# Recovery Guarantees for Distributed-OMP

**Chen Amiraz**
Weizmann Institute of Science

**Robert Krauthgamer**
Weizmann Institute of Science

**Boaz Nadler**
Weizmann Institute of Science

## Abstract

We study distributed schemes for high-dimensional sparse linear regression, based on orthogonal matching pursuit (`OMP`). Such schemes are particularly suited for settings where a central fusion center is connected to end machines, that have both computation and communication limitations. We prove that under suitable assumptions, distributed-`OMP` schemes recover the support of the regression vector with communication per machine linear in its sparsity and logarithmic in the dimension. Remarkably, this holds even at low signal-to-noise-ratios, where individual machines are unable to detect the support. Our simulations show that distributed-`OMP` schemes are competitive with more computationally intensive methods, and in some cases even outperform them.

## 1 INTRODUCTION

Sparse linear regression is a fundamental problem in machine learning, statistics and signal processing. Indeed, sparsity is a natural and widely applied modeling assumption in high dimensional settings. A sparsity assumption gives rise to the variable selection problem, of identifying a small subset of variables which are most informative for a given prediction problem. Here, we consider the popular sparse linear regression model with random noise,

$$y = \mathbf{x}^\top \boldsymbol{\theta} + \sigma\xi, \tag{1}$$

where $y \in \mathbb{R}$ is the response, $\mathbf{x} \in \mathbb{R}^d$ is a vector of explanatory variables, $\boldsymbol{\theta} \in \mathbb{R}^d$ is the unknown vector of regression coefficients, $\xi \in \mathbb{R}$ is a standard normal random variable, i.e., $\xi \sim \mathcal{N}(0,1)$, and $\sigma > 0$ is the noise level. We consider a high dimensional setting $d \gg 1$ with a vector $\boldsymbol{\theta}$ of sparsity $K = \|\boldsymbol{\theta}\|_0 \ll d$. In a centralized setting, given $N$ samples from (1), common tasks are to accurately estimate $\boldsymbol{\theta}$ as well as its support $\mathcal{S} = \mathrm{supp}(\boldsymbol{\theta}) = \{i \mid \theta_i \neq 0\}$. Many methods have been proposed and analyzed to solve these tasks, including combinatorial algorithms, linear programming, greedy approaches and regularization schemes (Mallat and Zhang, 1993; Tibshirani, 1996; Chen et al., 2001; Miller, 2002; Candes and Tao, 2005; Tropp and Gilbert, 2007; Blumensath and Davies, 2008; Needell and Tropp, 2009; Dai and Milenkovic, 2009; Fletcher et al., 2009; Bertsimas et al., 2016; Hastie et al., 2020; Amir et al., 2021).

In various contemporary applications, the data is stored across multiple machines. Moreover, due to communication or privacy constraints, the data at each machine cannot be sent to other machines in the network. Such cases bring about various distributed learning problems, see Wimalajeewa and Varshney (2017); Jordan et al. (2019) and references therein.

A common distributed setting, which we also consider here, consists of $M$ machines connected in a star topology to a fusion center, with each machine having for simplicity an equal number of samples, $n = N/M$. For the sparse model (1), some distributed methods attempt to recover the centralized solution that would have been computed by the fusion center, if it had access to all $N = nM$ samples of the $M$ machines. Examples include optimization-based methods (Mateos et al., 2010; Ling and Tian, 2011; Mota et al., 2011; Ling et al., 2012; Fosson et al., 2016; Smith et al., 2018; Scaman et al., 2019; SarcheshmehPour et al., 2023), Bayesian approaches (Makhzani and Valaee, 2013; Khanna and Murthy, 2016), and greedy schemes (Sundman et al., 2012; Li et al., 2015; Patterson et al., 2014; Han et al., 2015; Chouvardas et al., 2015). These methods are in general communication intensive, as they are iterative and may require many rounds to converge. A simpler single round divide-and-conquer scheme, is for each machine to send its own estimate of $\boldsymbol{\theta}$ and for the fusion center to average these estimates. For a wide range of problems, the resulting estimator

Proceedings of the 27$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

has a risk comparable to that of the centralized solution (Rosenblatt and Nadler, 2016; Wang et al., 2017; Jordan et al., 2019; Liu et al., 2023). For the sparse linear regression model (1), Lee et al. (2017) and Battey et al. (2018) proposed a single round distributed debiased-Lasso scheme, and proved that under suitable conditions it achieves the same error rate as the centralized solution. Yet, these debiased-Lasso methods have two limitations: (i) the communication per machine is at least linear in $d$; and (ii) the computational costs are considerable, as each machine has to solve $d+1$ Lasso problems. Barghi et al. (2021) and Fonseca and Nadler (2023) proposed debiased-Lasso methods with much less communication, where each machine sends to the center only the indices of its few largest coordinates.

We consider distributed estimation of the sparse vector $\boldsymbol{\theta}$ in the model (1), under the following setting: The $M$ end machines have both limited processing power and a restricted communication budget. This is motivated by modern applications where end machines are computationally weak, but collect high dimensional data. For example, in spectrum sensing, a network of sensors continuously monitor and collect high dimensional data, and repeatedly need to estimate the current vector $\boldsymbol{\theta}$. In these settings, computationally intensive methods such as debiased Lasso may be infeasible or prohibitively slow. In addition, under communication constraints, regardless of computational considerations, most of the above methods are not applicable in high dimensions, as their communication per machine is at least linear in $d$.

As the quantity of interest $\boldsymbol{\theta}$ is $K$-sparse with $K \ll d$, this gives rise to the following challenge: develop a scheme that accurately estimates the vector $\boldsymbol{\theta}$ with number of operations per machine linear in $d$ and communication *sublinear* in $d$, and derive theoretical guarantees for it. Here we focus on accurately estimating the support of $\boldsymbol{\theta}$. Indeed, as discussed in Battey et al. (2018); Fonseca and Nadler (2023), given an accurate estimate of the support, an additional single round of communication allows distributed estimation of $\boldsymbol{\theta}$ with the same error rate as in the centralized setting.

A natural base algorithm for machines with low computational resources is Orthogonal Matching Pursuit (`OMP`), as it is one of the fastest methods for sparse recovery (Chen et al., 1989; Pati et al., 1993; Mallat and Zhang, 1993). Several distributed-`OMP` schemes, which are computationally fast and incur little communication, were proposed in Duarte et al. (2005); Wimalajeewa and Varshney (2013); Sundman et al. (2014). In terms of theory, Tropp (2004); Tropp and Gilbert (2007); Cai and Wang (2011) derived guarantees for exact support recovery by `OMP` at a single machine, both with and without noise. However, their results do not extend to a distributed setting if the signal-to-noise ratio (SNR) at each machine is low. To the best of our knowledge, the only work to derive support recovery guarantees for distributed-`OMP` methods is by Wimalajeewa and Varshney (2014). However, their analysis is restricted to a noise-less compressed-sensing setting, with samples $\mathbf{x}_i$ that are random and independent across machines, and their proofs rely heavily on the symmetry between all non-support variables. Another related work, by Amiraz et al. (2022), considered distributed sparse mean estimation. This problem can be viewed as a special case of distributed sparse linear regression where the design matrices are orthogonal, whereas we study incoherent design matrices. This distinction is crucial, because similarly to Wimalajeewa and Varshney (2014), the proof of Amiraz et al. (2022) is symmetry-based, and is inapplicable in our framework.

Our key contribution is the derivation of a recovery guarantee for a distributed-`OMP` scheme, see Theorem 4.1. Remarkably, our guarantee holds even at low SNRs, where each individual machine fails to recover the support. The main challenge in our analysis is that the samples $\mathbf{x}_i$, assumed deterministic, may be similar (or even identical) across machines. Hence, at low SNRs, several machines might send the *same* incorrect support variable to the fusion center. Deriving a theoretical guarantee in this case requires a different and more delicate analysis than that of previous works. Specifically, to bound the probability that a non-support variable is sent to the fusion center we use recent lower bounds on the maximum of correlated Gaussian random variables (Lopes and Yao, 2022).Thus, our analysis goes significantly beyond the limitations of previous works by providing theoretical guarantees in a more general setting, where the design matrices may be correlated, deterministic or even structured, and for noisy signals. Furthermore, our analysis provides insight into how distributed-`OMP` methods can achieve exact support recovery even at low SNRs where individual machines fail to do so.

To complement our theoretical analysis, we compare via simulations the support-recovery success of several algorithms including distributed-`OMP` and debiased Lasso schemes (Lee et al., 2017; Battey et al., 2018; Barghi et al., 2021). In addition we compare to distributed sure independence screening (`SIS`) schemes (Fan and Lv, 2008), which are also suitable for computationally weak machines. In distributed `SIS` schemes, each machine first excludes variables weakly correlated to the response, and then estimates the sparse vector $\boldsymbol{\theta}$ on the remaining ones via any appropriate algorithm. In our experiments we considered smoothly

clipped absolute deviation (SCAD) (Fan and Li, 2001) and OMP. Our simulations show that, as expected, the best performing scheme is debiased Lasso, but at the expense of significantly higher communication and computational costs. Interestingly, in comparison to a communication-restricted thresholded variant of debiased Lasso, distributed-OMP methods perform comparably, and in some cases even outperform it, while being orders of magnitude faster.

The rest of the paper is structured as follows. In Section 2 we formulate the distributed sparse linear regression problem. Section 3 presents the distributed-OMP algorithms that the paper focuses on. Our main theoretical contributions are outlined in Section 4. To support our theoretical results, we present and discuss several simulations in Section 5. We conclude the paper with a discussion in Section 6.

**Notation** We use the standard $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ notation to hide constants independent of the problem parameters and $\tilde{O}(\cdot)$ to hide terms polylogarithmic in $d$. For functions $f, g$, the notations $f = o(g)$ and $f \ll g$ mean that $f(d)/g(d) \to 0$ as $d \to \infty$. We say that an estimator $\hat{S}$ achieves exact support recovery with high probability if $\Pr\left[\hat{S} = \mathcal{S}\right] \to 1$ as both $d \to \infty$ and the number of machines $M = M(d) \to \infty$ at a suitable rate. The smallest integer larger than or equal to $x$ is denoted $\lceil x \rceil$. The set of integers $1, 2, \ldots, M$ is denoted as $[M]$. For a standard Gaussian $Z \sim \mathcal{N}(0,1)$, the complement of its cumulative distribution function is $\Phi^c(t) = \Pr[Z > t]$. We denote the inner product of two vectors $\mathbf{u}, \mathbf{v}$ by $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$.

## 2 PROBLEM SETUP

We consider linear regression with a sparse coefficient vector in a distributed setting, where $M$ machines are connected in a star topology to a fusion center. Each machine $m \in [M]$ holds $n$ samples from the sparse regression model (1), i.e., a design matrix $\mathbf{X}^{(m)} \in \mathbb{R}^{n \times d}$ and a response vector $\mathbf{y}^{(m)} \in \mathbb{R}^n$, related via

$$\mathbf{y}^{(m)} = \mathbf{X}^{(m)} \boldsymbol{\theta} + \sigma \boldsymbol{\xi}^{(m)}, \tag{2}$$

where $\boldsymbol{\xi}^{(m)} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $\sigma$ is the unknown noise level. While the $M$ machines may have the same or similar design matrices, their noises $\boldsymbol{\xi}^{(m)}$ are assumed to be independent. We assume $\boldsymbol{\theta}$ is $K$-*sparse*, namely $\|\boldsymbol{\theta}\|_0 = |\mathrm{supp}(\boldsymbol{\theta})| = K$, with the value of $K$ known to the center.

The problem we consider is exact recovery of the support of $\boldsymbol{\theta}$, which is a standard goal in sparse linear regression, and has been widely studied in both non-distributed and distributed settings. We study this problem under the constraints that the $M$ machines have limited computational resources and limited communication with the fusion center. This setting is relevant in various applications including distributed compressed sensing and sensor networks.

**Algorithm 1:** OMP_Step

**input :** $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \mathbb{R}^n$, support set $S$
**output:** support index $j$
compute $\hat{\boldsymbol{\theta}} = \mathrm{argmin}_{\mathbf{z} \in \mathbb{R}^d, \mathrm{supp}(\mathbf{z}) = S} \|\mathbf{y} - \mathbf{X}\mathbf{z}\|_2$
compute residual $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}$
output index $j = \arg\max \left\{ \frac{|\langle \mathbf{x}_i, \mathbf{r} \rangle|}{\|\mathbf{x}_i\|} : i \in [d] \right\}$

## 3 DISTRIBUTED OMP SCHEMES

OMP-based schemes are popular for sparse support recovery, and are highly attractive in distributed settings where computation and communication are limited. We consider two distributed OMP schemes to estimate the support of $\boldsymbol{\theta}$. Both schemes use the following subroutine, denoted OMP_Step, which performs a single step of the OMP algorithm, and outputs a new variable to be added to the current support set. As outlined in Algorithm 1, given a matrix $\mathbf{X}$, a vector $\mathbf{y}$, and a current support set $S$, the subroutine computes $\hat{\boldsymbol{\theta}}$, the least squares approximation of $\boldsymbol{\theta}$ on the support $S$ and its residual vector $\mathbf{r}$. It then outputs an index $j \in [d]$ whose column $\mathbf{x}_j$ has maximal correlation with $\mathbf{r}$. A key property of OMP_Step is the orthogonality of the residual to the columns of $\mathbf{X}$ in the set $S$. Hence, the output of OMP_Step is a new index $j \notin S$.

The simplest distributed OMP method is for each machine to separately run OMP for $K$ steps and send its $K$ locally-computed indices to the fusion center. The center estimates the support of $\boldsymbol{\theta}$ by the $K$ indices that received the largest number of votes. To cope with low-SNR regimes where the top $K$ indices at individual machines may not include all support indices, we propose a variant where each machine runs OMP for a *larger* number of steps and thus sends a support of size $L > K$. This scheme, which we call Distributed OMP (D-OMP), is outlined in Algorithm 2.

A second scheme, which we call Distributed Joint OMP (DJ-OMP), computes the support set one index at a time, using $K$ communication rounds. Starting with an empty support set $S_0 = \emptyset$, at each round $t = 1, \ldots, K$, the center sends the current set $S_{t-1}$ to the $M$ machines. Then, each machine calls OMP_Step and sends the resulting index $j^{(m,t)}$ to the center. At the end of each round, the center adds to the support set an index $j_t$ that received the most votes,

**Algorithm 2:** `Distributed OMP (D-OMP)`

**At each machine** $m = 1, \ldots, M$
  **input :** $\mathbf{X}^{(m)} \in \mathbb{R}^{n \times d}$, $\mathbf{y}^{(m)} \in \mathbb{R}^n$, integer $L$
  **output:** message $S_L^{(m)}$ to center
  initialize $S_0^{(m)} = \emptyset$
  **for** *round* $t = 1, \ldots, L$ **do**
    $j^{(m,t)} = \texttt{OMP\_Step}\left(\mathbf{X}^{(m)}, \mathbf{y}^{(m)}, S_{t-1}^{(m)}\right)$
    update support set $S_t^{(m)} = S_{t-1}^{(m)} \cup \left\{j^{(m,t)}\right\}$
  **end**
  send $S_L^{(m)}$ to the center
**At the fusion center**
  **input :** messages $\left\{S_L^{(m)}\right\}_{m \in [M]}$, sparsity $K$
  **output:** estimated support $S$
  for each index $j \in [d]$, calculate the number of votes it received $\mathbf{v}_j = \sum_{m \in [M]} \mathbb{1}\left\{j \in S_L^{(m)}\right\}$
  sort indices by number of votes, $\mathbf{v}_{\pi(1)} \geq \cdots \geq \mathbf{v}_{\pi(d)}$
  return $K$ indices with most votes $S = \{\pi(1), \ldots, \pi(K)\}$

**Algorithm 3:** `Distributed Joint OMP`

initialize $S_0 = \emptyset$
**for** *round* $t = 1, \ldots, K$ **do**
  **At each machine** $m = 1, \ldots, M$
    $j^{(m,t)} = \texttt{OMP\_Step}\left(\mathbf{X}^{(m)}, \mathbf{y}^{(m)}, S_{t-1}\right)$
    send index $j^{(m,t)}$ to fusion center
  **At the fusion center**
    **input :** messages $j^{(m,t)}$, sparsity $K$
    calculate number of votes for each index $j$, $\mathbf{v}_j^{(t)} = \sum_{m \in [M]} \mathbb{1}\left\{j = j^{(m,t)}\right\}$
    find most voted index $j_t = \text{argmax}_j \mathbf{v}_j^{(t)}$
    add $j_t$ to support set $S_t = S_{t-1} \cup \{j_t\}$
    send $j_t$ to all machines
    if $t = K$ output $S_K$
**end**

$S_t = S_{t-1} \cup \{j_t\}$. After $K$ rounds, the center outputs the support set $S_K$. Since `OMP_Step` outputs an index not in the current set $S_{t-1}$, at each round $t$ of `DJ-OMP`, a new index is indeed added by the center, $j_t \notin S_{t-1}$. This scheme is outlined in Algorithm 3.

**Computation and Communication Complexity.** Let us first analyze the number of operations in a single execution of `OMP_Step`. Given a support set $S$, computing $\hat{\boldsymbol{\theta}}$ via least squares involves multiplying a $|S| \times n$ matrix by its transpose, and then inverting the resulting $|S| \times |S|$ matrix. Next, finding the index $j$ most correlated to the residual requires $d$ inner products of vectors in $\mathbb{R}^n$. For $|S|$ sufficiently small, say $o\left(d^{1/3}\right)$, the computational cost of `OMP_Step` is dominated by the latter step whose cost is $O(nd)$.

We now compare the two schemes `DJ-OMP` and `D-OMP` with $L = K$. In terms of computational complexity, in both schemes each machine performs the same number of operations. Thus, for $K = o\left(d^{1/3}\right)$ their computational complexity per machine is $O(ndK)$. In terms of communication, in both schemes each machine sends (and in `DJ-OMP` also receives) a total of $K$ indices, and so the communication per machine is $O(K \log d)$ bits. The main difference is that `D-OMP` performs a single round, whereas `DJ-OMP` performs $K$ rounds. Hence, `DJ-OMP` requires synchronization and is slower in comparison to `D-OMP`.

**Related Works.** Various distributed-`OMP` methods were proposed in the past decade. Wimalajeewa and Varshney (2013) considered the same `D-OMP` scheme as we do, with $L = K$. In addition, they proposed a `DC-OMP` algorithm, which is similar to `DJ-OMP`. In `DC-OMP`, at each round, instead of adding just one index to the support, the fusion center adds all indices that received at least two votes. A distributed-`OMP` approach for a different setting where each machine has its own regression vector $\boldsymbol{\theta}^{(m)}$ was proposed by Sundman et al. (2014). In their setting, the support sets of the $M$ vectors $\boldsymbol{\theta}^{(m)}$ are assumed to be similar, and the $M$ machines are connected in a general topology without a fusion center.

# 4 THEORETICAL RESULTS

Despite their simplicity, to the best of our knowledge, distributed-`OMP` schemes lack rigorous mathematical support and only limited theoretical results have been derived for them. Wimalajeewa and Varshney (2014) proved a support recovery guarantee for `DC-OMP`, but only in a restricted noise-less compressed-sensing setting, where the entries of the design matrices are all random and i.i.d. across machines. In contrast, in this section we derive a support recovery guarantee for `DJ-OMP`, under a more general setting, where the design matrices are deterministic and potentially structured, and the responses $y$ are noisy. Specifically, we prove in Theorem 4.1 that if the SNR is high enough (the non-zero entries of $\boldsymbol{\theta}$ are sufficiently large in absolute value), then with high probability `DJ-OMP` recovers the support set $\mathcal{S}$. Remarkably, the SNR required by our theorem is well below that required for individual machines to succeed. Its proof appears in Appendix A.

Towards formally stating our result, we first review known recovery guarantees for `OMP` on a single machine, and mathematically define the SNR in our prob-

lem.

**Distributed Coherence Condition.** The coherence of a matrix $\mathbf{A}$ with columns $\mathbf{a}_j$ is defined as

$$\mu(\mathbf{A}) = \max_{i \neq j} \frac{|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2}. \tag{3}$$

A matrix $\mathbf{A}$ satisfies the *Mutual Incoherence Property* (MIP) with respect to a sparsity level $K$ if

$$\mu(\mathbf{A}) < \frac{1}{2K-1}. \tag{4}$$

A fundamental result by Tropp (2004) is that in an ideal noise-less setting ($\sigma = 0$), the MIP condition (4) is sufficient for exact support recovery by `OMP`.

