Accepted Manuscript

Sparse regression and support recovery with $\mathbb{L}_2$-Boosting algorithms

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

PII: S0378-3758(14)00120-7
DOI: http://dx.doi.org/10.1016/j.jspi.2014.07.006
Reference: JSPI 5304

To appear in: *Journal of Statistical Planning and Inference*

Received date: 22 February 2013
Revised date: 14 April 2014
Accepted date: 7 July 2014

Please cite this article as: Champion, M., Cierco-Ayrolles, C., Gadat, S., Vignes, M., Sparse regression and support recovery with $\mathbb{L}_2$-Boosting algorithms. J. Statist. Plann. Inference (2014), http://dx.doi.org/10.1016/j.jspi.2014.07.006

We investigate the prediction consistency and support recovery of L2 Boosting.

We extend these results to a high dimensional statistical framework.

We investigate the behaviour of such algorithms in a multivariate settings.

Sparse regression and support recovery with $\mathbb{L}_2$-Boosting algorithms

Magali Champion[1&2], Christine Cierco-Ayrolles[2], Sébastien Gadat[1] & Matthieu Vignes[2]

[1] *Institut de Mathématiques de Toulouse - Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse, Cedex 9, France*
{magali.champion,sebastien.gadat@math.univ-toulouse.fr}
[2] *INRA, UR875 MIA-T, F-31326 Castanet-Tolosan, France*
{christine.cierco,matthieu.vignes@toulouse.inra.fr}

Abstract

This paper focuses on the analysis of $\mathbb{L}_2$-Boosting algorithms for linear regressions. Consistency results were obtained for high-dimensional models when the number of predictors grows exponentially with the sample size n. We propose a new result for Weak Greedy Algorithms that deals with the support recovery, provided that reasonable assumptions on the regression parameter are fulfilled. For the sake of clarity, 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 that assumptions on the restricted isometry of the representation and on the sparsity of the model are fulfilled. The interest of these two algorithms is demonstrated on various datasets.

Keywords Boosting, regression, sparsity, high-dimension.

1 Introduction

Context of our work This paper presents a study of *Weak Greedy Algorithms* (WGA) and statistical $\mathbb{L}_2$-Boosting procedures derived from these WGA. These methods are dedicated to the approximation or estimation of several parameters that encode the relationships between input variables X and any response Y through a noisy linear representation $Y = f(X) + \varepsilon$, where ε models the amount of noise in the data. We assume that f may be linearly spanned on a predefined dictionary of functions $(g_j)_{j=1\dots p}$:

$$f(x) = \sum_{j=1}^{p} a_j g_j(x). \tag{1}$$

We aim at recovering unknown coefficients $(a_j)_{j=1\dots p}$

when one n-sample $(X_k, Y_k)_{k=1\dots n}$ is observed in the high-dimensional paradigm. Moreover, we are also interested in extending the boosting methods to the multi-task situation described in [HTF09]: Y is described by m coordinates $Y = (Y^1, \dots, Y^m)$, and each one is modelled by a linear relationship $Y^i = f^i(X) + \varepsilon^i$. These relationships are now parametrised through the family of unknown coefficients $(a_{i,j})_{1\le i\le m, 1\le j\le p}$. In both univariate or multivariate situations, we are primarily interested in the recovery of the structure (*i.e.* non-zero elements) of the matrix $A = (a_{i,j})_{1\le i\le m, 1\le j\le p}$, when a limited amount of observations n is available compared to the large dimension p of the feature space. In brief, the goal is to identify significant relationships between variables X and Y. We formulate this paradigm as a feature selection problem: we seek relevant elements of the dictionary $(g_j(X))_{j=1\dots p}$ that explain (in)dependencies in a measured dataset.

Feature selection algorithms can be split into three large families: exhaustive subset exploration, subspace methods, and forward algorithms with shrinkage. The exhaustive search suffers from an obvious hard combinatorial problem (see [Hoc83]), and subspace methods such as [Gad08] are generally time-consuming. In contrast, forward algorithms are fast, and shrinkage of greedy algorithms aims to reduce overfitting in stepwise subset selection (see [HTF09]). However, as pointed out by [ST07], *collinearities* may confuse greedy stepwise algorithms and subsequent estimates, which is not the case for the two other families of methods. Another main difficulty in our setting is that we often cope with *high-dimensional* situations where thousands of variables can be measured and where, at most, only a few hundred measures are manageable. For example, this is the case when dealing with biological network reconstruction, a problem that can be cast in a multivariate variable selection framework to decipher which regulatory relationships between entities actually dictate the evolution of the system [VVA$^+$11, OM12]. Several strategies were proposed to circumvent these hindrances in a statistical framework. Among them, in addition to a control on the isometry effect of the matrix X, the leading assumption of the *sparsity* of the solution A leads to satisfactory estimations. All the more, it is a quite reasonable hypothesis in terms of the nature of some practical problems. We clarify this notion of sparsity and give bounds for the applicability of our results. Note that Wainwright [Wai09] and Verzelen [Ver12b] established the limit of the statistical estimation feasibility of latent structures in random matrices with Gaussian noise and Gaussian Graphical Model frameworks, respectively.

Related works Among the large number of recent advances on linear regression within a sparse univariate setting, we focused our point of view and investigate the use of Weak Greedy Algorithms for estimating regression parameters of Equation (1). Since the pioneering works of Schapire [Sch90] and Schapire and Freund [SF96], there has been an abundant literature on *Boosting algorithms* (as an example, see [BY10] for a review). Friedman [FHT00] gave a statistical view of Boosting and related it to the maximisation of the likelihood in a logistic regression scenario (see [Rid99]). Subsequent papers also proposed algorithmic extensions (*e.g.*, a functional gradient descent algorithm with $\mathbb{L}_2$ loss function, [BY03]). For prediction or classification purposes, boosting techniques were shown to be particularly suited to large dataset problems. Indeed, just like the Lasso [Tib96] and the Dantzig Selector [CT07], which are two classical methods devoted to solving regression problems, Boosting uses variable selection, local optimisation and shrinkage. Even though Lasso, Dantzig and Elastic net ([ZH05]) estimators are inspired by penalised M-estimator methods and appear to be different from the greedy approach, like boosting methods, it is worthwhile to observe that, from an algorithmic point of view, these methods are very similar in terms of their practical implementation. Their behaviour is stepwise and based on correlation computed on the predicted residuals. We refer to [MRY07] for an extended comparison of such algorithms.

In a multivariate setting, some authors such as [LPvdGT11] or [OWJ11] use the geometric structure of an $\mathbb{L}_1$ ball derived from the Lasso approach . Others adopt a model selection strategy (see [SAB11]). Some authors also propose to use greedy algorithms such as Orthogonal Matching Pursuit developed in ([ER11]) or Basis Pursuit ([GN06]). More recently, due to their attractive computational properties and to their ability to deal with high-dimensional predictors, Boosting algorithms have been adapted and applied to bioinformatics for microarray data analysis as well as for gene network inference ([Büh06] and [ADH09]).

Organisation of the paper The works of [Tem00] and [TZ11] provide estimates of the rate of the approximation of a function by means of greedy algorithms, which inspired our present work. Section 2 is dedicated to Weak Greedy Algorithms. We first recall some key results needed for

our purpose. Section 2.1 may be omitted by readers familiar with such algorithms. In Section 2.2, we then provide a description of the behaviour of the $\mathbb{L}_2$-*Boosting algorithm* in reasonable noisy situations and in Section 2.3, we obtain a new result on support recovery. In Section 3, we describe two new extensions of this algorithm, referred to as *Boost-Boost algorithms*, dedicated to the *multi-task* regression problem. We also establish consistency results under some mild sparsity and isometry conditions. Section 4 is dedicated to a comparison of the performances of the Boosting methodology we propose with several approaches (Bootstrap Lasso [Bac08], Random Forests [Bre01] and remMap [PZB$^+$10] on several simulated datasets. The features of these datasets allow us to conclude that the two new Boosting algorithms are competitive with other state-of-the art methods, even when the theoretical assumptions of our results are challenged. For the sake of clarity, driving components of the proofs are given in the main text, whereas detailed proofs of theoretical results are presented in the Appendix of the paper.

2 Greedy algorithms

In this section, we describe some essential and useful results on greedy algorithms that 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 in this section are derived from those of Temlyakov [Tem00] and adapted to our particular setting. We slightly enrich the presentation by adding some supplementary shrinkage parameters, which offers additional flexibility in the noisy setting. It will in fact be necessary to understand the behaviour of the WGA with shrinkage to show the statistical consistency of the Boosting method.

2.1 A review of the Weak Greedy Algorithm (WGA)

Let H be a Hilbert space and $\|.\|$ denote 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:

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

Greedy algorithms generate iterative approximations of any $f \in H$, using a linear combination of elements of $\mathcal{D}$. Consistent with the notations of [Tem00], let $G_k(f)$ (as opposed to $R_k(f)$) denote the approximation of f (as opposed to the residual) at step k of the algorithm. These quantities are linked through the following equation:

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

At step k, we select an element $\varphi_k \in \mathcal{D}$, 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 into the model. Though the elements φ_k selected by (2) along the algorithm may not be uniquely defined, the convergence of the algorithm is still guaranteed by our next results. The second shrinkage parameter γ is the standard step-length parameter of the Boosting algorithm. It avoids a binary add-on, and actually smoothly inserts the new predictor into 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 the algorithm convergence rate. However, we decided to only consider the simplest version of WGA, because these improvements generally disappear in the noisy framework from a theoretical point of view (see [Büh06]).

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:

$$|\langle \varphi_k, R_{k-1}(f)\rangle| \geq \quad \nu \ \max_{g\in\mathcal{D}} \ |\langle g, R_{k-1}(f)\rangle|\,, \tag{2}$$

Step 2 Compute the current approximation and residual:

$$G_k(f) = G_{k-1}(f) + \gamma\langle R_{k-1}(f), \varphi_k\rangle\varphi_k \quad \text{and} \quad R_k(f) = R_{k-1}(f) - \gamma\langle R_{k-1}(f), \varphi_k\rangle\varphi_k. \tag{3}$$

end for

Following the arguments developed in [Tem00], 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*

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

then, for a suitable constant C_B that 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 in the noisy regression framework

This section aims at extending the previous results to several noisy situations. We present a noisy version of WGA, and we clarify the consistency result of [Büh06] by careful considerations on the empirical residuals instead of the theoretical ones (which are in fact unavailable; see Remark 1).

2.2.1 Noisy Boosting algorithm

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

$$\forall i = 1 \dots n, \qquad Y_i = f(X_i) + \varepsilon_i, \qquad \text{where} \qquad f = \sum_{j=1}^{p_n} a_j g_j. \tag{4}$$

The Hilbert space, $\mathbb{L}_2(P) := \{f, \|f\|^2 = \int f^2(x)dP(x) < \infty\}$, 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. Let us define the empirical inner product $\langle , \rangle_{(n)}$ as:

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

The empirical WGA is analogous to the coupled Equations (2) and (3), replacing $\langle , \rangle$ by the empirical inner product $\langle , \rangle_{(n)}$.

