Bernoulli **22**(2), 2016, 1227–1277
DOI: 10.3150/14-BEJ691

Greedy algorithms for prediction

ALESSIO SANCETTA

Department of Economics, Royal Holloway University of London, Egham Hill, Egham TW20 0EX, UK. E-mail: asancetta@gmail.com; url: https://sites.google.com/site/wwwsancetta/

In many prediction problems, it is not uncommon that the number of variables used to construct a forecast is of the same order of magnitude as the sample size, if not larger. We then face the problem of constructing a prediction in the presence of potentially large estimation error. Control of the estimation error is either achieved by selecting variables or combining all the variables in some special way. This paper considers greedy algorithms to solve this problem. It is shown that the resulting estimators are consistent under weak conditions. In particular, the derived rates of convergence are either minimax or improve on the ones given in the literature allowing for dependence and unbounded regressors. Some versions of the algorithms provide fast solution to problems such as Lasso.

Keywords: Boosting; forecast; Frank–Wolfe Algorithm; Hilbert space projection; Lasso; regression function

1. Introduction

The goal of this paper is to address the problem of forecasting in the presence of many explanatory variables or individual forecasts. Throughout the paper, the explanatory variables will be referred to as regressors even when they are individual forecasts that we wish to combine or basis functions, or in general elements in some dictionary.

The framework is the one where the number of regressors is often large relatively to the sample size. This is quite common in many fields, for example, in macroeconomic predictions (e.g., Stock and Watson [64,65,66]). Moreover, when there is evidence of structural breaks, it is not always possible to use the full sample without making further assumptions. Indeed, it is often suggested to forecast using different sample sizes in an effort to mitigate the problem (e.g., Pesaran *et al.* [56], Pesaran and Picks [57]). When doing so, we still need to make sure that the forecasts built using smaller sample sizes are not too noisy.

For these reasons, it is critical to consider procedures that allow us to select and/or combine variables in an optimal way when the data are dependent. It is clear that in large-dimensional problems, variable selection via information criteria is not feasible, as it would require the estimation of a huge number of models. For example, if we are considering 100 regressors, naive model selection of a model with only 10 variables (i.e., an order of magnitude lower) would require estimation and comparison of $\binom{100}{10}$ models, which is in the order of billions.

This paper considers greedy algorithms to do automatic variable selection. There are many references related to the algorithms considered here (e.g., Bühlmann [15], Barron *et al.* [6], Huang, Cheang and Barron [40], Bühlmann and van de Geer [17]). These existing results are not applicable to standard prediction problems, as they assume i.i.d. random variable with bounded regressors and in some case bounded error terms.

Greedy algorithms have been applied to time series problems both in a linear and non-linear context (e.g., Audrino and Bühlmann [3,4], Audrino and Barone-Adesi [2], amongst others). However, to the author's knowledge, in the linear case, only Lutz and Bühlmann [49] derive consistency under strong mixing. There, no rates of convergence are given. (See Audrino and Bühlmann [4], for the non-linear case, again where no rates are given.) The above references only consider Boosting. It is known that other greedy algorithms possess better convergence rates (e.g., Barron *et al.* [6]). Here, only linear predictions are considered. Of course, when the regressors are a basis for some function space, the results directly apply to series estimators, hence, non-linear prediction (e.g., Mallat and Zhang [50], Daubechies, Defrise and De Mol [27], Barron *et al.* [6], Bühlmann and van de Geer [17], Sancetta [63], for more details along these lines).

To be precise, this paper shall consider greedy algorithms and provide rates of convergence which are best possible for the given set up or considerably improve on the existing ones, even under dependence conditions. The first algorithm is the L_2-Boosting studied by Bühlmann [15], also known as Projection Pursuit in signal processing (e.g., Mallat and Zhang [50]) and Pure Greedy Algorithm in approximation theory (e.g., DeVore and Temlyakov [29]). As mentioned above, it is routinely used in many applications, even in time series problems. The second algorithm is known as Orthogonal Greedy Algorithm (OGA) in approximation theory (e.g., DeVore and Temlyakov [29], Temlyakov [68]), and has also been studied in the statistical literature (Barron *et al.* [6]). It is the one the most resembles OLS estimation. The OGA is also reviewed in Bühlmann and van de Geer [17], where it is called Orthogonal Matching Pursuit (see also Zhang [81], Cai and Wang [23], for recent results). The third algorithm is a version of the Hilbert space projection algorithm studied by Jones [43] and Barron [5] with the version studied in this paper taken from Barron *et al.* [6], and called the Relaxed Greedy Algorithm (RGA). Adding a natural restriction to the RGA, the algorithm leads to the solution of the Lasso problem, which appears to be relatively new (see Sancetta [63]). This constrained version will be called Constrained Greedy Algorithm (CGA). Finally, closely related to the CGA is the Frank–Wolfe Algorithm (FWA) (see Frank and Wolfe [35], and Clarkson [26], Jaggi [42], and Freund, Grigas and Mazumder [36], for recent results). This selection seems to span the majority of known algorithms used in applied work.

The general problem of variable selection is often addressed relying on penalized estimation with an l_1 penalty. Greedy algorithms can be related to Lasso as they both lead to automatic variable selection. Algorithms that use a penalty in the estimation will not be discussed here. It is well known (Friedman *et al.* [37]) that the Lasso solution can be recovered via Pathwise Coordinate Optimization (a stagewise recursive algorithm), using the results of Tseng [71] (see also Daubechies, Defrise and De Mol [27], for related results). On the other hand, Huang, Cheang and Barron [40] have extended the RGA to the case of a Lasso penalty. (For recent advances on asymptotics for Lasso, the reader may consult Greenshtein and Ritov [39], Bunea, Tsybakov and Wegkamp [19], van de Geer [76], Huang, Cheang and Barron [40], Zhang [79], Belloni and Chernozhukov [10], Belloni *et al.* [9], amongst others.) Another related approach for variable selection under sparsity and design matrix constraints is via linear programming (e.g., Candes and Tao [24]).

One related question which is also considered here is the one of persistence as defined by Greenshtein and Ritov [39] and explored by other authors (e.g., Greenshtein [38], Bühlmann and van de Geer [17], Bartlett, Mendelson and Neeman [7]). This problems is of interest in a prediction context and relates to the idea of pseudo true value. Loosely speaking, one is interested in finding the largest class of linear models relative to which the estimator is still optimal in some sense. Here, it is shown that for mixing data, persistence holds for the class of linear models as large as the ones considered in Greenshtein and Ritov [39] and Bartlett, Mendelson and Neeman [7].

The focus of the paper is on prediction and consistency of the forecasts. Asymptotic normality of the estimators is not derived due to the weak conditions used (e.g., see Bühlmann [16], Nickl and van de Geer [53], van de Geer *et al.* [75], Zhang and Zhang [80] for results on statistical significance for high-dimensional, sparse models, under different estimation procedures and assumptions).

The paper is structured as follows. The remainder of this section defines the estimation setup, the objectives and the conditions to be used. Two different sets of dependence conditions are used: beta mixing, which gives the best convergence rates, and more general conditions allowing for non-mixing data and possibly long memory data. Section 2 starts with a summary of existing results comparing them with some of the ones derived here. The actual statement of all the results follows afterward. With the exception of the PGA, it is shown that the algorithms can achieve the minimax rate under beta mixing. However, for the PGA, the rates derived here considerably improve on the ones previously obtained. The algorithms are only reviewed later on in Section 2.3. The reader unfamiliar with these algorithms can browse through Section 2 right after Section 1 if needed. A discussion of the conditions and examples and applications of the results are given in Section 2.4. In particular, Section 2.4.3 gives examples of applications to long memory achieving convergence rates as good or better than the ones derived by other authors under i.i.d. observations, though requiring the population Gram matrix of the regressors to have full rank. In Section 3, details on implementation are given. Section 3 contains remarks of practical nature including vectorized versions of the algorithms, which are useful when implemented in scripting languages such as R and Matlab. This section also gives details on finite sample performance via simulation examples to complement the theoretical results. For example, the simulations in Section 3.3 show that – despite the slower rates of convergence – the PGA seems to perform particularly well when the signal to noise is low (see also Bühlmann and van de Geer [17], Section 12.7.1.1). The proofs are all in Section 4. Section 4 contains results on the approximation properties of the algorithms that can be of interest in their own. For example, a simple extension of the result in DeVore and Temlyakov [29] to statistical least square estimation is given in order to bound the approximation error of the PGA (L_2-Boosting). Moreover, it is also shown that the complexity of the PGA grows sub-linearly with the number of iteration, hence compensating this way for the higher approximation error (Lemma 8 in Section 4). This observation appears to be new and it is exploited when considering convergence under non-mixing data.

1.1. Estimation setup

There are possibly many more regressors than the sample size. However, most of the regressors are not needed or useful for prediction, for example, they may either be zero or have a progressively decreasing importance. This means that most of the regressors are redundant. Redundancy is formally defined in terms of a bound on the absolute sum of the regression coefficients. In particular, let $\mathcal{X}$ be a set of regressors of cardinality K, possibly much larger than the sample size n and growing with n if needed. Then the focus is on the linear regression function $\mu(x) = \sum_{k=1}^{K} b_k x^{(k)}$ where $\sum_{k=1}^{K} |b_k| \leq B < \infty$, and $x^{(k)}$ is the kth element in x. As B increases, the class of functions representable by μ becomes larger (e.g., when $\mathcal{X}$ is a set of functions whose linear span is dense in some space of functions). The same remark is valid when K grows with n, as for sieve estimators. The absolute summability of the regression coefficients is standard (e.g., Bühlmann [15], Barron *et al.* [6]). This restriction is also used in compressed sensing, where a signal with no noise admits a sparse representation in terms of a dictionary (e.g., Temlyakov [67], Chapter 5). Nevertheless, high-dimensional statistics also considers the problem of consistency when $B \to \infty$ at the rate $\mathrm{o}(\sqrt{n/\ln K})$ (e.g., Greenshtein and Ritov [39], Greenshtein [38], Bühlmann and van de Geer [17], Bartlett, Mendelson and Neeman [7]). Here, it is shown that Greedy algorithms are consistent in this situation when the data are dependent and the regressors are not necessarily bounded.

Notational details and conditions are introduced next. Given random variables Y and X, interest lies in approximating the conditional regression function $\mathbb{E}[Y|X] = \mu_0(X)$, with the linear regression $\mu(X) := \sum_{k=1}^{K} b_k X^{(k)}$, where $\sum_{k=1}^{K} |b_k| \leq B$, and $X^{(k)}$ denotes the kth element of X. Hence, μ_0 does not need to be linear. (Most of the literature, essentially, considers the case when the true regression function $\mu_0 \in \mathcal{L}(B)$ with Barron *et al.* [6] being one of the few exceptions.) Let $\{Y_i, X_i\colon\ i = 1, 2, \ldots, n\}$ be possibly dependent copies of Y, X. Define the empirical inner product

$$\langle Y, X^{(k)}\rangle_n := \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{(k)} \quad \text{and} \quad |X^{(k)}|_n^2 := \langle X^{(k)}, X^{(k)}\rangle_n.$$

To make sure that the magnitude of the regression coefficients is comparable, assume that $|X^{(k)}|_n^2 = 1$. This is a standard condition that also simplifies the discussion throughout (e.g., Bühlmann [15], Barron *et al.* [6]). In practice, this is achieved by dividing the original variables by $|X^{(k)}|_n$. Throughout, it is assumed that the variables have unit $|\cdot|_n$ norm. This also implies that $\mathbb{E}|X^{(k)}|_n^2 = 1$. Denote by

$$\mathcal{L}(B) := \left\{\mu\colon\ \mu(X) = \sum_{k=1}^{K} b_k X^{(k)}, \sum_{k=1}^{K} |b_k| \leq B, X \in \mathcal{X}\right\},$$

the space of linear functions on $\mathcal{X}$ with l_1 coefficients bounded by B. It follows that $\mathcal{L}(B)$ is a Hilbert space under the inner product $\langle X^{(k)}, X^{(l)}\rangle = \mathbb{E}X^{(k)}X^{(l)}$ as well as the empirical inner product $\langle X^{(k)}, X^{(l)}\rangle_n$. Also, let $\mathcal{L} := \bigcup_{B<\infty} \mathcal{L}(B)$ be the union of the above spaces for finite B. The goal it to estimate the regression function when the true expectation is replaced by the empirical one, that is, when we use a finite sample of n observations $\{Y_i, X_i\colon\ i = 1, 2, \ldots, n\}$. As already mentioned, B is only known to be finite, and this is a standard set up used elsewhere (e.g., Bühlmann [15], Barron *et al.* [6]). Moreover, μ_0 does not need to be an element of $\mathcal{L}(B)$ for any finite B.

Results are sometimes derived using some restricted eigenvalue condition on the empirical Gram matrix of the regressors also called compatibility condition (e.g., Bühlmann and van de Geer [17], for a list and discussion). For example, the minimum eigenvalue of the empirical Gram matrix of any possible m regressors out of the K possible ones, is given by

$$\rho_{m,n} := \inf\left\{\frac{|\sum_{k=1}^{K} X^{(k)} b_k|_n^2}{\sum_{k=1}^{K} |b_k|^2} \colon \sum_{k=1}^{K} \{b_k \neq 0\} = m\right\}, \tag{1}$$

where $\{b_k \neq 0\}$ is the indicator function of a set (e.g., Zhang [81], and many of the references on Lasso cited above; see also the isometry condition in Candes and Tao [24]). The above condition means that the regressors are approximately orthogonal, and typical examples are frames (e.g., Daubechies, Defrise and De Mol [27]). This condition is usually avoided in the analysis of convergence rates of greedy algorithms. Note that unless one uses a fixed design for the regressors, (1) is random. In this paper, m usually refers to the number of iterations or greedy steps at which the algorithm is stopped. The population counterpart of (1) will be denoted by ρ_m, that is,

$$\rho_m := \inf\left\{\frac{\mathbb{E}|\sum_{k=1}^{K} X^{(k)} b_k|_n^2}{\sum_{k=1}^{K} |b_k|^2} \colon \sum_{k=1}^{K} \{b_k \neq 0\} = m\right\}. \tag{2}$$

When m is relatively small, ρ_m plus an $o_p(1)$ term can be used to bound $\rho_{m,n}$ from below (e.g., Loh and Wainwright [48]; see also Nickl and van de Geer [53]). Eigenvalue restrictions will be avoided here under mixing dependent conditions. However, under non-mixing and possibly long memory conditions, the convergence rates can deteriorate quite quickly. Restricting attention to the case $\rho_m > 0$ allows one to derive more interesting results.

Throughout the paper, the following symbols are used: $\lesssim$ and $\gtrsim$ indicate inequality up to a multiplicative finite absolute constant, $\asymp$ when the left-hand side and the right-hand side are of the same order of magnitude, $\wedge$ and $\vee$ are min and max, respectively, between the left-hand side and the right-hand side.

1.2. Objective

To ease notation, let $|\cdot|_2 = (\mathbb{E}|\cdot|^2)^{1/2}$ and define

$$\gamma(B) := \inf_{\mu \in \mathcal{L}(B)} |\mu - \mu_0|_2 \tag{3}$$

to be the approximation error of the best element in $\mathcal{L}(B)$, and let μ_B be the actual minimizer. Since for each $B < \infty$ the set $\mathcal{L}(B)$ is compact, one can replace the inf with min in the above display. The approximation can improve if B increases. For simplicity, the notation does not make explicit the dependence of the approximation error on K, as K is the same for all the algorithms, while B can be different for the CGA and FWA, as it will be shown in due course.

Let X' be a random variable distributed like X but independent of the sample. Let $\mathbb{E}'$ be expectation w.r.t. X' only. The estimator from any of the greedy algorithms will be denoted by F_m. The bounds are of the following kind:

$$\left(\mathbb{E}'\left|\mu_0\left(X'\right) - F_m\left(X'\right)\right|^2\right)^{1/2} \lesssim \operatorname{error}(B, K, n, m) + \operatorname{algo}(B, m) + \gamma(B) \tag{4}$$

for any B in some suitable range, where relates to the B in the approximation μ_B from (3). The possible values of B depend on the algorithm. For the PGA, OGA and RGA, $B < \infty$, that is, the algorithms allow to approximate any function in $\mathcal{L}$, the union of $\mathcal{L}(B)$ for any $B > 0$. The CGA and FWA restrict $B \leq \bar{B}$ which is a user specified parameter. This gives direct control of the estimation error. The results for the CGA and FWA will be stated explicitly in $\bar{B}$, so that $\bar{B} \to \infty$ is allowed. The term $\gamma(B)$ is defined in (3), while

$$\operatorname{algo}(B, m)^2 \gtrsim |Y - F_m|_n^2 - \inf_{\mu \in \mathcal{L}(B)} |Y - \mu|_n^2$$

defines an upper bound for the error due to estimating using any of the algorithms rather than performing a direct optimization. It could be seen as part of the approximation error, but to clearly identify the approximation properties of each algorithm, $\operatorname{algo}(B, m)$ is explicitly defined. Finally, the term $\operatorname{error}(B, K, n, m)$ is the estimation error.

1.3. Approximation in function spaces

When $\mu_0 \notin \mathcal{L}$, the approximation can be large. This is not to say that functions in $\mathcal{L}$ cannot represent non-linear functions. For example, the set of regressors $\mathcal{X}$ could include functions that are dense in some set, or generally be a subset of some dictionary (e.g., Mallat and Zhang [50], Barron *et al.* [6], Sancetta [63]).

Consider the framework in Section 2.3 of Barron *et al.* [6]. Let μ_0 be a univariate function on $[0, 1]$, that is, μ_0 is the expectation of Y conditional on a univariate variable with values in $[0, 1]$. Suppose $\mathcal{D}$ is a dictionary of functions on $[0, 1]$, and denote its elements by g. Suppose that μ_0 is in the closure of functions admitting the representation $\mu(x) = \sum_{g \in \mathcal{D}} b_g g(x)$, where $\sum_{g \in \mathcal{D}} |b_g| \leq B$; b_g are coefficients that depend on the functions g. Examples include sigmoid functions, polynomials, curvelet, frames, wavelets, trigonometric polynomials, etc. Since $\mathcal{D}$ might be infinite or too large for practical applications, one considers a subset $\mathcal{X} \subset \mathcal{D}$, which is a dictionary of K functions on $[0, 1]$. Then $\mu_0(x) = \sum_{g \in \mathcal{X}} b_g g(x) + \sum_{g \in \mathcal{D} \setminus \mathcal{X}} b_g g(x)$. Assuming that $|\sum_{g \in \mathcal{D} \setminus \mathcal{X}} b_g g(x)|_2 \lesssim K^{-\alpha}$ for some $\alpha > 0$, the approximation error decreases as one expands the dictionary. Examples for non-orthogonal dictionaries are discussed in Barron *et al.* [6]. However, to aid intuition, one can consider Fourier basis for smooth enough functions to ensure that $\sum_{g \in \mathcal{D}} |b_g| < \infty$. If $\mathcal{X}$ is large enough, one may expect the second summation to have a marginal contribution.

Hence, with abuse of notation, the result of the present paper cover the aforementioned problem, where the functions $g \in \mathcal{X}$ are then denoted by $\{x^{(k)}\colon k = 1, 2, \ldots, K\}$; here $x \in [0, 1]$, while each $g(x)$ is denoted by $x^{(k)}(x)$, so that $x^{(k)}$ is not the kth entry in x but a function of x (the kth element in a dictionary). As mentioned in the Introduction, this paper does not make any distinction whether $\mathcal{X}$ is a set of explanatory variables or functions (in general a dictionary), so it also covers problems addressed in compress sensing with error noise.

1.4. Conditions

The theoretical properties of the algorithms are a function of the dependence conditions used. At first, absolute regularity is used. This allows to obtain results as good as if the data were independent (e.g., Chen and Shen [25]). However, for some prediction problems, absolute regularity might not be satisfied. Hence, more general dependence conditions shall be used. Generality comes at a big cost in this case.

Some notation is needed to recall the definition of absolute regularity. Suppose that $(W_i)_{i\in\mathbb{Z}}$ is a stationary sequence of random variables and, for any $d \geq 0$, let $\sigma(W_i\colon i \leq 0)$, $\sigma(W_i\colon i \geq d)$ be the sigma algebra generated by $\{W_i\colon i \leq 0\}$ and $\{W_i\colon i \geq d\}$, respectively. For any $d \geq 0$, the beta mixing coefficient $\beta(d)$ for $(W_i)_{i\in\mathbb{Z}}$ is

$$\beta(d) := \mathbb{E} \sup_{A\in\sigma(W_i\colon i\geq d)} \bigl|\Pr\bigl(A|\sigma(W_i\colon i \leq 0)\bigr) - \Pr(A)\bigr|$$

(see Rio [60], Section 1.6, for other equivalent definitions). The sequence $(W_i)_{i\in\mathbb{Z}}$ is absolutely regular or beta mixing if $\beta(d) \to 0$ for $d \to \infty$.

Throughout, with slight abuse of notation, for any $p > 0$, $|\cdot|_p^p = \mathbb{E}|\cdot|^p$ is the L_p norm (i.e., do not confuse $|\cdot|_n$ with $|\cdot|_p$). Moreover, $\mu_0(X) := \mathbb{E}[Y|X]$ is the true regression function, $Z := Y - \mu_0(X)$ is the error term, $\Delta(X) = \mu_B(X) - \mu_0(X)$ is the approximation residual (recall that μ_B is the best L_2 approximation to μ_0 in $\mathcal{L}(B)$).

The asymptotics of the greedy algorithms are studied under the following conditions.

Condition 1. $\max_k |X^{(k)}|_n^2 = 1$, $\max_k |X^{(k)}|_2 = 1$.

