Sparse multi-task regression with $\mathbb{L}_2$-Boosting algorithm

Magali Champion, Christine Cierco-Ayrolles, Sébastien Gadat, Matthieu Vignes

Abstract

We are interested in the analysis of $\mathbb{L}_2$-Boosting algorithms for linear regressions. Some consistency result has already been proved for high-dimensional models, when the number of predictor grows exponentially with the sample size n. We propose a new result for Weak Greedy Algorithms, which deals with the support recovery, provided reasonable assumptions on the regression parameter. To clarify all the proofs, we also present some results in the deterministic case. Finally, we propose two multi-task versions of $\mathbb{L}_2$-Boosting for which we can extend these stability results provided assumptions on the sparsity of the model.

Keywords: Boosting, Regression, Sparsity, High dimension.

1. Introduction

2. Greedy algorithms

In this section, we describe some essential and useful results on greedy algorithms which build approximations of any functional data f by stepwise iterations. In the deterministic case (*i.e.* noiseless setting), we will refer to 'approximations' of f. In the noisy case, these approximations of f will be designated as 'sequential estimators'. Results on Weak Greedy Algorithms of this section are deduced from Temlyakov [1] to our particular setting. We slightly enrich the presentation by adding some supplementary shrinkage parameters, which offers additional flexibility in the noisy setting. Indeed, it will be necessary to understand the behaviour of the WGA with shrinkage to show statistical consistency of Boosting method.

2.1. Reminders on Weak Greedy Algorithm (WGA)

Let H be an Hilbert space, and $\|.\|$ denotes its associated norm, which is derived from the inner product $\langle , \rangle$ on H. We define a *dictionary* as a (finite) subset $\mathcal{D} = (g_1, \ldots, g_p)$ of H, which satisfies

$$\forall g_i \in \mathcal{D}, \quad \|g_i\| = 1 \quad \text{and} \quad \overline{\text{Span}\, \mathcal{D}} = H.$$

The coherence of such dictionary $\mathcal{D}$, $\rho_{\mathcal{D}}$, associated to the inner product in H is then defined as:

$$\rho_{\mathcal{D}} = \max_{1 \leq i \neq j \leq p} |\langle g_i, g_j \rangle|.$$

Of course, when the dictionary is orthogonal, the coherence is null, which is an extreme case. In this work, we are primarily interested in non-orthogonal dictionaries, since it is a common, if not universal, setting of real high dimensional data sets. Since no confusion can occur in the sequel, the subscript $\mathcal{D}$ is omitted and the coherence is denoted ρ.

Greedy algorithms generate iterative approximations of any $f \in H$, using linear combination of elements of $\mathcal{D}$. Mimicking notations of [1], denote $G_k(f)$ (resp. $R_k(f)$) the approximation of f (resp. the residual) at step k of the algorithm. These quantities are linked with the following equation:

$$R_k(f) = f - G_k(f).$$

Algorithm 1 Weak Greedy Algorithm (WGA)

Require: function f, $(\nu, \gamma) \in (0,1]^2$ (shrinkage parameters), k_{up} (number of iterations.)

Initialisation: $G_0(f) = 0$ and $R_0(f) = f$.

for $k = 1$ to k_{up} **do**

Step 1 Select φ_k in $\mathcal{D}$ such that:

$$|\langle \varphi_k, R_{k-1}(f) \rangle| \geq \quad \nu \ \max_{g \in \mathcal{D}} \ |\langle g, R_{k-1}(f) \rangle|, \tag{1}$$

Step 2 Compute the current approximation and residual:

$$\begin{aligned} G_k(f) &= G_{k-1}(f) + \gamma \langle R_{k-1}(f), \varphi_k \rangle \varphi_k \\ R_k(f) &= R_{k-1}(f) - \gamma \langle R_{k-1}(f), \varphi_k \rangle \varphi_k, \end{aligned} \tag{2}$$

end for

At step k, we select $\varphi_k \in \mathcal{D}$ which provides a sufficient amount of information on residual $R_{k-1}(f)$. The first shrinkage parameter ν stands for a tolerance towards the optimal correlation between the current residual and any dictionary element. It offers some flexibility in the choice of the new element plugged in the model. Even if elements φ_k such that Equation (1) is satisfied may not be uniquely defined, the convergence of the algorithm is guaranteed by our next results. The second shrinkage parameter γ is the standard step-length parameter of Boosting algorithm. It avoids a binary add-on, and actually smoothly inserts the new predictor in the approximation of f. Refinements of WGA including an adaptive choice of ν or γ with the iteration k, or a barycentre average between $G_{k-1}(f)$ and $\langle R_{k-1}(f), \varphi_k \rangle \varphi_k$ may improve algorithm convergence rate. We decide to only consider the simplest version of WGA, because in the noisy framework, these improvements generally disappear from a theoretical point of view (see [2]).

Following the arguments developed in [1], we can extend their results and obtain a polynomial approximation rate:

Theorem 2.1 (Temlyakov, 2000). *Let $B > 0$ and assume that $f \in \mathcal{A}(\mathcal{D}, B)$, where*

$$\mathcal{A}(\mathcal{D}, B) = \left\{ f = \sum_{j=1}^{p} a_j g_j, \quad with \quad \sum_{j=1}^{p} |a_j| \leq B \right\},$$

then for a suitable constant C_B which only depends on B:

$$\|R_k(f)\| \leq C_B (1 + \nu^2 \gamma(2-\gamma)k)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}.$$

2.2. Stability of the Boosting algorithm for noisy regression

This section aims at extending previous results to noisy cases.We present a noisy version of WGA, and we clarify the consistency result of [2] by careful considerations on the empirical residuals instead of theoretical ones (which are indeed unavailable, see Remark 1).

2.2.1. Noisy Boosting algorithm

We consider an unknown $f \in H$, and we observe some i.i.d. variables $(X_i, Y_i)_{i=\{1\dots n\}}$, with arbitrary distributions, and we cast the following regression model on the dictionary $\mathcal{D}$:

$$\forall i = 1 \dots n, \qquad Y_i = f(X_i) + \varepsilon_i, \qquad \text{where} \qquad f = \sum_{j=1}^{p_n} a_j g_j. \tag{3}$$

The Hilbert space $\mathbb{L}_2(P) := \{f, \|f\|^2 = \int f^T(x) f(x) dP(x) < \infty\}$, is endowed with the inner product $\langle f, g \rangle = \int f^T(x) g(x) dP(x)$, where P is the unknown law of the random variables X. We define the empirical WGA, that is analogised to coupled equations (1) and (2), by replacing $\langle , \rangle$ by the empirical inner product $\langle , \rangle_{(n)}$, defined as:

$$\forall (h_1, h_2) \in H, \quad \langle h_1, h_2 \rangle_{(n)} := \frac{1}{n} \sum_{i=1}^{n} h_1(X_i) h_2(X_i) \text{ and } \|h_1\|^2_{(n)} := \frac{1}{n} \sum_{i=1}^{n} h_1(X_i)^2.$$

Remark 1. *The theoretical residual $\hat{R}_k(f) = \boldsymbol{f} - \hat{G}_k(f)$ cannot be used for the WGA (see Equations (4) and (5)) even with the empirical inner product, since f is not observed. Hence, only the observed residuals at step k, $Y - \hat{G}_k$, can be used in the algorithm. This point is not totally clear in the initial work of [2], since notations used in its proofs are read as if $\hat{R}_k(f) = f - \hat{G}_k(f)$ was available. We write explicit and correct proofs in Section Appendix A.2*

Algorithm 2 Noisy Weak Greedy Algorithm

Require: Observations $(X_i, Y_i)_{i=\{1\dots n\}}$, $\gamma \in (0,1]$ (shrinkage parameter), k_{up} (number of iterations).

Initialisation: $\hat{G}_0(f) = 0$.

for $k = 1$ to k_{up} **do**

Step 1: Select $\varphi_k \in \mathcal{D}$ such that:

$$|\langle Y - \hat{G}_{k-1}(f), \varphi_k \rangle_{(n)}| = \max_{1 \le j \le p_n} |\langle Y - \hat{G}_{k-1}(f), g_j \rangle_{(n)}|. \tag{4}$$

Step 2: Compute the current approximation and residual:

$$\hat{G}_k(f) = \hat{G}_{k-1}(f) + \gamma \langle Y - \hat{G}_{k-1}(f), \varphi_k \rangle_{(n)} \varphi_k. \tag{5}$$

end for

2.2.2. Stability of the Boosting algorithm

We will use in the sequel the two following notations: for any sequences $(a_n)_{n\ge 0}$ and $(b_n)_{n\ge 0}$ and a random sequence $(X_n)_{n\ge 0}$, $a_n = \underset{n\to+\infty}{O}(b_n)$ means that a_n/b_n is a bounded sequence, and $X_n = \underset{n\to+\infty}{o_P}(1)$ means that $\forall \varepsilon > 0$, $\lim_{n\to+\infty} \mathbb{P}(|X_n| \ge \varepsilon) = 0$. We recall here needed standard assumptions on high dimensional models.

Hypotheses $\mathbf{H_1}$

$\mathbf{H_{1-1}}$ For any $g_j \in \mathcal{D}$: $\mathbb{E}[g_j(X)^2] = 1$ and $\sup_{1\le j\le p_n, n\in\mathbb{N}} \|g_j(X)\|_\infty < \infty$.

$\mathbf{H_{1-2}}$ The number of predictors p_n satisfies $p_n = \underset{n\to+\infty}{O}\left(\exp(Cn^{1-\xi})\right)$, with $\xi \in (0,1)$ and $C > 0$.

$\mathbf{H_{1-3}}$ $(\varepsilon_i)_{i=1\dots n}$ are i.i.d centred variables in $\mathbb{R}$, independent from $(X_i)_{i=1\dots n}$, satisfying $\mathbb{E}|\varepsilon|^t < \infty$, for some $t > \frac{4}{\xi}$, where ξ is given in H1-2.

$\mathbf{H_{1-4}}$ The sequence $(a_j)_{1\le j\le p_n}$ satisfies: $\sup_{n\in\mathbb{N}} \sum_{j=1}^{p_n} |a_j| < \infty$.

Remark 2. *Assumption* $\mathbf{H_{1-1}}$ *is clearly satisfied for compactly supported real polynomials, or Fourier expansion with trigonometric polynomials. Assumption* $\mathbf{H_{1-2}}$ *bounds the high dimensional setting and states that* $\log(p_n)$ *should be at the most of the same order as* n*. Assumption* $\mathbf{H_{1-3}}$ *is on the nature of the noise, which must be centred with at least a bounded second moment. It is required to apply a uniform law of large numbers and is satisfied for a great number of distributions, such as Gaussian or Laplace ones. Last assumption* $\mathbf{H_{1-4}}$ *is a sparsity hypothesis on the unknown signal. It is trivially satisfied when the decomposition* $(a_j)_{j=1\dots p_n}$ *of* f *is bounded and has a fixed sparsity index: Card*$\{i|a_i \neq 0\} \le S$.