Algorithm 2 Noisy Weak Greedy Algorithm

Require: Observations $(X_i, Y_i)_{i=\{1...n\}}$, $\gamma \in (0,1]$ (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:

$$|\langle Y - \hat{G}_{k-1}(f), \varphi_k \rangle_{(n)}| = \max_{1 \leq j \leq p_n} |\langle Y - \hat{G}_{k-1}(f), g_j \rangle_{(n)}|. \tag{5}$$

Step 2: Compute the current approximation and residual:

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

end for

Remark 1 *The theoretical residual $\hat{R}_k(f) = \boldsymbol{f} - \hat{G}_k(f)$ cannot be used for the WGA (see Equations (5) and (6)) 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 so clear in the initial work of [Büh06], since notations used in its proofs are read as if $\hat{R}_k(f) = f - \hat{G}_k(f)$ was available. More explicit proofs are provided in Section A.2.*

2.2.2 Stability of the Boosting algorithm

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

Hypothesis $\mathbf{H_{dim}}$

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

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

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

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

Assumption $\mathbf{H_{dim-1}}$ is clearly satisfied for compactly supported real polynomials or Fourier expansions with trigonometric polynomials. Assumption $\mathbf{H_{dim-2}}$ bounds the high dimensional setting and states that $\log(p_n)$ should be, at the most, on the same order as n. Assumption $\mathbf{H_{dim-3}}$ specifies the noise and especially the size of its tail distribution. It must be centred with at least a bounded second moment. This hypothesis is required to apply the uniform law of large numbers and is satisfied by a great number of distributions, such as Gaussian or Laplace ones. The last assumption $\mathbf{H_{dim-4}}$ is a sparsity hypothesis on the unknown signal. It is trivially satisfied when the decomposition $(a_j)_{j=1...p_n}$ of f is bounded and has a fixed sparsity index:

Card $\{i|a_i \neq 0\} \leq S$. Note that it could be generalised to $\sum_{j=1}^{p_n} |a_j| \underset{n\to+\infty}{\longrightarrow} +\infty$ at the expense of additional restrictions on ξ and p_n (see Equation (19) in Appendix A.2).

We then formulate the first important result of the Boosting algorithm, obtained by [Büh06], which represents a *stability result.*

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

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

We only give the outline of the proof here. Details can be found in the Appendix. A straightforward calculation shows that the theoretical residuals are updated as:

$$\hat{R}_k(f) = \hat{R}_{k-1}(f) - \gamma\langle \hat{R}_{k-1}(f), \varphi_k\rangle_{(n)}\varphi_k - \gamma\langle \varepsilon, \varphi_k\rangle_{(n)}\varphi_k. \tag{7}$$

The proof then results from the study of a *phantom* algorithm, which reproduces the behaviour of the deterministic WGA. In this algorithm, the inner product $\langle , \rangle$ replaces its empirical counterpart, and the (random) sample-driven choice of dictionary element $(\varphi_k)_{k\geq 0}$ is governed by Equation (5) of Algorithm 2. The phantom residuals are initialised by $\tilde{R}_0(f) = \hat{R}_0(f) = f$ and satisfy the following equation at step k:

$$\tilde{R}_k(f) = \tilde{R}_{k-1}(f) - \gamma\langle \tilde{R}_{k-1}(f), \varphi_k\rangle\varphi_k, \tag{8}$$

where φ_k is chosen using Equation (5). On the one hand, we establish an analogue of Equation (2) for φ_k, which allows us to apply Theorem 2.1 to the phantom residual $\tilde{R}_k(f)$. On the other hand, we provide an upper bound for the difference between $\hat{R}_k(f)$ and $\tilde{R}_k(f)$. The proof then results from a careful balance between these two steps.

2.3 Stability of support recovery

2.3.1 Ultra-high dimensional case

This paragraph presents our main results in the univariate case for the ultra-high dimensional case. We prove the stability of the support recovery with the noisy WGA. Provided that assumptions on the amplitude of the active coefficients of f and the structure of the dictionary are fulfilled, the WGA exactly recovers the support of f with high probability. This result is related to the previous work of [Tro04] and [Zha09] for recovering sparse signals using Orthogonal Matching Pursuit.

To state the theorem, we denote D as the $n \times p$ matrix whose columns are the p elements $(g_1, ..., g_p)$ of the dictionary $\mathcal{D}$. In the following text, $D_{\mathcal{S}}$ will be the matrix D restricted to the elements of $\mathcal{D}$ that are in $\mathcal{S} \subset [\![1, p]\!]$. Since $D_{\mathcal{S}}$ is not squared and therefore not invertible, D^+ is written as its pseudo-inverse. If we denote $\mathcal{S}$ as the support of f and S as its cardinality, we can then make the following assumptions.

Hypothesis $\mathbf{H_S}$: The matrix $D_{\mathcal{S}}$ satisfies:

$$\max_{j\notin\mathcal{S}} \|D_{\mathcal{S}}^+ g_j\|_1 < 1.$$

This assumption is also known as the exact recovery condition (see [Tro04]). It will ensure that only active coefficients of f can be selected along the iterations of Algorithm 2 (noisy Boosting algorithm).

Hypothesis $\mathbf{H_{RE^-}}$: A $\lambda_{min} > 0$ independent of n exists so that:

$$\inf_{\beta, \mathrm{Supp}(\beta) \subset \mathcal{S}} \|D\beta\|^2 / \|\beta\|^2 \geq \lambda_{min}.$$

λ_{min} of Assumption $\mathbf{H_{RE^-}}$ is the smallest eigenvalue of the restricted matrix ${}^tD_{\mathcal{S}}D_{\mathcal{S}}$. Assumption $\mathbf{H_{RE^-}}$ stands for the restricted isometry condition [CT05] or the sparse eigenvalue condition (e.g., [Zha09] and [ZY06]). Remark that our assumption is different from that of [Zha09] since we assume that $\forall j, \|g_j\| = 1$. For more details about this assumption, see Section 3.2.

Hypothesis $\mathbf{H_{SNR}}$: Elements $(a_j)_{1 \leq j \leq p_n}$ satisfy:

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

Note that the greater the number of variables is, the larger the value of active coefficients of f are and the more restrictive Assumption $\mathbf{H_{SNR}}$ is (see Section 2.3.2 below).

Theorem 2.3 (Support recovery) *(i) Assume that Hypotheses $\mathbf{H_{dim}}$ and $\mathbf{H_S}$ hold. Then, with high probability, only active coefficients are selected by Equation (5) along iterations of Algorithm 2.*

(ii) Moreover, if Hypotheses $\mathbf{H_{RE^-}}$ and $\mathbf{H_{SNR}}$ hold with a sufficiently small $\kappa < \kappa^*$ (κ^* only depending on γ), then Algorithm 2 fully recovers the support of f with high probability.

Similar results are already known for other algorithms devoted to sparse problems (see [GN06] for Basis Pursuit algorithms, and [Tro04], [CJ11] or [Zha09] for Orthogonal Matching Pursuit (OMP)). It is also known for other signal reconstruction algorithms [OWJ11], [CW11], [Zha09], which also rely on a sparsity assumption. Our assumption is stronger than the condition obtained by [Zha09] since active coefficients should be bounded from below by a power of $\log(n)^{-1}$ instead of $\log(p)^{1/2} n^{-1/2}$ in Theorem 4 of [Zha09]. However, obtaining optimal conditions on active coefficients is not straightforward and beyond the scope of this paper. The *weak* aspect of WGA seems harder to handle compared to the treatment of OMP (for example) because the amplitude of the remaining coefficients on active variables has to be recursively bound from one iteration to the next, according to the size of shrinkage parameters.

Let $\rho := \max_{1 \leq i \neq j \leq n} |\langle g_i, g_j \rangle|$ be the coherence of the dictionary $\mathcal{D}$. For non-orthogonal dictionaries, which are common settings of real high-dimensional datasets, the coherence is non-null. A sufficient condition to obtain the support recovery result would then be $\rho(2S-1) < 1$, where $S := |\mathcal{S}|$ is the number of non-null coordinates of f, combined with $\mathbf{H_{SNR}}$. However, it should be observed that this assumption is clearly more restrictive than $\mathbf{H_{RE^-}}$ when the number of predictors p_n becomes large.

In summary, a trade-off between signal sparsity, dimensionality, signal-to-noise ratio and sample size has to be reached. We provide explicit constant bounds for results on similar problems. Very interesting discussions can be found in [Wai09] (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 section on *Sparsity and ultra-high dimensionality* of [Ver12b].

2.3.2 High dimensional case

In this paragraph, we restrict our study to high-dimensional models, where the number of predictors should be, at the most, on the same order of n: $p_n = \underset{n \to +\infty}{O}(n^a)$ with $a > 0$. Then, provided that Assumption $\mathbf{H^+_{SNR}}$ below is fulfilled, Theorem 2.3 still holds.

Hypothesis $\mathbf{H^+_{SNR}}$: Elements $(a_j)_{1\leq j\leq p_n}$ satisfy:

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

Indeed, following the proof of Theorems 2.2 and 2.3, our assumption about the size of p_n implies that $\zeta_n = \mathcal{O}_P(\exp(-n^{1-\xi}))$ in the uniform law of large numbers (Lemma A.1), where ξ is given by $\mathbf{H_{dim-3}}$. The number of iterations of Algorithm 2 is then allowed to grow with n since $k_n := Cn^\beta$, with $\beta < 1-\xi$, which ensures that $\left(\frac{5}{2}\right)^{k_n}\zeta_n$ is small enough. The decrease of the theoretical residuals $(\|\hat{R}_k\|^2)_k$ is finally on the order of $Cn^{-\beta\alpha}$, where C depends on the shrinkage parameters γ and ξ, although α depends on the rate of approximation of the boosting ($\alpha = (2-\gamma)/(2(6-\gamma))$). Now Theorem 2.3 follows with $\kappa < \kappa^* := \beta\alpha/2$.

As a consequence, in the high-dimensional case, Assumption $\mathbf{H^+_{SNR}}$ is less restrictive than Assumption $\mathbf{H_{SNR}}$ and Algorithm 2 converges faster and can easily recover even small active coefficients of the true function f.

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

In this section, our purpose is to extend the above algorithm and results to the multi-task situation. The main focus of this work lies in the choice of the optimal task to be boosted. We therefore propose a new algorithm that follows the initial spirit of iterative Boosting (see [Sch99] for further details) and the multi-task structure of f. We first establish an approximation result in the deterministic setting and we then extend the stability results of Theorems 2.2 and 2.3 to the so called Boost-Boost algorithm for noisy multi-task regression.

3.1 Multi-task Boost-Boost algorithms

Let $H_m := H^{\otimes m}$ denote 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:

$$\forall i \in [\![1,m_n]\!], \qquad f^i = \sum_{j=1}^{p_n} a_{i,j}g_j. \tag{9}$$

A canonical extension of WGA to the multi-task problem can be computed as follows (Algorithm 3).

In the multi-task framework at step k, it is crucial to choose the coordinate from among the residuals that 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 the 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, relying on either the Residual L^2 norm - Equation (10) - or on the $\mathcal{D}$-Correlation sum - Equation (11).

Note that this latter choice is rather different from the choice proposed in [GN06], which uses the multichannel energy and sums the correlations of each coordinate 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 [CT07] (page 2316).