Condition 2. *The sequence $(X_i, Z_i)_{\in\mathbb{Z}}$ is stationary absolutely regular with beta mixing coefficients $\beta(i) \lesssim \beta^i$ for some $\beta \in [0, 1)$ and $\mathbb{E}|Z|^p < \infty$ for some $p > 2$, $\max_{k\leq K} |X^{(k)}|$ is bounded, and the approximation residual $\Delta(X) = \mu_B(X) - \mu_0(X)$ is also bounded. Moreover, $1 < K \lesssim \exp\{Cn^a\}$, for some absolute constant C and $a \in [0, 1)$.*

Bounded regressors and sub-Gaussian errors are the common conditions under which greedy algorithms are studied. Condition 2 already weakens this to the error terms only possessing a $p > 2$ moment. However, restricting attention to bounded regressors can be limiting. The next condition replaces this with a moment condition.

Condition 3. *The sequence $(X_i, Z_i)_{\in\mathbb{Z}}$ is stationary absolutely regular with beta mixing coefficients $\beta(i) \lesssim \beta^i$ for some $\beta \in [0, 1)$ and*

$$\mathbb{E}\bigl|ZX^{(k)}\bigr|^p + \mathbb{E}\bigl|X^{(k)}\bigr|^{2p} + \mathbb{E}\bigl|\Delta(X)X^{(k)}\bigr|^p < \infty, \tag{5}$$

for some $p > 2$. Moreover, $1 < K \lesssim n^\alpha$ for some $\alpha < (p - 2)/2$ (with p as just defined).

Note that in the case of independent random variables, one could relax the moment condition to $p \geq 2$. Recall that μ_0 is not restricted to be in $\mathcal{L}(B)$. Only the resulting estimator will be. The expectation of $\Delta(X)$ is the bias.

There are examples of models that are not mixing (e.g., Andrews [1], Bradley [13]). For example, the sieve bootstrap is not mixing (Bickel and Bühlmann [11]). It is important to extend the applicability of the algorithms to such case. The gain in generality leads to a considerably slower rate of convergence than the i.i.d. and beta mixing case. This is mostly due to the method of proof. It is not known whether the results can be improved in such cases. Dependence is now formalized by the following.

Condition 4. *Denote by $\mathbb{E}_0$ the expectation conditional at time* 0 (*w.r.t. the natural filtration of the random variables*). *Recall that $|\cdot|_p := (\mathbb{E}|\cdot|^p)^{1/p}$. The sequence $(X_i, Z_i)_{\in\mathbb{Z}}$ is stationary, and for some $p \geq 2$,*

$$d_{n,p} := \max_k \sum_{i=0}^{n} \frac{(|\mathbb{E}_0 Z_i X_i^{(k)}|_p + |\mathbb{E}_0[(1-\mathbb{E})|X_i^{(k)}|^2]|_p + |\mathbb{E}_0[(1-\mathbb{E})\Delta(X_i)X_i^{(k)}]|_p)}{(i+1)^{1/2}} < \infty$$

for any n.

Note that the dependence condition is in terms of mixingales and for weakly dependent data, $\sup_n d_{n,p} < \infty$ when the pth moment exists, under certain conditions. The general framework allows us to consider data that might be strongly dependent (long memory), when $d_{n,p} \to \infty$ (see Example 4 for some details).

2. Algorithms

The algorithms have already appeared elsewhere, and they will be reviewed in Section 2.3. All the algorithms studied here do achieve a global minimum of the empirical risk. This minimum might not be unique if the number of variables are larger than the sample size. Moreover, the convergence rates of the algorithms to the global minimum can differ. The reader unfamiliar with them, can skim through Section 2.3 before reading the following. In particular, the PGA has the slowest rate, while all the others have a faster rate which is essentially optimal (see Lemmas 4, 5, 6 and 7, for the exact rates used here; see DeVore and Temlyakov [29], and Barron *et al.* [6], for discussions on optimality of convergence rates). The optimal rate toward the global minimum is $m^{-1/2}$ under the square root of the empirical square error loss, where m is the number of greedy iterations. For the PGA the convergence rate of the approximation error of the algorithm, algo(B, m), is only $m^{-1/6}$, without requiring the target Y to be itself an element of $\mathcal{L}(B)$, that is, a linear function with no noise (Lemma 4). For functions in $\mathcal{L}(B)$, Konyagin and Temlyakov [45] improved the rate to $m^{-11/62}$, while Livshitz and Temlyakov [47] show a lower bound $m^{-0.27}$. Hence, the approximation rate of the PGA is an open question. The slow rate of the PGA (L_2-Boosting) has led Barron *et al.* [6] to disregard it. While the approximating properties of the PGA are worse than the other algorithms, its finite sample properties tend to be particularly good in many cases (e.g., Section 3.3). An overview of how the present results add to the literature and further details are summarized next.

2.1. Comparison with existing results

There are many results on greedy algorithms under different conditions. Table 1 summarizes and compares some of these results. For each algorithm the most interesting results from the present paper are presented first. The symbols used to describe the conditions are defined in the glossary of symbols at the end of this section.

Table 1. Comparison of results

Algorithm/author/conditions	Rates
PGA	
M$(X;b)$, M$(Z;p)$, D$(X,Z;\beta^n)$, K(E), L_2	$(\frac{\ln K}{n})^{1/8}$
M$(X,Z;g)$, D$(X,Z;$NM$)$, K(P), X, L_2	$(\frac{d^2_{n,\bar{p}}}{n})^{(1-\epsilon)/8}$
Bühlmann and van de Geer [17]	
M$(X;b)$, M$(Z;g)$, D$(X,Z;iid)$, K(E), L_{2n}	$(\frac{\ln K}{n})^{(1-\epsilon)/16}$
Lutz and Bühlmann [49]	
M$(X,Z;p)$, D$(X,Z;n^\alpha)$, K(E), L_2	o(1)
OGA	
M$(X;b)$, M$(Z;p)$, D$(X,Z;\beta^n)$, K(E), L_2	$(\frac{\ln K}{n})^{1/4}$
M$(X,Z;g)$, D$(X,Z;$NM$)$, K(P), X, L_2	$(\frac{d^2_{n,\bar{p}}}{n})^{(1-\epsilon)/6}$
Bühlmann and van de Geer [17]	
M$(X;b)$, M$(Z;g)$, D$(X,Z;iid)$, K(E), L_2	$(\frac{1}{n})^{1/6} \vee (\frac{\ln K}{n})^{1/4}$
Barron *et al.* [6]	
M$(X,Z;b)$, D$(X,Z;iid)$, K(P), EL_2	$(\frac{\ln K}{n})^{1/4}$
Zhang [81]	
M$(X,Z;b)$, D$(X,Z;iid)$, X, L_{2n}, $+$	$(\frac{K_0}{n})^{1/2}$
RGA	
M$(X;b)$, M$(Z;p)$, D$(X,Z;\beta^n)$, K(E), L_2	$(\frac{\ln K}{n})^{1/4}$
M$(X,Z;g)$, D$(X,Z;$NM$)$, K(P), X, L_2	$(\frac{d^2_{n,\bar{p}}}{n})^{(1-\epsilon)/6}$
Barron *et al.* [6]	
M$(X,Z;b)$, D$(X,Z;iid)$, K(P), EL_2	$(\frac{\ln K}{n})^{1/4}$
CGA and FWA	
M$(X;b)$, M$(Z;p)$, D$(X,Z;\beta^n)$, K(E), L_2, $+$	$(\frac{\ln K}{n})^{1/4}$
M$(X,Z;g)$, D$(X,Z;$NM$)$, K(P), L_2, $+$	$(\frac{d^2_{n,\bar{p}}}{n})^{(1-\epsilon)/4}$

2.1.1. *Glossary for Table 1*

2.1.1.1. *Moments.* M (variable; moment type); moment types: $p =$ moments, refer to paper for exact p, $g =$ sub-Gaussian tails, $b =$ bounded random variables; for example, $\mathrm{M}(X, Z; b)$ means that both X and Z are bounded.

2.1.1.2. *Dependence.* D (variable, dependence type); dependence types are all stationary: $iid =$ i.i.d. or just independence, $\alpha^n/\beta^n =$ geometric alpha/beta mixing, $n^\alpha/n^\beta =$ polynomial alpha/beta mixing; NM = non-mixing; see paper for details on the polynomial rate and how it relates to moments.

2.1.1.3. *K.* K (growth rate); number of regressors K: $P = n^a$ for any $a < \infty$, $E = \exp\{Cn^a\}$ for $a \in [0, 1)$, $C < \infty$.

2.1.1.4. *Design matrix.* X if conditions are imposed on the design matrix, for example, compatibility conditions, otherwise, no symbol is reported.

2.1.1.5. *Loss function.* $L_2 = L_2$ loss as in the l.h.s. of (4) and results holding in probability, $EL_2 =$ same as L_2 but results holding in L_1 (i.e., take a second expectation w.r.t. to the sample data), $L_{2n} =$ empirical L_2 loss.

2.1.1.6. *Additional remarks on glossary.* The true function μ_0 is assumed to be in $\mathcal{L}(B)$ for some finite B. When rates are included, ϵ is understood to be a positive arbitrarily small constant. Also, $\bar{p}$ in $d_{n,\bar{p}}$ refers to a large p depending on ϵ and K, with exact details given in Corollary 5. In some cases, conditions may not fit exactly within the classification given above due to minor differences, in which case they may still be classified within one group. The symbol $+$ is used to denote additional conditions which can be found in the cited paper. For Zhang [81], K_0 represents the true number of non-zero coefficients and it is supposed to be small. For the CGA and the FWA, the symbol $+$ refers to the fact that the user pre-specifies a $\bar{B} < \infty$ and constrains estimation in $\mathcal{L}(B)$ with $B \leq \bar{B}$, and it also assumes that $\mu_0 \in \mathcal{L}(\bar{B})$. The results in the paper are more general, and the restrictions in Table 1 are for the sake of concise exposition and comparison.

2.1.2. *Comments*

Table 1 only provides upper bounds. Interest would also lie in deriving lower bound estimates (e.g., Donoho and Johnstone [30], Birgé and Massart [12], Tsybakov [72], and Bunea, Tsybakov and Wegkamp [20], for such rates for certain nonparametric parametric problems; see also Tsybakov [73], Chapter 2, for a general discussion on lower bounds). The results in Tsybakov [72] and Bunea, Tsybakov and Wegkamp [20] provide minimax rates and explicit estimators for certain function classes which exactly apply in the present context. Suppose that the error term Z is Gaussian, the regressors X are bounded and an i.i.d. sample is available. Let μ_n be any estimator in $\mathcal{L}(B)$. From Theorem 2 in Tsybakov [72], one can deduce that

$$\sup_{\mu\in\mathcal{L}(B)} |\mu-\mu_n| \gtrsim \begin{cases} B\sqrt{\dfrac{K}{n}}, & \text{if } K \lesssim \sqrt{n}, \\ B\left(\dfrac{\ln K}{n}\right)^{1/4}, & \text{if } K \gtrsim \sqrt{n}. \end{cases}$$

This results is also useful to understand the difference between the result derived by Zhang [81] for the OGA and usual results for Lasso under sparsity. In these cases, the target function is in a much smaller class than $\mathcal{L}(B)$, that is, μ_0 is a linear function with a small number of K_0 non-zero regression coefficients. Within this context, one can infer that the result from Zhang [81] is the best possible (e.g., use Theorem 3 in Tsybakov [72]).

Under mixing conditions, the convergence rates for the OGA, RGA, CGA and FWA are optimal. Table 1 shows that the results in Barron *et al.* [6] for the OGA and RGA are also optimal, but require i.i.d. bounded regressors and noise. The convergence rates for the PGA are not optimal, but considerably improve the ones of Bühlmann and van de Geer [17] also allowing for unbounded regressors and dependence.

2.2. Statement of results

2.2.1. *Mixing data*

In the following, when some relation is said to hold in probability, it means it holds with probability going to one as $n \to \infty$. Also, note that the linear projection of μ_0 onto the space spanned by the regressors is in $\mathcal{L}$ (the union of the $\mathcal{L}(B)$ spaces) because the number of regressors K is finite. Hence, let

$$B_0 := \arg\inf_{B>0} \gamma(B) \tag{6}$$

be the absolute sum of the coefficients in the unconstrained linear projection of μ_0 onto the space spanned by the regressors ($\gamma(B)$ as in (3)). Of course, K is allowed to diverge to infinity with n, if needed, which in consequence may also imply B_0 in (6) can go to infinity.

Theorem 1. *Under Condition 1 and either Conditions 2 or 3,*

$$\left(\mathbb{E}'\left|\mu_0(X') - F_m(X')\right|^2\right)^{1/2} \lesssim \mathrm{error}(B, K, n, m) + \mathrm{algo}(B) + \gamma(B) \tag{7}$$

in probability, where

$$B \in \begin{cases} [B_0, \infty), & \textit{for the PGA, OGA, RGA}, \\ (0, \bar{B}], & \textit{for the CGA and FWA}, \end{cases} \tag{8}$$

where

$$\mathrm{error}(B,K,n,m)=\begin{cases}\sqrt{\dfrac{m\ln K}{n}}, & \text{for the PGA, OGA, RGA,}\\ \bar{B}\left(\dfrac{\ln K}{n}\right)^{1/4}, & \text{for the CGA and FWA,}\end{cases} \tag{9}$$

$$\mathrm{algo}(B,m)=\begin{cases}B^{1/3}m^{-1/6}, & \text{for the PGA,}\\ Bm^{-1/2}, & \text{for the OGA and RGA,}\\ \bar{B}m^{-1/2}, & \text{for the CGA and FWA.}\end{cases} \tag{10}$$

Remark 1. When $B_0 \le \bar{B}$, asymptotically, the CGA and FWA impose no constraint on the regression coefficients. In this case, these algorithms also satisfy (7) with (8) as for the OGA and RGA. While B_0 is unknown, this observation will be used to deduce Corollary 2. Also note that (7) for the PGA, OGA and RGA is minimized by $B = B_0$.

Theorem 1 allows one to answer several questions of interest about the algorithms. Note that $\mathrm{error}(B,K,n,m)$ in (9) does not depend on B, as a consequence of the method of proof; it will depend on B for some of the other results. The next two results will focus on two related important questions. One concerns the overall convergence rates of the estimator when the true function $\mu_0 \in \mathcal{L}$, that is, μ_0 is linear with absolutely summable coefficients. The other concerns the largest linear model in reference of which the estimator is optimal in a square error sense (i.e., persistence in the terminology of Greenshtein and Ritov [39], or traditionally, this is termed consistency for the linear pseudo true value). Rates of convergence are next. These rates directly follow from Theorem 1, using the fact that $B < \infty$ and equating $\mathrm{error}(B,K,n,m)$ with $\mathrm{algo}(B,m)$ and solving for m.

Corollary 1. *Under the conditions of Theorem 1, if*

$$m \text{ satisfies }\begin{cases}\asymp\left(\dfrac{n}{\ln K}\right)^{3/4}, & \text{for the PGA,}\\ \asymp\sqrt{\dfrac{n}{\ln K}}, & \text{for the OGA and RGA,}\\ \gtrsim\sqrt{\dfrac{n}{\ln K}}, & \text{for the CGA and FWA}\end{cases}$$

then, in probability

$$\left(\mathbb{E}'\left|\mu_0(X') - F_m(X')\right|^2\right)^{1/2} \lesssim \begin{cases}\left(\dfrac{\ln K}{n}\right)^{1/8}, & \text{for the PGA if } \mu_0\in\mathcal{L},\\ \left(\dfrac{\ln K}{n}\right)^{1/4}, & \text{for the OGA and RGA if } \mu_0\in\mathcal{L},\\ \bar{B}\left(\dfrac{\ln K}{n}\right)^{1/4}, & \text{for the CGA and FWA if } \mu_0\in\mathcal{L}(\bar{B}).\end{cases}$$

The CGA and FWA achieve the minimax rate under either Conditions 2 or 3 if $\mu_0 \in \mathcal{L}(\bar{B})$ as long as the number of iterations m is large enough. The drawback in fixing $\bar{B}$ is that if $\mu_0 \in \mathcal{L}(B)$ with $\bar{B} < B$, there can be an increase in bias. This can be avoided by letting $\bar{B} \to \infty$ with the sample size. The following can then be used to bound the error when $\bar{B} < B$, if $\mu_0 \in \mathcal{L}(B)$ (Sancetta [63]).

Lemma 1. *Let $\mu \in \mathcal{L}(B)$ for some $B < \infty$. Then*

$$\inf_{\mu' \in \mathcal{L}(B')} \left|\mu - \mu'\right|_2 \leq \max\left\{B - B', 0\right\}.$$

The bounds are explicit in $\bar{B}$ so that one can let $\bar{B} \to \infty$ if needed and apply Lemma 1 to show that the approximation error goes to zero if $\mu_0 \in \mathcal{L}(B)$ for some bounded B.

Next, one can look at the idea of persistence, which is also related to consistency of an estimator for the pseudo true value in the class of linear functions. Adapting the definition of persistence to the set up of this paper, the estimator F_m is persistent at the rate $B \to \infty$ if

$$\mathbb{E}'\left|Y' - F_m\left(X'\right)\right|^2 - \inf_{\mu \in \mathcal{L}(B)} \mathbb{E}'\left|Y' - \mu\left(X'\right)\right|^2 = \mathrm{o}_p(1), \tag{11}$$

where X' and Y' are defined to have same marginal distribution as the X_i's and Y_i's, but independent of them. Directly from Theorem 1 deduce the following.

Corollary 2. *Let $\bar{B} = B$ for the CGA and FWA. Under the conditions of Theorem* 1, (11) *holds if $m \to \infty$ such that $m = \mathrm{o}(n/\ln K)$ and $B = \mathrm{o}(\sqrt{m})$ for all algorithms.*

2.2.2. *Non-mixing and strongly dependent data*

In the non-mixing case, the rates of convergence of the estimation error can quickly deteriorate. Improvements can then be obtained by restricting the population Gram matrix of the regressors to be full rank. The next result does not restrict ρ_m.

Theorem 2. *Under Conditions* 1 *and* 4, (7) *holds in probability, with*

$$\operatorname{error}(B, K, n, m) = \left(\frac{d_{n,p}^2 K^{4/p}}{n}\right)^{1/4} \times \begin{cases} \left(B + m^{1/2}\right), & \text{for the PGA}, \\ (B + m), & \text{for the OGA, RGA}, \\ \bar{B}, & \text{for the CGA and FWA} \end{cases}$$

and

$$B \in \begin{cases} (0, \infty), & \text{for the PGA, OGA, RGA}, \\ (0, \bar{B}], & \text{for the CGA and FWA} \end{cases}$$

and $\operatorname{algo}(B, m)$ *as in* (10).

Unlike $\operatorname{error}(B, K, n, m)$ in (9) which did not depend on B, the above is derived using a different method of proof and does depend on B. Also note the different restriction on B. Letting $\mu_0 \in \mathcal{L}$, one obtains the following explicit convergence rates.

Corollary 3. *Suppose that*

$$m \text{ satisfies } \begin{cases} \asymp \left(\dfrac{n}{d_{n,p}^2 K^{4/p}}\right)^{3/8}, & \text{for the PGA,} \\ \asymp \left(\dfrac{n}{d_{n,p}^2 K^{4/p}}\right)^{1/12}, & \text{for the OGA and RGA,} \\ \gtrsim \left(\dfrac{n}{d_{n,p}^2 K^{4/p}}\right)^{1/8}, & \text{for the CGA and FWA.} \end{cases} \tag{12}$$

Under the conditions of Theorem 2, in probability,

$$\left(\mathbb{E}'\left|\mu_0(X') - F_m(X')\right|^2\right)^{1/2} \lesssim \begin{cases} \left(\dfrac{d_{n,p}^2 K^{4/p}}{n}\right)^{1/16}, & \text{for the PGA if } \mu_0 \in \mathcal{L}, \\ \left(\dfrac{d_{n,p}^2 K^{4/p}}{n}\right)^{1/12}, & \text{for the OGA and RGA if } \mu_0 \in \mathcal{L}, \\ \left(\dfrac{d_{n,p}^2 K^{4/p}}{n}\right)^{1/4}, & \text{for the CGA and FWA if } \mu_0 \in \mathcal{L}(\bar{B}). \end{cases}$$

The results are close to the lower bound $\mathrm{O}(n^{-1/4})$ only for the CGA and FWA under weak dependence, not necessarily mixing data (i.e., $\sup_n d_{n,p} < \infty$) and variables with moments of all orders (i.e., p arbitrary large). Now, restrict attention to $\rho_K > 0$, that is, ρ_m in (2) with $m = K$. This is equivalent to say that the population Gram matrix of the regressors has full rank. In this case, the results for the PGA, OGA and RGA can be improved. By following the proofs in Section 4, it is easy to consider ρ_m going to zero as $m \to \infty$, but at the cost of extra details, hence these case will not be reported here.

Theorem 3. *Suppose that* $\rho_K > 0$. *Under Conditions 1 and 4, for the PGA, OGA and RGA, (7) holds in probability with*

$$\operatorname{error}(B, K, n, m) = \left(m + m^{1/2} B\right)\left(\frac{d_{n,p}^2 K^{4/p}}{n}\right)^{1/2}$$

for any positive B, *and* $\operatorname{algo}(B, m)$ *as in (10), as long as* $\operatorname{error}(B, K, n, m) + \operatorname{algo}(B, m) = \mathrm{o}(1)$.

The above theorem leads to much better convergence rates.