We formulate then the first important result of Boosting algorithm, obtained by [2], which stands for a *stability result*.

Theorem 2.2 (Consistency of WGA). *Consider Algorithm 2 presented above and assume that Hypotheses* $\mathbf{H_1}$ *are fulfilled. Then, there exists a sequence* $k_n := C\log(n)$*, with* $C < \xi/4\log(3)$*, such that:*

$$\mathbb{E}\|f - \hat{G}_{k_n}(f)\|_{(n)}^2 = \underset{n\to+\infty}{o_P}(1).$$

We only give here the outline of proof, details can be found in the Appendix section. A straightforward calculus shows that theoretical residuals would be updated as

$$\hat{R}_k(f) = \hat{R}_{k-1}(f) - \gamma\langle\hat{R}_{k-1}(f), \varphi_k\rangle_{(n)}\varphi_k - \gamma\langle\varepsilon, \varphi_k\rangle_{(n)}\varphi_k. \tag{6}$$

The proof of stability results from the study of a *phantom* algorithm, which reproduces the behaviour of the deterministic version of the algorithm, the inner product $\langle , \rangle$ being replaced by its empirical counterpart, and the (random) sample-driven choice of $(\varphi_k)_{k\geq 0}$ is governed by the random algorithm defined according to Equation (4). We thus consider a semi-population algorithm which works with the deterministic inner product and the random coordinates and element of dictionary selected by the random WGA. The phantom residuals are initialised by $\tilde{R}_0(f) = \hat{R}_0(f) = f$ and satisfy at step k:

$$\tilde{R}_k(f) = \tilde{R}_{k-1}(f) - \gamma\langle\tilde{R}_{k-1}(f), \varphi_k\rangle\varphi_k, \tag{7}$$

where φ_k is chosen using Equation (4). The proof is then broken down in two steps. On the one hand, we would establish an analogue of equation (1) for φ_k which can allow us to apply Theorem 2.1 to the phantom residual $\tilde{R}_k(f)$. On the other hand, we give an upper bound for the difference between $\hat{R}_k(f)$ and $\tilde{R}_k(f)$.

2.3. Stability of support recovery

This paragraph presents our main result in the univariate case. We prove the stability of support recovery as built by the noisy WGA. We prove that the WGA exactly recovers the support of the function with high probability, if we assume an amplitude condition on active coefficients (see hypothesis $\mathbf{H_2}$ below). Interestingly, this result is related to the sparsity of f, *i.e* the number of its non-null coordinates and assumptions on its order of magnitude with respect to ρ. Next assumption also deals with the regression parameter range. A minimal bound for the value of the active coefficients in the decomposition of f is needed to derive a consistency result of the support estimate. If we denote $\mathcal{S}$ the support of f, the next assumption is stated as follows:

Hypothesis $\mathbf{H_2}$: Given ξ defined in $\mathbf{H_{1-2}}$, elements $(a_j)_{1\leq j\leq p_n}$ satisfy:

$$\exists\kappa \in (0,1), \quad \forall j \in \mathcal{S}, \qquad |a_j| \geq n^{-\kappa\xi}.$$

Remark that the greater the number of variables, the larger the value of ξ and the less restrictive Assumption $\mathbf{H_2}$. A constraint on shrinkage parameter γ is also needed to obtain the following result.

Theorem 2.3 (Support recovery). *i) Assume* $\rho(2S-1) < 1$ *and hypothesis* $\mathbf{H_1}$*, there exists a maximal shrinkage parameter* γ^**, such that, for all* $0 < \gamma < \gamma^*$ *in Equation (5), with high probability only active coefficients are selected by Equation (4) along iterations of Algorithm 2.*

ii) Moreover, if hypothesis $\mathbf{H_2}$ *holds with a sufficiently small* $\kappa < \kappa^*$ *(with* κ^* *depending on* S *and* γ*), then Algorithm 2 fully recovers the support of* f *with high probability.*

Point i) of Theorem 2.3 shows that along the iterations of Algorithm2, no false positive elements are introduced in the support of f.

Concerning point ii), related results are known for other algorithms devoted to sparse problems (see for instance [3] for Basis Pursuit algorithms, and [4], [5], or [6] for Orthogonal Matching Pursuit (OMP)). The link between coherence and sparsity for greedy algorithms has already been pointed out by several authors (see for instance [7] or [4] and references therein). It is already known for other signal reconstruction algorithms [8], [9], [6], which also rely on a sparsity assumption. Regarding the condition obtained by [6], our assumption is stronger since active coefficients should be bounded from below by $n^{-\kappa\xi}$ instead of $\log(p)^{1/2}n^{-1/2}$ in Theorem 4 of [6]. Our result may not be optimal but optimal conditions on active coefficients are beyond the scope of this paper. The *weak* aspect of WGA seems harder to handle, compared to the treatment of OMP (for instance) because one has to recursively bound the amplitude of the remaining coefficients on active variables from one iteration to the next according to the size of shrinkage parameters.

Observe that γ^* can be chosen equal to or smaller than 13/18, (see the proof of Theorem 2.3). As for the comprehensive support estimation part, the size of κ^* is made explicit in the proof of Theorem 2.3 (see Section Appendix A.3). It is dictated by the level of noise in the data. When the constraint on κ is not satisfied, it is still possible to show that only correct variables are selected by any WGA. If S becomes large, κ must be chosen close to 0 to obtain a support recovery result. It thus implies a restrictive bound condition on the amplitude of active coefficients. In a similar way, if κ is fixed, then Theorem 2.3 exhibits a permitted maximum size for sparsity S, for which we guarantee the exact recovery of the support with high probability.

In summary, a trade-off between signal sparsity, dimensionality, signal-to-noise ratio and sample size has to be reached. We give explicit constant bounds for results on similar problems. Interesting discussions can be found in [10] (see their Theorems 1 and 2 for sufficient and necessary conditions for an *exhaustive search decoder* to succeed with high probability in recovering a function support) and in the *Sparsity and ultra-high dimensionality* Section of [11].

3. A new $\mathbb{L}_2$-Boosting algorithm for multi-task situations

In this section, our purpose is to extend the algorithm and the results presented above to the multi-task situation. The main focus of this work resides in

the choice of the optimal task to be boosted. Hence, we propose a new algorithm which follows the initial spirit of iterative Boosting (see [12] for further details) and the multi-task structure of f. We first establish an approximation result in the deterministic setting and then we extend stability results of Theorems 2.2 and 2.3 for the so called Boost-Boost algorithm for noisy multi-task regression.

3.1. Multi-task Boost-Boost algorithms

Let us denote $H_m := H^{\otimes m}$ the Hilbert space obtained by m-tensorisation with the inner product:

$$\forall (f, \tilde{f}) \in H_m^2, \qquad \langle f, \tilde{f} \rangle_{H_m} = \sum_{i=1}^{m} \langle f^i, \tilde{f}^i \rangle_H .$$

Given any dictionary $\mathcal{D}$ on H, each element $f \in H_m$ will be described by its m coordinates $f = (f^1, \dots f^m)$, where each f^i is spanned on $\mathcal{D}$, with unknown coefficients:

$$\forall i \in [\![1, m_n]\!], \qquad f^i = \sum_{j=1}^{p_n} a_{i,j} g_j . \tag{8}$$

A canonical extension of WGA to the multi-task problem is described by Algorithm 3.

In the multi-task framework at step k, it is crucial to choose in the residuals the coordinate which is meaningful and thus *most* needs improvement, as well as the best regressor $\varphi_k \in \mathcal{D}$. The main idea is to focus on coordinates that are still poorly approximated. We introduce a new shrinkage parameter $\mu \in (0, 1]$. It allows a tolerance towards the optimal choice of the coordinate to be boosted either relying on the Residual L^2 norm -Equation (9)- or on the $\mathcal{D}$-Correlation sum -Equation (10).

Note that this latter choice is rather different from the choice proposed in [3], which uses the multichannel energy and it sums the correlations of each coordinates of the residuals to any element of the dictionary. Comments on pros and cons of minimising the Residual L^2 norm or the $\mathcal{D}$-Correlation sum viewed as the correlated residual can be found in [13] (page 2316). Although [13] advocates for a final advantage for the $\mathcal{D}$-Correlation sum alternative, we also consider the Residual L^2 norm which seems natural since it relies on the norm of the residuals themselves instead of the sum of information gathered by individual regressors on each residuals. Moreover, conclusions of [13] are more particularly focused on an orthogonal design matrix. The noisy WGA for the multi-task problem is described by Algorithm 4 where we replace the inner product $\langle ., . \rangle$ by the empirical inner product $\langle ., . \rangle_{(n)}$.
We use coupled criteria of Equations (9) and (11) in the Residual L^2 norm Boost-Boost algorithm, while we use criteria of Equations (10) and (11) in its $\mathcal{D}$-Correlation sum counterpart.

Algorithm 3 Boost-Boost algorithm

Require: $f = (f^1, ..., f^m)$, $(\gamma, \mu, \nu) \in (0,1]^3$ (shrinkage parameters), k_{up} (number of iterations).

Initialisation: $G_0(f) = 0_{H_m}$ and $R_0(f) = f$.

for $k = 1$ to k_{up} **do**

Step 1: Select f^{i_k} according to:

$$\|R_{k-1}(f^{i_k})\|^2 \geq \mu \max_{1\leq i\leq m} \|R_{k-1}(f^i)\|^2, \qquad \text{[Residual } L^2 \text{ norm]} \tag{9}$$

or to

$$\sum_{j=1}^{p} \langle R_{k-1}(f^{i_k}), g_j\rangle^2 \geq \mu \max_{1\leq i\leq m} \sum_{j=1}^{p} \langle R_{k-1}(f^i), g_j\rangle^2, \text{[}\mathcal{D}\text{-Correlation sum]} \tag{10}$$

Step 2: Select $\varphi_k \in \mathcal{D}$ such that:

$$|\langle R_{k-1}(f^{i_k}), \varphi_k\rangle| \geq \nu \max_{1\leq j\leq p} |\langle R_{k-1}(f^{i_k}), g_j\rangle|, \tag{11}$$

Step 3: Compute the current approximation:

$$\begin{aligned} &G_k(f^i) = G_{k-1}(f^i), \quad \forall i \neq i_k, \\ &G_k(f^{i_k}) = G_{k-1}(f^{i_k}) + \gamma\langle R_{k-1}(f^{i_k}), \varphi_k\rangle\varphi_k. \end{aligned} \tag{12}$$

Step 4: Compute the current residual: $R_k(f) = f - G_k(f)$.

end for

3.2. Approximation Results in the deterministic setting

We consider the sequence of functions $(R_k(f))_k$ recursively built according to our Boost-Boost Algorithm 3 either with the choice (9) or (10). Since $\overline{\text{Span}\,\mathcal{D}} = H$, for any $f \in H_m$, each f^i can be decomposed in H, and we denote S^i, the minimal amount of sparsity for such a representation. We then prove a first approximation result.

