Sparse recovery of Diffusion Dynamics: Handling High-Dimensionality in Repeated Short Trajectories

Elise Bayraktar[1] **Charlotte Dion-Blanc**[1,2]

[1]LPSM, Sorbonne Université, France
[2]CMAP, École Polytechnique, France

Abstract

High-dimensional stochastic differential equations encode complex interaction structures within their drift component. We propose a novel approach to estimate this drift from independent high-frequency trajectory data observed over a short time horizon. Each trajectory is modelled as the solution of a Brownian-driven stochastic differential equation, while the number of time points within each path tends to infinity. We further assume that the drift function governing the dynamics can be expressed as a linear combination of a growing number of Lipschitz basis functions. To promote accurate recovery of the underlying dynamics under sparsity constraints, we propose a Lasso-regularised likelihood criterion. Under suitable regularity conditions, we establish convergence rates for the resulting estimator and emphasise how they depend on the dimensional parameters of the problem, in particular on the number of observed trajectories. We assess the performance of the estimator on synthetic datasets, both from an estimation and a generative perspective. Finally, we illustrate the practical relevance of the approach on a real-world climate dataset, highlighting its ability to perform variable selection.

1 INTRODUCTION

Stochastic differential equations (SDEs) provide a probabilistic framework for learning continuous-time dynamics from noisy observations. In dimension $d \geq 1$, they allow one to capture rich multivariate interactions while supporting statistically grounded inference and generative modelling. In this work, we assume that the observations are generated by the following equation,

$$\mathrm{d}X_t = -b_\theta(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \qquad X_0 = 0, \quad (1)$$

where W is a d-dimensional standard Brownian motion defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. The drift $b_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is driven by an unknown parameter $\theta \in \Theta \subset \mathbb{R}^p$, and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the diffusion coefficient.

Moreover, in order to tackle the major challenge that is the estimation of the drift in this framework, we assume that the *true* parameter θ_0 of dimension p is sparse. We investigate the estimation of θ_0 from N independent and identically distributed (i.i.d.) paths observed at n discrete times on the finite time interval $[0, 1]$. The quantities p, d, n and N go to infinity. We propose a Lasso (Least Absolute Shrinkage and Selection Operator) estimator of θ_0 [Tibshirani, 1996], which performs both variable selection and parameter estimation.

Related work. The drift estimation of SDEs has been widely studied assuming ergodicity and a long observation window. In this framework, the Lasso estimator has been studied e.g. by Amorino et al. [2025] both for a linear model and for the Ornstein-Uhlenbeck process, while De Gregorio et al. [2025] addressed drift and volatility estimation using a different choice of penalty, the Elastic-Net.

In many real-world applications, data naturally come as multiple short trajectories rather than a single long realisation. This calls for estimation and proof techniques that exploit repetition across paths instead of relying on asymptotic stationarity. Few references are available, recently Nakakita [2025] focused on the Ornstein-Uhlenbeck model from continuous observations, extending the ergodic results of Ciołek et al. [2025].

The benefits of regularisation have also been observed for Neural Networks. Taheri et al. [2021] introduced "scale regularisation" to induce sparsity in neural networks and obtain better theoretical guarantees. This type of Neural Networks with sparsity built-in has been applied to diffusion generative models by Taheri and Lederer [2025]. Similarly, sparse Neural Networks are used by Zhao et al. [2025] to

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:469–490.

estimate the drift of a diffusion process over a compact set from high-frequency observations, generalising the work of Oga and Koike [2024] in the ergodic framework.

Our Contribution. This paper studies the theoretical guarantees and generative performance of the Lasso estimator in the regime of repeated short trajectories. Our main contributions are as follows:

- **Oracle Inequality and Error Bounds:** We derive an oracle inequality for the Lasso estimator (Theorem 3.1) which holds on the intersection of high-probability sets. Based on this result, we establish bounds on the estimation error in both l_1 and l_2 norms.
- **Inference from Repeated Trajectories:** Unlike Amorino et al. [2025], our analysis does not rely on the ergodic properties of the process. Instead, we use concentration inequalities across the N independent trajectories to control the probabilistic events. We provide the convergence rate of the Lasso estimator, together with conditions on the dimension parameters n, N, p and d.
- **Diffusion Coefficient:** A key theoretical addition is to allow for a non-constant diffusion coefficient $\sigma(x)$, while previous works such as Amorino et al. [2025] assume constant volatility. We prove that the theoretical results hold under an ellipticity condition on σ.
- **Numerical Validation:** We demonstrate the efficacy of our proposed method through an extensive numerical study, establishing that the Lasso estimator outperforms Sparse Neural Networks [Zhao et al., 2025] in generative performance. Furthermore, we provide a practical application using climate data to highlight the model's real-world utility and motivation.

Organisation of the paper. In Section 2, we present the assumptions and the Lasso estimator. Section 3 contains the main results, including the control of the probability sets and error bounds. Finally, Section 4 presents a full numerical study on synthetic data, evaluating both estimation error and generative quality compared to the Sparse Neural Networks [Zhao et al., 2025]. We also present an illustration of variable selection on climate data from NOAA Physical Sciences Laboratory.

2 MODEL AND ESTIMATION METHOD

In the sequel, we assume that we repeatedly observe the solution of Equation (1) on the time interval $[0, 1]$ at equidistant discrete times. We denote the observations as

$$(X_{t_i}^{(j)})_{i=0,\dots,n-1},\ j \in [N],\ t_i := i\Delta_n = i/n,$$

where the notation $[N]$ refers to the integers from 1 to N. The increments of the process are denoted by

$$\Delta X_i^{(j)} := X_{t_i}^{(j)} - X_{t_{i-1}}^{(j)}.$$

2.1 ASSUMPTIONS AND NOTATIONS

We consider the case of the generalised linear drift, that is

$$b_\theta := \phi_0 + \sum_{k=1}^{p} \theta_k \phi_k. \tag{2}$$

The isolation of ϕ_0 serves as a geometric reference point for the dynamics. By fixing the scale and form of the autonomous drift ϕ_0, the coefficients θ_k can be interpreted as the sensitivities or coupling strengths of the system to auxiliary coordinates. This model, previously considered by Amorino et al. [2025] and Ciołek et al. [2025], is flexible and allows for the capture of a wide range of functions. For example, if ϕ_k is an Hilbert-basis then the class of drift functions that we consider approximates the space $L^2(\mathbb{R}^d)$.

Assumption 1 (Lipschitz)**.** *For all $0 \le k \le p$, the functions $\phi_k : \mathbb{R}^d \to \mathbb{R}^d$ are Lipschitz, and we denote by L_k their Lipschitz constant which are uniformly bounded in p and d.*

Assumption 2 (Sparsity)**.** *We assume that the true parameter of the model given in Equation* (1)*, that we denote by θ_0, satisfies the sparsity constraint:*

$$\|\theta_0\|_0 := \#\{1 \le k \le p, \quad \theta_{0,k} \neq 0\} = s. \tag{3}$$

Remark 2.1. *Assumption 1 implies that the drift function $b_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz, and we denote by L its Lipschitz norm. In particular $L \le L_0 + \sum_{k=1}^{p} \theta_k L_k$, and due to Assumption 2 the drift at the true parameter b_{θ_0} is Lipschitz with Lipschitz constant $L_0 + s \max_k \theta_{0,k} L_k$ bounded in p and d.*

Assumption 3 (Ellipticity)**.** *The matrix $\sigma\sigma^T$ is positive and satisfies a uniform ellipticity condition: there exists a constant $c_1 \ge 1$ such that*

$$\forall x \in \mathbb{R}^d, \quad \frac{1}{c_1} I_d(x) \le \sigma\sigma^T(x) \le c_1 I_d(x). \tag{4}$$

Remark 2.2. *Assumption 3 implies that $\sigma\sigma^T$ and $(\sigma\sigma^T)^{-1} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are spectrally bounded in the sense that there exists $C_\sigma > 0$ such that*

$$\sup_x \|\sigma\sigma^T(x)\|_2 + \sup_x \|(\sigma\sigma^T)^{-1}(x)\|_2 \le C_\sigma,$$

where the spectral norm is defined for $A \in \mathbb{R}^{d\times d}$ by $\|A\|_2 := \sup_{\|x\|_2 \le 1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^T A)}$.

Assumption 4. *The diffusion σ is of class $\mathcal{C}^2$, bounded, has bounded first derivatives, and the second derivatives have at most polynomial growth: for some positive constant c and some $q \in \mathbb{N}^*$ we have*

$$\forall x \in \mathbb{R}^d, \quad \|\sigma(x)\|_2 + \|\nabla\sigma(x)\|_2 \le cd^2,$$
$$\|\partial^2_{x_k x_l}\sigma(x)\|_2 \le c(1 + \|x\|_2^q).$$

Note that Assumption 1 and Assumption 4 imply that Equation (1) admits a unique strong solution (X_t), which is a homogeneous continuous Markov process.

We define the cone $\mathcal{C}(s,c)$ of approximately sparse vectors,

$$\mathcal{C}(s,c) := \{x \in \mathbb{R}^p \backslash \{0\}, \quad \|x\|_1 \leq (1+c)\|x_{|\mathcal{I}_s(x)}\|_1\},$$

where $\mathcal{I}_s(x)$ is the set of the s largest elements of x. In particular, if x is s-sparse, then $\|x\|_1 = \|x_{|\mathcal{I}_s(x)}\|_1$ and $x \in \mathcal{C}(s,c)$ for all $c \geq 0$.

Assumption 5 (Identifiability)**.** *There exists $\ell > 0$ such that for all $\theta, v \in \mathbb{R}^p$ satisfying $\|\theta\|_0 = s$ and $\theta - v \in \mathcal{C}(s, 3 + 4/\gamma)$ for an arbitrary γ, and for all $n \in \mathbb{N}^*$*

$$\frac{(\theta-v)^T \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n (\sigma^{-1}\Phi)^T(\sigma^{-1}\Phi)(X_{t_{i-1}})\right](\theta-v)}{\|\theta-v\|_2^2} > \ell,$$

where $\Phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \in \mathbb{R}^{d\times p}$ is the function whose k-th column is the function ϕ_k.

The identifiability assumption is comparable to the one considered in Ciołek et al. [2025], modified as this work assumes stationarity. It is satisfied for orthonormal bases with compact support such as the wavelet bases, as the transition density of the process is bounded away from zero on any compact domain.

Remark 2.3. *Let us emphasise that the considered model encompasses the Ornstein–Uhlenbeck model. In this case $p = d^2$ and $\phi_k(X) = E_{i,j}X$ where $k = (i,j)$ and $E_{i,j}$ is a $d \times d$ matrix with a single non-zero entry equal to 1 at coordinate (i,j). In this setting, Assumption 5 is verified when X is stationary, or more generally whenever $\mathbb{E}(\int_0^1 X_s X_s^T ds)$ is non-degenerate. This unified treatment is one of the main advantages of our approach, while the existing literature such as Amorino et al. [2025] develop separate arguments for the Ornstein–Uhlenbeck and the general linear model.*

Finally, due to the non-constant diffusion coefficient, we need to introduce the following weighted empirical norm $\|f\|_{\sigma,n,N} := \sqrt{\langle f,f\rangle_{\sigma,n,N}}$ with

$$\langle f,g\rangle_{\sigma,n,N} = \frac{1}{nN}\sum_{i=1}^n\sum_{j=1}^N \left[(\sigma^{-1}f)(X_{t_{i-1}}^{(j)})\right]^T(\sigma^{-1}g)(X_{t_{i-1}}^{(j)}).$$

2.2 ESTIMATOR

If the full trajectory of the N observations of the process $\{(X_t^{(j)})_{0\leq t\leq 1}\}_{j\in[N]}$ is observed continuously on the time interval $[0,1]$, we have an explicit scaled negative log-likelihood function using Girsanov's theorem, see e.g. Kutoyants [2004],

$$L_N(\theta) := \frac{1}{N}\sum_{j=1}^N\Big[\int_0^1 \left[(\sigma^{-1}b_\theta)(X_s^{(j)})\right]^T\sigma(X_s^{(j)})^{-1}\mathrm{d}X_s^{(j)} + \frac{1}{2}\int_0^1 \left\|\sigma(X_s^{(j)})^{-1}b_\theta(X_s^{(j)})\right\|_2^2\mathrm{d}s\Big].$$

We consider the least squares contrast based on the approximation of the likelihood. It generalises the one considered in Denis et al. [2020] and Amorino et al. [2025].

$$R_{n,N}(\theta) := \frac{1}{N}\sum_{i=1}^n\sum_{j=1}^N\Big\|\sigma(X_{t_{i-1}}^{(j)})^{-1}\Delta X_i^{(j)} + \Delta_n\sigma(X_{t_{i-1}}^{(j)})^{-1}b_\theta(X_{t_{i-1}}^{(j)})\Big\|_2^2.$$

We define the Lasso estimator $\widehat{\theta}$ as

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta}\{R_{n,N}(\theta) + \lambda\|\theta\|_1\}, \tag{5}$$

where $\lambda > 0$ is the regularisation parameter. The constant λ will be chosen as a function of N, n, p and d. The larger λ is, the stronger the penalty ($\|\theta\|_1$), and hence the sparser the estimator $\widehat{\theta}$.