Corollary 4. *Suppose that*

$$m \asymp \begin{cases} \left(\dfrac{n}{d_{n,p}^2 K^{4/p}}\right)^{3/4}, & \text{for the PGA,} \\ \left(\dfrac{n}{d_{n,p}^2 K^{4/p}}\right)^{1/3}, & \text{for the OGA and RGA.} \end{cases} \tag{13}$$

Under the conditions of Theorem 3,

$$\left(\mathbb{E}'\left|\mu_0(X') - F_m(X')\right|^2\right)^{1/2} \lesssim \begin{cases} \left(\dfrac{d_{n,p}^2 K^{4/p}}{n}\right)^{1/8}, & \textit{for the PGA if } \mu_0 \in \mathcal{L}, \\ \left(\dfrac{d_{n,p}^2 K^{4/p}}{n}\right)^{1/6}, & \textit{for the OGA and RGA if } \mu_0 \in \mathcal{L}. \end{cases}$$

Under non-mixing dependence, deterioration in the rate of convergence due to K becomes polynomial rather than the logarithmic one of Theorem 1. On the positive side, the dependence condition used is very simple and can be checked for many models (e.g., Doukhan and Louhichi [31], Section 3.5, Dedecker and Doukhan [28], for examples and calculations). Interesting results can be deduced when the regressors have a moment generating function. Then the rates of convergence can be almost as good if not better than the ones derived by other authors assuming i.i.d. data, though only when $\rho_K > 0$ holds.

Corollary 5. *Suppose that X and Z have moments of all order and $K \lesssim n^\alpha$ for some $\alpha \in \mathbb{N}$. Under Conditions* 1 *and* 4, *choosing m as in* (13) *for the PGA*, *OGA and RGA and as in* (12) *for the CGA and FWA*, *for any $\epsilon \in (0,1)$*, *and $p = 4\alpha/\epsilon$*,

$$\left(\mathbb{E}'\left|\mu_0(X') - F_m(X')\right|^2\right)^{1/2} \lesssim \begin{cases} \left(\dfrac{d_{n,p}^2}{n}\right)^{(1-\epsilon)/8}, & \textit{for the PGA if } \mu_0 \in \mathcal{L} \textit{ and } \rho_K > 0, \\ \left(\dfrac{d_{n,p}^2}{n}\right)^{(1-\epsilon)/6}, & \textit{for the OGA and RGA if } \mu_0 \in \mathcal{L} \textit{ and } \rho_K > 0, \\ \left(\dfrac{d_{n,p}^2}{n}\right)^{(1-\epsilon)/4}, & \textit{for the CGA and FWA if } \mu_0 \in \mathcal{L}(\bar{B}) \end{cases}$$

in probability.

2.3. Review of the algorithms

The algorithms have been described in several places in the literature. The following sections review them. The first two algorithms are boosting algorithms and they are reviewed in Bühlmann and van de Geer [17]. The third algorithm has received less attention in statistics despite the fact that it has desirable asymptotic properties (Barron *et al.* [6]). The fourth algorithm is a constrained version of the third one and further improves on it in certain cases. The fifth and last algorithm is the basic version of the Frank–Wolfe [35] algorithm.

2.3.1. *Pure Greedy Algorithm* (*a.k.a. L_2-Boosting*)

Boosting using the L_2 norm is usually called L_2-Boosting, though some authors also call it Pure Greedy Algorithm (PGA) in order to stress its origin in the approximation theory literature (e.g.,

```
Set:
$m \in \mathbb{N}$
$F_0 := 0$
$\nu \in (0, 1]$
For: $j = 1, 2, \ldots, m$
    $s(j) := \arg\max_k |\langle Y - F_{j-1}, X^{(k)} \rangle_n|$
    $g_j(X) := \langle Y - F_{j-1}, X^{s(j)} \rangle_n X^{s(j)}$
    $F_j := F_{j-1} + \nu g_j(X)$
```

Figure 1. PGA (L_2-Boosting).

Barron *et al.* [6]), and this is how it will be called here. The term matching pursuit is also used by engineers (e.g., Mallat and Zhang [50]). Figure 1 recalls the algorithm. There, $\nu \in (0, 1]$ is the shrinkage parameter and it controls the degree of greediness in the algorithm. For example, as $\nu \to 0$ the algorithm in Figure 1 converges to Stagewise Linear Regression, a variant of the LARS algorithm that has striking resemblance to Lasso (Efron *et al.* [34], for details). In order to avoid ruling out good regressors that are correlated to $X^{s(m)}$ ($s(m)$ as defined in Figure 1 and $X^{s(m)} = X^{(s(m))}$ throughout to ease notation) one chooses ν smaller then 1, usually 0.1 (Bühlmann [15]).

The PGA recursively fits the residuals from the previous regression to the univariate regressor that reduces the most the residual sum of the squares. At each step j, the algorithm solves $\min_{k,b} |Y - F_{j-1} - X^{(k)} b|_n^2$. However, the coefficient can then be shrunk by an amount $\nu \in (0, 1)$ in order to reduce the degree of greediness. The resulting function F_m is an element of $\mathcal{L}(B_m)$ for some $B_m = \mathrm{O}(m^{1/2})$ (Lemma 8). The algorithm is known not to possess as good approximation properties as the other algorithms considered in this paper. However, this is compensated by B_m not growing too fast, hence, also the estimation error does not grow too fast.

2.3.2. *Orthogonal Greedy Algorithm* (*a.k.a. Orthogonal Matching Pursuit*)

The Orthogonal Greedy Algorithm (OGA) (e.g., Barron *et al.* [6]) is also known as Orthogonal Matching Pursuit. Figure 2 recalls that the OGA finds the next regressor to be included based on the same criterion as for PGA, but at each m iteration, it re-estimates the regression coefficients

```
Set:
$m \in \mathbb{N}$
$F_0 := 0$
For: $j = 1, 2, \ldots, m$
    $s(j) := \arg\max_k |\langle Y - F_{j-1}, X^{(k)} \rangle_n|$
    $P_X^j :=$ OLS operator on $\mathrm{span}\{X^{s(1)}, X^{s(2)}, \ldots, X^{s(j)}\}$
    $F_j := P_X^j Y$
```

Figure 2. OGA (Orthogonal Matching Pursuit).

by OLS using the selected regressors. For convenience, the OLS projection operator is defined by P_X^m where the m stresses that one is only using the regressors included up to iteration m, that is, $P_X^m Y = \sum_{k=1}^m b_{kn} X^{s(k)}$ for OLS coefficients b_{kn}'s. Hence, in some circumstances, the OGA is too time consuming, and may require the use of generalized inverses when regressors are highly correlated. However, Pati, Rezaiifar and Krishnaprasad [54] give a faster algorithm for its estimation.

2.3.3. *Relaxed Greedy Algorithm*

The Relaxed Greedy Algorithm (RGA) is a less popular method, which however has the same estimation complexity of the PGA. It is reviewed in Figure 3. The RGA updates taking a convex combination of the existing regression function with the new predictor. The RGA does not shrink the estimated coefficient at each step, but does shrink the regression from the previous iteration $j-1$ by an amount $1-w_j$, where $w_j = j^{-1}$. Other weighting schemes such that $w_j \in (0,1)$ and $w_j = \mathrm{O}(j^{-1})$ can be used and the results hold as they are (see Remark 2.5 in Barron *et al.* [6]). The weight sequence $w_j = j^{-1}$ produces an estimator that has the simple average structure $F_m = \sum_{j=1}^m (\frac{j}{m}) g_j(X)$.

The RGA is advocated by Barron *et al.* [6], as it possesses better theoretical properties than PGA (L_2-Boosting) and it is simpler to implement than the OGA. At each stage j, the algorithm solves $\min_{k,b} |Y - (1-w_j)F_{j-1} + w_j X^{(k)} b|_n^2$. It is possible to also consider the case where w_j is not fixed in advance, but estimated at each iteration. In Figure 3, one just replaces the line defining $s(j)$ with

$$\bigl[s(j), w_j\bigr] := \arg \max_{k \le K, w \in [0,1]} \bigl|\bigl\langle Y - (1-w)F_{j-1}, X^{(k)}\bigr\rangle_n\bigr|. \tag{14}$$

The asymptotic results hold as they are, as in this case, the extra optimization can only reduce algo(m, B), the error in the algorithm. The same remark holds for the next algorithms.

2.3.4. *Constrained greedy and Frank–Wolfe Algorithms*

The Constrained Greedy Algorithm (CGA) is a variation of the RGA. It is used in Sancetta [63] in a slightly different context. The Frank–Wolfe Algorithm (FWA) (Frank and Wolfe [35]; see

```
Set:
$m \in \mathbb{N}$
$F_0 := 0$
For: $j = 1, 2, \dots, m$
$w_j = 1/j$
$s(j) := \arg\max_k |\langle Y - (1-w_j)F_{j-1}, X^{(k)}\rangle_n|$
$g_j(X) := \langle Y - (1-w_j)F_{j-1}, X^{s(j)}\rangle_n X^{s(j)}$
$F_j := (1-w_j)F_{j-1} + g_j(X)$
```

Figure 3. RGA.

```
CGA
Set:
$m \in \mathbb{N}$
$F_0 := 0$
$\bar{B} < \infty$
For: $j = 1, 2, \dots, m$
    $w_j = 1/j$
    $s(j) := \arg\max_k |\langle Y - (1 - w_j)F_{j-1}, X^{(k)}\rangle_n|$
    $b_j := \frac{1}{w_j}\langle Y - (1 - w_j)F_{j-1}, X^{s(j)}\rangle_n$
    $g_j(X) := \mathrm{sign}(b_j)(|b_j| \wedge \bar{B})X^{s(j)}$
    $F_j := (1 - w_j)F_{j-1} + w_j g_j(X)$
```

```
FWA
Set:
$m \in \mathbb{N}$
$F_0 := 0$
$\bar{B} < \infty$
For: $j = 1, 2, \dots, m$
    $w_j := 2/(1 + j)$
    $s(j) := \arg\max_k |\langle Y - F_{j-1}, X^{(k)}\rangle_n|$
    $b_j := \bar{B}\,\mathrm{sign}(\langle Y - F_{j-1}, X^{s(j)}\rangle_n)$
    $g_j(X) := b_j X^{s(j)}$
    $F_j := (1 - w_j)F_{j-1} + w_j g_j(X)$
```

Figure 4. CGA and FWA.

Clarkson [26], Jaggi [42], Freund, Grigas and Mazumder [36], for recent results on its convergence) is a well-known algorithm for the optimization of functions under convex constraints. Figure 4 review the algorithms. The two algorithms are similar, though some notable differences are present. The FWA chooses at each iteration the regressor that best fits the residuals from the previous iteration model. Moreover, the regression coefficient is chosen as the value of the constraint times the sign of the correlation of the residuals with the chosen regressor. On the other hand, the difference of the CGA from the RGA is that at each step the estimated regression coefficient is constrained to be smaller in absolute value than a pre-specified value $\bar{B}$. When the function one wants to estimate is known to lie in $\mathcal{L}(1)$, the algorithm is just a simplified version of the Hilbert Space Projection algorithm of Jones [43] and Barron [5] and have been studied by several authors for estimation of mixture of densities (Li and Barron [46], Rakhlin, Panchenko and Mukherjee [59], Klemelä [44], Sancetta [62]).

At each step j, the CGA solves $\min_{k,|b|\le\bar{B}} |Y - (1 - w_j)F_{j-1} + w_j X^{(k)} b|_n^2$. Under the square loss with regression coefficients satisfying $\sum_{k=1}^K |b_k| \le \bar{B}$, the FWA reduces to minimization of the linear approximation of the objective function, minimized over the simplex, that is, $\min_{k,|b|\le\bar{B}} \langle bX^{(k)}, F_{j-1} - Y\rangle_n$ with update of F_j as in Figure 4. Despite the differences, both the CGA and the FWA lead to the solution of the Lasso problem. In particular, the regression coefficients are the solution to the following problem:

$$\min_{b_1,b_2,\dots,b_K} \left| Y - \sum_{k=1}^K b_k X^{(k)} \right|_n, \qquad \text{such that } \sum_{k=1}^K |b_k| \le \bar{B}.$$

The above is the standard Lasso problem due to Tibshirani [69]. In particular, CGA and FWA solve the above problem as $m \to \infty$,

$$|Y - F_m|_n^2 \le \inf_{\mu\in\mathcal{L}(\bar{B})} \left|Y - \mu(X)\right|_n^2 + \frac{\bar{B}^2}{m}$$

(Lemma 6 and 7, in Section 4, where for simplicity, only the weighting schemes as in Figure 4 are considered). The complexity of the estimation procedure is controlled by $\bar{B}$. This parameter can be either chosen based on a specific application, or estimated via cross-validation, or splitting the sample into estimation and validation sample.

The CGA and FGA also allows one to consider the forecast combination problem with weights in the unit simplex, by minor modification. To this end, for the CGA let

$$g_j(X) := \big[(b_j \wedge 1) \vee 0\big] X^{s(j)}, \tag{15}$$

so that $\bar{B} = 1$ and the estimated b_j's parameters are bounded below by zero. For the FWA change,

$$s(j) := \arg\max_k \langle Y - F_{j-1}, X^{(k)} \rangle_n; \, b_j := \bar{B} \operatorname{sign}\big(\langle Y - F_{j-1}, X^{s(j)} \rangle_n\big) \vee 0,$$

where one does not use the absolute value in the definition of $s(j)$. (This follows from the general definition of the Frank–Wolfe Algorithm, which simplifies to the algorithm in Figure 4 when $\sum_{k=1}^{K} |b_k| \le \bar{B}$.) Hence, the resulting regression coefficients are restricted to lie on the unit simplex.

As for the RGA, for the CGA and FWA it is possible to estimate w_j at each greedy step. For the CGA, this requires to change the line defining $s(j)$ with (14). Similarly, for the FWA, one adds the following line just before the definition of F_j:

$$w_j = \arg \min_{w \in [0,1]} \big|Y - (1 - w)F_{j-1} + w g_j(X)\big|_n^2.$$

These steps can only reduce the approximation error of the algorithm, hence, the rates of convergence derived for the fixed sequence w_j are an upper bound for the case when w_j is estimated at each step.

2.4. Discussion

2.4.1. *Objective function*

The objective function is the same one used in Bühlmann [15], which is the integrated square error (ISE), where integration is w.r.t. the true distribution of the regressors (note that the expectation is w.r.t. X' only). This objective function is zero if the (out of sample) prediction error is minimized (recall that $\mu_0(X) = \mathbb{E}[Y|X]$), and for this reason it is used in the present study. Under this objective, some authors derive consistency, but not explicit rates of convergence (e.g., Bühlmann [15], Lutz and Bühlmann [49]). An exception is Barron *et al.* [6], who derive rates of convergence for the mean integrated square error. Rates of convergence of greedy algorithms are usually derived under a weaker norm, namely the empirical L_2 norm and the results hold in probability (e.g., Bühlmann and van de Geer [17], and references therein). This is essentially equivalent to assuming a fixed design for the regressors. The empirical L_2 norm has been used to show consistency of Lasso, hence deriving results under this norm allows one to compare to Lasso in a more explicit way. Convergence of the empirical L_2 norm does not necessarily guarantee that the prediction error is minimized, asymptotically.

Example 1. Let $F_m(X) = \sum_{k=1}^K X^{(k)} b_{kn}$ be the output of one of the algorithms, where the subscript n is used to stress that b_{kn} depends on the sample. Also, let $Z := Y - \mu_0(X)$, and $\mu_0(X) = \sum_{k=1}^K X^{(k)} b_{k0}$, where the b_{k0}'s are the true coefficients. Control of the empirical L_2 norm only requires control of Control of $\langle Z, \sum_{k=1}^K X^{(k)}(b_{kn} - b_{k0})\rangle_n$ (e.g., Lemma 6.1 in Bühlmann and van de Geer [17]) and this quantity tends to be $\mathrm{O}_p(m \ln K/n)$ under regularity conditions. On the other hand, control of the L_2 norm (i.e., ISE) also requires control of $(1 - \mathbb{E})|\sum_{k=1}^K X^{(k)}(b_{kn} - b_{k0})|_n^2$. Sufficient conditions for this term to be $\mathrm{O}_p(m \ln K/n)$ are often used, but in important cases such as dependent non-mixing random data, this does not seem to be the case anymore. Hence, this term is more challenging to bound and requires extra care (see van de Geer [74], for results on how to bound such a term in an i.i.d. case).

2.4.2. *Dependence conditions*

Absolute regularity is convenient, as it allows to use decoupling inequalities. In consequence, the same rate of convergence under i.i.d. observations holds under beta mixing when the mixing coefficients decay fast enough. Many time series models are beta mixing. For example, any finite order ARMA model with i.i.d. innovations and law absolutely continuous w.r.t. the Lebesgue measure satisfies geometric mixing rates (Mokkadem [51]). Similarly, GARCH models and more generally models that can be embedded in some stochastic recursive equations are also beta mixing with geometric mixing rate for innovations possessing a density w.r.t. the Lebesgue measure (e.g., Basrak, Davis and Mikosch [8], for details: they derive the results for strong mixing, but the result actually implies beta mixing). Many positive recurrent Markov chains also satisfy geometric absolute regularity (e.g., Mokkadem [52]). Hence, while restrictive, the geometric mixing rate of Conditions 2 and 3 is a convenient condition satisfied by common time series models.

In Condition 3, (5) is used to control the moments of the random variables. The geometric mixing decay could be replaced with polynomial mixing at the cost of complications linking the moments of the random variables (i.e., (5)) and their mixing coefficients (e.g., Rio [60], for details).

Condition 4 only controls dependence in terms of some conditional moments of the centered random variables. Hence, if the dependence on the past decreases as we move towards the future, the centered variables will have conditional moment closer and closer to zero. On the other hand, Conditions 2 and 3 control dependence in terms of the sigma algebra generated by the future and the past of the data. This is much stronger than controlling conditional expectations, and computation of the resulting mixing coefficients can be very complicated unless some Markov assumptions are made as in Mokkadem [51,52] or Basrak, Davis and Mikosch [8] (see Doukhan and Louhichi [31], for further discussion and motivation).

2.4.3. *Examples for Conditions 3 and 4*

To highlight the scope of the conditions and how to establish them in practice, consider a simple non-trivial example.

Example 2. Let $\mu_0(X) = g(X^{(k)}; k \le K)$, where g satisfies

$$\bigl|g\bigl(x^{(k)}; k \le K\bigr) - g\bigl(z^{(k)}; k \le K\bigr)\bigr| \lesssim \sum_{k=1}^K \lambda_k \bigl|x^{(k)} - z^{(k)}\bigr|$$

for $\sum_{k=1}^{K} \lambda_k \leq 1$, $\lambda_k \geq 0$ and $g(x^{(k)}; k \leq K) = 0$ when $x^{(k)} = 0$ for all $k \leq K$. Since $K \to \infty$ with n, it is natural to impose this condition which is of the same flavor as $\sum_{k=1}^{K} |b_k| \leq B$ in the linear model. Suppose that $(Z_i)_{i\in\mathbb{Z}}$ is a sequence of independent random variables ($Z = Y - \mathbb{E}[Y|X]$) with finite p moments, and independent of the regressors $(X_i)_{i\in\mathbb{Z}}$. The regressors admit the following vector autoregressive representation, $X_i = HW_i$, where H is a $K \times L$ matrix with positive entries and rows summing to one; $W_i = AW_{i-l} + \varepsilon_i$, A is a diagonal $L \times L$ matrix with entries less than one in absolute values, and $(\varepsilon_i)_{i\in\mathbb{Z}}$ is a sequence of i.i.d. L dimensional random variables with finite $2p$ moments, that is, $\mathbb{E}|\varepsilon_{i,k}|^{2p} < \infty$, where $\varepsilon_{i,k}$ is the kth entry in ε_i. Throughout, the K dimensional vectors are column vectors.

If one takes $L = K$ and H to be diagonal, $X_i = W_i$. As $K \to \infty$, the process is not necessarily mixing. Hence, one is essentially required to either keep L fixed or impose very restrictive structure on the innovations in order to derive mixing coefficients. Luz and Bühlmann [15] consider vector autoregressive models (VAR) with the dimension of the variables increasing to infinity. They then assume that the model is strongly mixing. However, it is unclear that a VAR of increasing dimensionality can be strongly mixing. The mixing coefficients of functions of independent random variables are bounded above by the sum of the mixing coefficients of the individual variables (e.g., Theorem 5.1 in Bradley [14]). If the number of terms in the sum goes to infinity (i.e., K in the present context, q in Luz and Bühlmann [15]), such VAR may not be strongly mixing. Even using a known results on Markov chain, it is not possible to show that VAR models with increasing dimension are mixing without very restrictive conditions on the innovations (e.g., condition iii in Theorem $1'$ in Mokkadem [51]).

Restrictions such as A being diagonal or $(X_i)_{i\in\mathbb{Z}}$ and $(Z_i)_{i\in\mathbb{Z}}$ being independent are only used to simplify the discussion, so that one can focus on the standard steps required to establish the validity of the conditions in Example 2. The above model can be used to show how to check Conditions 3 and 4 and how Condition 3 can fail.

Lemma 2. *Consider the model in Example* 2. *Suppose that* ε_i *has a density w.r.t. the Lebesgue measure and* L *is bounded. Then Condition* 3 *is satisfied.*

Lemma 3. *Consider the model in Example* 2. *Suppose that* $\varepsilon_{i,k}$ *only takes values in* $\{-1, 1\}$ *with equal probability for each* k, $L = K$ *and* H *is the identity matrix* (*i.e.*, $X_i = W_i$), *while all the rest is as in Example* 2. *Then Condition* 3 *is not satisfied*, *but Condition* 4 *is satisfied.*

The proof of these two lemmas – postponed to Section 4.6 – shows how the conditions can be verified.

The next examples provides details on the applicability of Condition 4 to possibly long memory processes. In particular, the goal is to show that Corollary 5 can be applied. In consequence, new non-trivial models and conditions are allowed. In these examples, the rates of convergence implied by Corollary 5 are comparable to, or better than the ones in Bühlmann and van de Geer [17] which require i.i.d. observations. However, one needs to restrict attention to regressors whose population Gram matrix has full rank ($\rho_K > 0$). The following only requires stationarity and ergodicity of the error terms.

