Online Orthogonal Matching Pursuit

El Mehdi Saad, Gilles Blanchard, Sylvain Arlot

Online Orthogonal Matching Pursuit

El Mehdi Saad*, Gilles Blanchard*, Sylvain Arlot*

Abstract

Greedy algorithms for feature selection are widely used for recovering sparse high-dimensional vectors in linear models. In classical procedures, the main emphasis was put on the sample complexity, with little or no consideration of the computation resources required. We present a novel online algorithm: Online Orthogonal Matching Pursuit (OOMP) for online support recovery in the random design setting of sparse linear regression. Our procedure selects features sequentially, with one pass over data, alternating between allocation of samples only as needed to candidate features, and optimization over the selected set of variables to estimate the regression coefficients. Theoretical guarantees about the output of this algorithm are proven and its computational complexity is analysed.

1 Introduction

In the context of large scale machine learning, one often deals with massive data-sets and a considerable number of features. While processing such large data-sets, one is often faced with scarce computing resources. The adaptability of online learning algorithms to such constraints made them very popular in the machine learning community.

In the current work we address the problem of online feature selection, i.e support recovery algorithms restricted to a single training pass over the available data. This setting is particularly relevant when the system cannot afford several passes throughout the training set: for example, when dealing with massive amounts of data or when memory or processing resources are restricted, or when data is not stored but presented in a stream.

Suppose that there exists a vector $\beta^* \in \mathbb{R}^d$ with $\|\beta^*\|_0 = s^* \leq d$ such that the response variable y is generated according to the linear model $y = \langle x, \beta^* \rangle + \epsilon$, where ϵ satisfies $\mathbb{E}[\epsilon|x] = 0$, let $S^* = \mathrm{supp}\,(\beta^*)$. Throughout the article, we consider that the feature vector x is random, and we assume that $|y| < 1$ and $\|x\|_\infty < M$ almost surely for a known constant $M > 0$. The straightforward formulation of sparse regression using a l_0- pseudo-norm constraint is computationally intractable. This challenge motivated the rise of many computationally tractable procedures whose statistical validity has been established under additional assumptions such as the Irrepresentable Condition (IC) and Restricted Isometry Property (RIP).

Many algorithms have been proposed for support recovery, the most popular procedures use a convex relaxation with the l_1-norm (LASSO based algorithms, Tibshirani [1996]), and greedy procedures such as Orthogonal Matching Pursuit algorithm (OMP, Mallat and Zhang

*Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, Université Paris-Saclay.

[1993]), where features are selected sequentially. In this paper, we develop a novel online variant of OMP. Theoretical guarantees about OMP on support recovery were developed by Zhang [2011], under the IC+RIP assumption, and many variants have been developed Blumensath and Davies [2008], Combettes and Pokutta [2019], where different optimization procedures are used instead of ordinary least squares. However, the computational complexity remains of the order $\mathcal{O}(nd)$ for one variable selection step and $\mathcal{O}(s^*nd)$ for total support recovery, with a sample size satisfying $n = \Omega\left(\max\left(s^*, \frac{1}{\min\{|\beta_i^*|^2, \beta_i^* \neq 0\}}\right)\right)$ for exact support recovery with a high probability guarantee. A drawback of these procedures, besides the need to perform multiple passes over the training set, is that the sample size, hence the computational complexity of every step, depends on $(\min\{|\beta_i^*|, \beta_i^* \neq 0\})^{-1}$. Intuition suggests that recovery of the larger coefficients of β^* should be possible with less data and hence less computational complexity. We propose a feature selection procedure that is consistent with this intuition.

If the support size s^* is known, the proposed algorithm (OOMP) halts after recovering all features in S^*. Otherwise, it relies on some external criterion (such as a runtime budget), whenever halted, the procedure returns a set of features guarantees to belong to S^* with high probability. Moreover, we show that support recovery is achieved in finite time and provide a control on the computational complexity necessary to attain this goal.

1.1 Main contributions

This paper is about the design and analysis of support recovery for linear models in the online setting. We make the following contributions:

- We design a general modular procedure, where the learner can use any black-box optimization algorithm combined with an approximate best arm identification approach, provided those procedures come with suitable guarantees. We show that at any interruption time, it is guaranteed with high probability that the set of selected features S satisfies: $S \subseteq S^*$.
- We instantiate the general design using a variant of the stochastic gradient descent for the optimization and a LUCB-type (Lower Upper Confidence Bound) procedure for approximate best arm selection. The proposed algorithm has the advantage of being adapted to the streaming setting (i.e. requiring only one pass over data).
- A prior knowledge on the support size s^* or the magnitude of the smallest coefficient: $\min\{|\beta_i^*|, \beta_i^* \neq 0\}$, is not necessary to run the procedure. We show that OOMP recovers the support S^* in finite time and provide a control on the runtime necessary to achieve this objective.
- We compare the runtime required for support recovery using OOMP (C^{OOMP}) with the corresponding runtime using batch version OMP (C^{OMP}). We show that when $d > (s^*)^3$, it always holds $C^{\text{OOMP}} = \mathcal{O}(C^{\text{OMP}} \log^2\left(C^{\text{OMP}}\right))$, and when the coefficients of β^* have a different order of magnitude, C^{OOMP} can be much smaller than C^{OMP}. We provide some examples (such as polynomially decaying coefficients) to illustrate the gain in computational complexity of OOMP with respect to OMP.
- OMP was shown to require less data than Lasso for *support recovery* (Zhang [2009]). We consider the streaming sparse regression algorithm (SSR) presented in Steinhardt et al.

[2014], which is conceptually related to Lasso, as a benchmark to compare OOMP with l_1-regularization type algorithms. We prove that when $d > (s^*)^3$, OOMP outperforms SSR in terms of computational complexity.

Organization In section 2, we present high level ideas and key properties which underpin greedy feature selection principles such as the Orthogonal Matching Pursuit algorithm (in the batch as well as in the online setting). We then extend this idea and design a general Online OMP procedure which is built using two black-box procedures (namely **Optim** and **Try-Select**) in Section 3. Then, we instantiate this general procedure using Algorithms 5 for **Optim** and 6 for **Try-Select** in Section 4. Finally, we state theoretical guarantees about the output of the presented algorithm and provide a control on its runtime complexity. The last section presents simulations using synthetic data.

1.2 Notations used

Throughout the paper, we use the notation $[n] = \{1, \dots, n\}$. We denote by d the total input space dimension (total number of features), and s^* denotes the cardinality of the set S^* of features to be recovered. For a vector $\gamma \in \mathbb{R}^d$ and $F \subseteq [d]$, we denote $\gamma_{i:F}$ the coordinate of γ corresponding to the i-th element of F ranked in increasing order, and γ_F the vector of $\mathbb{R}^{|F|}$ such that $(\gamma_F)_i := \gamma_{i:F}$. Similarly, for a matrix $M \in \mathbb{R}^{d\times d}$ we denote M_F the matrix in $\mathbb{R}^{|F|\times|F|}$ obtained by restricting the matrix M to the lines and columns with indices in F. For a random vector $x \in \mathbb{R}^d$, a random variable $y \in \mathbb{R}$ and $F \subseteq [d]$ we denote $\mathrm{Cov}(x_F, y)$ the vector in $\mathbb{R}^{|F|}$ defined by $\mathrm{Cov}(x_F, y)_i = \mathrm{Cov}(x_{i:F}, y), \forall i \in [|F|]$. We denote Σ the covariance matrix of x. For $\beta \in \mathbb{R}^d$ let us denote $\mathcal{R}(\beta) = \mathbb{E}_{(x,y)}[(y - \langle x, \beta\rangle)^2]$ the (population) squared risk function.

The prefix S refers to results presented in the supplementary material.

2 Batch OMP and oracle version

We start with recalling the standard batch OMP (Algorithm 1) for reference. Then we will introduce an "oracle" version when the data is random, which will serve as a guide for constructing the online algorithm.

2.1 Batch OMP

Given a batch measurement matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a response vector $\boldsymbol{Y} \in \mathbb{R}^n$, at each iteration, OMP picks a variable that has the highest empirical correlation (in absolute value) with the ordinary linear least squares regression residue of the response variable with respect to features selected in the previous iterations. The algorithm stops when the maximum correlation is below a given threshold η.

Each iteration of Algorithm 1 comprises a selection procedure, where one selects a feature based on its correlation with the current residuals, and an optimization procedure, in this

Algorithm 1 OMP($\boldsymbol{X}$,$\boldsymbol{Y}$,η)

$S = \emptyset$, $\bar{\beta} = 0$
while true **do**
  $\hat{i} \leftarrow \text{argmax}_{j \notin S} |\boldsymbol{X}_{\cdot j}^t (\boldsymbol{Y} - \boldsymbol{X}\bar{\beta})|$.
  if $|\boldsymbol{X}_{\cdot \hat{i}}^t (\boldsymbol{Y} - \boldsymbol{X}\bar{\beta})| < \eta$ **then**
    Break
  else
    $S \leftarrow S \cup \{\hat{i}\}$
    $\bar{\beta} \leftarrow \underset{\text{supp}(\beta) \subseteq S}{\text{argmin}} \|\boldsymbol{X}\beta - \boldsymbol{Y}\|^2$
  end if
end while
return: S, $\bar{\beta}$.

Algorithm 2 Oracle OMP

Input: integer $s^* (\infty \text{ if unknown}), \mu \in [0, 1)$.
Let $S = \emptyset$.
while $|S| < s^*$ **do**
  Let $\beta^S = \underset{\text{supp}(\beta) \subseteq S}{\text{argmin}} \mathbb{E}_{(x,y)}[(y - \langle x, \beta \rangle)^2]$
  Let $Z_i^S = \mathbb{E}[x_i (y - \langle x, \beta^S \rangle)], (i = 1, \ldots, d)$.
  Select i^* such that:
    $Z_{i^*}^S \in [\mu \max_{j \in [d] \setminus S} Z_j^S, \max_{j \in [d] \setminus S} Z_j^S]$
  if $Z_{i^*}^S = 0$ **then Break**
  $S \leftarrow S \cup \{i^*\}$
end while
Output S.
On interrupt: return S.

case the ordinary least squares, where one optimizes the squared loss function over the space spanned by the set of selected features, and determines the new residuals for the next iteration.

2.2 Oracle OMP

To understand why OMP works, we consider the setting where the data is random and present an "oracle" (or population) version of OMP in order to give an insight about the core principle of its selection strategy, which we will adapt to the streaming setting. Throughout this work we assume the following on the generating distribution of feature vector and noise:

Assumption 1. $\mathbb{E}[x] = 0$, $y = \langle \beta^*, x \rangle + \epsilon$, *and the noise variable satisfies* $\mathbb{E}[\epsilon | x] = 0$.

Let us introduce the following classical assumption in support recovery literature, which appears in Tropp [2004], Zhao and Yu [2006] and Zhang [2009] as the irrepresentable condition (IC). Consider a subset $S \subseteq [d]$ and denote

$$\mu_S = \max_{j \in [d] \setminus S} \|\Sigma_S^{-1} \text{Cov}(x_S, x_j)\|_1.$$

Assumption 2 (Irrepresentable condition, IC)**.** *For all $S \subseteq [d]$ such that $|S| = s^*$,*

$$0 \leq \mu_S < 1.$$

Remark: The assumption $\mu_{S^*} < 1$ is often used for exact support recovery, it was shown in Zhang [2009] that it is a *necessary* condition for the consistency of batch OMP feature selection.

Consider for a subset $S \subseteq S^*$:

$$\beta^S \in \underset{\text{supp}(\beta) \subseteq S}{\text{argmin}} \mathcal{R}(\beta).$$

We define the covariance between the oracle residuals with each feature as:

$$Z_i^S := \mathbb{E}[x_i(y - \langle x, \beta^S \rangle)], i = 1, \ldots, d. \tag{1}$$

The selection criterion used in oracle OMP relies on the quantities Z_i^S, thanks to the following lemma:

Lemma 2.1. *Suppose Assumptions 1 and 2 hold. For any $S \subseteq S^*$, we have (with the convention $\max \emptyset = 0$):*

$$\max_{j \notin S^*} |Z_j^S| \leq \mu_{S^*} \max_{i \in S^* \setminus S} |Z_i^S|. \tag{2}$$

Algorithm 2 presents the resulting procedure, called Oracle version of OMP. In order to ease notations will use μ instead of μ_{S^*} in the remainder of this paper.

Remarks:
• A similar result was used in Zhang [2009] for the case of fixed design with random noise, where it was shown that either the empirical counterparts of Z_i^S are small, or they satisfy an inequality analogous to (2).
• The right-hand side of (2) can be written as $\max_{i \in S^*} |Z_i^S|$, since $Z_i^S = 0$ for all $i \in S$.
• This lemma shows in particular that under Assumptions 1-2, if $S \subseteq S^*$ and $\max_i |Z_i^S| > 0$, then $\max_{i \notin S^*} |Z_i^S| < \max_{i \in S^*} |Z_i^S|$. Hence, unless $S^* = S$, picking the feature with the largest population correlation $|Z_i^S|$ guarantees that this feature belongs to S^*.
• In the oracle setting, the algorithm stops as soon as $\max_i |Z_i^S| = 0$, since Lemma 2.1 guarantees that $S = S^*$ then. In the batch setting with a finite amount n of available data, the algorihm stops when the maximum empirical correlation is too small and and cannot guarantee $\max_i |Z_i^S| > 0$ due to estimation error. The threshold for stopping then depends on estimation error, hence on n, see Zhang [2009].

3 Online OMP

3.1 Settings

In a computation-resources-constrained setting, one aims at using the least possible queries of data points and features in order to gain in computational and memory efficiency. For a data point $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, define $z \in \mathbb{R}^{d+1}$ by: $z_{[d]} = x$ and $z_{d+1} = y$.

Algorithm 3 Online OMP(δ, s^*)

Input: $s^*(\infty$ if unknown), $\delta \in (0,1)$
Input: $\mu \in (0,1), \rho > 0$ (globals)
Let $S = \emptyset$.
while $|S| < s^*$ **do**
  $U \leftarrow$ **Select**$(S, \frac{\delta}{2(|S|+1)(|S|+2)}, 1)$
  $S \leftarrow S \cup U$
end while
Return: S
On interrupt: return S

In this paper, we focus on the the streaming data setting were one-pass over data is performed, as summarized above:

The algorithm queries quantities through: **query-new**(F), which takes as input $F \subseteq [d+1]$ and outputs the partial observation z_F of a fresh data point independent from all previously queried quantities. One call to **query-new**(F) has a time complexity of $\mathcal{O}(|F|)$.

In what follows, we will split algorithms into subroutines and assume that the input of each subroutine only depends on the result of past queries. This ensures that all the new data accessed by a subroutine can be considered as i.i.d. conditionally to its input. More formally, let us denote by $\mathcal{F}_n$ the σ-algebra generated by all queried quantities up to the n^{th} **query-new** query, and let N be the (possibly random) number of queries made before the call to the current subroutine. Mathematically, N is a stopping time; and, conditional to $\mathcal{F}_N$ the K next calls to **query-new** produce an i.i.d. sequence of (possibly partially observed) data points. We always assume that the input to each subroutine is $\mathcal{F}_N$-measurable. Below we will analyse each subroutine for a fixed input and derive probabilities with respect to the queried (i.i.d.) data; in the global flow of the algorithm, under the above assumption the same probabilistic bounds will hold conditional to $\mathcal{F}_N$.

3.2 Algorithm

Online OMP (Algorithm 3) selects variables sequentially. In its general form, Algorithm 4 (**Select**) consists of two sub-routines: **Optim** and **Try-Select**. The first provides an approximation of the regression coefficients for features in S. The latter is an approximate best arm identification strategy which uses the output of **Optim** and queries data points in order to try to select feature i, such that Z_i^S is large enough (Lemma 2.1 shows that such a feature is in S^*). We now describe how **Optim** and **Try-Select** operate:

Optim sub-routine: is assumed to be a black-box optimization procedure such that for any fixed subset $S \subseteq [d]$, positive number ξ and $\delta \in (0,1)$, **Optim**(S, δ, ξ) queries fresh data points through **query-new**$(S \cup \{d+1\})$ and outputs an approximation $\tilde{\beta}^S$ for β^S. We say that **Optim** satisfies the *optimization confidence property* if

$$\mathbb{P}\left[\mathcal{R}\left(\tilde{\beta}^S\right) - \mathcal{R}\left(\beta^S\right) > \xi \,\middle|\, S, \delta, \xi\right] \leq \delta, \tag{3}$$

Algorithm 4 Select(S,δ,ξ)

[Globals: $\mu \in (0,1), \rho \in (0,1)$]
$\tilde{\beta} \leftarrow$ **Optim**(S, δ, ξ)
$(U, \text{Success}) \leftarrow$ **Try-Select**$(S, \delta, \tilde{\beta}, \xi)$
if $\neg$Success **then**
  Return: **Select**$(S, \delta/2, \xi/4)$
else
  return U
end if

where the probability is with respect respect to the data queried during the procedure, for any fixed input (S, δ, ξ).

Try-Select sub-routine: Given a set of selected features S, an (approximate) regression coefficients vector $\tilde{\beta}^S$ and a confidence bound ξ (on $\tilde{\beta}^S$), **Try-Select**$(S, \delta, \tilde{\beta}^S, \xi)$ queries fresh data points to approximate Z_i^S defined by (1) for $i \in [d] \setminus S^*$ and either returns `Success=False`, or `Success=True` along with a set U of new selected features.

We say that **Try-Select** satisfies the *selection property* if for any (fixed) input $(S, \delta, \tilde{\beta}^S, \xi)$, it holds for the (random) output $(\texttt{Success}, U)$:

$$\text{provided } S \subseteq S^* \text{ and } \mathcal{R}\left(\tilde{\beta}^S\right) - \mathcal{R}\left(\beta^S\right) \leq \xi, \text{it holds:}$$
$$\mathbb{P}\Big[\overline{A}(\texttt{Success}, U) \,\Big|\, S, \delta, \tilde{\beta}^S, \xi\Big] \leq \delta,$$
$$\text{where } \overline{A}(\texttt{Success}, U) := \left\{\texttt{Success} = \texttt{True}; \exists i \in U : \mu_{S^*} \max_{j \in S^* \setminus S} \left|Z_j^S\right| \geq \left|Z_i^S\right|\right\}, \quad (4)$$

where the probability is with respect to all data queries made by **Try-Select** for fixed input. This implies in particular that $U \subset S^* \setminus S$ with probability $1 - \delta$, by Lemma 2.1 (and in particular, with the convention $\max \emptyset = 0$, the probability of returning $\texttt{Success} = \texttt{True}$ when $S = S^*$ is less than δ).

