Journal of Machine Learning Research 26 (2025) 1-54 Submitted 7/21; Revised 3/25; Published 3/25

High-Dimensional L_2-Boosting: Rate of Convergence

Ye Luo KURTLUO@HKU.HK
Hong Kong University Business School
The University of Hong Kong
Hong Kong

Martin Spindler MARTIN.SPINDLER@UNI-HAMBURG.DE
Institute for Statistics
University of Hamburg
Germany

Jannis Kueck KUECK@DICE.HHU.DE
Düsseldorf Institute for Competition Economics
Heinrich Heine University Düsseldorf
Germany

Editor: Corinna Cortes

Abstract

Boosting is one of the most significant developments in machine learning. This paper studies the rate of convergence of L_2-Boosting in a high-dimensional setting under early stopping. We close a gap in the literature and provide the rate of convergence of L_2-Boosting in a high-dimensional setting under approximate sparsity and without beta-min condition. We also show that the rate of convergence of the classical L_2-Boosting depends on the design matrix described by a sparse eigenvalue condition. To show the latter results, we derive new, improved approximation results for the pure greedy algorithm, based on analyzing the revisiting behavior of L_2-Boosting. These results might be of independent interest. Moreover, we introduce so-called "restricted L_2-Boosting". The restricted L_2-Boosting algorithm sticks to the set of the previously chosen variables, exploits the information contained in these variables first and then only occasionally allows to add new variables to this set. We derive the rate of convergence for restricted L_2-Boosting under early stopping which is close to the convergence rate of Lasso in an approximate sparse, high-dimensional setting without beta-min condition. We also introduce feasible rules for early stopping, which can be easily implemented and used in applied work. Finally, we present simulation studies to illustrate the relevance of our theoretical results and to provide insights into the practical aspects of boosting. In these simulation studies, L_2-Boosting clearly outperforms Lasso. An empirical illustration and the proofs are contained in the Appendix.

Keywords: Boosting, L_2-Boosting, restricted L_2-Boosting, early stopping, high-dimensional regression, rate of convergence, approximate sparsity, approximation theory

1. Introduction

In this paper we consider L_2-Boosting algorithms for regression which are coordinatewise greedy algorithms that estimate the target function under L_2 loss. Boosting algorithms represent one of the major advances in machine learning and statistics in recent years. Freund and Schapire's AdaBoost algorithm for classification (Freund and Schapire (1997)) has attracted much attention in the machine learning community as well as in statistics. Many variants of the AdaBoost algorithm have been introduced and proven to be very competitive in terms of prediction accuracy in a variety of applications with a strong resistance to overfitting. Boosting methods were originally proposed as ensemble methods, which rely on the principle of generating multiple predictions and majority voting (averaging) among the individual classifiers (Bühlmann and Hothorn (2007)). An important step in the analysis of boosting algorithms was Breiman's interpretation of boosting as a gradient descent algorithm in function space, inspired by numerical optimization and statistical estimation (Breiman (1996), Breiman (1998)). Building on this insight, Friedman et al. (2000) and Friedman (2001) embedded boosting algorithms into the framework of statistical estimation and additive basis expansion. This also enabled the application of boosting for regression analysis. Boosting for regression was proposed by Friedman (2001), and then Bühlmann and Yu (2003) defined and introduced L_2-Boosting. An extensive overview of the development of boosting and its manifold applications is given in the survey of Bühlmann and Hothorn (2007).

In the high-dimensional setting there are two important but unsolved problems on L_2-Boosting. First, the convergence rate of the L_2-Boosting, in particular under early stopping, has not been thoroughly analyzed. Second, the pattern of the variables selected at each step of L_2-Boosting is unknown. In this paper, we show that these two problems are closely related. We establish results on the sequence of variables that are selected by L_2-Boosting. We call a step of L_2-Boosting "revisiting" if the variable chosen in this step has already been selected in previous steps. We analyze the revisiting behavior of L_2-Boosting, i.e., how often L_2-Boosting revisits. We then utilize these results to derive an upper bound of the rate of convergence of the L_2-Boosting.[1] We show that frequency of revisiting, as well as the convergence speed of L_2-Boosting, depend on the structure of the design matrix, namely on the minimal and maximal restricted eigenvalue.

Moreover, we introduce the so–called "restricted L_2-Boosting", another variant of the classical boosting algorithm. The restricted L_2-Boosting algorithm sticks to the set of the previously chosen variables, exploits the information contained in these variables first and then only occasionally allows to add new variables to this set. We show that the convergence rate is close to that of Lasso and achieves the same rate as orthogonal L_2-Boosting, which was derived in Kueck et al. (2023), but without the need for orthogonal projections which are costly to calculate. It should be noted, that all results are shown in an approximate sparse, high-dimensional setting without beta-min condition. We also introduce feasible, data-driven rules for early stopping for both algorithms, which can be easily implemented and used in applied work.

1. Without analyzing the sequence of variables selected at each step of L_2-Boosting, only much weaker results on convergence speed of L_2-Boosting are available based on DeVore and Temlyakov (1996) and Livshitz and Temlyakov (2003).

Compared to Lasso, boosting uses a somewhat unusual penalization scheme. The penalization is done by "early stopping" to avoid overfitting in the high-dimensional case. In the low-dimensional case, L_2-Boosting without stopping converges to the ordinary least squares (OLS) solution. In a high-dimensional setting, early stopping is key for avoiding overfitting and for the predictive performance of boosting. We give new stopping rules that are simple to implement and also works very well in practical settings as demonstrated in the simulation studies. We prove that such a stopping rule achieves the best bound obtained in our theoretical results. In a deterministic setting, which is when there is no noise or error term in the model, boosting methods are also known as greedy algorithms (the pure greedy algorithm (PGA) and the orthogonal greedy algorithm (OGA)). In signal processing, L_2-Boosting is essentially the same as the matching pursuit algorithm of Mallat and Zhang (1993). We will employ the abbreviations BA (L_2-Boosting algorithm), oBA (orthogonal L_2-Boosting algorithm) and resBA (restricted L_2-Boosting algorithm) for the stochastic versions we analyze. The rate of convergence of greedy algorithms has been analyzed in DeVore and Temlyakov (1996) and Livshitz and Temlyakov (2003). Temlyakov (2011) is an excellent survey of recent results on the approximation theory of greedy approximation. To the best of our knowledge, with an additional assumption on the design matrix, we establish the first results on revisiting in the deterministic setting and greatly improve the existing results of DeVore and Temlyakov (1996). These results, which are available in the appendix, are essential for our analysis of L_2-Boosting, but might also be of interest in their own right.

As mentioned above, boosting for regression was introduced by Friedman (2001). L_2-Boosting was defined in Bühlmann and Yu (2003). Its numerical convergence, consistency, and statistical rates of convergence of boosting with early stopping in a low-dimensional setting were obtained in Zhang and Yu (2005). Consistency in prediction norm of L_2-Boosting in a high-dimensional setting was first proved in Bühlmann (2006). The numerical convergence properties of boosting in a low-dimensional setting are analyzed in Freund et al. (2016). The orthogonal Boosting algorithm in a statistical setting under different assumptions is analyzed in Ing and Lai (2011). The rates for the PGA and OGA are obtained in Barron et al. (2008). For results on orthogonal boosting and modifications, we refer to the excellent survey by Lai and Yuan (2021). In this paper we consider linear basis functions. Classification and regression trees, and the widely used neural networks, involve non-linear basis functions. We hope that our results can serve as a starting point for the analysis of non-linear basis functions which is left for future research.

The structure of this paper is as follows: In Section 2, the L_2-Boosting algorithm (BA/PGA) is defined together with its modifications, the restricted-L_2-Boosting algorithm (resBA/resPGA) and the orthogonalized version (oBA), which serves as reference. In Section 3, we present a new approximation result for the pure greedy algorithm (PGA) and an analysis of the revisiting behavior of the boosting algorithm. In Section 4, we provide the main results of our analysis, namely an analysis of the boosting algorithm and some of its variants. The proofs together with some details of the new approximation theory for PGA are provided in the appendix. Section 5 contains a simulation study that offers some insights into the methods and also provides some guidance for stopping rules in applications. Additional empirical examples can be found in the appendix. Section 6 provides concluding remarks.

Notation: Let z and y be n-dimensional vectors. Define $||z||$ to be the Euclidean norm, and $||z||_{2,n} := \sqrt{\mathbb{E}_n[z^2]}$ to be the empirical L_2-norm with $\mathbb{E}_n[z] = 1/n \sum_{i=1}^n z_i$. Define $< \cdot, \cdot >_n$ to be the inner product defined by: $< z, y >_n = 1/n \sum_{i=1}^n z_i y_i$. For a random variable X, $\mathbb{E}[X]$ denotes its expectation. The correlation between the random variables X and Y is denoted by $corr(X, Y)$. We use the notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We also use the notation $a \precsim b$ to mean $a \leq cb$ for some constant $c > 0$ that does not depend on n; and $a \precsim_P b$ to mean $a = \mathcal{O}_P(b)$. For a vector $\beta \in \mathbb{R}^p$, $supp(\beta)$ denotes the set of indices of which the corresponding element in β is not zero. Further, given a set of indices $T \subset \{1, \ldots, p\}$, we denote by β_T the vector in which $\beta_{T_j} = \beta_j$ if $j \in T$, $\beta_{T_j} = 0$ if $j \notin T$.

2. L_2-Boosting with componentwise least squares

To define the boosting algorithm for linear models, we consider the following regression setting:

$$y_i = x_i'\beta + r_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

with vector $x_i = (x_{i,1}, \ldots, x_{i,p_n})$ consisting of p_n predictor variables, β a p_n-dimensional coefficient vector and a random, mean-zero error term ε_i, $\mathbb{E}[\varepsilon_i | x_i] = 0$. We allow the dimension of the predictors p_n to grow with the sample size n, and to be even larger than the sample size, i.e., $dim(\beta) = p_n > n$. We impose an approximate sparsity condition. This means that there is a large set of potential variables, but the number of relevant variables, which can grow with the sample size, denoted by s_n, is small compared to the sample size, i.e. $\|\beta\|_0 = s_n < n$. The random variable r_i denotes the approximation error of the sparse model. In the following, we will drop the dependence of s_n and p_n on the sample size in the notation and denote it by s and p if no confusion will arise. X denotes the $n \times p$ design matrix where the single observations x_i form the rows. X_j denotes the jth column of design matrix, and $x_{i,j}$ the jth component of the vector x_i. We consider a fixed design for the regressors. We assume that the regressors are standardized with mean zero and variance one, i.e., $\mathbb{E}_n[x_{i,j}] = 0$ and $\mathbb{E}_n[x_{i,j}^2] = 1$ for $j = 1, \ldots, p$,

The basic principle of boosting can be described as follows. We follow the interpretation of Breiman (1998) and Friedman (2001) of boosting as a functional gradient descent optimization (minimization) method. The goal is to minimize a loss function, e.g., an L_2 loss or the negative log-likelihood function of a model, by an iterative optimization scheme. In each step, the (negative) gradient which is used in every step to update the current solution is modelled and estimated by a parametric or nonparametric statistical model, the so-called base learner. The fitted gradient is used for updating the solution of the optimization problem. A strength of boosting, besides the fact that it can be used for different loss functions, is its flexibility with regard to the base learners. We then repeat this procedure until some stopping criterion is met. The literature has developed many different forms of boosting algorithms. In this paper, we consider L_2-Boosting with componentwise linear least squares, as well as two variants. All three are designed for regression analysis. "L_2" refers to the loss function, which is the sum of squares of the residuals $Q_n(\beta) = \sum_{i=1}^n (y_i - x_i'\beta)^2$ typical in regression analysis. In this case, the gradient equals the residuals. "Componentwise linear least squares" refers to the base learners. We fit the gradient (i.e. residuals) against each regressor (p univariate regressions) and select the predictor/variable which correlates most

highly with the gradient/residual, i.e., decreases the loss function most, and then update the estimator in this direction. We next update the residuals and repeat the procedure until some stopping criterion is met. In this paper, we focus on the "classical" L_2-Boosting, which was introduced in Friedman (2001) and refined in Bühlmann and Yu (2003) for regression analysis, and two modifications: restricted-L_2-Boosting and orthogonal L_2-Boosting. Further, we also provide some results for post-L_2-Boosting which is defined in Comment 1. As far as we know, post-L_2-Boosting has not yet been defined and analyzed in the literature.

2.1 L_2-Boosting

For L_2-Boosting with componentwise least squares, the algorithm is given below.

Algorithm 1 (L_2-Boosting (BA/PGA))

(1) Start/Initialization: $\beta^0 = 0$ *(p-dimensional vector), set maximum number of iterations* m_{stop} *and set iteration index* m *to* 0.

(2) At the $(m+1)^{th}$ *step, calculate the residuals* $U_i^m = y_i - x_i'\beta^m$.

(3) For each predictor variable $j = 1, \ldots, p$, *calculate :*

$$\gamma_j^m := \frac{\sum_{i=1}^n U_i^m x_{i,j}}{\sum_{i=1}^n x_{i,j}^2} = \frac{< U^m, x_j >_n}{\mathbb{E}_n[x_{i,j}^2]}.$$

Select the variable j^m *that is the most correlated with the residuals*[2], *i.e.,*

$$\max_{1 \le j \le p} |\gamma_j^m|.$$

(4) Update the estimator: $\beta^{m+1} := \beta^m + \gamma_{j^m}^m e_{j^m}$ *where* e_{j^m} *is the* j^m*th index vector.*

(5) Increase m *by one. If* $m < m_{stop}$, *continue with (2); otherwise stop.*

For simplicity, write γ^m for the value of $\gamma_{j^m}^m$ at the m^{th} step. The act of stopping is crucial for boosting algorithms, as stopping too late or never stopping leads to overfitting and therefore some kind of penalization is required. A suitable solution is to stop early, i.e., before overfitting takes place. "Early stopping" can be interpreted as a form of penalization. Similar to Lasso, early stopping might induce a bias through shrinkage. A potential way to decrease the bias is by "post-L_2-Boosting" which is defined in Comment 1 below. In general, during the run of the boosting algorithm, it is possible that the same variable is selected at different steps, which means the variable is revisited. This revisiting behavior is key to the analysis of the rate of convergence of L_2-Boosting. In the next section, we will analyze the revisiting properties of boosting in more detail.

Remark 1 *Post-L_2-Boosting is a variant of L_2-Boosting, namely, a post-model selection estimator that applies ordinary least squares (OLS) to the model selected by L_2-Boosting in the first step. To define this estimator formally, we make the following definitions:*

2. Equivalently, which fits the gradient best in a L_2-sense.

$T := supp(\beta)$ and $\hat{T} := supp(\beta^{m^*})$, the support of the true model and the support of the model estimated by L_2-Boosting as described above with stopping at m^*. A superscript C denotes the complement of the set with regard to $\{1, \ldots, p\}$. In the context of Lasso, OLS after model selection was analyzed in Belloni and Chernozhukov (2013). Given the above definitions, the post-model selection estimator or OLS post-L_2-Boosting estimator will take the form

$$\tilde{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \; Q_n(\beta) : \beta_j = 0 \quad \text{for each} \quad j \in \hat{T}^C. \tag{2}$$

In this paper, we will not focus on post-L_2-Boosting.

2.2 Restricted L_2-Boosting

In this section, we introduce the so-called restricted L_2-Boosting algorithm. The motivation is that the variable selection behavior of the "pure" greedy algorithm, introduced in the section before, is challenging to analyze. Related to this and even more important, the pure greedy algorithm also has a provably slow rate of convergence in general driven by challenging singular cases, as highlighted by the work of Temlyakov (2011). These cases result in hard to handle variable selection patterns. In general, boosting first selects the relevant variables and then the irrelevant variables. But the overall pattern is difficult to analyze, because the selection of relevant and irrelevant variables can alternate, in particular at an advanced stage of the algorithm when the correlation of the relevant variables with the remainder is small due to previous extraction. The restricted L_2-Boosting algorithm sticks to the set of the already chosen variables for some time and exploits the information contained in these variables, i.e. extracts the correlation of these variables with the remainder until new variables are selected and added to the "consideration" set. The algorithm is given in the following way:

Algorithm 2 (restricted L_2-Boosting Algorithm (resBA/resPGA))

(1) Start/Initialization: $\beta^0 = 0$ (p-dimensional vector) and set iteration index m to 0.

(2) At the $(m+1)^{th}$ step, calculate the residuals $U_i^m = y_i - x_i'\beta^m$. Define the set of variables being selected as $\hat{T}^m$.

(3) Set $T^* := \{1, \ldots, p\}$ if $l_m = 0$ and $T^* := \hat{T}^m$ if $l_m = 1$, for $l_m \in \{0, 1\}$ being a sequence of integers indexed by m.[3]

(4) For each predictor variable $j \in T^*$, calculate :

$$\gamma_j^m := \frac{\sum_{i=1}^n U_i^m x_{i,j}}{\sum_{i=1}^n x_{i,j}^2} = \frac{< U^m, x_j >_n}{\mathbb{E}_n[x_{i,j}^2]}.$$

Select the variable j^m that is the most correlated with the residuals , i.e.,

$$\max_{j \in T^*} |\gamma_j^m|.$$

(5) Update the estimator: $\beta^{m+1} := \beta^m + \gamma_{j^m}^m e_{j^m}$ where e_{j^m} is the j^mth index vector.

3. This criterion is essentially restricting the algorithm within the set of already selected variables.

(6) Stopping Criterion: Stop at m^* which is defined as the first m with

$$\|U^{m+1}\|_{2,n}^2/\|U^m\|_{2,n}^2 > 1 - C_U \frac{\log(2p/\alpha)}{n}$$

when $l_m = 0$ where C_U and α are defined in Theorem 17.

Remark 2 *A related version of restricted L_2-Boosting algorithm is the iterated post-L_2-Boosting algorithm where at certain iterations during the algorithm a projection step is performed on the already selected variables (Kueck et al. (2023)). By projecting the residuals on the selected variables, all correlation/information of the variables is taken out. The restricted L_2-Boosting algorithm avoids these projection steps which are computationally expensive.*

2.3 Orthogonal L_2-Boosting

A variant of the L_2-Boosting algorithm is orthogonal Boosting (oBA) or the orthogonal greedy algorithm in its deterministic version. Only the updating step is changed: after each selection step, an orthogonal projection of the response variable is conducted on all the variables which have been selected up to this point. The advantage of this method is that any variable is selected at most once in this procedure, while in the previous version the same variable might be selected at different steps which makes the analysis far more complicated. More formally, the method can be described as follows by modifying step (4) in Algorithm 1:

Algorithm 3 (Orthogonal L_2-Boosting (oBA))

(1) Start/Initialization: $\beta^0 = 0$ (p-dimensional vector), set maximum number of iterations m_{stop} and set iteration index m to 0.

(2) At the $(m+1)^{th}$ step, calculate the residuals $U_i^m = y_i - x_i'\beta^m$.

(3) For each predictor variable $j = 1, \ldots, p$, calculate :

$$\gamma_j^m := \frac{\sum_{i=1}^n U_i^m x_{i,j}}{\sum_{i=1}^n x_{i,j}^2} = \frac{< U^m, x_j >_n}{\mathbb{E}_n[x_{i,j}^2]}.$$

Select the variable j^m that is the most correlated with the residuals , i.e.,

$$\max_{1 \le j \le p} |\gamma_j^m|.$$

(4) Update the estimator: $\beta^{m+1} = (X_{T^{m+1}}{}' X_{T^{m+1}})^{-1} X_{T^{m+1}}{}' y$.

(5) Increase m by one. If $m < m_{stop}$, continue with (2); otherwise stop.

Remark 3 *Orthogonal L_2-Boosting (and also post-L_2-Boosting) requires, to be well-defined, that the number of selected variables be smaller than the sample size. This is enforced by our stopping rule, as we will see later.*

2.4 Comparison

We have introduced different variants of L_2-Boosting algorithms. Among them, L_2-Boosting is the most widely used algorithm in empirical applications. As we will see later, the rate of convergence is slower than the Lasso rate as there are cases for which the convergence, even in the noiseless case, is too slow. In signal processing and related fields, the noiseless version of the orthogonal Boosting ("orthogonal matching pursuit") is also very popular. It has been shown that for orthogonal L_2-Boosting fast convergence rates with a feasible, data-driven stopping criterion in high dimensions can be obtained even in the case with noise (see Kueck et al. (2023) without beta-min condition and Stankewitz (2024) under stronger assumptions like beta-min condition and normality). The restricted L_2-Boosting can be considered as a middle ground between both. It only uses the boosting steps (without orthogonal projection) but, as we will show, it achieves the same rate of convergence as orthogonal L_2-Boosting. By restricting to the set of already selected variables for some steps, the information of the variables is partialled-out of the target variable y and in the limit it converges to an orthogonal projection step, as used in orthogonal L_2-Boosting. Orthogonal L_2-Boosting might be interpreted as post-L_2-Boosting where the refit takes place after each step. By only selectively/occasionally adding new variables, where only standard boosting steps are used, the rate of convergence is improved compared to L_2-Boosting. Restricted L_2-Boosting can be considered as an approximation to orthogonal Boosting but without orthogonal projections. Therefore, it offers advantageous computational properties as orthogonal projections are costly to calculate. The algorithm only employs the default boosting steps, which are based on univariate regressions and are fast and easy to compute. Therefore, the proposed restricted L_2-Boosting algorithm has excellent theoretical properties and is computationally superior to orthogonal L_2-Boosting. It is also computationally superior to iterated L_2-Boosting, where projection steps are imposed from time to time, since projections can be dispensed at all.