Theorem 3.1 (Convergence of the Boost-Boost Algorithm). *Let $f = (f^1, \dots f^m) \in H_m$ such that, for any coordinate i, $f^i \in \mathcal{A}(\mathcal{D}, B)$.*

i) There exists a suitable constant C_B which only depends on B: the approximations provided by the Residual L^2 norm Boost-Boost algorithm satisfy, for all $k \geq m$

$$\forall i \in [\![1, m]\!], \ \|R_k(f^i)\| \leq C_B \mu^{-\frac{1}{2}} \nu^{-\frac{\nu(2-\gamma)}{2+\nu(2-\gamma)}} (\gamma(2-\gamma))^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}} \left(\frac{k}{m}\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}.$$

ii) Assume $\rho S(1+\nu^{-1}) < 1+\rho$, there exists a suitable constant $C_{\rho,S,B}$ such that the approximations provided by the $\mathcal{D}$-Correlation sum Boost-Boost

Algorithm 4 Noisy Boost-Boost algorithm

Require: Observations $(X_i, Y_i)_{i=1,\dots,n}$, $\gamma \in (0,1]$ (shrinkage parameter), k_{up} (number of iterations).

Initialisation: $\hat{G}_0(f) = 0_{H_m}$.

for $k = 1$ to k_{up} **do**

Step 1: Select i_k according to:

$$\|Y^{i_k} - \hat{G}_{k-1}(f^{i_k})\|^2_{(n)} = \max_{1\leq i\leq m} \|Y^i - \hat{G}_{k-1}(f^i)\|^2_{(n)}, [\text{Residual } L^2 \text{ norm}]$$

or to

$$\sum_{j=1}^{p} \langle Y^{i_k} - \hat{G}_{k-1}(f^{i_k}), g_j\rangle^2_{(n)} = \max_{1\leq i\leq m} \sum_{j=1}^{p} \langle Y^i - \hat{G}_{k-1}(f^i), g_j\rangle^2_{(n)}, [\mathcal{D}\text{-Correlation sum}]$$

Step 2: Select $\varphi_k \in \mathcal{D}$ such that:

$$|\langle Y^{i_k} - \hat{G}_{k-1}(f^{i_k}), \varphi_k\rangle_{(n)}| = \max_{1\leq j\leq p} |\langle Y^i - \hat{G}_{k-1}(f^i), g_j\rangle_{(n)}|,$$

Step 3: Compute the current approximation:

$$\begin{aligned} &\hat{G}_k(f^i) = \hat{G}_{k-1}(f^i), \quad \forall i \neq i_k, \\ &\hat{G}_k(f^{i_k}) = \hat{G}_{k-1}(f^{i_k}) + \gamma \langle Y^{i_k} - \hat{G}_{k-1}(f^{i_k}), \varphi_k\rangle_{(n)} \varphi_k. \end{aligned}$$

end for

algorithm satisfy, for all $k \geq m$

$$\forall i \in [\![1, m]\!], \ \ \|R_k(f^i)\| \leq C_{\rho,S,B} \mu^{-\frac{1}{2}} \nu^{-\frac{\nu(2-\gamma)}{2+\nu(2-\gamma)}} (\gamma(2-\gamma))^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}} \left(\frac{k}{m}\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}.$$

Remark 3. *Remark that this theorem recovers the classical condition in noisy setting,* $\rho(2S-1) < 1$ *when* $\nu = 1$. *We can discuss on the added value brought by the Residual* L^2 *norm Boost-Boost algorithm. Comparing to* m *naive runs of standard WGA on each coordinates of the residuals, the proposed algorithm is efficient when the coordinates of the residuals are unbalanced, i.e. when few columns possess most of the information to be predicted. In the opposite, when WGA is applied to well balanced tasks, there is no clear advantage to use Residual* L^2 *norm Boost-Boost algorithm.*

3.3. Stability of the Boost-Boost algorithms for noisy multi-task regression

We establish a theoretical convergence result for these two versions of multi-task WGA. We first state few assumptions adapted to the multi-task setting.

Hypotheses $\mathbf{H_1^{Mult}}$

$\mathbf{H_{1-1}^{Mult}}$ For any $g_j \in \mathcal{D}$: $\mathbb{E}[g_j(X)^2] = 1$ and $\sup_{1\leq j\leq p_n, n\in\mathbb{N}} \|g_j(X)\|_\infty < \infty$.

$\mathbf{H_{1-2}^{Mult}}$ There exist $\xi \in (0,1), C > 0$ such that the number of predictors and tasks (p_n, m_n) satisfy

$$p_n \vee m_n = \underset{n\to+\infty}{O}\left(\exp(Cn^{1-\xi})\right).$$

$\mathbf{H_{1-3}^{Mult}}$ $(\varepsilon_i)_{i=1\dots n}$ are i.i.d centred in $\mathbb{R}^{m_n}$, independent of $(X_i)_{i=1\dots n}$ such that for some $t > \frac{4}{\xi}$, where ξ is defined in $\mathbf{H_{1-2}^{Mult}}$, $\sup_{1\leq j\leq m_n, n\in\mathbb{N}} \mathbb{E}|\varepsilon^j|^t < \infty$.

Moreover, the variance of ε^j does not depend on j: $\forall (j,\tilde{j}) \in [\![1, m_n]\!]^2$, $\mathbb{E}|\varepsilon^j|^2 = \mathbb{E}|\varepsilon^{\tilde{j}}|^2$.

$\mathbf{H_{1-4}^{Mult}}$ The sequence $(a_{i,j})_{1\leq j\leq p_n, 1\leq i\leq m_n}$ satisfies: $\sup_{n\in\mathbb{N}, 1\leq i\leq m_n} \sum_{j=1}^{p_n} |a_{i,j}| < \infty$.

Remark that a critical change appears in Hypothesis $\mathbf{H_{1-3}^{Mult}}$. Indeed, each tasks should be of equal variances. We thus need to normalise the data before applying the Boost-Boost algorithms.

Hence, we can derive a result on the consistency of the Residual L^2 norm Boost-Boost algorithm. This extend the result of Theorem 2.2 for univariate WGA.

Theorem 3.2 (Consistency of the Boost-Boost Residual L^2 norm). *Assume that Hypotheses $\mathbf{H_1^{Mult}}$ are fulfilled and that $\rho(2S-1) < 1$, then there exists a sequence $k_n := C\log(n)$, with $C < \xi/4\log(3)$, such that:*

$$\forall i \in [\![1, m_n]\!], \quad \mathbb{E}\|f^i - \hat{G}_{k_n}(f^i)\|_{(n)}^2 = \underset{n\to+\infty}{o_P}(1).$$

As regards the Boost-Boost algorithm defined with the sum of correlations, if the number of predictors p_n satisfies a more restrictive assumption than $\mathbf{H_{1-2}^{Mult}}$, we prove a similar result.

Theorem 3.3 (Consistency of the Boost-Boost $\mathcal{D}$-Correlation sum). *Assume that Hypotheses $\mathbf{H_1^{Mult}}$ are fulfilled, with $p_n = \underset{n\to+\infty}{O}(n^{\xi/4})$, and suppose that $\rho(2S-1) < 1$, then there exists a sequence $k_n := C\log(n)$ with $C < \xi/8\log(3)$ such that:*

$$\forall i \in [\![1, m_n]\!], \quad \mathbb{E}\|f^i - \hat{G}_{k_n}(f^i)\|_{(n)}^2 = \underset{n\to+\infty}{o_P}(1).$$

Remark that Assumption $\mathbf{H_{1-2}^{Mult}}$ includes the very high dimensional case, whilst Theorem 3.3 with one more restrictive assumption, is theoretically limited to the high dimensional case.

We can also obtain a consistency result for the support of the Boost-Boost algorithms.

Hypothesis $\mathbf{H_2^{Mult}}$: Given ξ defined in $\mathbf{H_{1-2}^{Mult}}$, elements $(a_{i,j})_{1\leq i\leq m_n, 1\leq j\leq p_n}$ satisfy:

$$\exists \kappa \in (0,1), \forall i \in [\![1, m_n]\!] \quad \forall j \in \mathcal{S}^i, \qquad |a_{i,j}| \geq n^{-\kappa\xi}.$$

Theorem 3.4 (Support recovery). *Assume* $\rho(2S-1) < 1$ *and Assumptions* $\mathbf{H_1^{Mult}}$ *are fulfilled, then the two propositions hold.*

i) There exists a maximal shrinkage parameter γ^**, such that, for all* $0 < \gamma < \gamma^*$ *in Equation (12), with high probability only active coefficients are selected along iterations of Algorithm 4.*

ii) Moreover, if Assumption $\mathbf{H_2^{Mult}}$ *holds with a sufficiently small* $\kappa < \kappa^*$ *(with* κ^* *depending on* S *and* γ*), then both Boost-Boost procedures fully recover the support of* f *with high probability.*

4. Application to simulated data

This section is dedicated to simulation studies to assess practical performances of our method in light of expected theoretical results. The toy set we use is a class of challenging univariate, or multi-task, noisy linear data sets with different characteristics. They are simulated according to a linear modelling $Y = XA + \varepsilon$, where Y is a $n \times m$ response matrix, X is a $n \times p$ observation matrix, ε is additional Gaussian noise and A is the parameter that encodes relationships to be inferred. We used $n = 100$ samples and all data sets are replicated 10 times each. Covariates are generated according to a multi-task Gaussian distribution with covariance matrix $\forall i, X_i \sim \mathcal{N}(0, 10I_p)$. Errors are generated according to a multi-task normal distribution with an identity covariance matrix (except the last data set) and non-zero A-coefficients are drawn according to a $\mathcal{N}(0,1)$ distribution.

We change the value of parameters p (number of predictor variables), m (number of responses), s (sparsity of each row of A). To assess the performance of our approach, we use on the one hand error of prediction, which is definedn as $\|Y - XA\|^2_{(n)}$. On the other hand, we compute the number of false positive and false negative parameters, *i.e* inferred by mistake and missed coefficients. The maximal number of iterations of Boosting algorithm was set to 20 with a shrinkage factor γ equal to 0.2.