3 THEORETICAL STUDY OF THE LASSO ESTIMATOR

3.1 HIGH PROBABILITY RANDOM SETS

To study the behaviour of the Lasso estimator, our first objective is to derive an oracle inequality. To this end, we need to control two error terms: a martingale component driven by the Brownian motion and an error arising from the time-discretisation of the continuous process.

We first control the stochastic noise term. To this end, we define the set $\mathcal{T}_{\mathrm{mart}}$ on which the martingale error is dominated by the regularisation parameter λ:

$$\mathcal{T}_{\mathrm{mart}} := \Big\{\Big\|\frac{1}{nN}\sum_{i=1}^n\sum_{j=1}^N \Phi(X_{t_{i-1}}^{(j)})^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)}) \times\Big(\int_{t_{i-1}}^{t_i}\sigma(X_s^{(j)})\mathrm{d}W_s^{(j)}\Big)\Big\|_\infty \leq \frac{\lambda}{4}\Big\}.$$

We then turn to the approximation error. When discretising the process, a bias appears involving the variation of the drift between t_{i-1} and t_i. We introduce the set $\mathcal{T}_{\mathrm{disc}}$, on which it is negligible:

$$\mathcal{T}_{\mathrm{disc}} := \Big\{\Big\|\frac{1}{nN}\sum_{i=1}^n\sum_{j=1}^N \Phi(X_{t_{i-1}}^{(j)})^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)}) \times\Big(\int_{t_{i-1}}^{t_i}\big(b_{\theta_0}(X_s^{(j)}) - b_{\theta_0}(X_{t_{i-1}}^{(j)})\big)\mathrm{d}s\Big)\Big\|_\infty \leq \frac{\lambda}{4}\Big\}.$$

Finally, we introduce the set $\mathcal{T}_{\text{comp}}$, which corresponds to the *compatibility condition*, common in the Lasso literature, [see *e.g.* Bühlmann and Van De Geer, 2011].

$$\mathcal{T}_{\text{comp}} := \left\{ \inf_{\substack{\theta\in\mathbb{R}^p:\|\theta\|_0=s,\\ v\in\mathbb{R}^p:\theta-v\in\mathcal{C}(s,3+4/\gamma)}} \frac{\|b_\theta - b_v\|_{\sigma,n,N}}{\|\theta - v\|_2} \geq k \right\}$$

for a constant $k > 0$ that will be chosen later. Intuitively, this condition ensures that the active components are not too correlated, so that distinct parameter values do not yield nearly identical drifts. It is fundamental for Lasso estimators as it links the prediction error (the numerator) to the estimation error (the denominator), which is crucial for deriving the typical oracle inequality.

3.2 ORACLE INEQUALITY

The main result of this section is the following oracle inequality, which controls the prediction error. We denote in particular the drift estimator:

$$b_{\widehat{\theta}} = \phi_0 + \sum_{k=1}^{p} \widehat{\theta}_k \phi_k.$$

It satisfies the following result.

Theorem 3.1. *Let $\theta \in \Theta$ such that $\|\theta\|_0 = s$. Then, for any $\gamma > 0$, we have on $\mathcal{T}_{\text{mart}} \cap \mathcal{T}_{\text{disc}} \cap \mathcal{T}_{\text{comp}}$*

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \leq (1+\gamma)\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + \frac{4\lambda^2 s(2+\gamma)^2}{k^2\gamma\Delta_n^2},$$

where λ is the tuning parameter of the Lasso estimator and k is the constant on $\mathcal{T}_{\text{comp}}$.

The proof of this result is detailed in Section B.2. This oracle inequality shows that, on a certain random set, the Lasso estimator achieves a prediction error that is, up to a multiplicative factor $(1 + \gamma)$ and a remainder term, as small as that of the best (s)-sparse approximation, thereby illustrating its near-oracle performance under sparsity.

Since the oracle inequality holds on the intersection $\mathcal{T}_{\text{mart}} \cap \mathcal{T}_{\text{disc}} \cap \mathcal{T}_{\text{comp}}$, the remainder of this section is dedicated to showing that this event occurs with high probability.

To this end, we define the set $\Omega_{n,N}$ on which the observations $(X^{(j)}_{t_{i-1}})_{1\leq i\leq n,\ 1\leq j\leq N}$ are bounded by a quantity depending on d, n and N,

$$\Omega_{n,N} := \left\{ \max_{1\leq i\leq n,\ 1\leq j\leq N} \|X^{(j)}_{t_{i-1}}\|_2 < \sqrt{d}\log(nN) \right\}. \tag{6}$$

Intuitively, the factor $\sqrt{d}$ in the bound reflects the order of magnitude of the norm of a Gaussian vector in $\mathbb{R}^d$. Besides, $\log(nN)$ is chosen to control the maximum over the $n \times N$ points, given the exponential decay of the process tails. We now state that this set has high-probability.

Theorem 3.2. *Under Assumptions 3 and 4 and for $\Omega_{n,N}$ defined in* (6)*, we have that for n large enough,*

$$\mathbb{P}(\Omega^c_{n,N}) \leq \frac{1}{nN},$$

where for any event B, we denote by $B^c := \Omega\backslash B$ its complement.

The last step is to be able to control the probabilities of the random sets $\mathcal{T}_{\text{mart}}$, $\mathcal{T}_{\text{disc}}$, and $\mathcal{T}_{\text{comp}}$. For a fixed value of $\varepsilon \in (0, 1)$, let:

$$\begin{aligned}
\lambda_1 &:= C_1 \frac{\log(nN)}{n}\sqrt{\frac{d}{N}}\sqrt{\log(2p) + \log(1/\varepsilon)},\\
\lambda_2 &:= C_2 \frac{sd\log(nN)^2}{n^{3/2}}\sqrt{\log(p) + \log(1/\varepsilon)},\\
N_1 &:= c_1 d^2 p^2 \log(nN)^4 \frac{(5+4/\gamma)^4}{(\ell - k^2)^2}\Big[\log(2/\varepsilon)\\
&\qquad + \log\big(21^{2s}[p^{2s} \wedge (ep/2s)^{2s}]\big)\Big],\\
N_2 &:= \frac{1}{n\varepsilon}.
\end{aligned}$$

Remark 3.3. *When comparing our constants to those in Amorino et al. [2025], λ_1 corresponds to $\max(\lambda_{1,1}, \lambda_{1,2})$. We observe that the dependence on the dimension d is improved from $d^{3/4}$ to $\sqrt{d}$, and the rate benefits from the number of repetitions N. The term λ_2 (discretisation error) maintains the same order as in previous work. However, a key difference appears in N_1, which is of order d^2p^2 compared to T_1 of order d^2. This stems from the proof techniques: we use Hoeffding's inequality on $\Phi \in \mathbb{R}^{d\times p}$, in contrast to the concentration inequality used in Amorino et al. [2025].*

We are now ready to provide a lower bound for the probability of $\mathcal{T}_{\text{mart}} \cap \mathcal{T}_{\text{disc}} \cap \mathcal{T}_{\text{comp}}$, ensuring that the oracle inequality holds with high probability.

Theorem 3.4. *Under Assumptions 1 - 5, for a fixed value of $\varepsilon \in (0, 1/4)$, for all $\lambda > \max\{\lambda_1, \lambda_2\}$ and $N > \max\{N_1, N_2\}$ we have*

$$\mathbb{P}(\mathcal{T}_{\text{mart}} \cap \mathcal{T}_{\text{disc}} \cap \mathcal{T}_{\text{comp}}) \geq 1 - 4\varepsilon.$$

Proof. Denoting by $A = \mathcal{T}_{\text{mart}} \cap \mathcal{T}_{\text{disc}} \cap \mathcal{T}_{\text{comp}}$ we decompose the probability as follows

$$\begin{aligned}
\mathbb{P}(A^c) &= \mathbb{P}(A^c \cap \Omega_{n,N}) + \mathbb{P}(A^c \cap \Omega^c_{n,N})\\
&= \mathbb{P}(A^c|\Omega_{n,N})\mathbb{P}(\Omega_{n,N}) + \mathbb{P}(A^c|\Omega^c_{n,N})\mathbb{P}(\Omega^c_{n,N})\\
&\leq \mathbb{P}(A^c|\Omega_{n,N}) + \mathbb{P}(\Omega^c_{n,N}).
\end{aligned}$$

Then, the result stems from Theorem 3.2 with Theorems B.3, B.4 and B.5, presented in Section B.3. We deduce that

$$\mathbb{P}(A^c) \leq 3\varepsilon + \frac{1}{nN} \leq 4\varepsilon.$$

□

3.3 CONVERGENCE RATES

As a consequence of Theorem 3.4, we obtain the following error bounds.

Corollary 3.5. *We assume Assumptions 1 - 5, and we recall that $\|\theta_0\|_0 = s$. For a fixed value of $\varepsilon \in (0, 1/4)$, for all $\lambda > \max\{\lambda_1, \lambda_2\}$, $N > \max\{N_1, N_2\}$ and $0 < k < \sqrt{\ell}$ (where ℓ is given in Assumption 5) we have with probability at least $1 - 4\varepsilon$ that*

$$\|\widehat{\theta} - \theta_0\|_2^2 \leq \frac{4\lambda^2 n^2 s(2+\gamma)^2}{k^4\gamma},$$

and

$$\|\widehat{\theta} - \theta_0\|_1 \leq \frac{8\lambda n s(1+\gamma)(2+\gamma)}{k^2\gamma^{3/2}}.$$

Proof. From Theorem 3.4, we can work on $\mathcal{T}_{\text{mart}} \cap \mathcal{T}_{\text{disc}} \cap \mathcal{T}_{\text{comp}} \cap \Omega_{n,N}$ with probability at least $1 - 4\varepsilon$. If we apply the oracle inequality to $\theta = \theta_0$, we have

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|_{\sigma,n,N}^2 \leq \frac{4\lambda^2 n^2 s(2+\gamma)^2}{k^2\gamma}.$$

On $\mathcal{T}_{\text{comp}}$ we have $\|b_{\widehat{\theta}} - b_{\theta_0}\|_{\sigma,n,N} \geq k\|\widehat{\theta} - \theta_0\|_2$ and

$$\|\widehat{\theta} - \theta_0\|_2^2 \leq \frac{4\lambda^2 n^2 s(2+\gamma)^2}{k^4\gamma}.$$

The l_1 bound is obtained using that $\widehat{\theta} - \theta_0 \in \mathcal{C}(s, 3 + 4/\gamma)$ and Cauchy-Schwarz inequality

$$\begin{aligned}\|\widehat{\theta} - \theta_0\|_1 &\leq (4 + \frac{4}{\gamma})\|\widehat{\theta}|_{S(\theta_0)} - \theta_0\|_1 \\ &\leq \frac{4(\gamma+1)}{\gamma}\sqrt{s}\|\widehat{\theta} - \theta_0\|_2 \\ &\leq \frac{8\lambda n s(1+\gamma)(2+\gamma)}{k^2\gamma^{3/2}},\end{aligned}$$

where $S(\theta_0)$ is the support of θ_0, and we used the sparsity condition from Assumption 2. □

Finally, we present asymptotic convergence rates depending on whether the martingale or the discretisation error dominates.

Corollary 3.6. *Assume Assumptions 1 - 5 and the sparsity condition $\|\theta_0\|_0 = s$. For a fixed value of $\varepsilon \in (0, 1/4)$, for all $N > \max\{N_1, N_2\}$ and $0 < k < \sqrt{\ell}$ we have with probability at least $1 - 4\varepsilon$ that*

1. *If $\log(nN)s\frac{\sqrt{dN}}{\sqrt{n}} \to 0$, then the discretisation error is negligible and*

$$\|\widehat{\theta} - \theta_0\|_2 = O_{\mathbb{P}}\left(\frac{\log(nN)\sqrt{ds\log(p)}}{k^2\sqrt{N}}\right).$$

2. *If $\log(nN)s\frac{\sqrt{dN}}{\sqrt{n}} \to \infty$, then the discretisation error dominates and*

$$\|\widehat{\theta} - \theta_0\|_2 = O_{\mathbb{P}}\left(\frac{d}{k^2}\log(nN)^2\sqrt{\frac{s^3\log(p)}{n}}\right).$$

Let us notice that our results parallel those obtained in the ergodic framework by Amorino et al. [2025]. As the number of repetitions N is analogous to the time horizon T in their analysis, our estimator achieves the same convergence rates, up to logarithmic terms. Besides, the order $(s/N)^{-1/2}$ is the same order found in Nakakita [2025] for the special case of Ornstein-Uhlenbeck process with repeated continuous observations.

4 NUMERICAL ILLUSTRATION

In this section, we present numerical results to confirm our theoretical findings. First, we evaluate the estimator on simulated data, and observe how the estimation error (in l_2 norm) decreases as the number of observed trajectories N increases, particularly for high-dimensional parameters. Then, we compare the generative performance of our approach against the Sparse Neural Network proposed by Zhao et al. [2025], using the Wasserstein distance and Maximum Mean Discrepancy (MMD). Finally, we apply our method on real climate data. By applying our estimator to the stochastic regression model proposed by De Gregorio et al. [2025], we aim to explain which climate indices are predictors of the global mean temperature.

4.1 SYNTHETIC DATA

We simulate the d-dimensional diffusion process solution of (1) for the choice of drift function

$$b_\theta(x) = x + \sum_{k=1}^{p}\theta_k\cos((k+1)x), \qquad x \in \mathbb{R}^d,$$

where the cosine function is applied component-wise. The trajectories are simulated using an Euler-Maruyama scheme with 5000 discretisation steps.