Example 3. Let $(Z_i)_{i\in\mathbb{Z}}$ be a stationary ergodic sequence with moments of all orders, and suppose that $(X_i)_{i\in\mathbb{Z}}$ is i.i.d., independent of the Z_i's, and with zero mean and moments of all orders and such that $\rho_K > 0$. Moreover, suppose that $\mu_0 \in \mathcal{L}$. By independence of X_i and the Z_i's, and the fact the that X_i's are i.i.d. mean zero, it follows that $\mathbb{E}_0 Z_i X_i^{(k)} = 0$. Similarly, $\mathbb{E}_0(1-\mathbb{E})|X_i^{(k)}|^2 = 0$ for $i > 0$. Finally, given that $\mu_0 \in \mathcal{L}$, $\Delta(X) = \mu_B(X) - \mu_0(X) = 0$ by choosing B large enough so that $\mu_B = \mu_0$. Hence, this implies that $\sup_n d_{n,p} < \infty$ in Corollary 5, though for the CGA and FWA it is necessary to assume $\mu_0 \in \mathcal{L}(\bar{B})$ and not just $\mu_0 \in \mathcal{L}$, or just $\mu_0 \in \mathcal{L}$ but $\bar{B} \to \infty$.

Remarkably, Example 3 shows that if the regressors are i.i.d., it is possible to achieve results as good as the ones derived in the literature only assuming ergodic stationary noise. The next example restricts the noise to be i.i.d., but allows for long memory Gaussian regressors and still derives convergence rates as fast as the ones of Example 3.

Example 4. *Let $X_i^{(k)} = \sum_{l=0}^{\infty} a_{lk}\varepsilon_{i-l,k}$, where $(\varepsilon_{i,k})_{i\in\mathbb{Z}}$ is a sequence of i.i.d. standard Gaussian random variables, and $a_{0k} = 1$, $a_{lk} = l^{-(1+\epsilon)/2}$ $\epsilon \in (0,1]$ for $l > 0$. Also, suppose that $(X_i)_{i\in\mathbb{Z}}$ is independent of $(Z_i)_{i\in\mathbb{Z}}$, which is i.i.d. with moments of all orders. It is shown in Section 4.7 that for this* MA(∞) *model with Gaussian errors,*

$$\left|\mathbb{E}_0(1-\mathbb{E})\left|X_i^{(k)}\right|^2\right|_p \lesssim i^{-(1+\epsilon)}$$

when $i > 0$. Hence, in Condition 4 $\sup_n d_{n,p} < \infty$ for any $p < \infty$, and in consequence, one can apply Corollary 5 if $\rho_K > 0$ and the true function is in $\mathcal{L}$ or $\mathcal{L}(\bar{B})$ for the CGA and FWA. For the CGA and FWA, $\rho_K = 0$ is allowed.

3. Implementation and numerical comparison

3.1. Vectorized version

Vectorized versions of the algorithms can be constructed. These versions make quite clear the mechanics behind the algorithms. The vectorized versions are useful when the algorithms are coded using scripting languages or when n and K are very large, but $K = \mathrm{o}(n)$. In this case, the time dimension n could be about $\mathrm{O}(10^7)$ or even $\mathrm{O}(10^8)$ and the cross-sectional dimension $K = \mathrm{O}(10^3)$. The memory requirement to store a matrix of doubles of size $10^7 \times 10^3$ is in excess of 70 gigabytes, often too much to be stored in RAM on most desktops. On the other hand, sufficient statistics such as $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$ ($\mathbf{X}$ being the $n \times K$ matrix of regressors and $\mathbf{Y}$ the $n \times 1$ vector of dependent variables and the subscript T stands for transpose) are manageable and can be updated through summation.

Figure 5 shows vectorized versions of the algorithms. Of course, it is always assumed that the regressors have been standardized, that is, $\mathrm{diag}(\mathbf{X}^T\mathbf{X}/n) = I_K$, the identity matrix, where $\mathrm{diag}(\cdot)$ stands for the diagonal matrix constructed from the diagonal of its matrix argument. The symbol 0_K is the K dimensional vector of zeros, while for other vector quantities, the subscript denotes the entry in the vector, which are assumed to be column vectors.

PGA	OGA	RGA
Set:		
$C = \mathbf{X}^T\mathbf{Y}/n$	$C = \mathbf{X}^T\mathbf{Y}/n$	$C = \mathbf{X}^T\mathbf{Y}/n$
$D = \mathbf{X}^T\mathbf{X}/n$	$D = \mathbf{X}^T\mathbf{X}/n$	$D = \mathbf{X}^T\mathbf{X}/n$
$b = 0_K$	$b = 0_K$	$b = 0_K$
$\nu \in (0, 1)$		
For: $j = 1, 2, \dots, m$		
$A = C - Db$	$A = C - Db$	$A = C - (1 - \frac{1}{j})Db$
$s(j) = \arg\max_{k \le K} \lvert A_k \rvert$	$s(j) = \arg\max_{k \le K} \lvert A_k \rvert$	$s(j) = \arg\max_{k \le K} \lvert A_k \rvert$
$a = 0_K$	P_X^j as in Figure 2	$a = 0_K$
$a_{s(j)} = A_{s(j)}$	$b = P_X^j Y$	$a_{s(j)} = A_{s(j)}$
$b = b + \nu a$		$b = (1 - \frac{1}{j})b + \frac{1}{j}a$

CGA	FWA
Set:	
$C = \mathbf{X}^T\mathbf{Y}/n$	$C = \mathbf{X}^T\mathbf{Y}/n$
$D = \mathbf{X}^T\mathbf{X}/n$	$D = \mathbf{X}^T\mathbf{X}/n$
$b = 0_K$	$b = 0_K$
$\bar{B} < \infty$	$\bar{B} < \infty$
For: $j = 1, 2, \dots, m$	
$A = C - (1 - \frac{1}{j})Db$	$A = C - Db$
$s(j) = \arg\max_{k \le K} \lvert A_k \rvert$	$s(j) = \arg\max_{k \le K} \lvert A_k \rvert$
$a = 0_K$	$a = 0_K$
$a_{s(j)} = \mathrm{sign}(A_{s(j)})(j \lvert A_{s(j)} \rvert \wedge \bar{B})$	$a_{s(j)} = \mathrm{sign}(A_{s(j)})\bar{B}$
$b = (1 - \frac{1}{j})b + \frac{1}{j}a$	$b = (1 - \frac{2}{1+j})b + \frac{2}{1+j}a$

Figure 5. Vectorized versions of the algorithms.

3.2. Choosing the number of iterations

In order to achieve the bounds in the theorem, m needs to be chosen large enough for the algorithm to perform well in terms of approximation error (see Lemmas 4, 5 and 6). Nevertheless, an excessively large m can produce poor results as shown in the theorems with the exception of CGA and FWA. In consequence, guidance on the number of iterations m is needed. The number of regressors can be left unconstrained in many situations, as long as the dependence is not too strong. The number of iterations can be chosen following results in the literature. Suppose the F_m estimator in the algorithm can be represented as $F_m(X) = P^m Y$ for some suitable projection operator P^m. Then one may choose the number of iterations according the following AIC criterion:

$$\ln\bigl(\lvert Y - F_m(X)\rvert_n^2\bigr) + 2\mathrm{df}\bigl(P^m\bigr)/n,$$

where $\mathrm{df}(P^m)$ are the degrees of freedom of the prediction rule P^m, which are equal to the sum of the eigenvalues of P^m, or equivalently they are equal to the trace of the operator. Bühlmann [15] actually suggests using the modified AIC based on Hurvich, Simonoff and Tsai [41]:

$$\ln\left(\left|Y - F_m(X)\right|_n^2\right) + \frac{1 + \mathrm{df}(P^m)/n}{1 - (\mathrm{df}(P^m) + 2)/n}.$$

For ease of exposition, let $\mathbf{X}_m$ be the $n \times m$ matrix of selected regressors and denote by $\mathbf{X}_m^{s(j)}$ the jth column of $\mathbf{X}_m$. For the PGA, Bühlmann and Yu [18] show that the degrees of freedom are given by the trace of

$$\mathcal{B}_m := I_n - \prod_{j=1}^{m}\left(I_n - \nu \frac{\mathbf{X}_m^{s(j)}(\mathbf{X}_m^{s(j)})'}{(\mathbf{X}_m^{s(j)})'\mathbf{X}_m^{s(j)}}\right),$$

where I_n is the n dimensional identity matrix.

The trace of the hat matrix $\mathcal{B}_m := \mathbf{X}_m(\mathbf{X}_m^T\mathbf{X}_m)^{-1}\mathbf{X}_m^T$ gives the degrees of freedom for the OGA, that is, $\mathrm{Trace}(\mathcal{B}_m) = m$.

Unfortunately, the projection matrix of the RGA is complicated and the author could not find a simple expression. Nevertheless, the degrees of freedom could be estimated (e.g., Algorithm 1 in Jianming [78]).

Choice of $\bar{B}$ is equivalent to the choice of the penalty constant in Lasso. Hence, under regularity conditions (Zou *et al.* [82], Tibshirani and Taylor [70]) the degrees of freedom of the CGA and FWA are approximated by the number of non-zero coefficients or the rank of the population Gram matrix of the selected variables. Alternatively, one has to rely on cross-validation to choose m for the PGA, OGA, RGA and $\bar{B}$ for the CGA and FWA.

3.3. Numerical results

To assess the finite performance of the algorithms a comprehensive set of simulations is carried out for all the algorithms. It is worth mentioning that the CGA and FWA are equivalent to Lasso, hence, conclusions also apply to the Lasso, even though the conditions used for consistency are very different.

For each Monte Carlo set up, 100 simulations are run, where the sample size is $n = 20{,}100$. Consider the model

$$Y_i = \sum_{k=1}^{K} X_i^{(k)} b_k + Z_i, \qquad X_i^{(k)} = \sum_{s=0}^{S} \theta_s \varepsilon_{i-s,k}, \qquad Z_i = \frac{\kappa}{\sigma}\sum_{s=0}^{S}\theta_s\varepsilon_{i-s,0},$$

where $K = 100$, $\kappa^2 = \frac{\mathrm{Var}(\sum_{k=1}^{K} X_i^{(k)} b_k)}{\mathrm{Var}(\sum_{s=0}^{S}\theta_s\varepsilon_{i-s,0})}$, so that $\sigma^2 \in \{8, 0.25\}$ is the signal to noise ratio, corresponding roughly to an R^2 of $0.89, 0.2$. The innovations $\{(\varepsilon_{i,k})_{i\in\mathbb{Z}}\colon k = 0, 1, \dots, K\}$ are collections of i.i.d. standard normal random variables. For $k, l > 0$, $\mathbb{E}\varepsilon_{i,k}\varepsilon_{i,l} = \omega^{|k-l|}$ with $\omega = \{0, 0.75\}$ with convention $0^0 = 1$, that is, a Toeplitz covariance matrix. Moreover, $\mathbb{E}\varepsilon_{i,0}\varepsilon_{i,k} = 0$ for any $k > 0$. Finally, $\{\theta_s\colon s = 0, 1, \dots, S\}$ is as follows:

Case ID: $\theta_0 = 1$ and $\theta_s = 0$ if $s > 0$;
Case WD: $\theta_s = (0.95)^s$ with $S = 100 + n$;
Case SD: $\theta_s = (s+1)^{-1/2}$ with $S = 1000 + n$.

In other words, the above model allows for time dependent Z_i's and X_i's as well for correlated regressors (when $\omega > 0$). However, the X and the Z are independent by construction. By different choice of regression coefficients b_k's, it is possible to define different scenarios for the evaluation of the algorithms. These are listed in the relevant subsections below. For each different scenario, the mean integrated square error (MISE) from the simulations is computed: that is, the Monte Carlo approximation of $\mathbb{E}[\mathbb{E}'|\mu_0(X') - F_m(X')|^2]$. Standard errors were all relatively small, so they are not reported, but available upon requests together with more detailed results.

The number of greedy steps m or the bound $\bar{B}$ were chosen by a cross-validation method for each of the algorithms (details are available upon request). Hence, results also need to be interpreted bearing this in mind, as cross-validation can be unstable at small sample sizes (e.g., Efron [33], see also Sancetta [61], for some simulation evidence and alternatives, amongst many others). Moreover, cross-validation is usually inappropriate for dependent data, often leading to larger than optimal models (e.g., Burman and Nolan [22], Burman, Chow and Nolan [21], for discussions and alternatives). Nevertheless, this also allows one to assess how robust is the practical implementation of the algorithms. Given the large amount of results, Section 3.8 summarizes the main conclusions.

3.4. Low-dimensional model

The true regression function has coefficients $b_k = 1/3$ for $k = 1, 2, 3$, and $b_k = 0$ for $k > 3$.

3.5. High-dimensional small equal coefficients

The true regression function has coefficients $b_k = 1/K$, $k \le K$.

3.6. High-dimensional decaying coefficients

The true regression function has coefficients $b_k = k^{-1}$, $k \le K$.

3.7. High-dimensional slowly decaying coefficients

The true regression function has coefficients $b_k = k^{-1/2}$, $k \le K$.

3.8. Remarks on numerical results

Results from the simulations are reported in Tables 2–5. These results show that the algorithms are somehow comparable, within a ±10% relative performance. Overall, the PGA (L_2-Boosting) is robust and often delivers the best results despite the theoretically slower convergence rates.

On the other hand, the performance of the OGA is somehow disappointing given the good theoretical performance. Table 2 shows that the OGA can perform remarkably well under very

Table 2. MISE: low-dimensional, $K = 100$

	$n = 20$					$n = 100$				
(ω, σ^2)	PGA	OGA	RGA	CGA	FWA	PGA	OGA	RGA	CGA	FWA
					Case ID					
$(0, 8)$	0.40	0.51	0.36	0.36	0.40	0.08	0.03	0.09	0.09	0.09
$(0, 0.20)$	0.59	0.87	0.93	0.75	0.77	0.47	0.52	0.49	0.44	0.44
$(0.75, 8)$	0.25	0.39	0.26	0.36	0.35	0.09	0.15	0.07	0.13	0.13
$(0.75, 0.25)$	0.86	1.20	1.29	1.00	1.14	0.50	0.45	0.49	0.48	0.47
					Case WD					
$(0, 8)$	1.65	2.06	1.56	1.51	1.52	0.67	0.68	0.54	0.56	0.54
$(0, 0.20)$	2.81	2.95	2.97	3.49	3.01	3.07	4.01	2.82	2.93	2.95
$(0.75, 8)$	1.25	2.21	1.24	1.35	1.32	0.87	1.18	0.79	0.85	0.89
$(0.75, 0.25)$	4.36	4.29	4.56	5.34	5.28	4.43	5.55	4.18	4.45	4.56
					Case SD					
$(0, 8)$	1.26	1.63	1.26	1.24	1.25	0.50	0.50	0.43	0.42	0.41
$(0, 0.20)$	2.31	2.36	2.55	2.61	2.53	2.20	2.72	2.16	2.15	2.14
$(0.75, 8)$	0.88	1.82	0.91	0.98	1.00	0.63	0.86	0.58	0.58	0.58
$(0.75, 0.25)$	3.28	3.37	3.58	3.88	4.14	3.13	3.74	3.05	3.11	3.12

Table 3. MISE: high-dimensional small coefficients, $K = 100$

	$n = 20$					$n = 100$				
(ω, σ^2)	PGA	OGA	RGA	CGA	FWA	PGA	OGA	RGA	CGA	FWA
					Case ID					
$(0, 8)$	0.10	0.12	0.11	0.10	0.10	0.08	0.10	0.08	0.08	0.09
$(0, 0.20)$	0.11	0.16	0.16	0.13	0.14	0.10	0.11	0.12	0.10	0.10
$(0.75, 8)$	0.20	0.27	0.17	0.17	0.14	0.09	0.12	0.08	0.09	0.09
$(0.75, 0.25)$	0.26	0.38	0.38	0.29	0.33	0.23	0.28	0.25	0.22	0.22
					Case WD					
$(0, 8)$	0.35	0.40	0.35	0.37	0.33	0.27	0.36	0.25	0.25	0.22
$(0, 0.20)$	0.50	0.56	0.53	0.59	0.52	0.53	0.68	0.51	0.54	0.56
$(0.75, 8)$	0.65	0.88	0.65	0.63	0.50	0.34	0.44	0.29	0.31	0.33
$(0.75, 0.25)$	1.28	1.28	1.34	1.58	1.50	1.27	1.62	1.22	1.32	1.37
					Case SD					
$(0, 8)$	0.28	0.30	0.28	0.28	0.26	0.22	0.29	0.21	0.21	0.19
$(0, 0.20)$	0.38	0.39	0.45	0.45	0.43	0.38	0.49	0.40	0.40	0.39
$(0.75, 8)$	0.51	0.70	0.51	0.50	0.43	0.25	0.37	0.24	0.26	0.27
$(0.75, 0.25)$	0.95	1.00	1.05	1.07	1.12	0.90	1.15	0.88	0.93	0.91

Table 4. MISE: high-dimensional decaying coefficients, $K = 100$

	$n = 20$					$n = 100$				
(ω, σ^2)	PGA	OGA	RGA	CGA	FWA	PGA	OGA	RGA	CGA	FWA
					Case ID					
(0, 8)	2.28	2.60	2.33	2.26	2.13	1.44	2.03	1.39	1.59	1.78
(0, 0.20)	2.42	3.61	3.72	3.02	3.12	2.23	2.48	2.56	2.22	2.25
(0.75, 8)	3.98	5.32	3.25	3.19	2.80	1.70	2.38	1.56	1.75	1.79
(0.75, 0.25)	5.51	7.89	8.18	6.27	7.02	4.46	5.49	5.07	4.33	4.37
					Case WD					
(0, 8)	7.80	8.72	7.91	8.01	7.28	5.34	7.23	4.60	4.53	4.43
(0, 0.20)	11.27	12.83	11.96	13.43	11.79	12.31	15.83	11.55	12.15	12.54
(0.75, 8)	13.01	17.84	12.66	12.10	10.22	6.65	8.78	5.86	6.30	6.62
(0.75, 0.25)	26.36	28.49	27.94	32.07	31.17	26.81	33.34	25.27	27.41	28.33
					Case SD					
(0, 8)	6.19	6.74	6.35	6.38	5.92	4.20	5.69	3.98	4.00	3.96
(0, 0.20)	8.95	8.86	10.40	10.45	10.04	8.81	10.91	9.14	9.06	8.91
(0.75, 8)	10.50	14.23	10.34	9.81	8.72	5.19	7.31	4.74	4.99	5.13
(0.75, 0.25)	19.90	21.25	22.46	23.48	24.58	19.10	24.36	18.51	19.45	19.07

Table 5. MISE: high-dimensional slow decay, $K = 100$

	$n = 20$					$n = 100$				
(ω, σ^2)	PGA	OGA	RGA	CGA	FWA	PGA	OGA	RGA	CGA	FWA
					Case ID					
(0, 8)	0.97	0.94	0.92	0.93	1.00	0.42	0.51	0.46	0.42	0.42
(0, 0.20)	1.34	1.95	2.05	1.67	1.69	1.01	0.95	1.05	0.97	0.99
(0.75, 8)	1.10	1.56	1.08	1.07	1.08	0.51	0.77	0.53	0.56	0.56
(0.75, 0.25)	2.28	3.22	3.50	2.70	3.03	1.54	1.71	1.68	1.47	1.50
					Case WD					
(0, 8)	3.54	4.31	3.49	3.52	3.41	1.88	2.22	1.62	1.70	1.68
(0, 0.20)	6.16	7.68	6.59	7.51	6.58	6.73	8.38	6.11	6.40	6.71
(0.75, 8)	4.48	6.73	3.99	4.03	3.89	2.46	3.33	2.21	2.44	2.55
(0.75, 0.25)	11.53	12.31	11.96	13.67	13.55	11.55	14.42	10.99	11.66	12.09
					Case SD					
(0, 8)	2.83	3.40	2.72	2.76	2.74	1.45	1.81	1.36	1.37	1.35
(0, 0.20)	5.04	4.81	5.82	5.89	5.60	4.81	5.90	4.85	4.84	4.69
(0.75, 8)	3.37	5.08	3.24	3.41	3.22	1.97	2.60	1.75	1.80	1.82
(0.75, 0.25)	8.51	8.82	9.57	10.07	10.51	8.12	10.00	7.98	8.11	8.16

special circumstance, that is, relatively large sample size ($n = 100$), time independent and uncorrelated regressors and high signal to noise ratio. To some extent, these are the conditions used by Zhang [81] to show optimality of the OGA.

The RGA, CGA and FWA provide good performance comparable to the PGA and in some cases better, especially when the signal to noise ration is higher. For example, Table 2 shows that these algorithms perform well as long as the regressors are either uncorrelated or the time dependence is low. Intuitively, time dependence leads to an implicit reduction of information, hence it is somehow equivalent to estimation with a smaller sample. This confirms the view that the PGA is usually the most robust of the methods.

While somehow equivalent, the FWA updates the coefficients in a slightly cruder way than the CGA. This seems to lead the FWA to have slightly different performance than the CGA in some cases, with no definite conclusion on which one is best. No attempt was made to use a line search for w_j (e.g., (14)) instead of the deterministic weights.

4. Proofs

The proof for the results requires first to show that the estimators nearly minimize the objective function $|Y - \mu(X)|_n^2$ for $\mu \in \mathcal{L}(B)$. Then uniform law of large numbers for $|Y - \mu(X)|_n^2$ with $\mu \in \mathcal{L}(B)$ or related quantities are established.