Appendix A. Stability results for Boosting algorithms

Appendix A.1. Concentration inequalities

We begin by reminding here some technical results. Lemma Appendix A.1, given in [2], provides a uniform law of large numbers, in order to compare inner products $\langle , \rangle_{(n)}$ and $\langle , \rangle$. It is useful for the proofs of theorems of Section 2.2.2 and 2.3, but does not call typical boosting arguments.

Lemma Appendix A.1. *Assume that Hypotheses* $\mathbf{H_1}$ *are fulfilled on dictionary* $\mathcal{D}$*,* f *and* ε*, with* $0 < \xi < 1$ *as given in* $\mathbf{H_{1-2}}$*, then:*

i) $\sup_{1 \le i,j \le p_n} |\langle g_i, g_j \rangle_{(n)} - \langle g_i, g_j \rangle| = \zeta_{n,1} = \mathcal{O}_P(n^{-\xi/2})$,

ii) $\sup_{1 \le i \le p_n} |\langle g_i, \varepsilon \rangle_{(n)}| = \zeta_{n,2} = \mathcal{O}_P(n^{-\xi/2})$,

iii) $\sup_{1 \le i \le p_n} |\langle f, g_i \rangle_{(n)} - \langle f, g_i \rangle| = \zeta_{n,3} = \mathcal{O}_P(n^{-\xi/2})$.

Denote $\zeta_n = \max\{\zeta_{n,1}, \zeta_{n,2}, \zeta_{n,3}, \zeta_{n,4}\} = \mathcal{O}_P(n^{-\xi/2})$. The following lemma (lemma 2 from [2]) also holds.

Lemma Appendix A.2. *Under Hypotheses* $\mathbf{H_1}$*, there exists a constant* $0 < C < +\infty$*, independent of* n *and* k*, such that on set* $\Omega_n = \{\omega, |\zeta_n(\omega)| < 1/2\}$*:*

$$\sup_{1\leq j\leq p_n} |\langle \hat{R}_k(f), g_j\rangle_{(n)} - \langle \tilde{R}_k(f), g_j\rangle| \leq C\left(\frac{5}{2}\right)^k \zeta_n.$$

Proof This lemma is given in [2], but their notations are confusing, since residuals $\hat{R}_k$ are used to compute φ_k instead of $Y - \hat{G}_k$ (see Remark 1 at the end of Section 2.2). Fortunately, we can generalise its application field using Lemma Appendix A.1. First, assume that $k = 0$. The desired inequality follows directly from point iii) of Lemma Appendix A.1. We now extend the proof by an inductive argument.

Denote $A_n(k,j) = \langle \hat{R}_k(f), g_j\rangle_{(n)} - \langle \tilde{R}_k(f), g_j\rangle$. Then, from the recursive relations of Equations (6) and (7), we obtain:

$$\begin{aligned}
A_n(k,j) &= \langle \hat{R}_{k-1}(f) - \gamma\langle \hat{R}_{k-1}(f), \varphi_k\rangle_{(n)}\varphi_k - \gamma\langle \varepsilon, \varphi_k\rangle_{(n)}\varphi_k, g_j\rangle_{(n)} \\
&\quad -\langle \tilde{R}_{k-1}(f) - \gamma\langle \tilde{R}_{k-1}(f), \varphi_k\rangle\varphi_k, g_j\rangle \\
&= A_n(k-1,j) - \gamma \underbrace{\langle \tilde{R}_{k-1}(f), \varphi_k\rangle(\langle \varphi_k, g_j\rangle_{(n)} - \langle \varphi_k, g_j\rangle)}_{=(I)} \\
&\quad -\gamma \underbrace{\langle \varphi_k, g_j\rangle_{(n)}(\langle \hat{R}_{k-1}(f), \varphi_k\rangle_{(n)} - \langle \tilde{R}_{k-1}(f), \varphi_k\rangle)}_{=(II)} \\
&\quad -\gamma \underbrace{\langle \varepsilon, \varphi_k\rangle_{(n)}\langle \varphi_k, g_j\rangle_{(n)}}_{=(III)}.
\end{aligned}$$

Expanding Equation (7) yields $\|\tilde{R}_k(f)\|^2 = \|\tilde{R}_{k-1}(f)\|^2 - \gamma(2-\gamma)\langle \tilde{R}_{k-1}(f), \varphi_k\rangle^2$. From the last equality, we deduce $\|\tilde{R}_k(f)\|^2 \leq \|\tilde{R}_{k-1}(f)\|^2 \leq \ldots \leq \|f\|^2$ and Lemma Appendix A.1 i) shows that

$$\sup_{1\leq j\leq p_n} |(I)| \leq \|\tilde{R}_{k-1}(f)\|\|\varphi_k\|\zeta_n \leq \|f\|\zeta_n.$$

Moreover,

$$\begin{aligned}
\sup_{1\leq j\leq p_n} |(II)| &\leq \sup_{1\leq j\leq p_n} |\langle \varphi_k, g_j\rangle_{(n)}| \sup_{1\leq j\leq p_n} |A_n(k-1,j)| \\
&\leq (\sup_{1\leq j\leq p_n} |\langle \varphi_k, g_j\rangle| + \zeta_n) \sup_{1\leq j\leq p_n} |A_n(k-1,j)| \\
&\leq (1+\zeta_n) \sup_{1\leq j\leq p_n} |A_n(k-1,j)|.
\end{aligned}$$

Finally, using i) and ii) from Lemma Appendix A.1:

$$\sup_{1\leq j\leq p_n} |(III)| \leq \sup_{1\leq j\leq p_n} |\langle \varphi_k, g_j\rangle_{(n)}| \sup_{1\leq j\leq p_n} |\langle \varepsilon^{i_k}, g_j\rangle_{(n)}| \leq (1+\zeta_n)\zeta_n.$$

Using our bounds on (I), (II) and (III), and $\gamma < 1$, we obtain on Ω_n

$$\begin{aligned}\sup_{1\leq j\leq p_n}|A_n(k,j)| &\leq \sup_{1\leq j\leq p_n}|A_n(k-1,j)| + \zeta_n\|f\| + (1+\zeta_n)\sup_{1\leq j\leq p_n}|A_n(k-1,j)| \\ &\quad +(1+\zeta_n)\zeta_n \\ &\leq \frac{5}{2}\sup_{1\leq j\leq p_n}|A_n(k-1,j)| + \zeta_n\left(\|f\|+\frac{3}{2}\right).\end{aligned}$$

A simple induction yields:

$$\begin{aligned}\sup_{1\leq j\leq p_n}|A_n(k,j)| &\leq \left(\frac{5}{2}\right)^k \underbrace{\sup_{1\leq j\leq p_n}|A_n(0,j)|}_{\leq\zeta_n} + \zeta_n\left(\|f\|+\frac{3}{2}\right)\sum_{\ell=0}^{k-1}\left(\frac{5}{2}\right)^\ell \\ &\leq \left(\frac{5}{2}\right)^k \zeta_n\left(1+\left(\sup_{n\in\mathbb{N}}\sum_{j=1}^{p_n}|a_j|+\frac{3}{2}\right)\sum_{\ell=1}^{\infty}\left(\frac{5}{2}\right)^{-\ell}\right),\end{aligned}$$

which ends the proof of i) by setting $C = 1+\left(\sup_{n\in\mathbb{N}}\sum_{j=1}^{p_n}|a_j|+\frac{3}{2}\right)\sum_{\ell=1}^{\infty}\left(\frac{5}{2}\right)^{-\ell}$.
□

Appendix A.2. Proof of consistency result

We aim then to apply Theorem 2.1 to the semi-population $\tilde{R}_k(f)$ version of $\hat{R}_k(f)$. This will be possible with high probability when $n \to +\infty$. We first observe that Lemma Appendix A.2 hold changing the theoretical residual $\hat{R}_k(f)$ by the observed residual $Y - \hat{G}_k(f)$ thanks to Lemma Appendix A.1 $ii)$. Hence, on set Ω_n, by definition of φ_k:

$$\begin{aligned}|\langle Y-\hat{G}_{k-1}(f),\varphi_k\rangle_{(n)}| &= \sup_{1\leq j\leq p_n}|\langle Y-\hat{G}_{k-1}(f),g_j\rangle_{(n)}| \\ &= \sup_{1\leq j\leq p_n}\left\{|\langle\tilde{R}_{k-1}(f),g_j\rangle| - C\left(\frac{5}{2}\right)^{k-1}\zeta_n\right\}. \quad \text{(A.1)}\end{aligned}$$

Applying Lemma Appendix A.2 again on set Ω_n, we have:

$$\begin{aligned}|\langle\tilde{R}_{k-1}(f),\varphi_k\rangle| &\geq |\langle Y-\hat{G}_{k-1}(f),\varphi_k\rangle_{(n)}| - C\left(\frac{5}{2}\right)^{k-1}\zeta_n \\ &\geq \sup_{1\leq j\leq p_n}|\langle\tilde{R}_{k-1}(f),g_j\rangle| - 2C\left(\frac{5}{2}\right)^{k-1}\zeta_n. \quad \text{(A.2)}\end{aligned}$$

Let $\tilde{\Omega}_n = \left\{\omega, \quad \forall k\leq k_n, \quad \sup_{1\leq j\leq p_n}|\langle\tilde{R}_{k-1}(f),g_j\rangle| > 4C\left(\frac{5}{2}\right)^{k-1}\zeta_n\right\}$. We deduce from Equation (A.2) the following inequality:

$$|\langle\tilde{R}_{k-1}(f),\varphi_k\rangle| \geq \frac{1}{2}\sup_{1\leq j\leq p_n}|\langle\tilde{R}_{k-1}(f),g_j\rangle|. \quad \text{(A.3)}$$

Consequently, on set $\Omega_n \cap \tilde{\Omega}_n$, we can apply Theorem 2.1 to family $(\tilde{R}_k(f^i))_k$, since it satisfies a WGA with constants $\tilde{\nu} = 1/2$.

$$\|\tilde{R}_k(f)\| \leq C_B \left(1 + \frac{1}{4}\gamma(2-\gamma)k\right)^{-\frac{2-\gamma}{2(6-\gamma)}}. \tag{A.4}$$