3. New Approximation Results for the Pure Greedy Algorithm

In approximation theory a key question is how fast functions can be approximated by greedy algorithms. Approximation theory is concerned with deterministic settings, i.e., the case without noise:

$$y_i = x_i'\beta, \quad i = 1, \dots, n. \tag{3}$$

Nevertheless, to derive rates for the L_2-Boosting algorithm in a stochastic setting, the corresponding results for the deterministic part play a key role. For example, the results in Bühlmann (2006) are limited by the result used from approximation theory, namely the rate of convergence of weak relaxed greedy algorithms derived in Temlyakov (2000). For the pure greedy algorithm, DeVore and Temlyakov (1996) establish a rate of convergence of $m^{-1/6}$ in the ℓ_2-norm, where m denotes the number of steps iterated in the PGA. This rate was improved to $m^{-11/62}$ in Konyagin and Temlyakov (1999), but Livshitz and Temlyakov (2003) established a lower bound of $m^{-0.27}$. The class of functions $\mathcal{F}$ which is considered in those papers is determined by general dictionaries $\mathcal{D}$ and given by

$$\mathcal{F} = \left\{ f \in \mathcal{H} : f = \sum_{k \in \Lambda} c_k w_k, w_k \in \mathcal{D}, |\Lambda| < \infty \quad \text{and} \quad \sum_{k \in \Lambda} |c_k| \leq M \right\},$$

where M is some constant, $\mathcal{H}$ denotes a Hilbert space, and the sequence (c_k) are the coefficients with regard to the dictionary $\mathcal{D}$. In this section, we discuss the approximation bound of the pure greedy algorithm where we impose an additional but widely used assumption on the Gram matrix $\mathbb{E}_n[x_i x_i']$ in high-dimensional statistics to tighten the bounds. We provide a new lemma on the revisiting behavior of the pure greedy algorithm and a new approximation result which is the core of this section. The proofs for this section and a detailed analysis of the revisiting behavior of the algorithm are moved to Appendix A and Appendix B. Let us define the restricted eigenvalue assumption which is also commonly used in the analysis of Lasso. To do this, consider

$$\Sigma(s, M) := \{A | dim(A) \leq s \times s, A \text{ is any diagonal submatrices of } M\},$$

for any square matrix M.

Definition 4 *The smallest and largest restricted eigenvalues are defined as*

$$\phi_s(s, M) := \min_{W \in \Sigma(s,M)} \phi_s(W),$$

and

$$\phi_l(s, M) := \max_{W \in \Sigma(s,M)} \phi_l(W).$$

$\phi_s(W)$ and $\phi_l(W)$ denote the smallest and largest eigenvalue of the matrix W.

Assummption A.1 (Sparse Eigenvalue (SE))
Consider the Gram matrix $\Sigma = \mathbb{E}_n[x_i x_i']$. Assume that all the elements on the diagonal of Σ are equal to one. We assume that there exist positive constants $c_\phi \leq 1$ and $C_\phi > 1$ such that

$$0 < c_\phi \leq \phi_s(s', \Sigma) \leq \phi_l(s', \Sigma) \leq C_\phi < \infty$$

holds for $s' \leq M_n$, where M_n is a sequence such that $M_n \to \infty$ with n and $M_n \geq C_M s \log(n)$, where C_M is a large enough fixed constant.

Remark 5 *This condition is a variant of the so-called "sparse eigenvalue condition", which is used for the analysis of the Lasso estimator. A detailed discussion of this condition is given in Belloni et al. (2010). Similar conditions, such as the restricted isometry condition or the restricted eigenvalue condition, have been used for the analysis of the Dantzig Selector (Candes and Tao (2007)) or the Lasso estimator (Bickel et al. (2009)). An extensive overview of different conditions on matrices and how they are related is given by van de Geer and Bühlmann (2009). Assuming that $\phi_s(m, E_n[x_i x_i']) > 0$, requires that all empirical Gram submatrices formed by any m components of x_i are positive definite. It is well-known that Condition SE is fulfilled for many designs of interest.*

Define $V^m = X\alpha^m$ as the residual for the PGA. Here, α^m is defined as the difference between the true parameter vector β and the approximation at the m^{th} step, β^m, i.e. $\alpha^m = \beta - \beta^m$. We would like to explore how fast V^m converges to 0. In our notation, $||V^{m+1}||_{2,n}^2 = ||V^m||_{2,n}^2 - (\gamma^m)^2$, therefore $||V^m||_{2,n}^2$ is non-increasing in m. As described in Algorithm 1, the sequence of variables selected in the PGA is denoted by $j^0, j^1, \ldots$. Define

$T^m := T \cup \{j^0, j^1, \ldots, j^{m-1}\}$ with $T := supp(\beta)$. Define $q(m) := |T^m|$ as the cardinality of T^m, $m = 0, 1, \ldots$. It is obvious that $q(0) = s$ and $q(m) \leq m + s$ where $s = \|\beta\|_0$ denotes the number of relevant regressors. It is essential to understand how PGA revisits the set of already selected variables. To analyze the revisiting behavior of the PGA, some definitions are needed to fix ideas.

Definition 6 *We say that the PGA is revisiting at the m^{th} step, if and only if $j^{m-1} \in T^{m-1}$. We define the sequence of labels $\mathcal{A} := \{A_1, A_2, ...\}$ with each entry A_i being either labelled as R(revisiting) or N(non-revisiting).*

Lemma 7 *Assume that assumption A.1 holds with $m < M_n$. Consider the sequence of steps $1, 2, \ldots, m$ with $||V^m||_{2,n}^2 > 0$. Denote $\mu_a(c) = 1 - \left(1 + \frac{1}{c}\right)^{-c}$ for any $c > 0$. Then, for any $\delta > 0$, the number of Rs in the sequence $\mathcal{A}$ at step m, denoted $R(m)$, must satisfy:*

$$|R(m)| \geq \frac{1 - (1+\delta)\mu_a(c_\phi)}{2 - (1+\delta)\mu_a(c_\phi)} m - \frac{(1+\delta)\mu_a(c_\phi)}{2 - (1+\delta)\mu_a(c_\phi)} q(0).$$

The lower bound stated in Lemma 7 has room for improvement, e.g., when $c_\phi = 1$, $|R(m)|/m = 1$ as it is shown in Lemma 20 in Appendix A, while we get $1/2$ in Lemma 7 as lower bounds of $|R(m)/m|$ as m becomes large enough. Deriving tight bounds is an interesting question for future research. More detailed properties of the revisiting behavior of L_2-Boosting are provided in the Appendix A.

With an estimated bound for the proportion of Rs in the sequence $\mathcal{A}$, we are now able to derive an upper bound for $||V^m||_{2,n}^2$. By Lemma 7, define $n_k^*(m) := \frac{m + \mu_a(c_\phi) q(k)}{2 - \mu_a(c_\phi)}$, where $n_k^*(m)(1+\delta)$ is an upper bound of $|q(m+k) - q(k)|$ as $q(m)$ is large enough, for any $\delta > 0$. Before we state the main result of this section, we present an other auxiliary lemma.

Lemma 8 *Assume that assumption A.1 holds. Consider the steps numbered as $k+1, \ldots, k+m$. Assume that $m + k < M_n$ and let $m = \lambda q(k)$ for a constant $\lambda > 0$. Define*

$$\zeta(c, \lambda) := \frac{\frac{c((1-\mu_a(c))\lambda - \mu_a(c))}{2+\lambda}}{\log\left(\frac{2+\lambda}{2-\mu_a(c)}\right)} + c$$

for all $\lambda \geq \frac{\mu_a(c)}{1-\mu_a(c)}$ and $c > 0$. Then, for any arbitrarily small $\delta > 0$ and $q(k)$ being large enough, there exists an arbitrarily small $\delta' > 0$ such that the following statement holds:

$$||V^{m+k}||_{2,n}^2 \leq ||V^k||_{2,n}^2 \left(\frac{q(k)}{q(k) + (1+\delta) n_k^*(m)}\right)^{\zeta(c_\phi, \lambda) - \delta'}.$$

Based on Lemma 8, we are able to develop our main results on the approximation theory of the pure greedy algorithm under L_2 loss.

Theorem 9 (Approximation Theory of PGA based on revisiting) *Assume that assumption A.1 holds. Define $\zeta^*(c) := \max_{\lambda \geq \frac{\mu_a(c)}{1-\mu_a(c)}} \zeta(c, \lambda)$ as a function of c. Then, for any $\kappa > 0$ and $m < M_n$, there exists a fixed constant $C > 0$ such that*

$$\frac{||V^m||_{2,n}^2}{||V^0||_{2,n}^2} \leq C \left(\frac{s}{m+s}\right)^{\zeta^*(c_\phi) - \kappa}$$

for m large enough.

Remark 10 *Our results stated in Theorem 9 depend on the lower bound of $|R(m)|/m$, which is the proportion of the Rs in the first m terms in the sequence $\mathcal{A}$. We conjecture that the convergence rate of PGA is close to exponential as $c \to 0$. Denote the actual proportion of R in the sequence $\mathcal{A}$ by $\psi(c)$, i.e.,*

$$|R(m)| \geq \psi(c)m - \psi_1(c)q(0),$$

where $\psi(c), \psi_1(c)$ are some constants depending on c. If $\psi(c) \to 1$, it is easy to show that

$$||V^m||_{2,n}^2 \precsim ||V^0||_{2,n}^2 \left(\frac{s}{s+m}\right)^{\zeta},$$

based on the proof of Theorem 9, for any arbitrarily large ζ. In general, further improvements in the convergence rate of PGA can be achieved by improving the lower bounds of $|R(m)|/m$. Table 1 gives different values of the SE constant c_ϕ for the corresponding values of $\zeta^*(c_\phi)$. The convergence rate of PGA and hence of L_2-Boosting is affected by the frequency of revisiting. Since different values of c_ϕ impose different lower bounds on the frequency of revisiting, different values of c_ϕ imply a different convergence rate of the process in our framework.

Table 1: Relation between c_ϕ and ζ^*

c_ϕ	$\zeta^*(c_\phi)$
1.0	1.19
0.9	1.04
0.8	0.89
0.7	0.76
0.6	0.63
0.5	0.51
0.4	0.40

Remark 11 *As already mentioned, the function $\zeta^*(c)$ defined in Theorem 9 is used to provide a general lower bound of the revisiting behavior which affects the rate of convergence of boosting algorithms. In some special cases, the PGA algorithm always selects a predictor from the true set T, indicating that $\zeta^*(c_\phi) \to \infty$, and therefore a stronger revisiting behavior can be shown. This is, for example, the case in a (near) orthogonal design, e.g., when X_j, $j = 1, \ldots, p$, are i.i.d. standard normally distributed or in an equi-correlated design, where $-1 < corr(X_i, X_j) = \rho < 1$ for all $i \neq j$. A formal discussion of the equi-correlated design with $\rho > 0$ is given in Appendix D. In both cases, we achieve the Lasso rate of convergence. But under more general designs, greedy algorithms can only achieve a slower rate of convergence (see Temlyakov (2011)), preventing one to achieve the Lasso rate of convergence, as we discuss in more detail in Section 4.*

4. Main Results

In this section, we provide the main results of our paper for L_2-Boosting (BA) and restricted L_2-Boosting (resBA) which were introduced in Section 2. To do this, we reconsider the high-

dimensional approximate sparse regression model in (1):

$$y_i = x_i'\beta + r_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

with vector $x_i = (x_{i,1}, \ldots, x_{i,p})$ consisting of p predictor variables, β a p-dimensional coefficient vector, and a random, mean-zero error term ε_i, $\mathbb{E}[\varepsilon_i|x_i] = 0$. The random variable r_i denotes the approximation error of the exact sparse model. In Assumption A.2, we provide the explicit conditions of the approximate sparse regression model, as, for example, also considered in Belloni et al. (2013) and Belloni et al. (2016). Another sparsity assumption used in the literature is, e.g., weak sparsity (Negahban et al. (2012); Ing (2020)), i.e. $\sum_{j=1}^{p} |\beta_j|^q \leq R_n$, where $0 < q \leq 1$, but in this paper we focus on the approximate sparse regression model.[4] In this section, the following assumptions are employed.

Assummption A.2 (Approximate Sparsity)

(i) $\|\beta\|_0 \leq s$

(ii) $\|r\|_{2,n} := r_n \leq \sqrt{\frac{C_r s \log(2p/\alpha)}{n}}$ *for some generic constant* $C_r > 0$ *and* $\alpha > 0$.

(iii) There exists a constant $K \geq 1$ *such that* $\|\beta\|^2 \leq n^{K-1}$.

Assummption A.3 (Error term)
For any α *small enough, with probability* $\geq 1 - \alpha$*, we have:*

$$\max_{1 \leq j \leq p} |\mathbb{E}_n[x_{i,j}\varepsilon_i]| \leq \sigma \sqrt{\frac{\log(2p/\alpha)}{n}} := \lambda_n.$$

In addition, we require that $\hat{\sigma}^2 := \mathbb{E}_n[\varepsilon_i^2]$ *satisfies that*

$$|\hat{\sigma}^2 - \sigma^2| \leq \omega\sigma^2,$$

for some small enough constant $\omega \in \left(0, \frac{1}{2}\right)$.

Remark 12 *The previous assumption is implied, for example, if the error terms are i.i.d.* $N(0, \sigma^2)$ *random variables. This in turn can be generalized/weakened to cases of non-normality by self-normalized random vector theory (de la Peña et al. (2009)) or the approach introduced in Chernozhukov et al. (2014).*

4.1 L_2-Boosting with Componentwise Least Squares

First, we analyze the classical L_2-Boosting algorithm with componentwise least squares. For this purpose, the approximation results which we derived in the previous section are key. While in the previous section the stochastic component was absent, in this section it is explicitly considered. The following definitions will be helpful for the analysis: U^m denotes the residuals at the m^{th} iteration, $U^m = y - X\beta^m = V^m + r + \varepsilon$. Here, β^m is the estimator at the m^{th} iteration. Again, we define the difference between the true and the estimated vector as $\alpha^m := \beta - \beta^m$. The prediction error is given by $V^m = X\alpha^m$. For boosting algorithms

4. The extension of our results to different sparsity assumptions is left for future work.

in the high-dimensional setting, it is essential to determine when to stop, i.e. the stopping criterion. In the low-dimensional case, stopping time is not important: the value of the objective function decreases and converges to the traditional OLS solution exponentially fast, as described in Bühlmann and Yu (2003). In the high-dimensional case, such fast convergence rates are usually not available: the residual ε can be explained by n linearly independent variables X_j. Thus, selecting more terms only leads to overfitting. Early stopping is comparable to the penalization in Lasso, which prevents one from choosing too many variables and hence overfitting. Similarly to Lasso, an (approximate) sparse structure will be needed for analysis. At each step, we minimize $||U^m||_{2,n}^2$ along the "most greedy" variable X_{j^m}. The following lemma establishes the main result of the convergence rate of L_2-Boosting.

Lemma 13 *Suppose assumptions A.1–A.3 hold and $s\log(p)/n \to 0$. Assume M_n is large enough, i.e.*

$$\log(M_n/s) + \left(\xi + \frac{1}{1+\zeta^*(c_\phi)}\right)\log\left(\frac{s\log(2p/\alpha)}{n||V^0||_{2,n}^2}\right) > 0$$

for some $\xi > 0$. Let m^*+1 be the first time that $||V^m||_{2,n} \le \eta\sqrt{m+s}\lambda_n$, where η is a constant large enough. Then, for any $\delta > 0$, with probability $\ge 1-\alpha$,

(1) it holds

$$m^* \precsim s\left(\frac{s\log(2p/\alpha)}{n||V^0+r||_{2,n}^2}\right)^{\frac{-1}{1+\zeta^*(c_\phi)-\delta}} \quad \text{and} \quad m^* < M_n; \tag{4}$$

(2) the prediction error $||V^{m^*+1}||$ satisfies:

$$||V^{m^*+1}||_{2,n}^2 \precsim ||V^0+r||_{2,n}^{\frac{2}{1+\zeta^*(c_\phi)-\delta}}\left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{\zeta^*(c_\phi)-\delta}{1+\zeta^*(c_\phi)-\delta}}. \tag{5}$$

Remark 14 *Lemma 13 shows that the convergence rate of the L_2-Boosting depends on the value of c_ϕ. For different values of c_ϕ, the lower bound of the proportion of revisiting ("R") in the sequence $\mathcal{A}$ should be different. Such lower bounds on the frequency of revisiting will naturally determine the upper bound for the deterministic component, which affects our results on the rate of convergence of L_2-Boosting. As $\zeta^*(c_\phi) \to \infty$, the statement* (5) *implies the usual Lasso rate of convergence.*

The bound of the approximation error $||V^m||_{2,n}^2$ stated in inequality (5) is obtained under an infeasible stopping criterion. Below we establish another result which employs the same convergence rate but with a feasible stopping criterion which can be implemented in empirical studies.

Theorem 15 *Suppose all conditions stated in Lemma 13 hold. Let $c_u > 4$ be a constant. Let m_1^*+1 be the first time that*

$$\frac{||U^m||_{2,n}^2}{||U^{m-1}||_{2,n}^2} > 1 - c_u\log(2p/\alpha)/n.$$

Then, with probability at least $1-\alpha$,

$$||V^{m_1^*+1}||_{2,n}^2 \precsim ||V^0||_{2,n}^{\frac{2}{1+\zeta^*(c_\phi)-\delta}} \left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{\zeta^*(c_\phi)-\delta}{1+\zeta^*(c_\phi)-\delta}}$$

for any $\delta > 0$.

Remark 16 *As we have already seen in the deterministic case, the rate of convergence depends on the constant* c_ϕ*. Table 2 shows for different values of* c_ϕ *the corresponding rates when* δ *is set to zero. Hence, the rates can be interpreted as upper bounds.*

Table 2: Relation between c and $\frac{\zeta^*(c)}{1+\zeta^*(c)}$

c	rate
1.0	0.54
0.9	0.51
0.8	0.47
0.7	0.43
0.6	0.39
0.5	0.34
0.4	0.29

4.2 Restricted L_2-Boosting

The following theorem establishes the main result of convergence rate of restricted L_2-Boosting.

Theorem 17 *Suppose Assumptions A.1–A.3 hold. Define a sequence of positive integers* m_k *as follows: For* $k = 1, 2, \ldots$*, define* L_k *as a positive integer, and WLOG., we let* $L_k = L_n$ *as a constant, and define* $F_k = \lfloor C_F |\hat{T}^{m_{k-1}}| \log(n) \rfloor$ *for some absolute constant* C_F *large enough*[5] *with*

$$m_k := m_{k-1} + L_k + F_k$$

where $m_0 := 0$*. As suggested in the restricted* L_2*-Boosting, Algorithm 2 runs on the full set of variables when* $l_m = 0$ *and the selected variables when* $l_m = 1$*. To this end, consider a sequence of indices* $l_m \in \{0, 1\}$*,* $m = 1, 2, \ldots$*, with* $l_m = 0$ *for* $m = m_k + 1, ..., m_k + L_k$*, and* $l_m = 1$ *for* $m = m_k + L_k + 1, ..., m_{k+1}$*,* $k \geq 0$*. Suppose that* $L_n < K_L\sqrt{s}$ *for some generic positive constant* $K_L > 0$*. For* $C_U > 4/c_\phi$ *defined in Algorithm 2, with probability* $\geq 1-\alpha$*, we have*

$$\mathbb{E}_n\left[\left(x_i'(\beta^{m^*} - \beta)\right)^2\right] \precsim \frac{s\log(n)\log(2p/\alpha)}{n}.$$

Remark 18 *The bound of the prediction error of restricted* L_2*-Boosting in Theorem 17 is obtained under a feasible stopping criterion which can be easily implemented in empirical*

5. For example, F_n can be chosen as $\ln^\xi(n)$ for some $\xi > 2$ for n large enough.

studies. The constant C_U has to be chosen by the practitioner. In contrast to the constant $c_u > 4$ in Theorem 15, C_u depends on the sparse eigenvalue c_ϕ which can be (pre)-estimated by data and C_u is chosen accordingly. It is worth noting that the restricted L_2-Boosting algorithm is purely gradient based. Hence, there is no need for orthogonal projections which are computationally costly in high-dimensional settings. Therefore, restricted L_2-Boosting combines two desirable properties: computational efficiency and very fast rate of convergence. In practice, one also needs to specify the sequence of indices $l_m \in \{0, 1\}$, $m = 1, 2, \ldots$, which depends on L_k and F_k and thus determines the so-called "consideration" set. As already outlined in Section 2.2, the restricted L_2-Boosting algorithm sticks to the set of the already chosen variables for some time (F_k iterations) and exploits the information contained in these variables until new variables are added to the "consideration" set. By restricting to the set of already selected variables for some iterations, the information of the variables is partialled-out of the target variable y and in the limit it converges to an orthogonal projection step, as used in orthogonal L_2-Boosting. The constant $L_k \geq 1$ is the number of iterations where we allow the algorithm to select new variables.

5. Simulation Study

In this section, we present the results of our simulation study. The goal of this exercise is to illustrate the relevance of our theoretical results in providing insights into the functionality of boosting and the practical aspects of boosting. In particular, we demonstrate that the stopping rules for early stopping we propose work reasonably well in the simulations and give guidance for practical applications. Moreover, the comparison with Lasso might also be of interest. First, we start with an illustrative example and later we present further results, in particular, for different designs and settings.