In our distributed setting, each machine $m$ has its own design matrix $\mathbf{X}^{(m)}$ with coherence $\mu^{(m)} = \mu(\mathbf{X}^{(m)})$. We denote their maximal coherence by

$$\mu_{\max} = \mu_{\max}(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) = \max_{m \in [M]} \mu^{(m)}. \tag{5}$$

We say that a set of matrices $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$ satisfies the max-MIP condition w.r.t. a sparsity level $K$ if

$$\mu_{\max} < \frac{1}{2K-1}. \tag{6}$$

Eq. (6) implies that all machines satisfy the MIP condition (4). Hence, in a noise-less setting, `OMP` at each machine will correctly recover the support of $\boldsymbol{\theta}$.

**Remark 4.1.** *Note that $\mu_{\max}$ depends on all $M$ design matrices at the $M$ machines. In general, if they are random then $\mu_{\max}$ will also be random, and will increase with $M$. However, the max-MIP condition* (6) *is not necessarily very restrictive. For example, the coherence of a matrix with random i.i.d. Gaussian entries is tightly concentrated around its mean. In this case, assuming max-MIP* (6) *instead of MIP* (4) *on a single machine is not significantly limiting.*

The coherence plays a key role for `OMP` recovery also in the presence of noise, as we discuss next.

**SNR Regime.** We formally define the SNR in our distributed setting. We then focus on an interesting regime, in which the SNR is sufficiently high for `OMP` to recover the support of $\boldsymbol{\theta}$ in a centralized setting, where the center has access to all the samples from all machines, and yet too low for `OMP` at a single machine to individually recover it. For a $K$-sparse vector $\boldsymbol{\theta} \in \mathbb{R}^d$, a matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with coherence $\mu$ whose columns have unit norm, and a noise level $\sigma$, define

$$\theta_{\text{crit}}(\mu, d, K, \sigma) = \frac{\sigma\sqrt{2\log d}}{1-(2K-1)\mu}. \tag{7}$$

Notice that $\theta_{\text{crit}}(\mu, d, K, \sigma)$ is well defined under the MIP condition (4).

As in previous works, to derive exact support recovery guarantees, we consider vectors $\boldsymbol{\theta}$ whose non-zero entries have magnitude lower bounded by $\theta_{\min}$, namely $\min_{k \in \mathcal{S}} |\theta_k| \geq \theta_{\min}$. For a matrix $\mathbf{A}$ with unit-norm columns, define the SNR as $r = \left(\frac{\theta_{\min}}{\theta_{\text{crit}}(\mu,d,K,\sigma)}\right)^2$. Near the value $r = 1$, `OMP` (at a single machine) exhibits a phase transition from failure to success of support recovery. If the SNR is slightly higher, i.e., $r > \left(1 + \sqrt{\frac{\log K}{\log d}}\right)^2$, then with high probability `OMP` exactly recovers the support $\mathcal{S}$ (Ben-Haim et al., 2010). In contrast, if the SNR is slightly lower, i.e., $r < \left(1 - \sqrt{\frac{\log K}{\log d}} - \mu\right)^2$, then there are matrices $\mathbf{A} \in \mathbb{R}^{n \times d}$ with coherence $\mu$ and $K$-sparse vectors $\boldsymbol{\theta} \in \mathbb{R}^d$ for which given $\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \sigma\boldsymbol{\xi}$, `OMP` fails with high probability to recover the support of $\boldsymbol{\theta}$. In addition, this occurs empirically for several common families of matrices $\mathbf{A}$ and vectors $\boldsymbol{\theta}$ (Amiraz et al., 2021).

In our distributed setting the matrices $\mathbf{X}^{(m)}$ are assumed to be deterministic and do not necessarily have unit-norm columns. However, (2) is equivalent to

$$\mathbf{y}^{(m)} = \tilde{\mathbf{X}}^{(m)} \tilde{\boldsymbol{\theta}}^{(m)} + \sigma\boldsymbol{\xi}^{(m)}, \tag{8}$$

where each column $\tilde{\mathbf{x}}_j^{(m)}$ of the matrix $\tilde{\mathbf{X}}^{(m)}$ is scaled to have unit norm, i.e., $\tilde{\mathbf{x}}_j^{(m)} = \mathbf{x}_j^{(m)} / \|\mathbf{x}_j^{(m)}\|$, and accordingly $\tilde{\theta}_j^{(m)} = \|\mathbf{x}_j^{(m)}\| \theta_j$. Clearly, the support of each $\tilde{\boldsymbol{\theta}}^{(m)}$ is identical to that of $\boldsymbol{\theta}$. We assume that for a suitable $\tilde{\theta}_{\min}$, the vector $\boldsymbol{\theta}$ satisfies that

$$\min_m \|\mathbf{x}_k^{(m)}\| \, |\theta_k| \geq \tilde{\theta}_{\min}, \quad \forall k \in \mathcal{S}. \tag{9}$$

Given the above discussion, in our distributed setting we define the SNR parameter $r$ as follows,

$$r = \left(\frac{\tilde{\theta}_{\min}}{\theta_{\text{crit}}(\mu_{\max}, d, K, \sigma)}\right)^2. \tag{10}$$

If $r > 1$ then $\tilde{\theta}_{\min} > \theta_{\text{crit}}(\mu^{(m)}, d, K, \sigma)$ at every machine $m \in [M]$, and hence in any single machine `OMP` would recover the support of $\boldsymbol{\theta}$ with high probability.

Next, consider a centralized setting where all $N = Mn$ samples are available to the fusion center. This setting corresponds to a response vector $\mathbf{y} \in \mathbb{R}^N$ and measurement matrix $\mathbf{X} \in \mathbb{R}^{N \times d}$ formed by stacking the vectors $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}$ and the rows of $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}$, respectively. In analogy to (10), to guarantee support recovery in this case, a sufficient condition is that the centralized SNR $r^{(c)} = \left(\frac{\tilde{\theta}_{\min}^{(c)}}{\theta_{\text{crit}}(\mu(\mathbf{X}),d,K,\sigma)}\right)^2 > 1$.

Here $\tilde{\theta}^{(c)}_{\min}$ is a value such that for all support indices $k \in \mathcal{S}$, $|\theta_k| \geq \tilde{\theta}^{(c)}_{\min}/\|\mathbf{X}_k\|$, where $\mathbf{X}_k$ is the $k$-th column of $\mathbf{X}$. Since $\|\mathbf{X}_k\| \geq \sqrt{M}\min_m \left\|\mathbf{x}_k^{(m)}\right\|$, then in a centralized setting OMP is guaranteed to succeed when $\sqrt{M}\tilde{\theta}_{\min} > \theta_{\text{crit}}(\mu(\mathbf{X}), d, K, \sigma)$. Given the definition (7) for $\theta_{\text{crit}}$, an SNR regime that is interesting to study in the distributed setting is

$$\frac{1}{M}\left(\frac{1-(2K-1)\mu_{\max}}{1-(2K-1)\mu(\mathbf{X})}\right)^2 < r < 1. \tag{11}$$

In this range, the SNR is sufficiently high for recovery in the centralized setting, but too low to guarantee recovery at individual machines. As we show next, for a subrange of the SNR values in Eq. (11), the DJ-OMP scheme can still achieve exact support recovery.

### 4.1 Support Recovery Guarantee

We present three assumptions for our recovery guarantee to hold. As OMP is based on dot products between the residual and normalized columns of the design matrices, we first introduce the following quantity that bounds how large these can be,

$$\delta = \delta\left(K, \mu_{\max}\right) = \tfrac{(K-1)\mu_{\max}^2}{1-(K-2)\mu_{\max}}. \tag{12}$$

As we show in Section A.5, under the max-MIP condition (6), $\delta \leq \mu_{\max}$. Our first assumption is that the number of machines is sufficiently large, with the dependence on $K$ encoded in the quantity $\delta$.

**Assumption 4.1.** *$M \geq M_c\left(d, K, \mu_{\max}, r\right)$, where*

$$M_c\left(d, K, \mu_{\max}, r\right) = K\left\lceil \frac{16\log d}{\Phi^c\left(\frac{(1-\sqrt{r})\sqrt{2\log d}}{\sqrt{1-\delta}(1-\mu_{\max})}\right)}\right\rceil. \tag{13}$$

In our analysis, we assume that $d \gg 1$ and that $\mu_{\max}$ is small. This implies that also $\delta$ is small and hence

$$M_c\left(d, K, \mu_{\max}, r\right) \approx K d^{\left(\frac{1-\sqrt{r}}{\sqrt{1-\delta}(1-\mu_{\max})}\right)^2}, \tag{14}$$

which follows from the approximation $\Phi^c(t) \approx e^{-t^2/2}$ and omitting $O(\log d)$ factors. Thus, larger SNR values (though still smaller than one), require fewer machines to guarantee support recovery.

To guarantee support recovery by DJ-OMP, we also need to upper bound the probability that a non-support index is sent to the fusion center. As described in the appendix, for this we use a recent result on the left tail of the maximum of correlated Gaussian random variables (Lopes and Yao, 2022). The SNR that guarantees recovery thus depends on a parameter $\epsilon = \epsilon(K, \mu_{\max})$, with smaller values of $\epsilon$ leading to a lower SNR. However, for our proof to work, $\epsilon$ cannot be arbitrarily small, and we set it as follows.

**Assumption 4.2.** *The scalar $\epsilon = \epsilon(K, \mu_{\max})$ satisfies*

$$\frac{\sqrt{\mu_{\max}+\delta}}{1+\sqrt{\mu_{\max}+\delta}} < \epsilon < 1. \tag{15}$$

Importantly, for $\mu_{\max}$ small, $\epsilon$ can be chosen to be as small as $O(\sqrt{\mu_{\max}})$. As detailed in the theorem below, this allows recovery at low SNRs.

Finally, we define a few quantities that characterize the lower bound we impose on the SNR $r$. Let

$$Q_0\left(d, K\right) = \tfrac{\log\left(88\sqrt{2}K\right)}{\log d}, \tag{16}$$

and define $Q_1\left(d, K, \mu_{\max}, \epsilon\right)$ and $Q_2\left(d, K, \mu_{\max}\right)$ by

$$Q_1 = \tfrac{1-(1-\mu_{\max})\sqrt{1-\delta}\left((1-\epsilon)\sqrt{1-\mu_{\max}}-\sqrt{Q_0}\right)}{1-2\mu_{\max}K\sqrt{1-\delta}\frac{1-\mu_{\max}}{1-(2K-1)\mu_{\max}}}, \tag{17}$$

$$Q_2 = \tfrac{\sqrt{2+2(\mu_{\max}+\delta)}\left(1+\sqrt{1-\delta}(1-\mu_{\max})\sqrt{Q_0}\right)}{\sqrt{1-\delta}(1-\mu_{\max})+\sqrt{2+2(\mu_{\max}+\delta)}}. \tag{18}$$

**Assumption 4.3** (SNR Condition)**.** *The SNR $r$ is lower bounded as follows*

$$\sqrt{r} \geq \begin{cases} Q_2 & (4K-1)\mu_{\max} - 2K\mu_{\max}^2 \geq 1 \\ \min(Q_1, Q_2) & \text{otherwise} \end{cases} \tag{19}$$

We can now state our support recovery guarantee. The following theorem shows that under the above assumptions, the DJ-OMP algorithm, which requires lightweight communication and computation, recovers the support of $\boldsymbol{\theta}$, with high probability.

**Theorem 4.1.** *Under the max-MIP condition* (6) *and Assumptions 4.1-4.3, for sufficiently large $d = d(\epsilon)$, with probability at least $1 - \frac{2}{d}\left(2^K - 1\right)$, DJ-OMP with $K$ rounds recovers the support of the $K$-sparse vector $\boldsymbol{\theta}$.*

Let us analyze the implications of the theorem when $K \ll d$ and $\mu_{\max}, \epsilon, \delta \ll 1$. In this case $Q_1 \approx \epsilon$ and $Q_2 \approx \frac{\sqrt{2}}{1+\sqrt{2}}$. Hence, Assumption 4.3 is approximately $r > (\min(Q_1, Q_2))^2 \approx \epsilon^2$ or $r \gtrsim \mu_{\max}$. Thus, there is a range of relatively low SNR values for which with a sufficiently large number of machines, DJ-OMP is guaranteed to recover the support, even though individual machines fail to do so.

**Remark 4.2.** *Several works considered distributed settings where each machine has a different vector $\boldsymbol{\theta}^{(m)}$, but they all share the same support $\mathcal{S}$ (Duarte et al., 2005; Ling and Tian, 2011; Ling et al., 2012; Wimalajeewa and Varshney, 2014; Li et al., 2015). Theorem 4.1 also holds in such cases, under the following condition on the vectors $\boldsymbol{\theta}^{(m)}$, instead of* (9)*,*

$$\min_{m\in[M]} \left\|\mathbf{x}_k^{(m)}\right\| \left|\theta_k^{(m)}\right| \geq \tilde{\theta}_{\min} \quad \forall k \in \mathcal{S}.$$

**Remark 4.3.** *Our approach can be extended to handle the case where the sparsity level $K$ is unknown. In this case, we may set a stopping criterion whereby the fusion center stops the communication rounds with the $M$ machines and returns its current support estimate if the number of votes for the most-voted index falls below a predefined threshold. Corollary B.1 shows that for a compressed sensing setting where each matrix entry is i.i.d. Bernoulli, the success probability is almost the same as in Theorem 4.1. The corollary and corresponding simulation results can be found in Appendix B.*

**Remark 4.4.** *The success probability in Theorem 4.1 is influenced by the inter-round dependency. It can be improved by variants of our basic scheme. For instance, allocating half of the machines to the first $K/2$ rounds and the rest to the remaining rounds boosts the success probability to $1 - 2^{K/2+1}/d$. Maximizing this approach by using fresh $M/K$ machines at each round increases the probability to $1 - 2K/d$. However, this requires a higher SNR to offset the reduced number of machines in each round. We believe that the success probability in Theorem 4.1 for the basic scheme may be improved to $1 - poly(K)/d$, but this remains an open question for future research.*

We now compare Theorem 4.1 to related works. Amiraz et al. (2022) studied distributed sparse mean estimation, which is a special case of distributed sparse linear regression where the design matrices are orthogonal. They designed low-communication distributed schemes that provably recover the support for a wide range of SNR values. However, their proofs rely on the design matrices being orthogonal, and do not generalize to incoherent matrices. Their schemes are single-round, essentially using the orthogonality to recover all $K$ support indices in parallel, in contrast to our `DJ-OMP` scheme which has $K$ iterations, and requires a careful analysis of error propagation. As mentioned above, Wimalajeewa and Varshney (2014) considered a compressed-sensing setting with incoherent random matrices whose entries are drawn i.i.d. from the same distribution, and with no noise ($\sigma = 0$). In both of these papers, a key property that greatly simplifies the analysis is that at all machines the probability for selecting a non-support index is the same for all $k \notin \mathcal{S}$. Our theorem shows that even without this symmetry between the non-support indices, distributed-`OMP` algorithms can achieve exact support recovery.

## 5 SIMULATION RESULTS

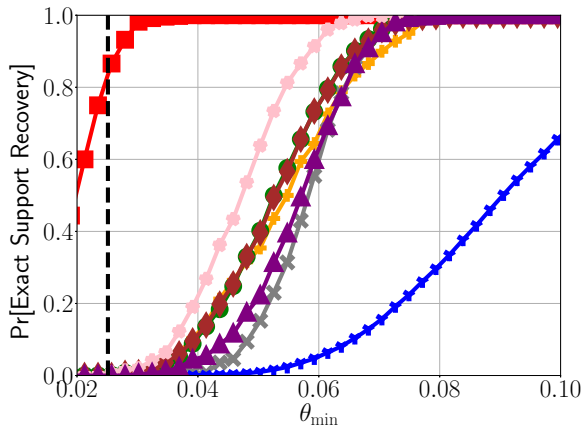
We compare experimentally the following algorithms, which have different computation and communication costs (see Table 1): (i) `Deb-Lasso` where each machine computes a debiased-Lasso estimate of $\boldsymbol{\theta} \in \mathbb{R}^d$ and sends it to the center. The center averages these $M$ vectors and returns its top $K$ indices (Lee et al., 2017; Battey et al., 2018); (ii) `Deb-Lasso-K`, a variant of Barghi et al. (2021), where each machine sends the top $K$ indices of its debiased-Lasso estimate; (iii) `SIS-SCAD-K`, a distributed `SIS` scheme, where each machine performs variable screening followed by `SCAD` (Fan and Lv, 2008). It sends its resulting support set to the center, which selects the top $K$ indices by majority voting; (iv) `SIS-OMP-K`, another distributed `SIS` scheme where each machine estimates its support set using `OMP` on the remaining features; (v) `D-OMP` with $L = K$; (vi) `D-OMP` with $L = 2K$; and (vii) `DJ-OMP`. To illustrate the ability of `DJ-OMP` to recover the support when individual machines fail, for reference we also ran `OMP` on a single machine, ignoring the data in all other $M - 1$ machines. Note that while `OMP`-based schemes are essentially parameter free (beyond the sparsity $K$), in the debiased-Lasso schemes all machines need to know the noise level $\sigma$.

Table 1: Communication and Computation Costs

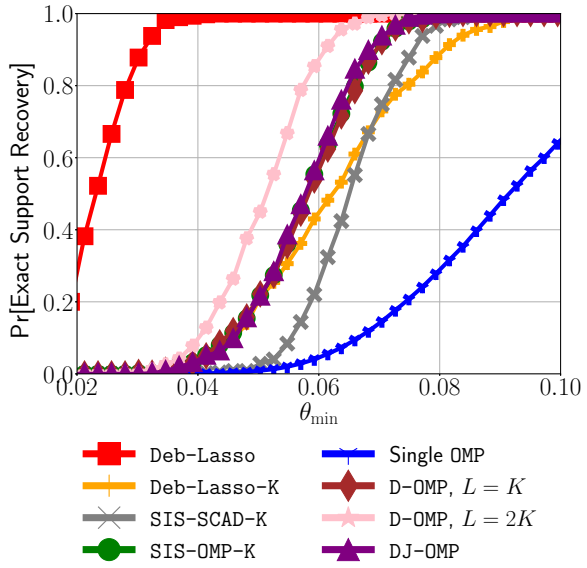
| **Algorithm** | **Communication cost** | **Computational cost,** $K \ll d^{1/3}$ |
|---|---|---|
| Single `OMP` | $\tilde{O}(K)$ | $O(ndK)$ |
| `Deb-Lasso` | $\tilde{O}(d)$[1] | solving $d+1$ Lasso optimization problems |
| `Deb-Lasso-K` | $\tilde{O}(K)$ | |
| `SIS-SCAD-K` | SNR dependent | $O(nd)$ |
| `SIS-OMP-K` | $\tilde{O}(K)$ | |
| `D-OMP`, $L=K$ | $\tilde{O}(K)$ | $O(ndK)$ |
| `DJ-OMP` | | |

We now describe the simulation setup. Each matrix $X^{(m)}$ is generated as follows. Each row is drawn independently from $\mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma$ is a Toeplitz matrix with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \alpha^{|i-j|}$ for $i \neq j$ for some $\alpha \in [0, 1)$. In all settings, we generate $M = 20$ such matrices, each containing $n = 2000$ samples. The noise level is $\sigma = 1$, and the vector $\boldsymbol{\theta}$ has a sparsity $K = 5$, with $\boldsymbol{\theta} = \theta_{\min} \cdot [1, -1.5, 2, -2.5, 3, 0, \ldots, 0]^\top$. The tuning parameter in the debiased-Lasso methods, which scales the $\ell_1$ term of each of the $d+1$ Lasso objectives, is set to $\lambda = 2\sigma\sqrt{\frac{\log d}{n}}$. We consider two settings both of dimension $d = 10000$. In Setting (a), $\alpha = 0$, i.e., all matrix entries are i.i.d. $\mathcal{N}(0, 1)$. In Setting (b), $\alpha = 0.1$, so the columns of $X^{(m)}$ are weakly correlated.

[1]For `Deb-Lasso`, each machine sends the vector $\hat{\theta}^{(m)}$ itself, so the $\tilde{O}(\cdot)$ notation hides the number of bits used for each quantized value.




(a) IID Design Matrices ($\alpha = 0$)




(b) Correlated Design Matrices ($\alpha = 0.1$)

Figure 1: Support Recovery as a Function of $\theta_{\min}$.

