Available online at www.sciencedirect.com

ScienceDirect

Journal of Approximation Theory 259 (2020) 105472

Journal of Approximation Theory

www.elsevier.com/locate/jat

Full Length Article

Recovery guarantees for polynomial coefficients from weakly dependent data with outliers

Lam Si Tung Ho[a,*], Hayden Schaeffer[b], Giang Tran[c], Rachel Ward[d]

[a] *Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada*
[b] *Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA, USA*
[c] *Department of Applied Mathematics, University of Waterloo, Waterloo, ON, Canada*
[d] *Department of Mathematics, University of Texas at Austin, Austin, TX, USA*

Received 17 December 2018; received in revised form 14 July 2020; accepted 10 August 2020
Available online 20 August 2020

Communicated by Simon Foucart

Abstract

Learning non-linear systems from noisy, limited, and/or dependent data is an important task across various scientific fields including statistics, engineering, computer science, mathematics, and many more. In general, this learning task is ill-posed; however, additional information about the data's structure or on the behavior of the unknown function can make the task well-posed. In this work, we study the problem of learning nonlinear functions from corrupted and weakly dependent data. The learning problem is recast as a sparse robust linear regression problem where we incorporate both the unknown coefficients and the corruptions in a basis pursuit framework. The main contribution of our paper is to provide a reconstruction guarantee for the associated ℓ_1-optimization problem where the sampling matrix is formed from weakly dependent data. Specifically, we prove that the sampling matrix satisfies the null space property and the stable null space property, provided that the data is compact and satisfies a suitable concentration inequality. We show that our recovery results are applicable to various types of weakly dependent data such as exponentially strongly α-mixing data, geometrically $\mathcal{C}$-mixing data, and uniformly ergodic Markov chain. Our theoretical results are verified via several numerical simulations.

Keywords: Recovery guarantee; Basis pursuit; Sparsity; Concentration inequalities; Function approximation; Weakly dependent data

* Corresponding author.
E-mail addresses: Lam.Ho@dal.ca (L.S.T. Ho), schaeffer@cmu.edu (H. Schaeffer), giang.tran@uwaterloo.ca (G. Tran), rward@math.utexas.edu (R. Ward).

https://doi.org/10.1016/j.jat.2020.105472

1. Introduction

In the past few decades, there has been a rapid growth of interest in automated learning from data across various scientific fields including statistics [52], engineering [30], computer science [23,39], mathematics, and many more. An overview of machine learning problems in a wide range of contexts (statistical learning theory, pattern recognition, system identification, deep learning, and so on) can be found in [1,6,19,20]. One of the main paradigms is to learn an unknown target function from a given collection of input–output pairs (supervised learning), which can be rephrased as the problem of finding an approximation of a multi-dimensional function. For example, in [34,35], the authors demonstrated a connection between approximation theory and regularization with feedforward multilayer networks. In general, learning a smooth function from data is ill-posed unless a priori information about either the data structure or the generating function is provided [15,33,50].

One of the well-known methods to make the learning problem well-posed is to exploit additional properties of the target function [18]. For example, if the target function depends only on a few active coordinates associated with a suitable random matrix, the function can be recovered from a small number of samples [15]. On the other hand, many well-known learning methods consider the target function in a particular function class (such as radial basis functions, multivariate polynomials, projection pursuit, feed-forward neural networks, and tensor product methods) and add a penalty (such as Tikhonov regularization or sparse constraints) to the associated parameter estimation problem. For example, an adaptive high-dimensional polynomial interpolation technique is presented in [7] and an optimal least square method is proposed in [9].

Recently, sparse models combined with data-driven methods have been investigated intensively for learning nonlinear partial differential equations, nonlinear dynamical systems, and graph-based networks. The model selection problem for dynamical systems from time series dates back to [11] where the authors investigate the concepts from dynamical system theory to recover the underlying structure from data. In [55], the authors construct a sampling matrix from the data matrix and its power to recover the ordinary differential equations and find an optimal Kronecker product representation for the governing equations. Furthermore, based on the observation that many governing equations have a sparse representation with respect to high-dimensional polynomial spaces, the authors in [4] developed the SINDy algorithm which uses that sampling matrix and a sequential least-square thresholding algorithm to recover the governing equations of some unknown dynamical systems. The convergence of the SINDy algorithm is provided in [57]. A group-sparse model was proposed in [44] to learn governing equations from a family of dynamical systems with bifurcation parameters. By exploiting the cyclic structure of many nonlinear differential equations, the authors in [46] proposed an approach to identify the active basis terms using fewer random samples (in some cases on the order of a few snapshots). For the noisy case, in [43] the authors use the integral formulation of the differential equation to reduce the effect of noise and identify the model from a smoother basis set. To learn a nonlinear partial differential equation from spatio-temporal dataset, the authors in [42] proposed a LASSO-based approach using a dictionary of partial derivatives. In [40], the authors developed an adaptive ridge-regression version of [4] for learning nonlinear PDE, while in [36] a hidden physics model based on Gaussian processes was presented. On the other hand, the data are often contaminated by noise, contain outliers, have missing values, or have a limited amount of samples. When the given data are limited, there are several works addressing learning problems ranging from sampling strategies in high-dimensional dynamics

using random initial conditions [45], to a weighted ℓ_1-minimization on the lower set [8,37], model predictive control using SINDy [24], and sample complexity reduction to linear time-invariant systems [14]. In [47], the authors proposed a method to approximate an unknown function from noise measurements via sequential approximation. Geometric methods, such as [28], can be used to approximate functions in high-dimensions when the data concentrate on lower-dimensional sets.

Regarding supervised learning analysis, the input data are assumed to be independent and identically distributed (i.i.d.). However, this assumption does not hold in many applications such as speech recognition, medical diagnosis, signal processing, computational biology, and financial prediction. Alternatively, for non-i.i.d. processes satisfying certain mixing conditions, various reconstruction results have been addressed in different contexts. The convergence rates of several machine learning algorithms have been studied for non-i.i.d. data. Examples include weighted average algorithm [12], least squares support vector machines (LS-SVMs) [21], and one-vs-all multiclass plug-in classifiers [13]. In [53], the authors discussed several mixing conditions for weakly dependent observations which guarantee the consistency and asymptotic normality for the nonlinear least squares estimator. Minimum complexity regression estimators for m-dependent observations and strongly mixing observations were proposed in [32] using certain Bernstein-type inequalities for weakly dependent observations. In [41], a conditionally i.i.d. model for pattern recognition was proposed, where the inputs are conditionally independent given the output labels. In [49], the authors proved that if the data-generating process satisfies a certain law of large number, the support vector machines are consistent. In [22], a Bernstein-type inequality for geometrically $\mathcal{C}$-mixing processes is established and applied to deduce an oracle inequality for generic regularized empirical risk minimization algorithms. Using a strong central limit theorem for chaotic data and compressed sensing results, the authors in [51] proved a reconstruction guarantee for sparse reconstruction of governing equations for three-dimensional chaotic systems with outliers. The common technique in the mentioned works is the application of either a central limit theorem or a suitable concentration inequality for the given data.

In this work, we study the problem of learning nonlinear functions from identically distributed (but not necessarily independent) data that are corrupted by outliers and/or contaminated by noise. By expressing the target function in the multivariate polynomial space, the learning problem is recast as a sparse robust linear regression problem where we incorporate both the unknown coefficients and the corruptions in a basis pursuit framework. The main contribution of our paper is to provide a reconstruction guarantee for the associated ℓ_1-optimization problem where the (augmented) sampling matrix is formed from the data matrix, its powers, and the identity matrix. Although the data may not be i.i.d., we prove that the sampling matrix satisfies the null space property, provided that the data are compact and satisfies a suitable concentration inequality. Consequently, the basis pursuit problem will be guaranteed to have a unique solution and be stable with respect to noise. Numerically, we use the well-known Douglas–Rachford algorithm to solve the corresponding optimization problem. In general, the algorithm using monomial bases may be numerically unstable when the degree of the polynomial is large. However, our simulations indicate that the proposed method works well in various situations.

The paper is organized as follows: In Section 2, we explain the problem setting. In Section 3, we first recall the theory from compressive sensing, then present the theoretical reconstruction guarantees. In Section 4, we state the recovery results for various types of data including i.i.d. data, exponentially strongly α-mixing data, geometrically $\mathcal{C}$-mixing data, and uniformly

ergodic Markov chain. The numerical implementations and results are described in Section 5. We discuss the conclusion and future works in Section 6.

2. Problem statement

We would like to learn a function $f : \mathbb{R}^d \to \mathbb{R}$, from m data points $\left(\mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\theta}^{(i)}, y^{(i)} = f(\mathbf{x}^{(i)}) + \varepsilon^{(i)}\right)_{i=1}^{m}$, where $\{\mathbf{u}^{(i)}\}$ is corrupted data, $\{\mathbf{x}^{(i)}\}$ is the uncorrupted part, $\{\boldsymbol{\theta}^{(i)}\}$ represents the corruption, and $\{\varepsilon^{(i)}\}$ denotes noise. We say that $\mathbf{u}^{(i)}$ is an outlier if the corruption $\boldsymbol{\theta}^{(i)}$ is non-zero. Assume that the function of interest f is a multivariate polynomial of degree at most p:

$$f(x_1, \ldots, x_d) = \sum_{|\alpha| = \alpha_1 + \cdots + \alpha_d \le p} c^{\alpha} x_1^{\alpha_1} \ldots x_d^{\alpha_d}.$$

Let $\mathbf{y} = [y^{(1)}, \ldots, y^{(m)}]^T$, $\boldsymbol{\varepsilon} = [\varepsilon^{(1)}, \ldots, \varepsilon^{(m)}]^T$, $\boldsymbol{\theta}$ be the matrix where the rows are $\boldsymbol{\theta}^{(i)}$, and U be the data matrix,

$$U = \begin{bmatrix} - & \mathbf{u}^{(1)} & - \\ - & \mathbf{u}^{(2)} & - \\ & \cdots & \\ - & \mathbf{u}^{(m)} & - \end{bmatrix} = \begin{bmatrix} u_1^{(1)} & \ldots & u_d^{(1)} \\ u_1^{(2)} & \ldots & u_d^{(2)} \\ \vdots & \vdots & \vdots \\ u_1^{(m)} & \ldots & u_d^{(m)} \end{bmatrix} = \begin{bmatrix} | & & | \\ U_1 & \cdots & U_d \\ | & & | \end{bmatrix}_{m \times d}.$$

Then we form the dictionary matrix $\Phi = \Phi_U$ from data,

$$\Phi_U = \begin{bmatrix} | & | & | & & | & | & | & & | & \\ 1 & U_1 & U_2 & \ldots & U_d & U_1^2 & U_1 U_2 & \cdots & U_d^2 & \cdots \\ | & | & | & & | & | & | & & | & \end{bmatrix}_{m \times r}, \tag{2.1}$$

where $r = \binom{p+d}{d}$ is the maximal number of d-multivariate monomials of degree at most p.