In the following simulations, the parameter $\theta_0 \in \mathbb{R}^p$ is generated with 80% sparsity, with non-zero coefficients uniformly distributed in $[1, 2]$, and we work in dimension $d = 10$. The process is observed at high frequency with $n = 500$, and we study the impact of N and p. We first consider the classical diffusion $\sigma = I_d$, which satisfies Assumptions 3 and 4. Some simulated trajectories of the process across its 10 dimensions are presented in Figure 1. We then consider the diffusion coefficient

$$\sigma(x) = \text{diag}\Big(\frac{x_i}{\sqrt{1+x_i^2}}\Big),$$

which is bounded but does not satisfy Assumption 3 as $\sigma(0) = 0_d$ is not invertible.

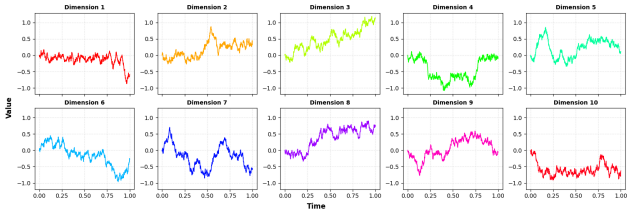


Figure 1: Simulated sample paths of the process for $\sigma = I_d$.

4.2 ESTIMATOR IMPLEMENTATION

Minimisation algorithm. We aim to minimise $\theta \mapsto R_{n,N}(\theta) + \lambda\|\theta\|_1$. The objective is the sum of two functions: the contrast function is convex and smooth, while the l_1 norm is convex but non-differentiable when a coordinate is equal to zero. This motivates the use of the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) proposed by Beck and Teboulle [2009] over other gradient descent algorithms. Compared to the standard ISTA, the rate of convergence is improved from $O(1/K)$ to $O(1/K^2)$ (with K the number of iterations) thanks to Nesterov's accelerated gradient descent method. This method computes the next iterate using a specific linear combination of the two previous points, instead of the last point. To compute the gradient step, the algorithm requires the Lipschitz constant of $\nabla R_{n,N}$, which we numerically approximate. We terminate the algorithm when the distance between iterates is below a threshold or after 200 iterations.

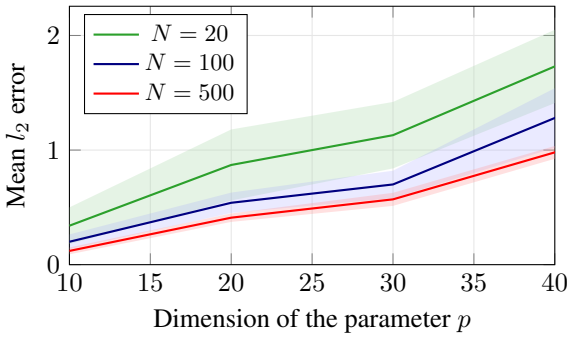
Choice of the penalisation. The key choice in the minimisation is the choice of the tuning parameter λ, which controls the l_1 penalisation and the number of active coefficients. The penalisation parameter λ is chosen from a thin grid (of size 30) using the Extended Bayesian Information Criteria (EBIC) introduced in Chen and Chen [2008]. For some $\gamma \in [0, 1]$

$$EBIC_\gamma(\lambda) = 2L_{\sigma,n,N}(\widehat{\theta}_\lambda) + |S(\widehat{\theta}_\lambda)| \log(nN) + 2\gamma \log\left(\binom{p}{|S(\widehat{\theta}_\lambda)|}\right),$$

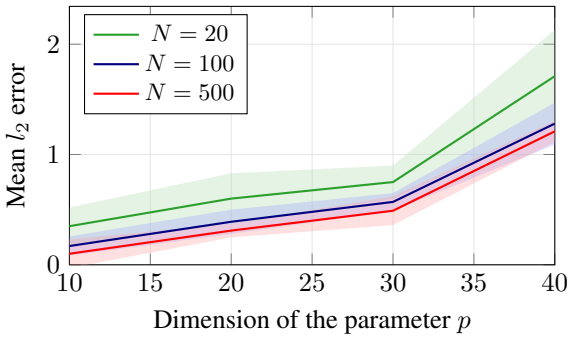
where $\widehat{\theta}_\lambda$ is the Lasso estimator with regularisation parameter λ, $|S(\widehat{\theta}_\lambda)|$ is the size of the support of $\widehat{\theta}_\lambda$ i.e. the number of non-zero coefficients, and $L_{\sigma,n,N}$ is the negative log-likelihood of the model. Compared to cross-validation, EBIC is less computationally expensive. As in Denis et al. [2026], we observe that EBIC performs better for parameter selection for large values of p.

4.3 ESTIMATION ERROR

To measure the accuracy of the Lasso estimator, we first measure the parameter estimation error in l_2 norm $\|\widehat{\theta} - \theta^*\|_2$. We study the behaviour of this error as the parameter dimension p (and therefore the number of non-zero coefficients s) and the number of observed trajectories N increase. For each value of $p \in \{10, 20, 30, 40\}$, a random value of θ_0 is selected and the errors are averaged over 30 iterations. The results are presented in Figure 2. We observe that the error decreases with a larger number of trajectories N, but increases as p and s grow, which is consistent with the theoretical convergence rates presented in Corollary 3.6. We remark that the estimator is consistent, despite $\sigma(x) = \text{diag}\left(\frac{x_i}{\sqrt{1+x_i^2}}\right)$ not satisfying the ellipticity condition required by our theoretical framework. Finally, we present a heat map of the estimated coefficients for $p = 20$ and $\sigma = I_d$ in Figure 3 to highlight the support recovery of the Lasso estimator.



(a) $\sigma = I_d$.



(b) $\sigma(x) = \text{diag}\left(\frac{x_i}{\sqrt{1+x_i^2}}\right)$

Figure 2: Mean error of the Lasso estimator ± one standard deviation for different values of N.

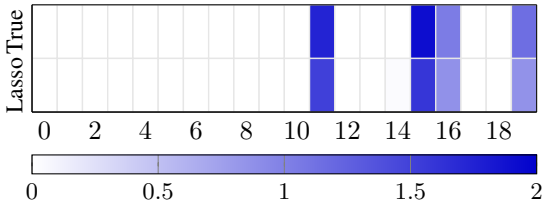


Figure 3: Heat map comparison of the sparse true parameter and the estimated parameter ($N = 100, p = 20, \sigma = I_d$).

4.4 GENERATIVE GAP

Benchmark. We benchmark the generative performance of our approach against the Sparse Neural Network (Sparse NN) proposed by Zhao et al. [2025], which also incorporates sparsity. The layers are of size $(d, 16, 16, 1)$ and the sparsity ratio is $s_{ratio} = 0.25$.

To evaluate the quality of the simulated paths, we generate $K = 500$ independent trajectories $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(K)}\}$ from the true SDE, and trajectories $\{\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \dots, \mathbf{Y}^{(K)}\}$ simulated using the drift estimated by either our Lasso estimator or the Sparse NN.

The fidelity of the generated paths is quantified using two metrics: the Wasserstein distance and the Maximum Mean Discrepancy (MMD), defined as follows.

Comparison metrics Our first metric to compare paths is the Wasserstein distance, defined for empirical distributions with uniform weights by

$$W_r(\mathbf{X}, \mathbf{Y}) = \min_{\pi} \left(\frac{1}{K} \sum_{j=1}^{K} \|\mathbf{X}^{(j)} - \mathbf{Y}^{(\pi(j))}\|_2^r \right)^{1/r},$$

where the minimum is over all permutations π of K elements, and the trajectories $\mathbf{X}^{(j)}, \mathbf{Y}^{(j)}$ can be seen as vectors in very high dimension $\mathbb{R}^{(n+1)d}$. We consider $r = 1$. In high dimensions, this formula is very computationally expensive, and requires complex methods such as the Hungarian algorithm in $O(K^3)$. In dimension one however, the Wasserstein distance has a closed-form formula. This motivated the introduction of the Sliced Wasserstein by Bonneel et al. [2015], computed as an average of linear projections. We use a Monte Carlo approximation over 500 directions.

The second metric is the Maximum mean discrepancy (MMD), which was originally proposed in Gretton et al. [2012] as a way to test if two distributions are the same. The MMD has become a standard benchmark in generative modelling to assess the fidelity of generated samples [see e.g. Li et al., 2015]. A probability distribution P is embedded into a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ via the mean embedding $\mu_P = \mathbb{E}_{\mathbf{X}\sim P}[k(\cdot, \mathbf{X})]$, for k a kernel. The Maximum mean discrepancy is then defined as $\text{MMD} = \|\mu_P - \mu_Q\|_{\mathcal{H}}$. For two sets of K trajectories $\mathbf{X}$ and $\mathbf{Y}$, we compute the unbiased estimator of the squared MMD proposed by Gretton et al. [2012] using a Gaussian kernel whose variance is determined via the median heuristic, after flattening the trajectories to vectors in $\mathbb{R}^{(n+1)d}$.

Numerical results. The generative fidelity of the estimated model compared to the Sparse Neural Network is illustrated in Figure 4 for $p = 30$ and $\sigma(x) = \text{diag}\left(x_i/\sqrt{1+x_i^2}\right)$. We observe that the Lasso estimator significantly outperforms the Sparse Neural Network in terms of Wasserstein distance and MMD. The gray line, labelled "True model", represents the Wasserstein distance computed between two independent sets of K trajectories generated from the true SDE. Note that this value is non-zero due to the finite sample size. Our estimator rapidly converges to this oracle bound. In particular, we note from Figure 4 that despite the violation of the ellipticity assumption and a large parameter dimension, the Lasso estimator accurately reconstructs the drift.

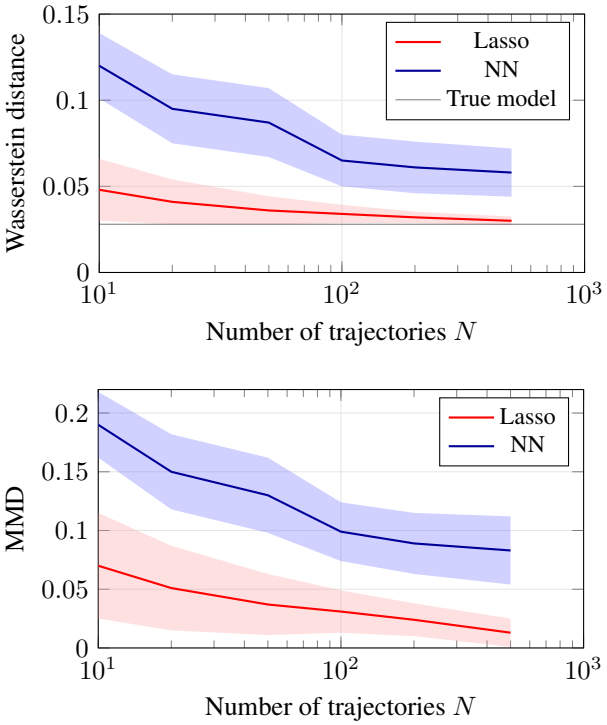


Figure 4: Performance of the Lasso estimator and the Sparse Neural Network for path generation ($p = 30$, $\sigma(x) = \text{diag}(x_i/\sqrt{1+x_i^2})$).

4.5 APPLICATION TO CLIMATE DATA

We conclude with an illustrative application to real climate data using the monthly climate indices provided by the NOAA Physical Sciences Laboratory (National Oceanic and Atmospheric Administration), available at `https://psl.noaa.gov/`. This example aims to show that our methodology is readily applicable to regression problems and can identify a meaningful subset of explanatory variables.

Dataset. Our target variable is the Global Mean Land/Ocean Temperature. We select diverse predictive in-

dices representing various physical components and geographic regions. We restrict our analysis to indices with complete records from 1951 to 2023. As seasonality dominates the data, we subtract the monthly average and normalise by the monthly standard deviation. Following the methodology in Chatterjee et al. [2012], we then remove the linear trend to filter out the long-term global warming.

We partition the data into 10-year segments and treat each decade as an independent realisation. This segmentation strikes a balance between the high-frequency sampling ($n = 120$) and the number of repetitions ($N = 7$). While the assumption of i.i.d. trajectories is a simplification of the true dynamics, the preliminary de-trending and de-seasonalisation steps mitigate the dependencies, making the approximation reasonable, see Figure 5.

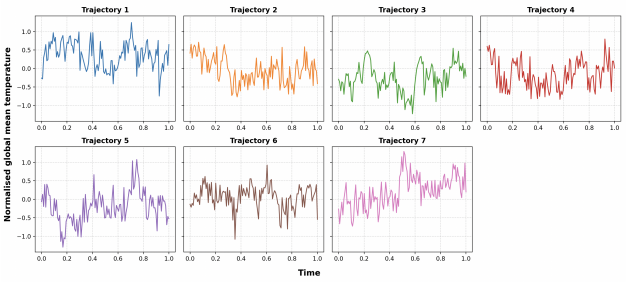


Figure 5: 10-year trajectories of the Global Temperature, after normalisation.

