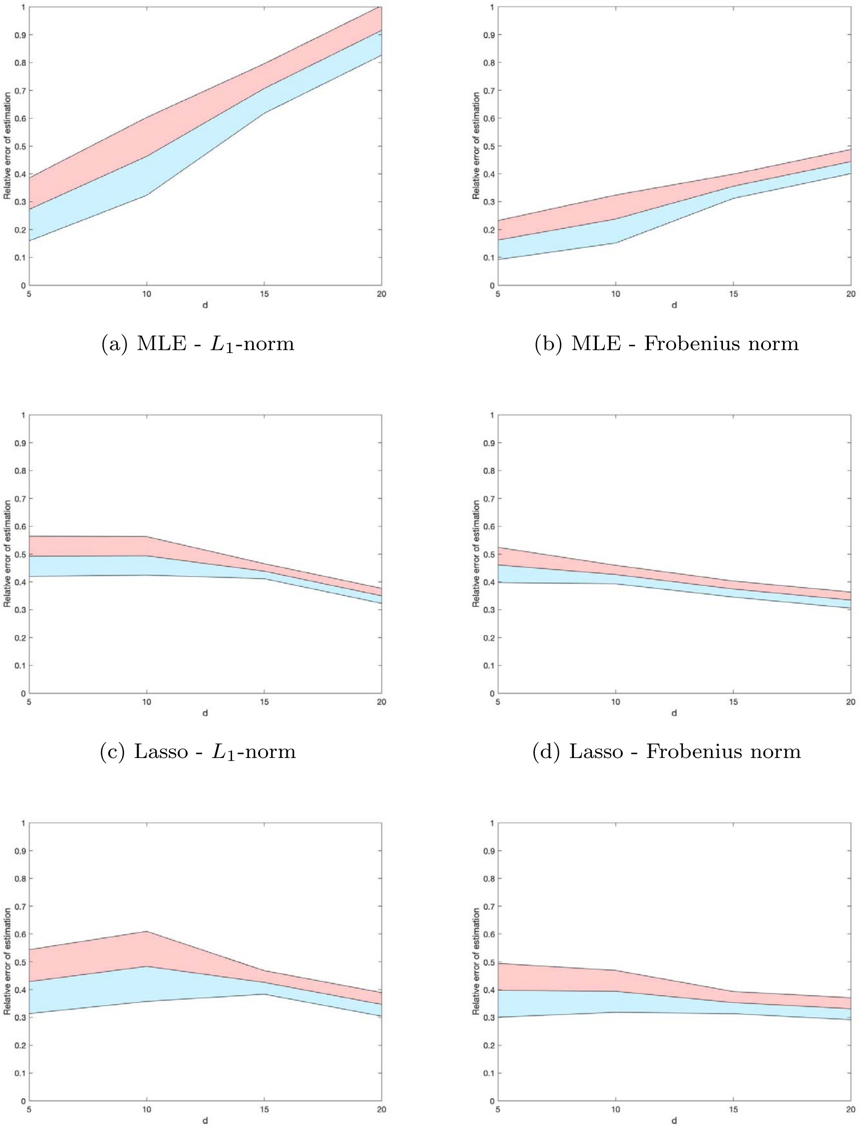
Our estimation methods are based on continuous observations of the the underlying process, which need to be discretised for numerical simulations. We will use 500000 discretisation points over the time interval $[0, T]$ with $T = 300$. Such approximation is sufficient for the illustration purpose, since further refinement of the grid does not lead to a significant improvement.

The selection of value of tuning parameter λ is made by cross-validation technique, using first 90% of observations as training set and last 10% as validation set. More precisely, in our simulations we utilise

$$\lambda_0 = \operatorname*{argmin}_{\lambda>0} \frac{\mathcal{L}_{[0.9T,T]}(\widehat{A}_L(\lambda))}{||\widehat{A}_L(\lambda)||_1},$$

where

$$\widehat{A}_L(\lambda) = \operatorname*{argmin}_{A\in\mathbb{R}^{d\times d}} (\mathcal{L}_{0.9T}(A) + \lambda\|A\|_1)$$