If **Try-Select** returns $\texttt{Success} = \texttt{False}$, this suggests that the bound ξ is not tight enough, i.e. that the prescribed precision ξ for the optimization part is insufficient to find a feature with the guarantee (4) holding with the required probability. In this case, using the doubling trick principle, **Select** is called recursively with the input $(S, \delta/2, \xi/4)$. Algorithm 4 presents the general form of the procedure **Select**.

If the cardinality $|S^*| = s^*$ is not known in advance, there is no stopping criterion and the procedure is run indefinitely. We assume that Online OMP will be interrupted externally by the user based on some arbitrary criterion, for example a limit on total computation time or other resource. In this case the current set S of selected features is returned. The next lemma ensures that at any interruption time, it is guaranteed with high probability that $S \subseteq S^*$.

Lemma 3.1. *Suppose that Assumptions 2 and 1 hold. Consider Algorithm 3 with the procedure* ***Select*** *given in Algorithm 4, assume that* ***Optim*** *satisfies the* optimization confidence property (3) *and that* ***Try-Select*** *satisfies the selection property* (4)*. Then when* ***OOMP****(δ, s^*) (Algorithm 3) is terminated, the variable S satisfies with probability at least $1 - 2\delta$: $S \subseteq S^*$.*

Remark: The above result only guarantees that the recovered features belong to the true support. We will see later in Lemma 5.1 that for the instantiations of **Try-Select** and **Optim** considered in the next section, unless the support S^* is completely recovered, the procedure **Select** finishes in finite time. Together with the previous lemma, this guarantees that the support S^* will be recovered in finite time with high probability, at which point **Select** will enter an infinite loop of recursive calls until interruption. In Section 5, we will derive quantitative bounds on the complexity for recovering the full support.

About the stopping rule: OOMP has access to a virtually infinite stream of data points, so unless it is halted externally by the user, the algorithm can (in principle) continue querying more data to search for potentially extremely small coefficients (in contrast to the batch setting where the amount of available data is limited). However it is possible, in every call of the procedure Try-Select, to communicate to the user an upper bound on the maximal magnitude of the remaining coefficients of variables in $S^* \setminus S$ (as shown in Section F). Therefore, the user can halt the procedure whenever that bound is small enough (alternatively, a threshold can be passed as an input to the algorithm and a corresponding stopping rule can be derived). We advocate an agnostic point of view where the user can decide for themselves when to halt the algorithm (based on the information on the magnitude of the remaining coefficients, but also possibly on limitations of the size of available data or computation time). Our recovery result guarantees that stopping at any time, the set of selected variables is (with high probability) a subset of S^*.

4 Instantiation of the Optimization procedure and Selection Strategy

In this section we provide an instantiation of **Try-Select** and **Optim** procedures.

4.1 Assumptions

In addition to the Irrepresentable Condition (IC) (Assumption 2) we will make an assumption of Restricted Isometry Property (RIP) Tropp [2004], Zhang [2009], Wainwright [2009] for the distribution of (x, y). Denote $\Lambda_S^{\min}$ and $\Lambda_S^{\max}$ the lowest and largest eigenvalue of Σ_S respectively.

Assumption 3. *[RIP] For all $S \subseteq [d]$ such that $|S| = s^*$, it holds $0 < \rho \leq \Lambda_S^{\min}, \Lambda_S^{\max} \leq L$.*

We also make the following assumption:

Assumption 4. *Assume that $|y| < 1$ and $\|x\|_\infty < M$ (a.s.).*

4.2 Instantiation of Optim and Try-Select

Recall that one call of the procedure **Select** results in successive calls of **Optim** and **Try-Select** until (at least) a feature is selected. Moreover, the quantities queried in a sub-routine call (either **Try-Select** or **Optim**) are independent from quantities queried during the execution of previous functions.

Algorithm 5 Optim $(S,\, \delta,\, \xi)$

Input: initial β_0, δ, ξ
Let $\tilde{\beta}_0 = \beta_0$, $\mathcal{X} = \mathcal{B}_{|S|}(0, \frac{2}{\sqrt{\rho}})$
$G \leftarrow 10\,|S|\,\frac{M^2}{\sqrt{\rho}} + 2\sqrt{|S|}M$
Let $T \leftarrow 21G^2 \log(1/\delta) \,/\, (\rho\xi)$
for $t \leftarrow 0, ..., T-1$ **do**
  $\eta_t \leftarrow \frac{2}{\rho(t+1)}$, $\nu_t \leftarrow \frac{2}{t+1}$
  $(X, Y) \leftarrow$ **query-new**$(S \cup \{d+1\})$
  $\gamma_{t+1} \leftarrow \beta_t - 2\eta_t(X^t\beta_t - Y)X$
  $\beta_{t+1} \leftarrow \Pi_{\mathcal{X}}(\gamma_{t+1})$
  //where $\Pi_{\mathcal{X}}$ is the projection operator on $\mathcal{X}$
  $\tilde{\beta}_{t+1} \leftarrow (1-\nu_t)\tilde{\beta}_t + \nu_t\beta_{t+1}$
end for
return $\tilde{\beta}_T$

Optimization procedure: We opted for the averaged stochastic gradient descent (Algorithm 5). High probability bounds on the output of this procedure were given in Harvey et al. [2019b]. We use this finding to build an optimization procedure satisfying the *optimization confidence property* (3) for an input (S, δ, ξ).

Proposition 4.1. *Let Assumptions 1,2, 3 and 4 hold. Then Algorithm 5 satisfies the optimization confidence property.*

Try-Select Strategy: Different approximate best arm identification strategies were developed in the literature. In this work, we opt for a LUCB-type strategy were we use some ideas from Mason et al. [2020]. We approximate Z_i^S by (i) replacing β^S by an approximation $\tilde{\beta}^S$ assumed to satisfy the condition $\mathcal{R}\left(\tilde{\beta}^S\right) - \mathcal{R}\left(\beta^S\right) \leq \xi$; (ii) replacing the expectation by an empirical counterpart using queried quantities. Given an i.i.d sequence $(X_h, Y_h), h \geq 1$, we define $\tilde{Z}_{i,n}^S(\tilde{\beta}^S)$ and $\tilde{V}_{i,n}(\tilde{\beta}^S)$ for $n \geq 2$, using $(X_h, Y_h), 1 \leq h \leq n$ written in matrix and vector form as $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{Y} \in \mathbb{R}^n$ by:

$$
\begin{aligned}
\tilde{Z}_{i,n}^S(\tilde{\beta}^S) &:= \frac{1}{n}\boldsymbol{X}_{.i}^t(\boldsymbol{X}\tilde{\beta}^S - \boldsymbol{Y}), i = 1, \ldots, d; \\
\tilde{V}_{i,n}(\tilde{\beta}^S) &:= \frac{1}{n(n-1)} \\
&\qquad \sum_{1\leq h,l\leq n} \left(\boldsymbol{X}_{i,h}(\boldsymbol{X}\tilde{\beta}^S - \boldsymbol{Y})_h - \boldsymbol{X}_{i,l}(\boldsymbol{X}\tilde{\beta}^S - \boldsymbol{Y})_l\right)^2; \\
\tilde{V}_{i,n}^+(\tilde{\beta}^S) &:= \max\left\{\tilde{V}_{i,n}(\tilde{\beta}^S); \frac{1}{1000}\frac{LM^2}{\rho}\right\}.
\end{aligned}
$$

Note that $\tilde{V}_{i,n}(\tilde{\beta}^S)^+$ represents a thresholded version of the empirical variance $\tilde{V}_{i,n}(\tilde{\beta}^S)$. Proposition 4.2 gives a concentration inequality for $\tilde{Z}_{i,n}^S$, using empirical Bernstein bounds **?**.

For $i \in [d] \setminus S, n \geq 2$ and $\delta \in (0,1)$, define $\tilde{B}(\tilde{\beta}^S) := M^2\|\tilde{\beta}^S\|_1 + M$ and:

$$\mathrm{conf}(i,n,\delta) := \sqrt{\frac{8\tilde{V}^+_{i,n}(\tilde{\beta}^S)\log(8dn^2/\delta)}{n}} + \frac{28\tilde{B}(\tilde{\beta}^S)\log(8dn^2/\delta)}{3(n-1)}. \tag{5}$$

Proposition 4.2. *Consider a fixed subset $S \subseteq S^*$ and put $k := |S|$. Suppose Assumptions 1, 2, 3 and 4 hold. Assume to be given a fixed $\tilde{\beta}^S \in \mathbb{R}^d$ with support S, satisfying $\mathcal{R}\left(\tilde{\beta}^S\right) - \mathcal{R}\left(\beta^S\right) \leq \xi$. For all $\delta \in (0,1)$, with probability at least $1-\delta$ it holds:*

$$\textit{for all } i \in [d] \setminus S, \textit{ and } n \geq 2: \quad |\tilde{Z}^S_{i,n}(\tilde{\beta}^S) - Z^S_i| \leq \frac{1}{2}\mathrm{conf}(i,n,\delta) + M\sqrt{\xi}. \tag{6}$$

Proposition 4.2 entails the following: conditionally to $S \subseteq S^*$, for all $\delta \in (0,1)$, with probability at least $1-\delta$: for all $i \in [d] \setminus S, n \geq 2$, the condition $2M\sqrt{\xi} < \mathrm{conf}(i,n,\delta)$ implies

$$|\tilde{Z}^S_{i,n} - Z^S_i| \leq \mathrm{conf}(i,n,\delta). \tag{7}$$

Provided inequality (7) holds true, and let $\hat{i} \in \mathrm{argmax}\{|\tilde{Z}^S_{i,n}| + \mathrm{conf}(i,n,\delta)\}$, then, if $j \in [d] \setminus S$ satisfies the following condition:

$$|\tilde{Z}^S_{j,n}| - \mathrm{conf}(j,n,\delta) \geq \mu\left(|\tilde{Z}^S_{\hat{i},n}| + \mathrm{conf}(\hat{i},n,\delta)\right), \tag{8}$$

then it holds that $\left|Z^S_j\right| > \mu \max_{i \in S^*}\left|Z^S_i\right|$ (see Lemma G.1 for a proof). Thus, in view of Proposition 4.2, under the above conditions, an algorithm selecting features j satisfying (8) satisfies the *selection property*.

Using this observation, we build Algorithm 6 as follows: the procedure repeatedly queries fresh data points (x,y) and updates the quantities $\tilde{Z}^S_{i,n}$ simultaneously for all $i \in [d] \setminus S$. After each iteration, we pick $\hat{i} \in \mathrm{argmax}\{|\tilde{Z}^S_{i,n}| + \ \mathrm{conf}(i,n,\delta)\}$ and we eliminate features for j which we are certain that $j \notin \mathrm{argmax}_i|Z^S_i|$ (i.e suboptimal features) with high probability through the test:

$$\left|\tilde{Z}^S_{j,n}\right| + \ \mathrm{conf}(j,n,\delta) < \left|\tilde{Z}^S_{\hat{i},n}\right| - \ \mathrm{conf}(\hat{i},n,\delta).$$

Moreover, we select features satisfying the condition (8). The procedure halts when the condition:

$$\left|\tilde{Z}^S_{\hat{i},n}\right| \leq \frac{2}{1-\mu}\ \mathrm{conf}(\hat{i},n,\delta)$$

is no longer satisfied. The algorithm then returns the set of selected features U. Lemma 5.1 shows that unless the support S^* is completely recovered, $U \neq \emptyset$ and the procedure halts in finite time almost surely. A concise version of **Try-Select** is given in Algorithm 6 (the detailed version is in Algorithm 8).

Algorithm 6 Try-Select (S, δ, $\tilde{\beta}$, ξ)

Input: S, δ, $\tilde{\beta}$, ξ $\quad$ {$\tilde{\beta}$ is of dim. $|S|$}
Output: S, Success
Let $v, Z,$ conf be d-arrays
{will store $\tilde{V}_{i,n}, \tilde{Z}^S_{i,n}$ and conf(i,n)}
$n \leftarrow 0$, $Z \leftarrow \mathbf{0}$, $v \leftarrow \mathbf{0}, U \leftarrow \emptyset, L \leftarrow [d+1] \setminus S$
while True **do**
  $n \leftarrow n+1$
  $(X, Y) \leftarrow$ **query-new**(L)
  for all $i \in \{1, \dots d\}$ **do**
    $Z[i] \leftarrow \frac{1}{n} X_i (Y - X_S^t \tilde{\beta}) + \frac{n-1}{n} Z[i]$
    Update $v[i]$
    conf$[i] \leftarrow$ conf(i,n)
  end for
  if $2M\sqrt{\xi} >$ min$_i$ conf$[i]$ **then**
    Success $\leftarrow$ False, **break**
  end if
  $\hat{i} \leftarrow \underset{i \in [d] \setminus S}{\text{argmax}} \{|Z[i]| + \text{conf}[i]\}$
  for all $i \in L \setminus \{d+1\}$ **do**
    if $|Z[i]| + \text{ conf}[i] \leq \left|Z[\hat{i}]\right| - \text{ conf}[\hat{i}]$ **then**
      $L \leftarrow L \setminus \{i\}$
    end if
    if $|Z[i]| - \text{ conf}[i] \geq \mu \left(\left|Z[\hat{i}]\right| + \text{ conf}[\hat{i}]\right)$ **then**
      $U \leftarrow U \cup \{i\}$
    end if
  end for
  if $|Z[\hat{i}]| > \frac{2}{1-\mu}$ conf$[\hat{i}]$ **then**
    Success $\leftarrow$ True, **break**
  end if
end while
return U, Success

5 Theoretical Guarantees and Computational Complexity Analysis

Consider one call of **Select**$(S, \delta, 1)$, for a fixed $S \subseteq S^*$. Lemma 5.1 below shows that, unless the support of S^* is totally recovered, the procedure **Select**$(S, \delta, 1)$ halts in finite time and updates S with a non-empty set of features.

Lemma 5.1. *Suppose Assumptions 1,2,3 and 4 hold. Consider one call of* ***Select***$(S, \delta, 1)$ *where* ***Try-Select*** *is given by Algorithm 6, and* ***Optim*** *is given by Algorithm 5. Denote by* τ

the stopping time where ***Select***$(S, \delta, 1)$ *updates* S *with the set of selected features* U *(i.e the subroutine* ***Try-Select*** *returns* U *and* `Success` $=$ `True`*), then :*

If $S \subsetneq S^*$*:* $\mathbb{P}(\tau < +\infty \text{ and } U \neq \emptyset) = 1$.

If $S = S^*$*:* $\mathbb{P}(\tau = +\infty) \geq 1 - 2\delta$.

Let $S \subsetneq S^*$ be a fixed subset and denote $k := |S|$. Recall that running **Select**$(S, \delta, 1)$ results in executing **Optim** and **Try-Select** alternatively (see Algorithm 4). Let us denote by $C^S_{\textbf{Optim}}$ the cumulative computational complexity of **Optim** when running **Select**$(S, \delta, 1)$ and by $C^S_{\textbf{Try-Select}}$ the cumulative computational complexity of **Try-Select** when running **Select**$(S, \delta, 1)$.

Theorem 5.2. *Suppose Assumptions 1, 2, 3 and 4 hold. Consider the procedure* ***Select*** *given by Algorithm 4,* ***Try-Select*** *given by Algorithm 6, and* ***Optim*** *as in Algorithm 5. Assume that* $S \subsetneq S^*$ *and denote* $k := |S|$*. Then* ***Select***$(S, \delta, 1)$ *selects a non-empty set of additional features* U *such that:*

$$\mathbb{P}\left(U \subset S^*\right) \geq 1 - 2\delta.$$

Moreover, the computational complexity of ***Select***$(S, \delta, 1)$ *subroutines* ***Optim*** *and* ***Try-Select*** *satisfy with probability at least* $1 - \delta$*:*

$$C^S_{\textbf{\textit{Optim}}} \leq \kappa k^3 \max\left\{\frac{1}{W_{i^*}^2}, \frac{\sqrt{k}}{W_{i^*}}\right\} \log\left(\frac{\bar{k}}{\delta W_{i^*}}\right);$$

$$C^S_{\textbf{\textit{Try-Select}}} \leq \kappa \sum_{i \in [d] \setminus S} \max\left\{\frac{1}{W_i^2}; \frac{\sqrt{\bar{k}}}{W_i}\right\} \log\left(\frac{d}{\delta W_{i^*}}\right) \log\left(\frac{\bar{k}}{W_{i^*}}\right);$$

where $i^* \in \operatorname{argmax}_{i \in S^* \setminus S} \left|Z_i^S\right|$*;* $W_i := \max((1-\mu)\left|Z_i^S\right|, \left|Z_{i^*}^S\right| - \left|Z_i^S\right|)$*;* $\bar{k} = \max\{1, k\}$ *and* κ *is a constant depending only on* ρ, L *and* M*.*

Theorem 5.2 provides high probability bounds on the computational complexity for a call to the procedure **Select**. A crucial point is that the complexity of the k-th step depends on the largest correlation $\left|Z_i^S\right|$ over the remaining (yet unselected) features, which in turn can be related to the average of the corresponding coefficients of β^* (see Lemma J.1). By contrast, due to the batch nature of OMP, its complexity is driven by the minimum coefficient of β^*, which determines the minimum amount of needed data for full recovery.

Let us introduce the following notation: let $\left(\beta_{(i)}\right)_{1 \leq i \leq s^*}$ be the coefficients of β^* ordered in decreasing sequence of magnitude. Let $\tilde{\beta}^2_{(s^*-k+1)}$ denote the average of the square of the k smallest non-zero coefficients of β^*: $\tilde{\beta}^2_{(s^*-k+1)} := \frac{1}{k}\sum_{i=s^*-k+1}^{s^*} \beta^2_{(i)}$.