To avoid cumbersome notation, for any functions of (Y, X), say f and g, write $\langle f, g\rangle_P := \int f(y,x)g(y,x)\,\mathrm{d}P(y,x)$ where P is the marginal distribution of (Y, X); moreover, $|f|_{P,2}^2 := \langle f, f\rangle_P$. In the context of the paper, this means that $|Y - \mu_n|_{P,2}^2 = \int |y - \mu_n(x)|^2\,\mathrm{d}P(y,x)$ for a possibly random function $\mu_n(x)$ (e.g., a sample estimator). Clearly, if $\mu_n = \mu$ is not random, $|Y - \mu|_{P,2}^2 = |Y - \mu|_2^2$. Consequently, the norm $|\cdot|_{P,2}^2$ means that $|\mu_n - \mu|_{P,2}^2 := \mathbb{E}'|\mu_n(X') - \mu(X')|^2$, where X' and $\mathbb{E}'$ are as defined just before (4).

For any $\mu(X) := \sum_{k=1}^K b_k X^{(k)} \in \mathcal{L}$, $|\mu|_{\mathcal{L}} = \sum_{k=1}^K |b_k|$ denotes the l_1 norm of the linear coefficients. Throughout, $R_m := (Y - F_m)$ denotes the residual in the approximation.

4.1. Approximation rates for the algorithms

The following provide approximation rates of the algorithms and show that the resulting minimum converges to the global minimum, which might not be unique, as the number of iterations m goes to infinity.

Lemma 4. *For the PGA, for any $\mu \in \mathcal{L}(B)$,*

$$|R_m|_n^2 \le \left|Y - \mu(X)\right|_n^2 + \left(\frac{4|Y|_n^4 B^2}{\nu(2-\nu)m}\right)^{1/3}.$$

Proof. Let $\tilde{R}_0 = \mu \in \mathcal{L}(B)$, and

$$\tilde{R}_m = \tilde{R}_{m-1} - \nu\left\langle X^{s(m)}, Y - F_{m-1}\right\rangle_n X^{s(m)}$$

so that $\tilde{R}_m \in \mathcal{L}(B_m)$, where $B_0 := B$,

$$B_m := B_{m-1} + \nu \left|\left\langle X^{s(m)}, Y - F_{m-1}\right\rangle_n\right|. \tag{16}$$

Also note that $\tilde{R}_m = R_m - (Y - \mu)$, where $R_m = Y - F_m$, $F_0 = 0$. Unlike R_0, $\tilde{R}_0$ has coefficients that are controlled in terms of B_m, hence, it will be used to derive a recursion for the gain at each greedy step. Hence, using these remarks,

$$|\tilde{R}_m|_n^2 = \langle \tilde{R}_m, \tilde{R}_m\rangle_n = \langle \tilde{R}_m, R_m\rangle_n - \langle \tilde{R}_m, Y - \mu\rangle_n \leq B_m \max_k \left|\left\langle X^{(k)}, R_m\right\rangle_n\right| - \langle \tilde{R}_m, Y - \mu\rangle_n$$

because $\tilde{R}_m \in \mathcal{L}(B_m)$, which, by definition of $X^{s(m+1)}$ implies

$$\begin{aligned}\left|\left\langle X^{(m+1)}, R_m\right\rangle_n\right| &\geq \frac{\langle \tilde{R}_m, \tilde{R}_m + Y - \mu\rangle_n}{B_m} = \frac{\langle \tilde{R}_m, R_m\rangle_n}{B_m} \\ &= \frac{\langle R_m, R_m\rangle_n - \langle R_m, Y - \mu\rangle_n}{B_m}\end{aligned}$$

using the definition of $\tilde{R}_m$ in the last equality. Then, by the scalar inequality $ab \leq (a^2 + b^2)/2$ the above becomes

$$\left|\left\langle X^{(m+1)}, R_m\right\rangle_n\right| \geq \frac{|R_m|_n^2 - |Y - \mu|_n^2}{2B_m}. \tag{17}$$

Note that the right-hand side is positive, if not, $|Y - F_m|_n^2 \leq |Y - \mu|_n^2$ and the lemma is proved (recall that $R_m = Y - F_m$). Now, note that $R_m = R_{m-1} - \nu\langle X^{s(m)}, R_{m-1}\rangle_n X^{s(m)}$, so that

$$\begin{aligned}|R_m|_n^2 &= |R_{m-1}|_n^2 + \nu^2\left|\left\langle X^{s(m)}, R_{m-1}\right\rangle_n\right|^2 - 2\nu\left|\left\langle X^{s(m)}, R_{m-1}\right\rangle_n\right|^2 \\ &= |R_{m-1}|_n^2 - \nu(2-\nu)\left|\left\langle X^{s(m)}, R_{m-1}\right\rangle_n\right|^2.\end{aligned}$$

The above two displays imply

$$|R_m|_n^2 \leq |R_{m-1}|_n^2 - \frac{\nu(2-\nu)}{4B_{m-1}^2}\left(|R_{m-1}|_n^2 - |Y - \mu|_n^2\right)^2.$$

Subtracting $|Y - \mu|_n^2$ on both sides, and defining $a_m := |R_m|_n^2 - |Y - \mu|_n^2$, and $\tau := \nu(2-\nu)/4$, the above display is

$$a_m \leq a_{m-1}\left(1 - \tau a_{m-1} B_{m-1}^{-2}\right). \tag{18}$$

The proof then exactly follows the proof of Theorem 3.6 in DeVore and Temlyakov [29]. For completeness, the details are provided. Define

$$\rho(R_m) := a_m^{-1/2}\left|\left\langle X^{s(m+1)}, R_m\right\rangle_n\right| \geq a_m^{1/2} B_m^{-1}. \tag{19}$$

Since $B_m \geq B_{m-1}$,

$$a_m B_m^{-2} \leq a_{m-1} B_{m-1}^{-2}\bigl(1 - \tau a_{m-1} B_{m-1}^{-2}\bigr) \leq \frac{1}{\tau m} \tag{20}$$

using Lemma 3.4 in DeVore and Temlyakov [29] in the second step in order to bound the recursion. Then (16) and (19) give

$$\begin{aligned} B_m &= B_{m-1}\bigl(1 + \nu\rho(R_{m-1}) a_{m-1}^{1/2} B_{m-1}^{-1}\bigr) \\ &\leq B_{m-1}\bigl(1 + \nu\rho(R_{m-1})^2\bigr). \end{aligned}$$

Multiply both sides of (18) by B_m, and substitute the lower bound (19) into (18), so that using the above display,

$$\begin{aligned} a_m B_m &\leq a_{m-1} B_{m-1}\bigl(1 + \nu\rho(R_{m-1})^2\bigr)\bigl(1 - \tau\rho(R_{m-1})^2\bigr) \\ &= a_{m-1} B_{m-1}\bigl(1 - \nu\tau\rho(R_{m-1})^4\bigr) \leq |Y|_n^2 B, \end{aligned}$$

where the last inequality follows after iterating because $1 - \nu\tau\rho(R_{m-1})^4 \in (0, 1)$ and substituting $B_0 = B$ and $a_0 = |Y|_n^2$. If $a_m > 0$, it is obvious that $1 - \nu\tau\rho(R_{m-1})^4 \in (0, 1)$. If this were not the case, the lemma would hold automatically at step m, by definition of a_m. Hence, by the above display together with (20),

$$a_m^3 = (a_m B_m)^2 a_m B_m^{-2} \leq \frac{4|Y|_n^4 B^2}{\nu(2-\nu)m}$$

using the definition of $\tau = \nu(2-\nu)/4$, so that $a_m \leq [4|Y|_n^4 B/(\nu(2-\nu)m)]^{1/3}$. □

The following bound for the OGA is Theorem 2.3 in Barron *et al.* [6].

Lemma 5. *For the OGA, for any* $\mu \in \mathcal{L}(B)$,

$$|R_m|_n^2 \leq \bigl|Y - \mu(X)\bigr|_n^2 + 4\frac{B^2}{m}.$$

The following Lemma 6 is Theorem 2.4 in Barron *et al.* ([6], equation (2.41)), where the CGA bound is inferred from their proof (in their proof set their β on page 78 equal to $w_k \bar{B}$ to satisfy the CGA constraint).

Lemma 6. *For the RGA, for any* $\mu \in \mathcal{L}(B)$,

$$|R_m|_n^2 \leq \bigl|Y - \mu(X)\bigr|_n^2 + \frac{B^2}{m}.$$

For the CGA the above holds with B *replaced by* $\bar{B}$ *in the above display and any* $\mu \in \mathcal{L}(\bar{B})$.

Lemma 7. *For the FWA, for any $\mu \in \mathcal{L}(\bar{B})$, and $m > 0$,*

$$|R_m|_n^2 \le \left|Y - \mu(X)\right|_n^2 + \frac{4\bar{B}^2}{m},$$

when $w_m = 2/(1+m)$.

Proof. From Jaggi ([42], equations (3)–(4), see also Frank and Wolfe [35]), for every $m = 1, 2, 3, \dots$, infer the first inequality in the following display:

$$\begin{aligned} |R_m|_n^2 - \left|Y - \mu(X)\right|_n^2 &\le (1 - w_m)\big(|R_{m-1}|_n^2 - \left|Y - \mu(X)\right|_n^2\big) \\ &\quad + w_m^2 \max_{\sum_{k=1}^K |b_k| \le \bar{B}, \sum_{k=1}^K |c_k'| \le \bar{B}} \left|\sum_{k=1}^K \big(b_k - b_k'\big) X^{(k)}\right|_n^2 \\ &\le (1 - w_m)\big(|R_{m-1}|_n^2 - \left|Y - \mu(X)\right|_n^2\big) + w_m^2 4\bar{B}^2 \max_{k \le K} \left|X^{(k)}\right|_n^2, \end{aligned}$$

where the second inequality follows because the maximum over the simplex is at one of the edges of the simplex. Moreover, $\max_{k \le K} |X^{(k)}|_n^2 = 1$ by construction. The result then follows by Theorem 1 in Jaggi [42] when $w_m = 2/(1+m)$. □

4.2. Size of the functions generated by the algorithms

The following gives a bound for the size of F_m in terms of the norm $|\cdot|_{\mathcal{L}}$; F_m is the function generated by each algorithm.

Lemma 8. *As $n \to \infty$, $\Pr(F_m \in \mathcal{L}(B_m)) \to 1$, where*:

PGA: $B_m \lesssim |Y|_2 m^{1/2}$;
OGA: $B_m \lesssim |Y|_2[(\frac{m}{\rho_{m,n}})^{1/2} \wedge m \wedge K]$ *with $\rho_{m,n}$ as in* (1);
RGA: $B_m \lesssim |Y|_2[(\frac{m}{\rho_{m,n}})^{1/2} \wedge m \wedge K]$ *with $\rho_{m,n}$ as in* (1), *as long as in Lemma* 6 $B^2/m = \mathrm{O}(1)$;
CGA and FWA: $B_m \le \bar{B}$.

Proof. Note that $F_m(X) = \sum_{k=1}^m b_k X^{s(k)}$, where to ease notation b_k does not make explicit the dependence on m. A loose bound for $|F_m|_{\mathcal{L}}$ is found by noting that

$$\left|\langle X^{(k)}, X^{(l)}\rangle_n\right| \le \max_k \left|\langle X^{(k)}, X^{(k)}\rangle_n\right| = 1,$$

so that each coefficient is bounded by $|Y|_n$. Since at the mth iteration we have at most m different terms and no more than K, $|F_m|_{\mathcal{L}} \le (m \wedge K)|Y|_n$. Given that $|Y|_n^2 = \mathrm{O}_p(1)$, one can infer the crude bound $|F_m|_{\mathcal{L}} = \mathrm{O}_p(m \wedge K)$. This is the worse case scenario, and can be improved for all the algorithms.

For the PGA, at the first iteration, $|b_1| := \max_k |\langle X^{(k)}, Y\rangle_n| \leq |Y|_n$, hence there is an $\alpha_1 \in [0,1]$ such that $|b_1| = \alpha_1^{1/2}|Y|_n$ (the root exponent is used to ease notation in the following steps). Then, by the properties of projections

$$|R_1|_n^2 = \left|Y - X^{s(1)}b_1\right|_n^2 = |Y|_n^2 - |b_1|^2 = |Y|_n^2(1-\alpha_1),$$

where the second inequality follows from $|X^{(k)}|_n^2 = 1$ for any k. By similar arguments, there is an $\alpha_2 \in [0,1]$ such that $|b_2| = \alpha_2^{1/2}|R_1|_n$ and $|R_2|_n^2 = |R_1|_n^2(1-\alpha_2)$. So by induction $|b_m| = \alpha_m^{1/2}|R_{m-1}|_n$ and $|R_m|_n^2 = |R_{m-1}|_n^2(1-\alpha_m)$. By recursion, this implies that

$$|b_m|^2 = \alpha_m(1-\alpha_{m-1})\cdots(1-\alpha_1)|Y|_n^2$$

and in consequence that

$$\sum_{k=1}^m |b_k| = \sum_{k=1}^m \alpha_k^{1/2}\prod_{l<k}(1-\alpha_l)^{1/2}|Y|_n,$$

where the empty product is 1. It is clear that if any $\alpha_k \in \{0,1\}$ for $k < m$ then $b_m = 0$, hence one can assume that all the α_k's are in $(0,1)$. The above display is maximized if $\alpha_l \to 0$ fast enough, as otherwise, the product converges to zero exponentially fast and the result follows immediately. Suppose that $\sum_{l=1}^\infty \alpha_l^2 < \infty$. Then, using the fact that $\ln(1-\alpha_l) = -\alpha_l + \mathrm{O}(\alpha_l^2)$,

$$\prod_{l<k}(1-\alpha_l) = \cdots = \exp\left\{\sum_{l=1}^{k-1}\ln(1-\alpha_l)\right\} = \exp\left\{-\sum_{l=1}^{k-1}\alpha_l + \mathrm{O}\left(\sum_{l=1}^{k-1}\alpha_l^2\right)\right\}$$
$$\asymp \exp\left\{-\sum_{l=1}^{k-1}\alpha_l\right\}.$$

The above converges exponentially fast to 0 if $\alpha_l \asymp l^{-\alpha}$ for $\alpha \in (0.5, 1)$. While the argument is not valid for $\alpha \in (0, 0.5]$, it is clear, that the convergence is even faster in this case. Hence, restrict attention to $\alpha = 1$, in which case, $\prod_{l<k}(1-\alpha_l) \asymp k^{-c}$ for some $c > 0$, that is, polynomial decay. On the other hand for $\alpha > 1$, the product converges. Hence, it must be the case that the maximum is achieved by setting $\alpha_l \asymp l^{-1}$ and assuming that the product converges. This implies that for the PGA,

$$\sum_{k=1}^m |b_k| \lesssim |Y|_n \sum_{k=1}^m \left(k^{-1}\right)^{1/2} \lesssim |Y|_n m^{1/2}.$$

Now, consider the OGA and the RGA. The following just follows by standard inequalities:

$$(\rho_{m,n}/m)^{1/2}\sum_{k=1}^m |b_k| \leq \rho_{m,n}^{1/2}\left(\sum_{k=1}^m |b_k|^2\right)^{1/2} \leq |F_m|_n. \tag{21}$$

For the OGA, by definition of the OLS estimator, $|F_m|_n \leq |Y|_n$ implying the result for the OGA using the above display and the crude bound. For the RGA, consider the case when $|F_m|_n$ is small and large, separately. If $|F_m|_n = \mathrm{o}_p(1)$, then clearly, $|F_m|_n = \mathrm{o}(|Y|_n)$, because Y is not degenerate. By this remark, the above display implies that

$$|F_m|_{\mathcal{L}} := \sum_{k=1}^{m} |b_k| = \mathrm{o}_p\big(\sqrt{m/\rho_{m,n}}|Y|_n\big)$$

and the result for the RGA would follow. Hence, one can assume that $|F_m|_n \gtrsim 1$ in probability, eventually as $m \to \infty$. In this case, by the approximating Lemma 6, if $B^2/m = \mathrm{O}(1)$,

$$|Y - F_m|_n^2 \leq |Y|_n^2 + \mathrm{O}(1)$$

which implies

$$|F_m|_n^2 \leq 2\langle Y, F_m\rangle_n + \mathrm{O}(1) \leq 2|Y|_n|F_m|_n + \mathrm{O}(1)$$

and in consequence

$$|F_m|_n \leq 2|Y|_n + \mathrm{O}\big(|F_m|_n^{-1}\big) = 2|Y|_n + \mathrm{O}_p(1)$$

by the fact that $|F_m|_n \gtrsim 1$, in probablity. Hence, using the above display together with (21), the result follows for the RGA as well.

For the CGA, the b_k's are all bounded in absolute value by $\bar{B}$. Since by construction, $F_m(X) = m^{-1}\sum_{k=1}^{m} b_k X^{s(k)}$, $|F_m|_{\mathcal{L}} \leq \bar{B}$. A similar argument holds for the FWA. □

It is natural to replace the random eigenvalue $\rho_{m,n}$ with the population one. This is achieved next.

Lemma 9. *Suppose Conditions 1 and 4 hold. Then $\rho_{m,n} \geq \rho_m - \mathrm{O}_p(d_{n,p} m K^{2/p} n^{-1/2})$ implying that if $d_{n,p} m K^{2/p} n^{-1/2} = \mathrm{o}(\rho_m)$, then $\rho_{m,n}^{-1} = \mathrm{O}_p(\rho_m^{-1})$.*

Proof. Note that

$$\rho_{m,n} = \inf_{|b|_0 \leq m, |b|_2 \leq 1} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{K} b_k X_i^{(k)}\right)^2, \qquad \rho_m = \inf_{|b|_0 \leq m, |b|_2 \leq 1} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\sum_{k=1}^{K} b_k X_i^{(k)}\right)^2,$$

where $|b|_0 = \sum_{k=1}^{K}\{b_k \neq 0\}$ and $|b|_2^2 = \sum_{k=1}^{K}|b_k|^2$, that is, the number of non-zero b_k's and their squared l_2 norm, respectively. By obvious manipulations, using the above display, and the definition of ρ_m,

$$\rho_{m,n} \geq \rho_m - \sup_{|b|_0 \leq m, |b|_2 \leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}(1 - \mathbb{E})\left(\sum_{k=1}^{K} b_k X_i^{(k)}\right)^2\right|,$$

hence it is sufficient to bound the r.h.s. of the above display. Using similar arguments as in the control of *II* in the proof of Lemma 15 in Section 4.4,

$$\begin{aligned}
&\mathbb{E} \sup_{|b|_0 \le m, |b|_2 \le 1} \left| \frac{1}{n} \sum_{i=1}^{n} (1-\mathbb{E}) \left(\sum_{k=1}^{K} b_k X_i^{(k)} \right)^2 \right| \\
&\quad \le m \mathbb{E} \max_{k,l \le K} \left| \frac{1}{n} \sum_{i=1}^{n} (1-\mathbb{E}) X_i^{(k)} X_i^{(l)} \right| \lesssim \frac{d_{n,p} m K^{2/p}}{\sqrt{n}},
\end{aligned}$$

and the first result follows. The second part is directly inferred from the first. □

4.3. Inequalities for dependent random variables

Two different inequalities will be needed depending on whether one assumes absolute regularity or mixingales. The following is suitable for beta mixing random variables. It is somewhat standard, but proved for completeness due to some adjustments to the present context.

Lemma 10. *Suppose that $\mathfrak{F}$ is a measurable class of functions with cardinality K. Let $(W_i)_{i\in\mathbb{Z}}$ be strictly stationary and beta mixing with mixing coefficients $\beta(i) \lesssim \beta^i$, $\beta \in [0,1)$. Suppose that for all $f \in \mathfrak{F}$, $\mathbb{E}|f(W_1)|^p < \infty$ for some $p > 2$. Then*

$$\mathbb{E} \max_{f \in \mathfrak{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} (1-\mathbb{E}) f(W_i) \right| \lesssim \sqrt{\ln K}$$

if $K \lesssim n^{\alpha}$ for some $\alpha < (p-2)/2$. If $\max_{f\in\mathfrak{F}} |f|$ is bounded, the result holds for $K \lesssim \exp\{n^{\alpha}\}$, $\alpha \in [0,1)$.