Consider now the set $\tilde{\Omega}_n^C = \left\{\omega, \quad \exists\, k \leq k_n \quad \sup_{1\leq j\leq p_n} |\langle \tilde{R}_{k-1}(f), g_j\rangle| \leq 4C\left(\frac{5}{2}\right)^{k-1}\zeta_n\right\}$.
If we denote $w_k = w_{k-1} + \gamma|\langle \tilde{R}_{k-1}(f), \varphi_k\rangle|$, with the initialisation $w_0 = 1$, following the proof of Theorem 2.1, we have:

$$\|\tilde{R}_k(f)\|^2 \leq \gamma^{-1} w_k \sup_{1\leq j\leq p_n} |\langle \tilde{R}_k(f), g_j\rangle|. \tag{A.5}$$

Moreover, a straightforward recursion combined with Cauchy-Schwarz's inequality and the fact that $\|\tilde{R}_k(f)\|$ is non-increasing show that

$$w_k \leq 1 + \gamma \sum_{\ell=1}^{k} \|\tilde{R}_{\ell-1}(f)\| \leq 1 + \gamma k \|f\|. \tag{A.6}$$

From Equations (A.5) and (A.6), we deduce that,

$$\|\tilde{R}_k(f)\|^2 \leq \gamma^{-1} 4C \left(\frac{5}{2}\right)^k \zeta_n (1 + \gamma k \|f\|). \tag{A.7}$$

Hence, on $(\Omega_n \cap \tilde{\Omega}_n) \cup \tilde{\Omega}_n^C$, by Equation (A.4) and (A.7),

$$\|\tilde{R}_k(f)\|^2 \leq C_B^2 \left(1 + \frac{1}{4}\gamma(2-\gamma)k\right)^{-\frac{2-\gamma}{6-\gamma}} + 4C\left(\frac{5}{2}\right)^k \zeta_n \gamma^{-1}(1 + \gamma k\|f\|). \tag{A.8}$$

To conclude, remark that $\mathbb{P}\left((\Omega_n \cap \tilde{\Omega}_n) \cup \tilde{\Omega}_n^C\right) \geq \mathbb{P}(\Omega_n) \underset{n\to+\infty}{\longrightarrow} 1$. Inequality (A.8) holds almost surely for all ω and for a sequence $k_n < (\xi/4\log(3))\log(n)$, which grows sufficiently slowly:

$$\|\tilde{R}_{k_n}(f)\| = o_P(1). \tag{A.9}$$

To finish the proof, let $k \geq 1$ and consider $A_k = \|\hat{R}_k(f) - \tilde{R}_k(f)\|$. By definition:

$$\begin{aligned} A_k &= \|\hat{R}_{k-1}(f) - \gamma\langle \hat{R}_{k-1}(f), \varphi_k\rangle_{(n)}\varphi_k - \gamma\langle \varepsilon, \varphi_k\rangle_{(n)}\varphi_k \\ &\quad - \left(\tilde{R}_{k-1}(f) - \gamma\langle \tilde{R}_{k-1}(f), \varphi_k\rangle\varphi_k\right)\| \\ &\leq A_{k-1} + \gamma|\langle Y - \hat{G}_{k-1}(f), \varphi_k\rangle_{(n)} - \langle \tilde{R}_{k-1}(f), \varphi_k\rangle|. \end{aligned} \tag{A.10}$$

Under Hypotheses $\mathbf{H_1}$, we deduce from Equation (A.10) the following inequality on Ω_n:

$$A_k \leq A_{k-1} + \gamma\left(C\left(\frac{5}{2}\right)^{k-1} + 1\right)\zeta_n. \tag{A.11}$$

Using $A_0 = 0$, we deduce recursively from Equation (A.11) that, on Ω_n, since $k := k_n$ grows sufficiently slowly:

$$A_{k_n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0. \tag{A.12}$$

Finally observe that $\|\hat{R}_{k_n}(f)\| \leq \|\tilde{R}_{k_n}(f)\| + A_{k_n}$. The conclusion holds using Equation (A.9) and (A.12).

Appendix A.3. Proof of support recovery

We now detail the proof of Theorem 2.3 which stands for the exact recovery of the support with high probability. Remind that we denote S ($\mathcal{S}$) the sparsity (the support) of f. We suppose that the current residuals could be decomposed on $\mathcal{D}$ as $\hat{R}_k(f) = \sum_{j=1}^{p_n} \theta_j^k g_j$, where $(\theta_j^k)_j$ is S_k-sparse, with support $\mathcal{S}_k$.

Proof of i): The aim of the first part of the proof is to show that along the iterations of Boosting, we only select elements of the support of f using Equation (4). Since $\mathcal{S}_0 = \mathcal{S}$, we only have to show that $(S_k)_{k\geq 0}$ is non-increasing, which implies that successive residual supports satisfy $\mathcal{S}_k \subset \mathcal{S}_{k-1}$. At the initial step $k = 0$, $S_0 = S$ and $\mathcal{S}_0 = \mathcal{S}$. The proof works now by induction, and we assume that $\mathcal{S}_{k-1} \subset \mathcal{S}$. Using the same outline of proof of Lemma Appendix A.2, we have:

$$\forall g_j \in \mathcal{D}, \qquad |\langle Y - \hat{G}_{k-1}(f), g_j\rangle_{(n)} - \langle \hat{R}_{k-1}(f), g_j\rangle| \leq C\zeta_n \left(\frac{5}{2}\right)^{k-1}. \tag{A.13}$$

Using Equation (A.13) and the decomposition of $\hat{R}_{k-1}(f)$ on $\mathcal{D}$, we can write:

$$\begin{aligned} \forall j \notin \mathcal{S}_{k-1}, \quad |\langle Y - \hat{G}_{k-1}(f), g_j\rangle_{(n)}| &\leq \rho\|\theta^{k-1}\|_1 + C\zeta_n \left(\frac{5}{2}\right)^{k-1} \\ &\leq \rho S_{k-1}\|\theta^{k-1}\|_\infty + C\zeta_n \left(\frac{5}{2}\right)^{k-1} \end{aligned} \tag{A.14}$$

Moreover, for $j \in \mathcal{S}_{k-1}$, such that $|\theta_j^{k-1}| = \|\theta^{k-1}\|_\infty$, we also have:

$$\begin{aligned} |\langle Y - \hat{G}_{k-1}(f), g_j\rangle_{(n)}| &\geq \|\theta^{k-1}\|_\infty - \rho(\|\theta^{k-1}\|_1 - \|\theta^{k-1}\|_\infty) - C\zeta_n \left(\frac{5}{2}\right)^{k-1} \\ &\geq (1 - \rho(S_{k-1} - 1))\|\theta^{k-1}\|_\infty - C\zeta_n \left(\frac{5}{2}\right)^{k-1}. \end{aligned} \tag{A.15}$$

Remind that element j is selected at step k following Equation (4). Hence, we deduce from Equations (A.14) and (A.15) that $j \in \mathcal{S}_k$ is in $\mathcal{S}_{k-1}$ if the following inequality is satisfied:

$$(1 - \rho(S_{k-1} - 1))\|\theta^{k-1}\|_\infty - C\zeta_n \left(\frac{5}{2}\right)^{k-1} \geq \rho S_{k-1}\|\theta^{k-1}\|_\infty + C\zeta_n \left(\frac{5}{2}\right)^{k-1},$$

which can be rewritten as:

$$(1-\rho(2S_{k-1}-1))\|\theta^{k-1}\|_\infty \geq 2C\zeta_n \left(\frac{5}{2}\right)^{k-1}. \tag{A.16}$$

Condition (A.16) deserves a special attention since θ^{k-1} is the decomposition of $\hat{R}_{k-1}(f)$ on dictionary $\mathcal{D}$ at step $k-1$. Thus $\|\theta^{k-1}\|_\infty \underset{k\to+\infty}{\longrightarrow} 0$. Hence, (A.16) is only valid for a limited number of iterations. Since $\zeta_n = \mathcal{O}_P(n^{-\xi/2})$, the condition of Equation (A.16) would result from the induction hypothesis (which implies that $\rho(2S_{k-1}-1) < 1$), if we have a sufficiently sharp control of $\frac{2C\zeta_n}{\|\theta^{k-1}\|_\infty}\left(\frac{5}{2}\right)^{k-1}$ all along the iterations of the Boosting algorithm, which are allowed to grow with n as $k_n := A\xi\log(n)$ with $A = 1/4\log(3)$ (see Theorem 2.2). This is possible, if we keep shrinkage parameter γ small enough.

More precisely, by using the definition of the update rule of Algorithm (7) and *ii*) of Lemma Appendix A.1, we consider at step k the index j such that $g_j = \varphi_k$:

$$|\theta_j^k - (\theta_j^{k-1} - \gamma\theta_j^{k-1})| \leq \gamma\zeta_n. \tag{A.17}$$

Let us define $u_k := \|\theta^k\|_\infty$, using Equation (A.17), we derive $u_{k+1} \geq u_k(1-\gamma) - \gamma\zeta_n$, and a comparison to an arithmetico-geometric sequence yields $u_k \geq (1-\gamma)^k(u_0+\zeta_n) - \zeta_n$. Hence, for all k, we obtain that with high probability, the next inequality holds

$$\frac{2C\zeta_n}{\|\theta^{k-1}\|_\infty}\left(\frac{5}{2}\right)^{k-1} \leq \frac{2Cn^{-\xi/2}}{(1-\gamma)^{k-1}(u_0+\zeta_n)-\zeta_n}\left(\frac{5}{2}\right)^{k-1}.$$

We now check that the right hand side of this inequality remains lower than $1-\rho(2S_{k-1}-1)$. The least favourable case (the largest attainable value of the right hand side) is achieved by the maximal number of iterations possible k_n. Moreover, $(1-\gamma)^{k_n}u_0$ is not allowed to be smaller than $n^{-\xi/2}\left(\frac{5}{2}\right)^{k_n-1}$. This last point is true provided that the shrinkage parameter γ is not too large. The maximal shrinkage parameter γ^* is found (by taking the arg max) while assuming that $\left(\frac{5}{2}\right)^{k_n-1} n^{-\xi/2} \leq (1-\gamma)^{k_n}$,. This allows us to conclude that (A.16) holds as soon as $\rho(2S_{k-1}-1) < 1$. This immediately implies that $\mathcal{S}_k \subset \mathcal{S}_{k-1}$ and $S_k \leq S_{k-1}$. Since k_n is equal to $A\xi\log(n)$ (see Theorem 2.2), the maximal shrinkage parameter is equal to $\gamma^* = 13/18$. □

Proof of ii): The second part of the proof consists in checking that, along the iterations of the Boosting algorithm, every correct element of the dictionary is chosen at least once.