Corollary 5.3. *Under the same assumptions as theorem 5.2. The computational complexity of* ***Select***$(S, \delta, 1)$ *subroutines* ***Optim*** *and* ***Try-Select*** *satisfy with probability at least* $1 - \delta$*:*

$$C^S_{\textbf{\textit{Optim}}} \leq \kappa \frac{k^3}{\tilde{\beta}^2_{(k+1)}} \log\left(\frac{\bar{k}}{\delta \tilde{\beta}^2_{(k+1)}}\right);$$

$$C^S_{\textbf{\textit{Try-Select}}} \leq \kappa \frac{d}{\tilde{\beta}^2_{(k+1)}} \log\left(\frac{\bar{k}}{\tilde{\beta}^2_{(k+1)}}\right) \log\left(\frac{d}{\delta \tilde{\beta}^2_{(k+1)}}\right);$$

where κ is a constant depending only on ρ, L, M, μ, and $\bar{k} = \max\{k, 1\}$.

We use bounds of corollary 5.3 to compare the computational complexity of OOMP with the computational complexity of OMP using the sample size prescribed by Zhang [2009] for full support recovery. Then, we compare OOMP with the SSR algorithm presented in Steinhardt et al. [2014] for streaming sparse regression, as a Lasso-type procedure. We use Theorem 8.2 in Steinhardt et al. [2014] to derive a sufficient sample size to achieve full support recovery.

We denote by C^{OOMP} the total runtime necessary for OOMP in order to recover the support completely, and denote by C^{OMP} and C^{SSR} the corresponding quantities for OMP and SSR respectively.

Corollary 5.4. *Under the same assumptions as theorem 5.2. If $d > (s^*)^3$, we have with probability at least $1 - \delta$:*

$$\frac{C^{OOMP}}{C^{OMP}} \leq \kappa \log^2\left(\frac{s^*}{\beta^2_{(s^*)}}\right) \frac{1}{s^*} \sum_{i=1}^{s^*} \frac{\beta^2_{(s^*)}}{\tilde{\beta}^2_{(i)}};$$
$$\frac{C^{OOMP}}{C^{SSR}} \leq \kappa \log^2\left(\frac{s^*}{\beta^2_{(s^*)}}\right) \frac{1}{(s^*)^2} \sum_{i=1}^{s^*} \frac{\beta^2_{(s^*)}}{\tilde{\beta}^2_{(i)}};$$

where κ is a constant depending only on ρ, L, M and μ.

Recall that we have $\forall i \in [s^*] : \beta^2_{(s^*)} \leq \tilde{\beta}^2_{(i)}$. Hence: $\frac{1}{s^*} \sum_{i=1}^{s^*} \frac{\beta^2_{(s^*)}}{\tilde{\beta}^2_{(i)}} \leq 1$, with equality only if all the square of the coefficients are equal. The SSR complexity bound have and additional factor $\frac{1}{s^*}$, the same factor appears when comparing the sample size used by OMP for support recovery n^{OMP} in Zhang [2009], with the corresponding quantity for Lasso n^{Lasso} in Zhao and Yu [2006]: $n^{\text{OMP}} = \mathcal{O}(\frac{n^{\text{Lasso}}}{s^*})$. Since our objective is support recovery, we will focus on the comparison between OOMP and OMP in the remainder of this paper.

In order to illustrate the advantage of OOMP over OMP, we consider the specific situation where the coefficients of β^* decay polynomially as: $\beta_i = \frac{1}{\sqrt{s^*}} \left(1 - \frac{i-1}{s^*}\right)^\gamma$, for $i \in S^*$ and $\beta_i = 0$ for $i \notin S^*$; with $\gamma \geq 0$ and we assume that $d > (s^*)^3$. Then we have, with probability at least $1 - \delta$:

$$\frac{C^{\text{OOMP}}}{C^{\text{OMP}}} \leq \kappa \frac{\log^2(s^*)}{(s^*)^{\min\{2\gamma, 1\}}}. \tag{9}$$

where κ is a constant depending only on ρ, L, M and μ. See section K for a proof of the results above. Thus, in a typical scenario of coefficient decay ($\gamma > 0$), OOMP reduces the complexity of OMP by a large factor (observe that the worst case in this scenario is $\gamma = 0$, i.e. when all coefficients all are of the same order, which is not the typical case in practice).

6 Simulations

In this section, we aim at comparing the computational complexities of OOMP and OMP. We denote n^{OMP} the sample size prescribed by- Zhang [2011] (recalled as Theorem K.2) to fully recover the support using OMP. We consider $C^{\text{OMP}} = s^* d n^{\text{OMP}} + (s^*)^2 n^{\text{OMP}}$ as a proxy for

the computational complexity of OMP. For OOMP, we use Lemma I.4 and evaluate C^{OOMP} as a function of the quantity of data points queried.

From a practical point of view, the number of iterations theoretically prescribed in the optimization procedure (the number T in Algorithm 5), and coming from Harvey et al. [2019b] is very pessimistic, due to the large numerical constant up to which the confidence bounds of the averaged stochastic gradient descent were developed. Taking this theoretical prescription to the letter resulted in the **Optim** step demanding an inordinate amount of data compared to **Try-Select**, while we expect the latter step to carry the larger part of the complexity burden due to the influence of the dimension d. For this reason, in our simulation we opted to significantly reduce this numerical constant, while ascertaining (since we know the ground truth) that the *optimization confidence property* (3) was still satisfied in practice in all simulations.

We generate samples (x_t, y_t) with each coordinate of x_t distributed as $\mathrm{Unif}\,[-B; B]$ with $B = 0.5$ and $y_t = \langle x_t, \beta^* \rangle + \epsilon_t$. We pick β^* to be a sparse vector with $s^* = \log_2(d)$ non zero coordinates and $\epsilon_t \sim \mathrm{Unif}\,([-\eta, \eta])$, where $\eta = 0.5$. We consider the case where the coefficients of β^* decay linearly: $\beta_i^* = \frac{1}{\sqrt{s^*}}\left(1 - \frac{i-1}{s^*}\right)$ for $i \in [s^*]$ and $\beta_i^* = 0$ if $i > s^*$. We consider two scenarios for the structure of the correlation matrix Σ: the orthogonal design $\Sigma_{\mathrm{orth}} = I_d$ and the power decay Toeplitz design, with parameter $\phi = 0.1$:

$$\Sigma_{\mathrm{Toeplitz}} = \begin{pmatrix} 1 & \phi & \cdots & \phi^{d-1} \\ \phi & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \phi \\ \phi^{d-1} & \cdots & \phi & 1 \end{pmatrix}$$

We run OOMP for $d \in \left\{2^2, 2^3, \ldots, 2^8\right\}$, we average the number of queried quantities over 20 runs and plot the ratio $\frac{C^{\mathrm{OOMP}}}{C^{\mathrm{OMP}}}$ in the logarithmic scale with base 2 as a function of $\log_2 d$ (Figure 1). We set $\delta = 0.1$. In all our simulation runs, the support S^* was correctly recovered. The results reported in Figure 1 show a significant reduction of the complexity between OOMP and OMP.

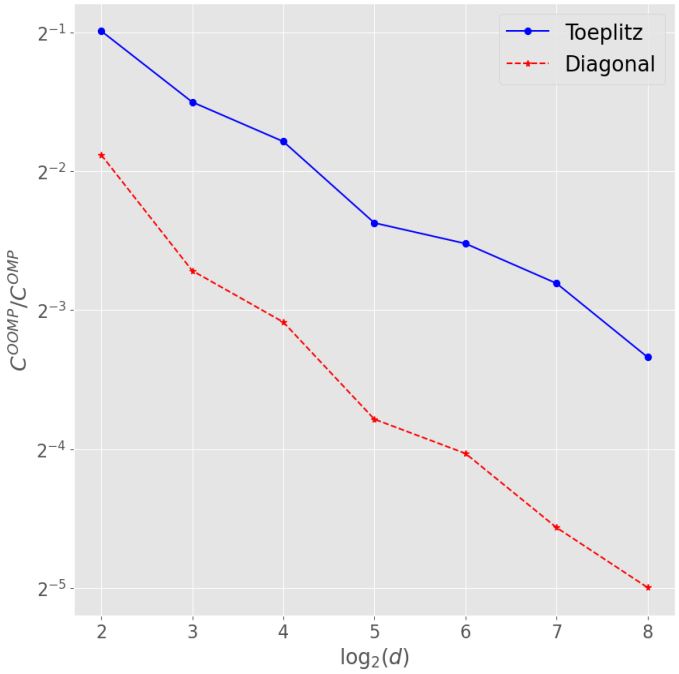

Figure 1: Comparison of computational complexities. The ratio $\frac{C^{\text{OOMP}.}}{C^{\text{OMP}.}}$ is plotted as a function of $\log_2(d)$ for both the Diagonal and Toeplitz covariance matrix.

References

Thomas Blumensath and Mike E Davies. Gradient pursuits. *IEEE Transactions on Signal Processing*, 56(6):2370–2382, 2008.

Cyrille Combettes and Sebastian Pokutta. Blended matching pursuit. In *Advances in Neural Information Processing Systems*, pages 2042–2052, 2019.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019a.

Nicholas JA Harvey, Christopher Liaw, and Sikander Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019b.

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all ϵ-good arms in stochastic bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL `http://www.cs.mcgill.ca/%7Ecolt2009/papers/012.pdf#page=1`.

Jacob Steinhardt, Stefan Wager, and Percy Liang. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.

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

Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55 (5):2183–2202, 2009.

Tong Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10(Mar):555–568, 2009.

Tong Zhang. Sparse recovery with orthogonal matching pursuit under rip. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

A Proof of Lemma 2.1

Suppose Assumptions 1 and 2 hold. For any subset $S \subseteq [d]$ define $\beta^S := \arg\min_{\mathrm{supp}(\beta)\subseteq S} \mathcal{R}(\beta)$, with $\mathcal{R}(\beta) = \mathbb{E}_{(x,y)}\left[(y - \langle x, \beta\rangle)^2\right]$.

Let us fix $S \subseteq S^*$, recall that $Z_i^S = \mathbb{E}\left[x_i(y - x^t\beta^S)\right]$; at first we only use the fact that the support S of β^S is a subset of S^*. We have, if $S^* \neq \emptyset$:

$$\begin{aligned}
\max_{i\in S^*} |Z_i^S| = \max_{i\in S^*} \left|\mathrm{Cov}\left(x_i, y - x^t\beta^S\right)\right| &= \max_{i\in S^*} \left|\mathrm{Cov}\left(x_i, x^t(\beta^{S^*} - \beta^S)\right)\right| \\
&= \max_{i\in S^*} \left|\mathbb{E}\left[x_i x^t(\beta^{S^*} - \beta^S)\right]\right| \\
&= \max_{i\in S^*} \left|\mathbb{E}\left[e_i^t x x^t(\beta^{S^*} - \beta^S)\right]\right| \\
&= \max_{i\in S^*} \left|e_i^t \Sigma\left(\beta^{S^*} - \beta^S\right)\right| \\
&= \left\|\Sigma\big(\beta^{S^*} - \beta^S\big)\right\|_\infty .
\end{aligned}$$

(The above remains true for $S^* = \emptyset$ with the convention $\max\emptyset = 0$). Recall that $S \subseteq S^*$, hence the support of β_S is included in S^*. Moreover by definition of β^{S^*}, its support is in S^*. Therefore, we have:

$$\max_{i\in S^*} |Z_i^S| = \|\Sigma_{S^*}\left(\beta^{S^*}_{S^*} - \beta^S_{S^*}\right)\|_\infty .$$

Let $v = \Sigma_{S^*}\left(\beta^{S^*}_{S^*} - \beta^{S}_{S^*}\right)$, and assume $v \neq 0$ (the case $v = 0$ is trivial). By definition of μ_{S^*}, we have for any $j \notin S^*$, using Assumption 2 and the previous display:

$$\begin{aligned}
\mu_{S^*} &= \max_{j \notin S^*} \left\| \Sigma^{-1}_{S^*} \mathrm{Cov}\left(x_{S^*}, x_j\right) \right\|_1 \\
&\geq \frac{\left|\mathrm{Cov}(x_{S^*}, x_j)^t \Sigma^{-1}_{S^*} v\right|}{\|v\|_\infty} \\
&= \frac{\left|\mathrm{Cov}(x_{S^*}, x_j)^t (\beta^{S^*}_{S^*} - \beta^{S}_{S^*})\right|}{\|v\|_\infty} \\
&= \frac{\left|\mathbb{E}\left[x_j x^t_{S^*} (\beta^{S^*}_{S^*} - \beta^{S}_{S^*})\right]\right|}{\|v\|_\infty} \\
&= \frac{\left|\mathbb{E}\left[x_j (y - x^t \beta^S)\right]\right|}{\|v\|_\infty} \\
&= \frac{|Z^S_j|}{\max_{i \in S^*} |Z^S_i|}.
\end{aligned}$$

We now use the actual definition of β^S, namely $\beta^S = \arg\min_{\mathrm{supp}(\beta) \subseteq S} \mathcal{R}(\beta)$, with $\mathcal{R}(\beta) = \mathbb{E}_{(x,y)}\left[(y - \langle x, \beta \rangle)^2\right]$. Since $\partial_i \mathcal{R}(\beta) = -2\mathbb{E}_{(x,y)}\left[x_i \left(y - \langle x, \beta \rangle\right)\right]$, we must have $0 = \partial_i \mathcal{R}(\beta^S) = -2Z^S_i$ for all $i \in S$. We conclude that $\max_{i \in S^*} |Z^S_i| = \max_{i \in S^* \setminus S} |Z^S_i|$ (including in the case $S = S^*$ where the latter right-hand side is 0 by convention), yielding the desired conclusion in conjunction with the last display.

B Technical Results

In this section we collect some technical results we will need for the proofs below. Recall that we assume the exact linear model:

$$y = \langle x, \beta^{S^*} \rangle + \epsilon,$$

with $\mathbb{E}[\epsilon | x] = 0$. In the result to come we restrict our attention to vectors β having support included in S for a fixed $S \subseteq S^*$ and denote $k := |S|$. Consequently we can with some abuse of notation assume that the ambient dimension is reduced to k (i.e $x \in \mathbb{R}^k$, $\beta^S \in \mathbb{R}^k$); let us denote by $\mathcal{R} : \mathbb{R}^k \to \mathbb{R}$ the loss function defined by: $\mathcal{R}(\beta) = \mathbb{E}[(y - x^t\beta)^2]$, $g : \mathbb{R}^k \to \mathbb{R}^k$ the gradient function defined by $g(\beta) = \nabla \mathcal{R}(\beta) = \mathbb{E}[2(x^t\beta - y)x]$ and for a sample (x, y) define: $\hat{g}_{(x,y)}(\beta) = 2(x^t\beta - y)x$. Denote by $\mathcal{B}_k\left(0, r\right)$ the closed ball centred at the origin with radius r in $\mathbb{R}^k$.

Lemma B.1. *Suppose Assumptions 3 and 4 hold. Considering the restrictions of functions $g, \hat{g}, \mathcal{R}$ to vectors β having support in S^* and reducing implicitly the ambient dimension to $s^* = |S^*|$, we have:*

1. *for any $S \subseteq S^*$: $\left\|\beta^S\right\|_2 \leq \frac{2}{\sqrt{\rho}}$.*
2. *$\forall \beta \in \mathcal{B}_k\left(0, \frac{2}{\sqrt{\rho}}\right)$: $\left\|\hat{g}_{(x,y)}(\beta)\right\|_2 \leq 4k\frac{M^2}{\sqrt{\rho}} + 2\sqrt{k}M$ (a.s).*

3. $\forall \beta \in \mathcal{B}_k\left(0, \frac{2}{\sqrt{\rho}}\right)$: $\|g(\beta)\|_2 \leq 4k\frac{M^2}{\sqrt{\rho}} + 2\sqrt{k}M$.

4. $\mathcal{R} : \mathbb{R}^k \to \mathbb{R}$ *is* ρ*-strongly convex.*

Proof. Recall that from Assumption 3, then the eigenvalues of the matrix Σ_{S^*} belong to $[\rho, L]$.

1. Since $\mathbb{E}[\epsilon|x] = 0$, and $y = x^t\beta^{S^*} + \epsilon$, we have for any $S \subseteq S^*$:

$$\mathbb{E}\left[\left(y - x^t\beta^S\right)^2\right] = \mathbb{E}\left[\left(x^t\left(\beta^{S^*} - \beta^S\right)\right)^2\right] + \mathbb{E}\left[\epsilon^2\right].$$

   By definition of β^S, it holds $\mathbb{E}\left[\left(y - x^t\beta^S\right)^2\right] \leq \mathbb{E}\left[y^2\right] \leq 1$, together with the above it gives:

$$\rho\|\beta^{S^*} - \beta^S\|_2^2 \leq \left(\beta^{S^*} - \beta^S\right)^t \Sigma_{S^*}\left(\beta^{S^*} - \beta^S\right) = \mathbb{E}\left[\left(x^t\left(\beta^{S^*} - \beta^S\right)\right)^2\right] \leq 1.$$

   In particular for $S = \emptyset$, we have: $\|\beta^{S^*}\|_2 \leq \frac{1}{\sqrt{\rho}}$. By the triangle inequality, for an arbitrary $S \subseteq S^*$:

$$\|\beta^S\|_2 \leq \frac{2}{\sqrt{\rho}}.$$

2. Let $\beta \in \mathcal{B}_k\left(0, \frac{2}{\sqrt{\rho}}\right)$, we have:

$$\begin{aligned}\|\hat{g}_{(x,y)}(\beta)\|_2 = \|2(x^t\beta - y)x\|_2 &\leq |2x^t\beta|\|x\|_2 + 2|y|\|x\|_2 \\ &\leq 2\|\beta\|_2\|x\|_2^2 + 2|y|\|x\|_2 \\ &\leq 2k\|x\|_\infty^2\|\beta\|_2 + 2\sqrt{k}\|x\|_\infty \\ &\leq 4k\frac{M^2}{\sqrt{\rho}} + 2\sqrt{k}M;\end{aligned}$$

   where we used: $\|x\|_2 \leq \sqrt{k}\|x\|_\infty$, and the assumptions $\|x\|_\infty \leq M$, $|y| \leq 1$.