Proof. Note that

$$\begin{aligned}
&\mathbb{E} \max_{f \in \mathfrak{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} (1-\mathbb{E}) f(W_i) \right| \\
&\quad \le \mathbb{E} \max_{f \in \mathfrak{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} (1-\mathbb{E}) f(W_i) \left\{ \max_{f\in\mathfrak{F}} |f(W_i)| \le M \right\} \right| \\
&\qquad + \mathbb{E} \max_{f \in \mathfrak{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} (1-\mathbb{E}) f(W_i) \left\{ \max_{f\in\mathfrak{F}} |f(W_i)| > M \right\} \right| \\
&\quad \le \mathbb{E} \max_{f \in \mathfrak{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} (1-\mathbb{E}) f(W_i) \left\{ \max_{f\in\mathfrak{F}} |f(W_i)| \le M \right\} \right| \\
&\qquad + 2\sqrt{n} \mathbb{E} \max_{f\in\mathfrak{F}} |f(W_i)| \left\{ \max_{f\in\mathfrak{F}} |f(W_i)| > M \right\} \\
&\quad =: I + II,
\end{aligned}$$

where in the last inequality one uses Minkowski's inequality. (Here, $\{\cdot\}$ is the indicator of a set.) By Hölder's inequality,

$$\begin{aligned} II &\leq 2\sqrt{n}\Big(\mathbb{E}\max_{f\in\mathfrak{F}}\big|f(W_i)\big|^p\Big)^{1/p}\Pr\Big(\max_{f\in\mathfrak{F}}\big|f(W_i)\big|>M\Big)^{(p-1)/p} \\ &\leq 2\sqrt{n}\Big(\mathbb{E}\max_{f\in\mathfrak{F}}\big|f(W_i)\big|^p\Big)^{1/p}K^{(p-1)/p}M^{-(p-1)} \\ &\lesssim \sqrt{n}KM^{-(p-1)} \end{aligned}$$

because by Markov inequality and the union bound,

$$\Pr\Big(\max_{f\in\mathfrak{F}}\big|f(W_i)\big|>M\Big)\lesssim KM^{-p},$$

while $\mathbb{E}\max_{f\in\mathfrak{F}}|f(W_i)|^p\lesssim K$ (e.g., Lemma 2.2.2 in van der Vaart and Wellner [77]). Hence, set $M=(\sqrt{n}K/\sqrt{\ln K})^{1/(p-1)}$ to ensure that $II=\mathrm{O}(\sqrt{\ln K})$. Pollard ([58], equation (8)) shows that if the W_i's are beta mixing, for any integer sequence $a_n=\mathrm{o}(n)$,

$$I\lesssim\sqrt{\ln K}|f|_{2\beta}\mathcal{E}\bigg(\frac{Ma_n\sqrt{2\ln K}}{|f|_{2\beta}\sqrt{n}}\bigg)+M\beta(a_n)\sqrt{n}, \tag{22}$$

where $\mathcal{E}$ is some positive increasing function such that $\lim_{x\to\infty}\mathcal{E}(x)=\infty$ and $|\cdot|_{2\beta}$ is the beta mixing norm introduced by Doukhan *et al.* [32] (see also Rio [60], equation (8.21)). The exact form of the norm is irrelevant for the present purposes, however, $|f|_{2\beta}\leq c_1<\infty$ for some constant c_1 under the condition on the mixing coefficients (e.g., Rio [60], p. 15). Since $\beta(a_n)\lesssim\beta^{a_n}$, for $a_n\asymp\ln n/\ln(1/\beta)$, and using the value for M set in II, deduce

$$I+II\lesssim\sqrt{\ln K}\mathcal{E}\bigg(\frac{c_2\ln n(\sqrt{n}K/\sqrt{\ln K})^{1/(p-1)}\sqrt{\ln K}}{\sqrt{n}}\bigg)+\sqrt{\ln K},$$

for some finite positive constant c_2. Substituting $K\asymp n^\alpha$ for any positive $\alpha<(p-2)/2$, the argument in the continuous increasing function $\mathcal{E}(\cdot)$ is bounded and the result follows. Notice that this choice of K also makes $M\beta(a_n)\sqrt{n}\lesssim 1$.

For the case of bounded $\max_{f\in\mathfrak{F}}|f|$, one can take M large enough, but finite so that $II=0$. Given that M is finite, $K\lesssim\exp\{n^\alpha\}$, and with a_n as before, (22) becomes

$$I\lesssim\sqrt{\ln K}|f|_{2\beta}\mathcal{E}\bigg(\frac{c_3\ln n\sqrt{n^\alpha}}{\sqrt{n}}\bigg)$$

for some finite constant c_3, and the argument of $\mathcal{E}(\cdot)$ is bounded because $\alpha<1$. Some tidying up gives the last result. □

The following is an extension of Burkhölder inequality to mixingales (see Peligrad, Utev and Wu [55], Corollary 1 for the exact constants).

Lemma 11. *Suppose that $(W_i)_{i\in\mathbb{Z}}$ is a mean zero stationary sequence of random variables. Let*

$$d_{n,p}(W) := \sum_{i=0}^{n}(i+1)^{-1/2}\left|\mathbb{E}[W_i|\mathcal{F}_0]\right|_p,$$

where $\mathcal{F}_0 := \sigma(W_i\colon i \le 0)$ is the sigma algebra generated by $(W_i\colon i \le 0)$. Then, for all $p \ge 2$, such that $|W_i|_p < \infty$,

$$\left|\sum_{i=1}^{n} W_i\right|_p \le C_p^{1/p} n^{1/2} d_{n,p}(W),$$

where for $p \in [2,4)$, $C_p \lesssim p^p$ while for $p \ge 4$, $C_p \lesssim (2p)^{p/2}$.

4.4. Uniform control of the estimator

Next, one needs a uniform control of the objective function. Recall that μ_B is the best approximation in $\mathcal{L}(B)$ to μ_0 in the L_2 sense.

Define

$$\mathcal{L}_0(B) := \left\{\mu\colon \mu(X) = \sum_{k=1}^{K} b_k X^{(k)}, \sum_{k=1}^{K}\{b_k \ne 0\} \le B\right\}.$$

These are linear functions with l_0 norm less or equal to B, that is, linear functions with at most B non-zero coefficients. The following is Lemma 5.1 in van de Geer [74] with minor differences. The proof is given for completeness.

Lemma 12. *Let $\mu' \in \mathcal{L}(B)$ be an arbitrary but fixed function and m a positive integer. Suppose that in probability, for some $\delta_1 \in (0,1)$ and $\delta_2, \delta_3 > 0$:*

1. $\sup_{\mu\in\mathcal{L}_0(2m)\colon |\mu|_2\le 1} |(1-\mathbb{E})|\mu|_n^2| \le \delta_1$,
2. $\sup_{\mu\in\mathcal{L}_0(2m)\colon |\mu|_2\le 1} |2(1-\mathbb{E})\langle Y-\mu', \mu\rangle_n| \le \delta_2$,
3. *the sequence $\mu_n \in \mathcal{L}_0(m)$ satisfies $|Y-\mu_n|_n^2 \le |Y-\mu'|_n^2 + \delta_3^2$,*
4. *the moment condition $\langle Y-\mu', \mu_n\rangle_P = 0$ holds.*

Then $|\mu_n - \mu'|_{P,2} \le (\delta_2+\delta_3)/(1-\delta_1)$ in probability (recall the definition of $|\cdot|_{P,2}$ at the beginning of Section 4).

Proof. Starting from the assumption

$$|Y-\mu_n|_n^2 \le \left|Y-\mu'\right|_n^2 + \delta_3^2,$$

by algebraic manipulations, $|\mu_n - \mu'|_n^2 \le 2\langle Y-\mu', \mu_n-\mu'\rangle_n + \delta_3^2$. Assume that $|\mu_n-\mu'|_{P,2} \ge \delta_3$ otherwise, there is nothing to prove. Hence, $\delta_3^2 \le \delta_3|\mu_n-\mu'|_{P,2}$. Also note that $\langle Y-\mu', \mu_n -$

$\mu'\rangle_P = 0$ by definition of μ' (point 4 in the statement). Adding and subtracting $|\mu_n - \mu'|_n^2$, and using the just derived bounds

$$\begin{aligned}
\left|\mu_n - \mu'\right|_{P,2}^2 &\leq \left|\mu_n - \mu'\right|_{P,2}^2 - \left|\mu_n - \mu'\right|_n^2 + 2\left(\left\langle Y - \mu', \mu_n - \mu'\right\rangle_n - \left\langle Y - \mu', \mu_n - \mu'\right\rangle_P\right) \\
&\quad + 2\left\langle Y - \mu', \mu_n - \mu'\right\rangle_P + \delta_3^2 \\
&\leq \left|\frac{|\mu_n - \mu'|_{P,2}^2 - |\mu_n - \mu'|_n^2}{|\mu_n - \mu'|_{P,2}^2}\right| \left|\mu_n - \mu'\right|_{P,2}^2 \\
&\quad + 2\left|\frac{\langle Y - \mu', \mu_n - \mu'\rangle_n - \langle Y - \mu', \mu_n - \mu'\rangle_P}{|\mu_n - \mu'|_{P,2}}\right| \left|\mu_n - \mu'\right|_{P,2} + \delta_3\left|\mu_n - \mu'\right|_{P,2}.
\end{aligned}$$

Given that μ_n and μ' are linear with at most m non-zero coefficients, then $\Delta\mu := (\mu_n - \mu')/|\mu_n - \mu'|_2$ is linear with at most $2m$-non-zero coefficients and $|\Delta\mu|_2 = 1$ by construction. Hence, in probability

$$\begin{aligned}
\left|\mu_n - \mu'\right|_{P,2}^2 &\leq \sup_{\Delta\mu \in \mathcal{L}_0(2m):\ |\mu|_2 \leq 1} \left|(1 - \mathbb{E})|\Delta\mu|_n^2\right| \left|\mu_n - \mu'\right|_{P,2}^2 \\
&\quad + \sup_{\Delta\mu \in \mathcal{L}_0(2m):\ |\mu|_2 \leq 1} \left|2(1 - \mathbb{E})\left\langle Y - \mu', \Delta\mu\right\rangle_n\right| \left|\mu_n - \mu'\right|_{P,2} + \delta_3\left|\mu_n - \mu'\right|_{P,2} \\
&\leq \delta_1\left|\mu_n - \mu'\right|_{P,2}^2 + (\delta_2 + \delta_3)\left|\mu_n - \mu'\right|_{P,2}.
\end{aligned}$$

Solving for $|\mu_n - \mu'|_{P,2}^2$ gives the result as long as $\delta_1 \in [0, 1)$. □

The next result is used to verify some of the conditions in the previous lemma.

Lemma 13. *Under Condition 1 and either Condition 2 or 3, for any arbitrary but fixed $\mu' \in \mathcal{L}$, and positive integer m, the following hold with probability going to one*:

1. $\sup_{\mu \in \mathcal{L}_0(m):\ |\mu|_2 \leq 1} |(1 - \mathbb{E})|\mu|_n^2| \lesssim \sqrt{\frac{m \ln K}{n}}$,
2. $\sup_{\mu \in \mathcal{L}_0(m):\ |\mu|_2 \leq 1} |(1 - \mathbb{E})\langle Y - \mu', \mu\rangle_n| \lesssim \sqrt{\frac{m \ln K}{n}}$.

Proof. Let $\mathcal{S}$ be an arbitrary but fixed subset of $\{1, 2, \dots, K\}$ with cardinality $|\mathcal{S}|$. Then, having fixed $\mathcal{S}$, $\mathfrak{F}_{\mathcal{S}} := \{\mu_{\mathcal{S}} := \sum_{k \in \mathcal{S}} b_k X^{(k)}\colon\ |\mu_{\mathcal{S}}|_2 \leq A\}$ is a linear vector space of dimension $|\mathcal{S}|$. In particular let $\Sigma_{\mathcal{S}}$ be the $m \times m$ dimensional matrix with entries $\{\mathbb{E}X^{(k)}X^{(l)}\colon\ k, l \in \mathcal{S}\}$, and $b_{\mathcal{S}}$ the m dimensional vector with entries $\{b_k\colon\ k \in \mathcal{S}\}$. Then $|\mu_{\mathcal{S}}|_2^2 = b_{\mathcal{S}}^T \Sigma_{\mathcal{S}} b_{\mathcal{S}} \geq 0$, where the superscript T stands for the transpose. In consequence, $\Sigma_{\mathcal{S}} = CC^T$ for some $m \times m$ matrix C. It follows that there is an isometry between $\mathfrak{F}_{\mathcal{S}}$ and $\{a \in \mathbb{R}^m\colon\ a = C^T b_{\mathcal{S}}\}$. Any vector a in this last set satisfies $a^T a = |\mu_{\mathcal{S}}|_2^2$, hence it is contained into the m dimensional sphere of radius A (under the Euclidean norm). By Lemma 14.27 in Bühlmann and van de Geer [55], such sphere has a δ cover of cardinality bounded by $(\frac{2A+\delta}{\delta})^m$ (under the Euclidean norm). Then note that the class of functions $\mathcal{L}_{02}(m, A) := \{\mu \in \mathcal{L}_0(m)\colon\ |\mu|_2 \leq A\} = \bigcup_{|\mathcal{S}| \leq m} \mathfrak{F}_{\mathcal{S}}$. Given that the union is over

$\sum_{s=1}^{m}\binom{K}{s} < mK^m$ number of elements, the covering number of $\mathcal{L}_{02}(m, A)$ is bounded above by $mK^m(\frac{2A+\delta}{\delta})^m$.

An argument in Loh and Wainwright ([48], proof of Lemma 15) allows one to replace the supremum over $\mathcal{L}_{02}(m, A)$ with the maximum over a finite set. Let $\{\mu^{(l)}: l = 1, 2, \dots, N\}$ be an L_2 1/3 cover for $\mathcal{L}_{02}(m, A)$, that is, for any $\mu \in \mathcal{L}_{02}(m, A)$ there is a $\mu^{(l)}$ such that $|\Delta\mu|_2 \le 1/3$, where $\Delta\mu := \mu - \mu^{(l)}$. An upper bound for the cardinality N of such cover has been derived above for arbitrary δ, so for $\delta = 1/3$, $N < mK^m(6A + 1)^m$. For a 1/3 cover, one has that $3\Delta\mu \in \mathcal{L}_{02}(m, A)$ or equivalently $\Delta\mu \in \mathcal{L}_{02}(m, A/3)$. This will be used next. By adding and subtracting quantities such as $(1-\mathbb{E})\langle\mu^{(l)}, \mu\rangle_n$ and using simple bounds, infer that (e.g., Loh and Wainwright [48], proof of Lemma 15),

$$\begin{aligned}
I &:= \sup_{\mu\in\mathcal{L}_{02}(m,A)}\left|(1-\mathbb{E})|\mu|_n^2\right| \\
&\le \max_{l\le N}\left|(1-\mathbb{E})\left|\mu^{(l)}\right|_n^2\right| + 2\sup_{\Delta\mu\in\mathcal{L}_{02}(m,A/3)}\max_{l\le N}\left|(1-\mathbb{E})\left\langle\mu^{(l)}, \Delta\mu\right\rangle_n\right| \\
&\quad + \sup_{\Delta\mu\in\mathcal{L}_{02}(m,A/3)}\left|(1-\mathbb{E})|\Delta\mu|_n^2\right| \\
&= \max_{l\le N}\left|(1-\mathbb{E})\left|\mu^{(l)}\right|_n^2\right| + \frac{2}{3}\sup_{\Delta\mu\in\mathcal{L}_{02}(m,A)}\max_{l\le N}\left|(1-\mathbb{E})\left\langle\mu^{(l)}, \Delta\mu\right\rangle_n\right| \\
&\quad + \frac{1}{9}\sup_{\Delta\mu\in\mathcal{L}_{02}(m,A)}\left|(1-\mathbb{E})|\Delta\mu|_n^2\right| \\
&= \max_{l\le N}\left|(1-\mathbb{E})\left|\mu^{(l)}\right|_n^2\right| + \frac{2}{3}\sup_{\mu\in\mathcal{L}_{02}(m,A)}\left|(1-\mathbb{E})|\mu|_n^2\right| \\
&\quad + \frac{1}{9}\sup_{\mu\in\mathcal{L}_{02}(m,A)}\left|(1-\mathbb{E})|\mu|_n^2\right|.
\end{aligned}$$

This implies that $I := \sup_{\mu\in\mathcal{L}_{02}(m,A)}|(1-\mathbb{E})|\mu|_n^2| \le \frac{9}{2}\max_{l\le N}|(1-\mathbb{E})|\mu^{(l)}|_n^2|$. By a similar argument,

$$\begin{aligned}
II &:= \sup_{\mu\in\mathcal{L}_{02}(m,A)}\left|(1-\mathbb{E})\left\langle Y-\mu', \mu\right\rangle_n\right| \\
&\le \max_{l\le N}\left|(1-\mathbb{E})\left\langle Y-\mu', \mu^{(l)}\right\rangle_n\right| + \sup_{\Delta\mu\in\mathcal{L}_{02}(m,A/3)}\left|(1-\mathbb{E})\left\langle Y-\mu', \Delta\mu\right\rangle_n\right| \\
&= \max_{l\le N}\left|(1-\mathbb{E})\left\langle Y-\mu', \mu^{(l)}\right\rangle_n\right| + \frac{1}{3}\sup_{\mu\in\mathcal{L}_{02}(m,A)}\left|(1-\mathbb{E})\left\langle Y-\mu', \mu\right\rangle_n\right|
\end{aligned}$$

implying $II := \sup_{\mu\in\mathcal{L}_{02}(m,A)}|(1-\mathbb{E})\langle Y-\mu', \mu\rangle_n| \le \frac{3}{2}\max_{l\le N}|(1-\mathbb{E})\langle Y-\mu', \mu^{(l)}\rangle_n|$. Hence, to bound I and II use the above upper bounds together with Lemma 10 and the upper bound for N ($N < mK^m(6A+1)^m$ with $A = 1$). □

The following is a modification of a standard crude result often used to derive consistency, but not convergence rates. However, for the CGA and FWA this will be enough to obtain sharp convergence rates independently of the number of iterations m. Recall $\mu_0(X) := \mathbb{E}[Y|X]$.

Lemma 14. *Let $\mu' \in \mathcal{L}$ be arbitrary, but fixed. Suppose that in probability, for some $\delta_1 \in (0, 1)$ and $\delta_2, \delta_3 > 0$, and for a positive B_m:*

1. $\sup_{\mu\in\mathcal{L}(B_m)} |(1-\mathbb{E})(|Y-\mu|_n^2 - |Y-\mu'|_n^2)| \le \delta_1$;
2. $|\mu' - \mu_0|_2^2 \le \delta_2$;
3. *the sequence $\mu_n \in \mathcal{L}(B_m)$ satisfies $|Y-\mu_n|_n^2 - |Y-\mu'|_n^2 \le \delta_3$.*

Then $|\mu_n - \mu_0|_{P,2} \le \sqrt{\delta_1+\delta_2+\delta_3}$ in probability.

Proof. By simple algebra, $|Y-\mu_n|_{P,2}^2 - |Y-\mu'|_{P,2}^2 = |\mu_n-\mu_0|_{P,2}^2 - |\mu'-\mu_0|_{P,2}^2$. Adding and subtracting $|Y-\mu_n|_n^2 - |Y-\mu'|_n^2$,

$$\begin{aligned}|\mu_n-\mu_0|_{P,2}^2 &\le \left|\mu'-\mu_0\right|_{P,2}^2 + \left[|Y-\mu_n|_{P,2}^2 - \left|Y-\mu'\right|_{P,2}^2\right] - \left[|Y-\mu_n|_n^2 - \left|Y-\mu'\right|_n^2\right] \\ &\quad + \left[|Y-\mu_n|_n^2 - \left|Y-\mu'\right|_n^2\right] \\ &\le \delta_2 + \left[|Y-\mu_n|_{P,2}^2 - \left|Y-\mu'\right|_{P,2}^2\right] - \left[|Y-\mu_n|_n^2 - \left|Y-\mu'\right|_n^2\right] + \delta_3,\end{aligned}$$

where the last step follows by points 2 and 3 in the lemma. However,

$$\begin{aligned}&\left(|Y-\mu_n|_{P,2}^2 - \left|Y-\mu'\right|_{P,2}^2\right) - \left(|Y-\mu_n|_n^2 - \left|Y-\mu'\right|_n^2\right) \\ &\le \sup_{\mu\in\mathcal{L}(B_m)} \left||Y-\mu|_{P,2}^2 - \left|Y-\mu'\right|_{P,2}^2 - \left(|Y-\mu|_n^2 - \left|Y-\mu'\right|_n^2\right)\right| \\ &= \sup_{\mu\in\mathcal{L}(B_m)} \left|(1-\mathbb{E})\left(|Y-\mu|_n^2 - \left|Y-\mu'\right|_n^2\right)\right| \le \delta_1,\end{aligned}$$

where the last inequality follows by assumption. Putting everything together the result follows. □

In what follows, define $\mathcal{L}_{01}(m, B) := \mathcal{L}_0(m) \cap \mathcal{L}_1(B)$, where $\mathcal{L}_1(B) = \mathcal{L}(B)$ the usual linear space of functions with absolute sum of coefficients bounded by B. The next result will be used to verify the conditions of the previous lemma in the case of the CGA and FWA but also as main ingredient to derive consistency rates for non-mixing data in a variety of situations.