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:

$$\|R_{k-1}(f^{i_k})\|^2 \geq \mu \max_{1 \leq i \leq m} \|R_{k-1}(f^i)\|^2, \qquad \text{[Residual } L^2 \text{ norm]} \tag{10}$$

or to

$$\sum_{j=1}^{p} \langle R_{k-1}(f^{i_k}), g_j \rangle^2 \geq \mu \max_{1 \leq i \leq m} \sum_{j=1}^{p} \langle R_{k-1}(f^i), g_j \rangle^2, \qquad [\mathcal{D}\text{-Correlation sum}] \tag{11}$$

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

$$|\langle R_{k-1}(f^{i_k}), \varphi_k \rangle| \geq \nu \max_{1 \leq j \leq p} |\langle R_{k-1}(f^{i_k}), g_j \rangle|, \tag{12}$$

Step 3: Compute the current approximation:

$$\begin{aligned} &G_k(f^i) = G_{k-1}(f^i), \quad \forall i \neq i_k, \\ &G_k(f^{i_k}) = G_{k-1}(f^{i_k}) + \gamma \langle R_{k-1}(f^{i_k}), \varphi_k \rangle \varphi_k. \end{aligned} \tag{13}$$

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

Although [CT07] tends toward a final advantage for the $\mathcal{D}$-Correlation sum alternative, we also consider the Residual L^2 norm that seems more natural. In fact, it relies on the norm of the residuals themselves instead of the sum of information gathered by individual regressors on each residual. Moreover, conclusions of [CT07] 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 (10) and (12) in the Residual L^2 norm Boost-Boost algorithm, whereas we use criteria of Equations (11) and (12) in its $\mathcal{D}$-Correlation sum counterpart.

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 with either choice (10) or (11). 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 as the minimal amount of sparsity for such a representation. We then prove a first approximation result provided that the following assumption is true.

Hypothesis $\mathbf{H_{RE^+}}$ A $\lambda_{max} < \infty$ independent of n exists so that:

$$\sup_{\beta, \text{Supp}(\beta) \subset \mathcal{S}} \|D\beta\|^2 / \|\beta\|^2 \leq \lambda_{max}.$$

λ_{max} of Assumption $\mathbf{H_{RE^+}}$ is the largest eigenvalue of the restricted matrix ${}^tD_{\mathcal{S}}D_{\mathcal{S}}$. Note that

$$\begin{aligned} \forall u \in \mathbb{R}^S, \quad {}^tu {}^tD_{\mathcal{S}} D_{\mathcal{S}} u = \|D_{\mathcal{S}} u\|^2 &\leq \|u\|^2 \sum_{j \in \mathcal{S}} \|g_j\|^2 \\ &\leq S\|u\|^2. \end{aligned} \tag{14}$$

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:

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

or to

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

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

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

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

Then, denote v as the eigenvector associated with the largest eigenvalue λ_{max} of ${}^t D_{\mathcal{S}} D_{\mathcal{S}}$. Equation (14) then makes it possible to write:

$${}^t v \lambda_{max} v \le S\|v\|^2,$$

which directly implies the following bound for λ_{max}: $\lambda_{max} \le S$. Then, if S is kept fixed independent from n, Assumption $\mathbf{H_{RE^+}}$ trivially holds.

On the other hand, if S is allowed to grow with n as $S/n \to_{n\to+\infty} l$, [BCT11] proves that the expected value of λ_{max} is also bounded for the special Wishart matrices:

$$\mathbb{E}(\lambda_{max}) \underset{n\to+\infty}{\longrightarrow} (1+\sqrt{l})^2.$$

Moreover, they show that fluctuations of λ_{max} around $\mathbb{E}\lambda_{max}$ are exponentially small with n, that is:

$$\mathbb{P}\left(\lambda_{max} > \mathbb{E}\lambda_{max} + \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 0, \quad \text{exponentially fast with } n.$$

In the case of matrices with subgaussian entries, with probability $1 - c\exp(-S)$, [Ver12a] also provides the following bound for λ_{min} and λ_{max}:

$$\sqrt{S/n} - c \le \lambda_{min} \le \lambda_{max} \le \sqrt{S/n} + c.$$

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

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

$$\sup_{1\leq i\leq m} \|R_k(f^i)\| \leq C_B \mu^{-\frac{1}{2}} \nu^{-\frac{\nu(2-\gamma)}{2+\nu(2-\gamma)}} \left(\gamma(2-\gamma)\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}} \left(\frac{k}{m}\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}.$$

(ii) Assume that Hypotheses $\mathbf{H_{RE^-}}$ and $\mathbf{H_{RE^+}}$ hold. A suitable constant $C_{\lambda_{min},B}$ then exists so that the approximations provided by the $\mathcal{D}$-Correlation sum Boost-Boost algorithm satisfy, for all $k \geq m$:

$$\sup_{1\leq i\leq m} \|R_k(f^i)\| \leq C_{\lambda_{min},B} \mu^{-\frac{1}{2}} \nu^{-\frac{\nu(2-\gamma)}{2+\nu(2-\gamma)}} \left(\gamma(2-\gamma)\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}} \left(\frac{k}{m}\right)^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}.$$

Remark 2 *Note first that Theorem 3.1 is a uniform result over the m_n coordinates. Then, note that Assumptions $\mathbf{H_{RE^-}}$ and $\mathbf{H_{RE^+}}$ are needed to obtain the second part of the theorem since we have to compare each coordinate of the residual with the coordinate chosen at step k. For the Residual L^2 norm Boost-Boost algorithm, this comparison trivially holds.*

We can discuss the added value brought by the Residual L^2 norm Boost-Boost algorithm. Compared to running m times standard WGA on each coordinate 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 contrast, when WGA is applied to well balanced tasks, there is no clear advantage to using the 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 the multi-task WGA. We first state several assumptions adapted to the multi-task setting.

Hypothesis $\mathbf{H^{Mult}_{dim}}$

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

$\mathbf{H^{Mult}_{dim-2}}$ $\xi \in (0,1), C > 0$ exist so that the number of predictors and tasks (p_n, m_n) satisfies

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

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

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

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

It should be noted that a critical change appears in Hypothesis $\mathbf{H^{Mult}_{dim-3}}$. Indeed, all tasks should be of equal variance. We thus need to normalise the data before applying the Boost-Boost algorithms.

We can therefore derive a result on the consistency of the Residual L^2 norm Boost-Boost algorithm. This extends the result of Theorem 2.2 for univariate WGA.

Theorem 3.2 (Consistency of the Boost-Boost Residual L^2 norm) *Assume that Hypotheses* $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim}}$, $\mathbf{H}_{\mathbf{RE}^-}$ *and* $\mathbf{H}_{\mathbf{RE}^+}$ *are fulfilled. A sequence* $k_n := C\log(n)$ *then exists, with* $C < \xi/4\log(3)$*, so that:*

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

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 $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim-2}}$, we prove a similar result.

Theorem 3.3 (Consistency of the Boost-Boost $\mathcal{D}$-Correlation sum algorithm) *Assume that Hypotheses* $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim}}$, $\mathbf{H}_{\mathbf{RE}^-}$ *and* $\mathbf{H}_{\mathbf{RE}^+}$ *are fulfilled, with* $p_n = \underset{n\to+\infty}{O}(n^{\xi/4})$*. A sequence* $k_n := C\log(n)$ *then exists with* $C < \xi/8\log(3)$ *so that:*

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

We concede that Assumption $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim-2}}$ includes the very high-dimensional case. Theorem 3.3 has a slightly more restrictive assumption and encompasses the high-dimensional perspective from a theoretical point of view.

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

Theorem 3.4 (Support recovery) *Assume Hypotheses* $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim}}$, $\mathbf{H}_{\mathbf{S}}$, $\mathbf{H}_{\mathbf{RE}^-}$ *and* $\mathbf{H}_{\mathbf{RE}^+}$ *are fulfilled, then the two propositions hold.*

(i) With high probability, only active coefficients are selected along iterations of Algorithm 4.

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

4 Numerical applications

This section is dedicated to simulation studies to assess the practical performances of our method. We compare it to existing methods, namely the Bootstrap Lasso [Bac08], Random Forests [Bre01] and the recently proposed remMap [PZB+10]. The aim of these applications is twofold. Firstly, we assess the performance of our algorithms in light of expected theoretical results and as compared to other state-of-the-art methods. Secondly, we demonstrate the ability of our algorithm to analyse datasets that have features encountered in real situations. Three types of data sets are used. The two first types are challenging multivariate, noisy, linear datasets with different characteristics, either uni-dimensional or multi-dimensional. The third type consists in a simulated dataset that mimics the behaviour of a complex biological system through observed measurements. Datasets and codes used in the experiments are available upon request from the authors.

First, we briefly present the competing methods. We then introduce the criteria we used to assess the merits of the different methods (including a numerically-driven stopping criterion). Datasets are precisely described in a dedicated paragraph. Finally, in the last paragraph, we discuss the obtained results. For the sake of convenience, we will shortcut the notation p_n to p as well as m_n to m in the sequel.

Algorithms and methods We used our two proposed Boost-Boost algorithms (denoted "$\mathcal{D}$-Corr" for the Boost-Boost $\mathcal{D}$-Correlation sum algorithm and "L^2 norm" for the Boost-Boost Residual L^2 norm algorithm) with a shrinkage factor $\gamma = 0.2$. When the number of responses m is set to 1, these two algorithms are similar to Algorithm 2 and will both be referred to as "WGA".

We compared them to a bootstrapped version of the Lasso, denoted "BootLasso" thereafter. The idea of this algorithm is essentially that of the algorithm proposed by Bach [Bac08]: it uses bootstrapped estimates of the active regression set based on a Lasso penalty. In [Bac08], only variables that are selected in every bootstrap are kept in the model, and actual coefficient values are estimated from a straightforward least square procedure. Due to high-dimensional settings and empirical observations, we slightly relaxed the condition for a variable to be selected: at a given penalty level, the procedure keeps a variable if more than 80% of bootstrapped samples lead to select it in the model. We computed a 5-fold cross-validation unknown parameter estimate. The R package `glmnet` v$1.9-5$ was used for the BootLasso simulations.

The second approach we used is a random forest algorithm [Bre01] in regression, known to be suited to reveal interactions in a dataset, denoted as "RForests". It consists in a set (the forest) of regression trees. The randomisation is combined into 'bagging' of samples and random selection of feature sets at each split in every tree. For each regression, predictors are ranked according to their importance, which computes the squared error loss when using a shuffled version of the variable instead of the original one. We filtered for variables that have a negative importance. Such variables are highly non-informative since shuffling their sample values leads to an increased prediction accuracy; this can happen for small sample sizes or if terminal leaves are not pruned at the end of the tree-building process. No stopping criterion is implemented since it would require storing all partial depth trees of the forest and would be very memory-consuming. However, in each forest, we artificially introduced a random variable made up of a random blend of values observed on any variable in the data for each sample. The rationale is that any variable that achieves a lower importance than this random variable is not informative and should be discarded from the model. For each forest, we repeated this random variable inclusion a hundred times. We selected a variable if its importance was at leaste 85 times out of 100 higher than that of the artificially introduced random variable, their importance could serve to rank them. We also computed a final prediction $\mathbb{L}_2$-error for the whole forest and model selection metrics associated with correctly predicted relationships. The R package `randomForest` v$4.6-7$ was used for the RForests simulations. Notice that the total running time for RForests is linear in the size of the output variables. Hence, when $m = 250$ (correlated covariates or correlated noise), the total running time is nearly 4 days. We hence present partial results in these two cases on a very limited number of networks (5).