Denote $\mathbf{c} = (c^{\alpha})_{|\alpha| \le p}$ the coefficient vector and $\mathbf{e} = \mathbf{y} - \Phi\mathbf{c}$, we can reformulate our problem as follows:

Find $(\mathbf{c}, \mathbf{e}) \in \mathbb{R}^r \times \mathbb{R}^m$ such that $\mathbf{y} = \Phi\mathbf{c} + \mathbf{e}$

Without corruptions and with arbitrary noise vector $\boldsymbol{\varepsilon}$, the problem is classically solvable by least squares regression once $m \ge r$. With corruptions, whose locations can be arbitrary but are unknown beforehand, if $m \ge r$ and at least $n = r$ of the m measurements are uncorrupted, then one could in theory do a regression on each of the $\binom{m}{n}$ subsets of n measurements and retain the set with the smallest error; however, this is an infeasible combinatorial algorithm. Thus, the convex relaxation of this combinatorial algorithm is a natural choice for reconstruction algorithm:

$$\min_{\mathbf{c}', \mathbf{e}'} \|\mathbf{e}'\|_1 \quad \text{subject to} \quad \mathbf{y} = \Phi\,\mathbf{c}' + \mathbf{e}'. \tag{2.2}$$

On the other hand, if the polynomial coefficients are sparse or the polynomial function can be approximated by a sparse polynomial, the learning problem can be recast as follows:

$$\min_{\mathbf{c}', \mathbf{e}'} \|\mathbf{c}'\|_1 + \|\mathbf{e}'\|_1 \quad \text{subject to} \quad \mathbf{y} = \Phi\,\mathbf{c}' + \mathbf{e}', \tag{2.3}$$

or, more generally, as the corrupted sensing problem [17,25,27],

$$\min_{\mathbf{c}', \mathbf{e}'} \|\mathbf{c}'\|_1 + \lambda\|\mathbf{e}'\|_1 \quad \text{subject to} \quad \mathbf{y} = \Phi\,\mathbf{c}' + \mathbf{e}'. \tag{2.4}$$

Table 1
Probability of exact recovery versus the maximum degree of polynomial in the unknown function (see Eq. (5.6)). The ambient dimension is $d = 3$, the maximum degree of the candidate polynomials is $p = 10$ for all runs (thus $r = 286$ monomial terms). The sparsity of the polynomial coefficients is equal to 2 and the sparsity of the vector e is 5. We measure the rate of recovery over 100 trials for various p.

Degree p:	2	3	5	8	10
Recovery, 15% sampling:	83	95	85	82	79
Recovery, 35% sampling:	98	98	99	99	99

Table 2
Probability of exact recovery versus the sparsity of the vector e (see Eq. (5.7)). The ambient dimension is $d = 10$, the maximum degree of the candidate polynomials is $p = 3$ (thus $r = 286$ monomial terms), and the sparsity of the polynomial coefficients is 5. The non-zero values in the vector e have uniformly random values in $[-H, H]$. In each test, the sampling rate is 50% and the parameter is set to $\lambda = 2$. Note that tuning λ may change the recovery rate.

Sparsity of the vector e	3	10	15
Recovery, $H = 0.5$	91	60	23
Recovery, $H = 2$	98	65	28
Recovery, $H = 10$	100	72	34

For the remainder of the paper, we denote the sparsity level of $\mathbf{c}$ by s_c, and the row-sparsity level of $\boldsymbol{\theta}$ by s_θ. In the noiseless case ($\boldsymbol{\varepsilon} = 0$), we have:

$$\|\mathbf{e}\|_0 = \#\{i : \boldsymbol{\theta}^{(i)} \neq 0\} \leq s_\theta.$$

Remark 2.1. In this paper, the distribution of input data is unknown. Therefore, it is difficult to construct an explicit and accurate formula for orthogonal polynomials with respect to this unknown distribution. Theoretically, we have shown that using the standard monomial basis leads to recovery guarantees under suitable conditions (see Theorem 3.4 for more details). In general, the algorithm using monomial bases may be numerically unstable when the degree of the polynomial is large. Numerically, we have demonstrated that the algorithm can recover polynomial coefficients well with the monomial basis of different degrees $p = 2, 3, 5, 8, 10$ (see Tables 1 and 2, for example).

3. Reconstruction guarantee analysis

Before presenting the properties of the matrix $[Id_m, \Phi]$ and theoretical guarantees for the corresponding ℓ_1-optimization problems, we first recall some results from compressive sensing including the null space property and the stable null space property (see [16] for a comprehensive overview).

3.1. Theory from compressive sensing

Definition 3.1. A matrix $A \in \mathbb{R}^{m \times N}$ is said to satisfy

- the null space property of order s if

$$\|v_S\|_1 < \frac{1}{2}\|v\|_1 \quad \text{for all} \quad v \in \ker A \backslash \{0\},$$

for any set $S \subset [N] := \{1, 2, \ldots, N\}$ with $card(S) \leq s$.

- the stable null space property of order s with constant $0 < \rho < 1$ if

$$\|v_S\|_1 \leq \frac{\rho}{\rho+1}\|v\|_1 \quad \text{for all} \quad v \in \ker A,$$

for any set $S \subset [N]$ with $card(S) \leq s$.

Proposition 3.2 (*Recovery Guarantee Given Null Space Property*)**.** *Given a matrix $A \in \mathbb{R}^{m\times N}$, every s-sparse vector $z^* \in \mathbb{R}^N$ with $y = Az^*$ is the unique solution of*

$$\min \|z\|_1, \quad \textit{subject to} \quad y = Az, \tag{3.1}$$

if and only if A satisfies the null space property of order s.

Proposition 3.3 (*Recovery Guarantee Given Stable Null Space Property*)**.** *Suppose a matrix $A \in \mathbb{R}^{m\times N}$ satisfies the stable null space property of order s with constant $0 < \rho < 1$. Then, for any $x \in \mathbb{R}^N$ with $y = Ax$, a solution $z^\#$ of the optimization problem (3.1) approximates the vector x with ℓ_1-error*

$$\|x - z^\#\|_1 \leq \frac{2(1+\rho)}{1-\rho} \inf_{\|w\|_0 \leq s} \|x - w\|_1.$$

The null space property for the matrix A, along with the existence of an s-sparse solution to the underdetermined system of equations, is a sufficient and necessary condition for sparse solutions of the NP hard minimization problem,

$$\min \|z\|_0, \quad \text{subject to} \quad y = Az,$$

to be exactly recovered via the ℓ_1-minimization (3.1). On the other hand, the stable null space property of the matrix A guarantees that any solution, sparse or not, can be recovered up to the error governed by its distance to s-sparse vectors.

3.2. Theoretical guarantees

We will show that if the uncorrupted data $\{\mathbf{x}^{(i)}\}$ satisfy an appropriate concentration inequality and their common distribution μ is *non-degenerate* (that is, if for any μ-measurable set B, $\mu(B) = 1$ implies B contains infinitely many elements), then the polynomial coefficients of the unknown function as well as the location of the outliers can be exactly recovered with high probability from the unique solution of the ℓ_1-minimization problem (2.3), provided that the output values y are exact. When the output values y contain dense noise, we show that every solution of the associated optimization problem can be approximated by a sparse solution under suitable assumptions.

To begin with, we will show that the matrix $[Id_m, \Phi_{m\times r}]$, where $\Phi_{m\times r} = \Phi_U$ is constructed from all monomials up to degree p, satisfies the null space property.

Theorem 3.4. *Fix $p \in \mathbb{N}$. Consider $\{\mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\theta}^{(i)}\}_{i=1}^m \subset \mathbb{R}^d$ where the uncorrupted data $\{\mathbf{x}^{(i)}\}$ are L^∞-bounded by a constant $B_\mathcal{X}$ and identically distributed according to a non-degenerate probability distribution μ, and the corruption $\{\boldsymbol{\theta}^{(i)}\}$ is L^∞-bounded by a constant B_Θ and s_θ-row sparse. Assume that $\{\mathbf{x}^{(i)}\}$ satisfies the following concentration inequality:*

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^m \varphi(\mathbf{x}^{(i)}) - \mathbb{E}[\varphi(\mathbf{x})]\right| \geq \zeta\right) \leq e^{-\kappa(\zeta,m)}, \tag{3.2}$$

for any $\zeta > 0$, any bounded Borel function φ, and some function $\kappa(\zeta, m)$. In addition, suppose there exists a constant $M_{\kappa,\delta,r}$ depending on κ, δ, r such that when $m > M_{\kappa,\delta,r}$, we have:

$$\kappa(m^{-\delta}, m) \geq 3\delta r \log m, \tag{3.3}$$

where $e = \exp(1)$, $r = \binom{p+d}{d}$, and $\delta > 0$ is some chosen constant.