3. Let $\beta \in \mathcal{B}_k\left(0, \frac{2}{\sqrt{\rho}}\right)$, we have:

$$\begin{aligned}\|g(\beta)\|_2 = \left\|\mathbb{E}\left[\hat{g}_{(x,y)}(\beta)\right]\right\|_2 &\leq \mathbb{E}\left[\|\hat{g}_{(x,y)}(\beta)\|_2\right] \\ &\leq 4k\frac{M^2}{\sqrt{\rho}} + 2\sqrt{k}M;\end{aligned}$$

   using the estimate of the previous point.

4. Recall that $\mathcal{R}$ is twice differentiable and its Hessian is given by $\mathbb{E}[xx^t] = \Sigma_{S^*} \geq \rho I_{s^*}$, therefore $\mathcal{R}$ is ρ-strongly convex.

□

C Proof of Lemma 3.1

Let us start by restating Lemma 3.1.

Lemma C.1. *Suppose that Assumptions 2 and 1 hold. Consider Algorithm 3 with the procedure* ***Select*** *given in Algorithm 4, assume that* ***Optim*** *satisfies the* optimization confidence property *and that* ***Try-Select*** *satisfies the* selection property*. Then when the* ***OOMP****(δ, s^*) (Algorithm 3) is terminated, the variable S satisfies with probability at least $1 - 2\delta$: $S \subseteq S^*$.*

Proof. First consider an idealized setting where the algorithm runs indefinitely. Let U_p denote the set of selected features at the p-th iteration of the main **while** loop of Algorithm 3. It can happen that the call to **Select** never terminates (this is actually the expected behaviour if all relevant features have been already discovered), so if $\bar{\tau}$ denotes the (random) last terminating iteration, we formally define $U_p = U_{\bar{\tau}}$ if $p > \bar{\tau}$ (this is of course irrelevant in practice but is just needed to always have a formally well defined U_p for all integers p). Denoting $S_p := \bigcup_{i=1}^{p} U_i$, we see that with this definition, for any integer $k \geq 1$:

$$\mathbb{P}\left(U_k \not\subset S^* | S_{k-1} \subseteq S^*\right) = \mathbb{P}\left(U_k \not\subset S^*; \bar{\tau} \geq k | S_{k-1} \subseteq S^*\right).$$

The event $\bar{\tau} \geq k$ implies that all iterations including the k^{th} one have terminated. Furthermore, the kth selection iteration then consisted in calling repeatedly the **Try-Select** with allowed error probability $\delta_{k,i} = (k(k+1)2^i)^{-1}\delta$ at the i-th call, until it returned `Success=true` (indicating termination of the k-th main selection iteration). Let us denote $B_{k,i}$ the event "the i-th call to **Optim** during the k-th selection iteration, if it took place, returned $\tilde{\beta}^S$ such that the optimization confidence property (3) holds", and $A_{k,i}$ the event "the i-th call to **Try-Select** during the k-th selection iteration, if it took place, returned `Success=true` and a subset of features $U \not\subset S^*$." It holds $\mathbb{P}(B_{k,i}^c | S_{k-1} \subseteq S^*) \leq \delta_{k,i}$ by the optimization confidence property, and $\mathbb{P}(A_{k,i} | S_{k-1} \subseteq S^*, B_{k,i}) \leq \delta_{k,i}$ by the selection property, so we have

$$\begin{aligned}
\mathbb{P}(U_k \not\subset S^*; \bar{\tau} \geq k | S_{k-1} \subseteq S^*) &\leq \mathbb{P}\left[\bigcup_{i=1}^{\infty} A_{k,i} \Big| S_{k-1} \subseteq S^*\right] \\
&\leq \sum_{i=1}^{\infty} \mathbb{P}(A_{k,i} | S_{k-1} \subseteq S^*) \\
&\leq \sum_{i=1}^{\infty} \mathbb{P}(A_{k,i} \cap B_{k,i} | S_{k-1} \subseteq S^*) + \mathbb{P}(B_{k,i}^c | S_{k-1} \subseteq S^*) \\
&\leq \sum_{i=1}^{\infty} \mathbb{P}(A_{k,i} | S_{k-1} \subseteq S^*, B_{k,i}) + \mathbb{P}(B_{k,i}^c | S_{k-1} \subseteq S^*) \\
&\leq 2 \sum_{i=1}^{\infty} \delta_{k,i}.
\end{aligned}$$

Now, the algorithm may be interrupted at a completely arbitrary time, and returns the

last active set $S = S_\tau$ for some $\tau \leq \bar{\tau}$. We then have

$$\begin{aligned}\mathbb{P}\left[S_\tau \nsubseteq S^*\right] \leq \mathbb{P}\left[\exists k \geq 1 : S_k \nsubseteq S^*\right] &\leq \mathbb{P}\left[\exists k \geq 1 : U_k \not\subset S^*; S_{k-1} \subseteq S^*\right] \\ &\leq \sum_{k\geq 1} \mathbb{P}\left[U_k \not\subset S^*; S_{k-1} \subseteq S^*\right] \\ &\leq \sum_{k\geq 1} \mathbb{P}\left[U_k \not\subset S^* | S_{k-1} \subseteq S^*\right] \\ &\leq 2\sum_{k,i=1}^{\infty} \delta_{k,i} = 2\delta.\end{aligned}$$

□

D Proof of Proposition 4.1

In this section we give high probability bounds on the output of the averaged stochastic gradient descent (ASGD, Algorithm 7). Theorem D.1 below is a slight modification of the main result in Harvey et al. [2019a], which consists in assuming that the error on the stochastic sub-gradients is bounded by a constant $G > 0$ instead of 1. We denote by $\Pi_{\mathcal{X}}$ the projection operator on $\mathcal{X} := \mathcal{B}\left(0, \frac{2}{\sqrt{\rho}}\right)$.

Algorithm 7 ASGD(T, β_0)

Input: initial β_0, T
for $t \leftarrow 0, ..., T-1$ **do**
  $\eta_t \leftarrow \frac{2}{\rho(t+1)}$, $\nu_t \leftarrow \frac{2}{t+1}$
  $(X, Y) \leftarrow$ **query-new**$(S \cup \{d+1\})$
  $\gamma_{t+1} \leftarrow \beta_t - 2\eta_t(X^t\beta_t - Y)X$
  $\beta_{t+1} \leftarrow \Pi_{\mathcal{X}}(\gamma_{t+1})$
  $\tilde{\beta}_{t+1} \leftarrow (1-\nu_t)\tilde{\beta}_t + \nu_t\beta_{t+1}$
end for
return $\tilde{\beta}_T$

We use the same notations as in Section B, we assume with some abuse of notation that the ambient dimension is reduced to $k := |S|$ (i.e $x \in \mathbb{R}^k$, $\beta^S \in \mathbb{R}^k$). We recall that we denote by $\mathcal{R} : \mathbb{R}^k \to \mathbb{R}$ the loss function defined by: $\mathcal{R}(\beta) = \mathbb{E}[(y - x^t\beta)^2]$, $g : \mathbb{R}^k \to \mathbb{R}^k$ the gradient function defined by $g(\beta) = \nabla\mathcal{R}(\beta) = \mathbb{E}[2(x^t\beta - y)x]$; in addition we consider $\hat{g}_n : \mathbb{R}^k \to \mathbb{R}^k$ defined by $\hat{g}_n(\beta) = 2((x_S^{(n)})^t\beta - y^{(n)})x^{(n)}$, where $(x^{(n)}, y^{(n)})$ are the output of the n^{th} call of **query-new** during Algorithm 5. Denote by $\mathcal{B}_k(0, r)$ the closed ball centred at the origin with radius r in $\mathbb{R}^k$.

Lemma B.1 shows that (under Assumptions 3-4), we have via the triangle inequality:

$$\|\hat{g}_{t+1}(\beta_t) - g(\beta_t)\| \leq 8k\frac{M^2}{\sqrt{\rho}} + 4\sqrt{k}M. \tag{10}$$

Where β_t are the iterates of Algorithm 5. We denote by G the upper bound in equation (10).

Theorem D.1. *Suppose Assumptions 3 and 4 hold. Let $\delta \in (0,1)$ and $S \subseteq S^*$ such that $S \neq \emptyset$. Denote by $\tilde{\beta}_T$ the output of ASGD$(T,0)$ (Algorithm 7).*

Then, with probability at least $1-\delta$ with respect to the samples queried during Algorithm 7:

$$\mathcal{R}(\tilde{\beta}_T) - \mathcal{R}(\beta^S) \leq \frac{21G^2 \log(1/\delta)}{\rho T},$$

where $G := 8k\frac{M^2}{\sqrt{\rho}} + 4\sqrt{k}M$.

The following corollary results by simply choosing T large enough such that the optimization confidence property is satisfied by Algorithm 7.

Corollary D.2. *Suppose assumptions Suppose Assumptions 3 and 4 hold. Let $\xi > 0, \delta \in (0,1)$. Consider algorithm 7 with inputs $(T,0)$ such that:*

$$T = \frac{21G^2 \log(1/\delta)}{\rho \xi},$$

where $k := |S|$ and $G := 8k\frac{M^2}{\sqrt{\rho}} + 4\sqrt{k}M$. Then the output $\tilde{\beta}_T$ satisfies with probability at least $1-\delta$:

$$\mathcal{R}(\tilde{\beta}_T) - \mathcal{R}(\beta^S) \leq \xi.$$

E Proof of Proposition 4.2

E.1 Technical Results

The following result is a straightforward modification of the empirical Bernstein inequality from Maurer and Pontil [2009], which consists in assuming that the random variables U_i belong to $[-B,B]$ for a $B > 0$, instead of $[0,1]$.

Lemma E.1. *Maurer and Pontil [2009] Let $U, U_1, \ldots, U_n$ be i.i.d. random variables with values in $[-B,B]$ and let $\delta > 0$. Then with probability at least $1-\delta$ we have:*

$$\left| \frac{1}{n}\sum_{i=1}^{n} U_i - \mathbb{E}[U] \right| \leq \sqrt{\frac{2V_n \ln(2/\delta)}{n}} + \frac{14B\ln(2/\delta)}{3(n-1)},$$

where:

$$V_n = \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (U_i - U_j)^2.$$

We are interested in applying the Lemma above to the quantities $\tilde{Z}^S_{i,n}$. Let (X,Y) be a queried sample, the following claim shows that the random variable $U := X_i(X^t\tilde{\beta}_S - Y)$ for $i \in [d]$, where X_i is the i^{th} feature X, satisfies the conditions of Lemma E.1.

Claim E.2. *Suppose Assumption 4 holds. Let (X, Y) be a sample, $\beta \in \mathbb{R}^d$ of support $S \subseteq [d]$ and such that $\|\beta\|_2 \leq \frac{2}{\sqrt{\rho}}$. Fix $i \in [d]$ and define $U = X_i \left(X^t\beta - Y\right)$. Then it holds almost surely:*

$$|U| \leq 2\sqrt{\frac{|S|}{\rho}}M^2 + M.$$

Proof. Using the Cauchy-Schwartz inequality, we have:

$$\begin{aligned} |U| &\leq |X_i| \left(\|X_S\|\|\beta\| + |Y|\right) \\ &\leq M\left(\sqrt{|S|}M\frac{2}{\sqrt{\rho}} + 1\right). \end{aligned}$$

□

Moreover, a straightforward calculation yields the result below.

Claim E.3. *Suppose Assumption 4 holds. Let (X, Y) be a sample, $\beta \in \mathbb{R}^d$ of support $S \subseteq [d]$. Fix $i \in [d]$ and define $U := X_i \left(X^t\beta - Y\right)$. Then it holds*

$$|U| \leq M^2\|\beta\|_1 + M.$$

Proof. We have:

$$\begin{aligned} |U| &\leq |X_i| \left(\|X\|_\infty \|\beta\|_1 + |Y|_\infty\right) \\ &\leq M\left(M\|\beta\|_1 + 1\right). \end{aligned}$$

□

E.2 Proof of Proposition 4.2

Consider an i.i.d sequence (X_h, Y_h). Let $n \geq 1$ and denote $(X_h, Y_h)_{1\leq h\leq n}$ in matrix and vector form as: $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{Y} \in \mathbb{R}^n$.

Let us first fix a set $S \subseteq S^*$, a feature $i \in [d] \setminus S$ and a vector $\beta \in \mathbb{R}^d$. Denote for all $j \in [n]$: $U_j := \boldsymbol{X}_{j,i}(\boldsymbol{X}_j^t\beta - \boldsymbol{Y}_j)$, where $\boldsymbol{X}_{j,i}$ is the i^{th} feature of the j^{th} sample $\boldsymbol{X}_j$. Recall that $\tilde{Z}_{i,n}^S(\beta) = \frac{1}{n}\sum_{j=1}^n U_j$ and $Z_i^S = \mathbb{E}_{(x,y)}[x_i\left(x^t\beta^S - y\right)]$. We have:

$$
\begin{aligned}
\left|\tilde{Z}^S_{i,n}(\beta) - Z^S_i\right| &= \left|\frac{1}{n}\sum_{j=1}^n U_j - \mathbb{E}_{(x,y)}\left[x_i\left(x^t\beta^S - y\right)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{j=1}^n U_j - \mathbb{E}_{(x,y)}\left[x_i\left(x^t\beta - y\right)\right]\right| + \left|\mathbb{E}_{(x,y)}\left[x_i\left(x^t\beta - y\right)\right] - \mathbb{E}_{(x,y)}\left[x_i\left(x^t\beta^S - y\right)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{j=1}^n U_j - \mathbb{E}_{(x,y)}\left[U_1\right]\right| + \left|\mathbb{E}_{(x,y)}\left[x_i x^t\left(\beta - \beta^S\right)\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{j=1}^n U_j - \mathbb{E}_{(x,y)}\left[U_1\right]\right| + M\left|\mathbb{E}_{(x,y)}\left[\left|x^t\left(\beta - \beta^S\right)\right|\right]\right| \\
&\leq \left|\frac{1}{n}\sum_{j=1}^n U_j - \mathbb{E}_{(x,y)}[U_1]\right| + M\sqrt{\mathcal{R}\left(\beta\right) - \mathcal{R}\left(\beta^S\right)}.
\end{aligned}
$$

Let us denote $\tilde{B}(\beta) := M^2\left\|\beta\right\|_1 + M$, and $\tilde{V}_n(\beta) := \frac{1}{n(n-1)}\sum_{1\leq p<q\leq n}(U_q - U_p)^2$. Since $(U_j)_{j\in[n]}$ are i.i.d and belong to $[-B, B]$ (Claim E.3, following from Assumption 3 and Lemma B.1 (i)), we have using Lemma E.1: for any $\delta \in (0,1)$, with probability at least $1 - \frac{\delta}{4dn^2}$:

$$
\left|\frac{1}{n}\sum_{j=1}^n U_j - \mathbb{E}_{(x,y)}\left[U_1\right]\right| \leq \sqrt{\frac{2\tilde{V}_n(\beta)\log(8dn^2/\delta)}{n}} + \frac{14\tilde{B}(\beta)\log(8dn^2/\delta)}{3(n-1)}. \tag{11}
$$

Now we apply a union bound over the sample size $n \geq 1$ and features $i \in [d]\setminus S$, we obtain: with probability at least $1 - \frac{\delta}{2}$, bound (11) holds for all n and i. To conclude, we choose $\beta = \tilde{\beta}^S$ and we use the risk bound (3) to have: with probability at least $1-\delta$:

$$
\forall i \in [d], \forall n \geq 1: \qquad \left|\tilde{Z}^S_{i,n}(\tilde{\beta}^S) - Z^S_i\right| \leq \sqrt{\frac{2\tilde{V}_n(\tilde{\beta}^S)\log(8dn^2/\delta)}{n}} + \frac{14\tilde{B}(\tilde{\beta}^S)\log(8dn^2/\delta)}{3(n-1)} + M\sqrt{\xi}.
$$

Recall:

$$
\tilde{V}^+_n(\beta) := \max\left(\tilde{V}_n(\beta), \frac{1}{1000}\frac{LM^2}{\rho}\right).
$$

Using the fact that $\tilde{V}_n(\beta) \leq \tilde{V}^+_n(\beta)$, combining with the above inequality we get the announced claim.

F Detailed algorithms for Try-Select

Algorithm 8 is a detailed version of Algorithm 6 (the shortened version in the main body of the paper).