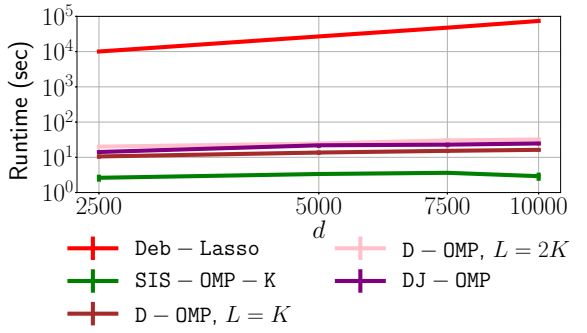



Figure 2: Runtime as a Function of $d$.

Further implementation details appear in Appendix D. Code that reproduces the results presented in this paper is publicly available on GitHub.[2]

Figure 1 illustrates the empirical success probability of the various algorithms as a function of $\theta_{\min}$ in the two settings outlined above. Formally, for an algorithm $A$,

$$p^A_{\text{success}}(\theta_{\min}) = \frac{1}{J}\sum_{j=1}^{J} \mathbb{1}\left\{S_j^A(\theta_{\min}) = \mathcal{S}\right\},$$

where $S_j^A(\theta_{\min})$ is the support set computed by algorithm $A$, for noise realization $j$ and lower bound $\theta_{\min}$ on the non-zero coefficients of $\boldsymbol{\theta}$, and $J$ is the total number of noise realizations, set to $J = 500$. The dashed vertical line in panel (a) is the lower bound $\theta_{\text{crit}}(\mu(\mathbf{X}), d, K, \sigma)$ of Eq. (7), above which in a centralized setting, `OMP` is guaranteed to recover the support with high probability. In panel (b), the MIP condition does not hold and the dashed line is not shown. Nonetheless, distributed schemes still succeed in this case.

Figure 1 reveals several phenomena. First, as anticipated, the performance of distributed-`OMP` algorithms is inferior to `Deb-Lasso`, which incurs much higher computational and communication costs. Second, in accordance with Theorem 4.1, distributed-`OMP` algorithms succeed at low SNR values, where `OMP` on a single machine fails with high probability. Third, `DJ-OMP`'s performance is comparable to `D-OMP` with $L = K$. For scenarios requiring one-shot communication, `D-OMP` with more steps, $L = 2K$ in this example, exceeds `DJ-OMP`'s performance, while incurring twice the communication cost, which is still much lower than $d$ if $K \ll d$. In Setting (a) where all entries of the matrices $\mathbf{X}^{(m)}$ are i.i.d. Gaussian, the performance of distributed-`OMP` algorithms is on par with the computationally demanding `Deb-Lasso-K`. Notably, in Setting (b) where the matrices $\mathbf{X}^{(m)}$ have correlated columns, distributed-`OMP` methods surpass `Deb-Lasso-K`. In the context of variable screening methods, for a wide range of SNR values, a single machine often misses the full support set during the screening step. Yet, incorporating voting schemes enables distributed support recovery. Similar to `Deb-Lasso-K`, `SIS-SCAD-K` matches the performance of distributed-`OMP` algorithms in Setting (a) but lags behind them in Setting (b). In all the studied settings, `SIS-OMP-K` performs similarly to `D-OMP` with $L = K$.

Figure 2 shows the runtime and error bars of several schemes, all implemented in Python, as a function of $d$ on a logarithmic scale. In this simulation, $\alpha = 0$

[2] https://github.com/ChenAttias/Distributed-OMP

and $\theta_{\min} = 0.1$ and we averaged over $J = 20$ realizations. The runtime of `Deb-Lasso-K` is similar to that of `Deb-Lasso`, and thus not shown. As seen in the figure, distributed-`OMP` methods are more than three orders of magnitude faster than `Deb-Lasso`. `SIS-OMP-K` achieves an additional improvement in runtime compared to distributed-`OMP` methods. A theoretical study of `SIS-OMP-K` is an interesting topic for future research.

Finally, in Appendix C we show empirically that the number of machines to recover the support scales as $M \approx d^{\beta}$ for some $\beta < 1$, in accordance with (14).

# 6 DISCUSSION

In distributed sparse linear regression, a fundamental theoretical aspect is determining SNR-dependent lower bounds on the communication required for exact support recovery. To the best of our knowledge, there are no such established lower bounds. This necessitates a nuanced exploration of communication requirements for exact support recovery under different SNRs. When the SNR is sufficiently high so that an individual machine can recover the support of $\boldsymbol{\theta}$, for example by `OMP`, the fusion center may recover the support $\mathcal{S}$ by contacting only one machine, incurring an incoming communication of only $O(K \log d)$ bits. Note that even in a noise-less setting, for the fusion center to recover the support, $K$ indices must be sent to the center, so $K \log d$ bits is a lower bound on the total required communication. On the other hand, when the SNR is low, distributed `Deb-Lasso` succeeds to recover the support of $\boldsymbol{\theta}$ but incurs a communication cost of $\tilde{O}(d)$ bits per machine, which might be prohibitive in high-dimensional settings.

We conjecture that at low-SNR values, no distributed algorithm can achieve exact support recovery with communication per machine $O(K \log d)$ bits. We note that for closely related problems, achieving the centralized minimax $\ell_2$ risk or the centralized prediction error is possible at low SNRs but requires a communication cost of $\Omega(d)$ bits (Shamir, 2014; Steinhardt and Duchi, 2015; Acharya et al., 2019; Barnes et al., 2020). Our work shows that for a range of SNR values between these two extremes, distributed-`OMP` algorithms do recover the support of $\boldsymbol{\theta}$ with communication per machine $O(K \log d)$. An interesting open question is to determine the optimal rate at which the required communication decreases as a function of the SNR by any distributed algorithm that achieves exact support recovery. Another interesting direction for future research is to characterize the tradeoff between communication costs and computational resources.

**Acknowledgements**

This research is supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center, and by a research grant from the Estate of Harry Schutzman. We thank the anonymous reviewers for their insightful comments and Rodney Fonseca for sharing with us his implementation of the `Deb-Lasso` and `Deb-Lasso-K` algorithms.

**References**

Jayadev Acharya, Chris De Sa, Dylan Foster, and Karthik Sridharan. Distributed learning with sublinear communication. In *International Conference on Machine Learning*, pages 40–50, 2019.

Tal Amir, Ronen Basri, and Boaz Nadler. The trimmed Lasso: Sparse recovery guarantees and practical optimization by the generalized soft-min penalty. *SIAM Journal on Mathematics of Data Science*, 3(3):900–929, 2021.

Chen Amiraz, Robert Krauthgamer, and Boaz Nadler. Tight recovery guarantees for orthogonal matching pursuit under Gaussian noise. *Information and Inference: A Journal of the IMA*, 10(2):573–595, 2021.

Chen Amiraz, Robert Krauthgamer, and Boaz Nadler. Distributed sparse normal means estimation with sublinear communication. *Information and Inference: A Journal of the IMA*, 11(3):1109–1142, 2022.

Hanie Barghi, Amir Najafi, and Seyed Abolfazl Motahari. Distributed sparse feature selection in communication-restricted networks. *arXiv preprint arXiv:2111.02802*, 2021.

Leighton Pate Barnes, Yanjun Han, and Ayfer Ozgur. Lower bounds for learning distributions under communication constraints via fisher information. *Journal of Machine Learning Research*, 21(236):1–30, 2020.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

Zvika Ben-Haim, Yonina C Eldar, and Michael Elad. Coherence-based performance guarantees for estimating a sparse vector under random noise. *IEEE Transactions on Signal Processing*, 58(10):5030–5043, 2010.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

Zygmunt Wilhelm Birnbaum. An inequality for Mill's ratio. *The Annals of Mathematical Statistics*, 13(2): 245–246, 1942.

Thomas Blumensath and Mike E Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5):629–654, 2008.

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

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Sheng Chen, Stephen A Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23 (4):493–507, 1952.

Symeon Chouvardas, Gerasimos Mileounis, Nicholas Kalouptsidis, and Sergios Theodoridis. Greedy sparsity-promoting algorithms for distributed learning. *IEEE Transactions on Signal Processing*, 63(6): 1419–1432, 2015.

Wei Dai and Olgica Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

Marco F Duarte, Shriram Sarvotham, Dror Baron, Michael B Wakin, and Richard G Baraniuk. Distributed compressed sensing of jointly sparse signals. In *Asilomar Conference on Signals, Systems and Computers, 2005.*, pages 1537–1541. IEEE, 2005.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911, 2008.

Alyson K Fletcher, Sundeep Rangan, and Vivek K Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, 2009.

Rodney Fonseca and Boaz Nadler. Distributed sparse linear regression under communication constraints. *arXiv preprint arXiv:2301.04022*, 2023.

Sophie M Fosson, Javier Matamoros, Carles Antón-Haro, and Enrico Magli. Distributed recovery of jointly sparse signals under communication constraints. *IEEE Transactions on Signal Processing*, 64(13):3470–3482, 2016.

Robert D Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

Puxiao Han, Ruixin Niu, and Yonina C Eldar. Modified distributed iterative hard thresholding. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3766–3770. IEEE, 2015.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or Lasso? analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Saurabh Khanna and Chandra R Murthy. Decentralized joint-sparse signal recovery: A sparse bayesian learning approach. *IEEE Transactions on Signal and Information Processing over Networks*, 3(1):29–45, 2016.

Yûsaku Komatu. Elementary inequalities for Mills' ratio. *Rep. Statist. Appl. Res. Un. Jap. Sci. Engrs*, 4:69–70, 1955.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1): 115–144, 2017.

Gang Li, Thakshila Wimalajeewa, and Pramod K Varshney. Decentralized and collaborative subspace pursuit: A communication-efficient algorithm for joint sparsity pattern recovery with sensor networks. *IEEE Transactions on Signal Processing*, 64(3):556–566, 2015.

Qing Ling and Zhi Tian. Decentralized support detection of multiple measurement vectors with joint sparsity. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2996–2999. IEEE, 2011.

Qing Ling, Zaiwen Wen, and Wotao Yin. Decentralized jointly sparse optimization by reweighted $\ell_q$ minimization. *IEEE Transactions on Signal Processing*, 61(5):1165–1170, 2012.

Zhan Liu, Xiaoluo Zhao, and Yingli Pan. Communication-efficient distributed estimation for high-dimensional large-scale linear regression. *Metrika*, 86(4):455–485, 2023.

Miles E Lopes and Junwen Yao. A sharp lower-tail bound for gaussian maxima with application to bootstrap methods in high dimensions. *Electronic Journal of Statistics*, 16(1):58–83, 2022.

Alireza Makhzani and Shahrokh Valaee. Distributed spectrum sensing in cognitive radios via graphical models. In *5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 376–379. IEEE, 2013.

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

Gonzalo Mateos, Juan Andrés Bazerque, and Georgios B Giannakis. Distributed sparse linear regression. *IEEE Transactions on Signal Processing*, 58 (10):5262–5276, 2010.

Alan Miller. *Subset selection in regression*. CRC Press, 2002.

João FC Mota, João MF Xavier, Pedro MQ Aguiar, and Markus Puschel. Distributed basis pursuit. *IEEE Transactions on Signal Processing*, 60(4): 1942–1956, 2011.

Deanna Needell and Joel A Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

Yagyensh Chandra Pati, Ramin Rezaiifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pages 40–44. IEEE, 1993.

Stacy Patterson, Yonina C Eldar, and Idit Keidar. Distributed compressed sensing for static and time-varying networks. *IEEE Transactions on Signal Processing*, 62(19):4931–4946, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Python Core Team. *Python: A dynamic, open source programming language*. Python Software Foundation, 2019. URL `https://www.python.org/`. Python version 3.8.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

Jonathan D Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

Diego F Saldana and Yang Feng. SIS: An R package for sure independence screening in ultrahigh-dimensional statistical models. *Journal of Statistical Software*, 83(2):1–25, 2018.

Yasmin SarcheshmehPour, Yu Tian, Linli Zhang, and Alexander Jung. Clustered federated learning via generalized total variation minimization. *IEEE Transactions on Signal Processing*, 71:4240–4256, 2023.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20:1–31, 2019.

Ohad Shamir. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Advances in Neural Information Processing Systems*, pages 163–171, 2014.

Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318): 626–633, 1967.

Virginia Smith, Simone Forte, Chenxin Ma, Martin Takáč, Michael I Jordan, and Martin Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18(230):1–49, 2018.

Jacob Steinhardt and John Duchi. Minimax rates for memory-bounded sparse linear regression. In *Conference on Learning Theory*, pages 1564–1587, 2015.

Dennis Sundman, Saikat Chatterjee, and Mikael Skoglund. A greedy pursuit algorithm for distributed compressed sensing. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2729–2732. IEEE, 2012.

Dennis Sundman, Saikat Chatterjee, and Mikael Skoglund. Distributed greedy pursuit algorithms. *Signal Processing*, 105:298–315, 2014.

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

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

Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Jialei Wang, Mladen Kolar, Nathan Srebro, and Tong Zhang. Efficient distributed learning with sparsity. In *International Conference on Machine Learning*, pages 3636–3645, 2017.

Thakshila Wimalajeewa and Pramod K Varshney. Cooperative sparsity pattern recovery in distributed networks via distributed-OMP. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5288–5292. IEEE, 2013.

Thakshila Wimalajeewa and Pramod K Varshney. OMP based joint sparsity pattern recovery under communication constraints. *IEEE Transactions on Signal Processing*, 62(19):5059–5072, 2014.

Thakshila Wimalajeewa and Pramod K Varshney. Application of compressive sensing techniques in distributed sensor networks: A survey. *arXiv preprint arXiv:1709.10401*, 2017.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]. See Section 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]. See Section 3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] Code is available: https://github.com/ChenAttias/Distributed-OMP

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]. See Section 4.
   (b) Complete proofs of all theoretical results. [Yes]. In the supplementary materials.
   (c) Clear explanations of any assumptions. [Yes]. See Section 4.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Code is available: https://github.com/ChenAttias/Distributed-OMP
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]. Error bars are not shown for figure 1 as they are smaller than the symbol size given the large number of simulations we performed.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]. In supplementary materials.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Materials

## A PROOFS

In this section we prove Theorem 4.1. For ease of presentation, in Section A.1 we state and prove Theorem A.1 which addresses the simpler case $K = 1$. The proof of Theorem 4.1 for the general case $K \geq 1$ appears in Section A.2. The proofs of various auxiliary lemmas appear in Sections A.3-A.6.

Towards proving both theorems, we first present a few preliminaries, state useful lemmas and outline the proof.

**Preliminaries.** Recall that `DJ-OMP` is an iterative algorithm, whereby at each round $t$, all $M$ machines call the subroutine `OMP_Step` with the same input set $S_{t-1}$. In principle, except at the first round where $S_0 = \emptyset$, this input set depends on all the data in all $M$ machines. This statistical dependency significantly complicates the analysis. Instead, as discussed below, in our proof we will analyze a single round of `DJ-OMP`, assuming all machines are provided with a *fixed* input set $s$.

Given an input set $s$ to the subroutine `OMP_Step`, each machine $m$ computes a sparse vector supported on $s$, i.e.,

$$\hat{\boldsymbol{\theta}}^{(m)} = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \left\| \mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{z} \right\|_2 \text{ s.t. } \operatorname{supp}(\mathbf{z}) = s. \tag{20}$$

Then, it calculates the corresponding residual vector

$$\mathbf{r}^{(m)} = \mathbf{y}^{(m)} - \mathbf{X}^{(m)}\hat{\boldsymbol{\theta}}^{(m)}. \tag{21}$$

Finally, each machine $m$ sends to the fusion center the index

$$j^{(m)} = \arg\max_{i\in[d]} |\langle \tilde{\mathbf{x}}_i^{(m)}, \mathbf{r}^{(m)} \rangle|, \tag{22}$$

where $\tilde{\mathbf{x}}_i^{(m)} = \frac{\mathbf{x}_i^{(m)}}{\left\|\mathbf{x}_i^{(m)}\right\|}$ is the $i$-th column of $\mathbf{X}^{(m)}$ divided by its norm.

As also described in Algorithm 3, given the messages sent by all $M$ machines, the fusion center computes a vector $\mathbf{v} \in \mathbb{R}^d$, where $\mathbf{v}_j$ counts the number of votes received by index $j$ for all $j \in [d]$. As discussed in the main text, indices in $s$ receive no votes and at each round a new index $j_{\text{center}}$ is chosen by the center,

$$j_{\text{center}} = j_{\text{center}}(s) = \arg\max_{j\in[d]\setminus s} \mathbf{v}_j.$$

Towards proving that with high probability $j_{\text{center}} \in \mathcal{S} \setminus s$, we define an additional quantity $\rho^{(m)} = \rho^{(m)}(s)$ that corresponds to the local SNR at machine $m$ given an input set $s$. Denote

$$\tilde{\theta}_{\max}^{(m)} = \tilde{\theta}_{\max}^{(m)}(s) = \max_{k\in\mathcal{S}\setminus s} \left\{ \left\| \mathbf{x}_k^{(m)} \right\| |\theta_k| \right\}. \tag{23}$$

Similar to the definition of $r$ in Eq. (10), we define

$$\rho^{(m)} = \rho^{(m)}(s) = \left( \frac{\tilde{\theta}_{\max}^{(m)}}{\theta_{\text{crit}}(\mu_{\max}, d, K, \sigma)} \right)^2, \tag{24}$$

where $\theta_{\text{crit}}$ is defined in Eq. (7). Where clear from the context and to simplify notation we will not write the dependence on the input set $s$ explicitly. Note that by its definition, for any input set $s$ that is strictly contained in $\mathcal{S}$, it follows that $\rho^{(m)} \geq r$. As discussed in Section 4, if $\rho^{(m)} > 1$, then with high probability machine $m$ would recover a support index, namely $j^{(m)} \in \mathcal{S} \setminus s$ (Amiraz et al., 2021). Therefore, in what follows, we consider a worst case scenario whereby $\rho^{(m)} \leq 1$ in all machines $m \in [M]$.

**Proof outline and lemmas.** For simplicity we prove the theorem assuming the number of machines is the smallest that still satisfies Assumption 4.1, namely $M = M_c\left(d, K, \mu_{\max}, r\right)$, with $M_c$ defined in Eq. (13). A larger number of machines would only increase the probability of exact support recovery. The main idea of the proof is to show that at each of the $K$ rounds, with high probability the center indeed chooses a support index. Specifically, consider a single round of `DJ-OMP` with a fixed input set $s \subset \mathcal{S}$. Then, for the center to choose an index $j_{\text{center}} \in \mathcal{S} \setminus s$, it suffices that there exists some support index $k \in \mathcal{S} \setminus s$ that received more votes than any non-support index, namely,

$$\mathbf{v}_k > \max_{j \notin \mathcal{S}} \mathbf{v}_j. \tag{25}$$

A sufficient condition for (25) to occur is that for some suitable threshold $t_c = t_c(s) > 0$, both

$$\mathbf{v}_k > t_c, \tag{26}$$

and

$$\max_{j \notin \mathcal{S}} \mathbf{v}_j < t_c. \tag{27}$$

As described below, our chosen threshold $t_c$ depends on the following quantity $F$, which provides a lower bound for the probability that a support index is sent to the center by one of the machines,

$$F\left(d, K, \mu_{\max}, r\right) = \frac{1}{2}\Phi^c\left(\frac{(1-\sqrt{r})\sqrt{2\log d}}{\sqrt{1-\delta}(1-\mu_{\max})}\right). \tag{28}$$

Note that by this definition, Eq. (13) can be rewritten as

$$M_c\left(d, K, \mu_{\max}, r\right) = K\left\lceil \frac{8\log d}{F\left(d, K, \mu_{\max}, r\right)} \right\rceil. \tag{29}$$

We will show that Eqs. (26) and (27) indeed hold with high probability with the following threshold

$$t_c = t_c(s) = \frac{\sum_{m\in[M]} F\left(d, K, \mu_{\max}, \rho^{(m)}(s)\right)}{M \cdot F\left(d, K, \mu_{\max}, r\right)} 4\log d, \tag{30}$$

where $r$, $\rho^{(m)}$ and $F$ are defined in Eqs. (10), (24), and (28) respectively. Note that $\rho^{(1)}, \ldots, \rho^{(M)}$ and $t_c$, which all depend also on the subset $s$, are not assumed to be known to the center and are only used in the proof.

The following Lemma A.1 provides a lower bound for the threshold $t_c$, which will be useful in our proofs. Its proof follows directly from the definition of $F$ in Eq. (28) and appears in Section A.3.