Then, when $m > M_{\kappa,\delta,r}$ and m satisfies

$$\begin{aligned} m &\geq \left(\max\{3 + 3B_{\mathcal{X}}^p, 4D^{-1}\}\right)^{1/\delta}, \\ m &> \frac{4 + 8s\,(1 + (B_{\mathcal{X}} + B_\Theta)^p)}{D}, \end{aligned} \tag{3.4}$$

the matrix $A = [Id_m, \Phi_{m\times r}]$, where $\Phi = \Phi_U$ is the dictionary matrix (2.1), satisfies the null space property of order $s \geq s_\theta$ with probability at least $(1 - m^{-\delta r})$. Here $D > 0$ is a constant depending only on p, d, and μ.

Proof. For each $c \in \mathcal{B} = \{v \in \mathbb{R}^r : \|v\|_1 = 1\}$, define $\varphi^c : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$\varphi^c(\mathbf{x}) = \left| \sum_{\alpha:|\alpha|\leq p} c^\alpha x_1^{\alpha_1} \dots x_d^{\alpha_d} \right|.$$

We first evaluate the lower bound for the summation $\sum_{i=1}^m \varphi^c(\mathbf{x}^{(i)})$. For any non-zero $c \in \mathbb{R}^r$, we have $\mathbb{E}[\varphi^c(\mathbf{x})] > 0$. Indeed, if $\mathbb{E}[\varphi^c(\mathbf{x})] = 0$, then $\varphi^c(\mathbf{x}) = 0$ μ-almost surely. Since μ is non-degenerate, there are infinitely many $\mathbf{x}$ such that $\varphi^c(\mathbf{x}) = 0$. This implies $c = 0$ which is a contradiction. Therefore, $\mathbb{E}[\varphi^c(\mathbf{x})] > 0$ for any $c \in \mathcal{B}$.

On the other hand, since the set $\mathcal{B}$ is compact and nonempty, we can apply the extreme value theorem for the continuous function $\mathbb{E}[\varphi^c(\mathbf{x})]$ to get the following bound:

$$\inf_{c\in\mathcal{B}} \mathbb{E}[\varphi^c(\mathbf{x})] \geq D > 0,$$

for some constant $D > 0$. Note that D depends on μ, d, and p.

According to a well-known result on the covering number (for example, see Appendix C.2, [16]), there exists a finite set of points $\mathcal{Q}$ in $\mathcal{B}$ of cardinality

$$|\mathcal{Q}| \leq \left(3\, m^\delta (1 + B_{\mathcal{X}}^p)\right)^r$$

such that

$$\max_{c\in\mathcal{B}} \min_{q\in\mathcal{Q}} \|c - q\|_1 \leq \frac{1}{m^\delta (1 + B_{\mathcal{X}}^p)}.$$

Applying the union bound on $\mathcal{Q}$ and using the assumption $\kappa(m^{-\delta}, m) \geq 3\delta r \log m$, we derive:

$$\Pr\left(\bigcup_{q\in\mathcal{Q}} \left\{ \left| \frac{1}{m} \sum_{i=1}^m \varphi^q(\mathbf{x}^{(i)}) - \mathbb{E}[\varphi^q(\mathbf{x})] \right| \geq m^{-\delta} \right\}\right) \leq m^{\delta r} (3 + 3B_{\mathcal{X}}^p)^r e^{-\kappa(m^{-\delta}, m)} \leq m^{-\delta r},$$

provided that

$$m \geq \left(3 + 3B_{\mathcal{X}}^p\right)^{1/\delta}. \tag{3.5}$$

Hence,

$$\Pr\left(\left\{ \max_{q\in\mathcal{Q}} \left| \frac{1}{m} \sum_{i=1}^m \varphi^q(\mathbf{x}^{(i)}) - \mathbb{E}[\varphi^q(\mathbf{x})] \right| \leq m^{-\delta} \right\}\right) \geq 1 - m^{-\delta r}.$$

Therefore, for any $q \in \mathcal{Q}$, we have:

$$\sum_{i=1}^{m} \varphi^q\left(\mathbf{x}^{(i)}\right) \geq m\mathbb{E}(\varphi^q(x)) - \left|\sum_{i=1}^{m} \varphi^q(\mathbf{x}^{(i)}) - m\mathbb{E}[\varphi^q(\mathbf{x})]\right| \geq mD - m^{1-\delta}, \tag{3.6}$$

with probability at least $(1 - m^{-\delta r})$.

For each $c \in \mathcal{B}$ there exists $q_c \in \mathcal{Q}$ so that $\|c - q_c\|_1 \leq \dfrac{1}{m^{\delta}(1 + B_{\mathcal{X}}^p)}$. Applying the Hölder's inequality for $\mathbf{x} = (x_j) \in \mathbb{R}^d$ with $\|\mathbf{x}\|_\infty \leq B_{\mathcal{X}}$, we obtain:

$$\left|\sum_{\alpha:|\alpha|\leq p} (c^\alpha - q_c^\alpha) \prod_{j=1}^{d} x_j^{\alpha_j}\right| \leq \|c - q_c\|_1 \max_{\alpha:|\alpha|\leq p} \prod_{j=1}^{d} \left|x_j^{\alpha_j}\right| \leq m^{-\delta}.$$

Combining with the inequality (3.6), we obtain

$$\begin{aligned}\sum_{i=1}^{m} \varphi^c\left(\mathbf{x}^{(i)}\right) &\geq \sum_{i=1}^{m} \varphi^{q_c}\left(\mathbf{x}^{(i)}\right) - \sum_{i=1}^{m} \left|\sum_{\alpha:|\alpha|\leq p} (c^\alpha - q_c^\alpha) \prod_{j=1}^{d} \left(x_j^{(i)}\right)^{\alpha_j}\right| \\ &\geq m(D - 2m^{-\delta}) \\ &\geq \frac{1}{2} mD,\end{aligned}$$

with probability at least $(1 - m^{-\delta r})$, provided that

$$m \geq \left(\frac{4}{D}\right)^{1/\delta}. \tag{3.7}$$

By linearity, we have in the same event,

$$\sum_{i=1}^{m} \varphi^c(\mathbf{x}^{(i)}) \geq \frac{1}{2} mD\, \|c\|_1, \quad \forall c \in \mathbb{R}^r \setminus \{0\}. \tag{3.8}$$

Next, we will estimate the lower bound for $\|\Phi c\|_1$, where $c \in \mathbb{R}^r \setminus \{0\}$. Denote $R = (R_1, \ldots, R_m)^T \in \mathbb{R}^m$, where R_i is defined as follows:

$$R_i = (\Phi c)_i - \sum_{\alpha:|\alpha|\leq p} c^\alpha \left(x_1^{(i)}\right)^{\alpha_1} \ldots \left(x_d^{(i)}\right)^{\alpha_d}, \quad i = 1, \ldots, m.$$

Applying the Hölder's inequality, we have

$$\begin{aligned}|(\Phi c)_i| &= \left|\sum_{\alpha:|\alpha|\leq p} c^\alpha \left(u_1^{(i)}\right)^{\alpha_1} \ldots \left(u_d^{(i)}\right)^{\alpha_d}\right| \leq \sum_{\alpha:|\alpha|\leq p} \left|c^\alpha \left(u_1^{(i)}\right)^{\alpha_1} \ldots \left(u_d^{(i)}\right)^{\alpha_d}\right| \\ &\leq \|c\|_1 \max_{\alpha:|\alpha|\leq p} \prod_{j=1}^{d} \left|\left(u_j^{(i)}\right)^{\alpha_j}\right| \leq \|c\|_1 \left(1 + (B_{\mathcal{X}} + B_\Theta)^p\right).\end{aligned} \tag{3.9}$$

Similarly, we have

$$\varphi^c(\mathbf{x}^{(i)}) \leq \|c\|_1 \left(1 + B_{\mathcal{X}}^p\right).$$

Therefore,

$$|R_i| \leq |(\Phi c)_i| + \varphi^c\left(\mathbf{x}^{(i)}\right) \leq 2\|c\|_1 (1 + (B_{\mathcal{X}} + B_\Theta)^p).$$

Since $\|\Theta\|_{2,0} \le s_\theta$, we deduce $\|R\|_0 \le s_\theta$ and

$$\|R\|_1 = \sum_{i=1}^{m} |R_i| \le 2s_\theta \|c\|_1 (1 + (B_{\mathcal{X}} + B_\Theta)^p). \tag{3.10}$$

Thus, in the event that (3.8) holds, we have combined with (3.10) that

$$\begin{aligned} \|\Phi c\|_1 &\ge \sum_{i=1}^{m} \varphi^c\left(\mathbf{x}^{(i)}\right) - \|R\|_1 \\ &\ge \|c\|_1 \left(\frac{1}{2} mD - 2s_\theta (1 + (B_{\mathcal{X}} + B_\Theta)^p)\right) \\ &\ge \frac{1}{4} mD \|c\|_1, \end{aligned} \tag{3.11}$$

provided moreover that

$$m \ge \frac{8s_\theta (1 + (B_{\mathcal{X}} + B_\Theta)^p)}{D} = \widetilde{C}.$$

Now, we are ready to verify the null space property condition for $A = [Id_m, \Phi_{m\times r}]$ in the event that (3.8) holds. Let $S \subset [m + r]$ be an arbitrary set of size s and $w \in \ker A \setminus \{\vec{0}\}$. Denote $\hat{c} \in \mathbb{R}^r$ to be the last r entries of w, and

$$S_1 = S \cap [m], \quad S_2 = (S \cap \{m + 1, \dots, m + r\}) - m \subset [r].$$

Since $w \in \ker A \setminus \{\vec{0}\}$, $\hat{c} \ne \vec{0}_r$ and $w = [-\Phi\hat{c}, \hat{c}]$. Using the inequality (3.9), we have

$$\|w_S\|_1 = \|\hat{c}_{S_2}\|_1 + \|(\Phi\hat{c})_{S_1}\|_1 \le \|\hat{c}\|_1 + \|(\Phi\hat{c})_{S_1}\|_1 \le \|\hat{c}\|_1 \left(1 + s\left(1 + (B_{\mathcal{X}} + B_\Theta)^p\right)\right).$$

On the other hand, using the inequality (3.11), we obtain

$$\|w\|_1 = \|\hat{c}\|_1 + \|\Phi\hat{c}\|_1 \ge \|\hat{c}\|_1 \left(1 + \frac{1}{4} mD\right).$$

Then when m satisfies (3.5), (3.7), and

$$m > \frac{4 + 8s(1 + (B_{\mathcal{X}} + B_\Theta)^p)}{D} > \widetilde{C}, \tag{3.12}$$

we have $\|w_S\|_1 < \frac{1}{2}\|w\|_1$, for any $w \in \ker A \setminus \{\vec{0}\}$. That completes our proof. □

Remark 3.5.