Lemma 15. *Suppose Condition 1. For any arbitrary, but fixed $\mu' \in \mathcal{L}_{01}(m, B)$, and $B_m < \infty$,*

$$\mathbb{E} \sup_{\mu\in\mathcal{L}_{01}(m,B_m):\ |\mu-\mu'|_2\le\delta} \left|(1-\mathbb{E})\left(\left|Y-\mu(X)\right|_n^2 - \left|Y-\mu'(X)\right|_n^2\right)\right| \lesssim \mathrm{error}(\delta),$$

where, under either Condition 2 *or* 3,

$$\text{error}(\delta) = \min\left\{\delta\sqrt{\frac{m}{\rho_{2m}}}, B + B_m\right\}\left(1 + \min\left\{\delta\sqrt{\frac{m}{\rho_{2m}}}, B + B_m\right\}\right)\left(\sqrt{\frac{\ln K}{n}}\right)$$

while under Condition 4,

$$\text{error}(\delta) = \min\left\{\delta\sqrt{\frac{m}{\rho_{2m}}}, B + B_m\right\}\left(1 + K^{1/p}\min\left\{\delta\sqrt{\frac{m}{\rho_{2m}}}, B + B_m\right\}\right)\left(\frac{d_{n,p}K^{1/p}}{\sqrt{n}}\right).$$

Proof. Note that $Y = \mu_0 + Z$, where Z is mean zero conditionally on X. Then, by standard algebra

$$\begin{aligned}
&(1-\mathbb{E})|Y-\mu|_n^2 - (1-\mathbb{E})\left|Y-\mu'\right|_n^2 \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} 2Z_i\bigl(\mu'(X_i) - \mu(X_i)\bigr) \\
&\qquad + \frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})\bigl(\mu(X_i) - \mu'(X_i)\bigr)\bigl(\mu(X_i) + \mu'(X_i) - 2\mu_0(X_i)\bigr) \\
&\quad =: I + II,
\end{aligned}$$

using the fact that $\mathbb{E}[Z|X] = 0$ in the equality. The two terms above can be bounded separately, uniformly in μ such that $|\mu - \mu'|_2 \le \delta$. First, let $\mu'(X) = \sum_{k=1}^{K} b'_k X_i^{(k)}$, where by definition of $\mathcal{L}_{01}(m, B)$, only m coefficients are non-zero. Note that for $\mu(X) = \sum_{k=1}^{K} b_k X^{(k)}$ in $\mathcal{L}_{01}(m, B)$, $(\mu' - \mu) \in \mathcal{L}_{01}(2m, B + B_m)$, because μ and μ' are arbitrary, hence do not need to have any variables in common for $2m \le K$ (recall that there are K variables $X^{(k)}$, $k \le K$). Define $c_k := \operatorname{sign}(b'_k - b_k)\sum_{k=1}^{K}|b'_k - b_k|$, $\lambda_k := (b'_k - b_k)/\sum_{k=1}^{K}|b'_k - b_k|$, where there are at most $2m$ non-zero λ_k's by the restriction imposed by $\mathcal{L}_{01}(2m, B + B_m)$. Hence,

$$\mu'(X) - \mu(X) = \sum_{k=1}^{K}\bigl(b'_k - b_k\bigr)X^{(k)} = \sum_{k=1}^{K}\lambda_k c_k X^{(k)},$$

with $|c_k| \le |\mu' - \mu|_{\mathcal{L}}$, and λ_k's in the $2m$ dimensional unit simplex. Given this restrictions, also note that

$$\sqrt{\frac{\rho_{2m}}{2m}}\sum_{k=1}^{K}\bigl|b'_k - b_k\bigr| \le \sqrt{\rho_{2m}\sum_{k=1}^{K}\bigl(b'_k - b_k\bigr)^2} \le \bigl|\mu' - \mu\bigr|_2$$

so that for any $\delta > 0$, $|\mu - \mu'|_2 \le \delta$ implies $|\mu' - \mu|_{\mathcal{L}} \le \delta\sqrt{2m/\rho_{2m}}$ or equivalently $|c_k| \le \min\{\delta\sqrt{2m/\rho_{2m}}, B + B_m\}$. Going from right to left, the above inequality is obtained from the Rayleigh quotient, and by bounding the l_1 norm by $\sqrt{2m}$ times the l_2 norm (e.g., use Jensen inequality of Cauchy–Schwarz). To ease notation, write $\sup_{|\mu-\mu'|_2\le\delta}$ for $\sup_{\mu\in\mathcal{L}_{01}(m,B):\ |\mu-\mu'|_2\le\delta}$.

Then, using the previous remarks, and also noting that the supremum over the unit simplex is achieved at one of the edges of the simplex,

$$\begin{aligned}\mathbb{E} \sup_{|\mu-\mu'|_2\le\delta} |I| &= 2\mathbb{E} \sup_{|\mu-\mu'|_2\le\delta} \left|\frac{1}{n}\sum_{i=1}^n Z_i\left(\sum_{k=1}^K \bigl(b'_k - b_k\bigr)X_i^{(k)}\right)\right| \\ &= 2\mathbb{E} \sup_{|\sum_{k=1}^K \lambda_k c_k X^{(k)}|_2\le\delta} \left|\sum_{k=1}^K \lambda_k c_k \frac{1}{n}\sum_{i=1}^n Z_i X_i^{(k)}\right| \\ &= 2\mathbb{E}\max_{k\le K} \sup_{|c_k|\le\min\{\delta\sqrt{2m/\rho_{2m}}, B+B_m\}} \left|c_k\frac{1}{n}\sum_{i=1}^n Z_i X_i^{(k)}\right| \\ &= 2\min\left\{\delta\sqrt{\frac{2m}{\rho_{2m}}}, B + B_m\right\}\mathbb{E}\max_{k\le K}\left|\frac{1}{n}\sum_{i=1}^n Z_i X_i^{(k)}\right|.\end{aligned}$$

Hence, it is sufficient to bound the expectation of the sequence $(Z_i X_i^{(k)})_{i\ge1}$, which is mean zero by construction. Under Conditions 2 or 3, $\mathbb{E}\max_{k\le K} |\frac{1}{n}\sum_{i=1}^n Z_i X_i^{(k)}| \lesssim \sqrt{\frac{\ln K}{n}}$, by Lemma 10, while under Condition 4,

$$\mathbb{E}\max_{k\le K}\left|\frac{1}{n}\sum_{i=1}^n Z_i X_i^{(k)}\right| \lesssim K^{1/p}\max_{k\le K}\left(\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_i X_i^{(k)}\right|^p\right)^{1/p} \lesssim \frac{d_{n,p}K^{1/p}}{\sqrt{n}}$$

by Lemma 11. To bound the terms in II, note that

$$\mu + \mu' - 2\mu_0 = \mu - \mu' + 2\bigl(\mu' - \mu_0\bigr).$$

Then, recalling $\Delta(X) := (\mu'(X) - \mu_0(X))$,

$$\begin{aligned}\mathbb{E} \sup_{|\mu-\mu'|_2\le\delta} |II| \le{}& \mathbb{E} \sup_{|\sum_{k=1}^K \lambda_k c_k X^{(k)}|_2\le\delta} \left|\frac{1}{n}\sum_{i=1}^n (1-\mathbb{E})\left(\sum_{k=1}^K \lambda_k c_k X_i^{(k)}\right)^2\right| \\ &+ \mathbb{E} \sup_{|\sum_{k=1}^K \lambda_k c_k X^{(k)}|_2\le\delta} \left|\frac{2}{n}\sum_{i=1}^n (1-\mathbb{E})\left(\sum_{k=1}^K \lambda_k c_k X_i^{(k)}\right)\Delta(X_i)\right| \\ =:{}& III + IV.\end{aligned}$$

Using arguments similar for the bound of I,

$$\begin{aligned}III &\le \mathbb{E} \sup_{|\sum_{k=1}^K \lambda_k c_k X^{(k)}|_2\le\delta, |\sum_{l=1}^K \lambda_l c_l X^{(l)}|_2\le\delta} \left|\sum_{k=1}^K \lambda_k c_k \sum_{l=1}^K \lambda_l c_l \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E}) X_i^{(k)} X_i^{(l)}\right| \\ &= \mathbb{E}\max_{k,l\le K} \sup_{|c_k|,|c_l|\le\min\{\delta\sqrt{2m/\rho_{2m}}, B+B_m\}} \left|c_k c_l \frac{1}{n}\sum_{i=1}^n (1-\mathbb{E}) X_i^{(k)} X_i^{(l)}\right|\end{aligned}$$

$$\leq \left(\min\left\{\delta\sqrt{\frac{2m}{\rho_{2m}}}, B + B_m\right\}\right)^2 \mathbb{E}\max_{k,l\leq K}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})X_i^{(k)}X_i^{(l)}\right|.$$

To finish the control of *III*, one can then proceed along the lines of the control of the *I* term:

$$\mathbb{E}\max_{k,l\leq K}\left|\frac{1}{n}\sum_{i=1}^{n}(1-\mathbb{E})X_i^{(k)}X_i^{(l)}\right| \lesssim \begin{cases}\sqrt{\ln K^2}, & \text{under Condition 2 or 3,}\\ \dfrac{d_{n,p}K^{2/p}}{\sqrt{n}}, & \text{under Condition 4.}\end{cases}$$

Similar arguments are used to bound *IV*. Putting these bounds together, and disregarding irrelevant constants, the result follows. □

4.5. Proof of theorems

Proof of Theorem 1. At first, prove the result for the PGA, OGA and RGA. The estimators satisfy $F_m \in \mathcal{L}_0(m)$. Hence, apply Lemma 12. Verify points 1–2 in Lemma 12, using Lemma 13, so that $\delta_1, \delta_2 \lesssim \sqrt{\frac{m\ln K}{n}}$ in Lemma 12. By Lemmas 4, 5 and 6, point 3 in Lemma 12 is verified with δ_3 proportional to $B^{1/3}m^{-1/6}$ for the PGA, $Bm^{-1/2}$ for the OGA and RGA with $\mu' = \mu_B$. Point 4 is satisfied by the remark around (6) for $B \geq B_0$ as required in (8). Hence, in probability, by the triangle inequality,

$$|\mu_0 - F_m|_{P,2} \lesssim \sqrt{\frac{m\ln K}{n}} + |\mu_0 - \mu_B|_2 + \text{algo}(B,m),$$

where $\text{algo}(B,m)$ is the appropriate error term in Lemmas 4, 5, 6.

For the CGA and FWA use Lemma 14 with $\mu' = \mu_B \in \mathcal{L}(B)$, $B = B_m = \bar{B}$, and $\mu_n = F_m$; recall μ_B is the minimizer in (3). In Lemma 14, $\delta_1 \lesssim \bar{B}\sqrt{\frac{\ln K}{n}}$ by Lemma 15 with $m = K$, so that $\mathcal{L}_{01}(m,B) = \mathcal{L}(B)$. By definition of μ_B, in Lemma 14, $\delta_2 = \gamma^2(\bar{B})$. Moreover, $\delta_3 \lesssim \bar{B}^2 m^{-1}$ by Lemmas 6 and 7. Hence, Lemma 14 is verified. □

The proof of Theorem 2 is next.

Proof of Theorem 2. By Lemma 8, $F_m \in \mathcal{L}(B_m)$ in probability, for some suitable B_m depending on the algorithm. The theorem then follows by an application of Lemma 14 with $\mu' = \mu_B \in \mathcal{L}(B)$ for arbitrary B, and $\mu_n = F_m$. In Lemma 14, $\delta_1 \lesssim (B + B_m)^2(\frac{d_{n,p}K^{2/p}}{\sqrt{n}})$ by Lemma 15. Then substitute B_m with the upper bounds given in Lemma 8. Finally, in Lemma 14, $\delta_2 = \gamma(B)$ and $\delta_3 = \text{algo}(B,m)$ by Lemmas 4, 5, or 6. Hence, Lemma 14 and the fact that $\sqrt{\delta_1 + \delta_2 + \delta_3} \leq \sqrt{\delta_1} + \sqrt{\delta_2} + \sqrt{\delta_3}$ imply the result. □

Theorem 3 relies on Theorem 3.4.1 in van der Vaart and Wellner [77], which is here recalled as a lemma for convenience, using the present notation and adapted to the current purposes.

Lemma 16. *Suppose that for any $\delta > \delta_n > 0$, and for $B_m \geq B$, and fixed function $\mu' \in \mathcal{L}_{0,1}(m, B)$:*

1. $\mathbb{E}|Y - \mu(X)|_n^2 - \mathbb{E}|Y - \mu'(X)|_n^2 \gtrsim \mathbb{E}|\mu(X) - \mu'(X)|_n^2$ *for any $\mu \in \mathcal{L}_{0,1}(m, B_m)$ such that $|\mu - \mu'|_2 \leq \delta$;*
2. $\mathbb{E}\sup_{\mu \in \mathcal{L}_{0,1}(m,B_m):\, |\mu-\mu'|_2 \leq \delta} |(1-\mathbb{E})|Y - \mu(X)|_n^2 - (1-\mathbb{E})|Y - \mu'(X)|_n^2| \lesssim \delta \frac{a_n}{n^{1/2}}$ *for some sequence $a_n = \mathrm{o}(n^{1/2})$;*
3. *there is a sequence r_n such that $r_n \lesssim \delta_n^{-1}$ and $r_n \lesssim \frac{n^{1/2}}{a_n}$;*
4. $\Pr(F_m \in \mathcal{L}_{0,1}(m, B_m)) \to 1$, *and* $|Y - F_m|_n^2 \leq |Y - \mu'(X)|_n^2 + \mathrm{O}_P(r_n^{-2})$.

Then $(\mathbb{E}|\mu_0(X') - F_m(X')|_n^2)^{1/2} \lesssim |\mu_0 - \mu'|_2 + r_n^{-1}$ *in probability.*

Here is the proof of Theorem 3.

Proof of Theorem 3. It is enough to verify the conditions in Lemma 16 and then show that one can replace the approximation error w.r.t. $\mu' \in \mathcal{L}_{0,1}(m, B)$ with the one w.r.t. μ_B. To verify point 1 in Lemma 16, restrict attention to μ such that $\mathbb{E}|\mu(X) - \mu'(X)|_n^2 \geq 4\mathbb{E}|\mu_0(X) - \mu'(X)|_n^2$. If this is not the case, the convergence rate (error) is proportional to $\mathbb{E}|\mu_0(X) - \mu'(X)|_n^2$ and Lemma 16 would apply trivially. Hence, suppose this is not the case. By standard algebra,

$$\begin{aligned}\mathbb{E}\big|Y - \mu(X)\big|_n^2 - \mathbb{E}\big|Y - \mu'(X)\big|_n^2 &= \mathbb{E}\big|\mu_0(X) - \mu(X)\big|_n^2 - \mathbb{E}\big|\mu_0(X) - \mu'(X)\big|_n^2 \\ &\geq \tfrac{1}{4}\mathbb{E}\big|\mu(X) - \mu'(X)\big|_n^2,\end{aligned}$$

where the inequality follows by problem 3.4.5 in van der Vaart and Wellner [77]. Hence, point 1 in Lemma 16 is satisfied. By construction, F_m has at most m non-zero coefficients. By this remark and Lemma 8, $F_m \in \mathcal{L}_{0,1}(m, B_m)$ with $B_m = \mathrm{O}_p(m^{1/2})$ for the PGA, $B_m = \mathrm{O}_p(m^{1/2}/\rho_m^{1/2})$ for the OGA, and $B_m = \mathrm{O}_p(m^{1/2}/\rho_{m,n}^{1/2})$ for the RGA if $\mathrm{algo}(B, m) = B^2/m = \mathrm{O}(1)$, which holds by the conditions in the theorem. The equality $\mathrm{algo}(B, m) = B^2/m$ follows by Lemma 6. By Lemma 9, if $d_{n,p} m K^{2/p} n^{-1/2} = \mathrm{o}(\rho_m)$, then $\rho_{m,n}^{-1} = \mathrm{O}_p(\rho_m^{-1})$. By the conditions in the theorem, $\rho_m > 0$, and $d_{n,p} m K^{2/p} n^{-1/2} \lesssim \mathrm{error}(B, K, n, m) = \mathrm{o}(1)$. Hence, infer that $B_m = \mathrm{O}_p(m^{1/2})$ for the OGA and RGA. Hence, point 2 in Lemma 16, is satisfied for any δ and $a_n = m^{1/2}(B + m^{1/2}) d_{n,p} K^{2/p}$ by Lemma 15, where

$$\mathrm{error}(\delta) \lesssim \delta m^{1/2}\big(B + m^{1/2}\big)\left(\frac{d_{n,p} K^{2/p}}{\sqrt{n}}\right)$$

using the fact that $\rho_{2m} > 0$ and $B_m = \mathrm{O}_p(m^{1/2})$.

It follows that point 3 in Lemma 16 is satisfied by $r_n = n^{1/2}/[m^{1/2}(B + m^{1/2}) d_{n,p} K^{2/p}]$. Moreover, by Lemma 4, 5 and 6

$$|Y - F_m|_n^2 \leq \big|Y - \mu'(X)\big|_n^2 + \mathrm{O}_p\big(u_n^{-2}\big),$$

with u_n^{-2} as given in the aforementioned lemmas because $\mu' \in \mathcal{L}_{0,1}(m, B) \subseteq \mathcal{L}(B)$. Since point 4 in Lemma 16 requires $u_n = \mathrm{O}(r_n)$, the actual rate of convergence is $u_n^{-1} \vee r_n^{-1} \leq u_n^{-1} + r_n^{-1}$ as stated in the theorem.

It is now necessary to replace the approximation error $\mathbb{E}|\mu_0(X) - \mu'(X)|_n^2$ with $\gamma(B) := \mathbb{E}|\mu_0(X) - \mu_B(X)|_n^2$. To this end, consider Lemmas 4, 5 and 6 with the empirical norm $|\cdot|_n$ replaced by $|\cdot|_{P,2}$. Going through the proof, the results are seen to hold as well with the same error rate (implicitly using Condition 1). Hence, note that, by standard algebra,

$$\left|Y - \mu'(X)\right|_{P,2}^2 - \mathbb{E}\left|Y - \mu_B(X)\right|_{P,2}^2 = \mathbb{E}\left|\mu_0(X) - \mu'(X)\right|_{P,2}^2 - \mathbb{E}\left|\mu_0(X) - \mu_B(X)\right|_{P,2}^2.$$

The above display together with the previous remark and Lemmas 4, 5 and 6 imply that

$$\mathbb{E}\left|\mu_0(X) - \mu'(X)\right|_2^2 \leq \mathbb{E}\left|\mu_0(X) - \mu_B(X)\right|_n^2 + \mathrm{O}\left(u_n^{-2}\right),$$

with u_n as defined above. Hence, Lemma 16 together with the above display gives the result which is valid for any B. □

4.6. Proof of Lemmas 1, 2 and 3

Proof of Lemma 1. If $B' \geq B$, the lemma is clearly true because $\mathcal{L}(B) \subseteq \mathcal{L}(B')$. Hence, assume $B' < B$. W.n.l.g. assume that $\sum_k |b_k| = B$, as $\mu \in \mathcal{L}(B)$. Let $\lambda_k = (|b_k|/B) \geq 0$, and $c_k = B(b_k/|b_k|)$. Then $\mu = \sum_k \lambda_k c_k X^{(k)}$. Define

$$\mu'' = \sum_k \lambda_k \left(\frac{B'}{B}\right) c_k X^{(k)}$$

and note that $\mu'' \in \mathcal{L}(B')$ by construction and $B'/B < 1$. Then

$$\begin{aligned}
\inf_{\mu' \in \mathcal{L}(B')} \left|\mu' - \mu\right|_2^2 &\leq \left|\sum_k \lambda_k c_k X^{(k)} - \sum_k \lambda_k \left(\frac{B'}{B}\right) c_k X^{(k)}\right|_2^2 \\
&= \left[1 - \left(\frac{B'}{B}\right)\right]^2 \sum_{k,l} \lambda_k c_k \lambda_l c_l \mathbb{E} X^{(k)} X^{(l)} \\
&\leq \left[1 - \left(\frac{B'}{B}\right)\right]^2 \left(\sum_k |\lambda_k c_k|\right)^2 = \left[1 - \left(\frac{B'}{B}\right)\right]^2 B^2,
\end{aligned}$$

where the second inequality follows using the fact that $|\mathbb{E}X^{(k)}X^{(l)}| \leq \mathbb{E}|X^{(k)}|^2 = 1$ and the last equality because $\sum_k |\lambda_k c_k| = \sum_k |b_k| = B$. □

Proof of Lemma 2. At first, show (5). By independence, $|ZX^{(k)}|_p = |Z|_p |X^{(k)}|_p < \infty$. Let A_{kl} be the (k,l) entry in A and similarly for H_{kl}. By stationarity, and the fact that A is diagonal, the lth entry in W_i is $W_{il} = \sum_{s=0}^{\infty} A_{ll}^s \varepsilon_{i-s,l}$, and by definition $X_i^{(k)} = \sum_{l=1}^{L} H_{kl} \sum_{s=0}^{\infty} A_{ll}^s \varepsilon_{i-s,l}$. Hence, by Minkowski inequality, and the fact that A_{ll}^s decays exponentially fast because less than one in absolute value, and the fact that $\{H_{kl}\colon l = 1, 2, \ldots, L\}$ is in the unit simplex, $|X^{(k)}|_{2p} \leq$

$\max_{l\le L}\sum_{s=0}^{\infty}|A_{ll}^{s}||\varepsilon_{i-s,l}|_{2p}<\infty$. Finally, by Hölder's inequality, the Lipschitz condition for g, and Minkowski inequality, for any $\mu_B(X)=\sum_{k=1}^{K}X^{(k)}b_k$,

$$\begin{aligned}\left|\Delta(X)X^{(k)}\right|_p &\le \left|\sum_{l=1}^{K}\left(\lambda_l+|b_l|\right)\left|X^{(l)}\right|\right|_{2p}\left|X^{(k)}\right|_{2p}\lesssim(1+B)\max_l\left|X^{(l)}\right|_{2p}\left|X^{(k)}\right|_{2p}\\ &\le(1+B)\max_k\left|X^{(k)}\right|_{2p}^{2}<\infty.\end{aligned}$$

This completes the proof of (5). Geometric absolute regularity follows using the fact that by construction, the mixing coefficients of $(X_i)_{i\in\mathbb{Z}}$ are equal to the mixing coefficients of $(W_i)_{i\in\mathbb{Z}}$ because X_i is just a linear transformation of W_i (i.e., the sigma algebras generated by the two processes are the same). The process $(W_i)_{i\in\mathbb{Z}}$ follows a L dimensional stationary AR(1) model with i.i.d. innovations having a density w.r.t. the Lebesgue measure. Hence, Theorem 1 in Mokkadem [51] says that the vector autoregressive process $(W_i)_{i\in\mathbb{Z}}$ is absolutely regular with geometrically decaying mixing coefficients as long as L is bounded. By independence, the sigma algebra generated by $(W_i)_{i\in\mathbb{Z}}$ and $(Z_i)_{i\in\mathbb{Z}}$ are independent. Then Theorem 5.1 Bradley [14] says that the mixing coefficient of $(W_i,Z_i)_{\in\mathbb{Z}}$ are bounded by the sum of the mixing coefficients of $(W_i)_{\in\mathbb{Z}}$ and $(Z_i)_{i\in\mathbb{Z}}$. Since the latter mixing coefficients are zero at any non-zero lags because of independence, geometric beta mixing follows, and Condition 3 holds. □