Finally, we compared our method to "remMap" (REgularized Multivariate regression for identifying MAster Predictors) that essentially produces sparse models with several very important regulatory variables in a high-dimensional setting. We refer to it as REM later in the paper. More specifically, REM uses an $\mathbb{L}_1$-norm penalty to control the overall sparsity of the coefficient matrix of the multivariate linear regression model. In addition, REM imposes a "group sparse" penalty, which is pasted from the group lasso penalty ([YL07]). This penalty puts a constraint on the $\mathbb{L}_2$ norm of regression coefficients for each predictor, which controls the total number of predictors entering the model and consequently facilitates the detection of so-called master predictors. We used the R package `remMap` v$0.1-0$ in our simulations. Parameter tuning was performed using the built-in cross-validation function. We varied parameters for DS1 and DS3 from 10^{-5} to 10^5 with a 10-fold multiplicative increment; for DS2, DS4 and DREAM datasets, the package could only run with parameters varying from 10^{-2} to 10^2. Lastly, in the very high-

dimensional settings of our scenarii ($p = m = 250$), the built-in cross-validation function of the `remMap` package wouldn't allow us to visit parameters outside the range 10^{-1} to 10^{1}, with over 24 hours of computation per network.

Performance assessment and stopping criterion An important issue when implementing a Boosting method, or any other model estimation procedure from a dataset, is linked to the definition of a stopping rule. It ideally guarantees that the algorithm ran long enough to provide informative conclusions without over-fitting the data. Cross-Validation (CV) or information criteria such as AIC or BIC address this issue. [LB06] presented a corrected AIC criterion. Firstly, the prediction error is required at each step and, secondly, the number of degrees of freedom of Boosting has to be evaluated. The latter is equal to the trace of a 'hat matrix' $\hat{H}$ (see [HTF09] or [BY03]). $\hat{H}$ is defined as the operator that enables the estimation $\hat{A}$ from the true parameter only. However, as pointed by [LB06], the computation of the hat matrix at step k has a complexity of $O(n^2p + n^3m^2k)$, and thus becomes not feasible if n, p or m are too large. For example, the computation of the hat matrix at the initialization of the algorithm (iteration $k = 1$) with $n = 100$ and $p = m = 250$ requires 6.10^8 operations, which takes around 7 seconds on an actual standard computer. Consequently, a typical run of the algorithm requires hundreds of iterations, which would last almost 10 hours just for selection purpose and is not reasonnable in practice.

We hence chose to use 5-fold cross-validation to assess the optimal number of iterations. Finally, it should be noted that cross-validation should be carefully performed, as pointed out by the erratum of [GWBV02]. It is imperative not to use the same dataset to both optimise the feature space and compute the prediction error. We refer the interested reader to the former erratum of [GWBV02] and several comments detailed in [AM02].

It should be noted that, in our simulation study, the cross-validation error E_{CV} decreases along the step of the Boosting algorithm while new variables are added in the model. The selected model was the one estimated after the first iteration that made the ratio of the total variation in the cross-validation error $|(E_{CV} - E_{min})/(E_{max} - E_{min})|$, where E_{max} and E_{min} are the maximal and the minimal values of the cross-validation error, below a 5% threshold.

The performances are measured through the normalised prediction error, also known as the mean square error:

$$MSE = \frac{1}{m}\sum_{i=1}^{m} \|Y^i - \hat{G}_{\hat{k}}(f^i)\|^2_{(n)},$$

where $\hat{G}_{\hat{k}}(f^i)$ denotes the approximation of coordinate i of f. We also report the rate of coefficients inferred by mistake (false positives, FP) and not detected (false negatives, FN).

First dataset We use two toy examples in both univariate ($m = 1$) and multi-task ($m = 5$ and $m = 250$) situations, with noisy linear datasets with different characteristics. They are simulated according to a linear modelling:

$$Y = XA + \varepsilon = f(X) + \varepsilon,$$

where Y is a $n \times m$ response matrix, X is a $n \times p$ observation matrix, ε is an additional Gaussian noise and A is the $p \times m$ S-sparse parameter matrix that encodes relationships to be inferred. Covariates are generated according to a multi-variate Gaussian distribution $\forall i, X_i \sim \mathcal{N}(0, I_p)$. Errors are generated according to a multi-variate normal distribution with an identity covariance matrix. Non-zero A-coefficients are set equal to 10 when $(p, m, S) = (250, 1, 5)$ and 1 for all other datasets.

In all our simulations, we always generate $n = 100$ observations; this situation corresponds to either moderate or very high-dimensional settings, depending on the number of explanatory variables (p) or on the number of response variables (m). Unless otherwise stated, all experiments are replicated 100 times and results are averaged over these replicates.

Prediction performances of tested methods are detailed in Table 1. In the first three simulation settings, when $m = 1$, the prediction performances of the Boosting algorithms are quite similar to those of the BootLasso and RF ones (see Table 1), but when the number of predictors is set to $1,000$, BootLasso results are poorer. REM seems to achieve a better prediction than other approaches, especially in the very high-dimensional settings ($p = 1,000$ while $m = 1$). This is still the case when $p = 250$ and $m = 5$ or 250. which .

(p,m,S)	(250,1,5)	(250,1,10)	(1000,1,20)	(250,5,50)	(250,250,1250)
WGA	0.21	0.23	0.42	$\emptyset$	$\emptyset$
$\mathcal{D}$-Corr	$\emptyset$	$\emptyset$	$\emptyset$	0.39	0.36
L^2 norm	$\emptyset$	$\emptyset$	$\emptyset$	0.40	0.38
BootLasso	0.30	0.28	0.78	0.31	0.40
RForests	0.18	0.25	0.49	0.41	0.20*
REM	0.33	0.18	0.08	0.21	0.19

Table 1: First dataset: MSE for the Boosting algorithms, with a shrinkage factor $\gamma = 0.2$, compared to the BootLasso, RForests and REM; the sample size n is set to 100. (*: for 5 simulated replicate data sets only as the running time for RForest was 4 days per network)

(p,m,S)	(250,1,5)		(250,1,10)		(1000,1,20)		(250,5,50)		(250,250,1250)	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
WGA	0.00	0.00	0.43	0.10	0.62	41.5	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
$\mathcal{D}$-Corr	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	0.84	3.42	0.10	0.65
L^2 norm	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	0.85	4.68	0.09	0.73
BootLasso	0.00	19.00	0.03	30.70	0.00	89.25	0.10	31.80	0.00	32.03
RForests	2.10	0.20	3.67	23.10	1.01	60.25	3.29	32.02	2.47*	2.76*
REM	0.58	0.00	1.49	0.00	5.53	6.65	2.66	0.00	2.35	0.00

Table 2: First dataset: Percentage of false positive FP coefficients and false negative FN coefficients for the Boosting algorithms, with a shrinkage factor $\gamma = 0.2$, compared to the BootLasso, RForests and REM; the sample size n is set to 100. (*: for 5 simulated replicate data sets only as the running time for RForest was 4 days per network)

Looking at the accuracy results of Table 2 at the same time is instructive: neither BootLasso nor RF succeed at recovering the structure of f, with the FN rate much higher than that of the $\mathbb{L}_2$-Boosting and REM approaches. In the moderately high-dimensional univariate setting $(p, m) = (250, 1)$, WGA and REM almost always recover the full model with few FP, while BootLasso and RF miss one third and one fourth of the correct edges, respectively. Figures in the high-dimensional univariate case $(p, m) = (1,000, 1)$ confirm this trend with a better precision for WGA, whereas REM achieves a better recall. This probably explains the much lower MSE for REM: the model selected in the REM framework is much richer and contains the vast majority of relevant relationships at the price of a low precision (just below 30%). In contrast, the model built by WGA is sparser with fewer FP, but misses some correct relationships.

We therefore empirically observe here that MSE is not too informative for feature selection, as reported by [HTF09], for example. The conclusion we can draw follow the same tendency in the high-dimensional multivariate settings $(p, m) = (250, 5)$ and $(p, m) = (250, 250)$. Again, REM is more comprehensive in retrieving actual edges, but it produces much more FP relationships than the multivariate boosting algorithms we presented.

In addition to the performance value, Fig. 1 represents the norm of each coordinate of the residual along the iterations of the Boost-Boost $\mathcal{D}$-Correlation sum algorithm when the number of predictors p is equal to 250 and the number of responses m is equal to 5 (A then includes $S = 50$ non-zero coefficients). Figure 1 shows that no residual coordinate is preferred along the iterations of the Boost-Boost $\mathcal{D}$-Correlation sum algorithm.

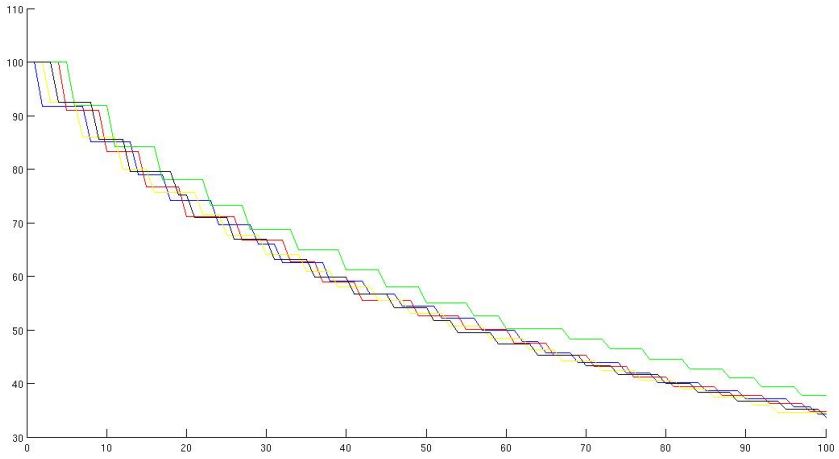

Figure 1: First dataset $(p, m, S) = (250, 5, 50)$: Norm of each coordinate of residuals along the first 100 iterations of the Boost-Boost $\mathcal{D}$-Correlation sum algorithm; the sample size n is set to 100.

Second dataset The following dataset stands for a more extreme situation. It is specifically designed to illustrate the theoretical results we presented on permissive sparsity and the lower bound of regression parameters. The idea is to consider a column-wise unbalanced design with highly correlated predictors or highly correlated noise coordinates (correlations can be as strong as ± 0.9). More precisely, we generated the second dataset with $p = 250$ and $m = 250$ as follows. For the first task (first column of X), we fixed 10 non-zero coefficients and set their value to 500. For each task from 2 to 241, we chose 10 coefficients and set their value to 1. The last 9 columns have respectively 9, 8, ... 1 non-zero coefficients, which are also set to 1. At last, we first generated in the first case some high correlations among covariates according to a multivariate Gaussian distribution with covariance matrix V so that $V_i^j = 0.9(-1)^{|i-j|}$. Then, we also generated some high correlations among the error terms according to the same multivariate Gaussian distribution with covariance V. Table 3 shows performances of the proposed algorithms on this dataset.