**Lemma A.1.** *Under the max-MIP condition* (6)*, for any fixed $s \subset \mathcal{S}$, the threshold $t_c = t_c(s)$ defined in Eq.* (30) *satisfies*

$$t_c \geq 4\log d. \tag{31}$$

The following Lemma A.2 states that if the expected number of votes for an index $k \in \mathcal{S} \setminus s$ is sufficiently high, then event (26) occurs with high probability. The next Lemma A.3 shows that if the expected number of votes for each non-support index $j \notin \mathcal{S}$ is sufficiently low, then event (27) occurs with high probability. These lemmas follow from Chernoff bounds and are proved in Section A.3 as well.

**Lemma A.2.** *Assume the max-MIP condition* (6) *holds. Fix $s \subset \mathcal{S}$, and let $t_c = t_c(s)$ be given by Eq.* (30)*. If $\mathrm{E}\left[\mathbf{v}_k\right] \geq 2t_c$ for some $k \in \mathcal{S} \setminus s$, then*

$$\Pr\left[\mathbf{v}_k \leq t_c\right] \leq \frac{1}{d}.$$

**Lemma A.3.** *Assume the max-MIP condition* (6) *holds. Fix $s \subset \mathcal{S}$, and let $t_c = t_c(s)$ be given by Eq.* (30)*. If for all non-support indices $j \notin \mathcal{S}$ it holds that $\mathrm{E}\left[\mathbf{v}_j\right] \leq \frac{t_c}{5}$ then*

$$\Pr\left[\max_{j\notin\mathcal{S}} \mathbf{v}_j \geq t_c\right] \leq \frac{1}{d}.$$

It remains to bound $\mathrm{E}\left[\mathbf{v}_j\right]$ from above for $j \in \mathcal{S} \setminus s$ and from below for $j \notin \mathcal{S}$. Towards this goal, denote by $p_j^{(m)}$ the probability that machine $m$ sends index $j$, namely

$$p_j^{(m)} = \Pr\left[j^{(m)} = j\right], \tag{32}$$

where $j^{(m)}$ is defined in Eq. (22).

Since $\mathrm{E}[\mathbf{v}_j] = \sum_m p_j^{(m)}$, it suffices to bound the probability $p_j^{(m)}$. For ease of presentation, we first derive these bounds for the case $K = 1$ in Section A.1, and then extend them to the general case $K \geq 1$ in Section A.2.

### A.1 Support recovery guarantee for sparsity $K = 1$

For completeness, we rewrite Assumptions 4.1-4.3 for this case. Since $K = 1$, by its definition in Eq. (12), $\delta(1, \mu_{\max}) = 0$. Hence, the quantity $F$ simplifies to

$$F(d, 1, \mu_{\max}, r) = \frac{1}{2}\Phi^c\left(\frac{1-\sqrt{r}}{1-\mu_{\max}}\sqrt{2\log d}\right), \tag{33}$$

and the quantity $M_c$ from Eq. (29) reduces to

$$M_c(d, 1, \mu_{\max}, r) = \left\lceil \frac{8\log d}{F(d, 1, \mu_{\max}, r)} \right\rceil. \tag{34}$$

Thus, for $K = 1$, Assumptions 4.1 and 4.2 read as follows:

**Assumption A.1.** $M \geq M_c(d, 1, \mu_{\max}, r)$.

**Assumption A.2.** *The parameter $\epsilon = \epsilon(\mu_{\max})$ satisfies*

$$\frac{\sqrt{\mu_{\max}}}{1+\sqrt{\mu_{\max}}} < \epsilon < 1. \tag{35}$$

The quantity $Q_0$ reduces to

$$Q_0(d, 1) = \frac{\log\left(88\sqrt{2}\right)}{\log d}. \tag{36}$$

In addition, the expressions for $Q_1$ and $Q_2$ simplify to

$$Q_1(d, 1, \mu_{\max}, \epsilon) = \frac{1-(1-\mu_{\max})\left((1-\epsilon)\sqrt{1-\mu_{\max}} - \sqrt{Q_0}\right)}{1-2\mu_{\max}}, \tag{37}$$

$$Q_2(d, 1, \mu_{\max}) = \frac{\sqrt{2+2\mu_{\max}}\left(1+(1-\mu_{\max})\sqrt{Q_0}\right)}{1-\mu_{\max}+\sqrt{2+2\mu_{\max}}}. \tag{38}$$

Finally, for $K = 1$, Assumption 4.3 on the SNR is:

**Assumption A.3** (SNR Condition)**.** *The SNR is sufficiently high,*

$$\sqrt{r} \geq \begin{cases} Q_2 & \mu_{\max} \geq 1/2 \\ \min(Q_1, Q_2) & \textit{otherwise} \end{cases} \tag{39}$$

**Theorem A.1.** *Under Assumptions A.1-A.3 and the max-MIP condition $\mu_{\max} < 1$, for sufficiently large $d = d(\epsilon)$, with probability at least $1 - 2/d$, a single round of* DJ-OMP *recovers the support of a 1-sparse vector $\boldsymbol{\theta}$.*

A few remarks are in place. First, note that when $K = 1$, D-OMP with $L = 1$ reduces to the same algorithm as DJ-OMP, and thus this result holds for this algorithm as well. Second, as mentioned in Section 4, when $\mu_{\max} \ll 1$ condition (39) roughly translates to $r \gtrsim \epsilon^2$, and hence $r \gtrsim \mu_{\max}$. Thus, there is a range of relatively low SNR values for which DJ-OMP succeeds to recover the support, even though the probability of any single machine to do so is very low.

#### A.1.1 Proof of Theorem A.1

When $K = 1$, only a single round is performed with an input set $s = \emptyset$. Thus it trivially holds that $s \subset \mathcal{S}$. In addition, OMP_Step simplifies to the following procedure. At each contacted machine $m$, the residual is simply the response vector, i.e., $\mathbf{r}^{(m)} = \mathbf{y}^{(m)}$. Thus, the index sent by machine $m$ to the fusion center is given by

$$j^{(m)} = \arg\max_{i \in [d]} |\langle \tilde{\mathbf{x}}_i^{(m)}, \mathbf{y}^{(m)} \rangle|. \tag{40}$$

Another simplification in the case $K = 1$ is that the support set contains only one index, which we denote by $k$, i.e., $\mathcal{S} = \{k\}$. To prove Theorem A.1, we derive a lower bound on the probability $p_k^{(m)}$ for the support index $k$ in the following Lemma A.4 and an upper bound on the probability $p_j^{(m)}$ for each non-support index $j \notin \mathcal{S}$ in the following Lemma A.5. Their proofs appear in Section A.4 and are based on a probabilistic analysis of the inner products between the response vector $\mathbf{y}^{(m)}$, which consists of signal and noise, and different columns $\tilde{\mathbf{x}}_i$.

**Lemma A.4.** *Assume that $\|\boldsymbol{\theta}\|_0 = K = 1$ and let $\mathcal{S} = \{k\} = \operatorname{supp}\{\boldsymbol{\theta}\}$. Further assume that the max-MIP condition (6) holds. For sufficiently large $d$, for each machine $m$,*

$$p_k^{(m)} \geq F\left(d, 1, \mu_{\max}, \rho^{(m)}\right), \tag{41}$$

*where $p_k^{(m)}$ and $F$ are defined in Eqs. (32) and (33) respectively.*

**Lemma A.5.** *Assume that $\|\boldsymbol{\theta}\|_0 = K = 1$ and let $\mathcal{S} = \operatorname{supp}\{\boldsymbol{\theta}\}$. Further assume that $\rho^{(m)}$ of Eq. (24) satisfies $\rho^{(m)} \leq 1$ for each machine $m$ and that the max-MIP condition (6) holds. If $\epsilon$ satisfies Assumption A.2, the SNR parameter $r$ satisfies Assumption A.3, and the dimension $d = d(\epsilon)$ is sufficiently large, then for each machine $m$ and each non-support index $j \notin \mathcal{S}$,*

$$p_j^{(m)} \leq \frac{F\left(d, 1, \mu_{\max}, \rho^{(m)}\right)}{11}. \tag{42}$$

We now formally prove Theorem A.1 by combining the above lemmas.

*Proof of Theorem A.1.* For simplicity, we assume that the number of machines is $M = M_c(d, 1, \mu_{\max}, r)$, since a larger number of machines would only increase the probability of successful support recovery. We first analyze the probability that event (26) occurs. By Lemma A.4, for the support index $k \in \mathcal{S}$, its expected number of votes is $\mathrm{E}[\mathbf{v}_k] = \sum_{m\in[M]} p_k^{(m)} \geq \sum_{m\in[M]} F\left(d, 1, \mu_{\max}, \rho^{(m)}\right)$. By the definitions of $t_c$ in Eq. (30) and $M_c$ in Eq. (34),

$$\mathrm{E}[\mathbf{v}_k] \geq \frac{M_c \cdot F(d, 1, \mu_{\max}, r)}{4 \log d} \cdot t_c = \left\lceil \frac{8 \log d}{F(d, 1, \mu_{\max}, r)} \right\rceil \frac{F(d, 1, \mu_{\max}, r)}{4 \log d} \cdot t_c \geq 2t_c.$$

By Lemma A.2, the event (26) occurs with probability at least $1 - 1/d$.

Next, we analyze the probability that event (27) occurs. Fix a non-support index $j \notin \mathcal{S}$. Since $\rho^{(m)} \leq 1$, then by Lemma A.5, its expected number of votes is $\mathrm{E}[\mathbf{v}_j] = \sum_{m\in[M]} p_j^{(m)} \leq \frac{1}{11} \sum_{m\in[M]} F\left(d, 1, \mu_{\max}, \rho^{(m)}\right)$. By the definitions of $t_c$ in Eq. (30) and $M_c$ in Eq. (34),

$$\mathrm{E}[\mathbf{v}_j] \leq \frac{1}{11} \left\lceil \frac{8 \log d}{F(d, 1, \mu_{\max}, r)} \right\rceil \frac{F(d, 1, \mu_{\max}, r)}{4 \log d} t_c < \frac{t_c}{5}.$$

The last inequality is justified as follows. Recall that $\lceil x \rceil \leq x + 1$ for all $x$. Thus,

$$\left\lceil \frac{8 \log d}{F(d, 1, \mu_{\max}, r)} \right\rceil \frac{F(d, 1, \mu_{\max}, r)}{4 \log d} \leq 2 + \frac{F(d, 1, \mu_{\max}, r)}{4 \log d}.$$

By the definition of $F$ in Eq. (33), it follows that $F(d, 1, \mu_{\max}, r) \leq 1$. Hence, when $d \geq 8$, then $\log d > 2$, and the term $\frac{F(d,1,\mu_{\max},r)}{4 \log d} \leq \frac{1}{8}$. Hence, by Lemma A.3, the event (27) occurs with probability at least $1 - 1/d$. A union bound completes the proof. □

### A.2 Proof of Theorem 4.1

We now prove that with high probability, DJ-OMP succeeds to recover the support of $\boldsymbol{\theta}$ with general sparsity level $K$. The proof relies on the following lemma, which bounds the probability that, given a fixed input set $s$, the center chooses an incorrect index at a single round of the algorithm.

**Lemma A.6.** *Let $s \subset [d]$ be a fixed set of indices given as input to a single round of DJ-OMP and denote by $j_{center}(s)$ the index chosen by the center at the end of this round. Under Assumptions 4.1-4.3 and the max-MIP condition (6), for sufficiently large $d = d(\epsilon)$, if $s \subset \mathcal{S}$ then the index $j_{center}(s)$ also belongs to the support set $\mathcal{S}$ with high probability. Specifically,*

$$\Pr[j_{center}(s) \notin \mathcal{S}] \leq 2d^{-1}. \tag{43}$$

First, let us show how Theorem 4.1 follows directly from Lemma A.6.

*Proof of Theorem 4.1.* Recall that DJ-OMP starts with $S_0 = \emptyset$, adds exactly one new index to the estimated support set at each round, and runs for exactly $K$ rounds. We denote by $S_1, S_2, \ldots, S_K$ the index sets found by the center after $t = 1, 2, \ldots, K$ distributed rounds of DJ-OMP, respectively.

Our goal is to upper bound the probability that $S_K$, the output of DJ-OMP after $K$ rounds, is not the true support set $\mathcal{S}$. To this end we decompose this failure probability according to the round at which the failure occurred,

$$\Pr\left[S_K \neq \mathcal{S}\right] = \sum_{t=1}^{K} \sum_{\substack{s_{t-1} \subset \mathcal{S} \\ |s_{t-1}|=t-1}} \Pr\left[j_t\left(s_{t-1}\right) \notin \mathcal{S} \text{ and } S_{t-1} = s_{t-1}\right].$$

Directly analyzing each of the terms above is challenging due to the statistical dependency between the set of indices found so far $S_{t-1}$, and the new index found in the current round. To overcome this, we use the inequality $\Pr[A \cap B] \leq \Pr[A]$, which gives

$$\Pr\left[S_K \neq \mathcal{S}\right] \leq \sum_{t=1}^{K} \sum_{\substack{s_{t-1} \subset \mathcal{S} \\ |s_{t-1}|=t-1}} \Pr\left[j_t(s_{t-1}) \notin \mathcal{S}\right].$$

Since now the set $s_{t-1}$ is fixed, we can bound each term via Lemma A.6. This gives

$$\Pr\left[S_K \neq \mathcal{S}\right] \leq \frac{2}{d} \sum_{t=1}^{K} \binom{K}{t-1} = \frac{2}{d}\left(2^K - 1\right),$$

which completes the proof. □

Next, we prove Lemma A.6. Since $s \subset \mathcal{S}$, we need to bound the probability $p_j^{(m)}$ of Eq. (32) for $j \in \mathcal{S} \setminus s$ and for $j \notin \mathcal{S}$. We shall do so using the following two lemmas. The first one, Lemma A.7, lower bounds a different quantity $q^{(m)}$ defined as the probability that the index sent by machine $m$ belongs to the support $\mathcal{S} \setminus s$,

$$q^{(m)} = q^{(m)}(s) = \Pr\left[j^{(m)} \in \mathcal{S} \setminus s\right]. \tag{44}$$

Lemma A.8 upper bounds $p_j^{(m)}$ for each $j \notin \mathcal{S}$. Their proofs appear in Section A.5.

**Lemma A.7.** *Assume that the max-MIP condition* (6) *holds. For each machine $m$, for sufficiently large $d$,*

$$q^{(m)} \geq F\left(d, K, \mu_{\max}, \rho^{(m)}\right), \tag{45}$$

*where $q^{(m)}$ and $F$ are defined in Eqs.* (44) *and* (28) *respectively.*

**Lemma A.8.** *Assume that $\rho^{(m)}$ of Eq.* (24) *satisfies $\rho^{(m)} \leq 1$ for each machine $m$ and that the max-MIP condition* (6) *holds. If $\epsilon$ satisfies Assumption 4.2, the SNR parameter $r$ satisfies Assumption 4.3, and the dimension $d = d(\epsilon)$ is sufficiently large, then for each machine $m$ and each non-support index $j \notin \mathcal{S}$,*

$$p_j^{(m)} \leq \frac{F\left(d, K, \mu_{\max}, \rho^{(m)}\right)}{11K}, \tag{46}$$

*where $p_j^{(m)}$ and $F$ are defined in Eqs.* (32) *and* (28) *respectively.*

We now formally prove Lemma A.6 by combining the above lemmas.

*Proof of Lemma A.6.* As mentioned above, for simplicity, we prove the lemma assuming that the number of machines is $M = M_c\left(d, K, \mu_{\max}, r\right)$, since a larger number of machines would only increase the probability of exact support recovery. We first analyze the probability that event (26) occurs. Since $s \subset \mathcal{S}$, the set of support indices not yet found is $\mathcal{S} \setminus s$. Let $\mathbf{v}(\mathcal{S} \setminus s) = \sum_{k \in \mathcal{S} \setminus s} \mathbf{v}_k$ be the total number of votes received for all these

support indices combined. By Lemma A.7, the expected number of votes is $\mathrm{E}\left[\mathbf{v}(\mathcal{S}\setminus s)\right] = \sum_{m\in[M]} q^{(m)} \geq \sum_{m\in[M]} F\left(d,K,\mu_{\max},\rho^{(m)}\right)$. By definition of $t_c$ in Eq. (30),

$$\mathrm{E}\left[\mathbf{v}(\mathcal{S}\setminus s)\right] \geq \frac{M_c\left(d,K,\mu_{\max},r\right)F\left(d,K,\mu_{\max},r\right)}{4\log d}\cdot t_c.$$

By definition of $M_c$ in Eq. (29),

$$\mathrm{E}\left[\mathbf{v}(\mathcal{S}\setminus s)\right] \geq K\left\lceil\frac{8\log d}{F\left(d,K,\mu_{\max},r\right)}\right\rceil\frac{F\left(d,K,\mu_{\max},r\right)}{4\log d}\cdot t_c \geq 2Kt_c.$$

By an averaging argument, there exists a support index $k\in\mathcal{S}\setminus s$ for which $\mathrm{E}[\mathbf{v}_k] \geq \frac{1}{|\mathcal{S}\setminus s|}\mathrm{E}\left[\mathbf{v}(\mathcal{S}\setminus s)\right] \geq 2t_c$. Thus, by Lemma A.2, the event (26) occurs with probability at least $1-1/d$.

Similarly to the proof of Theorem A.1, Lemmas A.3 and A.8 imply that the event (27) also occurs with probability at least $1-1/d$. The only change in the proof is that $M_c$ now has a factor of $K$, which cancels with the $1/K$ factor in Lemma A.8. A union bound completes the proof. □

### A.3 Proofs of Lemmas A.1, A.2 and A.3

We first prove Lemma A.1 and then use it to prove Lemmas A.2 and A.3.

*Proof of Lemma A.1.* By its definition in Eq. (28), the function $F$ is monotonic increasing in its fourth argument. Next, by Eq. (23), $\tilde{\theta}_{\min} \leq \tilde{\theta}^{(m)}_{\max}$, and thus $r \leq \rho^{(m)}$ for each $m\in[M]$. Hence,

$$\frac{1}{M}\sum_{m\in[M]}\frac{F(d,K,\mu_{\max},\rho^{(m)})}{F(d,K,\mu_{\max},r)} \geq 1$$

Inserting this inequality into the definition of $t_c$, in Eq. (30) concludes the proof. □

In the proofs below we use the following Chernoff bounds.

**Lemma A.9** (Chernoff (1952))**.** *Suppose $X_1,\ldots,X_d$ are independent Bernoulli random variables and let $X$ denote their sum. Then, for any $\phi\geq 0$,*

$$\Pr\left[X \geq (1+\phi)\,\mathrm{E}\left[X\right]\right] \leq e^{-\frac{\phi^2\mathrm{E}[X]}{2+\phi}}, \tag{47}$$

*and for any $0\leq\phi\leq 1$,*

$$\Pr\left[X \leq (1-\phi)\,\mathrm{E}\left[X\right]\right] \leq e^{-\frac{\phi^2\mathrm{E}[X]}{2}}. \tag{48}$$

Next, we introduce a few notations. Denote the indicator that machine $m$ sends index $k$ by $I_k^{(m)} = \mathbb{1}\left\{j^{(m)}=k\right\}$. The number of votes that $k$ receives is thus $\mathbf{v}_k = \sum_{m\in[M]} I_k^{(m)}$. Further denote $E_k = \mathrm{E}[\mathbf{v}_k]$. Recall that the noises $\left\{\boldsymbol{\xi}^{(m)}\right\}_{m\in[M]}$ are independent. Hence, for a fixed $s$, the indicators $\left\{I_k^{(m)}\right\}_{m\in[M]}$ are independent of each other. We now combine Lemmas A.9 and A.1 to prove Lemmas A.2 and A.3.