5.1 Illustrative Example

The goal of this section is to give an illustration of the different stopping criteria. We employ the following data generating process (dgp):[6]

$$y = 5x_1 + 2x_2 + 1x_3 + 0x_4 + \ldots + 0x_{10} + \varepsilon, \tag{6}$$

where $\varepsilon \sim N(0, 2^2)$ and $X = (X_1, \ldots, X_{10}) \sim N_{10}(0, I_{10})$ with I_{10} denoting the identity matrix of size 10×10. To evaluate the methods and, in particular, the stopping criteria, we conduct an analysis of both in-sample and out-of-sample mean squared error (MSE) defined in equation (8). For the out-of-sample analysis we draw a new observation for evaluation and calculation of the MSE. For the in-sample analysis we also repeat the procedure and form the average over all repetitions. In both cases we employ 60 repetitions. The sample size is $n = 20$. Hence, we have 20 observations to estimate 10 parameters. The results are presented in Figures 1 and 2. Both show how MSE depends on the number of steps of the boosting algorithm. We see that MSE first decreases with more steps, reaches its minimum and then starts to increase again due to overfitting. In both graphs the solution of the L_2-Boosting algorithm convergences to the OLS solution. We also indicate the MSE of Lasso estimators as horizontal lines (with cross-validated choice of the penalty parameter

6. In order to allow comparability the dgp is adopted from Bühlmann (2006).

and data-driven choice of the penalization parameter). In order to find a feasible stopping criterion, we have to rely on the in-sample analysis. Figure 2 reveals that the stopping criterion we introduced in the previous sections performs very well and even better than stopping based on a corrected AIC value which has been proposed in the literature as a stopping criterion for boosting. The average stopping steps of our criterion and the corrected AIC-based criterion (AICc) are presented by the vertical lines. On average our criterion stops earlier than the AICc based one. As our criterion performs better than the AICc, we will not report AICc results in the following subsection. For the restricted L_2-Boosting algorithm, similar patterns arise and are omitted.

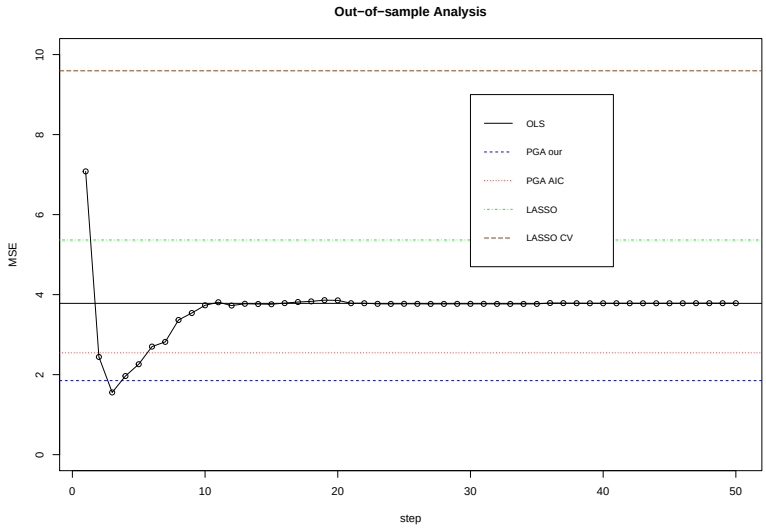

Figure 1: This figure shows the out-of-sample MSE of the L_2-Boosting algorithm depending on the number of steps. The horizontal lines show the MSE of OLS, Boosting and Lasso estimates.

5.2 Further Results

In this section, we present results for different designs and settings to give a more detailed comparison of the methods. We consider the linear model

$$y = \sum_{j=1}^{p} \beta_j x_j + \varepsilon, \tag{7}$$

with ε standard normal distributed and i.i.d.. For the coefficient vector β we consider two designs. First, we consider a sparse design, i.e., the first s elements of β are set equal to one, all other components to zero ($\beta = (1, \ldots, 1, 0, \ldots, 0)$). Then we consider a polynomial design in which the jth coefficient given by $1/j$, i.e., $\beta = (1, 1/2, 1/3, \ldots, 1/p)$. For the design matrix X, we consider two different settings: an "orthogonal" setting and a "correlated" setting. In the former setting, the entries of X are drawn as i.i.d. draws from a standard

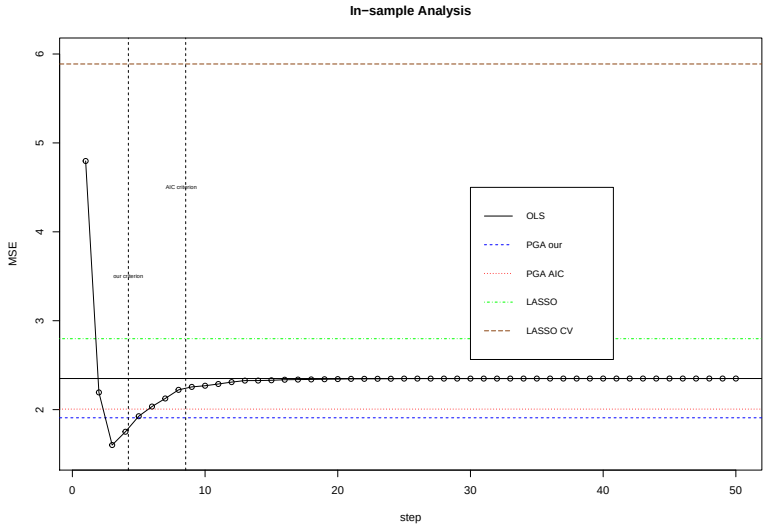

Figure 2: This figure shows the in-sample MSE of the L_2-Boosting algorithm depending on the number of steps. The horizontal lines show the MSE of OLS, Boosting and Lasso estimates.

normal distribution. In the correlated design, the x_i (rows of X) are distributed according to a multivariate normal distribution where the correlations are given by a Toeplitz matrix with factor 0.5 and alternating signs. To sum up, we have the following settings:

- X: "orthogonal" or "correlated"
- coefficient vector β: sparse design or polynomial decaying design
- $n = 100, 200, 400$
- $p = 100, 200$
- $s = 10$
- out-of-sample prediction size $n_1 = 50$
- number of repetitions $R = 500$

We consider the following estimators: L_2-Boosting with componentwise least squares, restricted L_2-Boosting, orthogonal L_2-Boosting and Lasso. For Lasso, we also consider the post-selection estimator ("p-Lasso"). Here, we consider a data-driven regressor-dependent choice for the penalization parameter (Belloni et al. (2012)) and cross-validation. Although cross-validation is very popular, it does not rely on established theoretical results and therefore we prefer a comparison with the formal penalty choice developed in Belloni et al.

Table 3: Simulation results: sparse, iid design (Boosting)

n	p	BA-oracle	BA-our	oBA-oracle	oBA-our	resBA-oracle	resBA-our
100	100	0.454	0.599	0.114	0.475	0.168	0.291
100	200	0.543	0.779	0.123	0.704	0.189	0.359
200	200	0.184	0.307	0.052	0.282	0.072	0.168
400	200	0.080	0.135	0.026	0.121	0.032	0.081
800	200	0.037	0.066	0.013	0.058	0.015	0.042

Table 4: Simulation results: sparse, iid design (Lasso)

n	p	Lasso	p-Lasso	Lasso-CV	p-Lasso-CV
100	100	0.88	0.70	0.54	0.94
100	200	1.02	1.30	0.72	1.02
200	100	0.29	0.28	0.22	0.43
200	200	0.37	0.39	0.30	0.44
400	100	0.13	0.11	0.09	0.18
400	200	0.16	0.20	0.14	0.27

(2012). For our boosting algorithms, we consider two stopping rules: "oracle" and a "data-dependent" stopping criterion ("our") which stops when

$$\frac{||U^m||_{2,n}^2}{||U^{m-1}||_{2,n}^2} = \frac{\hat{\sigma}_{m,n}^2}{\hat{\sigma}_{m-1,n}^2} > 1 - C\log(p)/n$$

for some constant C. Using this approach, boosting stops when the ratio of the estimated variances does not improve upon a certain amount any more. The oracle rule stops when the mean-squared-error (MSE), defined below, is minimized, which is not feasible in practical applications. The simulations were performed in R (R Core Team (2014)). For Lasso estimation the packages *hdm* by Chernozhukov et al. (2015) and *glmnet* by Jerome Friedman (2010) (for cross-validation) were used. The boosting procedures were implemented by the authors and the code is available upon request.[7] To evaluate the performance of the estimators, we use the MSE criterion. We estimate the models on the same data sets and use the estimators to predict 50 observations out-of-sample. The (out-of-sample) MSE is defined as

$$MSE = \mathbb{E}[(f(X) - f^m(X))^2] = \mathbb{E}[(X'(\beta - \beta^m))^2], \tag{8}$$

where m denotes the iteration at which we stop, depending on the employed stopping rule. The MSE is estimated by

$$\frac{1}{n_1}\sum_{i=1}^{n_1}[(f(x_i) - f^m(x_i))^2] = \frac{1}{n_1}\sum_{i=1}^{n_1}[(x_i'(\beta - \beta^m))^2] \tag{9}$$

for the out-of-sample predictions. The results of the simulation study are shown in Tables 3 – 10.

As expected, the oracle-based estimators clearly dominate in almost all cases, although our stopping criterion also gives very good results. Not surprisingly, given our theoretical results, both restricted L_2-Boosting and orthogonal L_2-Boosting outperform the standard L_2-Boosting in most cases. A comparison of restricted and orthogonal L_2-Boosting does

7. A R package is in preparation.

Table 5: Simulation results: sparse, correlated design (Boosting)

n	p	BA-oracle	BA-our	oBA-oracle	oBA-our	resBA-oracle	resBA-our
100	100	1.501	2.232	0.357	1.255	0.670	2.028
100	200	3.220	3.080	2.516	2.552	2.301	2.968
200	200	0.627	0.753	0.059	0.249	0.144	0.215
400	200	0.195	0.245	0.027	0.113	0.068	0.102
800	200	0.081	0.104	0.013	0.054	0.035	0.051

Table 6: Simulation results: sparse, correlated design (Lasso)

n	p	Lasso	p-Lasso	Lasso-CV	p-Lasso-CV
100	100	2.63	1.35	0.97	1.37
100	200	2.96	2.04	1.63	2.38
200	100	1.10	0.23	0.33	0.57
200	200	1.64	0.38	0.49	0.87
400	100	0.38	0.10	0.13	0.23
400	200	0.36	0.15	0.16	0.31

Table 7: Simulation results: polynomial, iid design (Boosting)

n	p	BA-oracle	BA-our	oBA-oracle	oBA-our	resBA-oracle	resBA-our
100	100	0.411	0.548	0.400	0.658	0.400	0.536
100	200	0.252	0.324	0.248	0.356	0.246	0.312
200	200	0.282	0.399	0.274	0.463	0.271	0.381
400	200	0.182	0.232	0.180	0.251	0.178	0.221
800	200	0.122	0.143	0.121	0.148	0.121	0.140

Table 8: Simulation results: polynomial, iid design (Lasso)

n	p	Lasso	p-Lasso	Lasso-CV	p-Lasso-CV
100	100	0.43	0.83	0.45	0.84
100	200	0.50	1.06	0.54	0.76
200	100	0.30	0.34	0.26	0.47
200	200	0.34	0.52	0.33	0.52
400	100	0.19	0.19	0.15	0.24
400	200	0.21	0.31	0.20	0.38

Table 9: Simulation results: polynomial, correlated design (Boosting)

n	p	BA-oracle	BA-our	oBA-oracle	oBA-our	resBA-oracle	resBA-our
100	100	0.242	0.421	0.228	0.506	0.223	0.380
100	200	0.251	0.531	0.244	0.670	0.238	0.489
200	200	0.193	0.312	0.171	0.337	0.165	0.269
400	200	0.149	0.203	0.113	0.188	0.111	0.161
800	200	0.102	0.126	0.078	0.111	0.077	0.100

Table 10: Simulation results: polynomial, correlated design (Lasso)

n	p	Lasso	p-Lasso	Lasso-CV	p-Lasso-CV
100	100	0.33	0.53	0.33	0.55
100	200	0.34	0.93	0.36	0.55
200	100	0.27	0.31	0.23	0.41
200	200	0.28	0.47	0.29	0.46
400	100	0.17	0.18	0.14	0.24
400	200	0.16	0.24	0.15	0.29

not provide a clear answer with advantages on both sides. It is worth noting that the post-Lasso estimator improves upon Lasso, but there are some exceptions, probably driven by overfitting. Cross-validation works very well in many settings. An important objective of the simulation study is to compare L_2-Boosting and Lasso. It seems that in the polynomial decaying setting, L_2-Boosting (particulary orthogonal L_2-Boosting with our stopping rule) dominates post-Lasso. This also seems true in the sparse i.i.d. setting. In the sparse correlated setting, they perform equally well overall. In summary, it seems that L_2-Boosting is a serious contender for Lasso in high-dimensional linear regression models as our new theoretical results propose.

6. Conclusion

Although boosting algorithms are widely used in research and industry, the analysis of their properties in high-dimensional settings has been quite challenging. In this paper, the rate of convergence for the L_2-Boosting algorithm and variants under early stopping in a high-dimensional setting are derived which has been a long-standing open problem until now. For the analysis of the L_2-Boosting algorithm, new approximation results are derived which improve on previous results and might be of independent interest. The rate of pure greedy algorithms can be slow in some pathological cases, as shown in Livshitz and Temlyakov (2003). Therefore, we introduce a new variant called restricted L_2-Boosting which achieves a faster rate of convergence that is comparable to the rate of convergence of Lasso. All results are derived without beta-min condition and under a feasible early stopping criterion. In this paper, we focus on linear basis functions. The analysis of nonlinear basis functions, like trees, is left for future research and the results in this paper might serve as a starting point.

Acknowledgments

We thank the editor Corinna Cortes for her support and two anonymous referees for their valuable comments which helped to improve the paper. We thank Chunrong Ai, Victor Chernozhukov, Jerry Hausman, Scott Kostyshak, Anna Mikusheva, Faming Liang, Whitney Newey, Philippe Rigollet, David Sappington and seminar participants at MIT and University of Florida for invaluable comments and discussions. Financial support by the Fritz Thyssen Stiftung (Az. 40.13.0.014) and Hong Kong RGC TRS (T32-615/24-R) is gratefully acknowledged.

Appendix A. A new approximation theory for PGA

A.1 Auxiliary lemmas on approximation theory of PGA

In this section of the appendix, we introduce preparatory results for a new approximation theory based on revisiting. These results are useful to prove Lemma 7. The proofs of these lemmas are provided in the next section. For any $m_1 \geq m$, define $L(m, m_1) = ||V^{m_1}||_{2,n}^2 / ||V^m||_{2,n}^2 \leq 1$. For any integers $q_1 > q$, define $\Delta(q, q_1) := \Pi_{j=0}^{q_1-q-1} \left(1 - \frac{c_\phi}{q+j}\right)$ with c_ϕ defined in Assumption A.1. It is easy to see that for any $k_1 > k$, $\Delta(k, k_1)/(k/k_1)^{c_\phi} < 1$ and

$$\Delta(k, k_1)/(k/k_1)^{c_\phi} \to 1 \tag{10}$$

as $k \to \infty$. First of all, we can establish the following naive bounds on $L(m, m_1)$:

Lemma 19 *Suppose $||V^m||_{2,n} > 0$, $m + 1 < M_n$ and $m_1 < M_n$. Under Assumption A.1, it holds*

a) For any m, $L(m, m + 1) \leq 1 - \frac{c_\phi}{q(m)}$;

b) For any $m \geq 0$, $m_1 > m$, $L(m, m_1) \leq \Delta(q(m), q(m) + m_1 - m)$.

The bound of $L(m, m_1)$ established in Lemma 19 is loose. To obtain better results on the convergence rate of $\|V^m\|_{2,n}^2$, the revisiting behavior of the PGA has to be analyzed in more detail. The revisiting behavior of PGA addresses the question when and how often variables are selected again which have already been selected before. When PGA chooses too many new variables, it leads on average to slower convergence rates and vice versa. The next results primarily focus on analyzing the revisiting behavior of the PGA. The following lemma summarizes a few basic facts of the sequence of A_i, $i \geq 1$.

Lemma 20 *Suppose $m < M_n$ and $m_1 < M_n$. Further, assume that Assumption A.1 is satisfied. It holds*

a) If $\mathbb{E}_n[x_i' x_i]$ is a diagonal matrix, i.e., $c_\phi = 1$, then there are only Rs in the sequence $\mathcal{A}$.

b) Define $N(m) := \{k | A_k = N, 1 \leq k \leq m\}$, the index set for the non-revisiting steps, and $R(m) := \{k | A_k = R, 1 \leq k \leq m\}$, the index set for the revisiting steps. Then $|R(m)| + |N(m)| = m$, $q(m) = |N(m)| + q(0)$, and $J_N(m) := \{j^k | k \in N(m)\}$ has cardinality equal to $|N(m)|$.

c) $L(0, m) \leq \Pi_{i=1}^{|N(m)|} \left(1 - \frac{c_\phi}{q(0)+i-1}\right) \times \left(1 - \frac{c_\phi}{q(m)}\right)^{|R(m)|}$, i.e., the sequence to maximize the upper bound of $L(0, m)$ stated above is $NN \ldots NRR \ldots R$. Consequently, the sequence $\{A_{m+1}, ..., A_{m_1}\}$ to maximize the upper bound of $L(m, m_1)$ for general $m_1 > m$ is also $NN \ldots NRR \ldots R$.

The proof of this lemma is obvious and hence omitted. Much more involved is the following result for characterizing the revisiting behavior.

Lemma 21 *Assume that Assumption A.1 holds and that $m < M_n$. Consider the sequence of steps $1, 2, \ldots, m$. Set $\mu_e(c_\phi) = (1-\exp(-1/c_\phi^2))$. Then, the number of Rs in the sequence $\mathcal{A}$ satisfies:*

$$|R(m)| \geq \frac{1-\mu_e(c_\phi)}{2-\mu_e(c_\phi)} m - \frac{\mu_e(c_\phi)}{2-\mu_e(c_\phi)} q(0).$$

Lemma 21 provides a lower bound of the proportions of Rs. It illustrates that the R spots occupy at least some significant proportion of the sequence $\mathcal{A}$, with the lower bound of the proportion depending on c_ϕ. In fact, such a result holds for arbitrary consecutive sequence $A_m, A_{m+1}, \ldots, A_{m+k}$, as long as $m+k < M_n$. In the main text, we further extend results stated in Lemma 21.

A.2 Proofs of lemmas in Appendix A.1

Proof [Proof of Lemma 19]
By definition,

$$||V^m||_{2,n}^2 = \sum_{j \in T^m} \alpha_j^m < V^m, X_j >_n = \sum_{j \in T^m} \alpha_j^m \|V^m\|_{2,n} corr(V^m, X_j).$$

Define

$$\rho_{j^m} := |\gamma^m| / \|V^m\|_{2,n} = |corr(V^m, X_{j^m})|$$

since $V^m = U^m$ in the deterministic case. Therefore,

$$\rho_{j^m} \sum_{j \in T^m} \left|\alpha_j^m\right| \geq \|V^m\|_{2,n},$$

i.e., $\rho_{j^m}^2 \left(\sum_{j \in T^m} |\alpha_j^m|\right)^2 \geq \|V^m\|_{2,n}^2$. By the Cauchy-Schwarz inequality,

$$\left(\sum_{j \in T^m} |\alpha_j^m|\right)^2 \leq q(m) ||\alpha^m||^2.$$

Therefore,

$$\rho_{j^m}^2 \geq \frac{c_\phi}{q(m)}$$

with c_ϕ defined in Assumption A.1 and

$$||V^{m+1}||_{2,n}^2 = ||V^m||_{2,n}^2 (1 - \rho_{j^m}^2) \leq \|V^m\|_{2,n}^2 \left(1 - \frac{c_\phi}{q(m)}\right), \tag{11}$$

i.e. $L(m, m+1) \leq 1 - \frac{c_\phi}{q(m)}$. The second statement follows from statement a) by iteration and the fact that $q(m'+1) \leq q(m') + 1$ for any $m' \geq 0$. ■

Proof [Proof of Lemma 21]
Define

$$\widetilde{N}(m) := \{l : j^l \notin T^0, j^l \text{ is only visited once within steps 1,2,...,m}\}$$

with $T^0 = T = supp(\beta)$. It is easy to see that $\widetilde{N}(m) \subset N(m)$ and $|\widetilde{N}(m)| \geq 2|N(m)| - m$ since we excluded T^0 in both $\tilde{N}$ and $N(m)$ and

$$|N(m)| = m - |R(m)|$$

and

$$|\widetilde{N}(m)| \geq m - 2|R(m)|$$

where $N(m) := \{k|A_k = N, 1 \leq k \leq m\}$ is the index set for the non-revisiting steps. For any j^l with $l \in \widetilde{N}(m)$, it holds $\alpha^m_{j^l} = -\gamma^l \; j^l \notin T^0$. If $|R(m)| \geq m/2$, then the statement of this lemma trivially holds. Therefore, we can assume that $\widetilde{N}(m)$ is non-empty. Hence,

$$||\alpha^m||^2 \geq \sum_{l \in \widetilde{N}(m)} (\gamma^{l-1})^2.$$

By the sparse eigenvalue condition A1,

$$\frac{1}{c_\phi}||V^m||^2_{2,n} \geq ||\alpha^m||^2 \geq \sum_{l \in \widetilde{N}(m)} (\gamma^{l-1})^2. \tag{12}$$

Note that by Lemma 19,

$$(\gamma^{l-1})^2 = ||V^{l-1}||^2_{2,n} - ||V^l||^2_{2,n} = ||V^{l-1}||^2_{2,n}\,(1 - L(l-1,l)) \geq \frac{c_\phi}{q(l-1)}||V^{l-1}||^2_{2,n}. \tag{13}$$