Algorithm 8 Try-Select (S, δ, $\tilde{\beta}$, ξ), Data Stream setting

Input: S, δ, $\tilde{\beta}$, ξ
Output: S, Success
let $n \leftarrow 0$ be the number of queried samples.
let $v \leftarrow 0$ be an array to store the quantities $\tilde{V}_{i,n}$.
let conf be an array to store the confidence bound values.
let Z be an array to store the quantities $\tilde{Z}^S_{i,n}$.
let $U \leftarrow \emptyset$ denote the set of selected variables.
let $L \leftarrow [d+1] \setminus S$ denote the set of candidate variables.
//BEGINNING OF INITIALIZATION
$n \leftarrow 1$
$(X, Y) \leftarrow$ **query-new**$([d+1])$
$\tilde{Z}_i \leftarrow X_i \left(Y - X_S^t \tilde{\beta}\right)$, for all $i \in [d] \setminus S$.
//INITIALIZATION FOR EMPIRICAL VARIANCE QUANTITIES
$s_i \leftarrow 0$, $m_i \leftarrow X_i$, for all $i \in [d] \setminus S$.
// END OF INITIALIZATION
while True **do**
  $(X, Y) \leftarrow$ **query-new**$([d+1])$
  $n \leftarrow n + 1$
  $\forall\ i$: $Z_i \leftarrow X_i \left(Y - X_S^t \tilde{\beta}\right)$
  $\forall\ i$: $\tilde{Z}_i \leftarrow \frac{1}{n} Z_i + \frac{n-1}{n} \tilde{Z}_i$.
  // UPDATING THE EMPIRICAL VARIANCE
  $\forall\ i$: $\text{temp}_i \leftarrow m_i$
  $\forall\ i$: $m_i \leftarrow m_i + (Z_i - m_i)/n_i$
  $\forall\ i$: $s_i \leftarrow s_i + (Z_i - \text{temp}_i) * (Z_i - m_i)$
  $\forall\ i$: $v_i \leftarrow s_i / (n_i - 1)$
  $\forall\ i$: $\text{conf}(i) \leftarrow \sqrt{\frac{8 v_i \log(8dn^2/\delta)}{n_i}} + \frac{28B \log(8dn^2/\delta)}{3(n_i - 1)}$
  if $2M\sqrt{\xi} > \min_i\{\text{conf}(i)\}$ **then**
    Success $\leftarrow$ False, **break**
  end if
  let $\hat{i} \leftarrow \underset{i \in [d] \setminus S}{\text{argmax}} \{|\tilde{Z}_i| + \text{conf}(i)\}$
  //COMMUNICATING AN UPPER BOUND ON THE MEAN OF THE NON-RECOVERED COEFFICIENTS
  Communicate: $\sqrt{\frac{L}{\rho^3}\left(|\tilde{Z}_{\hat{i}}| + \text{conf}(\hat{i})\right)}$
  for all $i \in L \setminus \{d+1\}$ **do**
    if $|Z_i| + \text{conf}(i) \leq \left|Z_{\hat{i}}\right| - \text{conf}(\hat{i})$ **then**
      $L \leftarrow L \setminus \{i\}$
    end if
    if $|Z_i| - \text{conf}(i) \geq \mu \left(\left|Z_{\hat{i}}\right| + \text{conf}(\hat{i})\right)$ **then**
      $U \leftarrow U \cup \{i\}$
    end if
  end for
  if $|\tilde{Z}_{\hat{i}}| > \frac{2}{1-\mu} \text{conf}(\hat{i})$ **then**
    Success $\leftarrow$ True, **break**
  end if
end while
return U, Success

On the upper bound of the mean of the non-recovered coefficients: The bound communicated through the command:

$$\textbf{Communicate:} \quad \sqrt{\frac{L}{\rho^3}\left(|\tilde{Z}_{\hat{i}}| + \mathrm{conf}(\hat{i})\right)}$$

Is a direct consequence of the bound in lemma J.1 along with proposition 4.2.

G Proof of the selection property

The proof that the proposed Algorithm 6 satisfies the selection property hinges on the following lemma:

Lemma G.1. *Let $S \subseteq S^*$ be fixed. Let $(\tilde{\beta}^S)$ be given. Assume there exists $n \geq 1$, $\hat{i}, j \in [d] \setminus S$ and positive numbers $(\varepsilon_i)_{i\in[d]\setminus S}$ are such that:*

$$\hat{i} \in \mathrm{Argmax}_{i\in[d]\setminus S}\{|\tilde{Z}^S_{i,n}| + \varepsilon_i\}; \tag{12}$$

$$\forall i \in [d] \setminus S : |\tilde{Z}^S_{i,n} - Z^S_i| \leq \varepsilon_i; \tag{13}$$

$$|\tilde{Z}^S_{j,n}| - \varepsilon_j \geq \mu\left(|\tilde{Z}^S_{\hat{i},n}| + \varepsilon_{\hat{i}}\right). \tag{14}$$

Then it holds $\left|Z^S_j\right| \geq \mu \max_{i\in S^*}\left|Z^S_i\right|$.

Proof. First assume $S \subsetneq S^*$. Let $i^* \in \mathrm{Argmax}_{i\in[d]\setminus S}\{|Z^S_i|\}$. We have:

(12) implies that:

$$|\tilde{Z}^S_{i^*,n}| + \varepsilon_{i^*} \leq |\tilde{Z}^S_{\hat{i},n}| + \varepsilon_{\hat{i}}$$

Moreover, using (13) twice along with (14):

$$\left|Z^S_j\right| \geq \left|\tilde{Z}^S_{j,n}\right| - \varepsilon_j \geq \mu\left(|\tilde{Z}^S_{\hat{i},n}| + \varepsilon_{\hat{i}}\right) \geq \mu\left(|\tilde{Z}^S_{i^*,n}| + \varepsilon_{i^*}\right) \geq \mu\left|Z^S_{i^*}\right|$$

In the case $S = S^*$, we have that $Z^S_i = 0$ for all i, Therefore the claimed conclusion holds. □

Since Proposition 4.2 ensures that (13) is satisfied with probability $1 - \delta$ (for $\varepsilon_i = \mathrm{conf}(i, n_i, \delta)$, and uniformly for all values of n_i), provided $2M\sqrt{\xi} < \mathrm{conf}(i, n, \delta)$ for all i, Algorithm 6, which checks the latter condition and selects j satisfying (14), satisfies the selection property.

H Proof of Lemma 5.1

Lemma 5.1 shows that the procedure **Select** given in Algorithm 4, where **Try-Select** is given by Algorithm 6 in the Data Stream setting and **Optim** given by Algorithm 5, finishes in finite time if $S \subsetneq S^*$ and with high probability doesn't select any feature if $S = S^*$.

We start by stating the two following technical claim.

Claim H.1. *Let Assumptions 1 and 2 hold, and $S \subsetneq S^*$. Then* $\max_{i\in[d]\setminus S}\{\left|Z_i^S\right|\} > 0$.

This claim is a direct consequence of Lemma J.1 (see the proof of this lemma in Section J).

Consider a set of i.i.d samples $(\boldsymbol{X}_j, \boldsymbol{Y}_j)_{j\in[n]}$, recall the following notation:

$$U_{i,j} := \boldsymbol{X}_{j,i}\left(\boldsymbol{X}_j^t\tilde{\beta}^S - \boldsymbol{Y}_j\right); \tag{15}$$

$$\tilde{Z}_{i,n}^S := \frac{1}{n}\sum_{j=1}^{n} U_{i,j}\,; \tag{16}$$

$$\tilde{V}_{i,n} := \frac{1}{n(n-1)}\sum_{1\le p<q\le n}\left(U_{i,p} - U_{i,q}\right)^2; \tag{17}$$

$$\tilde{V}_{i,n}^+ := \max\left(\tilde{V}_{i,n}, \frac{1}{1000}\frac{LM^2}{\rho}\right); \tag{18}$$

$$\tilde{B} := M^2\|\tilde{\beta}^S\|_1 + M; \tag{19}$$

$$\mathrm{conf}\,(i,n,\delta) := \sqrt{\frac{8\tilde{V}_{i,n}^+\log(2dn^2/\delta)}{n}} + \frac{28\tilde{B}\log(2dn^2/\delta)}{3(n-1)}. \tag{20}$$

Proof of Lemma 5.1. For the situation $S = S^*$, the argument is a repetition of the proof of Lemma 3.1 (only considered at the particular selection iteration k where $S_k = S^*$).

We now deal with the situation $S \subsetneq S^*$. We assume S to be fixed, denote $k = |S|$. As explained in the main body of the paper, the argument to follow, for fixed S, can be transposed directly as a reasoning conditional to $\mathcal{F}_{N_k}$, N_k being the number of data used before starting the k-th selection step, with a random S assumed to be $\mathcal{F}_{N_k}$-measurable.

Let $i^* := \mathrm{argmax}_{i\in[d]\setminus S}\{\left|Z_i^S\right|\}$ (a deterministic quantity). Proceeding by proof via contradiction, suppose that with positive probability, during the execution of **Select** $(S, \delta_k, 1)$, **Try-Select** either never finishes, or always returns `Success` = `False`. Assume for the rest of the argument that this event is satisfied. We can rule out the fact **Try-Select** never stops, since there is a stopping condition of the type $\mathrm{conf}(i, n, 2^{-p}\delta_k) < \mathrm{cst}$, which is eventually met since $n \to \infty$ during **Try-Select**, so that the left-hand side goes to zero and the right-hand-side constant is positive. Therefore, for all $p \ge 0$ representing the number of recursive calls, **Try-Select** returns `Success` = `False`, after having queried a (random) number n_p of data points, satisfying (see Algorithms 4 and 6) that

$$\begin{cases} 2M\sqrt{\dfrac{1}{4^p}} > \mathrm{conf}\left(i_p, n_p, \dfrac{\delta_k}{2^p}\right); \\ \dfrac{2}{1-\mu_{S^*}}\mathrm{conf}\left(i^*, n_p - 1, \dfrac{\delta_k}{2^p}\right) > \left|\tilde{Z}_{i^*,n_p-1}^S\right|. \end{cases} \tag{21}$$

Using the definition of conf in (20), the first inequality of (21) implies (using the fact that: $\tilde{B} > M$):

$$2M\sqrt{\frac{1}{4^p}} > \frac{28M\log\left(2^{p+1}dn_p^2/\delta_k\right)}{3(n_p-1)}.$$

This implies that $n_p \geq c2^p$ for some factor $c = c(M, \rho, k, d, \delta_k)$, and in particular that $\lim_{p\to\infty} n_p = +\infty$.

Now Claim E.2 shows that $\tilde{V}^+_{i^*,n}$ defined by (18) is bounded almost surely by a constant independent of p. Hence, from the definition (20):

$$\lim_{p\to\infty} \mathrm{conf}\left(i^*, n_p - 1, \frac{\delta_k}{2^{p+1}}\right) = 0.$$

We use the second inequality of (21) to conclude that $\lim_{p\to\infty} \left|\tilde{Z}^S_{i^*,n_p-1}\right| = 0$. By the contradiction hypothesis we assumed that this happens on an event of positive probability. On the other hand, since the variables $\tilde{Z}^S_{i^*,n}$ are averages of i.i.d. variables $(\xi_j)_{1\leq j\leq n}$, and n_p is a stopping time that is lower bounded by $c2^p$, Lemma H.2 implies that the variance of $\tilde{Z}^S_{i^*,n_p}$ goes to 0 as p grows, hence $\tilde{Z}^S_{i^*,n_p}$ converges in probability to $Z^S_{i^*}$. Finally, we have $\tilde{Z}^S_{i^*,n_p} = \frac{1}{n_p}\xi_p + \frac{n_p-1}{n_p}\tilde{Z}^S_{i^*,n_p-1}$, hence $\left|\tilde{Z}^S_{i^*,n_p} - \tilde{Z}^S_{i^*,n_p-1}\right| \leq \frac{2B}{n_p}$, so that $\tilde{Z}^S_{i^*,n_p-1}$ converges in probability to $Z^S_{i^*}$ as well. Therefore $\left|Z^S_{i^*}\right| = 0$, which contradicts the fact that $\max_i \left|Z^S_i\right| > 0$ (see Claim H.1).

We used the following result:

Lemma H.2. *Let $(M_n)_{n\geq 1}$ be a martingale with respect to the filtration $(\mathcal{F}_n)_{n\geq 1}$ and N be a stopping time. Let $U_n := M_n - M_{n-1}$, for $n \geq 1$ (putting $M_0 = \mathbb{E}[M_n]$). Assume $\mathbb{E}\left[U_n^2\right] \leq A^2$ for all $n \geq 1$, and that $N \geq n_0$ a.s. Then:*

$$\mathrm{Var}\left(\frac{M_N}{N}\right) \leq A^2\left(\frac{1}{n_0} + \sum_{i>n_0} i^{-2}\right).$$

Proof. Assume without loss of generality that $E[M_n] = 0 = M_0$. We have, using the fact that the event $\{N \geq j\} = \{N < j\}^c$ is $\mathcal{F}_{j-1}$-measurable since N is a stopping time:

$$\begin{aligned}
\mathbb{E}\left[M_N^2\right] = \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^{N} U_iU_j\right] &= \mathbb{E}\left[\frac{1}{N^2}\sum_{i,j=1}^{\infty} U_iU_j\mathbf{1}\{N \geq \max(i,j)\}\right] \\
&= \mathbb{E}\left[\frac{1}{N^2}\left(\sum_{i=1}^{\infty} U_i^2\mathbf{1}\{N \geq i\} + 2\sum_{i<j} U_iU_j\mathbf{1}\{N \geq j\}\right)\right] \\
&\leq \sum_{i=1}^{\infty} \max(n_0, i)^{-2}\mathbb{E}\left[U_i^2\right] + 2\sum_{i<j}\mathbb{E}\left[\frac{1}{N^2}\mathbf{1}\{N \geq j\}U_i\underbrace{\mathbb{E}\left[U_j|\mathcal{F}_{j-1}\right]}_{=0}\right] \\
&\leq A^2\sum_{i=1}^{\infty}\max(n_0, i)^{-2}.
\end{aligned}$$

□

Finally, the set of selected features U is not empty since the condition: $\left|\tilde{Z}_{\hat{i},n_p}\right| > \frac{2}{1-\mu}\,\mathrm{conf}(\hat{i}, n_p, \frac{\delta_k}{2^p})$ implies that the condition: $\left|\tilde{Z}_{\hat{i},n_p}\right| - \mathrm{conf}(\hat{i}, n_p, \frac{\delta_k}{2^p}) \geq \mu\left(\left|\tilde{Z}_{\hat{i},n_p}\right| + \mathrm{conf}(\hat{i}, , \frac{\delta_k}{2^p})\right)$ is satisfied. Therefore, U contains at least $\hat{i}$.

I Proof of Theorem 5.2

Theorem 5.2 states that **Select** $(S, \delta, 1)$ is guaranteed to select a feature in S^* with high probability if the support is not totally recovered. This part is directly implied by Lemma 3.1 and the fact that the proposed **Optim** and **Try-Select** subroutines satisfy the optimization confidence property and the selection property, respectively, as established previously.

More importantly, the theorem gives an upper bound on the cumulative computational complexity of the sub-routines **Try-Select** and **Optim**.

In what follows, following the same approach as in the rest of the paper, we concentrate on a specific selection iteration (call to **Select**) and consider $S \subsetneq S^*$ to be fixed. We start by stating some technical lemmas useful for the proof of this theorem.

I.1 Technical Result

The following concentration inequality is a simple modification of the inequality presented in Maurer and Pontil [2009] Theorem 10, which consists in assuming that variables $(U_{j,i})_{j\in[n]}$ defined below belong to $[-B, B]$ instead of $[0, 1]$.

Lemma I.1. *Consider a fixed $i \in [d] \setminus S$. Suppose Assumption 4 holds with $\boldsymbol{X}$ and $\boldsymbol{Y}$ being centred random variables. Consider a set of i.i.d. data points $(\boldsymbol{X}_j, \boldsymbol{Y}_j)_{j\in[n]}$. Let $\beta \in \mathbb{R}^d$ such that $\|\beta\|_2 \leq \frac{2}{\sqrt{\rho}}$ and $supp(\beta) \subseteq S$.*

Define for a sample $(\boldsymbol{X}_j, \boldsymbol{Y}_j)$: $U_{j,i} = \left|\boldsymbol{X}_{j,i}(\boldsymbol{X}_j^t\beta - \boldsymbol{Y}_j)\right|$, where $\boldsymbol{X}_{j,i}$ is the i^{th} feature of $\boldsymbol{X}_j$. Finally we define $\tilde{V}_{i,n}$ as:

$$\tilde{V}_{i,n} = \frac{1}{n(n-1)} \sum_{1\leq l<j\leq n} (U_{j,i} - U_{l,i})^2 . \tag{22}$$

We have in the samples $(\boldsymbol{X}_j, \boldsymbol{Y}_j)_{j\in[n]}$:

$$\mathbb{P}\left(\sqrt{\mathbb{E}\tilde{V}_{i,n}} > \sqrt{\tilde{V}_{i,n}} + B\sqrt{\frac{2\log(1/\delta)}{n-1}}\right) \leq \delta;$$
$$\mathbb{P}\left(\sqrt{\tilde{V}_{i,n}} > \sqrt{\mathbb{E}\tilde{V}_{i,n}} + B\sqrt{\frac{2\log(1/\delta)}{n-1}}\right) \leq \delta,$$

where $B = M + 2\sqrt{\frac{k}{\rho}}M^2$.

We refer to Maurer and Pontil [2009] Theorem 10, for a proof; recall that Claim E.2 shows that $|U_{j,i}| < B$ almost surely.

Claim I.2. *Let $i \in [d] \setminus S$. Under the same assumptions as in Lemma I.1, we have:*

$$\mathbb{E}\tilde{V}_{i,n} \leq 20\frac{LM^2}{\rho},$$

where the expectation is taken with respect to the sample $(\boldsymbol{X}_j, \boldsymbol{Y}_j)_{j\in[n]}$.

Proof. We have by a simple calculation:

$$\tilde{V}_{i,n} \leq \frac{2}{n} \sum_{j=1}^{n} U_{j,i}^2. \tag{23}$$

Hence:

$$\begin{aligned}
\mathbb{E}\big[\tilde{V}_{i,n}\big] &\leq 2\mathbb{E}_{(x,y)}[U_{1,i}^2] \\
&\leq 2M^2\mathbb{E}_{(x,y)}[(x^t\beta - y)^2] \\
&\leq 4M^2\mathbb{E}_{(x,y)}\left[\left(x^t\beta\right)^2 + y^2\right] \\
&\leq 4M^2\left(\beta^t\Sigma\beta + 1\right) \\
&\leq 4M^2\left(L\left\|\beta\right\|^2 + 1\right) \\
&\leq 4M^2\left(\frac{4L}{\rho} + 1\right) \\
&\leq 20\frac{LM^2}{\rho},
\end{aligned}$$

where we used the assumption that $\|\beta\|_2 \leq \frac{2}{\sqrt{\rho}}$ (Lemma B.1).