*Proof of Lemma A.2.* By the discussion above, we may apply the Chernoff bound (48) to the sum $\mathbf{v}_k$. Using the assumption $E_k \geq 2t_c$ and Lemma A.1, we obtain

$$\Pr\left[\mathbf{v}_k < t_c\right] \leq \Pr\left[\mathbf{v}_k < \tfrac{1}{2}E_k\right] \leq \exp\left(-E_k/8\right) \leq \exp\left(-t_c/4\right) \leq 1/d.$$

□

*Proof of Lemma A.3.* Fix $j\notin\mathcal{S}$ and let $\phi_j = \frac{t_c}{E_j} - 1$. The probability of interest is monotonically increasing in $E_j$. Hence, it suffices to prove the lemma for $E_j = t_c/5$. In this case $\phi_j = 4$, and $\phi_j/(2+\phi_j) = 2/3$. Applying the Chernoff bound (47) to the sum $\mathbf{v}_j$, we obtain

$$\Pr\left[\mathbf{v}_j > t_c\right] = \Pr\left[\mathbf{v}_j > (1+\phi_j)E_j\right] \leq \exp\left(-\frac{\phi_j^2}{2+\phi_j}E_j\right) \leq \exp\left(-\frac{8t_c}{15}\right).$$

By Lemma A.1, the above probability is smaller than $d^{-2}$, and by applying a union bound we conclude that

$$\Pr\left[\max_{j\notin\mathcal{S}} \mathbf{v}_j > t_c\right] \leq (d-K)\Pr\left[\mathbf{v}_j > t_c\right] \leq 1/d.$$

□

### A.4 Proofs of Lemmas A.4 and A.5

We begin with a few definitions and notations. For a set of indices $\mathcal{I}$, let $\mathbf{u}_{|\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ be the restriction of the vector $\mathbf{u}$ to $\mathcal{I}$. Similarly, for a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, let $\mathbf{A}_{|\mathcal{I}} \in \mathbb{R}^{n\times|\mathcal{I}|}$ be the restriction of the matrix $\mathbf{A}$ to the columns indexed by $\mathcal{I}$. Further denote by $\mathbf{A}^\dagger$ the Moore-Penrose pseudo inverse of the matrix $\mathbf{A}$, i.e., $\mathbf{A}^\dagger = \left(\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top$ and notice that $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}$. Lastly, recall that $\tilde{\mathbf{X}}^{(m)} \in \mathbb{R}^{n\times d}$ is the column-normalized matrix in machine $m$ and denote by $\mathbf{P}_{\mathcal{I}}^{(m)} \in \mathbb{R}^{n\times n}$ an orthogonal projection onto the span of $\tilde{\mathbf{X}}_{|\mathcal{I}}^{(m)}$, i.e.,

$$\mathbf{P}_{\mathcal{I}}^{(m)} = \tilde{\mathbf{X}}_{|\mathcal{I}}^{(m)}\left(\tilde{\mathbf{X}}_{|\mathcal{I}}^{(m)}\right)^\dagger. \tag{49}$$

For simplicity of notation, in Sections A.4-A.6 we fix a machine $m$ and thus omit the index $m$ from the proofs.

In our proofs we shall use classical tail bounds for the Gaussian distribution (Lemma A.10), a technical lemma regarding the Gaussian distribution, Lemma A.11, whose proof appears in Section A.6, and Lemma A.12, which bounds the left tail probability of the maximum of correlated Gaussian random variables (Lopes and Yao, 2022).

**Lemma A.10** (Gaussian tail bounds (Gordon, 1941))**.** *For any $t > 0$,*

$$\frac{t}{\sqrt{2\pi}(t^2+1)}e^{-t^2/2} \leq \Phi^c(t) \leq \frac{1}{\sqrt{2\pi}t}e^{-t^2/2}. \tag{50}$$

**Lemma A.11.** *For any $a, b \geq 0$,*

$$\Phi^c(a+b) < \sqrt{2}e^{-b^2/2}\Phi^c(a).$$

**Lemma A.12** ((Lopes and Yao, 2022))**.** *Let $(Z_1,\ldots,Z_d) \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$ where $\mathbf{\Sigma}_{ii} = 1$ for all $i \in [d]$ and $\mathbf{\Sigma}_{ij} \leq \eta < 1$ for some fixed $\eta > 0$ for all $i \neq j \in [d]$. Fix $\zeta \in (0,1)$. There is a constant $C > 0$ depending only on $(\eta,\zeta)$ such that*

$$\Pr\left[\max_{i\in[d]} Z_i < \zeta\sqrt{2(1-\eta)\log d}\right] \leq Cd^{-\frac{(1-\eta)(1-\zeta)^2}{\eta}}(\log d)^{\frac{1-\eta(2-\zeta)-\zeta}{2\eta}}. \tag{51}$$

To put Lemma A.12 in context, recall that the maximum of $d$ independent Gaussians is sharply concentrated at $\sqrt{2\log d}$. In general, for correlated Gaussian random variables, their maximum is lower. However, as the lemma shows, it is unlikely to be much lower than $\sqrt{2(1-\eta)\log d}$, where $\eta$ is an upper bound on the correlation. We use this result with $\eta = \mu_{\max}$ and $\zeta = 1-\epsilon$, where $\epsilon$ satisfies Assumption 4.2, in order to bound the probability that a non-support index is sent to the center and prove Lemma A.5.

Since here we are considering the case $K = 1$, the support of $\boldsymbol{\theta}$ is a single index $\mathcal{S} = \{k\}$. In this case, omitting the index of machine $m$, by Eq. (8) its response vector $\mathbf{y} = \mathbf{y}^{(m)}$ admits the following form

$$\mathbf{y} = \tilde{\theta}_k\tilde{\mathbf{x}}_k + \sigma\boldsymbol{\xi}. \tag{52}$$

Recall that by its definition in Eq. (23), $\tilde{\theta}_{\max} = \|\mathbf{x}_k\|\,|\theta_k| = |\tilde{\theta}_k|$. By Eq. (24) for $\rho$ and Eq. (7) for $\theta_{\text{crit}}$ with $K = 1$,

$$\tilde{\theta}_{\max} = \frac{\sigma\sqrt{2\rho\log d}}{1-\mu_{\max}}. \tag{53}$$

We now prove the lemmas.

*Proof of Lemma A.4.* Recall that $p_k$, defined in Eq. (32), is the probability that the support index $k$ is selected by `OMP_Step`. This occurs if out of all columns of $\tilde{\mathbf{X}}^{(m)}$, the $k$-th column has the highest correlation with the response vector. Hence, to prove the lemma we need to lower bound the probability of the following event,

$$|\langle\tilde{\mathbf{x}}_k,\mathbf{y}\rangle| \geq \max_{i\notin\mathcal{S}}|\langle\tilde{\mathbf{x}}_i,\mathbf{y}\rangle|. \tag{54}$$

where $\mathbf{y}$ is given by (52). To this end, we decompose the noise $\boldsymbol{\xi}$ in Eq. (52) as the sum of two components, the first $\boldsymbol{\xi}_{\|} = \mathbf{P}_k\boldsymbol{\xi} = \langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle \tilde{\mathbf{x}}_k$ is parallel to $\tilde{\mathbf{x}}_k$, namely $\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\|}\rangle = \langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle$, and the second $\boldsymbol{\xi}_{\perp} = \boldsymbol{\xi} - \boldsymbol{\xi}_{\|} = (\mathbf{I} - \mathbf{P}_k)\boldsymbol{\xi}$, is orthogonal to $\tilde{\mathbf{x}}_k$, i.e., $\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\perp}\rangle = 0$.

Next, we use this decomposition to bound the two terms in (54). Combining the expression (52) for $\mathbf{y}$, the decomposition of $\boldsymbol{\xi}$ and the fact that $\tilde{\theta}_{\max} = |\tilde{\theta}_k|$, the LHS of (54) can be bounded by

$$\begin{aligned} |\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle| &\geq \operatorname{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle = \operatorname{sign}\left(\tilde{\theta}_k\right)\left(\tilde{\theta}_k \langle \tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_k\rangle + \sigma\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle\right) \\ &= \tilde{\theta}_{\max} + \sigma \operatorname{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\|}\rangle . \end{aligned} \tag{55}$$

Similarly, the RHS of (54) can be bounded by

$$\begin{aligned} \max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y}\rangle| &= \max_{i\notin\mathcal{S}} \left|\tilde{\theta}_k \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k\rangle + \sigma\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\|} + \boldsymbol{\xi}_{\perp}\rangle\right| \\ &\leq \left(\tilde{\theta}_{\max} + \sigma|\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle|\right)\max_{i\notin\mathcal{S}}\{|\langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k\rangle|\} + \sigma\max_{i\notin\mathcal{S}}|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle| \\ &\leq \left(\tilde{\theta}_{\max} + \sigma\left|\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\|}\rangle\right|\right)\mu_{\max} + \sigma\max_{i\notin\mathcal{S}}|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle| . \end{aligned} \tag{56}$$

where the second step follows from the triangle inequality and the definitions of $\tilde{\theta}_{\max}$ and $\boldsymbol{\xi}_{\|}$, and the last step follows from the definition of $\mu_{\max}$. Combining Eq. (55) with Eq. (56) implies that a sufficient condition for (54) to hold is that

$$\max_{i\notin\mathcal{S}}|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle| \leq \operatorname{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\|}\rangle - \mu_{\max}\left|\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\|}\rangle\right| + \frac{\tilde{\theta}_{\max}}{\sigma}(1-\mu_{\max}) .$$

By Eq. (53), the above event may be written as

$$\max_{i\notin\mathcal{S}}|\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle| \leq \operatorname{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\|}\rangle - \mu_{\max}\left|\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_{\|}\rangle\right| + \sqrt{2\rho\log d}. \tag{57}$$

A key property is that $\boldsymbol{\xi}_{\|}$ and $\boldsymbol{\xi}_{\perp}$ are independent random variables. Hence, the left-hand side and right-hand side in the above inequality, which we denote by $A$ and $B$, respectively, are also independent random variables. Now, for any threshold $T \in \mathbb{R}$, with $A, B$ independent random variables,

$$\Pr[A \leq B] \geq \Pr[A \leq T \cap B \geq T] = \Pr[A \leq T]\cdot\Pr[B \geq T] . \tag{58}$$

Thus,

$$p_k \geq \Pr[A \leq T]\cdot\Pr]B \geq T] \tag{59}$$

and it suffices to lower bound these two probabilities.

In what follows we consider $T = \sqrt{2\log d}$. We begin with bounding $\Pr\left[A \leq \sqrt{2\log d}\right]$. Fix $i \notin \mathcal{S}$ and consider the quantity $\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle$. We may write $\tilde{\mathbf{x}}_i = \mathbf{P}_k\tilde{\mathbf{x}}_i + (\mathbf{I} - \mathbf{P}_k)\tilde{\mathbf{x}}_i$ Since $\boldsymbol{\xi}_{\perp} = (\mathbf{I} - \mathbf{P}_k)\boldsymbol{\xi}$, then $\langle \mathbf{P}_k\tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle = 0$, and $\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle = \langle (\mathbf{I} - \mathbf{P}_k)\tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle$ . Normalizing the inner product by the norm of $(\mathbf{I} - \mathbf{P}_k)\tilde{\mathbf{x}}_i$ yields a standard normal random variable $Z_i = \frac{\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}_{\perp}\rangle}{\|(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i\|_2} \sim \mathcal{N}(0,1)$. By the definition of $\mu_{\max}$,

$$\|(\mathbf{I} - \mathbf{P}_k)\tilde{\mathbf{x}}_i\|^2 = \tilde{\mathbf{x}}_i^T(\mathbf{I} - \mathbf{P}_k)\tilde{\mathbf{x}}_i = 1 - \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k\rangle^2 \geq \gamma_1^2,$$

where $\gamma_1 = \sqrt{1-\mu_{\max}^2}$. Hence,

$$\Pr[A \leq T] \geq \Pr\left[\max_{i\notin\mathcal{S}}|Z_i| \leq \frac{T}{\gamma_1}\right].$$

Since $\{Z_i\}_{i\notin\mathcal{S}}$ are jointly Gaussian, by (Šidák, 1967, Thm. 1), regardless of their covariance structure,

$$\Pr\left[\max_{i\notin\mathcal{S}}|Z_i| \leq \frac{T}{\gamma_1}\right] \geq \prod_{i\notin\mathcal{S}}\Pr\left[|Z_i| \leq \frac{T}{\gamma_1}\right].$$

Applying the Gaussian tail bound (50) with $T = \sqrt{2\log d}$,

$$\Pr\left[|Z_i| \leq \frac{\sqrt{2\log d}}{\gamma_1}\right] \geq 1 - \frac{\gamma_1}{\sqrt{\pi \log d}} d^{-1/\gamma_1^2}.$$

Combining the above three inequalities with Bernoulli's inequality $(1-a)^d \geq 1-da$ which holds for any $a \in [0,1]$, gives

$$\Pr\left[A \leq \sqrt{2\log d}\right] \geq \left(1 - \frac{\gamma_1}{\sqrt{\pi \log d}} d^{-1/\gamma_1^2}\right)^{d-1} \geq 1 - \frac{\gamma_1}{\sqrt{\pi \log d}} d^{1-1/\gamma_1^2} \geq \frac{1}{2}, \tag{60}$$

where the last inequality holds for sufficiently large $d$ and follows from noting that $0 < \gamma_1 \leq 1$.

We now bound $\Pr[B \geq T]$, where $B$ is the RHS of (57). Since $\tilde{\mathbf{x}}_k$ has unit norm, by the definition of $\boldsymbol{\xi}_\parallel$, then $Z = \langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_\parallel \rangle = \langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi} \rangle \sim \mathcal{N}(0,1)$. By the law of total probability,

$$\begin{aligned}
\Pr[B \geq T] &= \Pr\left[\operatorname{sign}\left(\tilde{\theta}_k\right) \langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_\parallel \rangle - \mu_{\max} \left|\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}_\parallel \rangle\right| \geq T - \sqrt{2\rho \log d}\right] \\
&\geq \Pr\left[\gamma_2 |Z| \geq T - \sqrt{2\rho \log d} \mid \operatorname{sign}(Z) = \operatorname{sign}\left(\tilde{\theta}_k\right)\right] \cdot \Pr\left[\operatorname{sign}(Z) = \operatorname{sign}\left(\tilde{\theta}_k\right)\right],
\end{aligned}$$

where $\gamma_2 = 1 - \mu_{\max}$. Since $Z$ is symmetric around zero, $\Pr\left[\operatorname{sign}(Z) = \operatorname{sign}\left(\tilde{\theta}_k\right)\right] = \frac{1}{2}$ and its magnitude is independent on its sign. Thus, for $T = \sqrt{2\log d}$,

$$\Pr\left[B \geq \sqrt{2\log d}\right] \geq \frac{1}{2} \Pr\left[\gamma_2 |Z| \geq \sqrt{2\log d} - \sqrt{2\rho \log d}\right] \geq \Phi^c\left(\frac{1-\sqrt{\rho}}{\gamma_2}\sqrt{2\log d}\right). \tag{61}$$

Inserting (60) and (61) with $\gamma_2 = 1 - \mu_{\max}$ into (59) and recalling the definition of $F$ in (33) completes the proof of Lemma A.4. □

*Proof of Lemma A.5.* Fix a non-support index $j \notin \mathcal{S}$. Recall that $p_j$, defined in Eq. (32), is the probability that index $j$ is selected by `OMP_Step`. This occurs if $j$ has the highest correlation with the response vector, i.e.,

$$p_j = \Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y} \rangle| > \max_{i \neq j} |\langle \tilde{\mathbf{x}}_i, \mathbf{y} \rangle|\right]. \tag{62}$$

In particular, for the $j$-th index to be chosen, the correlation of the $j$-th column with the response vector must exceed both that of the support column $k$, as well as that of any other non-support column $i \notin \{k, j\}$. Indeed, in what follows we separately upper bound

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y} \rangle| > \max_{i \notin \{k,j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y} \rangle|\right] \tag{63}$$

and

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y} \rangle| > |\langle \tilde{\mathbf{x}}_k, \mathbf{y} \rangle|\right], \tag{64}$$

and then use the following inequality to upper bound (62) by their minimum. Specifically, denote $A = |\langle \tilde{\mathbf{x}}_j, \mathbf{y} \rangle|$, $B = \max_{i \notin \{k,j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y} \rangle|$ and $C = |\langle \tilde{\mathbf{x}}_k, \mathbf{y} \rangle|$, then

$$\Pr[A > \max\{B, C\}] \leq \min\{\Pr[A > B], \Pr[A > C]\}. \tag{65}$$

For later use in both bounds, by the triangle inequality, the random variable $A$ can be upper bounded as follows

$$|\langle \tilde{\mathbf{x}}_j, \mathbf{y} \rangle| = \left|\tilde{\theta}_k \langle \tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_k \rangle + \sigma \langle \tilde{\mathbf{x}}_j, \boldsymbol{\xi} \rangle\right| \leq \tilde{\theta}_{\max}\mu_{\max} + \sigma |\langle \tilde{\mathbf{x}}_j, \boldsymbol{\xi} \rangle|. \tag{66}$$

We first bound (63). For each non-support index $i \notin \mathcal{S}$ such that $i \neq j$,

$$|\langle \tilde{\mathbf{x}}_i, \mathbf{y} \rangle| \geq \langle \tilde{\mathbf{x}}_i, \mathbf{y} \rangle = \tilde{\theta}_k \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k \rangle + \sigma \langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi} \rangle \geq -\tilde{\theta}_{\max}\mu_{\max} + \sigma \langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi} \rangle.$$

Combining this with Eq. (66), rearranging terms, and recalling the relation between $\theta_{\max}$ and $\rho$ in (53) yields

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y} \rangle| > \max_{i \notin \{k,j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y} \rangle|\right] \leq \Pr\left[|\langle \tilde{\mathbf{x}}_j, \boldsymbol{\xi} \rangle| + 2\mu_{\max}\frac{\sqrt{2\rho \log d}}{1-\mu_{\max}} > \max_{i \notin \{k,j\}} \langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi} \rangle\right]. \tag{67}$$

Next, we use the following inequality which holds for any pair of random variables $D, E$ and constant $T \in \mathbb{R}$,

$$\Pr[D > E] \leq \Pr[D \geq T] + \Pr[E < T]. \tag{68}$$

Applying this inequality with $T = (1-\epsilon)\sqrt{2(1-\mu_{\max})\log d}$ and $\epsilon \in (0,1)$ as in Eq. (35), we can upper bound (67) by

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \boldsymbol{\xi}\rangle| \geq a\sqrt{2\log d}\right] + \Pr\left[\max_{i\notin\{k,j\}} \langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}\rangle < (1-\epsilon)\sqrt{2(1-\mu_{\max})\log d}\right],$$

where

$$a = (1-\epsilon)\sqrt{1-\mu_{\max}} - \frac{2\mu_{\max}\sqrt{\rho}}{1-\mu_{\max}}.$$

Since $\tilde{\mathbf{x}}_j$ has unit norm, $\langle \tilde{\mathbf{x}}_j, \boldsymbol{\xi}\rangle \sim \mathcal{N}(0,1)$. Hence, the first term is bounded by

$$2\Phi^c\left(a\sqrt{2\log d}\right). \tag{69}$$

We now bound the second term. It involves the maximum of $d-2$ correlated Gaussians, whose covariance matrix $\Sigma$ has $\Sigma_{ii} = 1$ for all $i$, and $\Sigma_{ij} = \mathrm{Cov}\left(\langle \tilde{\mathbf{x}}_i, \boldsymbol{\xi}\rangle, \langle \tilde{\mathbf{x}}_l, \boldsymbol{\xi}\rangle\right) = \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_l\rangle \leq \mu_{\max}$. Hence, we can apply Lemma A.12 with $\eta = \mu_{\max}$ and $\zeta = 1-\epsilon$, which gives the following bound

$$C(d-2)^{-\frac{1-\mu_{\max}}{\mu_{\max}}\epsilon^2}\left(\log(d-2)\right)^{\frac{\epsilon-\mu_{\max}(1+\epsilon)}{2\mu_{\max}}}. \tag{70}$$

We now show that (69) is larger than (70), and thus

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > \max_{i\notin\{k,j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y}\rangle|\right] \leq 4\Phi^c\left(a\sqrt{2\log d}\right). \tag{71}$$

First note that if $\rho$ is sufficiently large such that $a \leq 0$, then (69) is larger than 1, and thus larger than (70). Otherwise, $a > 0$ and using the lower bound for the Gaussian tail of (50), we may lower bound (69) by $d^{-a^2-o(1)}$, where $o(1)$ hides factors that are asymptotically smaller than 1. The term (70) can be upper bounded by $d^{-b^2+o(1)}$, where $b = \sqrt{\frac{1-\mu_{\max}}{\mu_{\max}}}\epsilon$. Next, let us show that for a fixed $\epsilon > 0$, $b-a$ is positive and bounded away from 0. This, in turn, implies that for sufficiently large $d = d(\epsilon)$, (69) is larger than (70). Indeed, under condition (35), $\epsilon = \frac{\sqrt{\mu_{\max}}}{1+\sqrt{\mu_{\max}}} + \epsilon_0$ for some $\epsilon_0 > 0$. Thus, $b - a = \epsilon_0\sqrt{1-\mu_{\max}}\left(1 + \frac{1}{\sqrt{\mu_{\max}}}\right) + \frac{2\mu_{\max}\sqrt{\rho}}{1-\mu_{\max}}$, which is a sum of positive terms and hence bounded away from 0 as desired. Therefore, condition (35) implies that (63) can be bounded by (71).