It is sufficient to consider $j \in \mathcal{S}$ for the end of the proof, since we do not select an incorrect element, with high probability. Suppose that j is selected at step k, from inequality (A.17), we have $|\theta_j^k| \leq \gamma\zeta_n + (1-\gamma)|\theta_j^{k-1}|$. Hence, an other arithmetic-geometric comparison argument yields

$$|\theta_j^k| \leq (1-\gamma)^{s_j(k)}(|\theta_j^{k_0}| - \zeta_n) + \zeta_n, \tag{A.18}$$

where $s_j(k)$ denotes the number of times j is selected within the first k iterations of the Boosting and k_0 is the first step where j is selected.

We now end the proof by assuming that one element of $\mathcal{S}$ is never selected and we exhibit a contradiction. Thus, there exists $j_0 \in \mathcal{S}$, such that along the k_n iterations j_0 is never selected, which implies by Assumption $\mathbf{H_2}$ that: $|\theta^k_{j_0}| = |\theta^0_{j_0}| \geq n^{-\kappa\xi}$. Of course, there exists one element of $j_1 \in \mathcal{S}$, which is selected at least $\lfloor k_n/S \rfloor$ times ($\lfloor x \rfloor$ is the floor function evaluated in x, it is equal to largest integer not greater than x) and for the corresponding iteration k_1, we have

$$|\theta^{k_1}_{j_1}| \leq (1-\gamma)^{\lfloor k_n/S \rfloor}(\|\theta^0\|_\infty - \zeta_n) + \zeta_n.$$

Since j_1 is selected at step k_1, we have

$$\begin{aligned} |\theta^{k_1}_{j_1}| &\geq \gamma\left((1-\rho S)\|\theta^{k_1}\|_\infty - \zeta_n\left(\frac{5}{2}\right)^{k_1}\right) \geq \gamma\left((1-\rho S)|\theta^{k_1}_{j_0}| - \zeta_n\left(\frac{5}{2}\right)^{k_1}\right) \\ &\geq \gamma\left((1-\rho S)n^{-\kappa\xi} - \zeta_n\left(\frac{5}{2}\right)^{k_1}\right) \geq \gamma\left((1-\rho S)n^{-\kappa\xi} - \zeta_n\left(\frac{5}{2}\right)^{k_n}\right). \end{aligned}$$

We obtain the sought contradiction as soon as $n^{-\kappa\xi} \geq \left(\frac{5}{2}\right)^{A\log(n)\xi} n^{-\xi/2}$ and $(1-\gamma)^{\lfloor A\log(n)\xi/S \rfloor} \leq n^{-\kappa\xi}$, that is to say if $\kappa \leq \frac{1}{2} - A\log(\frac{5}{2}) \wedge \kappa^*(S)$ ($x \wedge y := min(x,y)$, with $\kappa^*(S) = A/S\log(1/(1-\gamma))$. For instance, the Boosting algorithm run with $A = 1/4\log(3)$ recovers the support with high probability if $\kappa^* \simeq 0.29/S$. This ends the proof of the support consistency. □

Appendix B. Proof of results for multi-task $\mathbb{L}_2$-Boosting algorithms

Appendix B.1. Proof of Theorem 3.1

We here break down in several steps the proof of Theorem 3.1. Remind that $\mathcal{D} = \{(g_j), 1 \leq j \leq p\}$ is a dictionary, with coherence ρ, which spans H. We set any $f = (f^1, \ldots, f^m) \in H_m$ such that $f^i \in \mathcal{A}(\mathcal{D}, B)$.

The first key remark is that, if we denote $s_i(k)$ the number of steps in which i is invoked until step k, for all $i \in [\![1, m]\!]$, we deduce from Theorem 2.1 that:

$$\forall k \geq 1, \quad \|R_{k-1}(f^i)\| \leq C_B(1 + \nu^2\gamma(2-\gamma)s_i(k-1))^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}. \tag{B.1}$$

The second key point of the proof consists in comparing $R_k(f^i)$ and $R_k(f^{i_k})$, where i_k is chosen using Equation (9) or (10). For the Boost-Boost Residual L^2 norm algorithm, this step is not pivotal since, using Equation (9),

$$\forall i \in [\![1, m]\!], \quad \|R_k(f^i)\| \leq \mu^{-1}\|R_k(f^{i_k})\|. \tag{B.2}$$

However, for the Boost-Boost $\mathcal{D}$-Correlation sum algorithm, we can prove the following lemma:

Lemma Appendix B.1. *Suppose that $\rho S(1+\nu^{-1}) < 1+\rho$, then one has for any k:*

$$\forall i \neq i_k, \quad \|R_{k-1}(f^i)\|^2 \leq \mu^{-1}\|R_{k-1}(f^{i_k})\|^2 \left(\frac{1+\rho(S-1)}{1-\rho(S-1)}\right)^3.$$

Proof Assume that each residual $R_k(f^i)$ is expanded on $\mathcal{D}$ at step k as: $R_k(f^i) = \sum_{j=1}^{p} \theta_{i,j}^k g_j$, where $(\theta_{i,j}^k)_{1\leq j\leq p}$ is S_k^i-sparse, with support $\mathcal{S}_k^i$. Remark that, along the iterations of the Boost-Boost algorithm, an incorrect element of the dictionary cannot be selected using Equation (11) (see Theorem 3.4 for some supplementary details). We observe then that assumption $\rho S(1+\nu^{-1}) < 1+\rho$ implies that at each step, each approximation is at most S-sparse. It trivially implies that $(S-1)\rho < 1$. We present an elementary lemma, proved by [14], which would be very useful until the end of the proof.

Lemma Appendix B.2. *Let $\mathcal{D} = (g_1, ..., g_p)$ a dictionary on H with coherence ρ.*

i) For any S-sparse family $(a_j)_{1\leq j\leq p}$, we have:

$$\left(\sum_{j=1}^{p} |a_j|^2\right)(1-\rho(S-1)) \leq \left\|\sum_{j=1}^{p} a_j g_j\right\|^2 \leq \left(\sum_{j=1}^{p} |a_j|^2\right)(1+\rho(S-1)).$$

ii) For any function f spanned on $\mathcal{D}$ as $f = \sum_{j=1}^{p} a_j g_j$, where $(a_j)_j$ is S -sparse, we have

$$\left(\sum_{j=1}^{p} |a_j|^2\right)^{1/2} (1-\rho(S-1)) \leq \left(\sum_{j=1}^{p} |\langle f, g_j\rangle|^2\right)^{1/2} \leq \left(\sum_{j=1}^{p} |a_j|^2\right)^{1/2} (1+\rho(S-1)).$$

Now, let $i \neq i_k$. By Lemma Appendix B.2 (r.h.s. of ii) and l.h.s. of i)) combined with condition $\rho(S-1) < 1$, we have

$$\sum_{j=1}^{p} |\langle R_{k-1}(f^{i_k}), g_j\rangle|^2 \leq \|R_{k-1}(f^{i_k})\|^2 \frac{(1+\rho(S-1))^2}{1-\rho(S-1)}. \tag{B.3}$$

Moreover Lemma Appendix B.2 again (l.h.s. of ii) and r.h.s. of i)) shows that

$$\forall 1 \leq i \leq m, \quad \sum_{j=1}^{p} |\langle R_{k-1}(f^i), g_j\rangle|^2 \geq \|R_{k-1}(f^i)\|^2 \frac{(1-\rho(S-1))^2}{1+\rho(S-1)}. \tag{B.4}$$

By definition of i_k (see Equation (10) in the Boost-Boost algorithm), we deduce that:

$$\begin{aligned} \forall i \in [\![1, m]\!], \qquad \sum_{j=1}^{p} |\langle R_{k-1}(f^{i_k}), g_j\rangle|^2 &\geq \mu \sum_{j=1}^{p} |\langle R_{k-1}(f^i), g_j\rangle|^2 \\ &\geq \mu \|R_{k-1}(f^i)\|^2 \frac{(1-\rho(S-1))^2}{1+\rho(S-1)}. \end{aligned} \tag{B.5}$$

The conclusion follows by using Equations (B.3) and (B.5). □

To conclude, we consider the Euclidean division of k by m: $k = mK + d$, where the remainder d is not greater than the divisor m. There exists a coordinate $i^* \in \{1 \dots m\}$, which is selected at least K times by Equation (9) or (10), hence $s_{i^*}(k) \geq K$. We also denote k^* the last step which selects i^* before step k. Since $\left(\|R_k(f^i)\|\right)_k$ is a non-increasing sequence along the iterations of the algorithm, by Equation (B.1), we have that:

$$\|R_{k-1}(f^{i^*})\| \leq \|R_{k^*-1}(f^{i^*})\| \leq C_B(1 + \nu^2\gamma(2-\gamma)(K-1))^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}. \quad \text{(B.6)}$$

The conclusion holds remarking that $\frac{k}{m} - 1 \leq K \leq \frac{k}{m}$ and $\nu < 1$, and using our bounds (B.2) for the Boost-Boost Residual L^2 norm algorithm, or Lemma Appendix B.1 for the Boost-Boost $\mathcal{D}$-Correlation sum algorithm.

Appendix B.2. Proof of Theorem 3.4

We begin this section by clarifying the proof of Theorem 3.4 since this result is needed to prove all others multi-task results. The proof rolls out in the same way as in section Appendix A.3. Our focus is on the choice of the regressor to add in the model whatever the column chosen to be regressed in the step before. So, in order to simplify, notations index i may be omitted and we can do exactly the same computations. Remark that the maximal shrinkage parameter allowed γ^* is equal to 13/18 for the Boost-Boost Residual L^2 norm algorithm, whereas $\gamma^* = 157/162$ for the Boost-Boost $\mathcal{D}$-Correlation algorithm, since the maximal number of iterations allowed is not exactly the same for the two algorithms (see section Appendix B.3 for more details).

Appendix B.3. Proof of Theorems 3.2 and 3.3

The proof of consistency results in the multi-task case rolls out as in Section Appendix A.2. Hence, we consider a semi-population version of the two Boost-Boost algorithms: let $(\tilde{R}_k(f))_k$ the phantom residuals, which is now living inH_m, initialised by $\tilde{R}_0(f) = f$, and satisfies at step k:

$$\begin{aligned} &\tilde{R}_k(f^i) = \tilde{R}_{k-1}(f^i) \ \text{ if } \ i \neq i_k, \\ &\tilde{R}_k(f^{i_k}) = \tilde{R}_{k-1}(f^{i_k}) - \gamma\langle \tilde{R}_{k-1}(f^{i_k}), \varphi_k\rangle\varphi_k, \end{aligned} \quad \text{(B.7)}$$

where (i_k, φ_k) is chosen according to Algorithm 4.