- Since $\|\Phi c\|_1 \ge \frac{1}{4} mD \|c\|_1$ for any $c \in \mathbb{R}^r \setminus \{0\}$ with probability $1 - m^{-\delta r}$, we conclude that the matrix Φ is of full column rank.
- From the proof, we also derive that if $s \ge r$, the matrix $[Id_m, \Phi_{m\times r}]$ satisfies the partial null space property of order $s - r$ (see [3], Definition 3.1).
- If we keep the conditions (3.5) and (3.7), and change the condition (3.12) to

$$m > \frac{4 + 4s(\rho + 1)(1 + (B_{\mathcal{X}} + B_\Theta)^p)}{\rho D}, \tag{3.13}$$

then $\|w_S\|_1 \le \frac{\rho}{\rho + 1}\|w\|_1$, for any $w \in \ker A$ and any set $S \subset [m+r]$ with $card(S) \le s$. It means A satisfies the stable null space property of order s.

Combining with the reconstruction results from compressed sensing (see Propositions 3.2 and 3.3), we immediately obtain the following reconstruction guarantees.

Theorem 3.6. *Fix $p \in \mathbb{N}$. Suppose we observe corrupted measurements*

$$\left(\mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\theta}^{(i)}, y^{(i)} = f(\mathbf{x}^{(i)}) + \varepsilon^{(i)}\right)_{i=1}^{m} \subset \mathbb{R}^d \times \mathbb{R},$$

where $\{\mathbf{x}^{(i)}\}$ and $\{\boldsymbol{\theta}^{(i)}\}$ satisfy the assumptions in Theorem 3.4, and f is a sparse multivariate polynomial with at most s_c monomial terms of degree at most p. Denote $\mathbf{y} = (y^{(i)})$, $s = s_c + s_\theta$, Φ_U to be the dictionary matrix (2.1), and $\mathbf{c}$ to be the unknown polynomial coefficients of f. The problem can be recast as

$$\mathbf{y} = \Phi\mathbf{c} + \mathbf{e},$$

for some $\mathbf{e} \in \mathbb{R}^m$.

(a) When $\boldsymbol{\varepsilon} = 0$, then $\operatorname{supp}\mathbf{e} = \{i : \boldsymbol{\theta}^{(i)} \neq 0\}$. Suppose $\|\mathbf{c}\|_0 + \|\mathbf{e}\|_0 \leq s$, then there is a constant $D > 0$ depending only on p, d, and μ, so that when m satisfies (3.4), the polynomial coefficients $\mathbf{c}$ of f as well as the vector $\mathbf{e}$ can be exactly recovered with probability $(1 - m^{-\delta r})$ from the unique solution to the ℓ_1-minimization problem:

$$\min_{\mathbf{c}',\mathbf{e}'} \|\mathbf{e}'\|_1 + \|\mathbf{c}'\|_1 \quad \text{subject to } \Phi\mathbf{c}' + \mathbf{e}' = \mathbf{y}.$$

(b) When $\boldsymbol{\varepsilon} \neq 0$ and is not necessarily sparse, if m satisfies (3.13), (3.5), and (3.7), a solution $(\mathbf{c}^\#, \mathbf{e}^\#)$ to the ℓ_1-minimization (2.3) approximates the true solution (c, e) with ℓ_1-error:

$$\|\mathbf{c} - \mathbf{c}^\#\|_1 + \|\mathbf{e} - \mathbf{e}^\#\|_1 \leq \frac{2(1+\rho)}{1-\rho}\left(\|\mathbf{c} - \mathbf{c}^*\|_1 + \|\mathbf{e} - \mathbf{e}^*\|_1\right),$$

where $[\mathbf{c}^*, \mathbf{e}^*]$ is the best s-term approximation (vector of s largest-magnitude entries) of $[\mathbf{c}, \mathbf{e}]$ and $\rho \in (0, 1)$ is the stable null space constant of the matrix $[Id_m, \Phi_U]$.

Remark 3.7.

- The partial ℓ_1-minimization problem in [51]

$$\min_{\mathbf{c}',\mathbf{e}'} \|\mathbf{e}'\|_1 \quad \text{subject to} \quad \mathbf{y} = \Phi\mathbf{c}' + \mathbf{e}',$$

  is a special case of problem (2.3) with $s = s_\theta \geq r$. In other words, given corrupted input–output data where the corruption measurements are s_θ-sparse, we can recover the polynomial function that fits the given data and detects the outliers correctly.
- The same result in Theorem 3.6 can be extended immediately to learn a system of high-dimensional polynomial functions $\mathbf{f} = (f_1, \ldots, f_n) \in \mathbb{R}^n$ with the same coefficient matrix, where each f_j is a multivariate polynomial of degree at most p:

$$\min_{\mathbf{c}'_j,\mathbf{e}'_j} \|\mathbf{c}'_j\|_1 + \|\mathbf{e}'_j\|_1 \quad \text{subject to} \quad \mathbf{y}_j = \Phi\mathbf{c}'_j + \mathbf{e}'_j, \quad 1 \leq j \leq n.$$

- By considering a slight modification of the matrix A, $\widetilde{A} = \left[\frac{1}{\lambda} Id_m, \Phi\right]$, we can verify that $\widetilde{A}$ also satisfies the null space property, provided that m is sufficiently large. Indeed, every $w \in \ker \widetilde{A} \setminus \{0\}$ can be written as $w = [-\lambda\Phi\hat{\mathbf{c}}, \hat{\mathbf{c}}]$. Then with the lower bound on

$\|\Phi \mathbf{c}\|_1$, we can immediately show $\|w_S\|_1 < \frac{1}{2}\|w\|_1$, provided that

$$\begin{aligned} m &\geq \left(\max\{3 + 3B_{\mathcal{X}}^p, 4D^{-1}\}\right)^{1/\delta}, \\ m &> \frac{8s\,(1 + (B_{\mathcal{X}} + B_{\Theta})^p)}{D}, \\ \lambda &> \frac{4}{mD - 8s\,(1 + (B_{\mathcal{X}} + B_{\Theta})^p)}. \end{aligned} \tag{3.14}$$

Hence, the corrupted compressed sensing problem

$$\min_{\mathbf{c}',\mathbf{e}'} \|\mathbf{c}'\|_1 + \lambda\|\mathbf{e}'\|_1, \quad \text{subject to } \mathbf{y} = \Phi\mathbf{c}' + \mathbf{e}',$$

will have a unique solution.

4. Recovery results for various types of data

In this section, we apply our results to several popular types of weakly dependent data. Indeed, we only need to verify that these types of data satisfy the required concentration inequality in Theorem 3.4. For the sake of simplicity, we state the recovery results for the noiseless case of $\mathbf{y}$ (i.e., when $\varepsilon^{(i)} = 0$).

4.1. Independent and identically distributed (i.i.d.) data

In [48], the authors provide the following Bernstein inequality for i.i.d. random variables:

Lemma 4.1. *If $\{\mathbf{x}^{(i)}\}$ are i.i.d. random variables with $|\varphi(\mathbf{x}^{(1)}) - \mathbb{E}(\varphi(\mathbf{x}^{(1)}))| \leq C_1$ a.s., then the following probability inequality holds for all $m \geq 1$:*

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m}\varphi(\mathbf{x}^{(i)}) - \mathbb{E}[\varphi(X)]\right| \geq \zeta\right) \leq 2\exp\left(-\frac{\zeta^2 m}{C_2 + C_3\zeta}\right), \tag{4.1}$$

where

$$C_2 = 2\mathbb{E}[\varphi^2(\mathbf{x}^{(1)})] - 2(\mathbb{E}[\varphi(\mathbf{x}^{(1)})])^2, \;\; C_3 = \frac{2}{3}C_1,$$

and φ is any bounded Borel function.

In this case, the function κ in the concentration inequality (3.2) is

$$\kappa(\zeta, m) = \frac{\zeta^2 m}{C_2 + C_3\zeta} - \log 2,$$

and satisfies the condition (3.3) for any constant $\delta \in \left(0, \frac{1}{2}\right)$, when m is large enough. Indeed, the condition on κ can be re-written as

$$r \leq \frac{1}{3\delta \log m}\left(\frac{m}{C_2 m^{2\delta} + C_3 m^{\delta}} - \log 2\right). \tag{4.2}$$

If the maximal polynomial degree p is fixed, the smaller δ is, the smaller m is needed to satisfy the inequality (4.2).

As a result, we have the following recovery result for i.i.d data.