□

Claim I.3. *Let $x \geq 1, c \in (0,1)$ and $y > 0$ such that:*

$$\frac{\log(x/c)}{x} > y. \tag{24}$$

Then:

$$x < \frac{2\log\left(\frac{1}{cy}\right)}{y}.$$

Proof. Inequality (24) implies

$$x < \frac{\log(x/c)}{y},$$

and further

$$\log(x/c) < \log(1/yc) + \log\log(x/c) \leq \log(1/yc) + \frac{1}{2}\log(x/c),$$

since it can be easily checked that $\log(t) \leq t/2$ for all $t > 0$. Solving and plugging back into the previous display leads to the claim. □

I.2 Proof of Theorem 5.2

It has already been established based on Lemma 3.1 that under Assumptions 1,2, 3 and 4, the set of features U selected by **Select**$(S, \delta, 1)$ belongs to S^* with high probability, and based on Lemma 5.1 that $U \neq \emptyset$.We therefore now focus on the control of the computational complexity.

Let $S \subsetneq S^*$ be a fixed subset and denote $k := |S|$. Recall that running **Select**$(S, \delta, 1)$ results in executing **Optim** and **Try-Select** alternatively until a condition is verified, implying

that at least one feature was selected (see Algorithm 4). We use the same notations as in Section 5 to denote the computational complexities of **Select**, **Try-Select** and **Optim**.

Lemma 5.1 shows that, unless interrupted, $\mathbf{Select}(S, \delta, 1)$ terminates in finite time. Therefore, the number of calls to **Optim** and **Try-Select** is finite. Let p denote this (random) number.

Let us adopt the following additional notations: For $q \in [p]$, let $m^{(q)}$ denote the number of samples queried during the q^{th} execution of **Optim**. Let, for $i \in [d] \setminus S$, $n_i^{(q)}$ denote the sample size used to compute $\tilde{Z}_i^S$ in the q^{th} execution of **Try-Select**.

The following lemma provides upper bounds for $C_{\mathbf{Optim}}$ and $C_{\mathbf{Try\text{-}Select}}$.

Lemma I.4. *Suppose Assumptions 3 and 4 hold. Let $S \subsetneq S^*$, we have almost surely:*

1. $C_{\textbf{\textit{Optim}}} \lesssim \sum_{q=1}^{p} m^{(q)} k$
2. $C_{\textbf{\textit{Try-Select}}} \lesssim \sum_{q=1}^{p} \sum_{i \in [d] \setminus S} n_i^{(q)}$,

where $\lesssim$ indicates inequality up to a numerical constant.

Proof. 1. **Optim** was instantiated using the averaged stochastic gradient descent (Algorithm 5), hence the computational complexity of the q^{th} call of **Optim** is upper bounded by $|S|m^{(q)}$ (up to a numerical constant). Therefore:

$$C_{\mathbf{Optim}} \lesssim \sum_{q=1}^{p} m^{(q)} k.$$

2. Consider the procedure **Try-Select** given in Algorithm 6. In one iteration, calling **query-new**(L) costs $\mathcal{O}(|L|)$. Once a sample (X, Y) is obtained, computing the residual $Y - X_S^t \tilde{\beta}$ costs $|S|$ and updating $\tilde{Z}, v_i$ and $\text{conf}(i)$ for all $i \in L$ costs $\mathcal{O}(|L|)$. Finally, selecting the feature i^* with the maximum $\{|\tilde{Z}_i| + \text{conf}(i)\}_{i \in L}$ costs $\mathcal{O}(|L|)$. The cost of the last two tests is $\mathcal{O}(|L|)$. Let $L_{q,t}$ denote the active set of features for the t-th iteration of **Try-Select** during its q-th call. We therefore have

$$C_{\mathbf{Try\text{-}Select}} \lesssim \sum_{q=1}^{p} \sum_{t=1}^{\infty} |L_{q,t}| = \sum_{q=1}^{p} \sum_{i \in [d] \setminus S} \sum_{t=1}^{\infty} \mathbf{1}\{i \in L_{q,t}\} = \sum_{q=1}^{p} \sum_{i \in [d] \setminus S} n_i^{(q)}.$$

□

In order to provide a control on the computational complexity of $C_{\mathbf{Select}}$, we need to derive a control on the (random) quantities p, $m^{(q)}$ and $n_i^{(q)}$ for $1 \leq q \leq p$ and $i \in [d] \setminus S$. In the remainder of this proof, κ will refer to a constant depending only on L, ρ and M. The value of κ may change from line to line.

Recall the definition:

$$\text{conf}\,(i, n, \delta) := \sqrt{\frac{8\tilde{V}_{i,n}^{+} \log\,(2dn^2/\delta)}{n}} + \frac{28\tilde{B} \log\,\big(2dn^2/\delta\big)}{3(n-1)}, \tag{25}$$

where $\tilde{B} := M + M^2\|\tilde{\beta}^S\|_1$ and $\tilde{V}_{i,n}^+$ is given by (22). Since conf(.) is a data-dependent function, the claim below provides a deterministic upper bound.

Claim I.5. *Suppose Assumption 4 holds with X and Y being centered random variables. Let $B_k := M + 2M^2\sqrt{\frac{k}{\rho}}$ and define:*

$$\overline{\text{conf}}\,(n,\delta) := 8\sqrt{\frac{LM^2\log\left(2dn^2/\delta\right)}{\rho n}} + \frac{27B_k\log\left(2dn^2/\delta\right)}{n}. \tag{26}$$

Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$ we have: $\forall i \in [d]\setminus S$, $\forall n \geq 2$:

$$\overline{\text{conf}}(n,\delta) \geq \text{conf}\,(i,n,\delta)\,.$$

Proof. Let $\delta \in (0,1)$. Lemma I.1 and Claim I.2 show that with probability at least $1-\delta$, $\forall i \in [d]\setminus S$, $n \geq 2$:

$$\sqrt{\tilde{V}_{i,n}} \leq \sqrt{8\frac{LM^2}{\rho}} + B_k\sqrt{\frac{2\log\left(2dn^2/\delta\right)}{n-1}}.$$

Moreover, recall that: $\tilde{B} = M^2\left\|\tilde{\beta}^S\right\|_1 + M$. Since $\tilde{\beta}^S \in \mathcal{B}_k\left(0, \frac{2}{\sqrt{\rho}}\right)$, we have: $\left\|\tilde{\beta}^S\right\|_1 \leq \sqrt{k}\left\|\tilde{\beta}^S\right\|_2 \leq 2\sqrt{\frac{k}{\rho}}$. Hence, we have almost surely: $\tilde{B} \leq B_k$. Using the bound on $\tilde{B}$ and on $\tilde{V}_{i,n}$ we obtain the conclusion. □

Let us denote $\delta_k := 1/(2(k+1)(k+2))$. At each iteration of OOMP (Algorithm 3), the procedure **Select** is called with inputs $(S, \delta_k, 1)$. Then **Select** is run following Algorithm 4 recursively until a condition, implying that at least an additional feature was selected, is verified. Thus, the inputs of the q^{th} call to **Select** are $(S, \delta_k/2^q, 1/4^q)$.

Computational complexity bounds:

We define the following key quantities: for $q \geq 1$, for $i \in [d]\setminus S$, let:

$$W_i := \max\left\{\frac{\left|Z_{i^*}^S\right| - \left|Z_i^S\right|}{4}; \frac{1-\mu}{3-\mu}\left|Z_i^S\right|\right\}, \tag{27}$$

and

$$\bar{n}_i^{(q)} := \min\left\{n > 0 : \overline{\text{conf}}\left(n, 2^{-q}\delta_k\right) < W_i\right\}, \tag{28}$$

where $i^* \in \text{argmax}_{i\in[d]}\left|Z_i^S\right|$.

The following argument proves the existence of $\bar{n}_i^{(q)}$: By assumption $S \subsetneq S^*$, Claim H.1 shows that $\left|Z_{i^*}^S\right| > 0$, thus $W_1 > 0$ as well. Definition 26 shows that $\overline{\text{conf}}(.,\delta)$ is strictly decreasing and converges to 0 when $n \to \infty$, which guarantees that $\bar{n}_i^{(q)}$ exists.

The technical result below gives an upper bound for $\bar{n}_i^q$:

Lemma I.6. *Let $i \in [d] \setminus S$ and $\bar{n}_i^{(q)}$ be defined by (28). Let W_i be the quantity defined by (27), We have:*

$$\bar{n}_i^{(q)} \leq \kappa \max\left\{\frac{1}{W_i^2}, \frac{\sqrt{k}}{W_i}\right\} \log\left(\frac{B_k d 2^q}{\delta_k W_i}\right),$$

where κ depends only on L, M and ρ, and $B_k := M + 2M^2\sqrt{\frac{k}{\rho}}$.

Proof. By definition of $\bar{n}_i^{(q)}$ we have:

$$\overline{\text{conf}}\left(\bar{n}_i^{(q)} - 1, 2^{-q}\delta_k\right) \geq W_i.$$

Using Definition 26 we have:

$$8\sqrt{\frac{LM^2 \log\left(2d(\bar{n}_i^{(q)} - 1)^2 2^q/\delta_k\right)}{\rho\left(\bar{n}_i^{(q)} - 1\right)}} + \frac{27B_k \log\left(2d(\bar{n}_i^{(q)} - 1)^2 2^q/\delta_k\right)}{\bar{n}_i^{(q)} - 1} \geq W_i.$$

Now, using the fact that $a + b > c \implies \max\{a, b\} > c/2$:

$$\begin{cases} \dfrac{\log\left(2d(\bar{n}_i^{(q)} - 1)2^q/\delta_k\right)}{\bar{n}_i^{(q)} - 1} \geq \dfrac{\rho}{256LM^2} W_i^2 \\ \qquad\qquad \text{or} \\ \dfrac{\log\left(2d(\bar{n}_i^{(q)} - 1)2^q/\delta_k\right)}{\bar{n}_i^{(q)} - 1} \geq \dfrac{1}{54B_k} W_i. \end{cases} \tag{29}$$

Now we use Claim I.3:

$$\begin{cases} \bar{n}_i^{(q)} - 1 \leq \dfrac{512LM^2}{\rho W_i^2} \log\left(\dfrac{128LM^2 d 2^q}{\rho\delta_k W_i^2}\right) \\ \qquad \text{or} \\ \bar{n}_i^{(q)} - 1 \leq \dfrac{108B_k}{W_i} \log\left(\dfrac{27B_k d 2^q}{\delta_k W_i}\right). \end{cases}$$

Finally, we upper bound $\bar{n}_i^{(q)}$ by the maximum of these bounds. □

For the rest of the proof, we upper bound the complexities of **Try-Select** and **Optim** using $\bar{n}_i^{(q)}$. The lemma below relates the quantities $n_i^{(q)}$ and $\bar{n}_i^{(q)}$.

Lemma I.7. *Under the assumptions of Theorem 5.2:*

$$\mathbb{P}\left(\forall q \leq p, \forall i \in [d] \setminus S : n_i^{(q)} \leq \bar{n}_i^{(q)} + 1\right) \geq 1 - 3\delta_k.$$

Proof. Let us fix $i \in [d] \setminus S$ and $q \in [p]$. We consider the iteration $n = n_i^{(q)} - 1$ during the q-th call of **Try-Select**, and let L denote the active set of features for this iteration.

Let $\hat{i} \in \text{argmax}_{j \in L} \left\{ \left| \tilde{Z}_{j,n} \right| + \text{ conf}(j, n, \delta_k 2^{-q}) \right\}$. We have by design of Algorithm 6 (since $n < n_i^{(q)}$):

$$\frac{2}{1-\mu} \text{conf}\left(\hat{i}, n, 2^{-q}\delta_k\right) > \left| \tilde{Z}^S_{\hat{i},n} \right|,$$

hence:

$$\frac{3-\mu}{1-\mu} \text{conf}\left(\hat{i}, n, 2^{-q}\delta_k\right) > \left| \tilde{Z}^S_{\hat{i},n} \right| + \text{conf}\left(\hat{i}, n, 2^{-q}\delta_k\right).$$

We therefore have (by definition of $\hat{i}$):

$$\frac{3-\mu}{1-\mu} \text{conf}\left(\hat{i}, n, 2^{-q}\delta_k\right) > \left| \tilde{Z}^S_{i,n} \right| + \text{conf}\left(i, n, 2^{-q}\delta_k\right). \tag{30}$$

As in the proof of Lemma 3.1, let us denote $B_{k,q}$ the event "the q-th call to **Optim** during the k-th selection iteration, if it took place, returned $\tilde{\beta}^S$ such that (3) holds" and recall that the optimization confidence property guarantees $\mathbb{P}\left[B^c_{k,q}\right] \leq \delta_k 2^{-q}$. Provided this control holds, recall that Proposition 4.2 shows that

$$\mathbb{P}\left(\forall m \geq 2, \forall j \in [d], \left| \tilde{Z}^S_{j,m} - Z^S_j \right| \leq \frac{1}{2} \text{conf}\left(j, m, 2^{-q}\delta_k\right) + M 2^{-q} \middle| B_{k,q} \right) \geq 1 - \delta_k 2^{-q}. \tag{31}$$

Let us denote by $A_{k,q}$ the event:

$$\forall m \geq 2, \forall j \in [d] \setminus S : \quad \left| \tilde{Z}^S_{j,m} - Z^S_j \right| \leq \text{conf}\left(j, m, 2^{-q}\delta_k\right) \tag{32}$$

Recall that at iteration n, we must have:

$$\forall i \in [d] \setminus S : \qquad \text{conf}(i, n, 2^{-q}\delta_k) \geq 2M 2^{-q},$$

thus (31) implies

$$\mathbb{P}\left(A_{k,q} \middle| B_{k,q} \right) \geq 1 - \delta_k 2^{-q}, \tag{33}$$

Using (30), we have:

$$\mathbb{P}\left(\frac{3-\mu}{1-\mu} \text{conf}\left(\hat{i}, n, 2^{-q}\delta_k\right) > \left| Z^S_i \right| \middle| B_{k,q} \right) \geq 1 - \delta_k 2^{-q}. \tag{34}$$

Using Claim I.5, it holds:

$$\mathbb{P}\left(\forall m \geq 2, \forall i \in [d] \setminus S : \overline{\text{conf}}\left(m, \delta_k 2^{-q}\right) > \text{conf}\left(i, m, \delta_k 2^{-q}\right) \right) \geq 1 - \delta_k 2^{-q}, \tag{35}$$

therefore, (34) gives:

$$\mathbb{P}\left(\overline{\text{conf}}\left(n, 2^{-q}\delta_k\right) > \frac{1-\mu}{3-\mu} \left| Z^S_i \right| \middle| B_{k,q} \right) \geq 1 - \delta_k 2^{-q}. \tag{36}$$

Let $i^* \in \text{argmax}_{j \in [d] \setminus S} \left| Z^S_j \right|$. Suppose that event $A_{k,q}$ is true. Let us show that $i^* \in L$. In fact, if $i^* \notin L$, we have by design of the procedure **Try-Select**: $\exists m < n$ and $\exists j \in [d] \setminus S$ such that:

$$\left| \tilde{Z}^S_{i^*,m} \right| + \text{ conf}(i^*, m, \delta_k 2^{-q}) < \left| \tilde{Z}^S_{j,m} \right| - \text{ conf}(j, m, \delta_k 2^{-q})$$

By definition of event $A_{k,q}$ in (32). We conclude that:

$$\left|Z_{i^*}^S\right| < \left|Z_j^S\right|,$$

which contradicts the definition of i^*. We therefore have: if $A_{k,q}$ is true then $i^* \in L$.

Moreover, by design of **Try-Select**:

$$\begin{aligned}\left|\tilde{Z}_{i,n}^S\right| + \text{conf}\left(i, n, \delta_k 2^{-q}\right) &\geq \left|\tilde{Z}_{\hat{i},n}^S\right| - \text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right) \\ &= \left|\tilde{Z}_{\hat{i},n}^S\right| + \text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right) - 2\text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right) \\ &\geq \left|\tilde{Z}_{i^*,n}^S\right| + \text{conf}\left(i^*, n, \delta_k 2^{-q}\right) - 2\text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right)\end{aligned}$$

Therefore:

$$\left|\tilde{Z}_{i,n}^S\right| - \text{conf}\left(i, n, \delta_k 2^{-q}\right) + 2\text{conf}\left(i, n, \delta_k 2^{-q}\right) \geq \left|\tilde{Z}_{i^*,n}^S\right| + \text{conf}\left(i^*, n, \delta_k 2^{-q}\right) - 2\text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right).$$

Since event $A_{k,q}$ is true, we upper bound the quantity : $\left|\tilde{Z}_{i,n}^S\right| - \text{conf}\left(i, n, \delta_k 2^{-q}\right)$, and lower bound the quantity: $\left|\tilde{Z}_{i^*,n}^S\right| + \text{conf}\left(i^*, n, \delta_k 2^{-q}\right)$. We obtain:

$$\left|Z_i^S\right| + 2\text{conf}\left(i, n, \delta_k 2^{-q}\right) \geq \left|Z_{i^*}^S\right| - 2\text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right).$$

As a conclusion, we have:

$$\mathbb{P}\left(\left|Z_i^S\right| + 2\text{conf}\left(i, n, \delta_k 2^{-q}\right) \geq \left|Z_{i^*}^S\right| - 2\text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right) \middle| B_{k,q}\right) \geq 1 - \delta_k 2^{-q},$$

which leads to:

$$\mathbb{P}\left(2\text{conf}\left(i, n, \delta_k 2^{-q}\right) + 2\text{conf}\left(\hat{i}, n, \delta_k 2^{-q}\right) \geq \left|Z_{i^*}^S\right| - \left|Z_i^S\right| \middle| B_{k,q}\right) \geq 1 - \delta_k 2^{-q}.$$

Finally, we use (35) to upper bound $\text{conf}\left(i, ., .\right)$ and $\text{conf}\left(\hat{i}, ., .\right)$ using $\overline{\text{conf}}(.)$:

$$\mathbb{P}\left(4\overline{\text{conf}}\left(n, \delta_k 2^{-q}\right) \geq \left|Z_{i^*}^S\right| - \left|Z_i^S\right| \middle| B_{k,q}\right) \geq 1 - \delta_k 2^{-q}. \tag{37}$$

We obtain, using (37) and (36):

$$\mathbb{P}\left(\overline{\text{conf}}\left(n, \delta_k 2^{-q}\right) \geq W_i \middle| B_{k,q}\right) \geq 1 - \delta_k 2^{-q}; \tag{38}$$

furthermore by definition of $\bar{n}_i^{(q)}$ (see (28)):

$$\overline{\text{conf}}\left(\bar{n}_i^{(q)}, \delta_k 2^{-q}\right) \leq W_i. \tag{39}$$

Using inequalities (38)-(39), we have:

$$\mathbb{P}\left(\overline{\text{conf}}\left(n_i^{(q)} - 1, \delta_k 2^{-q}\right) \geq \overline{\text{conf}}\left(\bar{n}_i^{(q)}, \delta_k 2^{-q}\right) \middle| B_{k,q}\right) \geq 1 - 2\delta_k 2^{-q}.$$

Denoting $D_{k,q}$ the event appearing above, we use $\mathbb{P}\left[D_{k,q}^c\right] \leq \mathbb{P}\left[D_{k,q}^c \cap B_{k,q}\right] + \mathbb{P}\left[B_{k,q}^c\right] \leq \mathbb{P}\left[D_{k,q}^c | B_{k,q}\right] + \mathbb{P}\left[B_{k,q}^c\right] \leq 2\delta_k 2^{-q}$ together with a union bound over $q \geq 1$ to get

$$\mathbb{P}\left(\forall q \leq p : \overline{\text{conf}}\left(n_i^{(q)} - 1, \delta_k 2^{-q}\right) \geq \overline{\text{conf}}\left(\bar{n}_i^{(q)}, \delta_k 2^{-q}\right)\right) \geq 1 - 3\delta_k.$$

The result follows from the fact that the function $n \to \overline{\text{conf}}(n, \delta)$ is decreasing for all $\delta \in (0, 1)$. □

In order to get an upper bound for the computational complexity of **Select**, we now develop a high probability bound on p (the total number of calls of **Try-Select** and **Optim** during one call of **Select** $(S, \delta_k, 1)$).