As previously, we aim at applying Theorem 3.1 to the phantom residuals. This will be possible if we can show an analogue of Equations (9) (for the Residual L^2 norm) or (10) (for the $\mathcal{D}$-Correlation sum) and (11). Remark that, from Theorem 3.4, sparsity of both residuals $\tilde{R}_k(f)$ and $\hat{R}_k(f)$ does not exceed S with high probability if we choose γ small enough in Equation (12).

We begin the proof by reminding Lemma Appendix A.1. In the multi-task case, this lemma can be easily extended as the following way:

Lemma Appendix B.3. *Assume that Hypotheses* $\mathbf{H_1^{Mult}}$ *are fulfilled on dictionary* $\mathcal{D}$*,* f *and* ε*, with* $0 < \xi < 1$ *as given in* $\mathbf{H_{1-2}^{Mult}}$*, then:*

i) $\sup_{1\leq i,j\leq p_n} |\langle g_i, g_j\rangle_{(n)} - \langle g_i, g_j\rangle| = \zeta_{n,1} = \mathcal{O}_P(n^{-\xi/2})$,

ii) $\sup_{1\leq i\leq p_n, 1\leq j\leq m_n} |\langle g_i, \varepsilon^j\rangle_{(n)}| = \zeta_{n,2} = \mathcal{O}_P(n^{-\xi/2})$,

iii) $\sup_{1\leq i\leq m_n, 1\leq j\leq p_n} |\langle f^i, g_j\rangle_{(n)} - \langle f^i, g_j\rangle| = \zeta_{n,3} = \mathcal{O}_P(n^{-\xi/2})$.

iv) $\sup_{1\leq i\leq m_n} |\|\varepsilon^i\|^2_{(n)} - \mathbb{E}(|\varepsilon^i|^2)| = \zeta_{n,4} = \mathcal{O}_P(n^{-\xi/2})$.

The first three points of Lemma Appendix B.3 are the same as i), ii) and iii) of Lemma Appendix A.1. The fourth point is something new, however, since it proof does not call typical boosting arguments, we don't state it here.

Denoting $\zeta_n = \max\{\zeta_{n,1}, \zeta_{n,2}, \zeta_{n,3}, \zeta_{n,4}\} = \mathcal{O}_P(n^{-\xi/2})$, we can show that Lemma Appendix A.2 is still true for the i_k-th coordinate of f. Moreover, let $i \neq i_k$. Since $\hat{R}_k(f^i) = \hat{R}_{k'}(f^i)$ for all $k' \leq k$ such that i_k is not selected between step k' and k (see Equation (12)), we can easily extend lemma Appendix A.2 to each coordinate of f:

$$\forall i \in [\![1, m_n]\!], \quad \sup_{1\leq j\leq p_n} |\langle \hat{R}_k(f^i), g_j\rangle_{(n)} - \langle \tilde{R}_k(f^i), g_j\rangle| \leq C\left(\frac{5}{2}\right)^k \zeta_n. \tag{B.8}$$

Using this extension of Lemma Appendix A.1, the same calculations of section Appendix A.2 can be done. Hence, considering the i_k-th coordinate of f chosen by Equations (9) or (10), on set Ω_n, inequality (A.3) also holds:

$$|\langle \tilde{R}(f^{i_k}), \varphi_k\rangle| \geq \frac{1}{2} \sup_{1\leq j\leq p_n} |\langle \tilde{R}_{k-1}(f^{i_k}), g_j|.$$

Consider now the Boost-Boost Residual L^2 norm algorithm. To obtain an analogue of (9), we need the following lemma, which compares norms of both residuals:

Lemma Appendix B.4. *Under Hypotheses* $\mathbf{H_1^{Mult}}$*, there exists a constant* $0 < C < +\infty$*, independent of* n *and* k*, such that on the set* $\Omega_n = \{\omega, |\zeta_n(\omega)| < 1/2\}$*:*

$$\forall i \in [\![1, m_n]\!], \quad |\|\hat{R}_{k-1}(f^i)\|^2_{(n)} - \|\tilde{R}_{k-1}(f^i)\|^2| \leq C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n.$$

Proof Consider the two residual sequences $(\hat{R}_k(f))_k$ and $(\tilde{R}_k(f))_k$, expanded on $\mathcal{D}$ as: $\hat{R}_{k-1}(f^i) = \sum_j \theta^k_{i,j} g_j$, and $\tilde{R}_{k-1}(f^i) = \sum_j \tilde{\theta}^k_{i,j} g_j$. Hence,

$$|\|\hat{R}_{k-1}(f^i)\|^2_{(n)} - \|\tilde{R}_{k-1}(f^i)\|^2| \leq \underbrace{|\sum_{j=1}^{p_n} \theta^k_{i,j} \left(\langle \hat{R}_{k-1}(f^i), g_j \rangle_{(n)} - \langle \tilde{R}_{k-1}(f^i), g_j \rangle \right)|}_{(I)}$$

$$+ \underbrace{|\sum_{j=1}^{p_n} \tilde{\theta}^k_{i,j} \left(\langle \hat{R}_{k-1}(f^i), g_j \rangle_{(n)} - \langle \tilde{R}_{k-1}(f^i), g_j \rangle \right)|}_{(II)}$$

$$+ \underbrace{|\sum_{j=1}^{p_n} \theta^k_{i,j} \langle \tilde{R}_{k-1}(f^i), g_j \rangle - \sum_{j=1}^{S} \tilde{\theta}^k_{i,j} \langle \hat{R}_{k-1}(f^i), g_j \rangle_{(n)}|}_{(III)}.$$

By Equation (B.8), we can provide two upper bounds for (I) and (II):

$$(I) \leq C \left(\frac{5}{2}\right)^{k-1} \sum_{j=1}^{p_n} |\theta^k_{i,j}| \zeta_n \quad \text{and} \quad (II) \leq C \left(\frac{5}{2}\right)^{k-1} \sum_{j=1}^{p_n} |\tilde{\theta}^k_{i,j}| \zeta_n.$$

Denoting $M := \max_{1 \leq j \leq S} \{|\theta^k_{i,j}|, |\tilde{\theta}^k_{i,j}|\}$, the following inequality holds for (I) and (II):

$$(I) \vee (II) \leq CMS \left(\frac{5}{2}\right)^{k-1} \zeta_n.$$

To conclude, using Lemma (Appendix B.3), one has:

$$(III) \leq \sum_{j=1}^{p_n} |\tilde{a}^k_{i,j}| \sum_{j'=1}^{p_n} |a^k_{i,j}| |\langle g_j, g_{j'} \rangle - \langle g_j, g_{j'} \rangle_{(n)}| \leq S^2 M^2 \zeta_n.$$

and the conclusion follows using our last bounds. □

Since Lemma Appendix B.4 is not directly applicable to the observed residual $Y - \hat{G}_k(f)$, the same calculation cannot be performed to obtain an analogue of Equation (9). However, we can compare the norm of the theoretical and observed residuals:

$$\begin{aligned} \forall i \in [\![1, m_n]\!], \quad \|Y^i - \hat{G}_{k-1}(f^i)\|^2_{(n)} &= \|\hat{R}_{k-1}(f^i) + \varepsilon^i\|^2_{(n)} \\ &= \|\hat{R}_{k-1}(f^i)\|^2_{(n)} + \|\varepsilon^i\|^2_{(n)} + 2\langle \hat{R}_{k-1}(f^i), \varepsilon^i \rangle_{(n)}. \end{aligned}$$

Note that, using Lemma Appendix B.3, we obtain: $|\langle \hat{R}_k(f^i), \varepsilon^i \rangle_{(n)}| \leq MS\zeta_n$, where M is defined in the proof of Lemma Appendix B.4. Hence, one has for all i:

$$\|\hat{R}_{k-1}(f^i)\|^2_{(n)} + \|\varepsilon^i\|^2_{(n)} - 2MS\zeta_n \leq \|Y^i - \hat{G}_{k-1}(f^i)\|^2_{(n)} \leq \|\hat{R}_{k-1}(f^i)\|^2_{(n)} + \|\varepsilon^i\|^2_{(n)} + 2MS\zeta_n. \tag{B.9}$$

Remind that $\mathbb{E}(|\varepsilon^i|^2)$ does not depend on i from Assumption $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{1-3}}$, and denote it by σ^2. Then, an application of Lemma Appendix B.3 iv) to Equation (B.9) yields

$$\|\hat{R}_{k-1}(f^i)\|^2_{(n)}+\sigma^2-(1+2MS)\zeta_n \leq \|Y^i-\hat{G}_{k-1}(f^i)\|^2_{(n)} \leq \|\hat{R}_{k-1}(f^i)\|^2_{(n)}+\sigma^2+(1+2MS)\zeta_n. \tag{B.10}$$

Hence, on Ω_n, by definition of i_k, Equation (B.10) and Lemma Appendix B.4, we can write:

$$\begin{aligned}
\|Y^{i_k} - \hat{G}_{k-1}(f^{i_k})\|^2_{(n)} &\geq \sup_{1\leq i\leq m_n} \|Y^i - \hat{G}_{k-1}(f^i)\|^2_{(n)} \\
&\geq \sup_{1\leq i\leq m_n} \left\{\|\hat{R}_{k-1}(f^i)\|^2_{(n)} + \sigma^2\right\} - (1+2MS)\zeta_n \\
&\geq \sup_{1\leq i\leq m_n} \left\{\|\tilde{R}_{k-1}(f^i)\|^2 + \sigma^2\right\} - C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n \\
&\quad -(1+2MS)\zeta_n. \qquad \text{(B.11)}
\end{aligned}$$

Using again the same calculus on set Ω_n:

$$\begin{aligned}
\|\tilde{R}_{k-1}(f^{i_k})\|^2 &\geq \|\hat{R}_{k-1}(f^{i_k})\|^2_{(n)} - C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n \\
&\geq \|Y^{i_k} - \hat{G}_{k-1}(f^{i_k})\|^2_{(n)} - \sigma^2 - (1+2MS)\zeta_n - C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n \\
&\geq \sup_{1\leq i\leq m_n} \left\{\|\tilde{R}_{k-1}(f^i)\|^2 + \sigma^2\right\} - \sigma^2 - 2(1+2MS)\zeta_n \\
&\quad -2C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n, \quad \text{by Equation (B.11).} \qquad \text{(B.12)}
\end{aligned}$$

We then obtain from Equation (B.12) that:

$$\|\tilde{R}_{k-1}(f^{i_k})\|^2 \geq \sup_{1\leq i\leq m_n} \|\tilde{R}_{k-1}(f^i)\|^2 - 2(1+2MS)\zeta_n - 2C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n. \tag{B.13}$$