Theorem 4.2. *Fix $p \in \mathbb{N}$. Suppose we observe corrupted measurements*

$$\left(\mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\theta}^{(i)}, y^{(i)} = f\left(\mathbf{x}^{(i)}\right)\right)_{i=1}^{m} \subset \mathbb{R}^d \times \mathbb{R},$$

where the uncorrupted data $\{\mathbf{x}^{(i)}\}$ are i.i.d. according to a non-degenerate distribution μ and L^∞-bounded by $B_{\mathcal{X}}$; the corruption $\{\boldsymbol{\theta}^{(i)}\}$ is L^∞-bounded by B_Θ and s_θ-row sparse; and f is a sparse multivariate polynomial with at most s_c monomials of degree at most p. Then, when m satisfies (3.4) and (4.2), the polynomial coefficients of the function f can be exactly recovered and the outliers can be successfully detected from the unique solution of (2.3) with high probability.

4.2. Exponentially strongly α-mixing data

We first recall the definition of α-mixing coefficients and a concentration inequality for α-mixing. For a stationary stochastic process $\{\mathbf{x}_t\}$, define (see [32,38])

$$\alpha(s) = \sup_{\substack{-\infty<t<\infty \\ A\in\sigma(\mathbf{x}_t^-), B\in\sigma(\mathbf{x}_{t+s}^+)}} |\Pr(A\cap B) - \Pr(A)\Pr(B)|,$$

where

$$\sigma(x_t^-) = \sigma(x_k \mid k\in\mathbb{Z}, k\le t), \quad \sigma(x_{t+s}^+) = \sigma(x_k \mid k\in\mathbb{Z}, k\ge t+s).$$

The stochastic process is said to be *exponentially strongly α-mixing* if

$$\alpha(s) \le \overline{\alpha}\exp(-c_\alpha s^\beta), \quad s\ge 1,$$

for some $\overline{\alpha} > 0$, $\beta > 0$, and $c_\alpha > 0$, where the constants β and c_α are assumed to be known. Note that strong mixing implies asymptotic independence over sufficiently large time.

In [32], the authors proved the following concentration inequality for exponentially strongly α-mixing:

Lemma 4.3. *If $\{\mathbf{x}^{(i)}\}$ are stationary exponentially strongly α-mixing with $|\varphi(\mathbf{x}^{(1)}) - \mathbb{E}(\varphi(\mathbf{x}^{(1)}))| \le C_0$ a.s., then the following probability inequality holds for all $m_\alpha \ge 2$ and for all $\zeta > 0$:*

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m}\varphi(\mathbf{x}^{(i)}) - \mathbb{E}[\varphi(\mathbf{x}^{(1)})]\right| \ge \zeta\right) \le C_1\exp\left(-\frac{\zeta^2 m_\alpha}{C_2 + C_3\zeta}\right), \tag{4.3}$$

where

$$m_\alpha := \left\lfloor \frac{m}{\lceil (8m/c_\alpha)^{1/(\beta+1)}\rceil}\right\rfloor = C_\alpha\, m^{\beta/(\beta+1)},$$

$$C_1 = 2(1+4e^{-2}\overline{\alpha}),\ C_2 = 2\mathbb{E}(\varphi^2(\mathbf{x}^{(1)})) - 2(\mathbb{E}(\varphi(\mathbf{x}^{(1)})))^2,\ C_3 = \frac{2}{3}C_0,$$

and φ is any bounded Borel function.

Hence the concentration inequality (3.2) is satisfied with

$$\kappa(\zeta, m) = \frac{\zeta^2 m_\alpha}{C_2 + C_3\zeta} - \log C_1.$$

Since

$$\kappa(m^{-\delta}, m) = \frac{m_\alpha}{C_2 m^{2\delta} + C_3 m^{\delta}} - \log C_1 \geq 3\delta r \log m, \tag{4.4}$$

for any $\delta \in \left(0, \frac{\beta}{2(\beta+1)}\right)$ when m is large enough, we have the recovery result for exponentially strongly α-mixing data.

Theorem 4.4. *Fix $p \in \mathbb{N}$. Suppose we observe corrupted measurements*

$$\left(\mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\theta}^{(i)}, y^{(i)} = f(\mathbf{x}^{(i)})\right)_{i=1}^{m} \subset \mathbb{R}^d \times \mathbb{R},$$

where the uncorrupted data $\{\mathbf{x}^{(i)}\}$ are stationary exponentially strongly α-mixing and L^∞-bounded by $B_{\mathcal{X}}$; the corruption $\{\boldsymbol{\theta}^{(i)}\}$ is L^∞-bounded by B_Θ and s_θ-row sparse; and f is a sparse multivariate polynomial with at most s_c monomials of degree at most p. If the stationary distribution μ of $\{\mathbf{x}^{(i)}\}$ is non-degenerate, then when m satisfies Eqs. (3.4) and (4.4), the polynomial coefficients of the function f can be exactly recovered and the outliers can be successfully detected from the unique solution of (2.3) with high probability.

4.3. Geometrically (time-reversed) $\mathcal{C}$-mixing data

The $\mathcal{C}$-mixing processes were introduced in [31] to exhibit many common dynamical systems that are not necessary α-mixing such as Lasota–Yorke maps, uni-modal maps, piecewise expanding maps in higher dimension. Moreover, the geometrically $\mathcal{C}$-mixing processes are strongly related to some well-known results on the decay of correlations for dynamical systems (see [22]).

Let $\{\mathbf{x}^{(i)}\}$ be an $\mathcal{X}$-valued stationary process on $(\Omega, \mathcal{A}, \mu)$. For a semi-norm $\|\cdot\|$ on a vector space of bounded measurable functions that satisfies $\|\exp(h)\| \leq \|\exp(h)\|_\infty \|h\|$, for every bounded measurable function $h : \mathcal{X} \to \mathbb{R}$, we define the $\mathcal{C}$-norm by $\|h\|_{\mathcal{C}} = \|h\|_\infty + \|h\|$.

Let $\mathcal{A}_1^i$ and $\mathcal{A}_{i+m}^\infty$ be the σ-algebras generated by $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(i)})$ and $(\mathbf{x}^{(i+m)}, \mathbf{x}^{(i+m+1)}, \ldots)$ respectively. Then, the $\mathcal{C}$-mixing coefficient is

$$\phi_{\mathcal{C}}(m) = \sup\Big\{\Big|\mathbb{E}[Zh(\mathbf{x}^{(i+m)})] - \mathbb{E}(Z)\mathbb{E}[h(\mathbf{x}^{(i+m)})]\Big| : i \geq 1,$$
$$Z \text{ is } \mathcal{A}_1^i\text{-measurable and } \|Z\|_1 \leq 1, \|h\|_{\mathcal{C}} \leq 1\Big\},$$

and the time-reversed $\mathcal{C}$-mixing coefficient is

$$\phi_{\mathcal{C},\mathrm{rev}}(m) = \sup\Big\{\Big|\mathbb{E}[Zh(X^{(i)})] - \mathbb{E}(Z)\mathbb{E}[h(X^{(i)})]\Big| : i \geq 1,$$
$$Z \text{ is } \mathcal{A}_{i+m}^\infty\text{-measurable and } \|Z\|_1 \leq 1, \|h\|_{\mathcal{C}} \leq 1\Big\}.$$

A sequence of random variables $\{\mathbf{x}^{(i)}\}$ is called geometrically (time-reversed) $\mathcal{C}$-mixing if

$$\phi_{\mathcal{C},(\mathrm{rev})}(m) \leq c\exp(-bm^\beta), \quad m \geq 1,$$

for some constants $b > 0, c \geq 0$, and $\beta > 0$. The following concentration inequality for stationary geometrically (time-reversed) $\mathcal{C}$-mixing process is a direct consequence of the Bernstein inequality presented in [22].

Lemma 4.5. *Let $\{\mathbf{x}^{(i)}\}_{i\geq 1}$ be a stationary geometrically (time reversed) $\mathcal{C}$-mixing process. Consider a function $\varphi : \mathcal{X} \to \mathbb{R}$ such that $\|\varphi\| \leq A$, $\|\varphi\|_\infty \leq B$, and $Var(\varphi(\mathbf{x}^{(1)})) \leq \sigma^2$. Then, for sufficient large m (see Equation 3.1 in [22]) we have*

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m}\varphi(\mathbf{x}^{(i)}) - \mathbb{E}[\varphi(\mathbf{x}^{(1)})]\right| \geq \zeta\right) \leq 4\exp\left(-\frac{m\zeta^2}{8(\log m)^{2/\beta}\left(\sigma^2 + \zeta B/3\right)}\right). \tag{4.5}$$

In this case, the concentration inequality (3.2) holds for

$$\kappa(\zeta, m) = \frac{m\zeta^2}{8(\log m)^{2/\beta}\left(\sigma^2 + \zeta B/3\right)} - \log 4, \tag{4.6}$$

and satisfies the condition (3.3) for any $\delta \in \left(0, \frac{1}{2}\right)$ when m is large enough. Hence, we have the recovery result for geometrically (time-reversed) $\mathcal{C}$-mixing data.

Theorem 4.6. *Fix $p \in \mathbb{N}$. Suppose we observe corrupted measurements*

$$\left(\mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\theta}^{(i)}, y^{(i)} = f(\mathbf{x}^{(i)})\right)_{i=1}^{m} \subset \mathbb{R}^d \times \mathbb{R},$$

where the uncorrupted data $\{\mathbf{x}^{(i)}\}$ are stationary geometrically (time-reversed) $\mathcal{C}$-mixing with respect to the semi-norm $\|h\| = \sup_{X\in\mathcal{X}} \|\nabla h(X)\|_1$ and L^∞-bounded by $B_{\mathcal{X}}$; the corruption $\{\boldsymbol{\theta}^{(i)}\}$ is L^∞-bounded by B_Θ and s_θ-row sparse; and f is a sparse multivariate polynomial with at most s_c monomials of degree at most p. If the stationary distribution μ of $\{\mathbf{x}^{(i)}\}$ is non-degenerate, then when m satisfies Eqs. (3.4) and (4.6), the polynomial coefficients of the function f can be exactly recovered and the outliers can be successfully detected from the unique solution of (2.3) with high probability.