Therefore, $(\gamma^{l-1})^2 \geq \frac{c_\phi}{q(l-1)}||V^m||^2_{2,n}$ for all $l \in \tilde{N}(m)$. Plugging this back into (12), we get:

$$\frac{1}{c_\phi}||V^m||^2_{2,n} \geq c_\phi \sum_{l \in \widetilde{N}(m)} \frac{1}{q(l-1)}||V^{l-1}||^2_{2,n}. \tag{14}$$

Since $l \in \widetilde{N}(m)$ are different integers with the maximum value of $q(l-1)$ being less than or equal to $q(m) = q(0) + |N(m)|$, it holds

$$\sum_{l \in \widetilde{N}(m)} \frac{1}{q(l-1)} \geq \sum_{l=1}^{|\widetilde{N}(m)|} \frac{1}{q(0) + |N(m)| - l} \geq \log((q(0)+|N(m)|)/(q(0)+|N(m)|-|\widetilde{N}(m)|)).$$

The inequality above implies that

$$\exp(1/c_\phi^2) \geq (q(0) + |N(m)|)/(q(0) + |N(m)| - |\widetilde{N}(m)|),$$

i.e., $|\widetilde{N}(m)| \leq (1 - \exp(-1/(c_\phi)^2))(q(0) + |N(m)|)$. Set

$$\mu_e(c_\phi) = (1 - \exp(-1/c_\phi^2)) \in (0, 1).$$

Since we know that $|\widetilde{N}(m)| \geq 2|N(m)| - m$, we immediately have:

$$|N(m)| \leq \frac{1}{2 - \mu_e(c_\phi)}(m + \mu_e(c_\phi)q(0))$$

and

$$|R(m)| \geq \frac{1 - \mu_e(c_\phi)}{2 - \mu_e(c_\phi)}m - \frac{\mu_e(c_\phi)}{2 - \mu_e(c_\phi)}q(0).$$

■

Appendix B. Proofs of main results in Section 3

Proof [Proof of Lemma 7]
First of all, WLOG, we can assume that $q(0)$ exceeds a large enough constant $Q(\delta)$. Otherwise, it can be assumed that the true parameter β contains some infinitesimal components such that $q(0) > Q(\delta)$. Let's revisit inequality (13). It holds

$$\sum_{l\in\widetilde{N}(m)} (\gamma^{l-1})^2 \geq \sum_{l\in\widetilde{N}(m)} ||V^{l-1}||^2 \frac{c_\phi}{q(l-1)}.$$

The right-hand side reaches its minimum when

$$\widetilde{N}(m) = \{m - |\widetilde{N}(m)| + 1, m - |\widetilde{N}(m)| + 2, ..., m\},$$

and for the step $m - |\widetilde{N}(m)| + l$, with $l = 1, 2, ..., |\widetilde{N}(m)|$, we have

$$q(m - |\widetilde{N}(m)| + l - 1) = q(m) - |\widetilde{N}(m)| + l - 1.$$

Hence, for any $\delta > 0$, and $q(0)$ large enough,

$$\begin{aligned}
(1+\delta)\sum_{l\in\widetilde{N}(m)} (\gamma^{l-1})^2 &\geq (1+\delta)\sum_{l\in\widetilde{N}(m)} ||V^{l-1}||_{2,n}^2 \frac{c_\phi}{q(l-1)} \\
&= (1+\delta)\sum_{l\in\widetilde{N}(m)} ||V^{m}||_{2,n}^2 \frac{1}{L(l-1,m)}\frac{c_\phi}{q(l-1)} \\
&\geq (1+\delta)||V^m||_{2,n}^2 c_\phi \sum_{l=1}^{|\widetilde{N}(m)|} \frac{1}{q(m-l)}\frac{1}{L(m-l,m)} \\
&\geq ||V^m||_{2,n}^2 c_\phi \sum_{l=1}^{|\widetilde{N}(m)|} \frac{1}{q(m)-l}\left(\frac{q(m)}{q(m)-l}\right)^{c_\phi} \\
&\geq c_\phi^2 \|\alpha^m\|^2 q(m)^{c_\phi} \sum_{k=q(m)-|\tilde{N}(m)|}^{q(m)-1} \left(\frac{1}{k}\right)^{1+c_\phi} \\
&\geq c_\phi \|\alpha^m\|^2 q(m)^{c_\phi}((q(m) - |\widetilde{N}(m)|)^{-c_\phi} - q(m)^{-c_\phi})
\end{aligned}$$

since

$$c_\phi \sum_{k=q(m)-|\tilde{N}(m)|}^{q(m)-1} \left(\frac{1}{k}\right)^{1+c_\phi} = ((q(m) - |\widetilde{N}(m)|)^{-c_\phi} - q(m)^{-c_\phi}),$$

$||V^m||_{2,n}^2 \geq c_\phi\|\alpha^m\|^2$ by Assumption A.1 and $||V^{l-1}||_{2,n}^2/||V^m||_{2,n}^2 = 1/L(l-1,m)$, while

$$L(m-l,m) \to \left(\frac{q(m)-l}{q(m)}\right)^{c_\phi}$$

as $q(m) - l \geq q(m) - |\tilde{N}(m)| \geq q(0) \to \infty$ by Lemma 19. Using $\|\alpha^m\|^2 \geq \sum_{l\in\widetilde{N}(m)}(\gamma^{l-1})^2$, we conclude

$$(1+\delta) \geq c_\phi q(m)^{c_\phi}((q(m) - |\widetilde{N}(m)|)^{-c_\phi} - q(m)^{-c_\phi}),$$

i.e,

$$|\widetilde{N}(m)| \leq q(m)\left[1-\left(1+\frac{1+\delta}{c_\phi}\right)^{-c_\phi}\right] \leq q(m)(1+\delta')\mu_a(c_\phi),$$

for some $\delta' > 0$, with $\delta' \to 0$ as $\delta \to 0$. Since we know that $|\widetilde{N}(m)| \geq 2|N(m)| - m$, we have analogous to the proof of Lemma 21,

$$|N(m)| \leq \frac{(|N(m)| + q(0))(1+\delta')\mu_a(c_\phi) + m}{2}$$

which implies

$$|N(m)| \leq \frac{q(0)(1+\delta')\mu_a(c_\phi) + m}{2 - (1+\delta')\mu_a(c_\phi)}$$

and

$$|R(m)| = m - |N(m)| \geq \frac{1-(1+\delta')\mu_a(c_\phi)}{2-(1+\delta')\mu_a(c_\phi)}m - \frac{(1+\delta')\mu_a(c_\phi)}{2-(1+\delta')\mu_a(c_\phi)}q(0).$$

■

Proof [Proof of Lemma 8]
Without loss of generality, we can assume that $k = 0$. We can also assume that $||V^0||_{2,n}^2 > 0$, because otherwise $||V^0||_{2,n}^2 = ||V^m||_{2,n}^2 = 0$ so that the conclusion already holds. Set

$$n_0 = |N(m)| \leq \frac{m + q(0)(1+\delta)\mu_a(c_\phi)}{2-(1+\delta)\mu_a(c_\phi)} \leq (1+\delta)\frac{m+\mu_a(c_\phi)q(0)}{2-\mu_a(c_\phi)} = (1+\delta)n_0^*(m)$$

for any $\delta > 0$ by Lemma 7 when $q(0)$ is large enough. Then, by Lemma 20,

$$||V^m||_{2,n}^2/||V^0||_{2,n}^2 \leq \Pi_{i=1}^{n_0}\left(1-\frac{c_\phi}{q(0)+i-1}\right)\left(1-\frac{c_\phi}{q(0)+n_0}\right)^{(m-n_0)}$$

where the right hand reaches its maximum when $n_0 = (1+\delta)n_0^*(m)$. When $q(0)$ is large enough, we know that for any $\delta > 0$,

$$\begin{aligned}\Pi_{i=1}^{(1+\delta)n_0^*(m)}\left(1-\frac{c_\phi}{q(0)+i-1}\right) &\leq (1+\delta)\left(\frac{q(0)}{q(0)+(1+\delta)n_0^*(m)}\right)^{c_\phi} \\ &= (1+\delta)\left(\frac{2-\mu_a(c_\phi)}{\lambda(1+\delta)+2+\delta\mu_a(c_\phi)}\right)^{c_\phi}\end{aligned}$$

where

$$\lambda = m/q(0) \tag{15}$$

and

$$\begin{aligned}\left(1-\frac{c_\phi}{q(0)+(1+\delta)n_0^*(m)}\right)^{m-(1+\delta)n_0^*(m)} &= \left(1-\frac{c_\phi}{q(0)\frac{2+(1+\delta)\lambda+\delta\mu_a(c_\phi)}{2-\mu_a(c_\phi)}}\right)^{q(0)\frac{((1-\delta)-\mu_a(c_\phi))\lambda-(1+\delta)\mu_a(c)}{2-\mu_a(c_\phi)}} \\ &\leq \exp\left(-\frac{c_\phi((1-\delta)-\mu_a(c_\phi))\lambda-(1+\delta)\mu_a(c_\phi))}{2+(1+\delta)\lambda+\delta\mu_a(c_\phi)}\right).\end{aligned}$$

Thus, for any $\delta > 0$, and for $q(0)$ large enough,

$$\begin{aligned}
&||V^m||_{2,n}^2/||V^0||_{2,n}^2 \\
&\leq (1+\delta)\left(\frac{2-\mu_a(c_\phi)}{\lambda(1+\delta)+2+\delta\mu_a(c_\phi)}\right)^{c_\phi}\exp\left(-\frac{c_\phi((1-\delta)-\mu_a(c_\phi))\lambda-(1+\delta)\mu_a(c_\phi))}{2+(1+\delta)\lambda+\delta\mu_a(c_\phi)}\right) \\
&= (1+\delta)\left(\frac{q(0)+(1+\delta)n_0^*(m)}{q(0)}\right)^{c_\phi}\exp\left(-\frac{c_\phi((1-\delta)-\mu_a(c_\phi))\lambda-(1+\delta)\mu_a(c_\phi))}{2+(1+\delta)\lambda+\delta\mu_a(c_\phi)}\right)
\end{aligned}$$

where

$$\frac{2-\mu_a(c_\phi)}{\lambda(1+\delta)+2+\delta\mu_a(c_\phi)} = \frac{q(0)+(1+\delta)n_0^*(m)}{q(0)}.$$

It is worth noting that the bound on the right-hand side does not depend on $q(0)$ or m only on λ. Hence, for some $\delta' > 0$ that is small enough that depends on any small enough $\delta > 0$, and for $q(0)$ large enough,

$$\begin{aligned}
||V^m||_{2,n}^2 &\leq ||V^0||_{2,n}^2(1+\delta)\left(\frac{q(0)}{q(0)+(1+\delta)n_0^*(m)}\right)^{c_\phi}\left(\frac{q(0)}{q(0)+(1+\delta)n_0^*(m)}\right)^{\zeta'(c_\phi,\lambda,\delta)} \\
&\leq ||V^0||_{2,n}^2\left(\frac{q(0)}{q(0)+(1+\delta)n_0^*(m)}\right)^{\zeta(c_\phi,\lambda)-\delta'}
\end{aligned}$$

with

$$\zeta'(c_\phi,\lambda,\delta) := \frac{\frac{c_\phi((1-\delta)-\mu_a(c_\phi))\lambda-(1+\delta)\mu_a(c_\phi))}{2+(1+\delta)\lambda+\delta\mu_a(c_\phi)}}{\log\left(\frac{q(0)}{q(0)+(1+\delta)n_0^*(m)}\right)}$$

and $\zeta(c_\phi,\lambda)$ defined in the statement of this lemma,

$$\zeta(c_\phi,\lambda) := \frac{\frac{c_\phi((1-\mu_a(c_\phi))\lambda-\mu_a(c_\phi))}{2+\lambda}}{\log\left(\frac{2+\lambda}{2-\mu_a(c_\phi)}\right)} + c_\phi.$$

■

Proof [Proof of Theorem 9]
As in the proof of Lemma 7, WLOG, we can assume that $q(0)$ exceeds a large enough constant $Q(\delta)$. Otherwise, we can consider the true parameter β contains some infinitesimal components such that $q(0) > Q(\delta)$. Let $\lambda^* \geq \frac{\mu_a(c)}{1-\mu_a(c)}$ be the maximizer of

$$\zeta(c,\lambda) := \frac{\frac{c((1-\mu_a(c))\lambda-\mu_a(c))}{2+\lambda}}{\log\left(\frac{2+\lambda}{2-\mu_a(c)}\right)} + c$$

given $c \in (0,1)$. Let $m_0 = 0$. Define the sequence $m_{i+1} = 1 + \lfloor m_i + \lambda^* q(m_i) \rfloor$ with

$$\lambda^* \leq \lambda_i := \frac{m_{i+1}-m_i}{q(m_i)} \leq \lambda^* + \frac{1}{q(m_i)} \leq \lambda^* + \frac{1}{\bar{q}}$$

and denote

$$n_i^* := \frac{m_{i+1} - m_i + \mu_a(c_\phi) q(m_i)}{2 - \mu_a(c_\phi)}.$$

By Lemma 8, for $q(m_i)$ large enough, e.g., $q(m_i) \geq \bar{q}$ for some $\bar{q}$ large enough, there exists $\delta > 0$ that is small enough, so that

$$\begin{aligned} \frac{||V^{m_{i+1}}||_{2,n}^2}{||V^{m_i}||_{2,n}^2} &\leq \left(\frac{q(m_i)}{q(m_i) + (1+\delta) n_i^*} \right)^{\zeta(c_\phi, \lambda_i) - \delta'} \\ &\leq \left(\frac{q(m_i)}{q(m_i) + (1+\delta) n_i^*} \right)^{\zeta(c_\phi, \lambda^*) - \delta''}, \end{aligned} \tag{16}$$

where $\delta'' := \delta' + \max_{\lambda \in [\lambda^*, \lambda^* + \frac{1}{\bar{q}}]} \frac{\partial \zeta(c_\phi, \lambda)}{\partial \lambda} \frac{1}{\bar{q}}$ is an arbitrarily small constant, as $\bar{q}$ is large enough, δ' is small enough, ζ is a continuously differentiable function with bounded derivatives, and $(1+\delta) n_i^* \geq q(m_{i+1}) - q(m_i)$. For those m with $q(m) \leq \bar{q}$, by (11) in the proof of Lemma 19, we have that

$$\frac{\|V^{m+1}\|_{2,n}^2}{\|V^m\|_{2,n}^2} \leq \left(1 - \frac{c_\phi}{q(m)} \right).$$

Therefore,

$$\begin{aligned} \frac{\|V^{m_{i+1}}\|_{2,n}^2}{\|V^{m_i}\|_{2,n}^2} &= \Pi_{m=m_i}^{m_{i+1}-1} \frac{\|V^{m+1}\|_{2,n}^2}{\|V^m\|_{2,n}^2} \\ &\leq \Pi_{m=m_i}^{m_{i+1}-1} \left(1 - \frac{c_\phi}{q(m)} \right) \\ &\leq \left(1 - \frac{c_\phi}{q(m_{i+1}-1)} \right)^{m_{i+1}-m_i} \\ &\leq \left(1 - \frac{c_\phi}{q(m_{i+1})} \right)^{m_{i+1}-m_i}. \end{aligned} \tag{17}$$

Define i^* as the index of i such that $q(m_{i^*-1}) < \bar{q}$ and $q(m_{i^*}) \geq \bar{q}$. By definition,

$$q(m_{i^*}) \leq q(m_{i^*-1}) + m_{i+1} - m_i \leq (1+\lambda^*) q(m_{i^*-1}) + 1 \leq (1+\lambda^*)\bar{q} + 1 =: \widetilde{q}. \tag{18}$$

Assume $i \geq i^*$. Since $(1+\delta) n_i^* \geq q(m_{i+1}) - q(m_i)$, it holds $q(m_i) \leq (1+\delta) \sum_{j=i^*}^{i-1} n_j^* + q(m_{i^*})$. If $i < i^*$, the statement follows by (17). As $\frac{q(m_i)}{q(m_i)+(1+\delta)n_i^*}$ is increasing in $q(m_i)$, we have that

$$\begin{aligned} \frac{q(m_i)}{q(m_i) + (1+\delta) n_i^*} &\leq \frac{(1+\delta) \sum_{j=i^*}^{i-1} n_j^* + q(m_{i^*})}{(1+\delta) \sum_{j=i^*}^{i-1} n_j^* + q(m_{i^*}) + (1+\delta) n_i^*} \\ &= \frac{(1+\delta) \sum_{j=i^*}^{i-1} n_j^* + q(m_{i^*})}{(1+\delta) \sum_{j=i^*}^{i} n_j^* + q(m_{i^*})}. \end{aligned} \tag{19}$$

By (19), we have that:

$$\begin{aligned}\frac{\|V^{m_i}\|_{2,n}^2}{\|V^0\|_{2,n}^2} &\leq \left(\Pi_{j=i^*}^{i-1}\frac{||V^{m_{j+1}}||_{2,n}^2}{||V^{m_j}||_{2,n}^2}\right)\left(\Pi_{j=0}^{i^*-1}\frac{||V^{m_{j+1}}||_{2,n}^2}{||V^{m_j}||_{2,n}^2}\right)\\ &\leq \left(\Pi_{j=i^*}^{i-1}\frac{q(m_j)}{q(m_j)+(1+\delta)n_j^*}\right)^{\zeta(c_\phi,\lambda^*)-\delta''}\left(1-\frac{c_\phi}{\widetilde{q}}\right)^{m_i^*}\\ &\leq \left(\frac{q(m_{i^*})}{q(m_i^*)+(1+\delta)\sum_{j=i^*}^{i-1}n_j^*}\right)^{\zeta(c_\phi,\lambda^*)-\delta''}\left(1-\frac{c_\phi}{\widetilde{q}}\right)^{m_i^*} \qquad (20)\end{aligned}$$

with $q(m_{j+1}) \leq \bar{q} \leq \tilde{q}$ when $j < i^*$. Note that by definition, we have

$$(1+\delta)\sum_{j=i^*}^{i-1} n_j^* = \frac{(1+\delta)}{2-\mu_a(c_\phi)}\left(m_i - m_{i^*} + \sum_{j=i^*}^{i-1}\mu_a(c_\phi)q(m_j)\right) \geq \frac{1}{2-\mu_a(c_\phi)}(m_i - m_{i^*}). \quad (21)$$

Plug in (21), we have that

$$\frac{\|V^{m_i}\|_{2,n}^2}{\|V^0\|_{2,n}^2} \leq C_\mu\left(\frac{q(m_{i^*})}{q(m_i^*)+m_i-m_{i^*}}\right)^{\zeta(c_\phi,\lambda^*)-\delta''}\left(1-\frac{c_\phi}{\widetilde{q}}\right)^{m_i^*} \qquad (22)$$

where $C_\mu := (2-\mu_a(c_\phi))^{\zeta(c_\phi,\lambda^*)-\delta''}$. By (18), we have

$$q(m_i^*) \leq \widetilde{q}.$$

If $m_i \geq 2m_{i^*}$, we conclude

$$\begin{aligned}\left(\frac{q(m_{i^*})}{q(m_i^*)+m_i-m_{i^*}}\right) &\leq \widetilde{q}\left(\frac{1}{q(m_i^*)+m_i-m_{i^*}}\right)\\ &\leq \widetilde{q}\left(\frac{2}{q(0)+m_i}\right). \qquad (23)\end{aligned}$$

Therefore,

$$\frac{\|V^{m_i}\|_{2,n}^2}{\|V^0\|_{2,n}^2} \leq C'_\mu\left(\frac{q(0)}{q(0)+m_i}\right)^{\zeta(c_\phi,\lambda^*)-\delta''}, \qquad (24)$$

where

$$C'_\mu = C_\mu\left(\frac{\widetilde{q}}{q(0)}\right)^{\zeta(c_\phi,\lambda^*)-\delta''} \leq C_\mu(2\widetilde{q})^{\zeta(c_\phi,\lambda^*)-\delta''}$$

is a fixed constant given $\bar{q}$ and δ''. If $m_i < 2m_{i^*}$, then,

$$\frac{\|V^{m_i}\|_{2,n}^2}{\|V^0\|_{2,n}^2} \leq \left(1-\frac{c_\phi}{\tilde{q}}\right)^{\frac{m_i}{2}} \leq \left(\frac{1}{1+m_i}\right)^{\zeta(c_\phi,\lambda^*)-\delta''} \leq \left(\frac{q(0)}{q(0)+m_i}\right)^{\zeta(c_\phi,\lambda^*)-\delta''}, \qquad (25)$$

for m_i being large enough, as the left hand side of the above inequality decays exponentially in m_i, while the right hand side decays only in fixed polynomial speed of m_i. Therefore, in either case, we have that

$$\frac{\|V^{m_i}\|_{2,n}^2}{\|V^0\|_{2,n}^2} \leq C''_\mu\left(\frac{q(0)}{q(0)+m_i}\right)^{\zeta(c_\phi,\lambda^*)-\delta''}, \qquad (26)$$

for some fixed $C''_\mu > 0$. For any $m > 0$, $m < M_0$, since $m_0, m_1, \ldots$ is an increasing sequence of positive integers, there exists i such that $m_i \leq m < m_{i+1}$. Thus, $\frac{m}{m_i} \leq \frac{m_{i+1}}{m_i} \leq (2 + \lambda^*)$. Therefore,

$$\frac{\|V^m\|_{2,n}^2}{\|V^0\|_{2,n}^2} \leq \frac{\|V^{m_i}\|_{2,n}^2}{\|V^0\|_{2,n}^2} C''_\mu \left(\frac{q(0)}{q(0)+m_i}\right)^{\zeta(c_\phi,\lambda^*)-\delta''} \leq \widetilde{C}_\mu \left(\frac{q(0)}{q(0)+m}\right)^{\zeta(c_\phi,\lambda^*)-\delta''}, \tag{27}$$

with

$$\widetilde{C}_\mu := C''_\mu (2+\lambda^*)^{\zeta(c_\phi,\lambda^*)-\delta''}$$

is a fixed constant. Lastly, by replacing δ'' with κ, the proof is completed.