Lemma I.8. *Suppose $p \geq 2$. Under the assumptions of Theorem 5.2, p satisfies the following inequality:*

$$\mathbb{P}\left(2^p \leq \kappa \max\left\{\frac{1}{W_{i^*}}; \sqrt{\frac{B_k}{W_{i^*}}}\right\}\right) \geq 1 - 3\delta_k,$$

where κ only depends on (ρ, L, M).

Proof. By definition of p, the procedure **Try-Select** returns `Success = False` in its call number $p - 1$. Then (see Algorithm 6) $\exists i \in [d] \setminus S$ such that:

$$2M\sqrt{\frac{1}{4^{p-2}}} > \text{conf}\left(i, n_i^{(p-1)}, \frac{\delta_k}{2^{p-2}}\right).$$

Using Definition 25 for conf, we deduce:

$$2M\sqrt{\frac{1}{4^{p-2}}} > \sqrt{\frac{8\tilde{V}^+_{i,n_i^{(p-1)}} \log\left(2^{p-1} d (n_i^{(p-1)} - 1)^2 / \delta_k\right)}{n_i^{(p-1)} - 1}}.$$

Recall that by definition of $\tilde{V}^+_{i,n_i}$, it holds

$$\tilde{V}^+_{i,n_i} \geq \frac{1}{10^3} \frac{LM^2}{\rho},$$

therefore

$$2M\frac{1}{2^{p-2}} > \frac{1}{11}\sqrt{\frac{LM^2}{\rho\left(n_i^{(p-1)} - 1\right)} \log\left(2^{p-1} d \left(n_i^{(p-1)} - 1\right)^2 / \delta_k\right)},$$

and finally

$$2^p \leq c\sqrt{\frac{\rho(n_i^{(p-1)} - 1)}{L \log\left(2^p d (n_i^{(p-1)} - 1)/\delta_k\right)}},$$

for c an absolute numerical constant.

Using Lemma I.7 along with the fact that the function $n \to n/\log(an)$ is non-decreasing for $a > 1$, we have:

$$\mathbb{P}\left(2^p \leq c\sqrt{\frac{\rho \bar{n}_i^{(p-1)}}{L\log\left(2^p d\bar{n}_i^{(p-1)}/\delta_k\right)}}\right) \geq 1 - 3\delta_k.$$

Recall from (29) that there is a numerical constant c' such that:

$$\frac{\log\left(d(\bar{n}_i^{(p-1)} - 1)2^q/\delta_k\right)}{\bar{n}_i^{(p-1)} - 1} \geq c' \max\left\{\frac{\rho}{LM^2}W_i^2; \frac{1}{B_k}W_i\right\}.$$

Finally, it is elementary to check that $\forall x \in [0, \left|Z_{i^*}^S\right|]$:

$$\begin{aligned}\max\left\{\frac{1}{4}\left(\left|Z_{i^*}^S\right| - x\right), \frac{1-\mu}{3-\mu}x\right\} &\geq \frac{3-\mu}{7-5\mu}\left|Z_{i^*}^S\right| \\ &\geq \frac{2}{7}W_{i^*}.\end{aligned}$$

Hence, taking $x = \left|Z_i^S\right|$ above, we get $W_i \geq \frac{2}{7}W_{i^*}$. As a conclusion, there exists a constant κ depending only on ρ, L and M such that:

$$\mathbb{P}\left(2^p \leq \kappa \max\left\{\frac{1}{W_{i^*}}; \sqrt{\frac{B_k}{W_{i^*}}}\right\}\right) \geq 1 - 3\delta_k.$$

□

Recall that we have: $C_{\textbf{Try-Select}} \lesssim \sum_{q=1}^p \sum_{i\in[d]\setminus S} n_i^{(q)}$ (Lemma I.4). Therefore, using Lemmas I.6, I.7 and I.8 above, we have with probability at least $1 - 3\delta_k$:

$$\begin{aligned}C_{\textbf{Try-Select}} &\lesssim \sum_{q=1}^p \sum_{i\in[d]\setminus S} n_i^{(q)} \\ &\lesssim \sum_{q=1}^p \sum_{i\in[d]\setminus S} \bar{n}_i^{(q)} \\ &\leq \sum_{q=1}^p \sum_{i\in[d]\setminus S} \kappa \max\left\{\frac{1}{W_i^2}, \frac{\sqrt{k}}{W_i}\right\} \log\left(\frac{B_k d 2^q}{\delta_k W_i}\right) \\ &\leq p\kappa \sum_{i\in[d]\setminus S} \max\left\{\frac{1}{W_i^2}, \frac{\sqrt{k}}{W_i}\right\} \log\left(\frac{B_k d 2^p}{\delta_k W_i}\right).\end{aligned}$$

In particular, Lemma I.8 shows that:

$$\mathbb{P}\left(2^p \lesssim \max\left\{\frac{1}{W_{i^*}}; \sqrt{\frac{B_k}{W_{i^*}}}\right\}\right) \geq 1 - 3\delta_k.$$

Hence, with probability at least $1 - 3\delta_k$:

$$\log(2^p) \leq \kappa \log\left(\frac{k}{W_{i^*}}\right).$$

We conclude after some elementary bounding that, with probability at least $1 - 6\delta_k$:

$$C_{\textbf{Try-Select}} \leq \kappa \sum_{i \in [d] \setminus S} \max\left\{\frac{1}{W_i^2}; \frac{\sqrt{k}}{W_i}\right\} \log\left(\frac{d}{\delta_k W_{i^*}}\right) \log\left(\frac{k}{W_{i^*}}\right),$$

where κ is a constant depending only on L, ρ and M.

Moreover, since the inputs of **Optim** at its q^{th} call when executing **Select** $(S, \delta_k, 1)$ are: $(S, \delta_k/2^q, 1/4^q)$. Hence, (by design of Algorithm 5) we have:

$$m^{(q)} \leq \kappa k^2 4^q \log\left(\frac{2^q}{\delta_k}\right), \tag{40}$$

where κ depends on L, M, and ρ. We therefore have:

$$\begin{aligned} C_{\textbf{Optim}} &\lesssim \sum_{q=1}^{p} k m^{(q)} \\ &\leq \sum_{q=1}^{p} \kappa k^3 2^{2q} \log\left(\frac{2^q}{\delta_k}\right) \\ &\leq \kappa k^3 2^{2(p+1)} \log\left(\frac{2^p}{\delta_k}\right). \end{aligned}$$

We conclude applying Lemma I.8: with probability at least $1 - 3\delta_k$,

$$C_{\textbf{Optim}} \leq \kappa k^3 \max\left\{\frac{1}{W_{i^*}^2}, \frac{\sqrt{k}}{W_{i^*}}\right\} \log\left(\frac{k}{\delta_k W_{i^*}}\right),$$

where κ is a factor depending only on L, M and ρ.

J Lower bound on the scores Z_i^S:

Let us denote $(\beta_{(i)}^{S^*})_i$ the reordered coefficients of β^{S^*}: $|\beta_{(1)}^{S^*}| \geq ... \geq |\beta_{(s^*)}^{S^*}|$. Lemma J.1 provides a lower bound for $\max_{i \in [d] \setminus S} \left|Z_i^S\right|$.

Lemma J.1. *Suppose Assumptions 1, 2, 3 and 4 hold. Assume that $S \subsetneq S^*$ and denote $k := |S|$, we have:*

$$\max_{i \in [d] \setminus S} \left|Z_i^S\right| \geq \sqrt{\frac{\rho^3}{L}} \frac{1}{\sqrt{s^* - k}} \left\|\beta^{S^*} - \beta^S\right\|_2 \geq \sqrt{\frac{\rho^3}{L}} \frac{1}{\sqrt{s^* - k}} \left\|\beta_{S^* \setminus S}^{S^*}\right\|_2.$$

In this section we prove Lemma J.1, we begin by presenting the following technical lemmas adapted from Zhang [2009] to fit the random design.

Claim J.2. *Suppose Assumptions 1 and 3 hold. Then for all $i \in [d]$: $\rho \leq \mathbb{E}[x_i^2] \leq L$.*

Claim J.2 is a direct consequence of Assumption 3 stating that the eigenvalues of Σ_S are lower bounded by ρ and upper bounded by L, and the observation that $\mathbb{E}\left[x_i^2\right]$ are the diagonal terms of Σ_S.

Lemma J.3. *Let x, y and z be real valued bounded and centered random variables, such that $\mathbb{E}\left[x^2\right] = 1$. We have:*

$$\inf_{\alpha \in \mathbb{R}} \mathbb{E}\left[(y + \alpha x - z)^2\right] = \mathbb{E}\left[(y-z)^2\right] - \frac{1}{\mathbb{E}\left[x^2\right]}\mathbb{E}\left[x\,(y-z)\right]^2.$$

Proof. The proof follows from simple algebra, the minimum is attained for $\alpha = -\frac{\mathbb{E}[x(y-z)]}{\mathbb{E}[x^2]}$. □

Lemma J.4. *Let Assumptions 1, 2, 3 and 4 hold, consider a fixed subset $S \subsetneq S^*$ and denote $k := |S|$. We have the following:*

$$\inf_{\alpha \in \mathbb{R}, i \in S^* \setminus S} \mathbb{E}\left[\left(x^t\beta^S + \alpha\beta_i^{S^*}x_i - y\right)^2\right] \leq \mathbb{E}\left[\left(x^t\beta^S - y\right)^2\right] - \frac{1}{s^* - k}\frac{\rho}{L}\mathbb{E}\left[\left(x^t(\beta^{S^*} - \beta^S)\right)^2\right].$$

Proof. Let $\eta \in \mathbb{R}$, we have:

$$\begin{aligned}
\min_{i \in S^* \setminus S} \mathbb{E}\left[\left(x^t\beta^S + \eta\beta_i^{S^*}x_i - y\right)^2\right] &\leq \frac{1}{s^* - k}\sum_{i \in S^* \setminus S}\mathbb{E}\left[\left(x^t\beta^S + \eta\beta_i^{S^*}x_i - y\right)^2\right] \\
&\leq \mathbb{E}\left[\left(x^t\beta^S - y\right)^2\right] + \frac{1}{s^* - k}\sum_{i \in S^* \setminus S}\eta^2\left(\beta_i^{S^*}\right)^2\mathbb{E}\left[x_i^2\right] \\
&\quad + \frac{1}{s^* - k}\sum_{i \in S^* \setminus S} 2\eta\beta_i^{S^*}\mathbb{E}\left[x_i\left(x^t\beta^S - y\right)\right].
\end{aligned}$$

Recall that optimality of β^S implies that for all $i \in S$: $\mathbb{E}\left[x_i\left(x^t\beta^S - y\right)\right] = 0$. Hence:

$$\begin{aligned}
\sum_{i \in S^* \setminus S}\beta_i^{S^*}\mathbb{E}\left[x_i\left(x^t\beta^S - y\right)\right] &= \sum_{i \in S^* \setminus S}\left(\beta_i^{S^*} - \beta_i^S\right)\mathbb{E}\left[x_i\left(x^t\beta^S - y\right)\right] \\
&= \sum_{i \in S^*}\left(\beta_i^{S^*} - \beta_i^S\right)\mathbb{E}\left[x_i\left(x^t\beta^S - y\right)\right] \\
&= \sum_{i \in S^*}\left(\beta_i^{S^*} - \beta_i^S\right)\mathbb{E}\left[x_i\left(x^t\beta^S - x^t\beta^{S^*}\right)\right] \\
&= \mathbb{E}\left[\left(\beta^{S^*} - \beta^S\right)^t x\left(x^t\beta^S - x^t\beta^{S^*}\right)\right] \\
&= \mathbb{E}\left[\left(x^t\left(\beta^{S^*} - \beta^S\right)\right)^2\right].
\end{aligned}$$

Therefore:

$$(s^*-k)\min_{i\in S^*\setminus S}\mathbb{E}\left[\left(x^t\beta^S+\eta\beta_i^{S^*}x_i-y\right)^2\right]$$
$$\leq (s^*-k)\,\mathbb{E}\left[\left(x^t\beta^S-y\right)^2\right]$$
$$+\eta^2\sum_{i\in S^*\setminus S}\mathbb{E}\left[x_i^2\right]\left(\beta_i^{S^*}-\beta_i^S\right)^2+2\eta\mathbb{E}\left[\left(x^t\left(\beta^{S^*}-\beta^S\right)\right)^2\right].$$

Optimizing over η we obtain:

$$\min_{\eta\in\mathbb{R},i\in S^*\setminus S}\mathbb{E}\left[\left(x^t\beta^S+\eta\left(\beta_i^{S^*}-\beta_i^S\right)x_i-y\right)^2\right]\leq\mathbb{E}\left[\left(x^t\beta^S-y\right)^2\right]-\frac{1}{s^*-k}\frac{\mathbb{E}\left[\left(x^t\left(\beta^{S^*}-\beta^S\right)\right)^2\right]^2}{\sum_{i\in S^*}\mathbb{E}\left[x_i^2\right]\left(\beta_i^{S^*}-\beta_i^S\right)^2}.$$

Observe that: $\mathbb{E}\left[\left(x^t\left(\beta^{S^*}-\beta^S\right)\right)^2\right]=\left\|\Sigma_{S^*}^{1/2}\left(\beta^{S^*}-\beta^S\right)\right\|_2^2\geq\rho\left\|\beta^{S^*}-\beta^S\right\|_2^2$. Moreover, $\mathbb{E}\left[x_i^2\right]\leq L$. We plug in this inequality into the above and obtain the announced conclusion. □

Now we prove Lemma J.1. Using Lemma J.3 we have:

$$\inf_{\alpha\in\mathbb{R},i\in S^*\setminus S}\mathbb{E}\left[\left(x^t\beta^S+\alpha\beta_i^{S^*}x_i-y\right)^2\right]=\mathbb{E}\left[(y-x^t\beta^S)^2\right]-\max_{i\in S^*\setminus S}\frac{1}{\left(\beta_i^{S^*}\right)^2\mathbb{E}\left[x_i^2\right]}\mathbb{E}\left[\beta_i^{S^*}x_i\left(x^t\beta^S-y\right)\right]^2,$$

which is equivalent to:

$$\max_{i\in S^*\setminus S}\frac{1}{\sqrt{\mathbb{E}\left[x_i^2\right]}}\mathbb{E}\left[x_i\left(x^t\beta^S-y\right)\right]=\left(\mathbb{E}\left[(y-x^t\beta^S)^2\right]-\inf_{\alpha\in\mathbb{R},i\in S^*\setminus S}\mathbb{E}\left[\left(x^t\beta^S+\alpha\left(\beta_i^{S^*}-\beta_i^S\right)x_i-y\right)^2\right]\right)^{1/2}$$

Using Lemma J.4, we have:

$$\max_{i\in S^*\setminus S}\frac{1}{\sqrt{\mathbb{E}\left[x_i^2\right]}}\mathbb{E}\left[x_i\left(x^t\beta^S-y\right)\right]\geq\left(\frac{1}{s^*-k}\frac{\rho}{L}\mathbb{E}\left[\left(x^t(\beta^{S^*}-\beta^S)\right)^2\right]\right)^{1/2}.\tag{41}$$

Now we use Claim J.2 and inequality (41):

$$\begin{aligned}\max_{i\in S^*\setminus S}\mathbb{E}\left[x_i\left(x^t\beta^S-y\right)\right]&\geq\max_{i\in S^*\setminus S}\sqrt{\frac{\rho}{\mathbb{E}\left[x_i^2\right]}}\mathbb{E}\left[x_i\left(x^t\beta^S-y\right)\right]\\&\geq\sqrt{\rho}\max_{i\in S^*\setminus S}\frac{1}{\sqrt{\mathbb{E}\left[x_i^2\right]}}\mathbb{E}\left[x_i\left(x^t\beta^S-y\right)\right]\\&\geq\sqrt{\rho}\left(\frac{1}{s^*-k}\frac{\rho}{L}\mathbb{E}\left[\left(x^t(\beta^S-\beta^{S^*})\right)^2\right]\right)^{1/2}\\&\geq\frac{\rho}{\sqrt{L}}\frac{1}{\sqrt{s^*-k}}\left\|\Sigma_{S^*}^{1/2}\left(\beta^{S^*}-\beta^S\right)\right\|_2\\&\geq\sqrt{\frac{\rho^3}{L}}\frac{1}{\sqrt{s^*-k}}\left\|\beta^{S^*}-\beta^S\right\|_2.\end{aligned}$$

The conclusion follows from the definition $Z_i^S=\mathbb{E}\left[x_i\left(x^t\beta^S-y\right)\right]$.