Model. To model the predictive relationship between climate indices, we consider the stochastic regression model proposed by De Gregorio et al. [2025]. We consider a system of d variables where the first $d-1$ components (the predictors) are modelled as independent Ornstein–Uhlenbeck processes with parameter 1. The target variable $X^{(d)}$ is then coupled to these predictors through a linear drift term. The full system is given by:

$$\begin{cases} \mathrm{d}X_t^{(i)} = -X_t^{(i)}\mathrm{d}t + \mathrm{d}W_t^{(i)}, \quad \text{for } 1 \le i < d, \\ \mathrm{d}X_t^{(d)} = -\left(X_t^{(d)} + \sum_{k=1}^{d-1} \theta_k X_t^{(k)}\right)\mathrm{d}t + \mathrm{d}W_t^{(d)}. \end{cases}$$

Variable selection and accuracy. Figure 6 illustrates the variable selection performed by our model. The Lasso estimator selects 5 predictors out of the 17 candidates.

First, we observe that the variables driving an increase in global temperatures, associated with negative coefficients, are Nino3.4 and the Western Hemisphere Warm Pool (WHWP). Nino3.4 represents the dominant mode of interannual variability [Trenberth, 1997], while the WHWP constitutes the second-largest region of warm water on Earth, acting as a major heat source [Wang and Enfield, 2001, 2003]. In addition, the model selects the North Atlantic

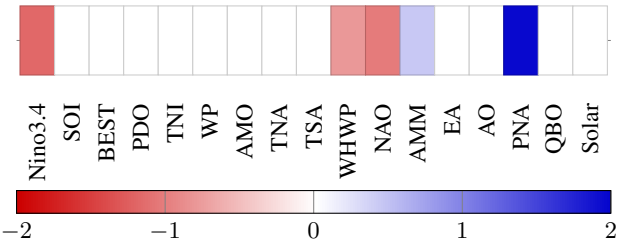


Figure 6: Estimated Lasso coefficients for the prediction of the Global Mean Temperature.

Oscillation (NAO) and the Pacific-North American (PNA) pattern. These indices capture the atmospheric variability, and dominate the winter climate variability over the Northern Hemisphere [Wallace and Gutzler, 1981]. Finally, the Atlantic Meridional Mode (AMM) [Chiang and Vimont, 2004] is selected, capturing the dynamics of the tropical Atlantic.

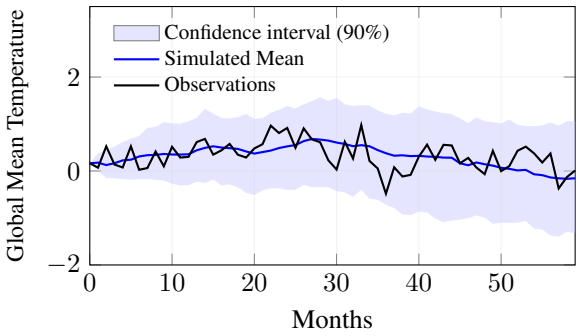


Figure 7: Comparison between observed and reconstructed Global Mean Temperature trajectories.

To assess the accuracy of our selected model, we perform a reconstruction experiment. Using the estimated drift parameters, we simulate 100 trajectories of the Global Mean Temperature conditional on the observed paths of the predictors. The results are displayed in Figure 7.

We observe that the simulated mean closely follows the trend of the actual observations. Furthermore, the true trajectory remains almost entirely within the confidence interval. This confirms that our model successfully captures the dynamics of the Global Mean Temperature.

As a last remark, we also fitted the maximum likelihood estimator for comparison. Because of its poor predictive performance, we did not include the corresponding figures. Still, we note that, as expected, it does not perform any variable selection, making the estimated coefficients difficult to interpret.

5 CONCLUSION AND LIMITATIONS

We introduced a sparse likelihood-based approach for learning drift dynamics from repeated short trajectories of SDEs. Our results clarify how replication can compensate for limited time horizons, both from theoretical and generative perspectives. To support reproducibility and further research, the code is made publicly available at `https://github.com/elisebayraktar/LASSO`. An interesting perspective is to investigate the case of Lévy processes, to extend the work of Dexheimer and Strauch [2024] to this modern framework of observations.

Acknowledgements

The work of Elise Bayraktar is funded by the 2022 DAE 103 EMERGENCE(S) - PROCECO project supported by Ville de Paris. This paper was completed while C. Dion-Blanc was affiliated with the Centre de Mathématiques Appliquées (CMAP) at Ecole Polytechnique. She is grateful to Sylvie Méléard for the opportunity to join her team, supported by the European Union (ERC, SINGER, 101054787). Both authors are grateful to Christophe Denis from Université Paris 1 for fruitful discussions.

References

Chiara Amorino, Francisco Pina, and Mark Podolskij. Sampling effects on lasso estimation of drift functions in high-dimensional diffusion processes. Electronic Journal of Statistics, 19(2):5068–5116, 2025.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci., 2(1):183–202, 2009.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. J. Math. Imaging Vision, 51(1):22–45, 2015.

Peter Bühlmann and Sara Van De Geer. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, 2011.

Soumyadeep Chatterjee, Karsten Steinhaeuser, Arindam Banerjee, Snigdhansu Chatterjee, and Auroop Ganguly. Sparse group lasso: Consistency and climate applications. Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012, pages 47–58, 04 2012.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. Biometrika, 95(3):759–771, 2008.

John C. H. Chiang and Daniel J. Vimont. Analogous pacific and atlantic meridional modes of tropical atmosphere–ocean variability. Journal of Climate, 17(21): 4143 – 4158, 2004.

Gabriela Ciołek, Dmytro Marushkevych, and Mark Podolskij. On Lasso estimator for the drift function in diffusion models. Bernoulli, 31(3):1811 – 1833, 2025.

Alessandro De Gregorio, Dario Frisardi, Stefano Iacus, and Francesco Iafrate. Adaptive elastic-net estimation for sparse diffusion processes. Statistical Inference for Stochastic Processes, 28(3):22, 2025.

Christophe Denis, Charlotte Dion, and Miguel Martinez. Consistent procedures for multiclass classification of discrete diffusion paths. Scandinavian Journal of Statistics, 47(2):516–554, 2020.

Christophe Denis, Charlotte Dion-Blanc, Romain E. Lacoste, and Laure Sansonnet. Lasso-type estimator and classification algorithm for high-dimensional multivariate hawkes processes. Scandinavian Journal of Statistics, 53 (2):919–968, 2026.

Niklas Dexheimer and Claudia Strauch. On Lasso and Slope drift estimators for Lévy-driven Ornstein-Uhlenbeck processes. Bernoulli, 30(1):88–116, 2024.

Avner Friedman. Partial differential equations of parabolic type. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1964.

Avner Friedman. Stochastic differential equations and applications. Vol. 1. Probability and Mathematical Statistics, Vol. 28. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975.

Emmanuel Gobet. LAN property for ergodic diffusions with discrete observations. Ann. Inst. H. Poincaré Probab. Statist., 38(5):711–737, 2002.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. J. Mach. Learn. Res., 13:723–773, 2012.

Yury A. Kutoyants. Statistical inference for ergodic diffusion processes. Springer Series in Statistics. Springer-Verlag London, 2004.

Yujia Li, Kevin Swersky, and Richard Zemel. Generative moment matching networks. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, page 1718–1727. JMLR.org, 2015.

Terry J. Lyons and Weian Zheng. On conditional diffusion processes. Proceedings of the Royal Society of Edinburgh: Section A Mathematics, 115(3–4):243–255, 1990.

Shogo Nakakita. Sparse estimation for the drift of high-dimensional Ornstein-Uhlenbeck processes with i.i.d. paths. arXiv preprint arXiv:2510.21505, 2025.

Akihiro Oga and Yuta Koike. Drift estimation for a multi-dimensional diffusion process using deep neural networks. Stochastic Processes and their Applications, 170:104240, 2024. ISSN 0304-4149.

Mahsa Taheri and Johannes Lederer. Regularization can make diffusion models more efficient. arXiv preprint arXiv:2502.09151, 2025.

Mahsa Taheri, Fang Xie, and Johannes Lederer. Statistical guarantees for regularized neural networks. Neural Networks, 142:148–161, 2021. ISSN 0893-6080.

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

Kevin E. Trenberth. The definition of el niño. Bulletin of the American Meteorological Society, 78(12):2771 – 2778, 1997.

Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

John M. Wallace and David S. Gutzler. Teleconnections in the geopotential height field during the northern hemisphere winter. Monthly Weather Review, 109(4):784 – 812, 1981.

Chunzai Wang and David B. Enfield. The tropical western hemisphere warm pool. Geophysical Research Letters, 28(8):1635–1638, 2001.

Chunzai Wang and David B. Enfield. A further study of the tropical western hemisphere warm pool. Journal of Climate, 16(10):1476 – 1493, 2003.

Yuzhen Zhao, Yating Liu, and Marc Hoffmann. Drift estimation for diffusion processes using neural networks based on discretely observed independent paths. arXiv preprint arXiv:2511.11161, 2025.

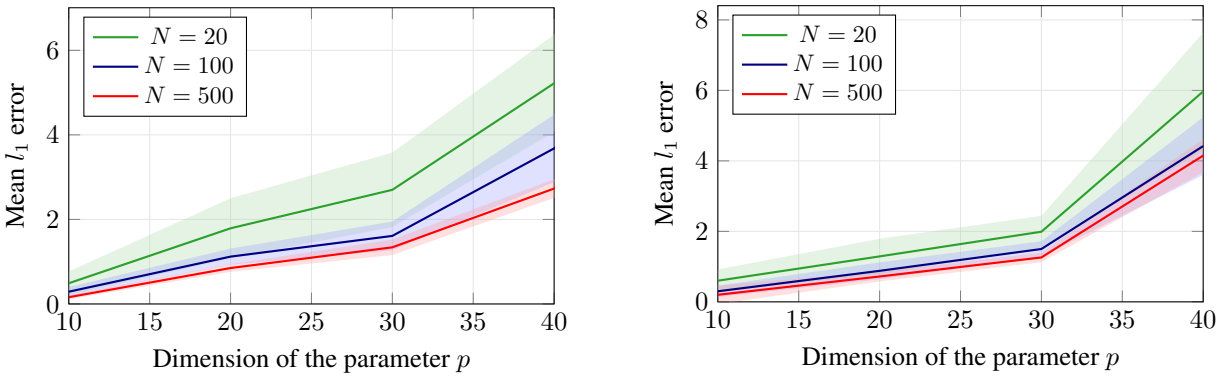
Sparse recovery of Diffusion Dynamics: Handling High-Dimensionality in Repeated Short Trajectories (Supplementary Material)

Elise Bayraktar[1] **Charlotte Dion-Blanc**[1,2]

[1]LPSM, Sorbonne Université, France
[2]CMAP, École Polytechnique, France

A ADDITIONAL NUMERICAL RESULTS

In this section, we provide additional numerical results on simulated data. We first present in Figure 8 the parameter estimation error in l_1 norm $\|\widehat{\theta} - \theta^*\|_1$, which measures the accuracy of the Lasso estimator. Similarly to what was observed in Figure 2, the error decreases with a larger number of trajectories N, but increases as p and s grow.



(a) $\sigma = I_d$.

(b) $\sigma(x) = \text{diag}(x_i/\sqrt{1+x_i^2})$

Figure 8: Mean error of the Lasso estimator ± one standard deviation for different values of N.

Figure 9 evaluates the generative fidelity of the estimated model compared to the Sparse Neural Network for $p = 10$ and $\sigma = I_d$. As in Figure 4, the Lasso estimator significantly outperforms the Sparse Neural Network in terms of Wasserstein distance and MMD.

The evolution of the l_2 error as a function of N is presented in Figure 10. For different values of p and of the volatiliyt coefficient, we have fitted a curve of the form $a\frac{\log(500x)}{\sqrt{x}} + c$, where c corresponds to the discretisation error with becomes dominant for large values of N. We see that our estimation error fits well with the predicted rate.

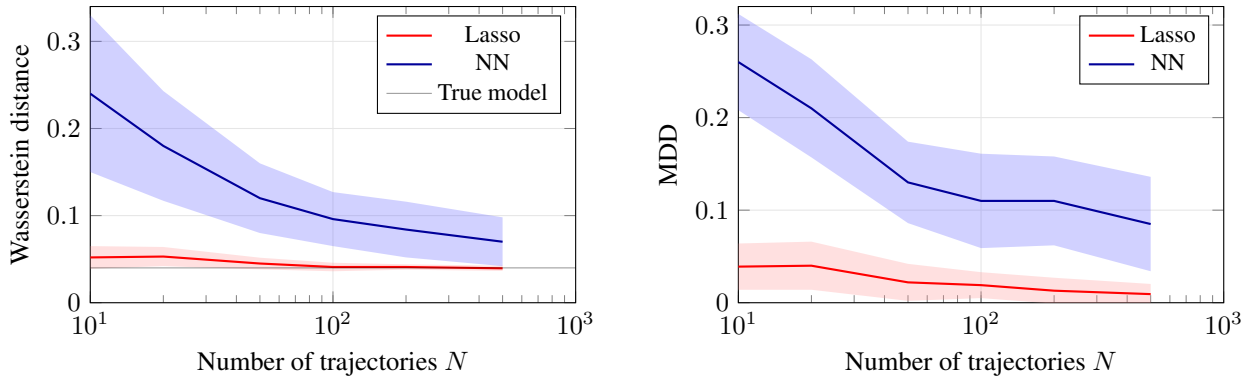
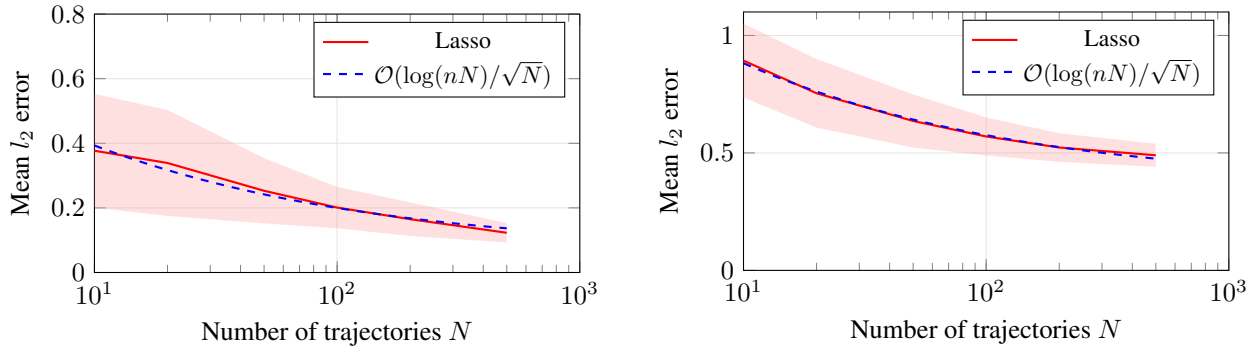


Figure 9: Performance of the Lasso estimator and the Sparse Neural Network for path generation ($p = 10$, $\sigma = I_d$).