4.4. Uniformly ergodic Markov chain

Let $\{\mathbf{x}^{(i)}\}$ be a Markov chain on $(\Omega, \mathcal{A})$ with a unique stationary distribution μ. We define:

$$P^k(x, A) = \Pr(\mathbf{x}_{k+i} \in A \mid \mathbf{x}_i = x).$$

The chain $\{\mathbf{x}^{(i)}\}$ is called *uniformly ergodic* if

$$\sup_x \|P^k(x, .) - \mu(.)\|_{TV} \to 0 \quad \text{as} \quad k \to \infty.$$

In this case, there exist a positive integer k_0, $\lambda > 0$, and a probability distribution ρ such that

$$P^{k_0}(x, A) \geq \lambda\rho(A), \quad \text{for all} \quad x \in \mathcal{X}, A \in \mathcal{A}.$$

We have the following concentration inequality for uniformly ergodic Markov chain [26]:

Lemma 4.7. *Let $\{\mathbf{x}^{(i)}\}$ be a stationary uniformly ergodic Markov chain. Then for any $\varphi : \mathcal{X} \to \mathbb{R}$ such that $\|\varphi\|_\infty \leq B$, any $\zeta > 0$, and $m \geq 1 + 3k_0B/(\lambda\zeta)$, we have*

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m}\varphi(\mathbf{x}^{(i)}) - \mathbb{E}[\varphi(\mathbf{x}^{(1)})]\right| \geq \zeta\right) \leq 2\exp\left[-\frac{m-1}{2}\left(\frac{\lambda}{k_0B}\zeta - \frac{3}{m-1}\right)^2\right]. \tag{4.7}$$

In this case, the concentration inequality (3.2) holds for

$$\kappa(\zeta, m) = \frac{m-1}{2}\left(\frac{\lambda}{k_0B}\zeta - \frac{3}{m-1}\right)^2 - \log 2.$$

Observe that

$$\begin{aligned}\kappa(m^{-\delta}, m) &= \frac{m-1}{2}\left(\frac{\lambda}{k_0 B}m^{-\delta} - \frac{3}{m-1}\right)^2 - \log 2 \\ &= \frac{\lambda^2(m-1)m^{-2\delta}}{2k_0^2B^2} - \frac{3\lambda m^{-\delta}}{k_0 B} + \frac{9}{2(m-1)} - \log 2 \\ &\geq 3\delta r \log m,\end{aligned} \tag{4.8}$$

for any $\delta \in \left(0, \frac{1}{2}\right)$ when m is large enough. Therefore, we have the recovery result for uniformly ergodic Markov chain data.

Theorem 4.8. *Fix $p \in \mathbb{N}$. Suppose we observe corrupted measurements*

$$\left(\mathbf{u}^{(i)} = \mathbf{x}^{(i)} + \boldsymbol{\theta}^{(i)}, y^{(i)} = f(\mathbf{x}^{(i)})\right)_{i=1}^{m} \subset \mathbb{R}^d \times \mathbb{R},$$

where the uncorrupted data $\{\mathbf{x}^{(i)}\}$ form a stationary uniformly ergodic Markov chain and L^∞-bounded by $B_\mathcal{X}$; the corruption $\{\boldsymbol{\theta}^{(i)}\}$ is L^∞-bounded by B_Θ and s_θ-row sparse; and f is a sparse multivariate polynomial with at most s_c monomials of degree at most p. If the stationary distribution μ of $\{\mathbf{x}^{(i)}\}$ is non-degenerate, then when m satisfies Eqs. (3.4) and (4.8), the polynomial coefficients of the function f can be exactly recovered and the outliers can be successfully detected from the unique solution of (2.3) with high probability.

5. Numerics and computational results

In this section, we verify the exact recovery of polynomial coefficients from data sampled from a stationary process with sparse random corruptions. For each of the examples, we use exponentially strong α-mixing data, although the other related processes would produce similar results. Our computational tests verify the recovery results from Theorem 3.6 as well as the method's dependence on parameters such as the sampling rate, polynomial degree, the sampling distribution of the corruption vector, and the sparsity of the corruption vector.

5.1. Algorithm

To solve the constrained optimization problem (2.4), we use the well-known Douglas–Rachford algorithm [10,29]. Eq. (2.4) can be written as:

$$\min_w \|w\|_1 \quad \text{subject to} \quad y = Aw, \tag{5.1}$$

with the new variable $w = [e, c]$ and the augmented matrix $A = [\lambda^{-1} Id_m,\ \Phi_{m\times r}]$. This can be relaxed to:

$$\min_w \|w\|_1 \quad \text{subject to} \quad \|y - Aw\|_2 \leq \sigma, \tag{5.2}$$

for some (non-negative) parameter σ. Using Eq. (5.2) gives the experiments more flexibility, while also coinciding with Eq. (5.1) when σ is sent to zero. Following the derivation from [46], let v be an auxiliary variable relaxing the constraints:

$$(w, v) \in \mathcal{D} := \{(w, v) |\ v = Aw\} \quad \text{and} \quad v \in B_\sigma(y) := \{v \mid \|y - v\|_2 \leq \sigma\}.$$

Using the indicator function for a set $\mathcal{S}$:

$$\mathbb{I}_{\mathcal{S}}(w) := \begin{cases} 0, & \text{if } w \in \mathcal{S} \\ \infty, & \text{if } w \notin \mathcal{S}, \end{cases}$$

Eq. (5.2) can be written as:

$$\min_{(w,v)} \; g_1(w, v) + g_2(w, v), \tag{5.3}$$

where $g_1(w, v) := \|w\|_1 + \mathbb{I}_{B_\sigma(y)}(v)$ and $g_2(w, v) := \mathbb{I}_{\mathcal{D}}(w, v)$. The Douglas–Rachford algorithm uses the proximal operator within its iterations. The proximal operator of a function g is defined as:

$$\text{prox}_{\gamma g}(z) := \underset{x}{\text{argmin}} \left\{ \frac{1}{2}\|z - x\|^2 + \gamma\, g(x) \right\},$$

where $\gamma > 0$ is a free parameter. As shown in [46], the proximal operators for g_1 and g_2 are as follows:

$$\text{prox}_{\gamma g_1}(w, v) = \left(S_\gamma(w),\ \text{proj}_{B_\sigma(y)}(w) \right),$$

where proj_{B_σ} is the Euclidean projection onto the ball:

$$\text{proj}_{B_\sigma(y)}(v) := \begin{cases} v, & \text{if } \|v - y\|_2 \le \sigma \\ y + \sigma \dfrac{v - y}{\|v - y\|_2}, & \text{if } \|v - y\|_2 > \sigma, \end{cases}$$

and the soft-thresholding function S (also known as *shrink*) is defined component-wise as:

$$S_\gamma(w_j) = \text{sign}(w_j)\, \max\left(|w_j| - \gamma, 0 \right).$$

The proximal operator for g_2 is defined by:

$$\text{prox}_{\gamma g_2}(w, v) = \left((Id_{m+r} + A^T A)^{-1}(w + A^T v),\ A(Id_{m+r} + A^T A)^{-1}(w + A^T v) \right).$$

Define $\text{rprox}_{\gamma g}(v) := 2\text{prox}_{\gamma g}(v) - v$, then the iterations for the Douglas–Rachford method are:

$$\begin{aligned} (\tilde{w}^{k+1}, \tilde{v}^{k+1}) &= \frac{1}{2}\left((\tilde{w}^k, \tilde{v}^k) + \text{rprox}_{\gamma g_2}\left(\text{rprox}_{\gamma g_1}\left(\tilde{w}^k, \tilde{v}^k \right) \right) \right), \\ (w^{k+1}, v^{k+1}) &= \text{prox}_{\gamma g_1}(\tilde{w}^{k+1}, \tilde{v}^{k+1}), \end{aligned} \tag{5.4}$$

and will converge to a minimizer of Eq. (5.2) for any $\gamma > 0$ [10]. Unless otherwise stated, in all of the computational examples we set $\gamma = 1$ and $\sigma = 10^{-10}$.

It is worth noticing that although the condition number $\kappa(I + A^T A)$ may be bad for monomial dictionaries, the effective condition number to solve $(I + A^T A)x = b$ may be manageable by looking at the space b is in (see Theorem 1 in [5]). Since we are looking for sparse solutions, b can be written over a subset of columns of $(I + A^T A)$, thus lowering the sensitivities. Nevertheless, the verification of the effective well-conditioning of the Douglas–Rachford algorithm is out of scope of this paper and we leave it for future work.

5.2. Computational results

Throughout the following examples, the uncorrupted data $\mathbf{x}^{(i)} \in \mathbb{R}^d$ are simulated as follows: First, we simulate a sequence of bounded i.i.d. random variables $\mathbf{z}^{(i)}$, $i = 1, 2, \ldots, m+3$. Then,

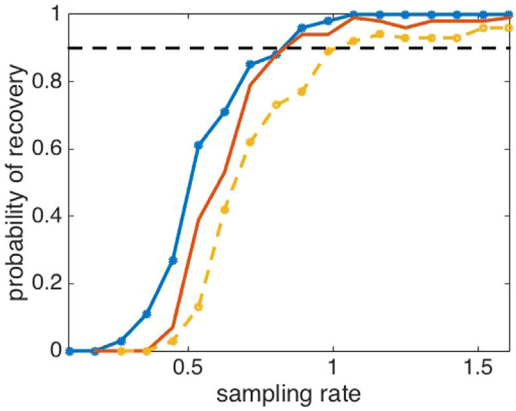

Fig. 1. Probability of exact recovery versus the sampling rate $\frac{m}{r}$ for the 5th order Eq. (5.5) with $d = 3$ ($r = 56$ monomial terms) and the sparsity of the polynomial coefficient vector equal to 3. We measure the rate of exact recovery over 100 trials, varying the sparsity of the vector e: 5 (blue solid line with dots), 10 (red solid line), and 12 (yellow dotted line). The horizontal line is the threshold to achieve 90% probability of success, which are all achieved with under-sampled data i.e. $m < r$. For large enough sampling rate (for example 150 samples), the probability of recovery approaches one.

we set

$$\mathbf{x}^{(i)} = \frac{\mathbf{z}^{(i)}}{16} + \frac{\mathbf{z}^{(i+1)}}{8} + \frac{\mathbf{z}^{(i+2)}}{4} + \frac{\mathbf{z}^{(i+3)}}{2}, \quad i = 1, 2, \ldots, m.$$

This leads to a sequence $(\mathbf{x}^{(i)})$ which is a stationary exponentially strongly α-mixing process.