K Computational Complexity Comparisons

K.1 Proof of Corollary 5.3:

Suppose Assumptions 1, 2, 3 and 4 hold. Consider the procedure **Select** given by Algorithm 4, **Try-Select** given by Algorithm 6, and **Optim** as in Algorithm 5. Assume that $S \subsetneq S^*$ and denote $k := |S|$. Using the result of theorem 5.2 we have with probability at least $1 - \delta$:

$$C^S_{\textbf{Optim}} \leq \kappa k^3 \max\left\{\frac{1}{Z^2_{i^*}}; \frac{\sqrt{k}}{Z_{i^*}}\right\} \log\left(\frac{\bar{k}}{\delta\,|Z_{i^*}|}\right);$$
$$C^S_{\textbf{Try-Select}} \leq \kappa d \max\left\{\frac{1}{Z^2_{i^*}}; \frac{\sqrt{\bar{k}}}{Z_{i^*}}\right\} \log\left(\frac{d}{\delta\,|Z_{i^*}|}\right) \log\left(\frac{\bar{k}}{|Z_{i^*}|}\right);$$

where $|Z_{i^*}| = \max_{i\in[d]}\{|Z_i|\}$, and κ is a constant depending on ρ, L, M and μ (for which the value may vary from line to line).

We plug-in the inequality of lemma J.1 and obtain:

$$C^S_{\textbf{Optim}} \leq \kappa k^3 \max\left\{\frac{s^* - k}{\left\|\beta^{S^*}_{S^*\setminus S}\right\|_2^2}; \frac{\sqrt{k(s^* - k)}}{\left\|\beta^{S^*}_{S^*\setminus S}\right\|_2}\right\} \log\left(\frac{\bar{k}}{\delta\left\|\beta^{S^*}_{S^*\setminus S}\right\|_2}\right);$$
$$C^S_{\textbf{Try-Select}} \leq \kappa d \max\left\{\frac{s^* - k}{\left\|\beta^{S^*}_{S^*\setminus S}\right\|_2^2}; \frac{\sqrt{k(s^* - k)}}{\left\|\beta^{S^*}_{S^*\setminus S}\right\|_2}\right\} \log\left(\frac{d}{\delta\left\|\beta^{S^*}_{S^*\setminus S}\right\|_2}\right) \log\left(\frac{\bar{k}}{\left\|\beta^{S^*}_{S^*\setminus S}\right\|_2}\right);$$

Hence, using the fact that $|S^* \setminus S| = s^* - k$ and the definition of $\tilde{\beta}_{(k+1)}$:

$$C^S_{\textbf{Optim}} \leq \kappa k^3 \max\left\{\frac{1}{\tilde{\beta}^2_{(k+1)}}; \frac{\sqrt{k}}{\tilde{\beta}_{(k+1)}}\right\} \log\left(\frac{\bar{k}}{\delta\tilde{\beta}^2_{(k+1)}}\right);$$
$$C^S_{\textbf{Try-Select}} \leq \kappa d \max\left\{\frac{1}{\tilde{\beta}^2_{(k+1)}}; \frac{\sqrt{k}}{\tilde{\beta}_{(k+1)}}\right\} \log^2\left(\frac{\bar{k}}{\delta\tilde{\beta}^2_{(k+1)}}\right);$$

The following claim concludes the proof:

Claim K.1. *Under the assumptions of theorem 5.2:*

$$\tilde{\beta}_{(k+1)} \leq \frac{1}{\sqrt{\rho s^*}}$$

Proof. We have by definition of $\tilde{\beta}_{(k+1)}$:

$$\begin{aligned}\tilde{\beta}^2_{(k+1)} &= \frac{1}{s^* - k} \sum_{i=k+1}^{s^*} \beta^2_{(i)} \\ &\leq \frac{s^* - k}{s^*} \quad \frac{1}{s^* - k} \sum_{i=k+1}^{s^*} \beta^2_{(i)} + \frac{k}{s^*} \quad \frac{1}{k} \sum_{i=1}^{k} \beta^2_{(i)} \\ &\leq \frac{1}{s^*} \sum_{i=1}^{s^*} \beta^2_{(i)} = \frac{1}{\rho s^*}\end{aligned}$$

□

K.2 Computational complexity of the Orthogonal Matching Pursuit

We consider OMP (Algorithm 1) as a benchmark and show that OOMP is more efficient in time complexity. OMP was initially derived under the fixed design setting presented below:

Let $\boldsymbol{X} = [x_1, \ldots, x_d] \in \mathbb{R}^{n \times d}$ an $n \times d$ data matrix and $\boldsymbol{Y} = [y_1, \ldots, y_n]$ a response vector generated according to the sparse model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta^{S^*} + \boldsymbol{\epsilon}.$$

Where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]$ is a zero mean random noise vector and $\text{support}(\beta^{S^*}) = S^*$. Define the following quantities:

$$\hat{\mu}_{S^*} = \max_{i \notin S^*} \left\| \left(\boldsymbol{X}^t_{S^*} \boldsymbol{X}_{S^*} \right)^{-1} \boldsymbol{X}^t_{S^*} \boldsymbol{x}_i \right\|_1,$$

and let $\hat{\rho}_{S^*}$ be the least eigenvalue of the empirical covariance matrix $\hat{\boldsymbol{\Sigma}}_{S^*} = \frac{1}{n} \boldsymbol{X}^t_{S^*} \boldsymbol{X}_{S^*}$.

OMP theoretical guarantees

Assumption 5. *Assume that:*

- $\hat{\mu}_{S^*} < 1$ *and* $\hat{\rho}_{S^*} > 0$.
- ϵ_i, *for* $i \in [1, n]$ *are i.i.d random variables bounded by* σ.

Theorem K.2 (Zhang [2009])**.** *Consider the OMP procedure (Algorithm 1), suppose Assumption 5 holds. Then for all* $\delta \in (0, 1)$, *if the sample size* n *satisfies:*

$$n \geq \frac{18\sigma^2 \log(4d/\delta)}{(1 - \hat{\mu}_{S^*})^2 \hat{\rho}^2_{S^*} \min_{i \in S^*} |\beta^{S^*}_i|^2}, \tag{42}$$

then the output of the procedure Algorithm 1 recovers $S = S^*$, *with probability at least* $1 - \delta$.

OMP computational complexity: We derive the computational complexity of OMP. Consider one iteration of Algorithm 1 and denote $k := |S|$. We assimilate the command:

$$i \leftarrow \text{argmax}_{j \notin S} |\boldsymbol{X}^t_{\cdot j}(\boldsymbol{Y} - \boldsymbol{X}\bar{\beta})| \tag{43}$$

to **Try-Select** and denote $C^{omp}_{\textbf{Try-Select},k}$ its computational complexity. Moreover, we assimilate the command:

$$\bar{\beta} \leftarrow \underset{\text{supp}(\beta) \subseteq S}{\text{argmin}} \|\boldsymbol{X}\beta - \boldsymbol{Y}\|^2 \tag{44}$$

to **Optim** and denote $C^{omp}_{\textbf{Optim},k}$ its computational complexity.We assume the OMP is run with n^{OMP} prescribed by Theorem K.2 for exact support recovery. We introduce the following additional notation: $a \simeq b$ if there exists numerical constants c_1 and c_2 such that: $a \leq c_1 b$ and $b \leq c_2 a$.

Lemma K.3. *Consider Algorithm 1 with inputs $(\boldsymbol{X}, \boldsymbol{Y}, \delta)$, and suppose assumption 5 holds. Then if n satisfies* (42) *we have:*

$$C^{omp}_{\textbf{\textit{Optim}},k} \simeq \frac{\sigma^2 k \log(d/\delta)}{(1-\hat{\mu}_{S^*})^2 \hat{\rho}^2_{S^*} \min_{i \in S^*} |\beta^{S^*}|^2};$$
$$C^{omp}_{\textbf{\textit{Try-Select}},k} \simeq \frac{\sigma^2 d \log(d/\delta)}{(1-\hat{\mu}_{S^*})^2 \hat{\rho}^2_{S^*} \min_{i \in S^*} |\beta^{S^*}|^2}.$$

Proof. Performing command (43) requires computing $\boldsymbol{X}^t\left(\boldsymbol{Y} - \boldsymbol{X}\bar{\beta}\right)$ and selecting the maximum of a list of (at most) d elements, thus $C^{omp}_{\textbf{Try-Select},k} \simeq dn^{\text{OMP}}$. Command (44) can be performed using a rank one update. Thus: $C^{omp}_{\textbf{Optim},k} \simeq kn^{\text{OMP}}$. To conclude we use Theorem K.2, which prescribes:

$$n^{\text{OMP}} = \frac{18\sigma^2 \log(4d/\delta)}{(1-\hat{\mu}_{S^*})^2 \hat{\rho}^2_{S^*} \min_{i \in S^*} |\beta^{S^*}_i|^2}.$$

□

Hence, the computational complexity for full support recovery using OMP satisfies:

$$C^{\text{OMP}} = \mathcal{O}\left(\frac{s^* d \log(d/\delta)}{\min_{i \in S^*}\{(\beta^*_i)^2\}}\right) \tag{45}$$

K.3 SSR computational complexity

SSR (Streaming Sparse Regression) is an online procedure guaranteed to perform well under similar conditions to the Lasso Steinhardt et al. [2014]. Theoretical guarantees show that if the number of iterations is large enough the support recovery is achieved with high probability.

Theorem 8.2 in Steinhardt et al. [2014] states that, the output vector $\hat{\beta}_T$ satisfies with probability at least $1 - 5\delta$, $\text{supp}(\hat{\beta}_T) \subseteq S^*$ and:

$$\left\|\hat{\beta}_T - \beta^*\right\|^2 = \mathcal{O}\left(\frac{(s^*)^2 \log(d \log(T)/\delta)}{T}\right), \tag{46}$$

where we used the bound $B \leq 6\sqrt{s^*}\frac{M^2}{\sqrt{\rho}}$. Hence, a sufficient condition to achieve the full support recovery $\text{supp}(\hat{\beta}_T) = S^*$ is : $\left\|\hat{\beta}_T - \beta^*\right\|^2 \leq \min_{i\in S^*}\{(\beta_i^*)^2\}$. Using (46) leads to the following bound on the number of iterations to recover all the support of β^*:

$$T = \mathcal{O}\left(\frac{(s^*)^2 \log(d/\delta)}{\min_{i\in S^*}\{(\beta_i^*)^2\}}\right)$$

One iteration of Algorithm 2 in Steinhardt et al. [2014] has a computational complexity of $\mathcal{O}(d)$. Hence, the total computational complexity for full support recovery C^{SSR} satisfies:

$$C^{\text{SSR}} = \mathcal{O}\left(\frac{(s^*)^2 d \log(d/\delta)}{\min_{i\in S^*}\{(\beta_i^*)^2\}}\right) \tag{47}$$

K.4 Proof of Corollary 5.4

Assuming that $d > (s^*)^3$, we have for every $S \subset S^*$: $C^S_{\textbf{Optim}} \leq C^S_{\textbf{Try-Select}}$. Hence, using corollary 5.3, we have:

$$C^{OOMP} \leq \kappa d \sum_{i=1}^{s^*} \frac{1}{\tilde{\beta}^2_{(s^*-i)}} \log\left(\frac{d}{\delta\beta^2_{(s^*)}}\right) \log\left(\frac{s^*}{\beta^2_{(s^*)}}\right) \tag{48}$$

We plug-in the bounds in (45) and (47):

$$C^{OOMP} \leq \kappa \sum_{i=1}^{s^*} \frac{\beta^2_{(s^*)}}{\tilde{\beta}^2_{(s^*-i)}} \log\left(\frac{d}{\delta\beta^2_{(s^*)}}\right) \log\left(\frac{s^*}{\beta^2_{(s^*)}}\right) \frac{C^{\text{OMP}}}{s^* \log(d/\delta)}. \tag{49}$$

$$C^{OOMP} \leq \kappa \sum_{i=1}^{s^*} \frac{\beta^2_{(s^*)}}{\tilde{\beta}^2_{(s^*-i)}} \log\left(\frac{d}{\delta\beta^2_{(s^*)}}\right) \log\left(\frac{s^*}{\beta^2_{(s^*)}}\right) \frac{C^{\text{SSR}}}{(s^*)^2 \log(d/\delta)}. \tag{50}$$

$$\tag{51}$$

Recall that:

$$\frac{\log\left(\frac{d}{\delta\beta^2_{(s^*)}}\right) \log\left(\frac{s^*}{\beta^2_{(s^*)}}\right)}{\log(d/\delta)} \leq \log^2\left(\frac{s^*}{\beta^2_{(s^*)}}\right).$$

We conclude that:

$$\frac{C^{\text{OOMP}}}{C^{\text{OMP}}} \leq \kappa \log^2\left(\frac{s^*}{\beta^2_{(s^*)}}\right) \frac{1}{s^*} \sum_{i=1}^{s^*} \frac{\beta^2_{(s^*)}}{\tilde{\beta}^2_{(i)}} \quad C^{\text{OMP}};$$

$$\frac{C^{\text{OOMP}}}{C^{\text{SSR}}} \leq \kappa \log^2\left(\frac{s^*}{\beta^2_{(s^*)}}\right) \frac{1}{(s^*)^2} \sum_{i=1}^{s^*} \frac{\beta^2_{(s^*)}}{\tilde{\beta}^2_{(i)}} \quad C^{\text{SSR}};$$

where κ is a constant depending only on L, M, ρ and μ.

K.5 A specific scenario: Polynomially decaying coefficients

We consider the case where the coefficients of β^* are given by

$$\beta_q^* = \frac{1}{\sqrt{s^*}} \left(1 - \frac{q-1}{s^*}\right)^\gamma, \quad \text{for } q \in [s^*], \tag{52}$$

with $\gamma > 0$. We omit the superscript $*$ to ease notations, in the remainder of this section, all the inequalities and equalities are up to factors depending only only on ρ, L, M and μ.

The following lemma provides a bound on the computational complexity of OOMP, OMP and SSR.

Lemma K.4. *Under the assumptions of Theorem 5.2, suppose that $d > (s^*)^3$ and the coefficients of β^* are given by* (52)*. Then with probability at least $1 - \delta$: If $\gamma \neq \frac{1}{2}$:*

$$C^{OOMP} \leq \kappa d \left\{ \frac{2\gamma\,(2\gamma+1)}{|2\gamma-1|} s^{2\gamma+1} + \frac{2\gamma+1}{|2\gamma-1|} s^2 \right\} \log\left(d/\delta\right) \log\left(s\right)$$
$$C^{OMP} \simeq d s^{2\gamma+2} \log(d/\delta)$$

If $\gamma = \frac{1}{2}$:

$$C^{OOMP} \leq \kappa d s^2 \log^2(s) \log\left(d/\delta\right)$$
$$C^{OMP} \simeq d s^3 \log(d/\delta)$$

Proof. Recall that $\tilde{\beta}^2_{(s-k+1)} = \frac{1}{k} \sum_{i=s-k+1}^{s} \beta_i^2$.

If $\gamma \neq \frac{1}{2}$:

$$\begin{aligned}
\sum_{k=0}^{s-1} \frac{1}{\tilde{\beta}^2_{(s-k)}} &= \sum_{k=0}^{s-1} \frac{s-k}{\sum_{q=k+1}^{s} \beta_q^2} \\
&\leq \sum_{k=0}^{s-1} \frac{s-k}{\frac{1}{s}\sum_{q=k+1}^{s} \left(1-\frac{q-1}{s}\right)^{2\gamma}} \\
&\leq \sum_{k=0}^{s-1} \frac{s^{2\gamma+1}\,(s-k)}{\sum_{q=1}^{s-k} q^{2\gamma}} \\
&\leq \sum_{k=0}^{s-1} \frac{s^{2\gamma+1}(s-k)}{\frac{1}{2\gamma+1}(s-k)^{2\gamma+1}} \\
&\leq (2\gamma+1)\sum_{k=0}^{s-1} \frac{s^{2\gamma+1}}{(s-k)^{2\gamma}} \\
&\leq (2\gamma+1)s\sum_{k=0}^{s-1} \left(1-\frac{k}{s}\right)^{-2\gamma} \\
&\leq (2\gamma+1)s^2\left(\frac{1}{s}\sum_{k=0}^{s-2} \left(1-\frac{k}{s}\right)^{-2\gamma} + s^{2\gamma-1}\right) \\
&\leq (2\gamma+1)s^2\left(\frac{1}{2\gamma-1}\left(\frac{1}{s^{1-2\gamma}}-1\right) + s^{2\gamma-1}\right).
\end{aligned}$$

If $\gamma = \frac{1}{2}$:

$$\begin{aligned}
\sum_{k=0}^{s-1} \frac{1}{\tilde{\beta}^2_{(s-k)}} &= \sum_{k=0}^{s-1} \frac{s-k}{\sum_{q=k+1}^{s} \beta_q^2} \\
&\leq \sum_{k=0}^{s-1} \frac{s-k}{\frac{1}{s}\sum_{q=k+1}^{s} \left(1-\frac{q-1}{s}\right)} \\
&\leq \sum_{k=0}^{s-1} \frac{s^2\,(s-k)}{\sum_{q=1}^{s-k} q} \\
&\leq \sum_{k=0}^{s-1} \frac{s^2(s-k)}{\frac{1}{2}(s-k)^2} \\
&\leq 2\sum_{k=0}^{s-1} \frac{s^2}{(s-k)} \\
&\leq s^2 \log(s),
\end{aligned}$$

which gives the result.

□

Using the lemma above, we conclude that, if $d > (s^*)^3$:

$$\frac{C^{\mathrm{OOMP}}}{C^{\mathrm{OMP}}} \leq \kappa \frac{\log^2(s)}{s^{\min\{2\gamma,1\}}}$$