Assumption $\mathbf{H_{SNR}}$ may not be fulfilled here, but we are interested in the robustness of the studied Boost-Boost algorithms in such a scenario. Results indicate that the Boost-Boost $\mathcal{D}$-Correlation sum algorithm and REM perform better overall. Their overall recall is quite poor (about $71.26 - 75.60\%$ of FN elements for REM and $74.20 - 83.36\%$ for the Boosting

algorithm). REM includes more irrelevant regressors in the model (with a rate of $4.38 - 7.07\%$ of FP elements for Boosting algorithms and $5.34 - 14.33\%$ for REM), probably because of the very high correlation levels between predictors or because of the intricate correlated noise we artifically added to the data. The latter seems indeed to be an even more challenging obstacle here. We recall here that in these 2 scenarii, a 1% in FP rate implies a difference of just over 600 falsely predicted edges. The algorithms we proposed were designed to deal partly with the correlation between responses when it's not too high and when the noise is not too high neither. It seems here that the correlated noise is a more difficult situation to tackle, perhaps only because of the choice we made to simulate it. The overall low recalls (or high FN rates) can be explained by the highly unbalanced design between columns as well. Moreover, Boosting algorithms and REM identify much richer models than BootLasso and RF do, quite beyond the $\frac{10}{2{,}455} \approx 0.41\%$ of TP in the first column whose coefficients dominate, even if their precision is not as good. On the opposite, RForest and BootLasso do tend to produce reliable coefficients (at least in identifying non-zero values) but at the price of a very poor coverage.

MSE are also quite high in this scenarii, mainly because the coefficient matrix includes many coefficients with values set to 500. Hence, the effect of imprecisely estimated coefficients can have quite a large impact on MSE values, even it is actually a true coefficient. $\mathcal{D}$- Correlation sum, L^2-norm and REM again achieve the best MSE among tested approaches, with REM taking the advantage again because of richer, less precise models.

	Correlated covariates			Correlated noises		
	FP (60,045)	FN (2,455)	MSE	FP (60,045)	FN (2,455)	MSE
$\mathcal{D}$-Corr	4.39	74.20	0.63	7.07	83.14	0.60
L^2 norm	4.38	74.50	0.63	6.94	83.38	0.61
BootLasso	0.81	77.21	0.82	0.76	87.64	1.21
RForests*	2.27	78.63	0.84	0.79	97.15	0.93
REM	5.35	71.26	0.62	14.33	75.60	0.47

Table 3: Second dataset: Percentage of false positive FP parameters (number of coefficients not to be predicted between brackets) and false negative FN parameters (number of coefficients to be predicted between brackets) and MSE for the Boosting algorithms, with a shrinkage factor $\gamma = 0.2$, compared with the BootLasso, RForests and REM; the sample size n is set to 100. We also indicate the number of edges to retrieve: $2,455$ and the number of potential FP: $250 * 250 - 2455 = 60,045$. (*: for 5 simulated replicate data sets only as the running time for RForest was 4 days per network).

Third dataset The last dataset mimics activation and inhibition relationships that exist between genes in the gene regulatory network of a living organism and is very close to a real data situation. This dataset, for which $p = 100$, is exactly the one that was provided by the DREAM Project [DRE] in their Challenge 2 on the "In Silico Network Challenge" (more precisely, the *InSilico_Size100_Multifactorial*). First, a directed network structure is chosen. Its features can be regarded as features of a biological network, *e.g.*, in terms of degree distribution. Coupled ordinary differential equations (ODEs) then drive the quantitative impact of gene expression on each other, the expression of a gene roughly representing its activity in the system. For example, if gene 1 is linked to gene 2 with a positive effect going from 1 to 2, then increasing the expression of gene 1 (as operator *do*, see [Pea09]) will increase the expression of gene 2. However, increasing the expression of gene 2 does not have a direct effect on gene 1. Lastly, the system of

ODEs is solved numerically to obtain steady states of the expression of the genes after technical and biological noises are created. We denote as A the $n \times p$ expression matrix of p genes for n individuals. This simulation process is highly non-linear compared to the first two scenarios described above.

The goal was to automatically retrieve network structure encoded in matrix A from data only. Samples were obtained by multifactorial perturbations of the network using GeneNetWeaver [SMF11] for simulations. A multifactorial perturbation is the simultaneous effect of numerous minor random perturbations in the network. It therefore measures a deviation from the equilibrium of the system. This could be seen as changes in the network due to very small environmental changes or genetic diversity in the population. Additional details and a discussion on the biological plausability (network structure, the use of chemical Langevin differential equations, system and experimental noise) of such datasets can be found in [MSMF09].

	FP (9,695)	FN (205)	MSE
$\mathcal{D}$-Corr	21.37	47.75	0.45
L^2 norm	18.98	50.20	0.50
BootLasso	1.40	77.93	0.32
RForests	7.68	68.98	0.20
REM	7.05	78.53	0.01

Table 4: Third dataset: Percentage of false positive FP parameters (number of coefficents not to be predicted between brackets) and false negative FN parameters (number of coefficients to be predicted between brackets) and MSE for the Boosting algorithms, with a shrinkage factor $\gamma = 0.2$, compared to the BootLasso, RForests and REM.

The results of tested methods on this last dataset are presented in Table 4. In this scenario, our two multivariate (we recall that $m = p = 100$) $\mathbb{L}_2$-Boosting algorithms both suffer from higher MSE. It also exhibits higher FP rates than other competing methods: $\approx 20\%$ vs. 1.4, 7.7 and 7.1% for BootLasso, RF and REM, respectively. Many FP coefficients may imply an increase in MSE, whereas the three other tested methods focus on fewer correct edges.

What can first be considered as a pitfall can be turned into a strength: recall can be close to (for L^2 norm) or even higher than ($\mathcal{D}$-Correlation sum) 50%, whereas other approaches reach 31% at best (RF). In other words, the $\mathcal{D}$-Correlation sum can retrieve more than half of the 205 edges to be predicted, at the price of producing more FP predictions, considered as noise from a model prediction perspective in the delivered list. RF is on average only able to grab 84 out of the 205 correct edges, but the prediction list is cleaner in a sense. A specifically designed variant of the RF approach that we tested was deemed the best performer for this challenge by the DREAM4 organisers [HTIWG10]. Our algorithm would have been ranked 2$^{\text{nd}}$.

For the sake of completeness, we computed the smallest and the largest eigenvalues of the restricted matrix ${}^tD_{\mathcal{S}}D_{\mathcal{S}}$, that are involved in the key Assumptions $\mathbf{H_{RE^-}}$ and $\mathbf{H_{RE^+}}$. We also provided the measured value of $\rho := \max_{j \notin \mathcal{S}} \|D_{\mathcal{S}}^{+} g_j\|_1$ of Assumption $\mathbf{H_S}$ in Table 5 for the three datasets, which quantifies the coherence of the dictionary: favourable situations correspond to small values of ρ, ideally lower than 1.

Regarding the first dataset, we obtain a larger value than 0 for λ_{min} and a moderate value of λ_{max}. This implies a reasonable value of $\lambda_{max}/\lambda_{min}$. This situation is thus acceptable according to the bound given by Equations (33) and (35) (see Appendix, Lemma B.2). Concerning Assumption $\mathbf{H_S}$, for each range of parameters on the first dataset, ρ is not very far from 1, which

	λ_{min}	λ_{max}	$\lambda_{max}/\lambda_{min}$	ρ
First data set $(p, m, S) = (250, 1, 5)$	75.31	130.43	1.73	0.82
First data set $(p, m, S) = (250, 1, 10)$	59.03	143.78	2.43	1.52
First data set $(p, m, S) = (1000, 1, 20)$	37.66	190.71	5.06	2.80
First data set $(p, m, S) = (250, 5, 50)$	49.14	157.20	3.20	1.71
First data set $(p, m, S) = (250, 250, 50)$	52.78	151.76	2.88	1.10
Second data set Correlated covariates	3.95	921.39	233.44	5.47
Correlated noises	41.12	181.49	4.41	1.88
Third data set	19.29	233.83	12.12	1.57

Table 5: Smallest and largest eigenvalue of the restricted matrix, ratio of these eigenvalues and computation of $\rho := \max_{j \notin \mathcal{S}} \|D_{\mathcal{S}}^{+} g_j\|_1$ for the three datasets.

explains the good numerical results. We have to particularly emphasize the first simulation study where $(p, m, S) = (250, 1, 5)$. With a coherence value ρ lower than 1, the WGA reaches to recover the true support of A. $\lambda_{max}/\lambda_{min}$ and ρ values for the second dataset support our numerical analysis (see Table 3) that shows that this is a very difficult dataset. This situation is clearly less favourable for the sparse estimation provided by our Boosting procedures than for the first data set. This is perhaps less visible for the second simulated setting, where additional noise was is correlated. Clearly in this latter case, hypothesis $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim-3}}$ is violated because the noise coordinates are not i.i.d. anymore. We however have no numerical indicator to quantify this.

For the last dataset, we can observe that $\mathbf{H_S}$ yields a moderate value of ρ but that the ratio of the restricted eigenvalues is quite large (compared to those obtained in the first dataset) and it is difficult to recover the support of the true network.

Taken together, this numerically shows that both $\mathbf{H_S}$ and $\mathbf{H_{RE^-}}$, $\mathbf{H_{RE^+}}$ are important to obtain good reconstruction properties. These assumptions then seem complementary and not redundant. However, the practical use of the proposed algorithms advocates a certain tolerance of the method towards divergence from the hypotheses that condition our theoretical results.

5 Concluding remarks

We studied WGA and established a support recovery result for solving linear regression in high-dimensional settings. We then proposed two statistically funded $\mathbb{L}_2$-Boosting algorithms derived thereupon in a multivariate framework. The algorithms were developped to sequentially estimate unknown parameters in a high-dimensional regression framework: significant possibly correlated regressor functions from a dictionnary need be identified, relative coefficients need be estimated and noise can disturb the observations. Consistency of two variants of the algorithms was proved in Theorem 3.2 for the L^2 norm variant and in Theorem 3.3 for the $\mathcal{D}$-Correlation sum variant. An important Support Recovery result (Theorem 3.4) under mild assumption on the sparsity of the regression function and on the restricted isometry of the X matrix then generalises the univariate result to the multi-task framework. Using the MSE of the model, we derived a simple

yet effective stopping criterion for our algorithms.

We then illustrated the proposed algorithms in a variety of simulated datasets in order to determine the ability of the proposed method to compete with state-of-the-art methods when the data is high-dimensional, noisy and the active elements can be unbalanced. Even if the algorithms we propose are not superior in all settings, we observed, for example, that they are very competitive in situations such as those of the DREAM4 In Silico Multifactorial Network Challenge. Without fine parameter tuning and with a very small computing time, our generic method would have ranked 2$^{\text{nd}}$ in this challenge. Moreover, it has the ability to quickly produce a rich prediction list of edges at an acceptable quality level, which might reveal novel regulatory mechanisms on real biological datasets.

Aknowledgements: Thanks are due to the anonymous reviewer, to the Associate Editor and to our colleagues whose suggestions greatly improved this manuscript.

A Appendix: Stability results for Boosting algorithms

A.1 Concentration inequalities

We begin by recalling some technical results. Lemma A.1, given in [Büh06], provides a uniform law of large numbers, in order to compare inner products $\langle , \rangle_{(n)}$ and $\langle , \rangle$. It is useful to prove the theorems of Section 2.2.2 and 2.3, and does not rely on boosting arguments.