Example 1. For the first example, we consider learning the function:

$$f(x) = 1 - 2x_1x_2x_3 + 5x_1^5, \tag{5.5}$$

where $\mathbf{x} \in \mathbb{R}^3$. In Fig. 1, we display the probability of exact recovery of the polynomial coefficients in Eq. (5.5) versus the sampling rate $\frac{m}{r}$, *i.e.* the number of rows versus columns in Φ. For this example, the matrix Φ contains all monomials up to fifth order in $\mathbb{R}^3$ for a total of 56 columns. We measure the probability of exact recovery by comparing the computed support set to the exact support set. We compute the success rate over 100 trials with random sparse outliers, varying the sparsity of the vector e between 5 (blue solid line with dots), 10 (red solid line), and 12 (yellow dotted line). In all examples, the sparsity of the vector e refers to the number of non-zero elements, i.e. s_θ. The vector e is drawn randomly for each trial from the normal distribution with standard deviation equal to 10. This value is chosen to roughly match the amplitude of the coefficients with respect to Φ, thus leading to a more challenging outlier-removal problem. The threshold to achieve 90% probability of success is marked by the horizontal dotted line. Note that the method is able to recovery the polynomial coefficients with 90% probability even when the matrix Φ is under-sampled ($m < r$.) As the sparsity of the vector e increases, the probability of exact recovery decreases, as expected. For large enough sampling rates, in particular after 150 samples, the calculated probability of recovery is nearly one.

Example 2. The second example investigates the recovery of the function:

$$f(x) = -1 - 2\,x_1^p, \tag{5.6}$$

for various $p > 1$ where $\mathbf{x} \in \mathbb{R}^3$. In Table 1, the probability of exact recovery versus the maximum degree of the polynomial is shown for two sampling rates: 15% and 35%. The maximum degree of the monomials used in the dictionary is set to 10 for all runs (for a total of 286 monomial terms). The sparsity of the polynomial coefficient vector is 2 and the sparsity of the vector e is set to 5 in this example. The vector e is drawn randomly for each trial from the normal distribution with standard deviation equal to 10. For this example, the matrix Φ is normalized column-wise before applying the algorithm in order to prevent potential bias due to the column-scaling as p increases. It was observed that normalization also helps with the numerical stability of the problem when using higher-order monomials.

For 15% sampling (the second row of Table 1), the computed probability of exact recovery is fairly stable (outside of the $p = 3$ case). One can observe a small decrease in the recovery rate between $p = 5$, $p = 8$, and $p = 10$. To test the stability in the high-recovery limit, we set the sampling rate to 35% (third row of Table 1). For 35% sampling, as p increases from 2 to 10 the recovery rate stays nearly the same (98% to 99%).

Example 3. In this example, we investigate the recovery of the function:

$$f(x) = -8.5 + 9.6\,x_1\,x_4 + 0.3\,x_2\,x_5 + 5.7\,x_1^3 + 1.9 x_3\,x_9^2, \tag{5.7}$$

under various conditions on the vector e. The non-zero values of the vector e are sampled uniformly from $[-H, H]$. We compute the probability of exact recovery versus the values of H and changes in the sparsity s_θ. The values of H are chosen to match the range of the magnitudes of the coefficients in Eq. (5.7). For this example, we use all monomials up to degree three in $\mathbb{R}^{10}$ for a total of 286 monomial terms. The sampling rate is fixed at 17.5% in order to highlight the variability between the recovery rates. In Table 2, we vary H between 0.5, 2, and 10 and the sparsity of the vector e between 3, 10, and 15. The scaling parameter λ is set to 2 for all experiments in Table 2. In all cases, as H increases, the recovery rate also increases. As the sparsity of the vector e increases, the recovery rate decreases, as expected. It was observed that the failures tended to occur on the vector $\mathbf{e}$, while the recovery of the polynomial coefficients was fairly stable to changes in H and to changes in the sparsity s_θ. In fact, it is possible, for certain cases, to recover the polynomial coefficients for relatively dense outlier vectors.

Example 4. This example details the recovery of the function:

$$f(x) = -1 + 2\,x_1^2 + 0.5\,x_5\,x_{20}, \tag{5.8}$$

when it is perturbed by the function $g(\mathbf{x}) = \epsilon \sin(2\pi x_1)$ (which is noise and is not part of the dictionary). We compute the probability of exact recovery of the polynomial coefficients in Eq. (5.8) versus various values of ϵ. We use all monomials up to degree two in $\mathbb{R}^{20}$ for a total of 231 monomial terms. The sampling rate is fixed at 21.7% and the sparsity of the vector e is set to 3. In Table 3, we can see that as ϵ increases, the recovery rate decreases. The average ℓ_1 error (over the successful trials) is stable over the various tests. For larger values of ϵ, it may be possible to recover the sparse coefficients by adjusting λ. For example, by lowering λ to 0.9, the recovery rate for $\epsilon = 10^{-3}$ becomes 94 and the average ℓ_1 error becomes 0.0141.

Table 3
Probability of exact recovery (of the polynomial coefficients) versus the sparsity of the vector e (see Eq. (5.8)). The ambient dimension is $d = 20$, the maximum degree of the candidate polynomials is 2 (thus $r = 231$ monomial terms), and the sparsity of the polynomial coefficients is equal to 3. In each test, the sampling rate is 21.7% and the parameter is set to $\lambda = 1$.

ϵ	0	10^{-5}	10^{-4}	10^{-3}
Recovery	99	98	98	85
ℓ_1 error	0.0144	0.0120	0.0104	0.0156

Table 4
Probability of exact recovery (of the polynomial coefficients in Eq. (5.9)) versus the sparsity ratio ρ_θ. The ambient dimension is $d = 3$, the maximum degree of the candidate polynomials is 5 (thus $r = 56$ monomial terms), and the sparsity of the polynomial coefficients is equal to 3. In each test, the parameter is set to $\lambda = 1$.

ρ_θ	0.25	0.5	0.75
$m = 200$	93	82	7
$m = 400$	97	94	44
$m = 600$	99	96	91

Example 5. The last example details the effect the sparsity of the vector e has on the recovery, in particular, when the sparsity is a fixed percentage of the total number of samples. Using the following function:

$$f(x) = 1 - 2x_1x_2x_3 + 5x_1^5, \tag{5.9}$$

we compute the probability of exact recovery of the polynomial coefficients in Eq. (5.9) versus various sparsity ratios and various sampling sizes. We denote the sparsity ratio by ρ_θ. The vector e is drawn randomly for each trial from the normal distribution with standard deviation equal to 15. In Table 4, we see that as ρ_θ increase, the recovery rate decreases. This also shows a limitation in the method's ability to recovery the polynomial from very few uncorrupted samples. For example, when $m = 200$ and $\rho_\theta = 0.75$, then there are only 50 non-corrupt samples to fit a polynomial with 56 free parameters. Since the 50 samples are weakly dependent, this may not be sufficient to determine the coefficients accurately. Also, this experiment shows that in the over-sampling regime, i.e. $m = 600$ and $\rho_\theta \leq 0.5$, the method gives the correct result fairly consistently (with probability larger than 0.9).

Remark 5.1. One approach for removing outliers from linear regression estimates is by thresholding the Cook's distance D_i, for data point i. Applying this approach to Example 5, with a threshold of $3\,\text{mean}(D_i)$ leads to zero probability of being able to located the corrupt samples. This leads to regression results with large errors. Alternatively, one may wish to vary the threshold for each problem; however, this will not lead to successful outlier removal for the data used in our experiments. For example, in the middle entry of Table 4 (i.e. $\rho_\theta = 0.5$ and $m = 400$) the minimum and maximum D_i for i in the corrupt set are $(4.3 \times 10^{-10},\ 5.5 \times 10^{-2})$, while the minimum and maximum D_i for i in the uncorrupted set are $(4.1 \times 10^{-10},\ 9.4 \times 10^{-4})$. This indicates that there is no value that would exactly separate the corrupt and uncorrupted data samples by thresholding Cook's distance.

6. Conclusion

Function approximation via ℓ_1-optimization is a useful technique for automated learning. There are many results on the behavior of the ℓ_1-solution when applied to i.i.d. data; however, theoretical results for dependent data is limited. The overall goal of this work is to show that under weaker conditions, exact and stable recovery is guaranteed. Specifically, we have shown that if the data is not independent but satisfies a suitable concentration inequality, one can provide a recovery guarantee for the learning function problem with corrupted data. Moreover, our proofs also show that the associated dictionary matrix generated from this type of data satisfies the null space property. It may be possible to weaken the requirements further while still preserving the core results. From numerical experiments, we observe that we need fewer number of measurements m (compared to the theoretical bounds) to recover the underlying function. It is likely that new theories are needed to incorporate the sparsity level of the target function to relax the conditions on m. Specifically, it would be interesting to investigate the compressed sensing setting for dependent data where the number of measurements m is less than r. In that direction, existing literature on the sparse linear regression problem in the compressive sensing setting considers only the ideal cases where the sampling matrix has i.i.d. Gaussian entries — see for example [17,25,27,54], or is formed from bounded orthonormal systems [2], randomly modulated unit-norm frames, or randomly subsampled orthonormal matrices [56]. However, it is often the case that sparsity (or sparsity with respect to some nice bounded orthonormal basis) may not hold. One of the benefits of our results is that they hold when the target function is indeed sparse with respect to monomials, approximately sparse (where sparsity helps with overfitting), or even dense.