We now bound (64). For the support index $k$, by (52),

$$|\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle| \geq \mathrm{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle = \mathrm{sign}\left(\tilde{\theta}_k\right)\left(\tilde{\theta}_k\langle \tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_k\rangle + \sigma\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle\right) = \tilde{\theta}_{\max} + \sigma\,\mathrm{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle.$$

Combining this with (66) and plugging $\tilde{\theta}_{\max}$ in Eq. (53), the probability (64) is upper bounded by

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \boldsymbol{\xi}\rangle| - \mathrm{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle > \sqrt{2\rho\log d}\right]. \tag{72}$$

We now upper bound this probability. Let $H = \langle \tilde{\mathbf{x}}_j, \boldsymbol{\xi}\rangle$, $G = \mathrm{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \boldsymbol{\xi}\rangle$ and $c = \sqrt{2\rho\log d}$.

For any pair of random variables $G, H$ and constant $c$,

$$\Pr[|H| - G > c] \leq \Pr[H - G > c] + \Pr[-H - G > c]. \tag{73}$$

By their definition, $H, G$ are jointly Gaussian with mean zero and covariance matrix

$$\begin{pmatrix} \sigma_H^2 & \sigma_{HG} \\ \sigma_{HG} & \sigma_G^2 \end{pmatrix}.$$

Hence, $H-G \sim \mathcal{N}(0, \sigma_H^2+\sigma_G^2-2\sigma_{HG})$ and $-H-G \sim \mathcal{N}(0, \sigma_H^2+\sigma_G^2+2\sigma_{HG})$. Similarly to the above discussion, the diagonal entries $\sigma_H^2 = \sigma_G^2 = 1$ and the off-diagonal entry $\sigma_{HG} = \text{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_j\rangle$. Since $|\sigma_{HG}| \leq \mu_{\max}$, then by (73),

$$\Pr\left[|H| - G > c\right] \leq \Phi^c\left(\frac{c}{\sqrt{2-2\sigma_{HG}}}\right) + \Phi^c\left(\frac{c}{\sqrt{2+2\sigma_{HG}}}\right) \leq 2\Phi^c\left(\frac{c}{\sqrt{2+2\mu_{\max}}}\right).$$

Inserting $c = \sqrt{2\rho\log d}$ yields

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > |\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle|\right] \leq 2\Phi^c\left(\sqrt{\frac{2\rho\log d}{2+2\mu_{\max}}}\right). \tag{74}$$

By Eq. (65), the probability (62) is at most the minimum between (71) and (74). By the monotonicity of the Gaussian CDF, it is upper bounded by

$$4\Phi^c\left(\max\left\{\left((1-\epsilon)\sqrt{1-\mu_{\max}} - \frac{2\mu_{\max}\sqrt{\rho}}{1-\mu_{\max}}\right), \sqrt{\frac{\rho}{2+2\mu_{\max}}}\right\}\sqrt{2\log d}\right). \tag{75}$$

Finally, to prove (42) of the lemma, we note that with $Q_1$ and $Q_2$ defined in Eqs. (37) and (38) respectively, by splitting to cases and applying some algebraic manipulations[3], condition (39) implies that

$$\frac{1-\sqrt{r}}{1-\mu_{\max}} + \sqrt{Q_0} < \max\left\{\left((1-\epsilon)\sqrt{1-\mu_{\max}} - \frac{2\mu_{\max}\sqrt{r}}{1-\mu_{\max}}\right), \sqrt{\frac{r}{2+2\mu_{\max}}}\right\}. \tag{76}$$

The definitions of $r$ and $\rho$ in Eqs. (10) and (24) imply that $\rho \geq r$. Thus, $\rho$ satisfies condition (39) and hence condition (76). The RHS of (76) is the same as the maximum in (75) above. Thus, (75) is upper bounded by

$$4\Phi^c\left(\left(\frac{1-\sqrt{\rho}}{1-\mu_{\max}} + \sqrt{Q_0}\right)\sqrt{2\log d}\right). \tag{77}$$

Since $\rho \leq 1$, we can apply Lemma A.11. Hence, by the definition of $Q_0$ in Eq. (36), and by the definition of $F$ in Eq. (33),

$$\begin{aligned} p_j &\leq 4\Phi^c\left(\left(\frac{1-\sqrt{\rho}}{1-\mu_{\max}} + \sqrt{Q_0}\right)\sqrt{2\log d}\right) \leq 4\sqrt{2}d^{-Q_0}\Phi^c\left(\frac{1-\sqrt{\rho}}{1-\mu_{\max}}\sqrt{2\log d}\right) \\ &= 4\sqrt{2}\frac{1}{88\sqrt{2}}\Phi^c\left(\frac{1-\sqrt{\rho}}{1-\mu_{\max}}\sqrt{2\log d}\right) = \frac{F\left(d,1,\mu_{\max},\rho\right)}{11}, \end{aligned}$$

which completes the proof of Lemma A.5. □

### A.5 Proof of Lemmas A.7 and A.8

We first make a few definitions and present a useful technical lemma. We begin by rewriting the residual $\mathbf{r}^{(m)}$ using the notations introduced in Section A.4. Recall that given an input support set $s$, each machine $m$ estimates its vector $\hat{\boldsymbol{\theta}}^{(m)}$ by solving the least squares problem (20). Thus, $\text{supp}\left(\hat{\boldsymbol{\theta}}^{(m)}\right) = s$ and

$$\hat{\boldsymbol{\theta}}^{(m)}_{|s} = \left(\mathbf{X}^{(m)}_{|s}\right)^{\dagger}\mathbf{y}^{(m)}.$$

[3]First, consider the case $\mu_{\max} \geq 1/2$. By the max-MIP condition (6), $\mu_{\max} < 1$, and hence the term $\frac{1-\mu_{\max}+\sqrt{2+2\mu_{\max}}}{(1-\mu_{\max})\sqrt{2+2\mu_{\max}}}$ is positive, and thus can multiply both sides of the inequality $\sqrt{r} > Q_2$ without altering its direction. Rearranging yields that the LHS of (76) is smaller than $\sqrt{\frac{r}{2+2\mu_{\max}}}$ and thus smaller than the RHS of (76). Now consider the case $\mu_{\max} < 1/2$. By (39), $\sqrt{r} > Q_1$ or $\sqrt{r} > Q_2$. By the same reasoning, the latter implies that the LHS of (76) is smaller than $\sqrt{\frac{r}{2+2\mu_{\max}}}$. Similarly, the term $\frac{1-2\mu_{\max}}{1-\mu_{\max}}$ is positive in this case, and thus multiplying the inequality $\sqrt{r} > Q_1$ by it and rearranging the terms implies that the LHS of (76) is smaller than $(1-\epsilon)\sqrt{1-\mu_{\max}} - \frac{2\mu_{\max}\sqrt{r}}{1-\mu_{\max}}$. Finally, the logical or relation between these conditions implies that the LHS of (76) is smaller than the maximum between the aforementioned terms.

Denote by $\tilde{\boldsymbol{\xi}}^{(m)}$ the projection of the noise $\boldsymbol{\xi}^{(m)}$ to the subspace orthogonal to the span of the columns of $\mathbf{X}^{(m)}_{|s}$, i.e., $\tilde{\boldsymbol{\xi}}^{(m)} = \left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\boldsymbol{\xi}^{(m)}$. Given that $s \subset \mathcal{S}$, the residual $\mathbf{r}^{(m)}$ defined in Eq. (21) can be written in the following form

$$\begin{aligned}
\mathbf{r}^{(m)} &= \mathbf{y}^{(m)} - \mathbf{X}^{(m)}\hat{\boldsymbol{\theta}}^{(m)} = \mathbf{y}^{(m)} - \mathbf{X}^{(m)}_{|s}\hat{\boldsymbol{\theta}}^{(m)}_{|s} = \left(\mathbf{I} - \mathbf{X}^{(m)}_{|s}\left(\mathbf{X}^{(m)}_{|s}\right)^{\dagger}\right)\mathbf{y}^{(m)} \\
&= \left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\mathbf{y}^{(m)} = \left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\left(\tilde{\mathbf{X}}^{(m)}\tilde{\boldsymbol{\theta}}^{(m)} + \sigma\boldsymbol{\xi}^{(m)}\right) \\
&= \left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\sum_{l\in\mathcal{S}\setminus s}\tilde{\theta}^{(m)}_l\tilde{\mathbf{x}}^{(m)}_l + \sigma\tilde{\boldsymbol{\xi}}^{(m)},
\end{aligned} \tag{78}$$

where $\tilde{\mathbf{X}}^{(m)}$ and $\tilde{\boldsymbol{\theta}}^{(m)}$ are the scaled versions of $\mathbf{X}^{(m)}$ and $\boldsymbol{\theta}$, as discussed after Eq. (8), and the last equality follows from the definition of $\mathbf{P}^{(m)}_s$ as a projection operator, so that $\left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\tilde{\mathbf{x}}^{(m)}_k = \mathbf{0}$ for any $k \in s$.

Denote by $K_d$ the size of the detected support set, i.e., $K_d = |s|$, and by $K_u$ the size of the undetected support set, i.e., $K_u = |\mathcal{S} \setminus s|$. Since $s \subset \mathcal{S}$, then $K_d + K_u = K$. Finally, we introduce the following quantity

$$\mu_s = \mu_s\left(K_d, \mu_{\max}\right) = \frac{K_d\mu^2_{\max}}{1 - (K_d - 1)\mu_{\max}}. \tag{79}$$

The following Lemma A.13 bounds the effect of the projection $\mathbf{I} - \mathbf{P}^{(m)}_s$ on the inner products and norms of columns of $\tilde{\mathbf{X}}^{(m)}$. Its proof appear in Appendix A.6.

**Lemma A.13.** *Assume that the max-MIP condition* (6) *holds and that $s \subset \mathcal{S}$. Then, the following inequalities hold for any $0 \le K_d \le K - 1$ and $1 \le K_u \le K$ such that $K_d + K_u = K$:*

1. *The quantity $\mu_s$ of Eq.* (79) *satisfies*

$$\mu_s \le \mu_{\max}, \tag{80}$$

*and*

$$K_u\left(\mu_{\max} + \mu_s\right) < K\mu_{\max}. \tag{81}$$

2. *For each index $i \notin s$,*

$$1 - \mu_s \le \left\|\left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\tilde{\mathbf{x}}^{(m)}_i\right\|^2_2 \le 1. \tag{82}$$

3. *For each pair of distinct indices $i \ne k$ such that $i, k \notin s$,*

$$\left|\left\langle\tilde{\mathbf{x}}^{(m)}_k, \left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\tilde{\mathbf{x}}^{(m)}_i\right\rangle\right| \le \mu_{\max} + \mu_s, \tag{83}$$

*and*

$$\left\|\left(\mathbf{I} - \mathbf{P}^{(m)}_s\right)\left(\mathbf{I} - \mathbf{P}^{(m)}_k\right)\tilde{\mathbf{x}}^{(m)}_i\right\|^2_2 \ge 1 - \mu^2_{\max} - \mu_s\left(1 + \mu_{\max}\right)^2. \tag{84}$$

*Furthermore,* $1 - \mu^2_{\max} - \mu_s\left(1 + \mu_{\max}\right)^2 > 0$.

For future use, notice that by its definition in Eq. (79), $\mu_s$ is an increasing function of $K_d$. Since $K_d \le K - 1$, then the quantity $\delta$ of Eq. (12) satisfies

$$\delta\left(K, \mu_{\max}\right) = \mu_s\left(K - 1, \mu_{\max}\right) \ge \mu_s\left(K_d, \mu_{\max}\right). \tag{85}$$

In addition, by Eq. (80), under max-MIP condition (6), $\delta \le \mu_{\max} < 1$, and hence the quantities in Section 4 are well defined.

For simplicity of notation, from now on we omit the dependence on the machine index $m$. Given the current estimated support set $s$, recall the definition of $\tilde{\theta}_{\max}$ in Eq. (23) and let $k \in \mathcal{S} \setminus s$ be an index for which

$$\|\tilde{\mathbf{x}}_k\| \cdot |\theta_k| = \tilde{\theta}_{\max} \tag{86}$$

(chosen arbitrarily in case of ties). By Eq. (24) for $\rho$ and Eq. (7) for $\theta_{\text{crit}}$,

$$\tilde{\theta}_{\max} = \frac{\sigma\sqrt{2\rho\log d}}{1-(2K-1)\mu_{\max}}. \tag{87}$$

We now prove Lemmas A.7 and A.8.

*Proof of Lemma A.7.* Recall that $q$, defined in Eq. (44), is the probability that some support index is selected by `OMP_Step`. A sufficient condition for this to occur is that the index $k$ defined in Eq. (86) has a higher correlation with the current residual than any non-support index $j \notin \mathcal{S}$. Thus, $q$ is lower bounded by the probability of the following event

$$|\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle| \geq \max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|. \tag{88}$$

Thus, to prove the lemma it suffices to lower bound the probability of event (88). Similarly to the proof of Lemma A.4, we decompose the noise $\tilde{\boldsymbol{\xi}}$ in Eq. (78) as the sum of two components, the first $\tilde{\boldsymbol{\xi}}_{\|} = \mathbf{P}_k\tilde{\boldsymbol{\xi}} = \left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}\right\rangle \tilde{\mathbf{x}}_k$ is parallel to $\tilde{\mathbf{x}}_k$, namely $\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle = \left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}\right\rangle$, and the second $\tilde{\boldsymbol{\xi}}_{\perp} = \tilde{\boldsymbol{\xi}} - \tilde{\boldsymbol{\xi}}_{\|} = (\mathbf{I}-\mathbf{P}_k)\,\tilde{\boldsymbol{\xi}}$, is orthogonal to $\tilde{\mathbf{x}}_k$, i.e., $\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\perp}\right\rangle = 0$.

Next, we use this decomposition to bound each of the terms in (88). By Eq. (78), for any index $i$,

$$\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle = \sum_{l\in\mathcal{S}\setminus s} \tilde{\theta}_l \langle \tilde{\mathbf{x}}_i, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle + \sigma\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}\right\rangle.$$

For the index $k$, $\|(\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_k\|_2^2 = \langle \tilde{\mathbf{x}}_k, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_k\rangle \geq 1-\mu_s$ by Eq. (82). For any other undetected support index $l \in \mathcal{S}\setminus\{s\cup k\}$, $|\langle \tilde{\mathbf{x}}_k, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle| \leq \mu_{\max}+\mu_s$ by (83) and $\left|\tilde{\theta}_l\right| \leq \tilde{\theta}_{\max}$ by its definition in Eq. (23). Combining these bounds with the definition of $\tilde{\boldsymbol{\xi}}_{\|}$ implies that the LHS of (88) can be bounded by

$$\begin{aligned}
|\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle| &\geq \operatorname{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle \\
&\geq \tilde{\theta}_{\max}\left(\langle \tilde{\mathbf{x}}_k, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_k\rangle - \sum_{l\in\mathcal{S}\setminus(s\cup\{k\})} |\langle \tilde{\mathbf{x}}_k, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle|\right) + \operatorname{sign}\left(\tilde{\theta}_k\right)\sigma\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}\right\rangle \\
&\geq \tilde{\theta}_{\max}\left(1-\mu_s-(K_u-1)\left(\mu_{\max}+\mu_s\right)\right) + \sigma\operatorname{sign}\left(\tilde{\theta}_k\right)\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle.
\end{aligned} \tag{89}$$

The RHS of (88) can be bounded by

$$\begin{aligned}
\max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle| &= \max_{i\notin\mathcal{S}}\left|\sum_{l\in\mathcal{S}\setminus s}\tilde{\theta}_l\langle \tilde{\mathbf{x}}_i, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle + \sigma\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}+\tilde{\boldsymbol{\xi}}_{\|}\right\rangle\right| \\
&\leq \tilde{\theta}_{\max}\max_{i\notin\mathcal{S}}\sum_{l\in\mathcal{S}\setminus s}|\langle \tilde{\mathbf{x}}_i, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle| + \sigma\max_{i\notin\mathcal{S}}\left|\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}\right\rangle\right| + \sigma\left|\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}\right\rangle\right|\max_{i\notin\mathcal{S}}|\langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k\rangle| \\
&\leq K_u\tilde{\theta}_{\max}\left(\mu_{\max}+\mu_s\right) + \sigma\max_{i\notin\mathcal{S}}\left|\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}\right\rangle\right| + \sigma\mu_{\max}\left|\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle\right|.
\end{aligned} \tag{90}$$

where the first step follows from Eq. (78) and the definitions of $\tilde{\boldsymbol{\xi}}_{\perp}$ and $\tilde{\boldsymbol{\xi}}_{\|}$, the second step follows from the triangle inequality and the definitions of $\tilde{\boldsymbol{\xi}}_{\|}$ and $\tilde{\theta}_{\max}$, and the last inequality follows from Eq. (83) and the definitions of $\mu_{\max}$ in Eq. (5). Combining Eq. (89) with Eq. (90) implies that a sufficient condition for (88) to occur is

$$\max_{i\notin\mathcal{S}}\left|\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}\right\rangle\right| \leq \operatorname{sign}\left(\tilde{\theta}_k\right)\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle - \mu_{\max}\left|\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle\right| + \frac{\tilde{\theta}_{\max}\left(1-2K_u\left(\mu_{\max}+\mu_s\right)+\mu_{\max}\right)}{\sigma}.$$

By Eq. (87) and by the inequality (81), a sufficient condition for the previous event to occur is

$$\max_{i\notin\mathcal{S}}\left|\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}\right\rangle\right| \leq \operatorname{sign}\left(\tilde{\theta}_k\right)\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle - \mu_{\max}\left|\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle\right| + \sqrt{2\rho\log d}. \tag{91}$$

As in the proof of Lemma A.4, denote the LHS of (91) by $A$, its RHS by $B$ and let $T = \sqrt{2\log d}$. By Eq. (58), it suffices to bound the probabilities of $A \le T$ and $B \ge T$.

We begin with bounding $\Pr[A \le T]$. Fix $i \notin \mathcal{S}$. By definition $\tilde{\boldsymbol{\xi}}_\perp = (\mathbf{I}-\mathbf{P}_k)\tilde{\boldsymbol{\xi}} = (\mathbf{I}-\mathbf{P}_k)(\mathbf{I}-\mathbf{P}_s)\boldsymbol{\xi}$. By the symmetry of projections, $\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_\perp \right\rangle = \langle \tilde{\mathbf{x}}_i, (\mathbf{I}-\mathbf{P}_k)(\mathbf{I}-\mathbf{P}_s)\boldsymbol{\xi}\rangle = \langle (\mathbf{I}-\mathbf{P}_s)(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i, \boldsymbol{\xi}\rangle$. Normalizing the inner product results in a standard normal random variable $Z_i = \frac{\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_\perp \right\rangle}{\|(\mathbf{I}-\mathbf{P}_s)(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i\|_2} \sim \mathcal{N}(0,1)$. By Eq. (84), $\|(\mathbf{I}-\mathbf{P}_s)(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i\|_2 \ge \gamma_1$, where $\gamma_1 = \sqrt{1-\mu_{\max}^2-\mu_s(1+\mu_{\max})^2}$. As in the proof of Lemma A.4, it follows that

$$\Pr[A \le T] \ge 1 - \frac{\gamma_1}{\sqrt{\pi \log d}} d^{-\frac{1}{\gamma_1^2}+1} \ge \frac{1}{2}, \tag{92}$$

where the last inequality holds for sufficiently large $d$ and follows from noting that $\gamma_1 \le 1$ by the max-MIP condition (6).