(a) MLE - L_1-norm

(b) MLE - Frobenius norm

(c) Lasso - L_1-norm

(d) Lasso - Frobenius norm

(e) Dantzig - L_1-norm

(f) Dantzig - Frobenius norm

Fig 2. *Relative error of maximum likelihood, Lasso and Dantzig estimators in L^1 and Frobenius norms depending on d. Middle line corresponds to the mean and coloured areas correspond to the standard deviation of the error over 10 independent simulations.*

In Figure 1 we demonstrate an example of the transition matrix $A_0 \in \mathbb{R}^{15\times 15}$ and the corresponding maximum likelihood, Lasso and Dantzig estimators. Instead of giving numerical values of the entries of A_0 we use a colour code to highlight the sparsity. We observe that MLE provides a good performance on the support, but it gives rather poor estimates outside the support. On the other hand, the superiority of the Lasso and Dantzig estimators, especially in terms of support recovery, is quite obvious even for relatively small dimension of matrix.

Figure 2 demonstrates the relative error of the maximum likelihood, Lasso and Dantzig estimators compared to the norm of the true matrix. We compute the relative error for dimensions $d = 5, \dots, 20$ and for L^1 and Frobenius norms. Figure 2 clearly shows the improvement of performance of penalized estimation methods with growth of the dimension d compared to the maximum likelihood estimation. Indeed, we observe that relative errors of maximum likelihood estimation grow linearly both in L^1 and Frobenius norms, while relative errors of Lasso and Dantzig estimators decay in d. The sparsity of the true parameter A_0 was chosen equal to $s = 0.3d^2$, which might explain the limiting behaviour of Lasso and Dantzig estimators when d is increasing. Finally, we observe that relative errors for Lasso and Dantzig estimators are practically equivalent, which is exactly in accordance with our theoretical results.

6. Proofs

6.1. Proof of Theorem 3.3

We first note the identity $\|VX\|_{L^2}^2 = \operatorname{tr}(V\widehat{C}_T V^\top)$. Replacing $\widehat{C}_T$ by its limit C_∞ we deduce the inequality $\operatorname{tr}(VC_\infty V^\top) \geq \mathfrak{k}_\infty > 0$ and therefore

$$\frac{\|VX\|_{L^2}^2}{\|V\|_2^2} = \frac{\operatorname{tr}(VC_\infty V^\top)}{\|V\|_2^2} - \frac{\operatorname{tr}(V(C_\infty - \widehat{C}_T)V^\top)}{\|V\|_2^2} \geq \mathfrak{k}_\infty - \frac{|\operatorname{tr}(V(C_\infty - \widehat{C}_T)V^\top)|}{\|V\|_2^2}. \tag{6.1}$$

Next, we introduce the set $\mathcal{K}(s) := \big\{V \in \mathbb{R}^{d\times d} \setminus \{0\} : \ \|V\|_0 \leq s\big\}$. As is shown in Lemma 6.1 it holds that

$$\sup_{V\in\mathcal{C}(s,c_0)} \frac{|\operatorname{tr}(V(C_\infty - \widehat{C}_T)V^\top)|}{\|V\|_2^2} \leq 3(c_0+2)^2 \sup_{V\in\mathcal{K}(2s)} \frac{|\operatorname{tr}(V(C_\infty - \widehat{C}_T)V^\top)|}{\|V\|_2^2}. \tag{6.2}$$

Thus, it suffices to consider $\mathcal{K}(s)$ instead of $\mathcal{C}(s,c_0)$ in the following discussion. Observing (6.1) we obtain that

$$\mathbb{P}\left(\inf_{V\in\mathcal{K}(s)} \frac{\|VX\|_{L^2}^2}{\|V\|_2^2} \geq \frac{\mathfrak{k}_\infty}{2}\right) \geq \mathbb{P}\left(\sup_{V\in\mathcal{K}(s)} \frac{|\operatorname{tr}(V(C_\infty - \widehat{C}_T)V^\top)|}{\|V\|_2^2} \leq \frac{\mathfrak{k}_\infty}{2}\right).$$