■

Appendix C. Proofs for the L_2-Boosting algorithm

C.1 Auxiliary lemmas for L_2-Boosting

The two lemmas below state several basic properties of the L_2-Boosting algorithm that will be useful in deriving the main results.

Lemma 22 *It holds*

$$||U^{m+1}||_{2,n}^2 = ||U^m||_{2,n}^2 - < U^m, X_{j^m} >_n^2 = ||U^m||_{2,n}^2 (1 - \rho^2(U^m, X_{j^m}))$$

and

$$||V^{m+1} + r||_{2,n}^2 = ||V^m + r||_{2,n}^2 - 2 < V^m + r, \gamma_{j^m}^m X_{j^m} >_n + (\gamma_{j^m}^m)^2,$$

where $\gamma_{j^m}^m =< U^m, X_{j^m} >_n$. *Moreover, since* $V^m = U^m - r - \varepsilon$,

$$\begin{aligned} ||V^{m+1} + r||_{2,n}^2 &= ||V^m + r||_{2,n}^2 - (2 < U^m, X_{j^m} >_n < U^m - \varepsilon, X_{j^m} >_n) + < U^m, X_{j^m} >_n^2 \\ &= ||V^m + r||_{2,n}^2 + 2\gamma_{j^m}^m < \varepsilon, X_{j^m} >_n - (\gamma_{j^m}^m)^2. \end{aligned}$$

Lemma 23 *Assume that Assumptions A.1-A.3 hold and* $m \leq M_n$. *Let* $Z_m = ||U^m||_{2,n}^2 - ||V^m + r||_{2,n}^2$. *Then, with probability* $\geq 1 - \alpha$ *and uniformly in* m, *it holds*

$$|Z_m - \sigma_n^2| \leq \left(2\sqrt{(1+\omega)C_r s} + 2\frac{\sqrt{m+s}}{\sqrt{c_\phi}}\|V^m\|_{2,n}\right)\lambda_n$$

with $\sigma_n^2 := ||\varepsilon||_{2,n}^2$. Lemma 23 bounds the difference between $||U^m||_{2,n}^2$ and $||V^m + r||_{2,n}^2$.

Proof [Proof of Lemma 23]
By Assumption A.3, we have that:

$$\begin{aligned}
|Z_m - ||\varepsilon||_{2,n}^2| &\leq 2 < \varepsilon, V^m + r >_n \\
&= 2 < \varepsilon, r + X\alpha^m >_n \\
&\leq 2\|\varepsilon\|_{2,n}\|r\|_{2,n} + 2| < \varepsilon, X\alpha^m >_n | \\
&\leq 2\sigma_n\sqrt{\frac{C_r s\log(2p/\alpha)}{n}} + 2\|\alpha^m\|_1\lambda_n \\
&\leq 2\sqrt{(1+\omega)C_r s}\lambda_n + 2\sqrt{m+s}\|\alpha^m\|_2\lambda_n \\
&\leq \left(2\sqrt{(1+\omega)C_r s} + 2\frac{\sqrt{m+s}}{\sqrt{c_\phi}}\|V^m\|_{2,n}\right)\lambda_n
\end{aligned}$$

as $|supp(\alpha^m)| \leq m + s$. ■

Next, we provide the lemmas that we require for proving our main result in Theorem 17. Define

$$U_T^o := y - \mathcal{P}_T y$$

as the residual of projection of y on X_T for any $T \subset \{1, 2, ..., p\}$. Further, we define $\mathcal{P}_{m,T}$ as the operator of performing m periods of PBA algorithm subject to the subset T, i.e.,

$$U_T^m := y - \mathcal{P}_{m,T} y = (I - \mathcal{P}_{m,T})U^0$$

with $U^0 := y$ and $V_T^m := U_T^0 - U_T^m$ denotes the approximation error. Lemma 24 measures a bound between the operator $\mathcal{P}_{m,T}$ and the projection operator $\mathcal{P}_T$. That said, $\mathcal{P}_{m,T}$ is an approximation of $\mathcal{P}_T$.

Lemma 24 (Iterations) *Suppose that Assumption A.1 holds. Suppose $T \subset \{1, 2, ..., p\}$ with $|T| \leq M_n$. Then, starting with $U^0 = y$, the L_2-Boosting algorithm that runs on the restricted subset T satisfies that:*

$$\|U^m - U_T^o\|_{2,n}^2 \leq \|y - U_T^o\|_{2,n}^2 \exp\left(-\frac{1}{|T|}c_\phi m\right). \tag{28}$$

Proof [Proof of Lemma 24] Given the L_2-Boosting algorithm,

$$\gamma_{j^m} := \max_{j\in T} |\frac{< U^m, X_{j^m} >_n}{\|X_{j^m}\|_{2,n}}| = | < U^m, X_{j^m} >_n |, \tag{29}$$

with j^m being the largest index. The approximation error can be written as follows:

$$V_T^m := U_T^0 - U_T^m = X_T\alpha_T^m.$$

We consider

$$U_T^m = V_T^m + \eta$$

where $\eta = U_T^m - V_T^m = U_T^o = y - \mathcal{P}_T y$ is orthogonal to all the column vectors in X_T. Note that η is never changing when m increases. As a result, we have that:

$$|\gamma_{j^m}|\|\alpha_T^m\|_1 \geq \sum_{j\in T} < U_T^m, \alpha_j^m X_{j^m} >_n = \|V_T^m\|_{2,n}^2. \tag{30}$$

Therefore,

$$|\gamma_{j^m}|^2 \geq \frac{\|V_T^m\|_{2,n}^4}{\|\alpha_T^m\|_1^2} \geq \frac{\|V_T^m\|_{2,n}^4}{|T|\|\alpha_T^m\|_2^2} \geq \frac{c_\phi \|V_T^m\|_{2,n}^4}{|T|\|V_T^m\|_{2,n}^2} = \frac{c_\phi \|V_T^m\|_{2,n}^2}{|T|}.$$

That is to say,

$$\|V_T^{m+1}\|_{2,n}^2 = \|U_T^{m+1}\|_{2,n}^2 - \|\eta\|_{2,n}^2 = \|U_T^m\|_{2,n}^2 - |\gamma_{j^m}|^2 - \|\eta\|_{2,n}^2 \leq \|V_T^m\|_{2,n}^2 \left(1 - \frac{c_\phi}{|T|}\right). \quad (31)$$

As a result,

$$\|V_T^m\|_{2,n}^2 \leq \|V_T^0\|_{2,n}^2 \left(1 - \frac{c_\phi}{|T|}\right)^m \leq \|V_T^0\|_{2,n}^2 \exp\left(-\frac{c_\phi}{|T|} m\right) \quad (32)$$

which proves the statement of the lemma. ■

Lemma 25 (Bounds on Residuals) *Assuming that assumptions A.1-A.3 hold, for any positive integer $M \leq M_n$, we have:*

$$\sup_{T \subset \{1,2,\ldots,p\}, |T| \leq M} \|\frac{1}{n} X_T^\intercal \varepsilon\|_{2,n}^2 \leq M\lambda_n^2,$$

$$\sup_{T \subset \{1,2,\ldots,p\}, |T| \leq M} \|\mathcal{P}_T \varepsilon\|_{2,n}^2 \leq \frac{1}{c_\phi} M\lambda_n^2,$$

$$\sup_{T \subset \{1,2,\ldots,p\}, |T| \leq M} \|\mathcal{P}_T(r+\varepsilon)\|_{2,n}^2 \leq 2\left(\frac{C_r s \log(2p/\alpha) + \sigma^2/c_\phi \log(2p/\alpha) M}{n}\right)$$

with probability $\geq 1 - \alpha$.

The proof is given in Lemma 1 in the appendix of Kueck et al. (2023).

Lemma 26 (lower bound on residuals) *Given Assumption A.1. Consider $U^m = X\alpha^m + r + \varepsilon$ with $\|\alpha^m\|_0 \leq M_n$. Then,*

$$|\gamma_{j^m}^m| \geq \sqrt{\frac{c_\phi}{\|\alpha^m\|_0}} \|X\alpha^m\|_{2,n} - \sqrt{\frac{C_r c_\phi s \log(2p/\alpha)}{n\|\alpha^m\|_0}} - \lambda_n$$

with probability $\geq 1 - \alpha$.

Proof [Proof of Lemma 26] Denote $\rho_m := \max_{j=1,2,\ldots,p} |corr(U^m, X_j)|$. It can be shown that $(\gamma_{j^m}^m)^2 = \rho_m^2 \|U^m\|_{2,n}^2$. We know that

$$\begin{aligned} < U^m, X\alpha^m >_n &\geq \|X\alpha^m\|_{2,n}^2 - \|r\|_{2,n}\|X\alpha^m\|_{2,n} - \lambda_n \|\alpha^m\|_1 \\ &\geq \|X\alpha^m\|_{2,n}^2 - \sqrt{\frac{C_r s \log(2p/\alpha)}{n}} \|X\alpha^m\|_{2,n} - \lambda_n \sqrt{\frac{\|\alpha^m\|_0}{c_\phi}} \|X\alpha^m\|_{2,n}. \end{aligned}$$

On the other hand, we have that:

$$\begin{aligned} < U^m, X\alpha^m >_n \leq & \sum_{j \in supp(\alpha)} \alpha_j^m < U, X_j >_n \leq \rho_m \|U^m\|_{2,n} \sum_{j \in supp(\alpha)} |\alpha_j^m| \\ & \leq |\gamma_{j^m}^m| \sqrt{\|\alpha^m\|_0} \|\alpha^m\|_2 \leq |\gamma_{j^m}^m| \sqrt{\frac{\|\alpha^m\|_0}{c_\phi}} \|X\alpha^m\|_{2,n}. \end{aligned}$$

Therefore, we have that:

$$|\gamma_{j^m}^m| \geq \sqrt{\frac{c_\phi}{|\|\alpha^m\|_0}} \|X\alpha^m\|_{2,n} - \sqrt{\frac{C_r c_\phi s \log 2p/\alpha}{n\|\alpha^m\|_0}} - \lambda_n. \tag{33}$$

■

C.2 Proofs of main results in Section 4

Proof [Proof of Lemma 13]
By assumption A.3, we know that $\lambda_n \geq \max_{1 \leq j \leq p} | < \varepsilon, X_j >_n |$ with probability $\geq 1 - \alpha$. According to our definition, $m^* + 1$ is the first time

$$||V^m||_{2,n} \leq \eta \sqrt{m + s} \lambda_n,$$

where η is a fixed positive constant that is large enough with $\eta > 2\sigma\sqrt{C_r}$. We know that in high-dimensional settings, $||U^m||_{2,n} \to 0$, so $||V^m||_{2,n}^2 \to \sigma^2$. Thus, by fixing p and n, such an m^* must exist. Therefore, we can consider any $m < \tilde{m} := (m^* + 1) \wedge M_n$. Note that $T^m := T_0 \cup \{j^0, j^1, \ldots, j^{m-1}\}$ and $V^m := X\alpha^m = X_{T^m}\alpha_{T^m}$. It holds

$$\begin{aligned} |\gamma_{j^m}^m| \|\alpha_{T^m}\|_1 \geq & < U^m, X_{T^m}\alpha_{T^m} >_n \\ & \geq \|V^m\|_{2,n}^2 - | < r, X_{T^m}\alpha_{T^m} > | - | < \varepsilon, X_{T^m}\alpha_{T^m} > | \\ & \geq \|V^m\|_{2,n}^2 - \|r\|_{2,n}\|V^m\|_{2,n} - \lambda_n \|\alpha_{T^m}\|_1. \end{aligned}$$

As we assume that $\|V^m\|_{2,n} > \eta\sqrt{m + s}\lambda_n$, with η being large enough, we have that $\|r\|_{2,n} \leq \frac{\sqrt{C_r}}{\sigma\eta}\|V^m\|_{2,n}$. Then,

$$\|V^m\|_{2,n}^2 - \|r\|_{2,n}\|V^m\|_{2,n} \geq \left(1 - \frac{\sqrt{C_r}}{\sigma\eta}\right) \|V^m\|_{2,n}^2$$

which implies that:

$$
\begin{aligned}
|\gamma_{j^m}| &\geq \frac{\left(1-\frac{\sqrt{C_r}}{\sigma\eta}\right)\|V^m\|_{2,n}^2}{\|\alpha_{\widetilde{T}^m}\|_1} - \lambda_n \geq \frac{\left(1-\frac{\sqrt{C_r}}{\sigma\eta}\right)\|V^m\|_{2,n}^2}{\sqrt{q(m)}\|\alpha_{\widetilde{T}^m}\|_2} - \lambda_n \\
&\geq \frac{\sqrt{c_\phi}\left(1-\frac{\sqrt{C_r}}{\sigma\eta}\right)}{\sqrt{q(m)}}\|V^m\|_{2,n} - \lambda_n \\
&\geq \frac{\sqrt{c_\phi}\left(1-\frac{\sqrt{C_r}}{\sigma\eta}-\frac{1}{\sqrt{c_\phi}\eta}\right)}{\sqrt{q(m)}}\|V^m\|_{2,n} \\
&\geq C_{V,1}\frac{\|V^m+r\|_{2,n}}{\sqrt{q(m)}} \geq C_{V,1}\left(\eta-\frac{\sqrt{C_r}}{\sigma}\right)\lambda_n
\end{aligned} \tag{34}
$$

where $\|V^m\|_{2,n} \geq \eta\sqrt{m+s}\lambda_n \geq \eta\sqrt{q(m)}\lambda_n$, and

$$
C_{V,1} := \sqrt{c_\phi}\left(1-\frac{\sqrt{C_r}}{\sigma\eta}-\frac{1}{\sqrt{c_\phi}\eta}\right)\frac{1}{1+\frac{\sqrt{C_r}}{\sigma\eta}}
$$

is a positive constant when η is large enough, and approaches to $\sqrt{c_\phi}$ as η goes to infinity. By Lemma 22, it holds that

$$
\begin{aligned}
||V^m+r||_{2,n}^2 - ||V^{m+1}+r||_{2,n}^2 &= ||U^m||_{2,n}^2 - ||U^{m+1}||_{2,n} - 2\gamma_{j^m}^m < X_{j^m}, \varepsilon >_n \\
&= (\gamma_{j^m}^m)^2 - 2\gamma_{j^m}^m < X_{j^m}, \varepsilon >_n .
\end{aligned} \tag{35}
$$

By equation (34), it follows that:

$$
\begin{aligned}
||V^m+r||_{2,n}^2 - ||V^{m+1}+r||_{2,n}^2 &\geq (\gamma_{j^m}^m)^2 - 2\gamma_{j^m}^m < X_{j^m}, \varepsilon >_n \\
&\geq (\gamma_{j^m}^m)^2 - 2|\gamma_{j^m}^m|\lambda_n \\
&\geq (\gamma_{j^m}^m)^2\left(1-\frac{2}{C_{V,1}\left(\eta-\frac{\sqrt{C_r}}{\sigma}\right)}\right) \\
&\geq C_{V,2}\|V^m+r\|_{2,n}^2,
\end{aligned}
$$

when η is large enough where

$$
C_{V,2} := \left(1-\frac{2}{C_{V,1}\left(\eta-\frac{\sqrt{C_r}}{\sigma}\right)}\right)C_{V,1}^2
$$

is arbitrarily close to c_ϕ when η is large enough. By the arguments in the proof of Lemma 21, when n is large enough, by (34) and $\|r\|_{2,n} \leq \frac{\sqrt{C_r}}{\sigma\eta}\|V^m\|_{2,n}$, we have that:

$$\begin{aligned}\frac{1}{c_\phi\left(1-\frac{\sqrt{C_r}}{\sigma\eta}\right)^2}\|V^m+r\|_{2,n}^2 &\geq \frac{1}{c_\phi}||V^m||_{2,n}^2 \\ &\geq \sum_{k\in\widetilde{N}(m)}(\gamma^{k-1})^2 \\ &\geq \sum_{k\in\widetilde{N}(m)}\frac{C_{V,1}^2}{q(k-1)}||V^{k-1}+r||_{2,n}^2\end{aligned}$$

which corresponds to (14) in Lemma 21 considering $\|V^m+r\|_{2,n}^2$ instead of $\|V^m\|_{2,n}^2$. Define

$$\psi = \max\left\{c_\phi - C_{V,2}, c_\phi - C_{V,1}^2/\left(1-\frac{\sqrt{C_r}}{\sigma\eta}\right)^2\right\}$$

which can be arbitrarily close to 0 as η is large enough. Thus, following the proof of Lemma 7 and Lemma 8, we can treat $c_\phi - \psi$ as the constant c_ϕ in Theorem 9. Therefore, results of Lemma 7 and Lemma 8 applies for $V^m + r$ with c_ϕ replaced by $c_\phi - \psi$. We conclude that

$$\frac{\|V^m+r\|_{2,n}^2}{\|V^0+r\|_{2,n}^2} \leq C\left(\frac{s}{s+m}\right)^{\zeta^*(c_\phi)-\psi'}, \tag{36}$$

for some arbitrarily small $\psi' > 0$ as η is large enough. By using that $\|V^m + r\|_{2,n} \geq \|V^m\|_{2,n} - \|r\|_{2,n}$, we have that

$$\left(\eta - \frac{\sqrt{C_r}}{\sigma}\right)\lambda_n\sqrt{m+s} \leq \|V^m+r\|_{2,n}$$

for all $m < \widetilde{m}$. It follows that

$$\left(\eta - \frac{\sqrt{C_r}}{\sigma}\right)^2\lambda_n^2(\widetilde{m}-1+s) \leq \widetilde{C}\|V^0+r\|_{2,n}^2\left(\frac{s}{s+\widetilde{m}-1}\right)^{\zeta^*(c_\phi)-\psi'}. \tag{37}$$

Therefore

$$\frac{s\log(p)}{n} \precsim ||V^0+r||_{2,n}^2\left(\frac{s}{s+\widetilde{m}-1}\right)^{\zeta^*(c_\phi)-\psi''+1},$$

where ψ'' can be arbitrarily close to 0 as m is large enough or equivalently,

$$\widetilde{m} \precsim s\left(\frac{s\log(2p/\alpha)}{n||V^0+r||_{2,n}^2}\right)^{-\frac{1}{1+\zeta^*(c_\phi)-\psi''}}.$$

By assumption,

$$\log(M_n/s) + \left(\xi + \frac{1}{1+\zeta^*(c_\phi)}\right)\log\left(\frac{s\log(2p/\alpha)}{n||V^0+r||_{2,n}^2}\right) > 0$$

for some $\xi > 0$. Thus, asymptotically,

$$\widetilde{m} = M_n \wedge (m^* + 1) < M_n,$$

i.e., $m^* + 1 < M_n$. Thus,

$$m^* \precsim s \left(\frac{s \log(2p/\alpha)}{n||V^0 + r||_{2,n}^2} \right)^{-\frac{1}{1+\zeta^*(c_\phi)-\psi''}} \tag{38}$$

and

$$||V^{m^*+1}||_{2,n}^2 \leq \eta\sqrt{m^* + 1 + s}\lambda_n \precsim ||V^0 + r||_{2,n}^{\frac{2}{1+\zeta^*(c_\phi)-\psi''}} \left(\frac{s \log(2p/\alpha)}{n} \right)^{\frac{\zeta^*(c_\phi)-\psi''}{1+\zeta^*(c_\phi)-\psi''}},$$

for any small $\psi'' > 0$ if η is large enough. ■

Proof [Proof of Theorem 15]
At the $(m_1^* + 1)^{th}$ step, we have:

$$||U^{m_1^*+1}||_{2,n}^2 > (1 - c_u \log(2p/\alpha)/n)||U^{m_1^*}||_{2,n}^2.$$

It follows that $(\gamma^{m_1^*})^2 < c_u \log(2p/\alpha)/n||U^{m_1^*}||_{2,n}^2$, while $(\gamma^m)^2 \geq c_u \log(2p/\alpha)/n||U^m||_{2,n}^2$ for all $m < m_1^*$. Consider the m^* defined in Lemma 13 as a reference point.