CRediT authorship contribution statement

Lam Si Tung Ho: Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing. **Hayden Schaeffer:** Conceptualization, Methodology, Investigation , Software, Formal analysis, Writing - original draft, Writing - review & editing. **Giang Tran:** Conceptualization, Methodology, Investigation, Writing - original draft, Writing - review & editing. **Rachel Ward:** Conceptualization, Methodology, Investigation, Writing - review & editing.

Acknowledgments

L.S.T.H. acknowledges the support of startup funds from Dalhousie University, the Canada Research Chairs program, and NSERC Discovery Grant RGPIN-2018-05447. H.S. acknowledges the support of AFOSR, FA9550-17-1-0125 and the support of NSF CAREER grant #1752116. G.T. acknowledges the support of NSERC Discovery Grant. R.W. acknowledges the support of NSF CAREER grant #1255631.

References

[1] Y.S. Abu-Mostafa, M. Magdon-Ismail, H.-T. Lin, Learning from Data, Vol. 4, AMLBook New York, NY, USA, 2012.

[2] B. Adcock, A. Bao, J. Jakeman, A. Narayan, Compressed sensing with sparse corruptions: Fault-tolerant sparse collocation approximations, 2017, arXiv preprint arXiv:1703.00135.

[3] A.S. Bandeira, K. Scheinberg, L.N. Vicente, On partial sparse recovery, 2013, arXiv preprint arXiv:1304.2809.

[4] S.L. Brunton, J.L. Proctor, J. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (15) (2016) 3932–3937.
[5] T.F. Chan, D.E. Foulser, Effectively well-conditioned linear systems, SIAM J. Sci. Stat. Comput. 9 (6) (1988) 963–969.
[6] V. Cherkassky, F.M. Mulier, Learning from Data: Concepts, Theory, and Methods, John Wiley & Sons, 2007.
[7] A. Chkifa, A. Cohen, C. Schwab, High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs, Found. Comput. Math. 14 (4) (2014) 601–633.
[8] A. Chkifa, N. Dexter, H. Tran, C. Webster, Polynomial approximation via compressed sensing of high-dimensional functions on lower sets, Math. Comp. 87 (311) (2018) 1415–1450.
[9] A. Cohen, G. Migliorati, Optimal weighted least-squares methods, SMAI J. Comput. Math. 3 (2017) 181–203.
[10] P.L. Combettes, J.-C. Pesquet, Proximal splitting methods in signal processing, in: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, 2011, pp. 185–212.
[11] J.P. Crutchfield, B.S. McNamara, Equation of motion from a data series, Complex Syst. 1 (1987) 417–452.
[12] N.V. Cuong, L.S.T. Ho, V. Dinh, Generalization and robustness of batched weighted average algorithm with v-geometrically ergodic Markov data, in: International Conference on Algorithmic Learning Theory, Springer, 2013, pp. 264–278.
[13] V. Dinh, L.S.T. Ho, N.V. Cuong, D. Nguyen, B.T. Nguyen, Learning from non-iid data: Fast rates for the one-vs-all multiclass plug-in classifiers, in: International Conference on Theory and Applications of Models of Computation, Springer, 2015, pp. 375–387.
[14] S. Fattahi, S. Sojoudi, Data-driven sparse system identification, 2018, arXiv preprint arXiv:1803.07753.
[15] M. Fornasier, K. Schnass, J. Vybiral, Learning functions of few arbitrary linear parameters in high dimensions, Found. Comput. Math. 12 (2) (2012) 229–262.
[16] S. Foucart, H. Rauhut, A Mathematical Introduction to Compressive Sensing, Birkhäuser Basel, 2013.
[17] R. Foygel, L. Mackey, Corrupted sensing: Novel guarantees for separating structured signals, IEEE Trans. Inform. Theory 60 (2) (2014) 1223–1247.
[18] J.H. Friedman, An overview of predictive learning and function approximation, in: From Statistics to Neural Networks, Springer, 1994, pp. 1–61.
[19] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, in: Springer series in Statistics New York, vol. 1, 2001.
[20] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, Vol. 1, MIT Press Cambridge, 2016.
[21] H. Hang, I. Steinwart, Fast learning from α-mixing observations, J. Multivariate Anal. 127 (2014) 184–199.
[22] H. Hang, I. Steinwart, A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning, Ann. Statist. 45 (2) (2017) 708–743.
[23] M.I. Jordan, T.M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260.
[24] E. Kaiser, J. Kutz, S.L. Brunton, Sparse identification of nonlinear dynamics for model predictive control in the low-data limit, 2017, arXiv preprint arXiv:1711.05501.
[25] S. Karmalkar, E. Price, Compressed sensing with adversarial sparse noise via L_1 regression, 2018, Preprint.
[26] I. Kontoyiannis, L.A. Lastras-Montano, S.P. Meyn, Relative entropy and exponential deviation bounds for general Markov chains, in: Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on, IEEE, 2005, pp. 1563–1567.
[27] X. Li, Compressed sensing and matrix completion with constant proportion of corruptions, Constr. Approx. 37 (1) (2013) 73–99.
[28] W. Liao, M. Maggioni, S. Vigogna, Learning adaptive multiscale approximations to data and functions near low-dimensional sets, in: Information Theory Workshop (ITW), 2016 IEEE, IEEE, 2016, pp. 226–230.
[29] P.-L. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators, SIAM J. Numer. Anal. 16 (6) (1979) 964–979.
[30] L. Ljung, System identification, in: Signal Analysis and Prediction, Springer, 1998, pp. 163–173.
[31] V. Maume-Deschamps, Exponential inequalities and functional estimations for weak dependent data: applications to dynamical systems, Stoch. Dyn. 6 (04) (2006) 535–560.
[32] D.S. Modha, E. Masry, Minimum complexity regression estimation with weakly dependent observations, IEEE Trans. Inform. Theory 42 (6) (1996) 2133–2145.
[33] E. Novak, H. Woźniakowski, Approximation of infinitely differentiable multivariate functions is intractable, J. Complexity 25 (4) (2009) 398–404.
[34] T. Poggio, F. Girosi, Networks for approximation and learning, Proc. IEEE 78 (9) (1990) 1481–1497.

[35] T. Poggio, F. Girosi, Regularization algorithms for learning that are equivalent to multilayer networks, Science 247 (4945) (1990) 978–982.
[36] M. Raissi, G.E. Karniadakis, Hidden physics models: Machine learning of nonlinear partial differential equations, J. Comput. Phys. 357 (2018) 125–141.
[37] H. Rauhut, R. Ward, Interpolation via weighted ℓ_1 minimization, Appl. Comput. Harmon. Anal. 40 (2) (2016) 321–351.
[38] M. Rosenblatt, A central limit theorem and a strong mixing condition, Proc. Natl. Acad. Sci. 42 (1) (1956) 43–47.
[39] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[40] S.H. Rudy, S.L. Brunton, J.L. Proctor, J.N. Kutz, Data-driven discovery of partial differential equations, Sci. Adv. 3 (4) (2017) e1602614.
[41] D. Ryabko, Pattern recognition for conditionally independent data, J. Mach. Learn. Res. 7 (Apr) (2006) 645–664.
[42] H. Schaeffer, Learning partial differential equations via data discovery and sparse optimization, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. 473 (2197) (2017) 20160446.
[43] H. Schaeffer, S.G. McCalla, Sparse model selection via integral terms, Phys. Rev. E 96 (2) (2017) 023302.
[44] H. Schaeffer, G. Tran, R. Ward, Learning dynamical systems and bifurcation via group sparsity, 2017, arXiv preprint arXiv:1709.01558.
[45] H. Schaeffer, G. Tran, R. Ward, Extracting sparse high-dimensional dynamics from limited data, SIAM J. Appl. Math. 78 (6) (2018) 3279–3295.
[46] H. Schaeffer, G. Tran, R. Ward, L. Zhang, Extracting structured dynamical systems using sparse optimization with very few samples, 2018, arXiv preprint arXiv:1805.04158.
[47] Y. Shin, K. Wu, D. Xiu, Sequential function approximation with noisy data, J. Comput. Phys. (2018).
[48] G.R. Shorack, J.A. Wellner, Empirical Processes with Applications to Statistics, Vol. 59, Siam, 2009.
[49] I. Steinwart, D. Hush, C. Scovel, Learning from dependent observations, J. Multivariate Anal. 100 (1) (2009) 175–194.
[50] A. Tikhonov, V.Y. Arsenin, John Wiley and Sons, Inc, 1977.
[51] G. Tran, R. Ward, Exact recovery of chaotic systems from highly corrupted data, Multiscale Model. Simul. 15 (3) (2017) 1108–1129.
[52] V.N. Vapnik, An overview of statistical learning theory, IEEE Trans. Neural Netw. 10 (5) (1999) 988–999.
[53] H. White, I. Domowitz, Nonlinear regression with dependent observations, Econometrica (1984) 143–161.
[54] W. Xu, E.-W. Bai, M. Cho, System identification in the presence of outliers and random noises: A compressed sensing approach, Automatica 50 (11) (2014) 2905–2911.
[55] C. Yao, E.M. Bollt, Modeling and nonlinear parameter estimation with kronecker product representation for coupled oscillators and spatiotemporal systems, Physica D 227 (1) (2007) 78–99.
[56] P. Zhang, L. Gan, C. Ling, S. Sun, Uniform recovery bounds for structured random matrices in corrupted compressed sensing, IEEE Trans. Signal Process. 66 (8) (2018) 2086–2097.
[57] L. Zhang, H. Schaeffer, On the convergence of the sindy algorithm, Multiscale Model. Simul. 17 (3) (2019) 948–972.