Lemma A.1 *Assume that Hypotheses* $\mathbf{H_{dim}}$ *are fulfilled on dictionary* $\mathcal{D}$*,* f *and* ε*, with* $0 < \xi < 1$ *as given in* $\mathbf{H_{dim-2}}$*, 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) $\sup_{1\leq i\leq p_n} |\langle g_i, \varepsilon\rangle_{(n)}| = \zeta_{n,2} = \mathcal{O}_P(n^{-\xi/2})$,

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

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 [Büh06]) also holds.

Lemma A.2 *Under Hypotheses* $\mathbf{H_{dim}}$*, a constant* $0 < C < +\infty$ *exists, independent of* n *and* k*, so 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 [Büh06], 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). It is nevertheless possible to generalise its application field using Lemma A.1. First, assume that $k = 0$. The desired inequality follows directly from point (iii) of Lemma 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, on the basis of the recursive relationships of Equations (7) and (8), we obtain:

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

Expanding Equation (8) 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 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,

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

Finally, using (i) and (ii) from Lemma 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

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

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

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 A.2 holds when replacing the theoretical residual $\hat{R}_k(f)$ with the observed residual $Y - \hat{G}_k(f)$, thanks to Lemma A.1 (ii). Hence, on the set Ω_n, by definition of φ_k:

$$\begin{aligned}|\langle Y - \hat{G}_{k-1}(f), \varphi_k\rangle_{(n)}| &= \sup_{1\leq j\leq p_n}|\langle Y - \hat{G}_{k-1}(f), g_j\rangle_{(n)}| \\ &= \sup_{1\leq j\leq p_n}\left\{|\langle\tilde{R}_{k-1}(f), g_j\rangle| - C\left(\frac{5}{2}\right)^{k-1}\zeta_n\right\}. \qquad (15)\end{aligned}$$

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

$$\begin{aligned}|\langle\tilde{R}_{k-1}(f), \varphi_k\rangle| &\geq |\langle Y - \hat{G}_{k-1}(f), \varphi_k\rangle_{(n)}| - C\left(\frac{5}{2}\right)^{k-1}\zeta_n \\ &\geq \sup_{1\leq j\leq p_n}|\langle\tilde{R}_{k-1}(f), g_j\rangle| - 2C\left(\frac{5}{2}\right)^{k-1}\zeta_n. \qquad (16)\end{aligned}$$

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 the following equality from Equation (16):

$$|\langle\tilde{R}_{k-1}(f), \varphi_k\rangle| \geq \frac{1}{2}\sup_{1\leq j\leq p_n}|\langle\tilde{R}_{k-1}(f), g_j\rangle|. \qquad (17)$$

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

$$\|\tilde{R}_k(f)\| \leq C_B \left(1 + \frac{1}{4}\gamma(2-\gamma)k\right)^{-\frac{2-\gamma}{2(6-\gamma)}}. \tag{18}$$

Now consider the set $\tilde{\Omega}_n^C = \left\{\omega, \quad \exists\, k \leq k_n \quad \sup_{1\leq j\leq p_n} |\langle \tilde{R}_{k-1}(f), g_j\rangle| \leq 4C\left(\frac{5}{2}\right)^{k-1}\zeta_n\right\}$. Note that:

$$\begin{aligned}
\|\tilde{R}_k(f)\|^2 &= \langle \tilde{R}_k(f), f - \gamma\sum_{j=0}^{k-1}\langle \tilde{R}_j(f), \varphi_j\rangle\varphi_j\rangle \\
&\leq \left(\sum_{j=1}^{p_n}|a_j| + \gamma\sum_{j=0}^{k-1}\left|\langle \tilde{R}_j(f), \varphi_j\rangle\right|\right)\sup_{1\leq j\leq p_n}\left|\langle \tilde{R}_k(f), g_j\rangle\right|.
\end{aligned}$$

Then, since $\|\tilde{R}_k(f)\|$ is non-increasing and by definition of $\tilde{\Omega}_n^C$, we deduce that on $\tilde{\Omega}_n^C$,

$$\|\tilde{R}_k(f)\|^2 \leq 4C\left(\frac{5}{2}\right)^k \zeta_n \left(\sum_{j=1}^{p_n}|a_j| + \gamma k\|f\|\right). \tag{19}$$

Hence, on $(\Omega_n \cap \tilde{\Omega}_n) \cup \tilde{\Omega}_n^C$, using Equations (18) and (19),

$$\|\tilde{R}_k(f)\|^2 \leq C_B^2\left(1 + \frac{1}{4}\gamma(2-\gamma)k\right)^{-\frac{2-\gamma}{6-\gamma}} + 4C\left(\frac{5}{2}\right)^k \zeta_n \left(\sum_{j=1}^{p_n}|a_j| + \gamma k\|f\|\right). \tag{20}$$

To conclude, note 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 (20) holds almost surely for all ω and for a sequence $k_n < (\xi/4\log(3))\log(n)$, which grows sufficiently slowly:

$$\|\tilde{R}_{k_n}(f)\| = o_P(1). \tag{21}$$

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

$$\begin{aligned}
A_k &= \|\hat{R}_{k-1}(f) - \gamma\langle \hat{R}_{k-1}(f), \varphi_k\rangle_{(n)}\varphi_k - \gamma\langle \varepsilon, \varphi_k\rangle_{(n)}\varphi_k - \left(\tilde{R}_{k-1}(f) - \gamma\langle \tilde{R}_{k-1}(f), \varphi_k\rangle\varphi_k\right)\| \\
&\leq A_{k-1} + \gamma|\langle Y - \hat{G}_{k-1}(f), \varphi_k\rangle_{(n)} - \langle \tilde{R}_{k-1}(f), \varphi_k\rangle|. \qquad (22)
\end{aligned}$$

Under Hypothesis $\mathbf{H_{dim}}$, we deduce the following inequality on Ω_n from Equation (22):

$$A_k \leq A_{k-1} + \gamma\left(C\left(\frac{5}{2}\right)^{k-1} + 1\right)\zeta_n. \tag{23}$$

Using $A_0 = 0$, we deduce recursively from Equation (23) that, on Ω_n, since $k := k_n$ grows sufficiently slowly:

$$A_{k_n} \xrightarrow[n\to+\infty]{\mathbb{P}} 0. \tag{24}$$

Finally, observe that $\|\hat{R}_{k_n}(f)\| \leq \|\tilde{R}_{k_n}(f)\| + A_{k_n}$. The conclusion holds using Equation (21) and (24).

A.3 Proof of support recovery

We now detail the proof of Theorem 2.3, which represents the exact recovery of the support with high probability. It should be recalled that we denote as S (respectively, $\mathcal{S}$) the sparsity (respectively, 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 (5). 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 as that of the proof of Lemma A.2, we have:

$$\forall g_j \in \mathcal{D}, \qquad |\langle Y - \hat{G}_{k-1}(f), g_j\rangle_{(n)} - \langle \hat{R}_{k-1}(f), g_j\rangle| \leq C\zeta_n \left(\frac{5}{2}\right)^{k-1}. \tag{25}$$

On the one hand, we deduce from Equation (25) below that:

$$\forall j \in \mathcal{S}_{k-1}, \quad |\langle Y - \hat{G}_{k-1}(f), g_j\rangle_{(n)}| \geq |\langle \hat{R}_{k-1}(f), g_j\rangle| - C\zeta_n \left(\frac{5}{2}\right)^{k-1}. \tag{26}$$

On the other hand, for $j \notin \mathcal{S}_{k-1}$, we also have:

$$|\langle Y - \hat{G}_{k-1}(f), g_j\rangle_{(n)}| \leq |\langle \hat{R}_{k-1}(f), g_j\rangle| + C\zeta_n \left(\frac{5}{2}\right)^{k-1}. \tag{27}$$

Now denote $M_k := \max_{j\in\mathcal{S}_{k-1}} |\langle \hat{R}_{k-1}(f), g_j\rangle|$ and $M_k^C := \max_{j\notin\mathcal{S}_{k-1}} |\langle \hat{R}_{k-1}(f), g_j\rangle|$. We recall that element j is selected at step k following Equation (5). Hence, we deduce from Equations (26) and (27) that $j \in \mathcal{S}_k$ is in $\mathcal{S}_{k-1}$ if the following inequality is satisfied:

$$M_k > M_k^C + 2C\zeta_n \left(\frac{5}{2}\right)^{k-1}. \tag{28}$$

The next step of the proof consists in comparing the two quantities M_k and M_k^C. Note that M and M^C can be rewritten as $\|{}^t D_{\mathcal{S}_{k-1}} \hat{R}_{k-1}(f)\|_\infty$ and $\|{}^t D_{\mathcal{S}_{k-1}^C} \hat{R}_{k-1}(f)\|_\infty$. Following the arguments of [Tro04], we have:

$$\begin{aligned} \frac{M_k}{M_k^C} &= \frac{\|{}^t D_{\mathcal{S}_{k-1}^C}\, {}^t D^+_{\mathcal{S}_{k-1}}\, {}^t D_{\mathcal{S}_{k-1}} \hat{R}_{k-1}(f)\|_\infty}{\|{}^t D_{\mathcal{S}_{k-1}} \hat{R}_{k-1}(f)\|_\infty} \\ &\leq \|{}^t D_{\mathcal{S}_{k-1}^C}\, {}^t D^+_{\mathcal{S}_{k-1}}\|_{\infty,\infty}, \end{aligned}$$

where $\|.\|_{q,q}$ is the subordonate norm of the space $(\mathbb{R}^q, \|.\|_q)$. In particular, the norm $\|.\|_{\infty,\infty}$ equals the maximum absolute row of its arguments, and we also have:

$$\frac{M_k}{M_k^C} \leq \|D^+_{\mathcal{S}_{k-1}} D_{\mathcal{S}_{k-1}^C}\|_{1,1} = \max_{j\notin\mathcal{S}_{k-1}} \|D^+_{\mathcal{S}_{k-1}} g_j\|_1.$$

Using Assumption $\mathbf{H_S}$ and the recursive assumption $\mathcal{S}_{k-1} \subset \mathcal{S}$, we obtain that $M_k > M_k^C$.

The end of the proof of (i) follows with Equation (28) for $k := k_n$ given by Theorem 2.2, which implies that $\zeta_n(5/2)^k \to 0$. □

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.

Assume that one element j_0 of $\mathcal{S}$ is never selected. Then, if we denote $\theta^k = (\theta_j^k)_{1\leq j\leq p_n}$ as the decomposition of $\hat{R}_{k-1}(f)$ on $\mathcal{D}$, we obtain:

$$\|\theta^k\|^2 = \sum_j (\theta_j^k)^2 \geq (\theta_{j_0}^k)^2 = a_{j_0}^2, \tag{29}$$

where a_{j_0} is the true coefficient of f associated with the element g_{j_0}.

Moreover, note that

$$\|\hat{R}_{k-1}(f)\|^2 = \|D\theta^k\|^2 \geq \lambda_{min}\|\theta^k\|^2, \tag{30}$$

with $\lambda_{min} := \inf_{\beta,\mathrm{Supp}(\beta)\subset\mathcal{S}} \|D\beta\|^2/\|\beta\|^2 > 0$ by Assumption $\mathbf{H_{RE^-}}$.