Case (a): Suppose $m_1^* < m^*$: By the proof of Lemma 13, $||V^m + r||_{2,n}^2$ is decreasing when $m \leq m_1^* + 1 \leq m^*$. By Lemma 23, it holds

$$||U^{m_1^*}||_{2,n}^2 \leq \sigma_n^2 + \left(2\sqrt{(1+\omega)C_r s} + 2\frac{\sqrt{m_1^* + s}}{\sqrt{c_\phi}}||V^{m_1^*}||_{2,n} \right) \lambda_n + \|V^{m_1^*} + r\|_{2,n}^2.$$

It follows that

$$\begin{aligned}(\gamma^{m_1^*})^2 &< c_u \log(2p/\alpha)/n||U^{m_1^*}||_{2,n}^2 \\ &< (1+\omega)c_u\lambda_n^2 + c_u \log(2p/\alpha)/n \left(2\sqrt{(1+\omega)C_r s} + 2\frac{\sqrt{m_1^* + s}}{\sqrt{c_\phi}}||V^{m_1^*}||_{2,n} \right) \lambda_n \\ &+ c_u \log(2p/\alpha)/n\|V^{m_1^*} + r\|_{2,n}^2. \end{aligned} \tag{39}$$

By inequality (34), we have that

$$|\gamma^{m_1^*}| \geq C_{V,1} \frac{\|V^{m_1^*} + r\|_{2,n}}{\sqrt{q(m_1^*)}},$$

for η large enough where $\|V^m\|_{2,n} \geq \eta\sqrt{m+s}\lambda_n$ holds for all $m \leq m^*$, including m_1^* by assumption that $m_1^* < m^*$. Combining this with inequality (39), we have that:

$$\begin{aligned}
&C_{V,1}^2 \frac{\|V^{m_1^*} + r\|_{2,n}^2}{m_1^* + s} \\
&\leq c_u\lambda_n^2(1+\omega) + 2c_u\lambda_n \frac{\log(2p/\alpha)}{n}\sqrt{(1+\omega)C_r s} \\
&\quad + 2c_u\lambda_n^3 \frac{\sqrt{m_1^*+s}}{\sigma^2\sqrt{c_\phi}}\|V^{m_1^*}\|_{2,n} \\
&\quad + c_u \log(2p/\alpha)/n\|V^{m_1^*} + r\|_{2,n}^2. \qquad (40)
\end{aligned}$$

We show that this leads to a contradiction and hence $m_1^* \geq m^*$. First, by Lemma 13, we know that

$$\sqrt{m^*+1+s}\lambda_n \precsim ||V^0 + r||_{2,n}^{\frac{2}{1+\zeta^*(c_\phi)-\psi''}} \left(\frac{s\log(2p/\alpha}{n}\right)^{\frac{\zeta^*(c_\phi)-\psi''}{1+\zeta^*(c_\phi)-\psi''}} \to 0 \qquad (41)$$

as n is large enough since $\frac{s\log(2p/\alpha)}{n} \to 0$ by assumption. As a result, for n large enough, we have that

$$\frac{1}{3}C_{V,1}^2 \frac{\|V^{m_1^*} + r\|_{2,n}^2}{m_1^* + s} > c_u \log(2p/\alpha)/n\|V^{m_1^*} + r\|_{2,n}^2 \qquad (42)$$

which corresponds to the third component in (40). Second, since $\|V^{m_1^*}\|_{2,n} \geq \eta\sqrt{m_1^*+s}\lambda_n$ by construction for η large enough, we have

$$\|V^{m_1^*} + r\|_{2,n}^2 \geq \left(1 - \frac{\sqrt{C_r}}{\sigma\eta}\right)^2 \|V^{m_1^*}\|_{2,n}^2$$

as shown in the proof of Lemma 13. Hence, for η being a large enough fixed constant and n large enough, we have that

$$\begin{aligned}
&\frac{1}{3}C_{V,1}^2\|V^{m_1^*} + r\|_{2,n}^2 \\
&\geq \frac{1}{3}C_{V,1}^2\|V^{m_1^*}\|_{2,n}^2\left(1 - \frac{\sqrt{C_r}}{\sigma\eta}\right)^2 \\
&\geq \|V^{m_1^*}\|_{2,n}\frac{1}{3}C_{V,1}^2\left(1 - \frac{\sqrt{C_r}}{\sigma\eta}\right)^2\sqrt{m_1^*+s}\lambda_n \\
&= 2\|V^{m_1^*}\|_{2,n}c_u/(\sigma^2\sqrt{c_\phi})\lambda_n^3\sqrt{m_1^*+s}\left(\sigma^2\sqrt{c_\phi}C_{V,1}^2\left(1 - \frac{\sqrt{C_r}}{\sigma\eta}\right)^2/(6c_u\lambda_n^2)\right) \\
&> 2c_u\lambda_n^3\frac{\sqrt{m_1^*+s}}{\sigma^2\sqrt{c_\phi}}\|V^{m_1^*}\|_{2,n} \qquad (43)
\end{aligned}$$

as $\sigma^2\sqrt{c_\phi}C_{V,1}^2\left(1-\frac{\sqrt{C_r}}{2\sigma\eta}\right)^2/(6c_u\lambda_n^2)\to\infty$. Third, for η large enough, we have that

$$\begin{aligned}\frac{1}{3}C_{V,1}^2\frac{\|V^{m_1^*}+r\|_{2,n}^2}{m_1^*+s} &\geq \frac{1}{3}C_{V,1}^2\eta^2\left(1-\frac{\sqrt{C_r}}{\sigma\eta}\right)^2\lambda_n^2\\ &> 2c_u\lambda_n^2(1+\omega)\\ &> c_u\lambda_n^2(1+\omega)+2c_u\lambda_n\frac{\log(2p/\alpha)}{n}\sqrt{(1+\omega)C_r s},\end{aligned}\tag{44}$$

as $2c_u\frac{\log(2p/\alpha)}{n}\sqrt{s}=2c_u\lambda_n/\sigma\sqrt{\frac{s\log(2p/\alpha)}{n}}=o(\lambda_n)$ since $\frac{s\log(2p/\alpha)}{n}\to 0$ for n large enough. Therefore, equations (42)–(44) lead to a contradiction with (40).

Case (b): We know that $m_1^*\geq m^*$: It follows that

$$(\gamma_{j^m}^m)^2\geq c_u\log(2p/\alpha)/n||U^m||_{2,n}^2$$

for all $m<m_1^*$. Since $||U^m||_{2,n}^2$ is a decreasing sequence, for δ small enough, there exists some m_2 such that $||U_2^m||_{2,n}^2>(1-\delta)\sigma_n^2$ for any $m\leq m_2$, and $||U^{m_2+1}||_{2,n}^2\leq(1-\delta)\sigma_n^2$. For δ small enough and $m\leq m_2\wedge m_1^*$, it holds

$$||V^{m+1}+r||_{2,n}^2-||V^m+r||_{2,n}^2=-(\gamma_{j^m}^m)^2+2\gamma_{j^m}^m<\varepsilon,X_{j^m}>_n\leq-(\gamma_{j^m}^m)^2+2\lambda_n|\gamma_{j^m}^m|$$

by Lemma 22. Since $c_u>4$, for δ small enough, it holds

$$|\gamma_{j^m}^m|^2\geq c_u\log(2p/\alpha)/n||U^m||_{2,n}^2\geq c_u(1-\delta)(1-\omega)\lambda_n^2>4\lambda_n^2$$

and thus $-(\gamma_{j^m}^m)^2+2\lambda_n|\gamma^m|<0$. Therefore, $\|V^m+r\|_{2,n}$ is strictly decreasing when $m\leq m_2$.

Case (b.1): Suppose $m_1^*<m_2$: By same arguments as in the proof of Lemma 13, we have

$$\|V^{m^*+1}+r\|_{2,n}\precsim_p||V^0+r||_{2,n}^{\frac{1}{1+\zeta^*(c_\phi)-\psi''}}\left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{\zeta^*(c_\phi)-\psi''}{2(1+\zeta^*(c_\phi)-\psi'')}}$$

for any $\psi''>0$ since $\|V^m+r\|_{2,n}$ is strictly decreasing for $m\leq m_2$. As

$$\|r\|_{2,n}\precsim\left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{1}{2}}\precsim\left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{\zeta^*(c_\phi)-\psi''}{2(1+\zeta^*(c_\phi)-\psi'')}},$$

it follows that

$$\begin{aligned}||V^{m_1^*+1}||_{2,n}&\leq||V^{m_1^*+1}+r||_{2,n}+\|r\|_{2,n}\leq||V^{m^*+1}+r||_{2,n}+\|r\|_{2,n}\\ &\precsim_p||V^0+r||_{2,n}^{\frac{1}{1+\zeta^*(c_\phi)-\psi''}}\left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{\zeta^*(c_\phi)-\psi''}{2(1+\zeta^*(c_\phi)-\psi'')}}+\left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{\zeta^*(c_\phi)-\psi''}{2(1+\zeta^*(c_\phi)-\psi'')}}\\ &\precsim_p||V^0+r||_{2,n}^{\frac{1}{1+\zeta^*(c_\phi)-\psi''}}\left(\frac{s\log(2p/\alpha)}{n}\right)^{\frac{\zeta^*(c_\phi)-\delta}{2(1+\zeta^*(c_\phi)-\psi'')}}\end{aligned}$$

which concludes the result.

Case (b.2): Suppose $m_1^* \geq m_2$: We show that this leads to a contradiction. First of all, we show that $m_2 \geq m^*$. We know that

$$||U^m||_{2,n}^2 = \sigma_n^2 + ||V^m + r||_{2,n}^2 + 2 < V^m + r, \varepsilon >_n$$

for any m. Since $||U^{m_2+1}||_{2,n}^2 \leq (1-\delta)\sigma_n^2$, it holds that

$$2 < V^{m_2+1} + r, \varepsilon >_n + ||V^{m_2+1} + r||_{2,n}^2 \leq -\delta\sigma_n^2. \tag{45}$$

Let's prove $m_2 \geq m^*$ by contradiction. Suppose $m_2 < m^*$, it follows that $m_2 + 1 \leq m^* + 1 < M_0$. Since we know that $||V^{m_2+1}||_{2,n} > \eta\sqrt{m_2 + 1 + s}\lambda_n$, it holds

$$||V^{m_2+1} + r||_{2,n}^2 > \left(1 - \frac{\sqrt{C_r}}{\sigma\eta}\right)^2 \|V^{m_2+1}\|_{2,n}^2.$$

Also, $| < r, \varepsilon >_n | \leq \|r\|_{2,n}\sigma_n \to 0$ as $n \to \infty$. Therefore, for η being a large enough fixed constant and n being large enough, we have that:

$$\begin{aligned} &||V^{m_2+1} + r||_{2,n}^2 + 2 < V^{m_2+1}, \varepsilon >_n \\ &\geq \left(1 - \frac{\sqrt{C_r}}{\sigma\eta}\right)^2 ||V^{m_2+1}||_{2,n}^2 - 2\frac{\sqrt{m_2 + 1 + s}}{\sqrt{c_\phi}}\lambda_n \|V^{m_2+1}\|_{2,n} > 0 \end{aligned}$$

and $2 < r, \varepsilon >_n > -\delta\sigma_n^2$ which leads to a contradiction with (45). Thus, it must hold that $m_2 \geq m^*$. Therefore,

$$||V^m + r||_{2,n}^2 \leq ||V^{m^*+1} + r||_{2,n}^2 \leq \left(1 + \frac{\sqrt{C_r}}{\sigma\eta}\right)^2 \eta(m^* + s + 1)\lambda_n^2$$

for any $m^* + 1 \leq m \leq m_2 + 1$. We also know that by assumption,

$$\begin{aligned} (\gamma^m)^2 &\geq c_u \log(2p/\alpha)/n \|U^m\|_{2,n}^2 \geq c_u \log(2p/\alpha)/n(1-\delta)\sigma_n^2 \\ &\geq c_u \log(2p/\alpha)/n(1-\delta)(1-\omega)\sigma^2 \\ &\geq c_u(1-\delta)(1-\omega)\lambda_n^2 \end{aligned}$$

for any $m \leq m_2 \leq m_1^*$. Since

$$||V^m + r||_{2,n}^2 - ||V^{m+1} + r||_{2,n}^2 = (\gamma^m)^2 - 2\gamma^m < X_{j^m}, \varepsilon >_n \geq c_{u_1}\lambda_n^2 > 0$$

for some constant $c_{u_1} > 0$ if $(1-\delta)(1-\omega)c_u/4 > 4$, it follows that

$$||V^m + r||_{2,n}^2 \geq ||V^{m+1} + r||_{2,n}^2 - c_{u_1}\lambda_n^2$$

for $m = m^*, m^* + 1, \ldots, m_2$. Consequently, $||V^{m_2+1} + r||_{2,n}^2 \leq ||V^{m^*+1} + r||_{2,n}^2$. By assumption, at the $(m_2 + 1)^{th}$ step, we know that $||U^{m_2+1}||_{2,n}^2 \leq (1-\delta)\sigma_n^2$. Hence,

$$2 < V^{m_2+1} + r, \varepsilon >_n + \|V^{m_2+1} + r\|_{2,n}^2 \leq -\delta\sigma_n^2. \tag{46}$$

However, $||V^{m_2+1}+r||_{2,n}^2 \leq ||V^{m^*+1}+r||_{2,n}^2$, and therefore, by Lemma 13,

$$2|< V^{m_2+1}+r, \varepsilon >_n| \leq ||V^{m_2+1}+r||_{2,n}\sigma_n \leq ||V^{m^*+1}+r||_{2,n}\sigma_n \to 0,$$

which contradicts (46) when n is large enough. Therefore, Case (b.2) can not happen, and the proof is concluded.

■

Proof [Proof of Theorem 17] WLOG., we can assume that $C_F \log n$ is an integer, so that F_k is a positive integer for $k = 1, 2, \ldots$. For some constant $C^* > 0$ that is large enough, define $k^* := \lfloor \frac{C^* s \log n}{L_n} \rfloor$. We first show that, if we run the algorithms forever, we have with probability $\geq 1-\alpha$:

$$\mathbb{E}_n\left[\left(x_i'(\beta^{m_{k^*}} - \beta_0)\right)^2\right] \precsim_p \frac{s\log(n)\log(p)}{n}$$

and

$$\mathbb{E}_n[\|\beta^{m_{k^*}} - \beta_0\|_2^2] \precsim_p \frac{s\log(n)\log(p)}{n}.$$

Define $\hat{T}^m$ as the set of variables that are selected by the end of time period m. It is worth noting that $|\hat{T}^{m_k}| \leq kL_k$ with $L_k = L_n$ for all $k \in \mathbb{Z}^+$ by construction. Define $\check{m}_k := m_{k-1} + L_k$. Therefore,

$$|\hat{T}^{m_k}| \leq kL_k \leq C^* s \log n,$$

for all $k \leq k^* := \lfloor \frac{C^* s \log n}{L_n} \rfloor$. As in Lemma 24, $\mathcal{P}_{m,T}$ denotes the operator of performing m periods of PBA algorithm subject to the subset T. Further, define $R_k := (\mathcal{P}_{\hat{T}^{\check{m}_k}} - \mathcal{P}_{F_k,\hat{T}^{\check{m}_k}})U^{\check{m}_k}$. By Lemma 24, we have that

$$\begin{aligned}\|R_k\|_{2,n}^2 &= \|(1-\mathcal{P}_{F_k,\hat{T}^{\check{m}_k}})U^{\check{m}_k} - (1-\mathcal{P}_{\hat{T}^{\check{m}_k}})U^{\check{m}_k}\|_{2,n}^2 \\ &\leq \|\mathcal{P}_{\hat{T}^{\check{m}_k}}U^{\check{m}_k}\|_{2,n}^2 \exp\left(-\frac{c_\phi F_k}{|\hat{T}^{\check{m}_k}|}\right) \leq n^{-2}\end{aligned}$$

for $F_k \geq \frac{2}{c_\phi}|\hat{T}^{\check{m}_k}|\log(n\cdot\|y\|_{2,n})$ starting at $U^{\check{m}_k}$. Note that R_k is a small term that could be neglected later with $\|y\|_{2,n} = \|U^0\|_{2,n}^2 \geq \|U^m\|_{2,n}^2$ being a decreasing sequence in m. Note that $U^{\check{m}_k} = y - x'\beta^{\check{m}_k}$, Therefore, $\mathcal{P}_{\hat{T}^{\check{m}_k}}U^{\check{m}_k} = \mathcal{P}_{\hat{T}^{\check{m}_k}}y$ in the projection of vector y on the columns of X with indices in $T^{\check{m}_k}$. As a result, we have

$$U^{m_k} = (I - P_{\hat{T}^{\check{m}_k}})U^{\check{m}_k} + R_k = (I - P_{\hat{T}^{\check{m}_k}})y + R_k \tag{47}$$

with $m_k = \check{m}_k + F_k$. Define $\widetilde{V}^k := V^{m_k} = X(\beta - \beta^{m_k})$ and $\widetilde{V}_o^k = X\beta - P_{\hat{T}^{\check{m}_k}}X\beta = (I - P_{\hat{T}^{\check{m}_k}})X_{\hat{T}^{\check{m}_k}}\beta_{\hat{T}^{\check{m}_k}}$. Therefore,

$$\widetilde{V}^k = X\beta - (y - U^{m_k}) = -(r+\varepsilon) + (I - P_{\hat{T}^{\check{m}_k}})y + R_k = (I - P_{\hat{T}^{\check{m}_k}})X\beta - P_{\hat{T}^{\check{m}_k}}(r+\varepsilon) + R_k$$

It implies that

$$\widetilde{V}^k = \widetilde{V}_o^k + R_k - P_{T^{\check{m}_k}}(r+\varepsilon). \tag{48}$$

By definition, $\|\widetilde{V}_o^k\|_{2,n}^2$ is a decaying sequence in k as $T^{\check{m}_k} \subset T^{\check{m}_{k+1}}$ holds for all $k \geq 0$. We need the following proposition:

Proposition 27 *Suppose all conditions in Theorem 17 hold. For M_n large enough, there exists an absolute constant $C_k > 0$ such that for some $k \leq \lfloor C_k s \log n / L_n \rfloor$ and $k(L_n + 1) < M_n$, we have that:*

$$\mathbb{E}_n[\big(x_i'(\beta^{m_k} - \beta)\big)^2] \precsim_p \frac{s \log(n) \log(p)}{n},$$

and

$$\|\beta^{m_k} - \beta\|_2^2 \precsim_p \frac{s \log(n) \log(p)}{n}.$$

Proof [Proof of Proposition 27] Assume that $k \leq k^*$ with $k^* = \lfloor \frac{C^* s \log n}{L_n} \rfloor$. Define $\widetilde{T}^{m_k} := T_0 \backslash \hat{T}^{m_k}$. That said, $\widetilde{T}^{m_k}$ is the set of variables that are in T_0 but not yet being selected by time m_k. Then, if $\widetilde{T}^{m_k} = \emptyset$, then, all the variables in T_0 are already selected. By definition,

$$V^{m_k} = y - U^{m_k} - X\beta = y - (I - \mathcal{P}_{F_k, \hat{T}^{\check{m}_k}}) U^{\check{m}_k} - X_{\hat{T}^{\check{m}_k}} \beta_{\hat{T}^{\check{m}_k}}.$$

It implies that:

$$\begin{aligned}
\|V^{m_k}\|_{2,n}^2 &= \|X_{\hat{T}^{\check{m}_k}} \beta_{\hat{T}^{\check{m}_k}} - y + (I - \mathcal{P}_{F_k, \hat{T}^{\check{m}_k}}) U^{\check{m}_k}\|_{2,n}^2 \\
&\leq 2\|X_{\hat{T}^{\check{m}_k}} \beta_{\hat{T}^{\check{m}_k}} - y + (I - \mathcal{P}_{\hat{T}^{\check{m}_k}}) U^{\check{m}_k}\|_{2,n}^2 + 2\|(\mathcal{P}_{\hat{T}^{\check{m}_k}} - \mathcal{P}_{F_k, \hat{T}^{\check{m}_k}}) U^{\check{m}_k}\|_{2,n}^2 \\
&\leq 2\|\mathcal{P}_{\hat{T}^{\check{m}_k}} (r_n + \varepsilon)\|_{2,n}^2 + 2\|(\mathcal{P}_{\hat{T}^{\check{m}_k}} - \mathcal{P}_{F_k, \hat{T}^{\check{m}_k}}) U^{\check{m}_k}\|_{2,n}^2 \\
&\leq \frac{(4C_r \log(2p/\alpha) + 4\sigma^2 C^*/c_\phi \log(2p/\alpha) \log(n)) s + 1}{n} \qquad (49)
\end{aligned}$$

by Lemma 25 since $\hat{T}^{\check{m}_k} \leq C^* s \log(n)$ and by Lemma 24 since $F_k > C_F |\hat{T}^{\check{m}_k}| \log n$ for some C_T large enough, which concludes the result. Now, suppose $\widetilde{T}^{m_k} \neq \emptyset$ for $k = 0, 1, ..., k^*$. Recall that

$$U^{m_k} := (I - \mathcal{P}_{\hat{T}^{\check{m}_k}}) y + R_k,$$

with $\|R_k\|_{2,n} \leq \frac{1}{n}$ and

$$V^{m_k} = (I - \mathcal{P}_{\hat{T}^{\check{m}_k}}) X\beta - \mathcal{P}_{\hat{T}^{\check{m}_k}} (r + \varepsilon) + R_k = (I - \mathcal{P}_{\hat{T}^{\check{m}_k}}) X_{\widetilde{T}^{\check{m}_k}} \beta_{\widetilde{T}^{\check{m}_k}}.$$

Suppose that $\|V^{m_k}\|_{2,n}^2 > \frac{C_V s \log(n) \log(2p/\alpha)}{n}$ for all $k \leq \frac{C_k s \log(n)}{L_n}$, where $C_k \leq C_M$ is a large enough constant. For C_V large enough, we have that $\|V^m\|_{2,n} > \sqrt{\frac{C_r c_\phi s \log(2p/\alpha)}{n}}$. By (48) and Lemma 25, it follows that

$$\begin{aligned}
\|\widetilde{V}_o^k\|_{2,n} &\geq \|V^{m_k}\|_{2,n} - \|\mathcal{P}_{\hat{T}^{\check{m}_k}} (r + \varepsilon)\|_{2,n} - \|R_k\|_{2,n} \\
&\geq \sqrt{\frac{C_V s \log(2p/\alpha) \log(n)}{n}} - \sqrt{\frac{2C_r s \log(2p/\alpha) + 2\sigma^2/c_\phi |\hat{T}^{\check{m}_k}| \log(2p/\alpha)}{n}} - \frac{1}{n} \\
&\geq \sqrt{\frac{C_V/2 s \log(2p/\alpha) \log(n)}{n}} \qquad (50)
\end{aligned}$$

given that $C_V > 2(2C_r + 2\sigma^2/c_\phi C_k + 1)$ is large enough fixed constant where we used that $|\hat{T}^{\check{m}_k}| \leq L_n k \leq C_k s \log(n)$. Define $\omega_n := \frac{c_\phi^2}{4L_n C_\phi^3}$. Next, we divide our discussions into two

cases:

Case (A1): $\sum_{t=m_{k-1}+1}^{\check{m}_k} \gamma_{j^t}^2 \geq \omega_n \|\widetilde{V}_o^{k-1}\|_{2,n}^2$. By definition, we have that:

$$\|U^{m_k}\|_{2,n}^2 \leq \|U^{\check{m}_k}\|_{2,n}^2 \leq \|U^{m_{k-1}}\|_{2,n}^2 - \sum_{t=m_{k-1}+1}^{\check{m}_k} \gamma_{j^t}^2 \leq \|U^{m_{k-1}}\|_{2,n}^2 - \omega_n \|\widetilde{V}_o^{k-1}\|_{2,n}^2. \quad (51)$$

Case (A2): $\sum_{t=m_{k-1}+1}^{\check{m}_k} \gamma_{j^t}^2 \leq \omega_n \|\widetilde{V}_o^{k-1}\|_{2,n}^2$. Define $\delta_{m,k} := \sum_{t=m_{k-1}+1}^{m} \gamma_{j^t} e_{j^t}$, where e_{j^t} is the unit vector that is equal to 1 only at the j^t-entry, and 0 otherwise, for $m \geq m_{k-1}$ and $m \leq \check{m}_{k-1}$. It holds

$$\|\delta_{m,k}\|_{2,n}^2 \leq L_n \sum_{t=m_{k-1}+1}^{m} \gamma_{j^t}^2 \leq L_n \omega_n \|\widetilde{V}_o^{k-1}\|_{2,n}^2. \quad (52)$$

As a result, by definition, for any $m = m_{k-1}, ..., \check{m}_k - 1$, we have that:

$$\begin{aligned}
|\gamma_{j^m}| \|\beta_{\widetilde{T}^{m_{k-1}}}\|_1 \geq & < U^m, X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}} >_n \\
= & \underbrace{< (I - \mathcal{P}_{\hat{T}^{m_{k-1}}}) X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}}, X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}} >_n}_{\Psi_{V,1}} \\
& + \underbrace{< (I - \mathcal{P}_{\hat{T}^{m_{k-1}}})(r + \varepsilon), X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}} >_n}_{\Psi_{V,2}} \\
& + \underbrace{< R_k, X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}} >_n}_{\Psi_{V,3}} \\
& + \underbrace{< X\delta_{m,k}, X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}} >_n}_{\Psi_{V,4}}
\end{aligned}$$

since

$$\begin{aligned}
U^m &= U^{m_k} + X\delta_{m,k} = (I - \mathcal{P}_{\hat{T}^{m_{k-1}}} + R_k)Y + X\delta_{m,k} \\
&= (I - \mathcal{P}_{\hat{T}^{m_{k-1}}}) X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}} + (I - \mathcal{P}_{\hat{T}^{m_{k-1}}})(r + \varepsilon) + R_k + X\delta_{m,k}.
\end{aligned}$$

For $\Psi_{V,1}$, we have that:

$$\Psi_{V,1} = \|(I - \mathcal{P}_{\hat{T}^{m_{k-1}}}) X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2. \quad (53)$$

There exists a $\delta \in \mathbb{R}^p$ with $supp(\delta) \subset \hat{T}^{m_{k-1}}$ such that $X\delta = \mathcal{P}_{\hat{T}^{m_{k-1}}} X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}}$. By definition, $\widetilde{T}^{m_{k-1}} \cap \hat{T}^{m_{k-1}} = \emptyset$, we have that:

$$\begin{aligned}
\Psi_{V,1} &= \|X\delta - X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2 \geq c_\phi (\|\beta_{\widetilde{T}^{m_{k-1}}}\|^2 + \|\delta\|^2) \\
&\geq c_\phi \|\beta_{\widetilde{T}^{m_{k-1}}}\|^2 \geq \frac{c_\phi}{C_\phi} \|X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}}\|^2. \quad (54)
\end{aligned}$$

By (50), we have that

$$\|X_{\widetilde{T}^{m_{k-1}}} \beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2 \geq \|\widetilde{V}_o^{k-1}\|_{2,n}^2 \geq \frac{C_V s \log(2p/\alpha) \log(n)}{2n}.$$

For $\Psi_{V,2}$, we have that:

$$\begin{aligned}|\Psi_{V,2}| &= | < (I - \mathcal{P}_{\hat{T}^{m_{k-1}}})r, X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}} >_n + < \varepsilon, X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}} >_n \\ &\quad - < \mathcal{P}_{\hat{T}^{m_{k-1}}}\varepsilon, X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}} >_n | \\ &\leq \|r\|_{2,n}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n} + \|\beta_{\widetilde{T}^{m_{k-1}}}\||\widetilde{T}^{m_{k-1}}|^{\frac{1}{2}}\lambda_n + \sqrt{\frac{|\hat{T}^{m_{k-1}}|}{c_\phi}}\lambda_n\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n} \\ &\leq \left(\sqrt{\frac{4C_r}{C_V\log(n)}} + \sqrt{\frac{\sigma^2 C_\phi}{C_V\log(n)}} + \sqrt{\frac{\sigma^2 C_k}{c_\phi C_V}}\right)\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2. \end{aligned} \tag{55}$$

For $\Psi_{V,3}$, we have that:

$$\Psi_{V,3} \geq -\|R_k\|_{2,n}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n} \geq -\sqrt{\frac{2}{sn\log(2p/\alpha)\log(n)C_V}}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2. \tag{56}$$

For $\Psi_{V,4}$, we have that:

$$\Psi_{V,4} \leq \|X\delta_{m,k}\|_{2,n}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n} \leq \sqrt{C_\phi L_n\omega_n}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2. \tag{57}$$

Combining (54), (55), (56) and (57), we have that:

$$\begin{aligned}|\gamma_{j^m}|\|\beta_{\widetilde{T}^{\check{m}_{k-1}}}\|_1 &\geq \Big(\frac{c_\phi}{C_\phi} - \sqrt{\frac{4C_r}{C_V\log(n)}} - \sqrt{\frac{\sigma^2 C_\phi}{C_V\log(n)}} - \sqrt{\frac{\sigma^2 C_k}{c_\phi C_V}} \\ &\quad - \sqrt{\frac{2}{sn\log(2p/\alpha)\log(n)C_V}} - \sqrt{C_\phi L_n\omega_n}\Big)\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2 \\ &\geq \frac{c_\phi}{4C_\phi}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2, \end{aligned} \tag{58}$$

given that $C_V > \frac{17C_\phi^2\sigma^2 C_k}{c_\phi^3}$ and n being large enough. As a result, we have that:

$$\begin{aligned}|\gamma_{j^m}| &\geq \frac{c_\phi}{4C_\phi}\frac{\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2}{\|\beta_{\widetilde{T}^{m_{k-1}}}\|_1} \geq \frac{c_\phi}{4C_\phi}\frac{\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2}{\sqrt{s}\|\beta_{\widetilde{T}^{m_{k-1}}}\|} \\ &\geq \frac{c_\phi^{3/2}}{4C_\phi\sqrt{s}}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}, \end{aligned} \tag{59}$$

where the last inequality follows from the sparse eigenvalue condition and $|\widetilde{T}^{m_{k-1}}| \leq s$. Therefore, we have that:

$$\sum_{m=m_{k-1}+1}^{\check{m}_{k-1}} |\gamma_{j^m}|^2 \geq \frac{L_n c_\phi^3}{16C_\phi^2 s}\|X_{\widetilde{T}^{m_{k-1}}}\beta_{\widetilde{T}^{m_{k-1}}}\|_{2,n}^2 \geq \frac{L_n c_\phi^3}{16C_\phi^2 s}\|\widetilde{V}_o^{k-1}\|_{2,n}^2. \tag{60}$$

Combining both cases (A1) and (A2), we have that:

$$\|U^{m_k}\|_{2,n}^2 \leq \|U^{m_{k-1}}\|_{2,n}^2 - \kappa_n\|\widetilde{V}_o^{k-1}\|_{2,n}^2 \tag{61}$$

where $\kappa_n := \min\left(\frac{c_\phi^2}{4L_nC_\phi^3}, \frac{L_nc_\phi^3}{16C_\phi^2 s}\right) \geq \kappa\frac{L_n}{s}$ with $\kappa := \min\left(\frac{c_\phi^2}{4K_L^2C_\phi^3}, \frac{c_\phi^3}{16C_\phi^2}\right)$ given that $L_n \leq K_L\sqrt{s}$ with K_L being a fixed constant, and hence, $\kappa > 0$ is a fixed constant. Define $q := \lceil \frac{s}{L_n} \rceil$ as the smallest integer $\geq \frac{s}{L_n}$.
WLOG., we can assume that $q = \frac{s}{L_n}$ for simplicity, i.e., $\frac{s}{L_n}$ is an integer. By applying (61) multiple times, we have that:

$$\|U^{m_{k+q}}\|_{2,n}^2 \leq \|U^{m_k}\|_{2,n}^2 - \kappa\frac{L_n}{s}\sum_{l=k}^{k+q-1}\|\widetilde{V}_o^l\|_{2,n}^2. \tag{62}$$

We again divide our analysis into two cases: Assume that $k \leq k+q \leq \frac{Cs\log(n)}{L_n}$.

Case (B1): $\|\widetilde{V}_o^{k+q-1}\|_{2,n}^2 > \frac{1}{2}\|\widetilde{V}_o^k\|_{2,n}^2$. It implies that $\|\widetilde{V}_o^l\|_{2,n}^2 \geq \frac{1}{2}\|\widetilde{V}_o^k\|_{2,n}^2$ for all $l = k, \ldots, k+q-1$ as

$$\|\widetilde{V}_o^l\|_{2,n}^2 = \|(I - \mathcal{P}_{\hat{T}^{\check{m}_l}})X\beta\|_{2,n}^2$$

is a decreasing sequence in l. It follows that

$$\begin{aligned}\|U^{m_{k+q}}\|_{2,n}^2 &\leq \|U^{m_k}\|_{2,n}^2 - \frac{\kappa L_n}{s}\sum_{l=k}^{k+q-1}\|\widetilde{V}_o^l\|_{2,n}^2 \\ &\leq \|U^{m_k}\|_{2,n}^2 - \frac{\kappa q L_n}{2s}\|\widetilde{V}_o^k\|_{2,n}^2 = \|U^{m_k}\|_{2,n}^2 - \frac{\kappa}{2}\|\widetilde{V}_o^k\|_{2,n}^2.\end{aligned}$$

Recall that by (47), we have:

$$U^{m_k} = \widetilde{V}_o^k + (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})(r+\varepsilon) + R_k.$$

It implies that

$$\begin{aligned}\|\widetilde{V}_o^{k+q}\|_{2,n}^2 \leq &\left(1 - \frac{\kappa}{2}\right)\|\widetilde{V}_o^k\|_{2,n}^2 + 2 < \widetilde{V}_o^k, (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})(r+\varepsilon) >_n \\ &- 2 < \widetilde{V}_o^{k+q}, (I - \mathcal{P}_{\hat{T}^{\check{m}_{k+q}}})(r+\varepsilon) >_n \\ &+ \|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}})(r+\varepsilon)\|_{2,n}^2 - \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k+q}}})(r+\varepsilon)\|_{2,n}^2 + \|R_k\|_{2,n}^2 - \|R_{k+q}\|_{2,n}^2 \\ &+ 2 < \widetilde{V}_o^k + (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})(r+\varepsilon), R_k >_n \\ &- 2 < \widetilde{V}_o^{k+q} + (I - \mathcal{P}_{\hat{T}^{\check{m}_{k+q}}})(r+\varepsilon), R_{k+q} >_n .\end{aligned}$$

By (55), replacing $k-1$ by k, for n large enough, we have

$$\begin{aligned}&2 < \widetilde{V}_o^k, (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})(r+\varepsilon) >_n - 2 < \widetilde{V}_o^{k+q}, (I - \mathcal{P}_{\hat{T}^{\check{m}_{k+q}}})(r+\varepsilon) >_n \\ &\leq 4\left(\sqrt{\frac{4C_r}{C_V\log(n)}} + \sqrt{\frac{\sigma^2C_\phi}{C_V\log(n)}} + \sqrt{\frac{\sigma^2C_k}{c_\phi C_V}}\right)\|X_{\widetilde{T}^{m_k}}\beta_{\widetilde{T}^{m_k}}\|_{2,n}^2 \\ &\leq \frac{4C_\phi}{c_\phi}\sqrt{\frac{2\sigma^2C_k}{c_\phi C_V}}\|\widetilde{V}_o^k\|_{2,n}^2 \leq \frac{\kappa}{12}\|\widetilde{V}_o^k\|_{2,n}^2\end{aligned} \tag{63}$$

given that $C_V > \frac{64 \cdot 36 \cdot 2 C_\phi^2 \sigma^2 C_k}{c_\phi^3 \kappa^2}$. Here we used that $\widetilde{V}_o^k := (1 - \mathcal{P}_{\hat{T}^{m_k}}) X\beta = (1 - \mathcal{P}_{\hat{T}^{m_k}}) X_{\widetilde{T}^{m_k}} \beta_{\widetilde{T}^{m_k}}$ as $\widetilde{T}^{m_k} = T_0 \backslash \hat{T}^{m_k}$. Again, we can find a δ with $supp(\delta) \subset \hat{T}^{m_k}$ such that $X\delta = \mathcal{P}_{\hat{T}^{m_k}} X_{\widetilde{T}^{m_k}} \beta_{\widetilde{T}^{m_k}}$. Hence, the supports of δ and $\beta_{\widetilde{T}^{m_k}}$ have no intersection and we conclude

$$\|\widetilde{V}_o^k\|_{2,n}^2 \geq c_\phi (\|\delta\|^2 + \|\beta_{\widetilde{T}^{m_k}}\|^2) \geq c_\phi \|\beta_{\widetilde{T}^{m_k}}\|^2 \geq \frac{c_\phi}{C_\phi} \|X_{\widetilde{T}^{m_k}} \beta_{\widetilde{T}^{m_k}}\|_{2,n}^2.$$

By Lemma 25 and (50), for n large enough, we have that:

$$\begin{aligned} & \|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}})(r + \varepsilon)\|_{2,n}^2 - \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k+q}}})(r + \varepsilon)\|_{2,n}^2 \qquad (64) \\ & = -\|P_{\hat{T}^{\check{m}_k}}(r + \varepsilon)\|_{2,n}^2 + \|\mathcal{P}_{\hat{T}^{\check{m}_{k+q}}}(r + \varepsilon)\|_{2,n}^2 \\ & \leq 2 \left(\frac{C_r s \log(2p/\alpha) + \sigma^2 / c_\phi \log(2p/\alpha) |\hat{T}^{\check{m}_{k+q}}|}{n} \right) \\ & \leq \frac{5 C_k \sigma^2}{c_\phi C_V} \|\widetilde{V}_o^k\|_{2,n}^2 \leq \frac{\kappa}{12} \|\widetilde{V}_o^k\|_{2,n}^2 \qquad (65) \end{aligned}$$

given that $C_V \geq \frac{60 C_k \sigma^2}{c_\phi \kappa}$. And for n large enough, for fixed C_V, by (50), we have that:

$$\begin{aligned} & \|R_k\|_{2,n}^2 - \|R_{k+q}\|_{2,n}^2 + 2 < \widetilde{V}_o^k + (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})(r + \varepsilon), R_k >_n \\ & - 2 < \widetilde{V}_o^{k+q} + (I - \mathcal{P}_{\hat{T}^{\check{m}_{k+q}}})(r + \varepsilon), R_{k+q} >_n \\ & \leq \frac{1}{n} + \frac{4}{n} \|\widetilde{V}_o^k\|_{2,n} + \frac{4}{n} \|r + \varepsilon\|_{2,n} \leq \frac{\kappa}{12} \|\widetilde{V}_n^k\|_{2,n}^2. \qquad (66) \end{aligned}$$

Combining (63), (65) and (66), we have that:

$$\|\widetilde{V}_o^{k+q}\|_{2,n}^2 \leq \left(1 - \frac{\kappa}{4}\right) \|\widetilde{V}_o^k\|_{2,n}^2. \qquad (67)$$

Hence, we have shown exponential decay of $\|\widetilde{V}_o^k\|_{2,n}^2$ in Case (B1).

Case (B2): $\|\widetilde{V}_o^{k+q}\|_{2,n}^2 \leq \frac{1}{2} \|\widetilde{V}_o^k\|_{2,n}^2$. This automatically proves exponential decay. Note that $\kappa \leq \frac{1}{4}$ as $C_\phi \geq c_\phi$. Combining Case (B1) and Case (B2), we have that

$$\|\widetilde{V}_o^{k+q}\|_{2,n}^2 \leq \left(1 - \frac{\kappa}{4}\right) \|\widetilde{V}_o^k\|_{2,n}^2.$$

Therefore, for any $kq \leq C_k s \log(n)$, we have that

$$\|\widetilde{V}_o^{kq}\|_{2,n}^2 \leq \|\widetilde{V}_o^0\|_{2,n}^2 \left(1 - \frac{\kappa}{4}\right)^k. \qquad (68)$$

Since $q = \frac{s}{L_n}$, let $k = \lfloor C_k L_n \log(n) \rfloor$. WLOG., assume that $C_k L_n \log(n)$ is an integer so that $k = C_k n \log(n)$. For n large enough and $C_k \geq \frac{4K}{L_n \kappa}$, we have that

$$\|\widetilde{V}_o^{kq}\|_{2,n}^2 \leq n^{C_k L_n \ln(1 - \frac{\kappa}{4})} \|\widetilde{V}_o^0\|_{2,n}^2 \leq n^{-C_k L_n \kappa / 4} n^{K-1} < \frac{C_V s \log(n) \log(p)}{4n} \qquad (69)$$

given that $\|\widetilde{V}_o^0\|_{2,n}^2 \leq n^{K-1}$ for some fixed constant $K > 0$. This leads to a contradictory to (50). Therefore, for given constant $C_k \geq \frac{4K}{L_n \kappa}$, there must exists $k \leq \frac{C_k s \log(n)}{L_n}$ such that

$$\|V^{m_k}\|_{2,n}^2 \leq \frac{C_V s \log(n) \log(2p/\alpha)}{n} \tag{70}$$

which concludes our proof. ■

By construction of Algorithm 2, we have that

$$\|U^{m+1}\|_{2,n}^2 = \|U^m\|_{2,n}^2 - \gamma_{j^m}^2$$

for all non-negative integers m. Therefore,

$$\frac{\|U^{m+1}\|_{2,n}^2}{\|U^m\|_{2,n}^2} = 1 - \frac{\gamma_{j^m}^2}{\|U^m\|_{2,n}^2}. \tag{71}$$

It implies that the algorithm will not stop if

$$\frac{\gamma_{j^m}^2}{\|U^m\|_{2,n}^2} \geq C_u \frac{\log(2p/\alpha)}{n}. \tag{72}$$

Suppose that the algorithm does not stop when $k \leq \frac{C' s \log n}{L_n} \leq k_M := \frac{C_M s \log n}{L_n}$ with C_M defined in Assumption 1. If $C' > C_k$, by Proposition 27, it holds

$$\mathbb{E}_n \left[\left(x_i'(\beta^{m_k} - \beta) \right)^2 \right] \precsim_p \frac{s \log(n) \log(p)}{n}$$

for some $k \leq C_k s \log n / L_n$. By Proposition 27, there also exists a positive constant C_V such that

$$\|\widetilde{V}^{k^*}\|_{2,n} \leq \frac{C_V s \log(2p/\alpha) \log(n)}{n} \tag{73}$$

with $k^* \leq C^* s \log n / L_n$. It implies that $|\hat{T}^{\check{m}_{k^*}}| = |\hat{T}^{m_{k^*}}| \leq C^* s \log n$. Therefore, by Lemma 25, we have that

$$\begin{aligned} \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}}) X\beta\|_{2,n} &\leq \|\widetilde{V}^{k^*}\|_{2,n} + \|\mathcal{P}_{\hat{T}^{\check{m}_{k^*}}}(r + \varepsilon)\|_{2,n} + \|R_{k^*}\|_{2,n} \\ &\leq \left(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1 \right) \sqrt{\frac{s \log n \log(2p/\alpha)}{n}}. \end{aligned} \tag{74}$$

It implies that for any $k \geq k^*$ and $k \leq k_M$, we have that

$$\begin{aligned} \|\widetilde{V}^k\|_{2,n} &\leq \|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}}) X\beta\|_{2,n} + \|\mathcal{P}_{\hat{T}^{\check{m}_k}}(r + \varepsilon)\|_{2,n} + \|R_k\|_{2,n} \\ &\leq \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}}) X\beta\|_{2,n} + \sqrt{\frac{2(C_r s \log(2p/\alpha) + k\sigma^2/c_\phi \log(2p/\alpha) s \log n)}{n}} + \frac{1}{n} \\ &\leq \left(C_{k^*} + 1 + 2(C_r + k\sigma^2/c_\phi)^{\frac{1}{2}} \right) \sqrt{\frac{s \log n \log(2p/\alpha)}{n}} \end{aligned}$$

where $C_k^* := (C_V^{\frac{1}{2}} + (2C_r + \sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1)$. By definition, with $\frac{s \log n \log(2p/\alpha)}{n} \to 0$, for n large enough, for arbitrarily small $\eta > 0$, we have that

$$\|U^{m_k}\|_{2,n} = \|(\varepsilon + r) + \widetilde{V}^k\|_{2,n} \geq \|(\varepsilon + r)\|_{2,n} - \|\widetilde{V}^k\|_{2,n} \geq \sqrt{(1-2\eta)\sigma^2} \tag{75}$$

and

$$\|U^{m_k}\|_{2,n} = \|(\varepsilon + r) + \widetilde{V}^k\|_{2,n} \leq \|(\varepsilon + r)\|_{2,n} + \|\widetilde{V}^k\|_{2,n} \leq \sqrt{(1+2\eta)\sigma^2}. \tag{76}$$