(a) $p = 10,\ \sigma = I_d$.

(b) $p = 30,\ \sigma(x) = \mathrm{diag}(x_i/\sqrt{1+x_i^2})$

Figure 10: Mean error of the Lasso estimator ± one standard deviation.

B PROOFS

B.1 TECHNICAL LEMMAS

Lemma B.1. *Under Assumptions 1 - 3, there exists a time $t_1 > 0$ and a constant $C > 0$ (independent of d) such that for all t_0, for all $t < t_1$, and for all $r > 1$,*

$$\mathbb{E}\left[\sup_{u\leq t}\left\|X^{(j)}_{t_0+u} - X^{(j)}_{t_0}\right\|_2^{2r}\right] \leq C^r\left(t^{2r}s^{2r}\left(\mathbb{E}\left[\left\|X^{(j)}_{t_0}\right\|_2^{2r}\right] + \max_k \left\|\phi_k(0)\right\|_2^{2r}\right) + t^r r^r d^r\right).$$

Proof. Using the inequality $|a+b|^p \leq 2^{p-1}(|a|^p + |b|^p)$ for $p \geq 1$, we first write

$$\left\|X^{(j)}_{t_0+u} - X^{(j)}_{t_0}\right\|_2^{2r} \leq 2^{2r-1}\left(\left\|\int_{t_0}^{t_0+u} b_{\theta_0}(X^{(j)}_v)\mathrm{d}v\right\|_2^{2r} + \left\|\int_{t_0}^{t_0+u}\sigma(X^{(j)}_v)\mathrm{d}W^{(j)}_v\right\|_2^{2r}\right).$$

The drift term is bounded using Hölder's inequality and the sparsity of the true parameter θ_0

$$\begin{aligned}\sup_{u\leq t}\left\|\int_{t_0}^{t_0+u} b_{\theta_0}(X^{(j)}_v)\mathrm{d}v\right\|_2^{2r} &\leq t^{2r-1}\int_{t_0}^{t_0+t}\left\|\sum_{k=0}^{p}\theta_{0,k}\phi_k(X^{(j)}_v)\right\|_2^{2r}\mathrm{d}v \\ &\leq t^{2r-1}s^{2r}\max_k|\theta_{0,k}|^{2r}\int_{t_0}^{t_0+t}\left\|\phi_k(X^{(j)}_v)\right\|_2^{2r}\mathrm{d}v.\end{aligned}$$

We decompose $\phi_k(X^{(j)}_v) = \phi_k(X^{(j)}_v) - \phi_k(X^{(j)}_{t_0}) + \phi_k(X^{(j)}_{t_0}) - \phi_k(0) + \phi_k(0)$, and recall that the ϕ_k are uniformly

Lipschitz from Assumption 1 and that $|a+b+c|^p \leq 3^{p-1}(|a|^p + |b|^p + |c|^p)$ for $p \geq 1$ to obtain

$$\mathbb{E}\left[\sup_{u\leq t}\left\|\int_{t_0}^{t_0+u} b_{\theta_0}(X_v^{(j)})\mathrm{d}v\right\|_2^{2r}\right] \tag{7}$$
$$\leq t^{2r}s^{2r}3^{2r-1}\max_k|\theta_{0,k}|^{2r}\left(L_k^{2r}\mathbb{E}\left[\sup_{u\leq t}\left\|X_{t_0+u}^{(j)}-X_{t_0}^{(j)}\right\|_2^{2r}\right]+L_k^{2r}\mathbb{E}\left[\left\|X_{t_0}^{(j)}\right\|_2^{2r}\right]+\|\phi_k(0)\|_2^{2r}\right).$$

The volatility term is a martingale. We first apply Hölder's inequality

$$\left\|\int_{t_0}^{t_0+u}\sigma(X_v^{(j)})\mathrm{d}W_v^{(j)}\right\|_2^{2r} \leq d^{r-1}\sum_{l=1}^d\left|\left(\int_{t_0}^{t_0+u}\sigma(X_v^{(j)})\mathrm{d}W_v^{(j)}\right)_l\right|^{2r}.$$

Using Burkholder's inequality on the one-dimensional martingale (with constant $C_{2r,Bur}\leq C\sqrt{r}$ for C independent of r) and Hölder's inequality, we have for any $1\leq l\leq d$

$$\begin{aligned}\mathbb{E}\left[\sup_{u\leq t}\left|\left(\int_{t_0}^{t_0+u}\sigma(X_v^{(j)})\mathrm{d}W_v^{(j)}\right)_l\right|^{2r}\right] &\leq C_{2r,Bur}^{2r}\mathbb{E}\left[\left(\int_{t_0}^{t_0+t}\sum_{m=1}^d\left(\sigma_{l,m}(X_v^{(j)})\right)^2\mathrm{d}v\right)^r\right]\\ &\leq C^{2r}r^rt^{r-1}\mathbb{E}\left[\int_{t_0}^{t_0+t}\left(\sum_{m=1}^d\left(\sigma_{l,m}(X_v^{(j)})\right)^2\right)^r\mathrm{d}v\right]\\ &\leq C^{2r}r^rt^rC_\sigma^{2r}. \end{aligned}\tag{8}$$

Combining equations (7) and (8), we have

$$\begin{aligned}\mathbb{E}\left[\sup_{u\leq t}\left\|X_{t_0+u}^{(j)}-X_{t_0}^{(j)}\right\|_2^{2r}\right] \leq& C_1^{2r}t^{2r}s^{2r}\mathbb{E}\left[\sup_{u\leq t}\left\|X_{t_0+u}^{(j)}-X_{t_0}^{(j)}\right\|_2^{2r}\right]\\ &+C_2^r\Big(t^{2r}s^{2r}(\mathbb{E}\left[\left\|X_{t_0}^{(j)}\right\|_2^{2r}\right]+\max_k\|\phi_k(0)\|_2^{2r})+t^rr^rd^r\Big),\end{aligned}$$

where the constants C_1 and C_2 do not depend on d or r. For a sufficiently small t_1 we have $C_1t_1s<1/2$, hence for $r\geq 1$ and $t<t_1$ we have $C_1^{2r}t^{2r}s^{2r}<C_1ts<1/2$. This allows us to conclude that $\exists t_1>0$, $\exists C>0$, $\forall t<t_1$, $\forall r>1$

$$\mathbb{E}\left[\sup_{u\leq t}\left\|X_{t_0+u}^{(j)}-X_{t_0}^{(j)}\right\|_2^{2r}\right]\leq C^r\Big(t^{2r}s^{2r}(\mathbb{E}\left[\left\|X_{t_0}^{(j)}\right\|_2^{2r}\right]+\max_k\|\phi_k(0)\|_2^{2r})+t^rr^rd^r\Big).$$

□

Lemma B.2. *Let X be a solution of*

$$\mathrm{d}X_t=-b(X_t)\mathrm{d}t+\sigma(X_t)\mathrm{d}W_t,\quad X_0=0.$$

Assume that b is Lipschitz and that σ satisfies Assumptions 3 and 4, then the transition density $p(t,x,y)$ of X satisfies

$$p(t,x,y)\leq\frac{C^{d/2}}{t^{d/2}}\exp(-\frac{\|x-y\|_2^2}{ct})\exp(ct\|x\|_2^2),$$

for some constant $C>0$ and $c>0$ independent of the dimension.

Proof. Similarly to the proof of Proposition 1.2 in Gobet [2002], we first consider the SDE where the drift coefficient is removed $X_t=X_0+\int_0^t\sigma(X_s)\mathrm{d}W_s$ where W is a Brownian motion under $\mathbb{P}^0$, and we denote by $p^0(t,x,y)$ its density. We will later prove that for some constant K and c independent of the dimension

$$p^0(t,x,y)\leq\frac{K^{d/2}}{t^{d/2}}\exp\Big(-\frac{\|x-y\|_2^2}{ct}\Big),\tag{9}$$

$$\|\nabla_xp^0(t,x,y)\|_2\leq\frac{K^{d/2}}{t^{(d+1)/2}}\exp\Big(-\frac{\|x-y\|_2^2}{ct}\Big).\tag{10}$$

We set

$$Z_t = \exp\left(-\int_0^t (\sigma^{-1}(X_s)b(X_s))^T \mathrm{d}W_s - \frac{1}{2}\int_0^t \|\sigma^{-1}(X_s)b(X_s)\|_2^2 \,\mathrm{d}s\right).$$

Under the Assumptions $\sigma^{-1}b$ has linear growth, ensuring that Z is a martingale. By Girsanov's theorem, we have:

$$p(t,x,y) = p^0(t,x,y)\mathbb{E}_x^0(Z_t|X_t = y). \tag{11}$$

To deal with $\mathbb{E}_x^0(Z_t|X_t = y)$, we use the law of the diffusion bridge from $X_0 = x$ to $X_t = y$ [see Lyons and Zheng, 1990] to transform the Brownian motion W into $W + \int \sigma(X_u)\frac{\nabla_x p^0(t-u,X_u,y)}{p^0(t,x,y)}\mathrm{d}u$. Hence $Z_t = 1 - \int_0^t Z_s\sigma^{-1}(X_s)b(X_s)\mathrm{d}W_s$ and

$$\mathbb{E}_x^0(Z_t|X_t = y) = 1 - \frac{1}{p^0(t,x,y)}\int_0^t \mathbb{E}_x^0[Z_s b(X_s).\nabla_x p^0(t-s,X_s,y)]\mathrm{d}s. \tag{12}$$

We now bound $|\mathbb{E}_x^0[Z_s b(X_s).\nabla_x p^0(t-s,X_s,y)]|$ using Hölder's inequality with conjugate exponents μ_1, μ_2. We use the following moment bound for Z (proven below): for any $\mu > 1$, there are some constants $T(d) > 0$, and $c > 0,\ K > 0$ independent of the dimension such that for all $t < T(d)$

$$\mathbb{E}_x^0(Z_t^\mu(1 + \|X_t\|_2)^\mu) \le K\exp(ct\|x\|_2^2)(1 + \|x\|_2)^\mu. \tag{13}$$

This, combined with equations (9) and (10) and the linear growth of b, allows us to write

$$\begin{aligned}
&|\mathbb{E}_x^0[Z_s b(X_s).\nabla_x p^0(t-s,X_s,y)]| \\
&\quad\le K^{d/2}\exp(cs\|x\|_2^2)(1+\|x\|_2)\left[\int_{\mathbb{R}^d}\frac{dz}{s^{d/2}(t-s)^{(d+1)\mu_2/2}}\exp\Big(-\frac{\|x-z\|_2^2}{cs} - \mu_2\frac{\|z-y\|_2^2}{c(t-s)}\Big)\right]^{1/\mu_2} \\
&\quad\le K^{d/2}\exp(ct\|x\|_2^2)(1+\|x\|_2)\frac{1}{t^{d/2\mu_2}(t-s)^{(d+1)/2-d/2\mu_2}}\exp\Big(-\frac{\|x-y\|_2^2}{c't}\Big).
\end{aligned}$$

We choose μ_2 close to 1 so that $(d+1)/2 - d/2\mu_2 < 1$ in order to ensure that the time integral converges. Coming back to (12), we get

$$\mathbb{E}_x^0(Z_t|X_t = y) \le 1 + \frac{K^{d/2}}{p^0(t,x,y)}\frac{\exp(ct\|x\|_2^2)(1+\|x\|_2)}{t^{(d-1)/2}}\exp(-\frac{\|x-y\|_2^2}{c't}),$$

and we conclude using (11), the lower bound on p^0(9) and that $\frac{\exp(ct\|x\|_2^2)\|x\|_2}{t^{(d-1)/2}} \le K\frac{\exp(c't\|x\|_2^2)}{t^{d/2}}$.