For a matrix $V \in \mathcal{K}(s)$ we denote its j-th row vector by v^j and $\mathbf{v} = \operatorname{vec}(V) \in \mathbb{R}^{d^2}$. Moreover, we define a symmetric random matrix $\mathcal{D}_C = \operatorname{id} \otimes (C_\infty - \widehat{C}_T) \in$

$\mathbb{R}^{d^2\times d^2}$. Then we deduce the identity

$$\frac{\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top)}{\|V\|_2^2} = \frac{\mathbf{v}^\top \mathcal{D}_C \mathbf{v}}{\|\mathbf{v}\|_2^2}. \tag{6.3}$$

According to Proposition 3.2 we obtain the following inequalities for any $x > 0$:

$$\mathbb{P}\left(\frac{|\mathbf{v}^\top \mathcal{D}_C \mathbf{v}|}{\|\mathbf{v}\|_2^2} \geq x\right) \leq \mathbb{P}\left(\frac{\sum_{j=1}^d |v^j(C_\infty - \widehat{C}_T)(v^j)^\top|}{\sum_{j=1}^d \|v^j\|_2^2} \geq x\right)$$

$$\leq \sum_{j=1}^d \mathbb{P}\left(\frac{|v^j(C_\infty - \widehat{C}_T)(v^j)^\top|}{\|v^j\|_2^2} \geq x\right) \leq 2d\exp\left(-TH_0(x)\right).$$

By Lemma 6.2 we conclude that

$$\mathbb{P}\left(\sup_{\mathbf{v}\in\mathbb{R}^{d^2}\setminus\{0\}:\ \|\mathbf{v}\|_0\leq s} \frac{|\mathbf{v}^\top \mathcal{D}_C \mathbf{v}|}{\|\mathbf{v}\|_2^2} \geq 3x\right) \leq 2d\Big(\frac{21ed^2}{s}\Big)^s \exp(-TH_0(x)).$$

We deduce from (6.3) that

$$\mathbb{P}\left(\inf_{V\in\mathcal{K}(s)} \frac{\|VX\|_{L^2}^2}{\|V\|_2^2} \geq 3x\right) \geq 1 - 2d\Big(\frac{21ed^2}{s}\Big)^s \exp\left(-TH_0(x)\right).$$

The latter statement together with (6.2) implies the inequality

$$\mathbb{P}\Big(\inf_{V\in\mathcal{C}(s,c_0)} \frac{\|VX\|_{L^2}^2}{\|V\|_2^2} \geq \frac{\mathfrak{k}_\infty}{2}\Big) \geq 1 - \epsilon_0,$$

for all $T \geq T_0(\epsilon_0, s, c_0)$, which completes the proof of Theorem 3.3.

6.2. Proof of Corollary 3.4

Let $e_{(i,j)} \in \mathbb{R}^{d\times d}$ be a matrix defined as $e_{(i,j)}^{kl} := 1_{(k,l)=(i,j)}$. We observe that

$$\left\{\|\mathrm{diag}(C_\infty - \widehat{C}_T)\|_\infty > \frac{\mathfrak{k}_\infty}{2}\right\} = \left\{\max_{1\leq j\leq d}\left|\mathrm{tr}(e_{(j,j)}(C_\infty - \widehat{C}_T)e_{(j,j)}^\top\right| > \frac{\mathfrak{k}_\infty}{2}\right\}$$

$$\subset \left\{\sup_{V\in\mathcal{C}(s,c_0)} \frac{\left|\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top\right|}{\|V\|_2^2} > \frac{\mathfrak{k}_\infty}{2}\right\}.$$

Furthermore,

$$|C_\infty^{ij} - \widehat{C}_T^{ij}| = \left|\mathrm{tr}(e_{(1,i)}(C_\infty - \widehat{C}_T)e_{(1,j)}^\top)\right|$$

$$\leq \frac{1}{2}\left|\mathrm{tr}((e_{(1,i)} + e_{(1,j)})(C_\infty - \widehat{C}_T)(e_{(1,i)} + e_{(1,j)})^\top)\right|$$