Proof of Lemma 3. By assumption $X_i=W_i$. Andrews [1] and Bradley [13] show that the AR(1) model as in the lemma is not strong mixing, hence is not absolutely regular and Condition 3 fails. Consider each term in the sum in Condition 4 separately. First, by independence, $|\mathbb{E}_0Z_iX_i^{(k)}|_p=0$ for any $i>0$. Second, using the infinite MA representation of the AR(1), the fact that the error terms are i.i.d., and then the triangle inequality

$$\begin{aligned}\left|\mathbb{E}_0(1-\mathbb{E})\left|X_i^{(k)}\right|^2\right|_p &=\left|\sum_{s,r=i}^{\infty}A_{kk}^{s}A_{kk}^{r}(1-\mathbb{E})\varepsilon_{i-s,k}\varepsilon_{i-r,k}\right|_p\\ &\le 2\sum_{s,r=i}^{\infty}\left|A_{kk}^{s}\right|\left|A_{kk}^{r}\right||\varepsilon_{i-s,k}\varepsilon_{i-r,k}|_p\lesssim A_{kk}^{2i}\end{aligned}$$

which is summable. Third, by the triangle inequality,

$$\begin{aligned}\left|\mathbb{E}_0(1-\mathbb{E})\Delta(X_i)X_i^{(k)}\right|_p &=\left|\mathbb{E}_0(1-\mathbb{E})\left(g\left(X_i^{(l)};l\le K\right)-\sum_{k=1}^{K}X_i^{(l)}b_l\right)X_i^{(k)}\right|_p\\ &\le\left|\mathbb{E}_0(1-\mathbb{E})g\left(X_i^{(l)};l\le K\right)X_i^{(k)}\right|_p+\left|\mathbb{E}_0(1-\mathbb{E})X_i^{(k)}\sum_{l=1}^{K}X_i^{(l)}b_l\right|_p\\ &=:I+II.\end{aligned}$$

Consider each term separately. Define $X_{i0}^{(k)} := \sum_{s=0}^{i-1} A_{kk}^s \varepsilon_{i-s,k}$ and $X_{i1}^{(k)} := \sum_{s=i}^{\infty} A_{kk}^s \varepsilon_{i-s,k}$, and note that $X_i^{(k)} = X_{i0}^{(k)} + X_{i1}^{(k)}$. By simple algebraic manipulations and repeated use of Minkowski inequality,

$$\begin{aligned} I &\leq \bigl|\mathbb{E}_0(1-\mathbb{E})g\bigl(X_{i0}^{(l)}; l \leq K\bigr)X_i^{(k)}\bigr|_p + \bigl|\mathbb{E}_0(1-\mathbb{E})\bigl(g\bigl(X_i^{(l)}; l \leq K\bigr) - g\bigl(X_{i0}^{(l)}; l \leq K\bigr)\bigr)X_i^{(k)}\bigr|_p \\ &\leq \bigl|(\mathbb{E}_0-\mathbb{E})g\bigl(X_{i0}^{(l)}; l \leq K\bigr)X_{i0}^{(k)}\bigr|_p + \bigl|(\mathbb{E}_0-\mathbb{E})g\bigl(X_{i0}^{(l)}; l \leq K\bigr)X_{i1}^{(k)}\bigr|_p \\ &\quad + \bigl|(\mathbb{E}_0-\mathbb{E})\bigl(g\bigl(X_i^{(l)}; l \leq K\bigr) - g\bigl(X_{i0}^{(l)}; l \leq K\bigr)\bigr)X_{i0}^{(k)}\bigr|_p. \end{aligned}$$

The first term on the right-hand side of the second inequality is zero by construction when taking expectations $(\mathbb{E}_0 - \mathbb{E})$. To bound the second term, note that by the properties of g, and Minkowski inequality,

$$\begin{aligned} \bigl|(\mathbb{E}_0-\mathbb{E})g\bigl(X_{i0}^{(l)}; l \leq K\bigr)X_{i1}^{(k)}\bigr|_p &\leq \left|(\mathbb{E}_0+\mathbb{E})\sum_{l\leq K}\lambda_l\bigl|X_{i0}^{(l)}\bigr|\bigl|X_{i1}^{(k)}\bigr|\right|_p \\ &\leq \left|\left(\sum_{l\leq K}\lambda_l\mathbb{E}_0\bigl|X_{i0}^{(l)}\bigr|\right)\bigl|X_{i1}^{(k)}\bigr|\right|_p + \left|\sum_{l\leq K}\lambda_l\mathbb{E}\bigl|X_{i0}^{(l)}\bigr|\mathbb{E}\bigl|X_{i1}^{(k)}\bigr|\right|_p \\ &\lesssim \bigl|X_{i1}^{(k)}\bigr|_p \lesssim \sum_{s=i}^{\infty}\bigl|A_{kk}^s\bigr| \end{aligned}$$

by the independence of $X_{i0}^{(k)}$ and $X_{i1}^{(k)}$ and the existence of p moments. The third term was bounded in a similar way. Hence, $I \lesssim \sum_{s=i}^{\infty}|A_{kk}^s| \lesssim |A_{kk}^i|$. Finally,

$$\begin{aligned} II &\leq B\max_l\bigl|\mathbb{E}_0(1-\mathbb{E})X_i^{(k)}X_i^{(l)}\bigr|_p \\ &\leq \left|\sum_{s,r=0}^{i-1}A_{kk}^sA_{ll}^r(\mathbb{E}_0-\mathbb{E})\varepsilon_{i-s,k}\varepsilon_{i-r,l}\right|_p + \left|\sum_{s,r=i}^{\infty}A_{kk}^sA_{ll}^r(\mathbb{E}_0-\mathbb{E})\varepsilon_{i-s,k}\varepsilon_{i-r,l}\right|_p \\ &\lesssim A_{kk}^{2i} \end{aligned}$$

as the first term is exactly zero. Clearly, both I and II are summable and the lemma is proved. □

4.7. Proof of Example 4

For simplicity, write ε_i in place of $\varepsilon_{i,k}$. By independence of the ε_i's and stationarity,

$$\begin{aligned} \mathbb{E}_0(1-\mathbb{E})\bigl|X_i^{(k)}\bigr|^2 &= \mathbb{E}_0(1-\mathbb{E})\sum_{s,r=0}^{\infty}a_sa_r\varepsilon_{i-s}\varepsilon_{i-r} = \sum_{s,r\geq i}a_sa_r\bigl[(1-\mathbb{E})\varepsilon_{i-s}\varepsilon_{i-r}\bigr] \\ &= \sum_{s\geq i}a_s^2(1-\mathbb{E})\varepsilon_{i-s}^2 + 2\sum_{r>s\geq i}a_sa_r\varepsilon_{i-s}\varepsilon_{i-r} =: I + II. \end{aligned}$$

For $i > 0$, define $\bar{a}_s = i^{(1+\epsilon)} a_s^2$ and $\bar{a} := \sum_{s \geq i} \bar{a}_s$ and note that $\bar{a}$ depends on i but is finite for any i because $i^{(1+\epsilon)} a_s^2 = i^{(1+\epsilon)} s^{-(1+\epsilon)} \leq 1$ for $s \geq i$ (recall the definition of a_s). Then, by the definition of $\bar{a}_s$ and then by Jensen inequality,

$$\begin{aligned}\mathbb{E}|I|^p &\leq \mathbb{E} \sum_{s \geq i} \left(\frac{\bar{a}_s}{\bar{a}} \right) \left| i^{-(1+\epsilon)} \bar{a} (1 - \mathbb{E}) \varepsilon_{i-s}^2 \right|^p \\ &= \sum_{s \geq i} \left(\frac{\bar{a}_s}{\bar{a}} \right) i^{-(1+\epsilon)p} \bar{a}^p \mathbb{E} \left| (1 - \mathbb{E}) \varepsilon_{i-s}^2 \right|^p \\ &\leq i^{-(1+\epsilon)p} \bar{a}^p \max_s \mathbb{E} \left| (1 - \mathbb{E}) \varepsilon_s^2 \right|^p\end{aligned}$$

because $\sum_{s \geq i} (\frac{\bar{a}_s}{\bar{a}}) = 1$ and the $\bar{a}_s \geq 0$. The above display implies that $|I|_p \lesssim i^{-(1+\epsilon)}$. It remains to bound II. For any random variable W such that $\mathbb{E} \exp\{|W|/\tau\} \leq 4$ for some $\tau > 0$, it is clear that

$$\mathbb{E}|W/\tau|^p \leq p! \bigl(\mathbb{E} \exp\bigl\{ |W|/\tau \bigr\} - 1 \bigr) \leq p! \times 3$$

using Taylor series expansion. This implies that $(\mathbb{E}|W|^p)^{1/p} \leq 3p\tau$ for such τ if it exists. Hence, apply this inequality to bound $\mathbb{E}|II|^p$. Noting that $\mathbb{E} \exp\{\tau^{-1}|II|\} \leq \mathbb{E} \exp\{\tau^{-1} II\} + \mathbb{E} \exp\{-\tau^{-1} II\}$, it is enough to bound $\mathbb{E} \exp\{\tau^{-1} II\}$. By Gaussianity, independence of the ε_i's, and the fact that $\exp\{\cdot\}$ is non-negative, letting $\mathbb{E}_i$ be expectation conditional on ε_i and its past,

$$\begin{aligned}\mathbb{E} \exp\bigl\{ \tau^{-1} II \bigr\} &= \prod_{r > s \geq i} \mathbb{E} \exp\bigl\{ \tau^{-1} 2 a_s \varepsilon_{i-s} a_r \varepsilon_{i-r} \bigr\} = \prod_{r > s \geq i} \mathbb{E} \exp\bigl\{ \mathbb{E}_{i-r} \bigl(\tau^{-1} 2 a_s \varepsilon_{i-s} a_r \varepsilon_{i-r} \bigr)^2 \bigr\} \\ &= \prod_{r > s \geq i} \mathbb{E} \exp\bigl\{ \bigl(\tau^{-1} 2 a_s a_r \bigr)^2 \varepsilon_{i-r}^2 \bigr\} = \prod_{r > s \geq i} \mathbb{E} \exp\left\{ 4 \bar{a}^2 i^{-2(1+\epsilon)} \tau^{-2} \frac{\bar{a}_s}{\bar{a}} \frac{\bar{a}_r}{\bar{a}} \varepsilon_{i-r}^2 \right\},\end{aligned}$$

where the last three steps use the properties of the moment generating function of a Gaussian random variable and the definition of $\bar{a}_s$ and $\bar{a}$, as used in the control of I. Hence, setting $\tau = 4 \bar{a} i^{-(1+\epsilon)}$, and recalling that $\bar{a}_s / \bar{a} \leq 1$ by construction, the above is then bounded by

$$\max_{r \geq i} \mathbb{E} \exp\left\{ \frac{\varepsilon_{i-r}^2}{4} \right\} = \int_{\mathbb{R}} \mathrm{e}^{z^2/4} \frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} \, \mathrm{d}z = \sqrt{2},$$

where the two equalities follow from the fact that ε_{i-r}^2 is a standard normal random variable, and then performing the integration. The above two display show that for $\tau = 4 \bar{a} i^{-(1+\epsilon)}$, $\mathbb{E} \exp\{\tau^{-1}|II|\} \leq \exp\{\tau^{-1} II\} + \exp\{-\tau^{-1} II\} \leq 2\sqrt{2} < 4$, which implies $|II|_p \lesssim \bar{a} i^{-(1+\epsilon)}$. The upper bounds for the L_p norms of I and II imply that $|\mathbb{E}_0 (1 - \mathbb{E}) |X_i^{(k)}|^2|_p \lesssim i^{-(1+\epsilon)}$.

References

[1] Andrews, D.W.K. (1984). Nonstrong mixing autoregressive processes. *J. Appl. Probab.* **21** 930–934. MR0766830

[2] Audrino, F. and Barone-Adesi, G. (2006). A dynamic model of expected bond returns: A functional gradient descent approach. *Comput*. *Statist*. *Data Anal*. **51** 2267–2277. MR2307500

[3] Audrino, F. and Bühlmann, P. (2003). Volatility estimation with functional gradient descent for very high-dimensional financial time series. *J*. *Comput*. *Finance* **6** 65–89.

[4] Audrino, F. and Bühlmann, P. (2009). Splines for financial volatility. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **71** 655–670. MR2749912

[5] Barron, A.R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans*. *Inform*. *Theory* **39** 930–945. MR1237720

[6] Barron, A.R., Cohen, A., Dahmen, W. and DeVore, R.A. (2008). Approximation and learning by greedy algorithms. *Ann*. *Statist*. **36** 64–94. MR2387964

[7] Bartlett, P.L., Mendelson, S. and Neeman, J. (2012). ℓ_1-regularized linear regression: Persistence and oracle inequalities. *Probab*. *Theory Related Fields* **154** 193–224. MR2981422

[8] Basrak, B., Davis, R.A. and Mikosch, T. (2002). Regular variation of GARCH processes. *Stochastic Process*. *Appl*. **99** 95–115. MR1894253

[9] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. MR3001131

[10] Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann*. *Statist*. **39** 82–130. MR2797841

[11] Bickel, P.J. and Bühlmann, P. (1999). A new mixing notion and functional central limit theorems for a sieve bootstrap in time series. *Bernoulli* **5** 413–446. MR1693612

[12] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J*. *Eur*. *Math*. *Soc*. (*JEMS*) **3** 203–268. MR1848946

[13] Bradley, R.C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics* (*Oberwolfach*, 1985). *Progr*. *Probab*. *Statist*. **11** (E. Eberlein and M.S. Taqqu, eds.) 165–192. Boston, MA: Birkhäuser. MR0899990

[14] Bradley, R.C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab*. *Surv*. **2** 107–144. MR2178042

[15] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann*. *Statist*. **34** 559–583. MR2281878

[16] Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

[17] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Heidelberg: Springer. MR2807761

[18] Bühlmann, P. and Yu, B. (2003). Boosting with the L_2 loss: Regression and classification. *J*. *Amer*. *Statist*. *Assoc*. **98** 324–339. MR1995709

[19] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron*. *J*. *Stat*. **1** 169–194. MR2312149

[20] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. *Ann*. *Statist*. **35** 1674–1697. MR2351101

[21] Burman, P., Chow, E. and Nolan, D. (1994). A cross-validatory method for dependent data. *Biometrika* **81** 351–358. MR1294896

[22] Burman, P. and Nolan, D. (1992). Data-dependent estimation of prediction functions. *J*. *Time Series Anal*. **13** 189–207. MR1168164

[23] Cai, T.T. and Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans*. *Inform*. *Theory* **57** 4680–4688. MR2840484

[24] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann*. *Statist*. **35** 2313–2351. MR2382644

[25] Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica* **66** 289–314. MR1612238

[26] Clarkson, K.L. (2010). Coresets, sparse greedy approximation, and the Frank–Wolfe algorithm. *ACM Trans*. *Algorithms* **6** Art. 63, 30. MR2760426

[27] Daubechies, I., Defrise, M. and De Mol, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm*. *Pure Appl*. *Math*. **57** 1413–1457. MR2077704

[28] Dedecker, J. and Doukhan, P. (2003). A new covariance inequality and applications. *Stochastic Process*. *Appl*. **106** 63–80. MR1983043

[29] DeVore, R.A. and Temlyakov, V.N. (1996). Some remarks on greedy algorithms. *Adv*. *Comput*. *Math*. **5** 173–187. MR1399379

[30] Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann*. *Statist*. **26** 879–921. MR1635414

[31] Doukhan, P. and Louhichi, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Process*. *Appl*. **84** 313–342. MR1719345

[32] Doukhan, P., Massart, P. and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann*. *Inst*. *H*. *Poincaré Probab*. *Statist*. **31** 393–427. MR1324814

[33] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J*. *Amer*. *Statist*. *Assoc*. **78** 316–331. MR0711106

[34] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann*. *Statist*. **32** 407–499. MR2060166

[35] Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Nav*. *Res*. *Logist*. *Q*. **3** 95–110. MR0089102

[36] Freund, R.M., Grigas, P. and Mazumder, R. (2013). AdaBoost and forward stagewise regression are first-order convex optimization methods. Preprint. Available at http://web.mit.edu/rfreund/www/FOM-nips2013-v85.0-non-nips.pdf.

[37] Friedman, J., Hastie, T., Höfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann*. *Appl*. *Stat*. **1** 302–332. MR2415737

[38] Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann*. *Statist*. **34** 2367–2386. MR2291503

[39] Greenshtein, E. and Ritov, Y. (2004). High-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 939–1105.

[40] Huang, C., Cheang, G.L.H. and Barron, A.R. (2008). Risk of penalized least squares, greedy selection and L1 penalization for flexible function libraries. Ph.D. Thesis, Yale Univ., ProQuest LLC, Ann Arbor, MI. Available at http://www.stat.yale.edu/~arb4/publications_files/RiskGreedySelectionAndL1penalization.pdf. MR2711791

[41] Hurvich, C.M., Simonoff, J.S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **60** 271–293. MR1616041

[42] Jaggi, M. (2013). Revisiting Frank–Wolfe: Projection-free sparse convex optimization. *J*. *Mach*. *Learn*. *Res*. *Workshop Conf*. *Proc*. **28** 427–435. Supplementary material available at http://jmlr.org/proceedings/papers/v28/jaggi13-supp.pdf.

[43] Jones, L.K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann*. *Statist*. **20** 608–613. MR1150368

[44] Klemelä, J. (2007). Density estimation with stagewise optimization of the empirical risk. *Mach*. *Learn*. **67** 169–195.

[45] Konyagin, S.V. and Temlyakov, V.N. (1999). Rate of convergence of pure greedy algorithm. *East J*. *Approx*. **5** 493–499. MR1738484

[46] Li, J.Q. and Barron, A.R. (2000). Mixture density estimation. In *Advances in Neural Information Processing Systems* (S.A. Solla, T.K. Leen and K.-R. Mueller, eds.) **12** 279–285. Cambridge, MA: MIT Press.

[47] Livshitz, E.D. and Temlyakov, V.N. (2003). Two lower estimates in greedy approximation. *Constr. Approx*. **19** 509–523. MR1998902

[48] Loh, P.-L. and Wainwright, M.J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist*. **40** 1637–1664. MR3015038

[49] Lutz, R.W. and Bühlmann, P. (2006). Boosting for high-multivariate responses in high-dimensional linear regression. *Statist. Sinica* **16** 471–494. MR2267246

[50] Mallat, S. and Zhang, Z. (1993). Matching pursuits with time–frequency dictionaries. *IEEE Trans. Signal Process*. **41** 3397–3415.

[51] Mokkadem, A. (1988). Mixing properties of ARMA processes. *Stochastic Process. Appl*. **29** 309–315. MR0958507

[52] Mokkadem, A. (1990). Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. H. Poincaré Probab. Statist*. **26** 219–260. MR1063750

[53] Nickl, R. and van de Geer, S. (2013). Confidence sets in sparse regression. *Ann. Statist*. **41** 2852–2876. MR3161450

[54] Pati, Y.C., Rezaiifar, R. and Krishnaprasad, P.S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of* 27*th Asilomar Conference on Signals, Systems and Computers* **1** 40–44. Pacific Grove, CA: IEEE.

[55] Peligrad, M., Utev, S. and Wu, W.B. (2007). A maximal $\mathbb{L}_p$-inequality for stationary sequences and its applications. *Proc. Amer. Math. Soc*. **135** 541–550. MR2255301

[56] Pesaran, M.H., Pettenuzzo, D. and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *Rev. Econ. Stud*. **73** 1057–1084. MR2260756

[57] Pesaran, M.H. and Pick, A. (2011). Forecast combination across estimation windows. *J. Bus. Econom. Statist*. **29** 307–318. MR2808603

[58] Pollard, D. (2002). Maximal inequalities via bracketing with adaptive truncation. *Ann. Inst. H. Poincaré Probab. Statist*. **38** 1039–1052. MR1955351

[59] Rakhlin, A., Panchenko, D. and Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM Probab. Stat*. **9** 220–229. MR2148968

[60] Rio, E. (2000). *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants*. *Mathématiques & Applications* (*Berlin*) [*Mathematics & Applications*] **31**. Berlin: Springer. MR2117923

[61] Sancetta, A. (2010). Bootstrap model selection for possibly dependent and heterogeneous data. *Ann. Inst. Statist. Math*. **62** 515–546. MR2608461

[62] Sancetta, A. (2013). A recursive algorithm for mixture of densities estimation. *IEEE Trans. Inform. Theory* **59** 6893–6906. MR3106872

[63] Sancetta, A. (2015). A nonparametric estimator for the covariance function of functional data. *Econometric Theory*. To appear.

[64] Stock, J.H. and Watson, M.W. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger* (R.F. Engle and H. White, eds.) 1–44. Oxford: Oxford Univ. Press.

[65] Stock, J.H. and Watson, M.W. (2003). How did leading indicator forecasts perform during the 2001 recession? *Federal Reserve Bank of Richmond Economic Quarterly* **89** 71–90.

[66] Stock, J.H. and Watson, M.W. (2004). Combination forecasts of output growth in a seven-country data set. *J. Forecast*. **23** 405–430.

[67] Temlyakov, V. (2011). *Greedy Approximation*. Cambridge: Cambridge Univ. Press. MR2848161

[68] Temlyakov, V.N. (2000). Weak greedy algorithms. *Adv. Comput. Math*. **12** 213–227. MR1745113

[69] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[70] Tibshirani, R.J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.* **40** 1198–1232. MR2985948

[71] Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109** 475–494. MR1835069

[72] Tsybakov, A.B. (2003). Optimal rates of aggregation. In *Proceedings of COLT*-2003. *Lecture Notes in Artificial Intelligence* 303–313. Heidelberg: Springer.

[73] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer. MR2724359

[74] van de Geer, S. (2014). On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Stat.* **8** 543–574. MR3211024

[75] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

[76] van de Geer, S.A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

[77] van de Vaart, A. and Wellner, J.A. (2000). *Weak Convergence and Empirical Processes*. New York: Springer.

[78] Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131. MR1614596

[79] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[80] Zhang, C.-H. and Zhang, S.S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

[81] Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.* **10** 555–568. MR2491749

[82] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967

Received August 2013 and revised December 2014