Proof of (13)**.** From Lemma B.1 $\forall q > 1$

$$\mathbb{E}_x^0((1+\|X_t\|_2)^q) \le K^q(tdq)^{q/2}(1+\|x\|_2)^q,$$

so, up to Hölder's inequality and choosing $T(d)$ so that $T(d)dq < 1$, we only need to upper bound $\mathbb{E}_x^0(Z_t^q)$. Fix $\lambda \ge 0$. As b is Lipschitz, there exists a constant $K > 0$, depending on the ellipticity constant c_1 from Assumption 3 but independent of the dimension, such that

$$\lambda\int_0^t \|\sigma^{-1}(X_s)b(X_s)\|_2^2\mathrm{d}s \le \lambda Kt\Big(\|x\|_2^2 + \sup_{s\in[0,t]}\|X_s - x\|_2^2\Big).$$

Taking the exponential and then the expectation, we obtain

$$\mathbb{E}_x^0\left[\exp\left(\lambda\int_0^t \|\sigma^{-1}(X_s)b(X_s)\|_2^2\mathrm{d}s\right)\right] \le \exp(\lambda Kt\|x\|_2^2)\cdot\mathbb{E}_x^0\left[\exp\left(\lambda Kt\sup_{s\in[0,t]}\|X_s - x\|_2^2\right)\right].$$

Using the Taylor expansion of the exponential function and the monotone convergence theorem, we write

$$\mathbb{E}_x^0\left[\exp\left(\lambda Kt\sup_{s\in[0,t]}\|X_s - x\|_2^2\right)\right] = \sum_{r=0}^\infty\frac{(\lambda Kt)^r}{r!}\mathbb{E}_x^0\left[\sup_{s\in[0,t]}\|X_s - x\|_2^{2r}\right].$$

Plugging the results of Lemma B.1, for t small there exists a constant $C > 0$ such that

$$\mathbb{E}_x^0\left[\exp\left(\lambda K t \sup_{s\in[0,t]}\|X_s - x\|_2^2\right)\right] \leq \sum_{r=0}^{\infty}\frac{(\lambda K t)^r}{r!}C^r\left(t^{2r}s^{2r}\|x\|_2^{2r} + t^r r^r d^r\right)$$
$$\leq \exp(\lambda K C t^3 s^2\|x\|_2^2) + \sum_{r=0}^{\infty}\frac{(\lambda K t)^r}{r!}C^r t^r r^r d^r.$$

For the second term, we use the fact that $r! \geq (r/e)^r$ hence

$$\sum_{r=0}^{\infty}\frac{(\lambda K t)^r}{r!}C^r t^r r^r d^r \leq \sum_{r=0}^{\infty}(\lambda K C t^2 d e)^r.$$

This is a geometric series, we choose $t \leq T(d)$ small enough such that $\lambda K C t^2 d e < 1$. Consequently, the series converges to a constant $A > 0$ independent of x, and for all $t \leq T(d)$

$$\mathbb{E}_x^0\left(\exp(\lambda\int_0^t\|\sigma^{-1}(X_s)b(X_s)\|_2^2\mathrm{d}s)\right) \leq A\exp(ct\|x\|_2^2). \tag{14}$$

In order to bound $\mathbb{E}_x^0(Z_t^q)$, we rewrite

$$Z_t^q = \exp\left(q\int_0^t(\sigma^{-1}(X_s)b(X_s))^T\mathrm{d}W_s - \frac{q}{2}\int_0^t\|\sigma^{-1}(X_s)b(X_s)\|_2^2\mathrm{d}s\right)$$
$$= \exp\left(q\int_0^t(\sigma^{-1}b(X_s))^T\mathrm{d}W_s - q^2\int_0^t\|\sigma^{-1}b(X_s)\|_2^2\mathrm{d}s\right)\exp\left((q^2-\frac{q}{2})\int_0^t\|\sigma^{-1}b(X_s)\|_2^2\mathrm{d}s\right).$$

We conclude using Cauchy–Schwarz inequality, where the first term is a martingale and has expectation equal to 1 and the second one is estimated by Equation (14).

Proof of (9)**.** Equation (9) and (10) are extensions of Equation (4.12) in Friedman [1975] by tracking the dependence of the dimension d in the constants. In the following, we give a brief sketch of the proof, following the work done by Friedman [1964]. We denote by $a = \sigma\sigma^T$, and we introduce the frozen transition density $Z(y,t;x,\tau)$, where the volatility is fixed at the starting point $\sigma(x)$. For $t > \tau$

$$Z(y,t;x,\tau) = \frac{1}{(2\pi)^{d/2}\det(a(x))^{1/2}(t-\tau)^{d/2}}\exp\Big(-\frac{(y-x)^T a^{-1}(x)(y-x)}{2(t-\tau)}\Big)$$
$$\leq \frac{K^{d/2}}{(t-\tau)^{d/2}}\exp\Big(-c\frac{\|x-y\|_2^2}{t-\tau}\Big), \tag{15}$$

where the constants $K > 0$ and $c > 0$ depend on the ellipticity constant c_1 from Assumption 3. We use the parametrix method to go from this first estimate to the density p^0, writing

$$p^0(t,x,y) = Z(y,t;x,0) + \int_0^t\int_{\mathbb{R}^d}Z(y,t;\eta,\tau)\Phi(\eta,\tau;x,0)\mathrm{d}\eta\mathrm{d}\tau, \tag{16}$$

where Φ is a solution of the Volterra equation

$$\Phi(\eta,\tau;x,0) = LZ(\eta,\tau;x,0) + \int_0^\tau\int_{\mathbb{R}^d}LZ(\eta,\tau;\xi,v)\Phi(\xi,v;x,0)\mathrm{d}\xi\mathrm{d}v,$$

with kernel

$$LZ(y,t;x,\tau) = \frac{1}{2}\sum_{i,j=1}^{d}[a_{ij}(y) - a_{ij}(x)]\frac{\partial^2 Z}{\partial y_i\partial y_j}(y,t;x,\tau).$$

We note that

$$\frac{\partial^2 Z}{\partial y_i\partial y_j}(y,t;x,\tau) = \left[\frac{(a^{-1}(x)(y-x))_i(a^{-1}(x)(y-x))_j}{(t-\tau)^2} - \frac{a^{-1}(x)_{ij}}{(t-\tau)}\right]Z(y,t;x,\tau).$$

In order to bound LZ, we use Assumption 4 to bound $a(y) - a(x)$, and we bound $\frac{\partial^2 Z}{\partial y_i \partial y_j}(y,t;x,\tau)$ using (15) and the ellipticity from Assumption 3. Some powers of $\frac{\|y-x\|_2^2}{t-\tau}$ are absorbed into the exponential, and the constants which are polynomial in d are negligible in front of $K^{d/2}$, hence

$$|LZ(y,t;x,\tau)| \leq \frac{K^{d/2}}{(t-\tau)^{(d+1)/2}} \exp\Big(-c' \frac{\|x-y\|_2^2}{t-\tau}\Big).$$

We next bound Φ by writing

$$\Phi(\eta,\tau;x,0) = \sum_{\nu=1}^{\infty} (LZ)_\nu(\eta,\tau;x,0),$$

where $(LZ)_\nu$ are the iterative convolutions of LZ. Following the same estimates as in Friedman [1964, Eq. 4.8], we get

$$|\Phi(\eta,\tau;x,0)| \leq \frac{K^{d/2}}{\tau^{(d+1)/2}} \exp\Big(-c\frac{\|\eta-x\|_2^2}{\tau}\Big). \tag{17}$$

Finally, substituting (17) and (15) back into (16) and applying Lemma 3 in Friedman [1964] to bound the integral, we conclude that there exist constants $K > 0$, $c > 0$ independent of the dimension such that

$$p^0(t,x,y) \leq \frac{K^{d/2}}{t^{d/2}} \exp\Big(-\frac{\|x-y\|_2^2}{ct}\Big).$$

□

B.2 PROOF OF THEOREM 3.1

Theorem 3.1. *Let $\theta \in \Theta$ such that $\|\theta\|_0 = s$. Then, for any $\gamma > 0$, we have on $\mathcal{T}_{\text{mart}} \cap \mathcal{T}_{\text{disc}} \cap \mathcal{T}_{\text{comp}}$*

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|_{\sigma,n,N}^2 \leq (1+\gamma)\|b_\theta - b_{\theta_0}\|_{\sigma,n,N}^2 + \frac{4\lambda^2 s(2+\gamma)^2}{k^2\gamma\Delta_n^2},$$

where λ is the tuning parameter of the Lasso estimator and k is the constant on $\mathcal{T}_{\text{comp}}$.

Proof. We first compute the contrast function

$$\begin{aligned} R_{n,N}(\theta) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N} \big\|\sigma(X_{t_{i-1}}^{(j)})^{-1}\Delta X_i^{(j)}\big\|_2^2 + \frac{2}{Nn}\sum_{i=1}^{n}\sum_{j=1}^{N}\big[\sigma(X_{t_{i-1}}^{(j)})^{-1}b_\theta(X_{t_{i-1}}^{(j)})\big]^T\sigma(X_{t_{i-1}}^{(j)})^{-1}\Delta X_i^{(j)} \\ + \Delta_n\|b_\theta - b_{\theta_0}\|_{\sigma,n,N}^2 - \Delta_n\|b_{\theta_0}\|_{\sigma,n,N}^2 + 2\Delta_n\langle b_\theta, b_{\theta_0}\rangle_{\sigma,n,N}. \end{aligned}$$

By definition of the Lasso estimator we have

$$R_{n,N}(\widehat{\theta}) - R_{n,N}(\theta) \leq \lambda\big(\|\theta\|_1 - \|\widehat{\theta}\|_1\big),$$

therefore for any $\theta \in \Theta$,

$$\Delta_n\|b_{\widehat{\theta}} - b_{\theta_0}\|_{\sigma,n,N}^2 \leq \Delta_n\|b_\theta - b_{\theta_0}\|_{\sigma,n,N}^2 + 2\langle b_\theta - b_{\widehat{\theta}}, \Delta X + \Delta_n b_{\theta_0}\rangle_{\sigma,n,N} + \lambda\big(\|\theta\|_1 - \|\widehat{\theta}\|_1\big). \tag{18}$$

Recalling the definition of X given in (1), we write

$$\begin{aligned} \langle b_\theta - b_{\widehat{\theta}}, \Delta X + \Delta_n b_{\theta_0}\rangle_{\sigma,n,N} = \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Big(b_\theta(X_{t_{i-1}}^{(j)}) - b_{\widehat{\theta}}(X_{t_{i-1}}^{(j)})\Big)^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\left(\int_{t_{i-1}}^{t_i}\sigma(X_s^{(j)})\mathrm{d}W_s^{(j)}\right) \\ + \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Big(b_\theta(X_{t_{i-1}}^{(j)}) - b_{\widehat{\theta}}(X_{t_{i-1}}^{(j)})\Big)^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\left(\int_{t_{i-1}}^{t_i}(b_{\theta_0}(X_{t_{i-1}}^{(j)}) - b_{\theta_0}(X_s^{(j)}))\mathrm{d}s\right). \end{aligned}$$

Let us introduce the two terms

$$A(\theta_1,\theta_2) := \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Big(b_{\theta_1}(X^{(j)}_{t_{i-1}}) - b_{\theta_2}(X^{(j)}_{t_{i-1}})\Big)^T(\sigma\sigma^T)^{-1}(X^{(j)}_{t_{i-1}})\left(\int_{t_{i-1}}^{t_i}(b_{\theta_0}(X^{(j)}_{t_{i-1}}) - b_{\theta_0}(X^{(j)}_s))\mathrm{d}s\right),$$

$$G(\theta_1,\theta_2) := \frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Big(b_{\theta_1}(X^{(j)}_{t_{i-1}}) - b_{\theta_2}(X^{(j)}_{t_{i-1}})\Big)^T(\sigma\sigma^T)^{-1}(X^{(j)}_{t_{i-1}})\left(\int_{t_{i-1}}^{t_i}\sigma(X^{(j)}_s)\mathrm{d}W^{(j)}_s\right).$$

This allows us to write (18) as

$$\Delta_n\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \le \Delta_n\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + 2A(\theta,\widehat{\theta}) + 2G(\theta,\widehat{\theta}) + \lambda\big(\|\theta\|_1 - \|\widehat{\theta}\|_1\big).$$

We observe that

$$|A(\theta,\widehat{\theta})| \le \|\theta-\widehat{\theta}\|_1\left\|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Phi(X^{(j)}_{t_{i-1}})^T(\sigma\sigma^T)^{-1}(X^{(j)}_{t_{i-1}})\left(\int_{t_{i-1}}^{t_i}(b_{\theta_0}(X^{(j)}_s) - b_{\theta_0}(X^{(j)}_{t_{i-1}}))\mathrm{d}s\right)\right\|_\infty,$$

$$|G(\theta,\widehat{\theta})| \le \|\theta-\widehat{\theta}\|_1\left\|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Phi(X^{(j)}_{t_{i-1}})^T(\sigma\sigma^T)^{-1}(X^{(j)}_{t_{i-1}})\left(\int_{t_{i-1}}^{t_i}\sigma(X^{(j)}_s)\mathrm{d}W^{(j)}_s\right)\right\|_\infty,$$

hence we obtain that, on the event $\mathcal{T}_{\mathrm{mart}}\cap\mathcal{T}_{\mathrm{disc}}$,

$$\Delta_n\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \le \Delta_n\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + \lambda\big(\|\theta\|_1 - \|\widehat{\theta}\|_1 + \|\theta-\widehat{\theta}\|_1\big).$$