Equation (30) deserves special attention since $(\|\hat{R}_{k-1}(f)\|)_k$ decreases with k. More precisely, Equations (20) and (23) of Section A.2 provide the following bound for $\|\hat{R}_{k-1}(f)\|$:

$$\|\hat{R}_{k-1}(f)\|^2 \leq (C\log(n))^{-\alpha},$$

where $\alpha := \frac{2-\gamma}{6-\gamma}$.

The sought contradiction is obtained using Assumption $\mathbf{H_{SNR}}$ in Equation (29) as soon as

$$\lambda_{min}\log(n)^{-2\kappa} \geq (C\log(n))^{-\alpha},$$

i.e., when $\kappa < \kappa^* := (2-\gamma)/2(6-\gamma)$. This ends the proof of the support consistency.

□

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

B.1 Proof of Theorem 3.1

We break down the proof of Theorem 3.1 into several steps here. It should be recalled that $\mathcal{D} = \{(g_j), 1 \leq j \leq p\}$ is a dictionary that spans H. We set any $f = (f^1, \ldots, f^m) \in H_m$ so that $f^i \in \mathcal{A}(\mathcal{D}, B)$.

The first key remark is that if we denote $s_i(k)$ as 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:

$$\forall k \geq 1, \quad \|R_{k-1}(f^i)\| \leq C_B(1 + \nu^2\gamma(2-\gamma)s_i(k-1))^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}. \tag{31}$$

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 (10) or (11). For the Boost-Boost Residual L^2 norm algorithm, this step is not pivotal since, using Equation (10):

$$\sup_{1\leq i\leq m} \|R_k(f^i)\| \leq \mu^{-1}\|R_k(f^{i_k})\|. \tag{32}$$

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

Lemma B.1 *Suppose that Assumptions $\mathbf{H_{RE^-}}$ and $\mathbf{H_{RE^+}}$ hold. Then, for any k:*

$$\sup_{1\leq i\leq m} \|R_{k-1}(f^i)\|^2 \leq \mu^{-1}\|R_{k-1}(f^{i_k})\|^2 \left(\frac{\lambda_{max}}{\lambda_{min}}\right)^3,$$

where λ_{min} and λ_{max} (given by Assumptions $\mathbf{H_{RE^-}}$ and $\mathbf{H_{RE^+}}$) are the smallest and the largest eigenvalues ${}^tD_{\mathcal{S}}D_{\mathcal{S}}$.

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 $\mathcal{S}_k^i$. Note that, along the iterations of the Boost-Boost algorithm, an incorrect element of the dictionary cannot be selected using Equation (12) (see Theorem 3.4 for some supplementary details). We observe then that Assumptions $\mathbf{H_{RE^-}}$ and $\mathbf{H_{RE^+}}$ imply that at each step, each approximation is at most S-sparse. We present an elementary lemma that will be very useful until the end of the proof.

Lemma B.2 *Let $\mathcal{D} = (g_1, ..., g_p)$ be a dictionary on H. Denote D as the matrix whose columns are the elements of $\mathcal{D}$, and for any $\mathcal{S} \subset [\![1, p]\!]$, $D_{\mathcal{S}}$ the matrix restricted to the elements of $\mathcal{D}$ that are in $\mathcal{S}$. Then, if we denote λ_{min} and λ_{max} as the smallest and the largest eigenvalues of the restricted matrix ${}^tD_{\mathcal{S}}D_{\mathcal{S}}$, the two propositions hold.*

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

$$\lambda_{min}\left(\sum_{j=1}^{p}|a_j|^2\right) \leq \left\|\sum_{j=1}^{p} a_j g_j\right\|^2 \leq \lambda_{max}\left(\sum_{j=1}^{p}|a_j|^2\right).$$

(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:

$$\lambda_{min}^2\left(\sum_{j=1}^{p}|a_j|^2\right)^{1/2} \leq \left(\sum_{j=1}^{p}|\langle f, g_j\rangle|^2\right)^{1/2} \leq \lambda_{max}^2\left(\sum_{j=1}^{p}|a_j|^2\right)^{1/2}.$$

Now, let $i \neq i_k$. By Lemma B.2 (right hand side -r.h.s.- of (ii) and left hand side -l.h.s.- of (i)) combined with Assumption $\mathbf{H_{RE^-}}$, we have:

$$\sum_{j=1}^{p}|\langle R_{k-1}(f^{i_k}), g_j\rangle|^2 \leq \|R_{k-1}(f^{i_k})\|^2\frac{\lambda_{max}^2}{\lambda_{min}}. \tag{33}$$

Moreover Lemma B.2 again (l.h.s. of (ii) and r.h.s. of (i)) and Assumption $\mathbf{H_{RE^+}}$ show that:

$$\forall 1 \leq i \leq m, \quad \sum_{j=1}^{p}|\langle R_{k-1}(f^{i}), g_j\rangle|^2 \geq \|R_{k-1}(f^{i})\|^2\frac{\lambda_{min}^2}{\lambda_{max}}. \tag{34}$$

By definition of i_k (see Equation (11) in the Boost-Boost algorithm), we deduce that:

$$\begin{aligned}\forall i \in [\![1, m]\!], \qquad \sum_{j=1}^{p}|\langle R_{k-1}(f^{i_k}), g_j\rangle|^2 &\geq \mu\sum_{j=1}^{p}|\langle R_{k-1}(f^{i}), g_j\rangle|^2 \\ &\geq \mu\|R_{k-1}(f^{i})\|^2\frac{\lambda_{min}^2}{\lambda_{max}}. \end{aligned}\tag{35}$$

The conclusion follows by using Equations (33) and (35). □

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. A coordinate $i^* \in \{1 \dots m\}$, that is selected at least K times by Equation (10) or (11) exists, hence $s_{i^*}(k) \geq K$. We also denote k^* as the last step that selects i^* before step k. Since $\left(\|R_k(f^i)\|\right)_k$ is a non-increasing sequence along the iterations of the algorithm, Equation (31) leads to:

$$\|R_{k-1}(f^{i^*})\| \leq \|R_{k^*-1}(f^{i^*})\| \leq C_B(1 + \nu^2\gamma(2-\gamma)(K-1))^{-\frac{\nu(2-\gamma)}{2(2+\nu(2-\gamma))}}. \tag{36}$$

The conclusion holds noting that $\frac{k}{m} - 1 \leq K \leq \frac{k}{m}$ and $\nu < 1$, and using our bounds (32) for the Boost-Boost Residual L^2 norm algorithm, or Lemma B.1 for the Boost-Boost $\mathcal{D}$-Correlation sum algorithm.

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 other multi-task results. The proof proceeds in the same way as in Section A.3. Our focus is on the choice of the regressor to add in the model, regardless of the column chosen to be regressed in the previous step. Therefore, in order to simplify notations, index i may be omitted and we can do exactly the same computations.

B.3 Proof of Theorems 3.2 and 3.3

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

$$\begin{aligned} &\tilde{R}_k(f^i) = \tilde{R}_{k-1}(f^i) \ \text{ if } \ i \neq i_k, \\ &\tilde{R}_k(f^{i_k}) = \tilde{R}_{k-1}(f^{i_k}) - \gamma \langle \tilde{R}_{k-1}(f^{i_k}), \varphi_k \rangle \varphi_k, \end{aligned} \tag{37}$$

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

As previously explained, we aim at applying Theorem 3.1 to the phantom residuals. This will be possible if we can show an analogue of Equations (10) (for the Residual L^2 norm) or (11) (for the $\mathcal{D}$-Correlation sum) and (12). Note that on the basis of Theorem 3.4, the 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 (13).

We begin the proof by recalling Lemma A.1. In the multi-task case, this lemma can be easily extended as follows:

Lemma B.3 *Assume that Hypotheses* $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim}}$ *are fulfilled on dictionary* $\mathcal{D}$*,* f *and* ε*, with* $0 < \xi < 1$ *as given in* $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim-2}}$*, 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 B.3 are the same as (i), (ii) and (iii) of Lemma A.1. The fourth point is something new. However, since its proof does not call for typical boosting arguments, we do not 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 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 $k' \leq k$ so that i_k is not selected between step k' and k (see Equation (13)), we can easily extend Lemma A.2 to each coordinate of f:

$$\sup_{1 \leq i \leq m_n} \sup_{1 \leq j \leq p_n} |\langle \hat{R}_k(f^i), g_j \rangle_{(n)} - \langle \tilde{R}_k(f^i), g_j \rangle| \leq C \left(\frac{5}{2}\right)^k \zeta_n. \tag{38}$$

Using this extension of Lemma A.1, the same calculations detailed in Section A.2 can be done. Hence, considering the i_k-th coordinate of f chosen by Equations (10) or (11), on the set Ω_n, inequality (17) also holds:

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

Now consider the Boost-Boost Residual L^2 norm algorithm. To obtain an analogue of (10), we need the following lemma, which compares the norms of both residuals:

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

$$\sup_{1\leq i\leq m_n} |\|\hat{R}_{k-1}(f^i)\|_{(n)}^2 - \|\tilde{R}_{k-1}(f^i)\|^2| \leq C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n.$$

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

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

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

Using Equation (38), we can provide two upper bounds for (I) and (II):

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

Denoting $M := \max_{1\leq j\leq S}\{|\theta_{i,j}^k|, |\tilde{\theta}_{i,j}^k|\}$, 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 (B.3), we have:

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

and the conclusion follows using our last bounds. □

Since Lemma 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 (10). However, we can compare the norm of the theoretical and observed residuals:

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

Note that, using Lemma 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 B.4. Hence, we have for all i:

$$\|\hat{R}_{k-1}(f^i)\|_{(n)}^2 + \|\varepsilon^i\|_{(n)}^2 - 2MS\zeta_n \leq \|Y^i - \hat{G}_{k-1}(f^i)\|_{(n)}^2 \leq \|\hat{R}_{k-1}(f^i)\|_{(n)}^2 + \|\varepsilon^i\|_{(n)}^2 + 2MS\zeta_n. \quad (39)$$

It should be recalled that $\mathbb{E}(|\varepsilon^i|^2)$ does not depend on i from Assumption $\mathbf{H}^{\mathbf{Mult}}_{\mathbf{dim-3}}$, and is denoted by σ^2. An application of Lemma B.3 (iv) to Equation (39) then yields:

$$\|\hat{R}_{k-1}(f^i)\|^2_{(n)}+\sigma^2-(1+2MS)\zeta_n \leq \|Y^i-\hat{G}_{k-1}(f^i)\|^2_{(n)} \leq \|\hat{R}_{k-1}(f^i)\|^2_{(n)}+\sigma^2+(1+2MS)\zeta_n. \quad (40)$$

Hence, on Ω_n, by definition of i_k, Equation (40) and Lemma B.4, we can write:

$$\begin{aligned}
\|Y^{i_k} - \hat{G}_{k-1}(f^{i_k})\|^2_{(n)} &\geq \sup_{1\leq i\leq m_n} \|Y^i - \hat{G}_{k-1}(f^i)\|^2_{(n)} \\
&\geq \sup_{1\leq i\leq m_n} \left\{\|\hat{R}_{k-1}(f^i)\|^2_{(n)} + \sigma^2\right\} - (1+2MS)\zeta_n \\
&\geq \sup_{1\leq i\leq m_n} \left\{\|\tilde{R}_{k-1}(f^i)\|^2 + \sigma^2\right\} - C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n \\
&\quad -(1+2MS)\zeta_n. \qquad (41)
\end{aligned}$$