$$+\frac{1}{2}\left|\mathrm{tr}(e_{(1,i)}(C_\infty - \widehat{C}_T)e_{(1,i)}^\top)\right| + \frac{1}{2}\left|\mathrm{tr}(e_{(1,j)}(C_\infty - \widehat{C}_T)e_{(1,j)}^\top)\right|$$
$$\leq 3 \sup_{V\in\mathcal{C}(s,c_0)} \frac{\left|\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top\right|}{\|V\|_2^2}$$

and hence

$$\left\{\|C_\infty - \widehat{C}_T\|_\infty > \frac{3\mathfrak{k}_\infty}{2}\right\} = \left\{\max_{1\leq i,j\leq d}\left|\mathrm{tr}(e_{(1,i)}(C_\infty - \widehat{C}_T)e_{(1,j)}^\top)\right| > \frac{3\mathfrak{k}_\infty}{2}\right\}$$
$$\subset \left\{\sup_{V\in\mathcal{C}(s,c_0)} \frac{\left|\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top\right|}{\|V\|_2^2} > \frac{\mathfrak{k}_\infty}{2}\right\}.$$

This completes the proof of Corollary 3.4.

6.3. Proof of Proposition 4.6

Since A satisfies the Dantzig constraint (2.9), we deduce by definition of the Dantzig estimator:

$$\begin{aligned}\|A\|_1 \geq \|\widehat{A}_{\mathrm{D}}\|_1 &= \|A - \delta_D(A)_{|\mathcal{A}}\|_1 + \|\delta_D(A)_{|\mathcal{A}^c}\|_1 \\ &\geq \|A\|_1 - \|\delta_D(A)_{|\mathcal{A}}\|_1 + \|\delta_D(A)_{|\mathcal{A}^c}\|_1,\end{aligned}$$

which proves part (i).

Now we show part (ii) of the proposition. Set $\delta := \widehat{A}_{\mathrm{L}} - \widehat{A}_{\mathrm{D}}$. Due to (2.8) we deduce

$$\begin{aligned}&\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 - \|(\widehat{A}_{\mathrm{D}} - A_0)X\|_{L^2}^2 \\ &= 2\mathrm{tr}\left(\left(\widehat{A}_{\mathrm{D}}\widehat{C}_T + \varepsilon_T - A_0\widehat{C}_T\right)\delta^\top\right) - 2\mathrm{tr}\left(\varepsilon_T\delta^\top\right) + \mathrm{tr}\left(\delta\widehat{C}_T\delta^\top\right) \qquad (6.4)\\ &= 2\mathrm{tr}\left(\left(\widehat{A}_{\mathrm{L}}\widehat{C}_T + \varepsilon_T - A_0\widehat{C}_T\right)\delta^\top\right) - 2\mathrm{tr}\left(\varepsilon_T\delta^\top\right) - \mathrm{tr}\left(\delta\widehat{C}_T\delta^\top\right).\end{aligned}$$

The Dantzig constraint (2.9) implies the inequality

$$\left|\mathrm{tr}\left(\left(\widehat{A}_{\mathrm{D}}\widehat{C}_T + \varepsilon_T - A_0\widehat{C}_T\right)\delta^\top\right)\right| \leq \|\widehat{A}_{\mathrm{D}}\widehat{C}_T + \varepsilon_T - A_0\widehat{C}_T\|_\infty\|\delta\|_1 \leq \lambda\|\delta\|_1,$$

and the same inequality holds for $\widehat{A}_{\mathrm{D}}$ being replaced by $\widehat{A}_{\mathrm{L}}$. On $\mathcal{E}(s,1)$ we have

$$\left|\mathrm{tr}\left(\varepsilon_T\delta^\top\right)\right| \leq \frac{\lambda}{2}\|\delta\|_1.$$

Furthermore, on $\left\{\|\widehat{A}_{\mathrm{L}}\|_0 \leq s\right\}$ it holds that $\delta \in \mathcal{C}(s,1)$ and we conclude from Theorem (3.3) that

$$\mathrm{tr}\left(\delta\widehat{C}_T\delta^\top\right) \geq \frac{\mathfrak{k}_\infty}{2}\|\delta\|_2^2.$$