We work here assuming that $\theta\in\Theta$ is such that $\|\theta\|_0 = s$, and let us denote $\mathcal{S}(\theta) = \mathrm{supp}(\theta)$. Then,

$$\begin{aligned}\|\theta\|_1 - \|\widehat{\theta}\|_1 + \|\theta-\widehat{\theta}\|_1 &= \|\theta\|_1 - \|\widehat{\theta}_{|S(\theta)}\|_1 - \|\widehat{\theta}_{|S(\theta)^c}\|_1 + \|\theta-\widehat{\theta}_{|S(\theta)}\|_1 + \|\widehat{\theta}_{|S(\theta)^c}\|_1\\ &\le 2\|\widehat{\theta}_{|S(\theta)} - \theta\|_1,\end{aligned}$$

hence

$$\Delta_n\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} + \lambda\|\widehat{\theta}-\theta\|_1 \le \Delta_n\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + 4\lambda\|\widehat{\theta}_{|S(\theta)} - \theta\|_1. \tag{19}$$

Finally, given a constant $\gamma>0$, we will consider two scenarios. On the one hand, if $4\lambda\|\widehat{\theta}_{|S(\theta)}-\theta\|_1 \le \gamma\Delta_n\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N}$ then we obtain from (19) that

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \le (1+\gamma)\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N}. \tag{20}$$

On the other hand, if $4\lambda\|\widehat{\theta}_{|S(\theta)} - \theta\|_1 > \gamma\Delta_n\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N}$ we obtain from (19) that

$$\|\theta-\widehat{\theta}\|_1 \le (4+\frac{4}{\gamma})\|\widehat{\theta}_{|S(\theta)} - \theta\|_1,$$

which implies that $\theta-\widehat{\theta}\in\mathcal{C}(s,3+4/\gamma)$. Applying Cauchy–Schwarz inequality in (19), we obtain that

$$\Delta_n\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} + \lambda\|\widehat{\theta}-\theta\|_1 \le \Delta_n\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + 4\lambda\sqrt{s}\|\widehat{\theta}-\theta\|_2.$$

On $\mathcal{T}_{\mathrm{comp}}$, we have $\|b_\theta - b_{\widehat{\theta}}\|_{\sigma,n,N} \ge k\|\widehat{\theta}-\theta\|_2$, hence

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \le \|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + 4\lambda\frac{\sqrt{s}}{k\Delta_n}(\|b_\theta - b_{\theta_0}\|_{\sigma,n,N} + \|b_{\theta_0} - b_{\widehat{\theta}}\|_{\sigma,n,N}).$$

Using the inequality $2ab \le \frac{a^2}{\varepsilon} + \varepsilon b^2$ twice with $a = \lambda\sqrt{s}$, $b = \|b_\theta - b_{\theta_0}\|_{\sigma,n,N}$ and then with $b = \|b_{\theta_0} - b_{\widehat{\theta}}\|_{\sigma,n,N}$, we obtain

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \le \|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + \frac{2}{k\Delta_n}\left(\varepsilon\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + \varepsilon\|b_{\theta_0} - b_{\widehat{\theta}}\|^2_{\sigma,n,N} + \frac{2\lambda^2 s}{\varepsilon}\right).$$

Rearranging the terms to isolate $\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N}$, we get

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \le \frac{k\Delta_n + 2\varepsilon}{k\Delta_n - 2\varepsilon}\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + \frac{4\lambda^2 s}{\varepsilon k\Delta_n}\frac{1}{1-\frac{2\varepsilon}{k\Delta_n}}.$$

In order for the first term to match with $(1+\gamma)$ from Equation (20) in the case $4\lambda\|\widehat{\theta}_{|S(\theta)} - \theta\|_1 \leq \gamma\Delta_n\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N}$, we choose $\varepsilon = \frac{k\gamma\Delta_n}{2(2+\gamma)}$ and obtain

$$\|b_{\widehat{\theta}} - b_{\theta_0}\|^2_{\sigma,n,N} \leq (1+\gamma)\|b_\theta - b_{\theta_0}\|^2_{\sigma,n,N} + \frac{4\lambda^2 s(2+\gamma)^2}{k^2\gamma\Delta_n^2},$$

which completes the proof. □

B.3 CONTROL OF THE RANDOM SET PROBABILITY

B.3.1 Proof of Theorem 3.2

Theorem 3.2. *Under Assumptions 3 and 4, and for $\Omega_{n,N}$ defined in* (6)*, we have that for n large enough,*

$$\mathbb{P}(\Omega^c_{n,N}) \leq \frac{1}{nN}.$$

Proof. We set $j \in [N]$ and $0 \leq t \leq 1$. We have from Lemma B.2 that the transition density of $X_t^{(j)}$ satisfies

$$p(t,x,y) \leq \frac{C^{d/2}}{t^{d/2}} \exp\Big(-\frac{\|x-y\|_2^2}{ct}\Big) \exp(ct\|x\|_2^2),$$

for some constant $C > 0$ and $c > 0$ independent of the dimension. We recall that for all $j \in [N]$, $X_0^{(j)} = 0$. For any $K > 0$

$$\begin{aligned}\mathbb{P}(\|X_t^{(j)}\|_2 > K) &= \int_{y\in\mathbb{R}^d, \|y\|_2>K} p(t,0,y)\mathrm{d}y \leq \int_{y\in\mathbb{R}^d,\|y\|_2>K} \frac{C^{d/2}}{t^{d/2}} \exp\Big(-\frac{\|y\|_2^2}{ct}\Big)\mathrm{d}y \\ &\leq C^{d/2}\mathbb{P}(\|Y\|_2 > K/\sqrt{ct}),\end{aligned}$$

where $Y \sim \mathcal{N}(0, I_d)$ is a standard Gaussian variable of dimension d, hence $\|Y\|_2^2$ has a chi-squared distribution with d degrees of freedom. For $0 < s < 1/2$, we obtain using Markov's inequality that for all $0 \leq t \leq 1$

$$\mathbb{P}(\|Y\|_2^2 > K^2/ct) \leq \mathbb{E}(e^{sX_d^2})e^{-sK^2/ct} \leq (1-2s)^{-d/2}e^{-sK^2/ct} \leq (1-2s)^{-d/2}e^{-sK^2/c}.$$

If $K^2/c > d$, this expression is minimised for $s = \frac{1}{2}(1 - \frac{dc}{K^2}) < 1/2$

$$\mathbb{P}(\|Y\|_2^2 > K^2/ct) \leq \exp\Big(-\frac{K^2/c - d}{2} + \frac{d}{2}\log(\frac{K^2}{cd})\Big).$$

Coming back to X, we write

$$\mathbb{P}(\|X_t^{(j)}\|_2 > K) \leq \exp(-\frac{K^2}{2c}) \exp\left(\frac{d}{2}\Big(1 + \log C + \log(\frac{K^2}{cd})\Big)\right). \tag{21}$$

If K^2/cd is sufficiently large, then $1 + \log C + \log(\frac{K^2}{cd}) < \frac{K^2}{2cd}$ and $\mathbb{P}(\|X_t^{(j)}\|_2 > K) \leq \exp(-K^2/4c)$.

We now define $\Omega_{n,N}$ of the form

$$\Omega_{n,N} = \left\{\max_{1\leq i\leq n,\ 1\leq j\leq N} \|X^{(j)}_{t_{i-1}}\|_2^2 < dA^2_{n,N}\right\},$$

and we detail the choice of $A_{n,N}$. Using Equation (21) for $K = \sqrt{d}A_{n,N}$, we have

$$\begin{aligned}\mathbb{P}(\Omega^c_{n,N}) &= \mathbb{P}\Big(\bigcup_{1\leq i\leq n}\bigcup_{1\leq j\leq N} \|X^{(j)}_{t_{i-1}}\|_2^2 \geq dA^2_{n,N}\Big) \\ &\leq \sum_{1\leq i\leq n}\sum_{1\leq j\leq N} \mathbb{P}\big(\|X^{(j)}_{t_{i-1}}\|_2^2 \geq dA^2_{n,N}\big) \\ &\leq nN\exp(-dA^2_{n,N}/4c).\end{aligned}$$

Choosing $A_{n,N} = \log(nN)$, we have for n large enough $\frac{d\log(nN)^2}{4c} > 2\log(nN)$ and

$$\mathbb{P}(\Omega_{n,N}^c) \leq \frac{1}{nN}.$$

□

B.3.2 The Martingale Set

Theorem B.3. *Under Assumptions 1 - 3, for a fixed value of $\varepsilon \in (0, 1/4)$ and for all $\lambda > \lambda_1$, it holds that*

$$\mathbb{P}(\mathcal{T}_{\text{mart}}^c | \Omega_{n,N}) \leq \varepsilon.$$

Proof. From now on, we are working on $\Omega_{n,N}$, defined in Equation (6). We want to upper bound

$$\begin{aligned}
\mathbb{P}(\mathcal{T}_{\text{mart}}^c) &= \mathbb{P}\Big(\Big\|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Phi(X_{t_{i-1}}^{(j)})^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\int_{t_{i-1}}^{t_i}\sigma(X_s^{(j)})\mathrm{d}W_s^{(j)}\Big\|_\infty > \frac{\lambda}{4}\Big) \\
&\leq p\max_k \mathbb{P}\Big(\Big|\frac{1}{N}\sum_{j=1}^{N} Z_k^{(j)}\Big| > \frac{\lambda}{4}\Big),
\end{aligned}$$

where $Z_k^{(j)}$ is the one-dimensional martingale

$$\begin{aligned}
Z_k^{(j)} &:= \Big(\frac{1}{n}\sum_{i=1}^{n}\Phi(X_{t_{i-1}}^{(j)})^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\int_{t_{i-1}}^{t_i}\sigma(X_s^{(j)})\mathrm{d}W_s^{(j)}\Big)_k \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{d}(\phi_k(X_{t_{i-1}}^{(j)}))_l\int_{t_{i-1}}^{t_i}\sum_{m=1}^{d}\big[(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\,\sigma(X_s^{(j)})\big]_{l,m}\mathrm{d}W_{m,s}^{(j)} \\
&= \frac{1}{n}\sum_{m=1}^{d} Y_m^{(j)},
\end{aligned}$$

with

$$Y_m^{(j)} := \sum_{i=1}^{n}\int_{t_{i-1}}^{t_i}\sum_{l=1}^{d}(\phi_k(X_{t_{i-1}}^{(j)}))_l\,\big[(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\,\sigma(X_s^{(j)})\big]_{l,m}\mathrm{d}W_{m,s}^{(j)}.$$

For a fixed j and conditionally on $(X_s^{(j)})_s$, the random variables $(Y_m^{(j)})_{1\leq m\leq d}$ are independent, centered and Gaussian, with

$$\text{Var}(Y_m^{(j)}|(X_s^{(j)})_s) = \sum_{i=1}^{n}\int_{t_{i-1}}^{t_i}\Big(\sum_{l=1}^{d}(\phi_k(X_{t_{i-1}}^{(j)}))_l\,\big[(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\,\sigma(X_s^{(j)})\big]_{l,m}\Big)^2\mathrm{d}s.$$

Using the independence to sum the variances, we have

$$\begin{aligned}
\text{Var}(Z_k^{(j)}|(X_s^{(j)})_s) &= \frac{1}{n^2}\sum_{m=1}^{d}\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i}\left(\sum_{l=1}^{d}(\phi_k(X_{t_{i-1}}^{(j)}))_l\,\big[(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\,\sigma(X_s^{(j)})\big]_{l,m}\right)^2\mathrm{d}s \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i}\sum_{m=1}^{d}\Big(\phi_k^T(X_{t_{i-1}}^{(j)})\,(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\,\sigma(X_s^{(j)})\Big)_m^2\mathrm{d}s \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i}\Big\|\phi_k^T(X_{t_{i-1}}^{(j)})\,(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\,\sigma(X_s^{(j)})\Big\|_2^2\mathrm{d}s.
\end{aligned}$$

We upper bound the conditional variance using that the functions ϕ_k are Lipschitz with constant L_k (uniformly bounded in p or d through Assumption 1) and the ellipticity of σ from Assumption 3

$$\begin{aligned}\mathrm{Var}(Z_k^{(j)}|(X_s^{(j)})_s) &\leq \frac{1}{n^2}\sum_{i=1}^{n}\int_{t_{i-1}}^{t_i}\left\|\phi_k(X_{t_{i-1}}^{(j)})\right\|_2^2\left\|(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\,\sigma(X_s^{(j)})\right\|_2^2\mathrm{d}s\\ &\leq \frac{n\Delta_n}{n^2}2\left(L_k^2\|X_{t_{i-1}}^{(j)}\|_2^2+\|\phi_k(0)\|_2^2\right)\,C_\sigma^4.\end{aligned}$$

As we are working on $\Omega_{n,N}$, we upper bound the sub-gaussian norm of the martingale $Z_k^{(j)}$

$$\|Z_k^{(j)}\|_{\psi^2}^2 \leq \frac{C}{n^2}n\Delta_n d\log(nN)^2.$$

Using sub-gaussian Hoeffding inequality [see *e.g.* Vershynin, 2018] on the average $\frac{1}{N}\sum_{j=1}^{N}Z_k^{(j)}$, we have

$$\mathbb{P}(\mathcal{T}_{\mathrm{mart}}^c) \leq 2p\exp\Big(-\frac{c\lambda^2Nn^2}{n\Delta_n d\log(nN)^2}\Big).$$