We now bound $\Pr[B \ge T]$, where $B$ is the RHS of Eq. (91). By definition of $\tilde{\boldsymbol{\xi}}_\parallel$, the inner product $\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_\parallel \right\rangle = \left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}} \right\rangle = \langle \tilde{\mathbf{x}}_k, (\mathbf{I}-\mathbf{P}_s)\boldsymbol{\xi}\rangle$. This random variable is equal in distribution to a Gaussian random variable $Z \sim \mathcal{N}\left(0, \|(\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_k\|_2^2\right)$. By Eq. (82), $\|(\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_k\|_2 \ge \sqrt{1-\mu_s}$. As in the proof of Lemma A.4,

$$\Pr[B \ge T] \ge \Phi^c\left(\frac{1-\sqrt{\rho}}{\gamma_2}\sqrt{2\log d}\right), \tag{93}$$

where $\gamma_2 = \sqrt{1-\mu_s}(1-\mu_{\max})$. Recall the definition of $\delta$ in Eq. (12). By Eq. (85), $\gamma_2 \ge \sqrt{1-\delta}(1-\mu_{\max})$. Combining this with the bounds (92) and (93) completes the proof of Lemma A.7. □

*Proof of Lemma A.8.* Fix a non-support index $j \notin \mathcal{S}$. Recall that $p_j$, defined in Eq. (32), is the probability that index $j$ is selected by `OMP_Step`. This occurs if $j$ has the highest correlation with the current residual, i.e.,

$$p_j = \Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| \ge \max_{i \in [d]\setminus s} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|\right].$$

Clearly, by taking the maximum over a subset of the indices $\mathcal{A} \subseteq [d]\setminus s$ that includes $j$, the probability can only be higher. Namely,

$$p_j \ge \Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| \ge \max_{i \in \mathcal{A}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|\right]. \tag{94}$$

Here we take $\mathcal{A}$ as the set of all non-support indices plus the index $k$, i.e., $\mathcal{A} = ([d]\setminus\mathcal{S}) \cup \{k\}$, where $k$ is defined in Eq. (86). Next, we separately upper bound

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| > \max_{i \notin \mathcal{S}\cup\{j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|\right] \tag{95}$$

and

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| > |\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle|\right] \tag{96}$$

and then upper bound $p_j$ using (65) with $A = |\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle|$, $B = \max_{i \notin (\mathcal{S}\cup\{j\})} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|$, and $C = |\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle|$.

For later use in both bounds, the random variable $A$ can be upper bounded as follows

$$\begin{aligned} |\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| &= \left| \sum_{l \in \mathcal{S}\setminus s} \tilde{\theta}_l \langle \tilde{\mathbf{x}}_j, (\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_l\rangle + \sigma \left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \right\rangle \right| \\ &\le \tilde{\theta}_{\max} \sum_{l \in \mathcal{S}\setminus s} |\langle \tilde{\mathbf{x}}_j, (\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_l\rangle| + \sigma \left|\left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \right\rangle\right| \le \tilde{\theta}_{\max} K_u (\mu_{\max}+\mu_s) + \sigma \left|\left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \right\rangle\right|, \end{aligned} \tag{97}$$

where the first equality follows from Eq. (78), the next inequality follows from the triangle inequality and the definition of $\tilde{\theta}_{\max}$ in Eq. (23), and the last inequality follow from (83). We now begin with event (95). By Eqs.

(83) and (81), for each non-support index $i \notin \mathcal{S}$ such that $i \neq j$,

$$\begin{aligned} |\langle \tilde{\mathbf{x}}_i, \mathbf{r} \rangle| &\geq \langle \tilde{\mathbf{x}}_i, \mathbf{r} \rangle = \sum_{l \in \mathcal{S} \setminus s} \tilde{\theta}_l \langle \tilde{\mathbf{x}}_i, (\mathbf{I} - \mathbf{P}_s) \tilde{\mathbf{x}}_l \rangle + \sigma \left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}} \right\rangle \\ &\geq -\tilde{\theta}_{\max} K_u (\mu_{\max} + \mu_s) + \sigma \left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}} \right\rangle. \end{aligned}$$

Combining the above bound with Eq. (97), rearranging the terms, recalling the relation between $\tilde{\theta}_{\max}$ and $\rho$ in (87) and applying inequality (81) yields

$$\Pr\left[ |\langle \tilde{\mathbf{x}}_j, \mathbf{r} \rangle| > \max_{i \notin \mathcal{S} \cup \{j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r} \rangle| \right] \leq \Pr\left[ \left| \left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \right\rangle \right| + \frac{2K\mu_{\max}\sqrt{2\rho \log d}}{1 - (2K-1)\mu_{\max}} > \max_{i \notin \mathcal{S} \cup \{j\}} \left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}} \right\rangle \right]. \tag{98}$$

As in the proof of Lemma A.5, applying (68) with $T = (1-\epsilon)\sqrt{2(1-\mu_{\max}) \log d}$ and $\epsilon \in (0,1)$ as in (15), we can upper bound (98) by

$$\Pr\left[ \left| \langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \rangle \right| \geq a\sqrt{2 \log d} \right] + \Pr\left[ \max_{i \notin \mathcal{S} \cup \{j\}} \langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}} \rangle < (1-\epsilon)\sqrt{2(1-\mu_{\max}) \log d} \right], \tag{99}$$

where

$$a = (1-\epsilon)\sqrt{1-\mu_{\max}} - \frac{2K\mu_{\max}\sqrt{\rho}}{1-(2K-1)\mu_{\max}}.$$

By the symmetry of projection matrices, $\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \rangle = \langle (\mathbf{I} - \mathbf{P}_s) \tilde{\mathbf{x}}_j, \boldsymbol{\xi} \rangle$. By Eq. (82), the norm $\|(\mathbf{I} - \mathbf{P}_s) \tilde{\mathbf{x}}_j\|_2 \leq 1$ and thus the first term in (99) is bounded by

$$2\Phi^c\left( a\sqrt{2 \log d} \right). \tag{100}$$

We now bound the second term in (99) using Lemma A.12 with $Z_i = \left\langle \frac{(\mathbf{I} - \mathbf{P}_s)\tilde{\mathbf{x}}_i}{\|(\mathbf{I} - \mathbf{P}_s)\tilde{\mathbf{x}}_i\|_2}, \boldsymbol{\xi} \right\rangle$. Towards this goal, notice that by Eq. (82), $\|(\mathbf{I} - \mathbf{P}_s) \tilde{\mathbf{x}}_i\|_2 \geq \sqrt{1-\mu_s}$. Thus, the second term in (99) is upper bounded by

$$\Pr\left[ \max_{i \notin \mathcal{S} \cup \{j\}} Z_i < \frac{(1-\epsilon)\sqrt{2(1-\mu_{\max}) \log d}}{\sqrt{1-\mu_s}} \right]. \tag{101}$$

Furthermore, by Eqs. (82) and (83), for each $i, l \notin \mathcal{S} \cup \{j\}$ such that $i \neq l$, $\mathrm{E}[Z_i Z_l] \leq \frac{\mu_{\max} + \mu_s}{1-\mu_s}$. Thus, we can apply Lemma A.12 with $\eta = \frac{\mu_{\max} + \mu_s}{1-\mu_s}$ and $\zeta = 1 - \epsilon$ to obtain that (101) is bounded by

$$C(d-K-1)^{-\frac{1-\mu_{\max}}{\mu_{\max}+\mu_s}\epsilon^2} \log^{\left(\frac{1-\mu_{\max}}{\mu_{\max}+\mu_s}\epsilon - 1\right)/2}(d-K-1). \tag{102}$$

Similarly to the proof of Lemma A.5, under condition (15) on $\epsilon$ and for sufficiently large $d = d(\epsilon)$, (100) is larger than (102), and thus

$$\Pr\left[ |\langle \tilde{\mathbf{x}}_j, \mathbf{r} \rangle| > \max_{i \notin \mathcal{S} \cup \{j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r} \rangle| \right] \leq 4\Phi^c\left( a\sqrt{2 \log d} \right). \tag{103}$$

We now turn to bounding (96). For the support index $k$, similarly to (89),

$$|\langle \tilde{\mathbf{x}}_k, \mathbf{r} \rangle| \geq \operatorname{sign}\left(\tilde{\theta}_k\right) \langle \tilde{\mathbf{x}}_k, \mathbf{r} \rangle \geq \tilde{\theta}_{\max}(1-\mu_s) - \tilde{\theta}_{\max}(K_u - 1)(\mu_{\max} + \mu_s) + \sigma \operatorname{sign}\left(\tilde{\theta}_k\right) \left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}} \right\rangle.$$

Combining the bound above with Eq. (97), recalling the relation between $\tilde{\theta}_{\max}$ and $\rho$ in (87) and applying the inequality (81) yields

$$\Pr\left[ |\langle \tilde{\mathbf{x}}_j, \mathbf{r} \rangle| > |\langle \tilde{\mathbf{x}}_k, \mathbf{r} \rangle| \right] \leq \Pr\left[ \left| \left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \right\rangle \right| - \operatorname{sign}\left(\tilde{\theta}_k\right) \left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}} \right\rangle > \sqrt{2\rho \log d} \right].$$

We now upper bound this probability. Let $H = \left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}} \right\rangle = \langle (\mathbf{I} - \mathbf{P}_s)\, \tilde{\mathbf{x}}_j, \boldsymbol{\xi} \rangle$ and $G = \operatorname{sign}\left(\tilde{\theta}_k\right) \left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}} \right\rangle = \operatorname{sign}\left(\tilde{\theta}_k\right) \langle (\mathbf{I} - \mathbf{P}_s)\, \tilde{\mathbf{x}}_k, \boldsymbol{\xi} \rangle$. Notice that by Eqs. (82), (83), and (80), $\sigma_H^2, \sigma_G^2 \le 1$. Combining Eqs. (12) and (85) implies that $\mu_s \le \delta$. Hence, $|\sigma_{HG}| \le \mu_{\max} + \mu_s \le \mu_{\max} + \delta$. Thus, as in the proof of Lemma A.5,

$$\Pr\left[|H| - G > c\right] \le 2\Phi^c\left(\sqrt{\frac{2\rho \log d}{2 + 2\left(\mu_{\max} + \delta\right)}}\right). \tag{104}$$

By Eq. (65), the probability (94) is at most the minimum between (103) and (104). By the monotonicity of the Gaussian CDF, (94) is upper bounded by

$$4\Phi^c\left(\max\left\{\sqrt{\frac{\rho}{2 + 2\left(\mu_{\max} + \delta\right)}}, \left(\left(1-\epsilon\right)\sqrt{1-\mu_{\max}} - \frac{2K\mu_{\max}\sqrt{\rho}}{1 - (2K-1)\mu_{\max}}\right)\right\}\sqrt{2\log d}\right). \tag{105}$$

Similarly to the proof of Lemma A.5, inserting the definitions of $Q_1$ and $Q_2$ in Eqs. (17) and (18) respectively, into Eq. (19) and rearranging various terms yields

$$\frac{1-\sqrt{r}}{\sqrt{1-\delta}\left(1-\mu_{\max}\right)} + \sqrt{Q_0} < \max\left\{\sqrt{\frac{r}{2 + 2\left(\mu_{\max} + \delta\right)}}, \left(\left(1-\epsilon\right)\sqrt{1-\mu_{\max}} - \frac{2K\mu_{\max}\sqrt{r}}{1 - (2K-1)\mu_{\max}}\right)\right\}. \tag{106}$$

The definitions of $r$ and $\rho$ in Eqs. (10) and (24) imply that $\rho \ge r$. Thus, $\rho$ satisfies Eq. (19) and hence Eq. (106). The RHS of Eq. (106) is the same as the maximum in Eq. (105) above. Thus, Eq. (105) is upper bounded by

$$4\Phi^c\left(\left(\frac{1-\sqrt{\rho}}{\sqrt{1-\delta}\left(1-\mu_{\max}\right)} + \sqrt{Q_0}\right)\sqrt{2\log d}\right). \tag{107}$$

By the assumption $\rho \le 1$, we can apply Lemma A.11. Hence, by the definition of $Q_0$ in Eq. (16), and by the definition of $F$ in Eq. (28),

$$\begin{aligned} p_j &\le 4\Phi^c\left(\left(\frac{1-\sqrt{\rho}}{\sqrt{1-\delta}\left(1-\mu_{\max}\right)} + \sqrt{Q_0}\right)\sqrt{2\log d}\right) \le 4\sqrt{2}d^{-Q_0}\Phi^c\left(\frac{1-\sqrt{\rho}}{\sqrt{1-\delta}\left(1-\mu_{\max}\right)}\sqrt{2\log d}\right) \\ &= 4\sqrt{2}\frac{1}{88\sqrt{2}K}\Phi^c\left(\frac{1-\sqrt{\rho}}{\sqrt{1-\delta}\left(1-\mu_{\max}\right)}\sqrt{2\log d}\right) = \frac{F\left(d, K, \mu_{\max}, \rho\right)}{11K}, \end{aligned}$$

which completes the proof of Lemma A.8. □

### A.6 Proofs of technical lemmas

*Proof of Lemma A.11.* Classical results by Birnbaum (1942) and Komatu (1955) are that for all $x \ge 0$, the following inequalities hold

$$\frac{2e^{-x^2/2}}{\sqrt{2\pi}\left(\sqrt{x^2+4}+x\right)} < \Phi^c\left(x\right) < \frac{2e^{-x^2/2}}{\sqrt{2\pi}\left(\sqrt{x^2+2}+x\right)}. \tag{108}$$

Hence,

$$\Phi^c\left(a+b\right) < \frac{2e^{-(a+b)^2/2}}{\sqrt{2\pi}\left(\sqrt{(a+b)^2+2}+a+b\right)},$$

and

$$\Phi^c\left(a\right) > \frac{2e^{-a^2/2}}{\sqrt{2\pi}\left(\sqrt{a^2+4}+a\right)}.$$

Combining the two yields the following

$$\Phi^c\left(a+b\right) < \frac{\left(\sqrt{a^2+4}+a\right)e^{-ab}}{\left(\sqrt{(a+b)^2+2}+a+b\right)}e^{-b^2/2}\Phi^c\left(a\right).$$

Notice that for any $a \ge 0$, the fraction in the above display is a decreasing function of $b$. Since $b \ge 0$, it suffices to note that $\frac{\left(\sqrt{a^2+4}+a\right)}{\left(\sqrt{a^2+2}+a\right)} \le \sqrt{2}$ for any $a \ge 0$. □

Towards proving Lemma A.13, we prove the following Lemma A.14, which bounds the inner product between vectors projected to the subspace orthogonal to $\tilde{\mathbf{X}}_{|s}$ under the assumption $s \subset \mathcal{S}$.

**Lemma A.14.** *Let $s \subset \mathcal{S}$ and denote $K_d = |s|$. Assume that $(K_d - 1)\,\mu_{\max} < 1$. Then, for any pair of vectors $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^n$*

$$\langle \mathbf{a}_1, \mathbf{a}_2 \rangle - \frac{\left|\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a}_1 \rangle \langle \tilde{\mathbf{x}}_j, \mathbf{a}_2 \rangle\right|}{1-(K_d-1)\,\mu_{\max}} \leq \langle \mathbf{a}_1, (\mathbf{I}-\mathbf{P}_s)\,\mathbf{a}_2 \rangle \leq \langle \mathbf{a}_1, \mathbf{a}_2 \rangle + \frac{\left|\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a}_1 \rangle \langle \tilde{\mathbf{x}}_j, \mathbf{a}_2 \rangle\right|}{1-(K_d-1)\,\mu_{\max}}. \tag{109}$$

*If in addition $\mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}$, then*

$$\|\mathbf{a}\|_2^2 - \frac{\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a} \rangle^2}{1-(K_d-1)\,\mu_{\max}} \leq \|(\mathbf{I}-\mathbf{P}_s)\,\mathbf{a}\|_2^2 \leq \|\mathbf{a}\|_2^2 - \frac{\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a} \rangle^2}{1+(K_d-1)\,\mu_{\max}}. \tag{110}$$

*Proof of Lemma A.14.* First, if $s = \emptyset$ then clearly $\langle \mathbf{a}_1, (\mathbf{I}-\mathbf{P}_s)\,\mathbf{a}_2 \rangle = \langle \mathbf{a}_1, \mathbf{a}_2 \rangle$ and both (109) and (110) trivially hold. Therefore, assume that $K_d \geq 1$. In this case

$$\langle \mathbf{a}_1, (\mathbf{I}-\mathbf{P}_s)\,\mathbf{a}_2 \rangle = \langle \mathbf{a}_1, \mathbf{a}_2 \rangle - \langle \mathbf{a}_1, \mathbf{P}_s \mathbf{a}_2 \rangle\,. \tag{111}$$

By definition of $\mathbf{P}_s$ in Eq. (49),

$$\langle \mathbf{a}_1, \mathbf{P}_s \mathbf{a}_2 \rangle = \left\langle \mathbf{a}_1, \tilde{\mathbf{X}}_{|s} \left(\tilde{\mathbf{X}}_{|s}^T \tilde{\mathbf{X}}_{|s}\right)^{-1} \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_2 \right\rangle = \left\langle \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_1, \left(\tilde{\mathbf{X}}_{|s}^T \tilde{\mathbf{X}}_{|s}\right)^{-1} \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_2 \right\rangle. \tag{112}$$

We now bound this term in absolute value. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, denote by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ its minimal and maximal eigenvalues, respectively. Consider $\mathbf{A} = \tilde{\mathbf{X}}_{|s}^T \tilde{\mathbf{X}}_{|s}$. Each of its entries $\mathbf{A}_{i,j}$ is an inner product $\langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j \rangle$ where $i, j \in s$. Hence, all of its diagonal entries are 1 and all of its off-diagonal entries are bounded in absolute value by $\mu_{\max}$. By the Gershgorin circle theorem, the eigenvalues of $\mathbf{A}$ lie in the interval $1 \pm (K_d - 1)\mu_{\max}$. Since $(K_d - 1)\,\mu_{\max} < 1$, all eigenvalues are strictly positive. Thus $\mathbf{A}$ is invertible, and the eigenvalues of $\mathbf{A}^{-1}$ satisfy

$$\frac{1}{1+(K_d-1)\,\mu_{\max}} \leq \lambda_{\min}\left(\mathbf{A}^{-1}\right) \leq \lambda_{\max}\left(\mathbf{A}^{-1}\right) \leq \frac{1}{1-(K_d-1)\,\mu_{\max}}. \tag{113}$$

Since the eigenvalues of $\mathbf{A}^{-1}$ are strictly positive, for any pair of vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$,

$$\lambda_{\min}\left(\mathbf{A}^{-1}\right) |\langle \mathbf{u}, \mathbf{v} \rangle| \leq \left|\left\langle \mathbf{u}, \mathbf{A}^{-1}\mathbf{v} \right\rangle\right| \leq \lambda_{\max}\left(\mathbf{A}^{-1}\right) |\langle \mathbf{u}, \mathbf{v} \rangle|\,.$$

Inserting $\mathbf{u} = \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_1$, $\mathbf{v} = \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_2$ and Eq. (112) yields

$$\lambda_{\min}\left(\mathbf{A}^{-1}\right) \left|\left\langle \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_1, \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_2 \right\rangle\right| \leq |\langle \mathbf{a}_1, \mathbf{P}_s \mathbf{a}_2 \rangle| \leq \lambda_{\max}\left(\mathbf{A}^{-1}\right) \left|\left\langle \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_1, \tilde{\mathbf{X}}_{|s}^T \mathbf{a}_2 \right\rangle\right|.$$

Combining the bounds in Eq. (113) with the decomposition $\tilde{\mathbf{X}}_{|s} \tilde{\mathbf{X}}_{|s}^T = \sum_{j\in s} \tilde{\mathbf{x}}_j \tilde{\mathbf{x}}_j^T$ gives

$$\frac{\left|\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a}_1 \rangle \langle \tilde{\mathbf{x}}_j, \mathbf{a}_2 \rangle\right|}{1+(K_d-1)\,\mu_{\max}} \leq |\langle \mathbf{a}_1, \mathbf{P}_s \mathbf{a}_2 \rangle| \leq \frac{\left|\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a}_1 \rangle \langle \tilde{\mathbf{x}}_j, \mathbf{a}_2 \rangle\right|}{1-(K_d-1)\,\mu_{\max}}. \tag{114}$$

When $\mathbf{a}_1 \neq \mathbf{a}_2$, the term $\langle \mathbf{a}_1, \mathbf{P}_s \mathbf{a}_2 \rangle$ can have an arbitrary sign. Thus, inserting the upper bound into Eq. (111) proves the inequality (109).