By definition, since $\|U^{m_k}\|_{2,n}$ is a decreasing sequence, we have that:

$$\gamma_{j^m}^2 \geq C_U \frac{\log(2p/\alpha)}{n} \|U^m\|_{2,n}^2 \geq C_U(1-2\eta)\sigma^2 \frac{\log(2p/\alpha)}{n}.$$

As a result, we have that:

$$\begin{aligned} \|U^{m_k}\|_{2,n}^2 &\leq \|U^{m_{k^*}}\|_{2,n}^2 - \sum_{m_k \leq m \leq m_{k^*}-1,\ l_m = 1} \gamma_{j^m}^2 \\ &\leq \|U^{m_{k^*}}\|_{2,n}^2 - (k - k^*) L_n C_U (1-2\eta)\sigma^2 \frac{\log(2p/\alpha)}{n}. \end{aligned} \tag{77}$$

Therefore, it implies that for any $k \geq k^*$, we have:

$$\begin{aligned} \|U^{m_k}\|_{2,n}^2 - \|U^{m_{k^*}}\|_{2,n}^2 &= \|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}})y + R_k\|_{2,n}^2 - \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})y + R_{k^*}\|_{2,n}^2 \\ &= \underbrace{\|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}})y\|_{2,n}^2 - \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})y\|_{2,n}^2}_{\Psi_1} \\ &\quad + \underbrace{2 < (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})y, R_k >_n - < (I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})y, R_{k^*} >_n}_{\Psi_2} \\ &\quad + \underbrace{\|R_k\|_{2,n}^2 - \|R_{k^*}\|_{2,n}^2}_{\Psi_3}. \end{aligned} \tag{78}$$

For Ψ_1, we have that:

$$\begin{aligned} \Psi_1 &= \|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}})y\|_{2,n}^2 - \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})y\|_{2,n}^2 \\ &= \underbrace{\|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}})X\beta\|_{2,n}^2 - \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})X\beta\|_{2,n}^2}_{\Psi_{11}} \\ &\quad \underbrace{-2 < (\mathcal{P}_{\hat{T}^{\check{m}_k}} - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})X\beta, (r + \varepsilon)) >_n}_{\Psi_{12}} \\ &\quad + \underbrace{\|\mathcal{P}_{\hat{T}^{\check{m}_{k^*}}}(r+\varepsilon)\|_{2,n}^2 - \|\mathcal{P}_{\hat{T}^{\check{m}_k}}(r+\varepsilon)\|_{2,n}^2}_{\Psi_{13}}. \end{aligned}$$

For Ψ_{11}, by (74), we have that:

$$\Psi_{11} \geq -\|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})X\beta\|_{2,n}^2 \geq -\left(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1\right)^2 \frac{s \log n \log(2p/\alpha)}{n}. \tag{79}$$

For Ψ_{12}, we know that

$$(\mathcal{P}_{\hat{T}^{\check{m}_k}} - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})X\beta = (I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})X\beta - (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})X\beta.$$

It is worth noting that $(I-\mathcal{P}_{\hat{T}^{\check{m}_k}})X\beta$ is a linear combination of columns of X with indices in $T_k := \hat{T}^{\check{m}_k} \cup T_0$ where $T_0 := supp(\beta)$. Therefore, we can find a $\zeta_k \in \mathbb{R}^p$ with $supp(\zeta_k) \subset T_k$, with $X\zeta_k = (I - \mathcal{P}_{\hat{T}^{\check{m}_k}})X\beta$. Since

$$\begin{aligned}\|(I - \mathcal{P}_{\hat{T}^{\check{m}_k}})X\beta\|_{2,n}^2 &\leq \|(I - \mathcal{P}_{\hat{T}^{\check{m}_{k^*}}})X\beta\|_{2,n}^2 \\ &\leq \left(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1\right)^2 \frac{s\log n \log(2p/\alpha)}{n},\end{aligned}$$

we have that

$$\|X\zeta_k\|_{2,n}^2 \leq \left(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1\right)^2 \frac{s\log n \log(2p/\alpha)}{n}. \tag{80}$$

By sparse eigenvalue condition in Assumption A.1, it implies that

$$\|\zeta_k\|_{2,n}^2 \leq \frac{1}{c_\phi}\left(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1\right)^2 \frac{s\log n \log(2p/\alpha)}{n}.$$

Therefore,

$$\begin{aligned}\Psi_{12} &= 2 < X\zeta_k - X\zeta_{k^*}, (r+\varepsilon) >_n \\ &= 2 < X\zeta_k - X\zeta_{k^*}, r >_n + 2 < \zeta_k - \zeta_{k^*}, X_{T_{k^*}}\varepsilon >_n \\ &\geq -2(\|X\zeta_k\|_{2,n} + \|X\zeta_{k^*}\|_{2,n})\|r\|_{2,n} - 2(\|\zeta_k\|_{2,n} + \|\zeta_{k^*}\|_{2,n})|T_{k^*}|^{\frac{1}{2}}\lambda_n \\ &\geq -4\sqrt{C_r}\left(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1\right)\frac{s\log(n)^{\frac{1}{2}}\log(2p/\alpha)}{n} \\ &\quad - \frac{4\sigma\sqrt{C^*}}{\sqrt{c_\phi}}\left(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1\right)\frac{s\log(n)\log(2p/\alpha)}{n}.\end{aligned} \tag{81}$$

For Ψ_{13}, by Lemma 25, we have that:

$$\Psi_{13} \geq -\|\mathcal{P}_{\hat{T}^{\check{m}_k}}(r+\varepsilon)\|_{2,n}^2 \geq -2\left(\frac{C_r s\log(2p/\alpha) + \sigma^2/c_\phi \log(2p/\alpha)kL_n}{n}\right). \tag{82}$$

Combining (79), (81) and (82), we have that:

$$\Psi_1 \geq -C_{\Psi_1}\frac{s\log n\log(2p/\alpha)}{n} - \frac{2\sigma^2}{c_\phi}\frac{kL_n\log(2p/\alpha)}{n} \tag{83}$$

where $C_{\Psi_1} := (4\sqrt{C_r} + 2C_r + 4\sigma\sqrt{C^*}/\sqrt{c_\phi} + (C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1))(C_V^{\frac{1}{2}} + (2C_r + 2\sigma^2/c_\phi C^*)^{\frac{1}{2}} + 1)$ is a fixed constant. For Ψ_2 and Ψ_3, since $\|R_k\|, \|R_{k^*}\| \leq \frac{1}{n}$, and by (76), we have that for n large enough:

$$\Psi_2 \leq -\frac{4(1+2\eta)\sigma}{n} \quad \text{and} \quad \Psi_3 \leq -\frac{2}{n^2}. \tag{84}$$

Therefore, we have that:

$$\|U^{m_k}\|_{2,n}^2 - \|U^{m_{k^*}}\|_{2,n}^2 \geq -C_{\Psi_1}\frac{s\log n\log(2p/\alpha)}{n} - \frac{\sigma^2}{c_\phi}\frac{kL_n\log(2p/\alpha)}{n} - \frac{4(1+2\eta)\sigma}{n} - \frac{2}{n^2}. \quad (85)$$

Recall that in equation (77), we have that

$$\|U^{m_k}\|_{2,n}^2 - \|U^{m_{k^*}}\|_{2,n}^2 \leq -(k-k^*)L_nC_U(1-2\eta)\sigma^2\frac{\log(2p\alpha)}{n}. \quad (86)$$

Therefore, combining (85) and (86), we have that

$$\begin{aligned}&(k-k^*)L_nC_U(1-2\eta)\sigma^2\frac{\log(2p\alpha)}{n}\\ &\leq C_{\Psi_1}\frac{s\log n\log(2p/\alpha)}{n} + \frac{\sigma^2}{c_\phi}\frac{kL_n\log(2p/\alpha)}{n} + \frac{4(1+2\eta)\sigma}{n} + \frac{2}{n^2}\end{aligned}$$

and by assumption that $C_U > \frac{4}{c_\phi}$, for small enough $\eta > 0$, it holds $C_U(1-2\eta) - \frac{4}{c_\phi} > 0$. Since $k^* \leq \frac{C^* s\log n}{L_n}$, it implies that for n large enough, we have that

$$k \leq \frac{k^*C_U(1-2\eta) + C_{\Psi_1}s\log n/L_n + 1}{C_U(1-2\eta) - \frac{4}{c_\phi}} \leq C'\frac{s\log n}{L_n}$$

where $C' := \frac{C^*C_U(1-2\eta)+C_{\Psi_1}+1}{C_U(1-2\eta)-\frac{4}{c_\phi}} s\log n/L_n$. By assumption that $C' \leq C_M$, all the analysis in the above that uses the sparse eigenvalue condition applies. Therefore, the algorithm must stop before m_k with $k \leq C'$. Therefore, when we stop at m^*, we have that

$$\gamma_{j^{m^*}} \leq \sqrt{C_U\frac{\log(2p/\alpha)}{n}}\|U^{m^*}\|_{2,n} \leq \sqrt{C_U\frac{\log(2p/\alpha)}{n}}(\|V^{m^*}\|_{2,n} + \|(r+\varepsilon)\|_{2,n}). \quad (87)$$

By Lemma 26, we have that

$$|\gamma_{j^{m^*}}| \geq \sqrt{\frac{c_\phi}{(|\hat{T}^{m^*}|+s)}}\|V^{m^*}\|_{2,n} - \sqrt{\frac{C_rc_\phi s\log 2p/\alpha}{n(|\hat{T}^{m^*}|+s)}} - \lambda_n.$$

It implies that

$$\begin{aligned}&\|V^{m^*}\|_{2,n}\left(\sqrt{c_\phi} - \sqrt{C_U\frac{(|\hat{T}^{m^*}|+s)\log(2p/\alpha)}{n}}\right)\\ &\leq \sqrt{C_U\frac{(|\hat{T}^{m^*}|+s)\log(2p/\alpha)}{n}}(2+2\eta)\sigma + \sqrt{\frac{C_rc_\phi s\log(2p/\alpha)}{n}} + \sqrt{(|\hat{T}^{m^*}|+s)}\lambda_n.\end{aligned}$$

Since $|\hat{T}^{m^*}| \leq C's\log(n)$ and $\frac{s\log(n)\log(p)}{n} \to 0$, we have that for n large enough,

$$\sqrt{\frac{c_\phi}{2}}\|V^{m^*}\|_{2,n} \leq \sqrt{\frac{s\log(n)\log(2p/\alpha)}{n}}\left(\sqrt{C_UC''}(2+2\eta)\sigma + \sqrt{C_rc_\phi} + \sigma C''\right)$$

for $C'' \leq C' + 1$. It follows that

$$\|V^{m^*}\|_{2,n}^2 \leq \frac{2}{c_\phi} \left(\sqrt{C_U C''}(2+2\eta)\sigma + \sqrt{C_r c_\phi} + \sigma C'' \right)^2 \frac{s \log(n) \log(2p/\alpha)}{n},$$

and

$$\|\beta - \beta^{m^*}\|^2 \leq \frac{2}{c_\phi^2} \left(\sqrt{C_U C''}(2+2\eta)\sigma + \sqrt{C_r c_\phi} + \sigma C'' \right)^2 \frac{s \log(n) \log(2p/\alpha)}{n}$$

follows by sparse eigenvalue condition which concludes the proof. ■

Appendix D. Discussion of the equi-correlated design

In the equi-correlated design, we assume to have a set of predictors $X_1, \ldots, X_p$, such that

$$Cov_n(X_i, X_j) = corr_n(X_i, X_j) = \rho \in (0, 1)$$

for $i \neq j$, where Cov_n represents the empirical covariance and $corr_n$ the empirical correlation, respectively. By assumption, all X_j, $j = 1, 2, \ldots, p$, are standardized with mean zero and variance one. For the case $Cov(X_i, X_j) = \rho$, the results are similar and can be well approximated by the case $Cov_n(X_i, X_j) = \rho$. To analyze the revisiting behavior of the PGA algorithm in model (3) in the main text, it is sufficient to consider $X\alpha^m = X(\beta - \beta^m)$, which is the approximation error at the m^{th} boosting step. It is worth noting that

$$Cov_n(X\alpha^m, X_j) = \sum_{i \in T^m} \alpha_i^m \rho, \text{ if } j \notin T^m, \tag{88}$$

$$Cov_n(X\alpha^m, X_j) = \sum_{i \in T^m} \alpha_i^m \rho + \alpha_j^m (1 - \rho), \text{ if } j \in T^m, \tag{89}$$

where $T^m := T \cup supp(\beta^m)$ with $T := supp(\beta)$. WLOG., one can assume that $\sum_{i \in T^m} \alpha_i^m \geq 0$. We divide our analysis into two cases:
Case (a): If $\sum_{i \in T^m} \alpha_i^m > 0$, there must be a $j \in T^m$ such that $\alpha_j^m > 0$. Due to Equations (88) and (89), this implies

$$|Cov_n(X\alpha^m, X_j)| > |Cov_n(X\alpha^m, X_l)|$$

for all $l \notin \hat{T}^m$ since $\alpha_j^m(1 - \rho) > 0$ and $\sum_{i \in T^m} \alpha_i^m \rho > 0$. Therefore, the algorithm must select one predictor from T^m.
Case (b): If $\sum_{i \in T^m} \alpha_i^m = 0$, the current approximation error $X\alpha^m$ has a correlation of 0 with all variables j that are not in T^m, see Equation (88). As a result, the algorithm must either **(b1)** select a predictor that is in T^m if not all α_i^m are zero for $i \in T^m$ or **(b2)** stop since all α_i^m are 0.
In summary, the algorithm will always tend to select one predictor from the existing T^m unless it stops as the approximation error is already zero. Therefore,

$$T = T^0 = T^1 = \cdots = T^m.$$

Hence, we have 100 percent of revisiting, indicating that $\zeta^*(c_\phi) \to \infty$.

Appendix E. Applications

In this section, we present applications from different fields to illustrate the boosting algorithms. We present applications to demonstrate how the methods work when applied to real data sets and, then compare these methods to related methods, i.e. Lasso. The focus is on making predictions which is an important task in many applications.

E.1 Riboflavin production

This application involves genetic data and analyzes the production of riboflavin. First, we describe the data set, then we present the results.

E.1.1 Data set

The data set has been provided by DSM (Kaiserburg, Switzerland) and was made publicly available for academic research in Bühlmann et al. (2014) (Supplemental Material). The real-valued response/dependent variable is the logarithm of the riboflavin production rate. The (co-)variables measure the logarithm of the expression level of $4,088$ genes ($p = 4,088$), which are normalized. This means that the covariables are standardized to have variance 1, and the dependent variable and the resources are "de-meaned", which is equivalent to including an unpenalized intercept. The data set consists of $n = 71$ observations which were hybridized repeatedly during a fed-batch fermentation process in which different engineered strains and strains grown under different fermentation conditions were analyzed. For further details we refer to Bühlmann et al. (2014), their Supplemental Material and the references therein.

E.1.2 Results

We analyze a data set on the production of riboflavin (vitamin B_2). We split the data set randomly into two samples: a training set and a testing set. We estimate the model with different methods on the training set and then use the testing set to calculate out-of-sample mean squared errors (MSE) in order to evaluate the predictive accuracy. The size of the training set was 60 and the remaining 11 observations were used for forecasting. The table below shows the MSE for different methods discussed in the previous sections.

Table 11: Results Riboflavin Production (out-of-sample MSE)

BA-our	oBA-our	Lasso	p-Lasso
0.3641	0.1080	0.1687	0.1539

All calculations were performed in R (R Core Team (2014)) with the package hdm (Chernozhukov et al. (2015)) and our own code. Replication files are available upon request. The results show, again, that orthogonal L_2-Boosting outperforms Lasso and post-Lasso in this application.

E.2 Predicting Test Scores

E.2.1 Data Set

Here, the task is to predict the final score in the subjects Mathematics and Portugese in secondary education. This is relevant, e.g., to identify students which need additional support to master the material. The data contains both student grades and demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics and Portuguese. The data set is made available at the UCI Machine Learning Repository and was contributed by Paulo Cortez. The main reference for the data set is Cortez and Silva (2008).

E.2.2 Results

We employed five-fold cross-validation to evaluate the predictive performance of the data set. The results remain stable when choosing a different number of folds. The data sets contain, for both test results, 33 variables, which are used as predictors. The data set for the Mathematics test scores contains 395 observations, the sample size for Portuguese is 649. The results confirm our theoretical derivations that boosting is comparable to Lasso.

Table 12: Prediction of education (out-of-sample MSE)

subject	BA-our	oBA-our	Lasso	p-Lasso
Mathematics	19.1	19.3	18.4	18.4
Protugese	8.0	7.9	7.8	7.8

References

Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. Approximation and learning by greedy algorithms. *Ann. Statist.*, 36(1):64–94, 02 2008. doi: 10.1214/009053607000000631. URL http://dx.doi.org/10.1214/009053607000000631.

A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012. ISSN 1468-0262. doi: 10.3982/ECTA9626. URL http://dx.doi.org/10.3982/ECTA9626.

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 05 2013. doi: 10.3150/11-BEJ410. URL http://dx.doi.org/10.3150/11-BEJ410.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010*, III:245–295, 2010. ArXiv, 2011.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 11 2013. ISSN 0034-6527. doi: 10.1093/restud/rdt044. URL https://doi.org/10.1093/restud/rdt044.

Alexandre Belloni, Victor Chernozhukov, and Ying Wei. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics*, 34 (4):606–619, 2016. ISSN 07350015. URL http://www.jstor.org/stable/44166593.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996. ISSN 0885-6125. doi: 10.1007/BF00058655. URL http://dx.doi.org/10.1007/BF00058655.

Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–824, 1998.

Peter Bühlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34 (2):559–583, 2006.

Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007. doi: 10.1214/07-STS242. URL http://dx.doi.org/10.1214/07-STS242. with discussion.

Peter Bühlmann and Bin Yu. Boosting with the L_2 Loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003. ISSN 01621459. URL http://www.jstor.org/stable/30045243.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014. doi: 10.1146/annurev-statistics-022513-115545. URL http://dx.doi.org/10.1146/annurev-statistics-022513-115545.

Emmanuel Candes and Terence Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pages 2313–2351, 2007.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564–1597, 2014.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. *hdm: High-Dimensional Metrics*, 2015. R package version 0.1.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. ISBN 978-3-540-85635-1. Limit theory and statistical applications.

R. A. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996. ISSN 1572-9044. doi: 10.1007/BF02124742. URL `http://dx.doi.org/10.1007/BF02124742`.

Robert M. Freund, Paul Grigas, and Rahul Mazumder. A new perspective on boosting in linear regression via subgradient optimization and relatives. *Annals of Statistics*, 2016.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28:337–407, 2000. doi: 10.1214/aos/1016218223. URL `http://projecteuclid.org/euclid.aos/1016218223`. with discussion.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 00905364. URL `http://www.jstor.org/stable/2699986`.

Ching-Kang Ing. Model selection for high-dimensional linear regression with dependent observations. *The Annals of Statistics*, 48(4):1959–1980, 2020.

Ching-Kang Ing and Tze Leung Lai. A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statistica Sinica*, 21(4):1473–1513, 2011.

Robert Tibshirani Jerome Friedman, Trevor Hastie. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01`.

S. V. Konyagin and V. N. Temlyakov. Rate of convergence of pure greedy algorithm. *East J. Approx.*, 5(4):493–499, 1999. ISSN 1310-6236.

Jannis Kueck, Ye Luo, Martin Spindler, and Zigan Wang. Estimation and inference of treatment effects with l2-boosting in high-dimensional settings. *Journal of Econometrics*, 234(2):714–731, 2023.

Tze Leung Lai and Hongsong Yuan. Stochastic Approximation: From Statistical Origin to Big-Data, Multidisciplinary Applications. *Statistical Science*, 36(2):291 – 302, 2021. doi: 10.1214/20-STS784. URL https://doi.org/10.1214/20-STS784.

E.D. Livshitz and V.N. Temlyakov. Two lower estimates in greedy approximation. *Constructive Approximation*, 19(4):509–523, 2003. ISSN 1432-0940. doi: 10.1007/s00365-003-0533-6. URL http://dx.doi.org/10.1007/s00365-003-0533-6.

S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *Trans. Sig. Proc.*, 41(12):3397–3415, Dec 1993. doi: 10.1109/78.258082. URL http://dx.doi.org/10.1109/78.258082.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538 – 557, 2012. doi: 10.1214/12-STS400. URL https://doi.org/10.1214/12-STS400.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

Bernhard Stankewitz. Early stopping for -boosting in high-dimensional linear models. *The Annals of Statistics*, 52(2):491 – 518, 2024. doi: 10.1214/24-AOS2356. URL https://doi.org/10.1214/24-AOS2356.

Vladimir Temlyakov. *Greedy Approximation*. Cambridge University Press, New York, NY, USA, 1st edition, 2011. ISBN 1107003377, 9781107003378.

V.N. Temlyakov. Weak greedy algorithms. *Advances in Computational Mathematics*, 12 (2):213–227, 2000. ISSN 1572-9044. doi: 10.1023/A:1018917218956. URL http://dx.doi.org/10.1023/A:1018917218956.

Sara A. van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, 3:1360–1392, 2009. doi: 10.1214/09-EJS506. URL http://dx.doi.org/10.1214/09-EJS506.

T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005. doi: 10.1214/009053605000000255. URL http://projecteuclid.org/euclid.aos/1123250222.