As $n\Delta_n = T = 1$ in our case, for a fixed value of $\varepsilon \in (0,1/4)$ we define

$$\lambda_1 = C_1\sqrt{\frac{n\Delta_n d\log(nN)^2}{n^2N}[\log(2p)+\log(1/\varepsilon)]} = C_1\frac{\log(nN)}{n}\sqrt{\frac{d}{N}}\sqrt{\log(2p)+\log(1/\varepsilon)},$$

and for all $\lambda \geq \lambda_1$

$$\mathbb{P}(\mathcal{T}_{\mathrm{mart}}^c|\Omega_{n,N}) \leq \varepsilon.$$

□

B.3.3 The Discretisation Set

Theorem B.4. *Under Assumptions 1 - 3, for a fixed value of $\varepsilon \in (0,1/4)$ and for all $\lambda > \lambda_2$, it holds that*

$$\mathbb{P}(\mathcal{T}_{\mathrm{disc}}^c|\Omega_{n,N}) \leq \varepsilon.$$

Proof. From now on, we are working on $\Omega_{n,N}$ (defined by (6)). We first rewrite

$$\begin{aligned}\mathbb{P}(\mathcal{T}_{\mathrm{disc}}^c) &= \mathbb{P}\left(\left\|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Phi(X_{t_{i-1}}^{(j)})^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\left(\sum_{l=1}^{p}\theta_{0,l}\int_{t_{i-1}}^{t_i}\big(\phi_l(X_s^{(j)})-\phi_l(X_{t_{i-1}}^{(j)})\big)\mathrm{d}s\right)\right\|_\infty > \frac{\lambda}{4}\right)\\ &= \mathbb{P}\left(\bigcup_{k=1}^{p}\left\{\left|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}\Phi(X_{t_{i-1}}^{(j)})^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\left(\sum_{l=1}^{p}\theta_{0,l}\int_{t_{i-1}}^{t_i}\big(\phi_l(X_s^{(j)})-\phi_l(X_{t_{i-1}}^{(j)})\big)\mathrm{d}s\right)\right|_k > \frac{\lambda}{4}\right\}\right)\\ &\leq p\max_k\mathbb{P}\left(\left|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}F_k^{i,j}\right| \geq \frac{\lambda}{4}\right),\end{aligned}$$

where

$$F_k^{i,j} := \phi_k(X_{t_{i-1}}^{(j)})^T(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\left(\sum_{l=1}^{p}\theta_{0,l}\int_{t_{i-1}}^{t_i}\big(\phi_l(X_u^{(j)})-\phi_l(X_{t_{i-1}}^{(j)})\big)\mathrm{d}u\right).$$

We will upper bound this probability using Markov's inequality, therefore we aim to upper bound $\mathbb{E}(|F_k^{i,j}|^r)$ (where $r>1$ will be specified later). We have using Cauchy-Schwarz inequality, the linear growth of ϕ_k from Assumption 1, and

Assumption 3

$$\begin{aligned}
\mathbb{E}(|F_k^{i,j}|^r) &\le \mathbb{E}\left[\left\|\phi_k(X_{t_{i-1}}^{(j)})\right\|_2^{2r}\right]^{1/2} \mathbb{E}\left[\left\|(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\Big(\sum_{l=1}^{p}\theta_{0,l}\int_{t_{i-1}}^{t_i}\big(\phi_l(X_u^{(j)})-\phi_l(X_{t_{i-1}}^{(j)})\big)\mathrm{d}u\Big)\right\|_2^{2r}\right]^{1/2}\\
&\le 2^{2r-1}\left[L_k^{2r}\mathbb{E}\left[\left\|X_{t_{i-1}}^{(j)}\right\|_2^{2r}\right]+\|\phi_k(0)\|_2^{2r}\right]^{1/2}\\
&\quad\times\mathbb{E}\left[\left\|(\sigma\sigma^T)^{-1}(X_{t_{i-1}}^{(j)})\right\|_2^{2r}\left\|\sum_{l=1}^{p}\theta_{0,l}\int_{t_{i-1}}^{t_i}\big(\phi_l(X_u^{(j)})-\phi_l(X_{t_{i-1}}^{(j)})\big)\mathrm{d}u\right\|_2^{2r}\right]^{1/2}\\
&\le C^r L_k^r\, d^{r/2}\log(nN)^r C_\sigma^r\, s^r \max_l|\theta_{0,l}|^r\,\mathbb{E}\left[\left\|\int_{t_{i-1}}^{t_i}\big(\phi_l(X_u^{(j)})-\phi_l(X_{t_{i-1}}^{(j)})\big)\mathrm{d}u\right\|_2^{2r}\right]^{1/2}\\
&\le C^r L_k^r C_\sigma^r\, d^{r/2}\log(nN)^r s^r\Delta_n^{(2r-1)/2}\max_l|\theta_{0,l}L_l|^r\left(\int_{t_{i-1}}^{t_i}\mathbb{E}\left[\left\|X_u^{(j)}-X_{t_{i-1}}^{(j)}\right\|_2^{2r}\right]\mathrm{d}u\right)^{1/2}. \qquad (22)
\end{aligned}$$

We bound the increments of the process using Lemma B.1

$$\begin{aligned}
\sup_{u\in[t_{i-1},t_i]}\mathbb{E}\left[\left\|X_u^{(j)}-X_{t_{i-1}}^{(j)}\right\|_2^{2r}\right] &\le C^r\Big(\Delta_n^{2r}s^{2r}(\mathbb{E}\left[\left\|X_{t_{i-1}}^{(j)}\right\|_2^{2r}\right]+\|\phi_k(0)\|_2^{2r})+\Delta_n^r r^r d^r\Big)\\
&\le C^r\Delta_n^r\Big(\Delta_n^r s^{2r}\log(nN)^{2r}+r^r d^r\Big)\le C^r\Delta_n^r\log(nN)^{2r}r^r d^r. \qquad (23)
\end{aligned}$$

We get by combining (22) and (23) that

$$\mathbb{E}(|F_k^{i,j}|^r)\le C^r\log(nN)^{2r}\,d^r s^r\Delta_n^{3r/2}r^{r/2},$$

therefore using Hölder's inequality

$$\mathbb{E}\Big(\Big|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}F_k^{i,j}\Big|^r\Big)\le(C\log(nN)^2 ds\Delta_n^{3/2}\sqrt{r})^r.$$

Using Markov's inequality, we deduce that

$$\mathbb{P}\Big(\Big|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}F_k^{i,j}\Big|\ge C\log(nN)^2 ds\Delta_n^{3/2}\sqrt{r}e\Big)\le\exp(-r),$$

which after a good choice of r can be written in the form

$$\mathbb{P}\left(\Big|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}F_k^{i,j}\Big|\ge u\right)\le\exp\left(-\Big(\frac{u}{eC\log(nN)^2 ds\Delta_n^{3/2}}\Big)^2\right).$$

We conclude that

$$\begin{aligned}
\mathbb{P}(\mathcal{T}_{\mathrm{disc}}^c) &\le p\max_k\mathbb{P}\left(\Big|\frac{1}{nN}\sum_{i=1}^{n}\sum_{j=1}^{N}F_k^{i,j}\Big|\ge\frac{\lambda}{4}\right)\\
&\le p\exp\Big(-\Big(\frac{\lambda/4}{eC\log(nN)^2 ds\Delta_n^{3/2}}\Big)^2\Big)\le p\exp\Big(-c\frac{\lambda^2}{\log(nN)^4 d^2 s^2\Delta_n^3}\Big).
\end{aligned}$$

Therefore, for a fixed value of $\varepsilon\in(0,1/4)$ we define

$$\lambda_2=C_2\log(nN)^2 sd\Delta_n^{3/2}\sqrt{\log(p)+\log(1/\varepsilon)}=C_2\frac{sd\log(nN)^2}{n^{3/2}}\sqrt{\log(p)+\log(1/\varepsilon)},$$

and for all $\lambda\ge\lambda_2$

$$\mathbb{P}(\mathcal{T}_{\mathrm{disc}}^c|\Omega_{n,N})\le\varepsilon.$$

□

B.3.4 Compatibility Condition

Theorem B.5. *Under Assumptions 1 - 3 and Assumption 5, for a fixed value of $\varepsilon \in (0, 1/4)$ and for all $N > N_1$, it holds that*

$$\mathbb{P}(\mathcal{T}_{\text{comp}}^c | \Omega_{n,N}) \leq \varepsilon.$$

Proof. From now on, we are working on $\Omega_{n,N}$ (defined by (6)). For $u \in \mathbb{R}^p$, we define

$$H_u = u^T \left(\frac{1}{N} \sum_{j=1}^{N} \frac{1}{n} \sum_{i=1}^{n} \left[(\sigma^{-1}\Phi)^T (\sigma^{-1}\Phi)(X_{t_{i-1}}^{(j)}) - \mathbb{E}\big[(\sigma^{-1}\Phi)^T (\sigma^{-1}\Phi)(X_{t_{i-1}}^{(j)}) \big] \right] \right) u.$$

Following the proof of Theorem 6.8. in Amorino et al. [2025] and using Assumption 5, we have

$$\begin{aligned}
\mathbb{P}\left(\inf_{\substack{\theta \in \mathbb{R}^p : \|\theta\|_0 = s, \\ v \in \mathbb{R}^p : \theta - v \in \mathcal{C}(s, 3+4/\gamma)}} \frac{\|b_\theta - b_v\|_{\sigma,n,N}}{\|\theta - v\|_2} \geq k \right) &\geq 1 - \mathbb{P}\left(\sup_{u \in \mathcal{C}(s,3+4/\gamma)} \frac{H_u}{\|u\|_2^2} \geq \ell - k^2 \right) \\
&\geq 1 - 21^{2s} \Big(p^{2s} \wedge (ep/2s)^{2s} \Big) \sup_{u \in \Theta} \mathbb{P}\left(\frac{H_u}{\|u\|_2^2} \geq \frac{\ell - k^2}{9(5 + \frac{4}{\gamma})^2} \right). \qquad (24)
\end{aligned}$$

The random function $\frac{H_u}{\|u\|_2^2}$ can be written as a sum of N independent random variables

$$\frac{H_u}{\|u\|_2^2} = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{\frac{1}{n} \sum_{i=1}^{n} u^T \left[(\sigma^{-1}\Phi)^T (\sigma^{-1}\Phi)(X_{t_{i-1}}^{(j)}) - \mathbb{E}\big[(\sigma^{-1}\Phi)^T (\sigma^{-1}\Phi)(X_{t_{i-1}}^{(j)}) \big] \right] u}{\|u\|_2^2} \right).$$

We recall that from Cauchy–Schwarz inequality $\forall M \in \mathbb{R}^{p \times p}$, $\forall u \in \mathbb{R}^p$, $u^T M u \leq \|u\|_2^2 \sqrt{\sum_k \sum_l M_{k,l}^2}$. From Assumption 1, the functions ϕ_k are Lipschitz. As we are working on $\Omega_{n,N}$ given in Equation (6), we have

$$\begin{aligned}
u^T (\sigma^{-1}\Phi)^T (\sigma^{-1}\Phi)(X_{t_{i-1}}^{(j)}) u &\leq \|u\|_2^2 \, \|\sigma^{-1}\Phi(X_{t_{i-1}}^{(j)})\|_2^2 \leq C_\sigma^2 \|u\|_2^2 \, \|\Phi(X_{t_{i-1}}^{(j)})\|_2^2 \\
&\leq C_\sigma^2 \|u\|_2^2 \sum_{k=1}^{p} \|\phi_k(X_{t_{i-1}}^{(j)})\|_2^2 \leq C \|u\|_2^2 pd \log(nN)^2.
\end{aligned}$$

Hence $\frac{H_u}{\|u\|_2^2}$ is a sum of independent, centered and bounded random variables, and we apply Hoeffding's inequality to obtain

$$\mathbb{P}\left(\frac{H_u}{\|u\|_2^2} \geq z \right) \leq 2 \exp\left(- \frac{c(zN)^2}{N(dp \log(nN)^2)^2} \right) \leq 2 \exp\left(- \frac{cz^2 N}{d^2 p^2 \log(nN)^4} \right).$$

Combining this upper bound with (24), we have proved that

$$\mathbb{P}(\mathcal{T}_{\text{comp}}) \geq 1 - 21^{2s} \Big(p^{2s} \wedge (ep/2s)^{2s} \Big) \, 2 \exp\Big(- \frac{cN}{d^2 p^2 \log(nN)^4} \frac{(\ell - k^2)^2}{(5 + 4/\gamma)^4} \Big).$$

Therefore, for a fixed value of $\varepsilon \in (0, 1/4)$ we define

$$\begin{aligned}
N_1 &= c_1 \frac{(5 + 4/\gamma)^4}{(\ell - k^2)^2} d^2 p^2 \log(nN)^4 \left[\log(2/\varepsilon) + \log\left(21^{2s} [p^{2s} \wedge (ep/2s)^{2s}] \right) \right] \\
&= c_1 d^2 p^2 \log(nN)^4 \frac{(5 + 4/\gamma)^4}{(\ell - k^2)^2} \left[\log(2/\varepsilon) + \log\left(21^{2s} [p^{2s} \wedge (ep/2s)^{2s}] \right) \right],
\end{aligned}$$

and for all $N \geq N_1$

$$\mathbb{P}(\mathcal{T}_{\text{comp}}^c | \Omega_{n,N}) \leq \varepsilon.$$

□