Let $\check{\Omega}^1_n = \left\{\omega, \quad \forall k \leq k_n \quad \sup_{1\leq i\leq m_n} \|\tilde{R}_{k-1}(f^i)\|^2 > 4\left(1+2MS+C\left(2\left(\frac{5}{2}\right)^{k-1}+S\right)S\right)\zeta_n\right\}$. We deduce from Equation (B.13) the following inequality on set $\Omega_n \cap \check{\Omega}^1_n$:

$$\|\tilde{R}_{k-1}(f^{i_k})\|^2 \geq \frac{1}{2} \sup_{1\leq i\leq m_n} \|\tilde{R}_{k-1}(f^i)\|^2.$$

Consider finally the Boost-Boost $\mathcal{D}$-Correlation sum algorithm. To obtain an analogue of Equation (10), the following lemma is needed:

Lemma Appendix B.5. *Under Hypotheses* $\mathbf{H_1^{Mult}}$*, there exists a constant* $0 < C < +\infty$*, independent of* n *and* k *such that, on set* $\Omega_n = \{\omega, |\zeta_n(\omega)| < 1/2\}$*:*

$$\forall i \in [\![1, m_n]\!], \quad \sup_{1\le j\le p_n} |\langle \hat{R}_k(f^i), g_j\rangle_{(n)}^2 - \langle \tilde{R}_k(f^i), g_j\rangle^2| \le C\left(\frac{5}{2}\right)^{2k} \zeta_n.$$

Proof Let $k \ge 1$, $i \in [\![1, m_n]\!]$. We have the following equality:

$$|\langle \hat{R}_k(f^i), g_j\rangle_{(n)}^2 - \langle \tilde{R}_k(f^i), g_j\rangle^2| = |\langle \hat{R}_k(f^i), g_j\rangle_{(n)} - \langle \tilde{R}_k(f^i), g_j\rangle||\langle \hat{R}_k(f^i), g_j\rangle_{(n)} + \langle \tilde{R}_k(f^i), g_j\rangle|, \tag{B.14}$$

where $|\langle \hat{R}_k(f^i), g_j\rangle_{(n)} - \langle \tilde{R}_k(f^i), g_j\rangle| \le C\left(\frac{5}{2}\right)^k \zeta_n$ by Equation (B.8).

Moroever, using recursive equation for $(\hat{R}_k(f^{i_k}))_k$, we can obtain the following bounds:

$$\begin{aligned}
\left|\langle \hat{R}_k(f^{i_k}), g_j\rangle_{(n)}\right| &\le \left|\langle \hat{R}_{k-1}(f^{i_k}), g_j\rangle_{(n)}\right| + \gamma\left|\langle \hat{R}_{k-1}(f^{i_k}), \varphi_k\rangle_{(n)}\langle \varphi_k, g_j\rangle_{(n)}\right| \\
&\quad + \gamma\left|\langle \varepsilon^{i_k}, \varphi_k\rangle_{(n)}\langle g_j, \varphi_k\rangle_{(n)}\right| \\
&\le \sup_{1\le j\le p_n}\left|\langle \hat{R}_{k-1}(f^{i_k}), g_j\rangle_{(n)}\right|\left(1 + \gamma|\langle \varphi_k, g_j\rangle_{(n)}|\right) + \gamma\zeta_n(1+\zeta_n) \\
&\le M_{k-1}^{i_k}(1 + \gamma(1+\zeta_n)) + \gamma\zeta_n(1+\zeta_n),
\end{aligned}$$

where $M_k^i := \sup_{1\le j\le p_n} |\langle \hat{R}_k(f^i), g_j\rangle_{(n)}|$. Remark that for $i \ne i_k$, $M_k^i = M_{k-1}^i$.

On Ω_n, we hence have for a suitable constant $C > 0$:

$$M_k^i \le M_{k-1}^i\left(1 + \frac{3}{2}\gamma\right) + C \ldots \le \left(1 + \frac{3}{2}\gamma\right)^k \left(\sup_{n\in\mathbb{N}} \sum_{j=1}^{p_n} |a_{i,j}| + \frac{3}{2}\right) + C. \tag{B.15}$$

By Equation (7), $\|\tilde{R}_k(f^i)\|$ is non-increasing. Hence $\|\tilde{R}_k(f^i)\| \le \|f^i\|$. Cauchy-Schwarz inequality allows us to write that:

$$\left|\langle \tilde{R}_k(f^i), g_j\rangle\right| \le \|\tilde{R}_k(f^i)\| \le \|f^i\|. \tag{B.16}$$

Hence, the conclusion holds using Equations (B.15) and (B.16) in Equation (B.14) for a large enough constant C. □

Observe that Lemma Appendix B.5 remains true, if we change the observed residual by the theoretical residual. Hence, on set Ω_n,

$$\begin{aligned}
\sum_{j=1}^{p_n} |\langle Y^{i_k} - \hat{G}_{k-1}(f^{i_k}), g_j\rangle_{(n)}|^2 &\ge \sup_{1\le i\le m_n} \sum_{j=1}^{p_n} |\langle Y^i - \hat{G}_{k-1}(f^i), g_j\rangle_{(n)}|^2 \\
&\ge \sup_{1\le i\le m_n} \sum_{j=1}^{p_n} \left(|\langle \tilde{R}_{k-1}(f^i), g_j\rangle_{(n)}|^2 - C\left(\frac{5}{2}\right)^{2(k-1)} \zeta_n\right) \\
&\ge \sup_{1\le i\le m_n} \sum_{j=1}^{p_n} |\langle \tilde{R}_{k-1}(f^i), g_j\rangle_{(n)}|^2 - Cp_n\left(\frac{5}{2}\right)^{2(k-1)} \zeta_n
\end{aligned} \tag{B.17}$$

Hence, Lemma Appendix B.5 again, on Ω_n:

$$\sum_{j=1}^{p_n} |\langle \tilde{R}_{k-1}(f^{i_k}), g_j \rangle|^2 \geq \sum_{j=1}^{p_n} |\langle Y^{i_k} - \hat{G}_{k-1}(f^{i_k}), g_j \rangle_{(n)}|^2 - C p_n \left(\frac{5}{2}\right)^{2(k-1)} \zeta_n$$

$$\geq \sup_{1 \leq i \leq m_n} \sum_{j=1}^{p_n} |\langle \tilde{R}_{k-1}(f^i), g_j \rangle|^2 - 2C p_n \left(\frac{5}{2}\right)^{2(k-1)} \zeta_n \text{ by Equation (B.17).} \tag{B.18}$$

Let $\check{\Omega}_n^2 = \left\{ \omega, \ \ \forall k \leq k_n \quad \sup_{1 \leq i \leq m_n} \sum_{j=1}^{p_n} |\langle \tilde{R}_{k-1}(f^i), g_j \rangle|^2 > 4C p_n \left(\frac{5}{2}\right)^{2(k-1)} \zeta_n \right\}$.
We deduce from Equation (B.18) the following inequality on $\Omega_n \cap \check{\Omega}_n^2$:

$$\sum_{j=1}^{p_n} |\langle \tilde{R}_{k-1}(f^{i_k}), g_j \rangle|^2 \geq \frac{1}{2} \sup_{1 \leq i \leq m_n} \sum_{j=1}^{p_n} |\langle \tilde{R}_{k-1}(f^i), g_j \rangle|^2.$$

Consequently, on $\Omega_n \cap \tilde{\Omega}_n \cap \check{\Omega}_n^1$ and $\Omega_n \cap \tilde{\Omega}_n \cap \check{\Omega}_n^2$, we can apply Theorem 3.1 to family $(\tilde{R}_k(f^i))_k$, since it satisfies a deterministic Boost-Boost algorithm with constants $\tilde{\mu} = 1/2$, $\tilde{\nu} = 1/2$, and has a bounded sparsity S.

Consider now the set $\left(\check{\Omega}_n^2\right)^C$. Using Equation (B.4), we get

$$\|\tilde{R}_k(f^i)\|^2 \leq \frac{1 + \rho(S-1)}{(1 - \rho(S-1))^2} \sum_{j=1}^{p_n} |\langle \tilde{R}_k(f^i), g_j \rangle|^2 \leq 4 \frac{1 + \rho(S-1)}{(1 - \rho(S-1))^2} C p_n \left(\frac{5}{2}\right)^{2k} \zeta_n.$$

On the set $\left(\check{\Omega}_n^1\right)^C$, we also have:

$$\|\tilde{R}_k(f^i)\|^2 \leq 4 \left(1 + 2MS + C \left(2 \left(\frac{5}{2}\right)^k + S\right) S\right) \zeta_n.$$

The end of the proof follows as in section Appendix A.2 by remarking that $\mathbb{P}\left((\Omega_n \cap \tilde{\Omega}_n) \cup \tilde{\Omega}_n^C \cup \check{\Omega}_n^C\right) \geq \mathbb{P}(\Omega_n) \underset{n \to +\infty}{\longrightarrow} 1$. Note that the conclusion holds for a sequence k_n which grows sufficiently slowly: for the Boost-Boost Residual L^2 norm algorithm, k_n is alllowed to grow as $(\xi / 4 \log(3)) \log(n)$ whereas k_n can only grow as $(\xi / 8 \log(3)) \log(n)$ for the Boost-Boost $\mathcal{D}$-Correlation sum algorithm.

References

[1] V. N. Temlyakov, Advances in Computational Mathematics 12 (2000) 213–227.

[2] P. Bühlmann, Annals of Statistics 34 (2006) 559–583.

[3] R. Gribonval, M. Nielsen, Advances in Computational Mathematics 28 (2006) 23–41.

[4] J. A. Tropp, IEEE Trans. Inform. Theory 50 (2004) 2231–2242.

[5] T. Cai, T. Jiang, Annals of Statistics 39 (2011) 1496–1525.

[6] T. Zhang, Journal of Machine Learning Research 10 (2009) 555–568.

[7] V. N. Temlyakov, P. Zheltov, Journal of Approximation Theory 163 (2011) 1134–1145.

[8] G. Obozinski, M. Wainwright, M. Jordan, Annals of Statistics 39 (2011) 1–17.

[9] T. Cai, L. Wang, IEEE Transactions on Information Theory 57 (2011) 4680–4688.

[10] M. Wainwright, IEEE Transactions on Information Theory 55 (2009) 5728–5741.

[11] N. Verzelen, Electronic Journal of Statistics 6 (2012) 38–90.

[12] R. E. Schapire, in: Computational learning theory (Nordkirchen, 1999), volume 1572 of *Lecture Notes in Comput. Sci.*, Springer, Berlin, 1999, pp. 1–10.

[13] E. Candes, T. Tao, Annals of Statistics 35 (2007) 2313–2351.

[14] D. N. Donoho, M. Elad, V. N. Temlyakov, Journal of Approximation Theory 147 (2007) 185–195.