We also have $\|\delta\|_1 \leq 2\|\delta_{|\mathrm{supp}(\widehat{A}_{\mathrm{L}})}\|_1 \leq 2\|\widehat{A}_{\mathrm{L}}\|_0^{1/2}\|\delta\|_2^2$. Observing the first identity of (6.4), putting the previous estimates together and using the inequality $2xy \leq ax^2 + y^2/2$ for $a > 0$, we obtain the following inequality

$$\|(\widehat{A}_{\mathrm{D}} - A_0)X\|_{L^2}^2 - \|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 \leq \frac{18}{\mathfrak{k}_\infty}\|\widehat{A}_{\mathrm{L}}\|_0\lambda^2.$$

On the other hand, applying the second identity of (6.4), we deduce that

$$\|(\widehat{A}_{\mathrm{L}} - A_0)X\|_{L^2}^2 - \|(\widehat{A}_{\mathrm{D}} - A_0)X\|_{L^2}^2 \leq \frac{18}{\mathfrak{k}_\infty}\|\widehat{A}_{\mathrm{L}}\|_0\lambda^2,$$

which completes the proof.

6.4. Some lemmas

In this subsection we present two results that can be easily deduced from Lemmas F.1, F.2 and F.3 from supplementary material of [1]. We state their proofs for the sake of completeness.

Lemma 6.1. *It holds that*

$$\sup_{V\in\mathcal{C}(s,c_0)} \frac{|\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top)|}{\|V\|_2^2} \leq 3(c_0+2)^2 \sup_{V\in\mathcal{K}(2s)} \frac{|\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top)|}{\|V\|_2^2}.$$

Proof. First, recall the definition of the set $\mathcal{C}(s,c_0)$ in (2.1) and denote the unit balls by $\mathbb{B}_q(r) := \{v \in \mathbb{R}^d : \|v\|_q \leq r\}$ for any $d \geq 1$ and $q \geq 0$, $r > 0$. Furthermore, we introduce the notation $\mathcal{K}(s) = \mathbb{B}_0(s) \cap \mathbb{B}_2(1)$ for $s \geq 1$. For any set P we denote its closure and convex hull by $\mathrm{cl}(P)$ and $\mathrm{conv}(P)$, respectively. By a direct application of Lemma F.1 from [1], we obtain the following approximation of cone sets by sparse sets: for any $S \subset \{1,\ldots,d\}$ with $|S| = s$ we get

$$\mathcal{C}(s,c_0) \cap \mathbb{B}_2(1) \subseteq \mathbb{B}_1\left((c_0+1)\sqrt{s}\right) \cap \mathbb{B}_2(1) \subseteq (c_0+2)\mathrm{cl}(\mathrm{conv}(\mathcal{K}(s))). \tag{6.5}$$

Next, by the statement of Lemma F.3 in [1] we have that

$$\sup_{V\in\mathrm{cl}(\mathrm{conv}(\mathcal{K}(s)))} |\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top)| \leq 3 \sup_{V\in\mathcal{K}(2s)} |\mathrm{tr}(V(C_\infty - \widehat{C}_T)V^\top)|. \tag{6.6}$$

Thus, (6.5) combined with (6.6) yields the proof. □

Lemma 6.2. *Let* $\boldsymbol{v} = \mathrm{vec}(V) \in \mathbb{R}^{d^2}$ *and* $\mathcal{D}_C = \mathrm{id} \otimes (C_\infty - \widehat{C}_T) \in \mathbb{R}^{d^2\times d^2}$. *Then it holds that*

$$\mathbb{P}\left(\sup_{\boldsymbol{v}\in\mathbb{R}^{d^2}\setminus\{0\}:\ \|\boldsymbol{v}\|_0\leq s} \frac{|\boldsymbol{v}^\top\mathcal{D}_C\boldsymbol{v}|}{\|\boldsymbol{v}\|_2^2} \geq 3x\right) \leq 2d\left(\frac{21ed^2}{s}\right)^s \exp(-TH_0(x)),$$

where the function H_0 *has been introduced in Proposition 3.2.*

Proof. Choose $U \subset \{1, \dots, d^2\}$ with $|U| = s$, and define

$$S_U = \left\{ \mathbf{v} \in \mathbb{R}^{d^2} : \|\mathbf{v}\|_2 \le 1, \ \mathrm{supp}(\mathbf{v}) \subseteq U \right\}.$$