Using the same calculus on the set Ω_n once again:

$$\begin{aligned}
\|\tilde{R}_{k-1}(f^{i_k})\|^2 &\geq \|\hat{R}_{k-1}(f^{i_k})\|^2_{(n)} - C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n \\
&\geq \|Y^{i_k} - \hat{G}_{k-1}(f^{i_k})\|^2_{(n)} - \sigma^2 - (1+2MS)\zeta_n - C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n \\
&\geq \sup_{1\leq i\leq m_n} \left\{\|\tilde{R}_{k-1}(f^i)\|^2 + \sigma^2\right\} - \sigma^2 - 2(1+2MS)\zeta_n \\
&\quad -2C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n, \quad \text{by Equation (41).} \qquad (42)
\end{aligned}$$

We then obtain from Equation (42) that:

$$\|\tilde{R}_{k-1}(f^{i_k})\|^2 \geq \sup_{1\leq i\leq m_n} \|\tilde{R}_{k-1}(f^i)\|^2 - 2(1+2MS)\zeta_n - 2C\left(2\left(\frac{5}{2}\right)^{k-1} + S\right) S\zeta_n. \qquad (43)$$

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 (43) the following inequality on set $\Omega_n \cap \check{\Omega}^1_n$:

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

Finally, consider the Boost-Boost $\mathcal{D}$-Correlation sum algorithm. To obtain an analogue of Equation (11), the following lemma is needed:

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

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

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

$$|\langle \hat{R}_k(f^i), g_j\rangle_{(n)}^2 - \langle \tilde{R}_k(f^i), g_j\rangle^2| = |\langle \hat{R}_k(f^i), g_j\rangle_{(n)} - \langle \tilde{R}_k(f^i), g_j\rangle||\langle \hat{R}_k(f^i), g_j\rangle_{(n)} + \langle \tilde{R}_k(f^i), g_j\rangle|, \tag{44}$$

where $|\langle \hat{R}_k(f^i), g_j\rangle_{(n)} - \langle \tilde{R}_k(f^i), g_j\rangle| \leq C\left(\frac{5}{2}\right)^k \zeta_n$ by Equation (38).

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

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

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

On Ω_n, we therefore have, as a suitable constant, $C > 0$:

$$M_k^i \leq M_{k-1}^i\left(1 + \frac{3}{2}\gamma\right) + C \ldots \leq \left(1 + \frac{3}{2}\gamma\right)^k\left(\sup_{n\in\mathbb{N}}\sum_{j=1}^{p_n}|a_{i,j}| + \frac{3}{2}\right) + C. \tag{45}$$

Using Equation (8), $\|\tilde{R}_k(f^i)\|$ is non-increasing and thus $\|\tilde{R}_k(f^i)\| \leq \|f^i\|$. The Cauchy-Schwarz inequality allows us to write that:

$$\left|\langle \tilde{R}_k(f^i), g_j\rangle\right| \leq \|\tilde{R}_k(f^i)\| \leq \|f^i\|. \tag{46}$$

The conclusion therefore holds using Equations (45) and (46) in Equation (44) for a large enough constant C. □

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

$$\begin{aligned}
\sum_{j=1}^{p_n}|\langle Y^{i_k} - \hat{G}_{k-1}(f^{i_k}), g_j\rangle_{(n)}|^2 &\geq \sup_{1\leq i\leq m_n}\sum_{j=1}^{p_n}|\langle Y^i - \hat{G}_{k-1}(f^i), g_j\rangle_{(n)}|^2 \\
&\geq \sup_{1\leq i\leq m_n}\sum_{j=1}^{p_n}\left(|\langle \tilde{R}_{k-1}(f^i), g_j\rangle_{(n)}|^2 - C\left(\frac{5}{2}\right)^{2(k-1)}\zeta_n\right) \\
&\geq \sup_{1\leq i\leq m_n}\sum_{j=1}^{p_n}|\langle \tilde{R}_{k-1}(f^i), g_j\rangle_{(n)}|^2 - Cp_n\left(\frac{5}{2}\right)^{2(k-1)}\zeta_n.
\end{aligned} \tag{47}$$

Using Lemma B.5 again on Ω_n:

$$\begin{aligned}
\sum_{j=1}^{p_n}|\langle \tilde{R}_{k-1}(f^{i_k}), g_j\rangle|^2 &\geq \sum_{j=1}^{p_n}|\langle Y^{i_k} - \hat{G}_{k-1}(f^{i_k}), g_j\rangle_{(n)}|^2 - Cp_n\left(\frac{5}{2}\right)^{2(k-1)}\zeta_n \\
&\geq \sup_{1\leq i\leq m_n}\sum_{j=1}^{p_n}|\langle \tilde{R}_{k-1}(f^i), g_j\rangle|^2 - 2Cp_n\left(\frac{5}{2}\right)^{2(k-1)}\zeta_n \quad \text{by Equation (47).}
\end{aligned} \tag{48}$$

Let $\check{\Omega}_n^2 = \left\{\omega, \quad \forall k \leq k_n \quad \sup_{1\leq i\leq m_n} \sum_{j=1}^{p_n} |\langle \tilde{R}_{k-1}(f^i), g_j\rangle|^2 > 4Cp_n \left(\frac{5}{2}\right)^{2(k-1)} \zeta_n \right\}$. On the basis of Equation (48), we can deduce 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.

Let us now consider the set $\left(\check{\Omega}_n^2\right)^C$. Using Equation (34), we obtain

$$\|\tilde{R}_k(f^i)\|^2 \leq \frac{\lambda_{max}}{\lambda_{min}^2} \sum_{j=1}^{p_n} |\langle \tilde{R}_k(f^i), g_j\rangle|^2 \leq 4 \frac{\lambda_{max}}{\lambda_{min}^2} Cp_n \left(\frac{5}{2}\right)^{2k} \zeta_n.$$

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 A.2 by noting 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 that grows sufficiently slowly: for the Boost-Boost Residual L^2 norm algorithm, k_n is allowed 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

[ADH09] S. Anjum, A. Doucet, and C.C. Holmes. A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics*, 25(22):2929–2936, 2009.

[AM02] C. Ambroise and G.J. McLachlan. Selection bias in gene extraction on the basis of mi- croarray gene-expression data. *Proceedings of the National Academy of Science*, 99:6562–6566, 2002.

[Bac08] F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the Twenty-fifth International Conference on Machine Learning*, pages 33–40, Helsinki, Finland, 2008. ACM.

[BCT11] J. D. Blanchard, C. Cartis, and J. Tanner. Compressed sensing: how sharp is the restricted isometry property? *SIAM Review*, 53(1):105–125, 2011.

[Bre01] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[Büh06] P. Bühlmann. Boosting for high-dimensional linear models. *Annals of Statistics*, 34(2):559–583, 2006.

[BY03] P. Bühlmann and B. Yu. Boosting with the L2-loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

[BY10] P. Bühlmann and B. Yu. Boosting. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, pages 69–74, 2010.

[CJ11] T. Cai and T. Jiang. Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Annals of Statistics*, 39(3):1496–1525, 2011.

[CT05] E. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[CT07] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[CW11] T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Transactions on Information Theory*, 57(7):4680–4688, 2011.

[DRE] Dream project. Organizers: Columbia university and IBM. Available: http//wiki.c2b2.columbia.edu/dream/index.php/The_DREAM_Project.

[ER11] Y. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Transactions on Information Theory*, 56:505–519, 2011.

[FHT00] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression. A statistical view of boosting (with discussion). *Annals of Statistics*, 28(2):337–407, 2000.

[Gad08] S. Gadat. Jump diffusion over feature space for object recognition. *SIAM J. Control Optim.*, 47(2):904–935, 2008.

[GN06] R. Gribonval and M. Nielsen. Beyond sparsity: recovering structured representations by L1 minimization and greedy algorithms. *Advances in Computational Mathematics*, 28(1):23–41, 2006.

[GWBV02] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. ene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[Hoc83] R. R. Hocking. Developments in linear regression methodology: 1959-1982. *Technometrics*, 25(6):219–249, 1983.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Springer edition, 2009.

[HTIWG10] V.A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9):e12776, 2010.

[LB06] R. W. Lutz and P. Bühlmann. Boosting for high multivariate responses in high dimensional linear regression. *Statistica Sinica*, 16(2):471–494, 2006.

[LPvdGT11] Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 39(4):2164–2204, 2011.

[MRY07] N. Meinshausen, G. Rocha, and B. Yu. Discussion: A tale of three cousins: Lasso, L_2Boosting and Dantzig. *Ann. Statist.*, 35(6):2373–2384, 2007.

[MSMF09] D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano. Generating realistic in silico gene networks for performance assesment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009.

[OM12] C.J. Oates and S. Mukherjee. Network inference and biological dynamics. *Annals of Applied Statistics*, 6(3):1209–1235, 2012.

[OWJ11] G. Obozinski, M.J. Wainwright, and M.I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 39:1–17, 2011.

[Pea09] J. Pearl. *Causality: Models, Reasoning and Inference*. 2nd ed. cambridge university press edition, 2009.

[PZB+10] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J.R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, 4(1):53–77, 2010.

[Rid99] G. Ridgeway. *Generalization of boosting algorithms and applications of Bayesian inference for massive datasets*. PhD thesis, University of Washington, 1999.

[SAB11] M. Solnon, S. Arlot, and F. Bach. Multi-task regression using minimal penalties. *Preprint*, pages 1–33, 2011.

[Sch90] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[Sch99] R. E. Schapire. Theoretical views of boosting. In *Computational learning theory (Nordkirchen, 1999)*, volume 1572 of *Lecture Notes in Comput. Sci.*, pages 1–10. Springer, Berlin, 1999.

[SF96] R.E. Schapire and Y. Freund. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco, 1996. Morgan Kaufman.

[SMF11] T. Schaffter, D. Marbach, and D. Floreano. Genenetweaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27(16):2263–70, 2011.

[ST07] T. Similä and J. Tikka. Input selection and shrinkage in multiresponse linear regression. *Comput. Statist. Data Anal.*, 52(1):406–422, 2007.

[Tem00] V. N. Temlyakov. Weak Greedy Algorithms. *Advances in Computational Mathematics*, 12(2,3):213–227, 2000.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[Tro04] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.

[TZ11] V. N. Temlyakov and P. Zheltov. On performance of greedy algorithms. *Journal of Approximation Theory*, 163(9):1134–1145, 2011.

[Ver12a] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices.* Compressed sensing. cambridge university press edition, 2012.

[Ver12b] N. Verzelen. Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[VVA+11] M. Vignes, J. Vandel, D. Allouche, N. Ramadan-Alban, C. Cierco-Ayrolles, T. Schiex, B. Mangin, and S. de Givry. Gene regulatory network reconstruction using Bayesian networks, the Dantzig selector, the lasso and their meta-analysis. *PLoS ONE*, 6(12), 2011.

[Wai09] M. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55:5728–5741, 2009.

[YL07] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2007.

[ZH05] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[Zha09] T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555–568, 2009.

[ZY06] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.