We now prove the inequality (110). Let $\mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}$. Since each of the terms in the two sums in Eq. (114) is positive, we may remove the absolute values, i.e.,

$$\frac{\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a} \rangle^2}{1+(K_d-1)\,\mu_{\max}} \leq \langle \mathbf{a}, \mathbf{P}_s \mathbf{a} \rangle \leq \frac{\sum_{j\in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a} \rangle^2}{1-(K_d-1)\,\mu_{\max}}. \tag{115}$$

Recall that since $(\mathbf{I}-\mathbf{P}_s)$ is a projection matrix, it is symmetric and idempotent. Thus,

$$(\mathbf{I}-\mathbf{P}_s)^T (\mathbf{I}-\mathbf{P}_s) = (\mathbf{I}-\mathbf{P}_s)(\mathbf{I}-\mathbf{P}_s) = (\mathbf{I}-\mathbf{P}_s)\,. \tag{116}$$

Hence,

$$\|(\mathbf{I}-\mathbf{P}_s)\,\mathbf{a}\|_2^2 = \langle \mathbf{a}, (\mathbf{I}-\mathbf{P}_s)\,\mathbf{a} \rangle = \|\mathbf{a}\|^2 - \langle \mathbf{a}, \mathbf{P}_s \mathbf{a} \rangle\,.$$

Inserting inequality (115) completes the proof of (110) and of Lemma A.14. □

*Proof of Lemma A.13.* We begin with proving inequalities (80) and (81). By the max-MIP condition (6), $1-(2K-1)\,\mu_{\max}>0$. Rearranging implies that

$$1-(K-2)\,\mu_{\max}>(K+1)\,\mu_{\max}.$$

Combining this with the bound on $\mu_s$ in (85) gives

$$\mu_s \le \frac{K-1}{K+1}\mu_{\max} \le \mu_{\max},$$

which proves (80). The max-MIP condition (6) implies that $1-(K-1)\,\mu_{\max}>0$. Using $K_d+K_u=K$ and rearranging yields $\frac{K_u\mu_{\max}}{1-(K_d-1)\mu_{\max}}<1$. Combining the definition of $\mu_s$ in (79) with this bound implies that

$$K_u\mu_s = K_d\frac{K_u\mu_{\max}^2}{1-(K_d-1)\mu_{\max}} < K_d\mu_{\max}.$$

Hence,

$$K\mu_{\max} = K_u\mu_{\max}+K_d\mu_{\max} > K_u\mu_{\max}+K_u\mu_s,$$

which proves (81).

We now prove the remaining inequalities using Lemma A.14, beginning with (82). Since $\mathbf{P}_s$ is a projection matrix, for any index $i\notin s$,

$$\|(\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_i\|_2^2 \le \|\tilde{\mathbf{x}}_i\|_2^2 = 1.$$

Recall that for any distinct pair of indices $i\neq j$, it holds that $0\le\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i\rangle^2\le\mu_{\max}^2$. By Eq. (110) with $\mathbf{a}=\tilde{\mathbf{x}}_i$,

$$\|(\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_i\|_2^2 \ge 1-\frac{\sum_{j\in s}\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i\rangle^2}{1-(K_d-1)\,\mu_{\max}} \ge 1-\frac{K_d\mu_{\max}^2}{1-(K_d-1)\,\mu_{\max}} = 1-\mu_s,$$

which concludes the proof of (82).

Next, we prove inequality (83). By the right inequality in (109) with $\mathbf{a}_1=\tilde{\mathbf{x}}_k$ and $\mathbf{a}_2=\tilde{\mathbf{x}}_i$,

$$\langle\tilde{\mathbf{x}}_k,(\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_i\rangle \le \langle\tilde{\mathbf{x}}_k,\tilde{\mathbf{x}}_i\rangle + \frac{\left|\sum_{j\in s}\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_k\rangle\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i\rangle\right|}{1-(K_d-1)\,\mu_{\max}}.$$

Thus, by the triangle inequality and by the definitions of $\mu_{\max}$ and $\mu_s$ in Eqs. (5) and (79) respectively,

$$\begin{aligned}|\langle\tilde{\mathbf{x}}_k,(\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_i\rangle| &\le |\langle\tilde{\mathbf{x}}_k,\tilde{\mathbf{x}}_i\rangle| + \frac{\left|\sum_{j\in s}\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_k\rangle\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i\rangle\right|}{1-(K_d-1)\,\mu_{\max}}\\ &\le \mu_{\max}+\frac{K_d\mu_{\max}^2}{1-(K_d-1)\,\mu_{\max}} = \mu_{\max}+\mu_s.\end{aligned}$$

Finally, we prove inequality (84). Recall that by the max-MIP condition (6), $\mu_{\max}<\frac{1}{K-1}$. For any distinct pair of indices $i\neq k$ such that $i,k\notin s$, Eq. (110) with $\mathbf{a}=(\mathbf{I}-\mathbf{P}_k)\,\tilde{\mathbf{x}}_i$ gives

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P}_s)\,(\mathbf{I}-\mathbf{P}_k)\,\tilde{\mathbf{x}}_i\|_2^2 &\ge \|(\mathbf{I}-\mathbf{P}_k)\,\tilde{\mathbf{x}}_i\|_2^2 - \frac{\sum_{j\in s}\langle\tilde{\mathbf{x}}_j,(\mathbf{I}-\mathbf{P}_k)\,\tilde{\mathbf{x}}_i\rangle^2}{1-(K_d-1)\,\mu_{\max}}\\ &= 1-\langle\tilde{\mathbf{x}}_k,\tilde{\mathbf{x}}_i\rangle^2 - \frac{\sum_{j\in s}\left(\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_i\rangle-\langle\tilde{\mathbf{x}}_k,\tilde{\mathbf{x}}_i\rangle\langle\tilde{\mathbf{x}}_j,\tilde{\mathbf{x}}_k\rangle\right)^2}{1-(K_d-1)\,\mu_{\max}}\\ &\ge 1-\mu_{\max}^2-\frac{K_d\left(\mu_{\max}+\mu_{\max}^2\right)^2}{1-(K_d-1)\,\mu_{\max}}\\ &= 1-\mu_{\max}^2-\mu_s(1+\mu_{\max})^2,\end{aligned}$$

which concludes the proof of Eq. (84). It remains to prove that

$$1-\mu_{\max}^2-\mu_s(1+\mu_{\max})^2>0.$$

First, let $K = 1$. This implies that $s = \emptyset$ and thus

$$\|(\mathbf{I} - \mathbf{P}_s)(\mathbf{I} - \mathbf{P}_k)\tilde{\mathbf{x}}_i\|_2 = \|(\mathbf{I} - \mathbf{P}_k)\tilde{\mathbf{x}}_i\|_2 \geq 1 - \mu_{\max}^2,$$

which is positive by the max-MIP condition (6). Now let $K > 1$. By the max-MIP condition (6), $\mu_{\max} < 1$ and $\frac{K_d \mu_{\max}}{1-(K_d-1)\mu_{\max}} < 1$. Thus,

$$\begin{aligned} 1 - \mu_{\max}^2 - \mu_s(1+\mu_{\max})^2 &> 1 - \mu_{\max}^2 - \mu_{\max}(1+\mu_{\max})^2 \\ &= 1 - \mu_{\max}\left(1 + 3\mu_{\max} + \mu_{\max}^2\right) \\ &> 1 - \mu_{\max}(1 + 4\mu_{\max}). \end{aligned}$$

Note that for each $K > 1$, it holds that $\mu_{\max} < \frac{K-1}{2}$. Thus,

$$1 - \mu_{\max}(1 + 4\mu_{\max}) > 1 - \mu_{\max}(1 + 2K - 2) > 0,$$

where the last inequality is another application of the max-MIP condition (6). □

# B UNKNOWN SPARSITY LEVEL

As mentioned in Remark 4.3, when the sparsity level $K$ is unknown, a threshold-based variant of DJ-OMP can be used to recover the support. In the following Corollary B.1, we prove that this variant, denoted DJ-OMP$^*$, succeeds with high probability in estimating both $K$ and the support of $\boldsymbol{\theta}$ when the design matrices are composed of i.i.d. Bernoulli entries. Note that the corollary assumes that $M > M_c$ and $M_c$ depends on $K$, however this is solely for the purpose of the proof. The corollary holds for a wide range of $M$ values and DJ-OMP$^*$ does not receive $K$ nor $M_c$ as input. After stating the corollary we present simulation results comparing DJ-OMP and DJ-OMP$^*$. The proof of Corollary B.1 completes this section.

**Corollary B.1.** *Denote by DJ-OMP$^*$ a variant of DJ-OMP in which the fusion center stops the communication rounds with the $M$ machines and outputs its current support set estimation when the number of votes for the most-voted index in the current round falls below a threshold of $\tilde{t}_c = 4\log d$. If the matrices $\mathbf{X}^{(m)}$ have i.i.d. Bernoulli $\pm\frac{1}{\sqrt{n}}$ entries and $n \geq 2(2K-1)^2\log(2Md^3)$, then under the conditions of Theorem 4.1, DJ-OMP$^*$ with $M_c \leq M \leq 2e^{-1}(d-K)\log d$ machines detects the correct support w.p. at least $1 - \frac{2^{K+1}}{d}$.*

Figure 3 compares the empirical support recovery probabilities of DJ-OMP and DJ-OMP$^*$ as a function of $\theta_{\min}$. The parameters are the same as those used for Figure 1(a), i.e., we generated $M = 20$ matrices of dimensions $n = 2000$ and $d = 10000$, with i.i.d. $\mathcal{N}(0,1)$ entries ($\alpha = 0$). The noise level is $\sigma = 1$, and the vector $\boldsymbol{\theta}$ has sparsity $K = 5$, with $\boldsymbol{\theta} = \theta_{\min} \cdot [1, -1.5, 2, -2.5, 3, 0, \ldots, 0]^\top$. Since assumption 4.1 does not hold and $M$ is small compared to the theoretical value $M_c$, the simulations use the threshold $\tilde{t}_c = 2$ for DJ-OMP$^*$, i.e., the center stops and returns its support set estimation once the top-voted index receives less than 2 votes. As the figure demonstrates, the success probability is not greatly affected by the use of a threshold-based stopping criterion that does not depend on $K$.

*Proof of Corollary B.1.* Denote by $B$ an event where the max-MIP condition (6) is not satisfied. We first show that this event occurs with probability at most $1/d$. For each machine $m \in [M]$, each entry of the design matrix $\mathbf{X}^{(m)}$ is an i.i.d. Bernoulli $\pm\frac{1}{\sqrt{n}}$ random variable. Thus, the inner product between two vectors $\mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)} \in \mathbb{R}^n$ where $i \neq j \in [d]$ is a sum of $n$ i.i.d. Bernoulli $\pm\frac{1}{n}$ random variables. In addition, by design each vector has unit $\ell_2$ norm. By Hoeffding's inequality (Hoeffding, 1963),

$$\forall t > 0, \quad \Pr\left[\left|\left\langle \mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)} \right\rangle\right| > t\right] < 2\exp\left(-nt^2/2\right).$$

The max-MIP condition (6) requires that the maximal inner product among all $M\binom{d}{2}$ pairs in all machines is bounded by $1/(2K-1)$. Combining the Hoeffding bound above with a union bound yields

$$\Pr[B] = \Pr\left[\max_{m\in[M]} \max_{i\neq j} \left|\left\langle \mathbf{x}_i^{(m)}, \mathbf{x}_j^{(m)} \right\rangle\right| > \frac{1}{2K-1}\right] < 2Md^2 \exp\left(-\frac{n}{2(2K-1)^2}\right) \leq \frac{1}{d}.$$

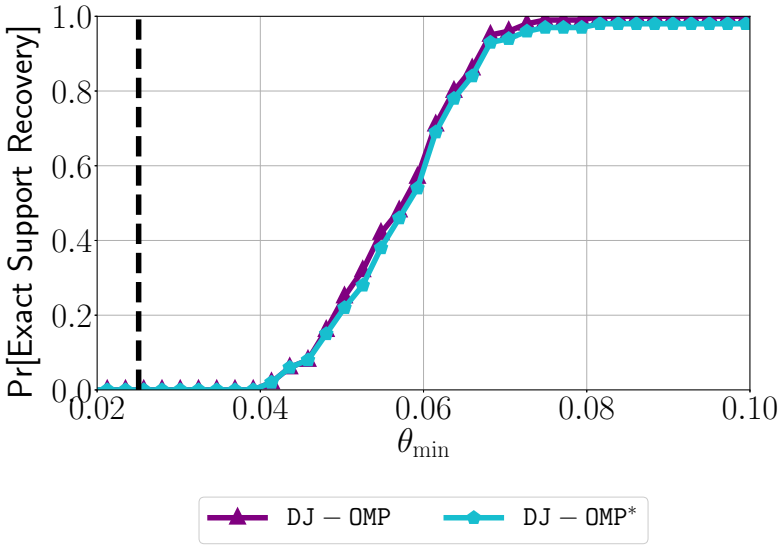



Figure 3: Support Recovery by DJ-OMP and DJ-OMP* as a Function of $\theta_{\min}$

We now show that when event $B$ does not occur, i.e., when the max-MIP condition (6) is satisfied, then DJ-OMP* recovers the support of $\boldsymbol{\theta}$ with probability at least $1 - \frac{2^{K+1}-1}{d}$. By Lemma A.1, the threshold $t_c$ of the proof of Theorem 4.1 satisfies $t_c \geq 4 \log d$. Thus, by essentially the same proof, with probability at least $1 - \frac{2^{K+1}-2}{d}$ the algorithm DJ-OMP* does not stop early, and recovers all $K$ support indices after $K$ rounds. It remains to show that the probability that it does not stop after $K$ rounds and adds another (non-support) element to its estimate is at most $\frac{1}{d}$.

Assume that $s = \mathcal{S}$. Hence, in each machine $m \in [M]$, similarly to Eq. (78), the residual $\mathbf{r}^{(m)}$ is

$$\begin{aligned} \mathbf{r}^{(m)} &= \left(\mathbf{I} - \mathbf{P}_{\mathcal{S}}^{(m)}\right) \mathbf{y}^{(m)} = \left(\mathbf{I} - \mathbf{P}_{\mathcal{S}}^{(m)}\right) \left(\mathbf{X}^{(m)} \boldsymbol{\theta}^{(m)} + \sigma \boldsymbol{\xi}^{(m)}\right) \\ &= \sigma \left(\mathbf{I} - \mathbf{P}_{\mathcal{S}}^{(m)}\right) \boldsymbol{\xi}^{(m)}. \end{aligned} \tag{117}$$

In other words, since the residual $\mathbf{r}^{(m)}$ is orthogonal to the set of vectors $\left\{\mathbf{x}_k^{(m)}\right\}_{k \in \mathcal{S}}$, it is composed only of (projected) noise. Recall that an index that has already been added to the support cannot be sent again. Thus, the probability $p_j^{(m)}$ that machine $m$ sends a fixed non-support index $j \notin \mathcal{S}$ is

$$p_j^{(m)} = \Pr\left[j = \arg\max_{i \notin \mathcal{S}} \left|\left\langle \mathbf{x}_i^{(m)}, \mathbf{r}^{(m)} \right\rangle\right|\right] = \Pr\left[j = \arg\max_{i \notin \mathcal{S}} \left|\left\langle \mathbf{x}_i^{(m)}, \left(\mathbf{I} - \mathbf{P}_{\mathcal{S}}^{(m)}\right) \boldsymbol{\xi}^{(m)} \right\rangle\right|\right]. \tag{118}$$

Since each entry of each matrix is i.i.d., then by symmetry considerations the above probability is uniform across the non-support indices, i.e., $p_j^{(m)} = \frac{1}{d-K}$. To bound the probability that a non-support index $j \notin \mathcal{S}$ receives more than $\tilde{t}_c$ votes, we use the following Chernoff bound. For a Binomial random variable $X \sim B(M, p)$, the multiplicative Chernoff bound (Chernoff, 1952) implies that

$$\forall \delta > 0, \quad \Pr\left[X \geq (1+\delta) Mp\right] \leq \left(\frac{e^{-\delta}}{(1+\delta)^{1+\delta}}\right)^{Mp}.$$

In the case $1 + \delta \geq 2e$, a simple calculation shows that

$$\forall t \geq 2e \cdot Mp, \quad \Pr\left[X \geq t\right] \leq 2^{-t}.$$

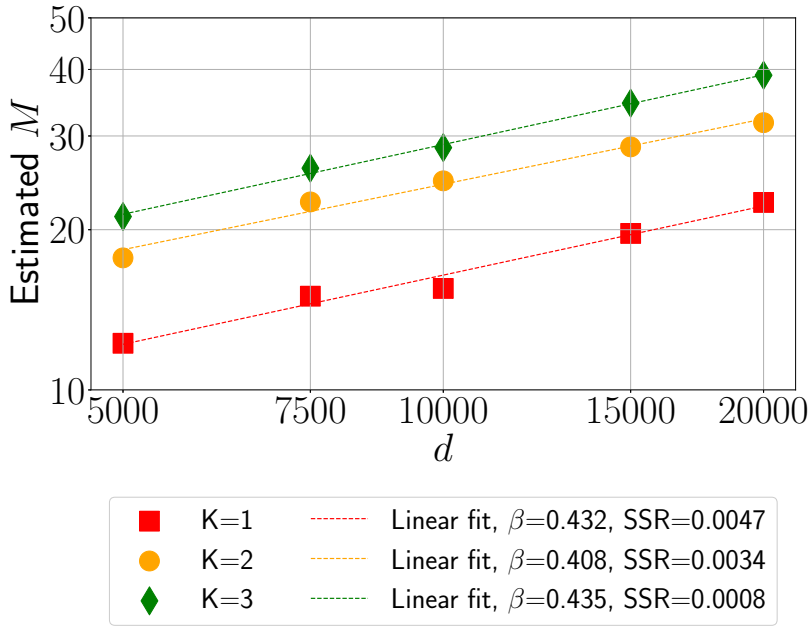



Figure 4: Number of Machines for Support Recovery by `DJ-OMP` vs. Dimension $d$

Since $p = \frac{1}{d-K}$, the assumption $M \leq 2e^{-1}(d-K)\log d$ implies that the above bound can be applied and hence the probability that a non-support index $j \notin \mathcal{S}$ receives more than $\tilde{t}_c$ votes can be bounded by

$$\Pr\left[\mathbf{v}_j \geq \tilde{t}_c\right] \leq 2^{-4\log d} = d^{-4\log 2}.$$

A union bound over all $d-K$ non-support indices $j \notin \mathcal{S}$ implies that the probability that the maximal number of votes for a non-support element is larger than $\tilde{t}_c$ is bounded by

$$\Pr\left[\max_{j\notin\mathcal{S}} \mathbf{v}_j \geq \tilde{t}_c\right] \leq (d-K)d^{-4\log 2} < d^{-1}.$$

Finally, a union bound concludes the proof. □

## C ADDITIONAL SIMULATION RESULTS

Theorem 4.1 holds under the max-MIP condition (6) and assumptions 4.1-4.3. However, in practice, `DJ-OMP` succeeds even if these assumptions are not met. For example, the max-MIP condition does not hold in the setting used in Figure 1(b), and thus none of the additional assumptions hold either. To examine assumption 4.1 further, we performed the following simulation, whose results are depicted in Figure 4. As described in Section 5, we generated matrices with i.i.d. Gaussian entries, i.e., $\alpha = 0$, with a fixed number of samples $n = 2000$, varying dimension $d$, varying number of machines $M$, and varying sparsity level $K$. In each simulation, the noise level is $\sigma = 1$, and each of the $K$ nonzero values of the sparse vector $\boldsymbol{\theta}$ equals $\theta_{\min} = 0.06$. We then used linear extrapolation to estimate for each dimension the number of machines needed to reach a given success probability, in our example 0.5, and displayed them on a logarithmic scale. In addition, we display a least-squares-based linear estimation of the relation between $\log(M)$ and $\log(d)$. The small resulting sum of squared residuals (SSR) support our result that the relationship is of the form $M = O(d^{\beta})$ for some $0 < \beta < 1$, even when the max-MIP condition does not hold, and in fact $\beta$ is empirically smaller than the exponent derived in Eq. (13). In addition, the estimated number of machines increases with $K$, which is also in accordance with Eq. (13). We obtained similar results when the matrices were slightly correlated, with slightly higher estimated number of machines.

# D IMPLEMENTATION DETAILS

The code used to generate the simulations in Section 5 was implemented in Python and was executed on an internal cluster (v3.8; Python Core Team, 2019, PSF licensed). For `SIS`-based methods, we used the SIS package by Saldana and Feng (2018), which was implemented using R statistical software (v4.0.3; R Core Team, 2023) and embedded into the Python code using the rpy2 package (https://rpy2.github.io/), all licensed by GPL-2 licenses. `Lasso`-based methods were implemented using the scikit-learn package by Pedregosa et al. (2011, BSD License). Other libraries that were used include NumPy (Harris et al., 2020, liberal BSD license), SciPy (Virtanen et al., 2020, BSD license), and Matplotlib (Hunter, 2007, BSD compatible license).