Then $\mathcal{K}(s) = \bigcup_{|U| \le s} S_U$. In what follows, we choose $\mathcal{A} = \{u_1, \dots, u_m\}$, which is a $\frac{1}{10}$-net of S_U. Lemma 3.5 of [29] guarantees that $|\mathcal{A}| \le 21^s$. Next, notice that for every $\mathbf{v} \in S_u$, there exists some $u_i \in \mathcal{A}$ such that $\|\Delta\mathbf{v}\| \le \frac{1}{10}$, where $\Delta\mathbf{v} = \mathbf{v} - u_i$. Then it holds

$$\begin{aligned}
\gamma &:= \sup_{\mathbf{v} \in S_U} |\mathbf{v}^\top \mathcal{D}_C \mathbf{v}| \\
&\le \max_i |u_i^\top \mathcal{D}_C u_i| + 2 \sup_{\mathbf{v} \in S_U} |\max_i u_i^\top \mathcal{D}_C (\Delta\mathbf{v})| + \sup_{v \in S_U} |(\Delta\mathbf{v})^\top \mathcal{D}_C (\Delta\mathbf{v})|.
\end{aligned}$$

Next, we use the fact that $10(\Delta\mathbf{v}) \in S_U$ which gives us in consequence

$$\sup_{\mathbf{v} \in S_U} |(\Delta\mathbf{v})^\top \mathcal{D}_C (\Delta\mathbf{v})| \le \frac{1}{100}\gamma$$

and

$$\begin{aligned}
&2 \sup_{\mathbf{v} \in S_U} |\max_i u_i^\top \mathcal{D}_C (\Delta\mathbf{v})| \\
&\le \frac{1}{10}\Big(\sup_{\mathbf{v} \in S_U} |(u_i + 10\Delta\mathbf{v})^\top \mathcal{D}_C (u_i + 10\Delta\mathbf{v})| \\
&\quad + \sup_{\mathbf{v} \in S_U} |u_i \mathcal{D}_C u_i| + \sup_{v \in S_U} |(10\Delta\mathbf{v})^\top \mathcal{D}_C (10\Delta\mathbf{v})| \Big) \\
&\le \frac{4}{10}\gamma + \frac{1}{10}\gamma + \frac{1}{10}\gamma
\end{aligned}$$

which implies that

$$\gamma \le 3 \max_i |u_i^\top \mathcal{D}_C u_i|.$$

Now, we take an union bound over all $u_i \in \mathcal{A}$ and combine it with inequality (3.1) from Proposition 3.2. Thus,

$$\mathbb{P}\left(\sup_{\mathbf{v} \in S_U} |\mathbf{v}^\top \mathcal{D}_C \mathbf{v}| \ge x \right) \le 2d \exp(-TH_0(x) + s \log 21).$$

Next, we take another union bound over $\binom{d^2}{s} \le \left(\frac{ed^2}{s}\right)^s$ choices of U. Thus,

$$\mathbb{P}\left(\sup_{\mathbf{v} \in \mathbb{R}^{d^2} \setminus \{0\} : \ \|\mathbf{v}\|_0 \le s} \frac{|\mathbf{v}^\top \mathcal{D}_C \mathbf{v}|}{\|\mathbf{v}\|_2^2} \ge 3x \right) \le 2d \Big(\frac{21ed^2}{s} \Big)^s \exp(-TH_0(x)),$$

which yields the proof. □

References

[1] Basu, S. and Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Annals of Statistics* **43** 1535-1567. MR3357870

[2] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705-1732. MR2533469

[3] Bolley, F., Cañizo, J. A. and Carrillo, J. A. (2011). Stochastic mean-field limit: non-Lipschitz forces and swarming. *Mathematical Models and Methods in Applied Sciences* **21** 2179-2210. MR2860672

[4] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics, Springer. MR2807761

[5] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313-2351. MR2382644

[6] Carmona, R. and Zhu, X. (2016). A probabilistic approach to mean field games with major and minor players. *Annals of Applied Probability* **26** 1535-1580. MR3513598

[7] Dicker, L., Li, Y. and Zhao, S. D. (2014). The Dantzig selector for censored linear regression models. *Statistica Sinica* **24** 251-275. MR3183683

[8] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348-1360. MR1946581

[9] Faugeras, O., Touboul, J. and Cessac, B. (2009). A constructive mean-field analysis of multi-population neural networks with random synaptic weights and stochastic inputs. *Frontiers in Computational Neuroscience* **3** 1-28.

[10] Fujimori, K. (2019). The Dantzig selector for a linear model of diffusion processes. *Statistical Inference for Stochastic Processes* **22** 475-498. MR3996026

[11] Gaïffas, S. and Matulewicz, G. (2019). Sparse inference of the drift of a high-dimensional Ornstein-Uhlenbeck process. *Journal of Multivariate Analysis* **169** 1-20. MR3875583

[12] Gregorio, A. D. and Iacus, S. (2012). Adaptive Lasso-type estimation for multivariate diffusion processes. *Econometric Theory* **28** 838-860. MR2959127

[13] Jackson, M. (2008). *Social and economic networks*. Princeton University Press, Princeton, NJ. MR2435744

[14] Jacod, J. and Protter, P. (2012). *Discretization of processes*. Stochastic Modelling and Applied Probability, Springer. MR2859096

[15] James, G., Radchenko, P. and Lv, J. (2009). Dasso: connections between the Dantzig selector and Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71** 127-142. MR2655526

[16] Kessler, M. and Rahbek, A. (2001). Asymptotic likelihood based inference for co-integrated homogenous Gaussian diffusions. *Scandinavian Jour-*

nal of Statistics **28** 455-470. MR1858411

[17] Küchler, U. and Sørensen, M. (1997). *Exponential families of stochastic processes.* Springer Series in Statistics, Springer. MR1458891

[18] Küchler, U. and Sørensen, M. (1999). A note on limit theorems for multivariate martingales. *Bernoulli* **5** 483-493. MR1693604

[19] Kutoyants, Y. A. (2004). *Statistical inference for ergodic diffusion processes.* Springer Series in Statistics, Springer. MR2144185

[20] McKean, H. P. (1966). Speed of approach to equilibrium for Kac's caricature of a Maxwellian gas. *Archive for Rational Mechanics and Analysis* **21** 343-367. MR0214112

[21] McKean, H. P. (1967). Propagation of chaos for a class of non-linear parabolic equations. *In* **7** 41-57. *Stochastic Differential Equations (Lecture Series in Differential Equations, Session Catholic University, Air Force Office of Scientific Research, Arlington.* MR0233437

[22] Nourdin, I. and Viens, F. G. (2009). Density formula and concentration inequalities with Malliavin calculus. *Electronic Journal of Probability* **14** 2287-2309. MR2556018

[23] Nualart, D. (2006). *The Malliavin calculus and related topics*, 2nd ed. Probability and Its Applications, Springer. MR2200233

[24] Periera, J. B. A. and Ibrahimi, M. (2014). Support recovery for the drift coefficient of high-dimensional diffusions. *IEEE Trnasactions of Information Theory* **60** 4026-4049. MR3225948

[25] Revuz, D. and Yor, M. (2005). *Continuous martingales and Brownian motion*, 3rd ed. A Series of Comprehensive Studies in Mathematics, Springer. MR1725357

[26] Spokoiny, V. (2012). Parametric estimation. Finite sample theory. *Annals of Statistics* **40** 2877-2909. MR3097963

[27] Spokoiny, V. (2017). Penalized maximum likelihood estimation and effective dimension. *Annales de l'Institut Henri Poincaré* **53** 1389-429. MR3606746

[28] Sznitman, A. S. (1991). Topics in propagation of chaos. In *École d'Été de Probabilités de Saint Flour XIX - 1989* (P. L. Hennequin, ed.) 165-251. volume 1464 of Lecture Notes in Mathematics, Springer, Berlin. MR1108185

[29] Vershynin, R. (2009). *Lectures in Geometric Functional Analysis.* available at romanv/papers/GFA-book/GFA-book.pdf.

[30] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* **101** 1418-1429. MR2279469