Sparse Estimation under Noise: Algorithms and Guarantees

Chen Amiraz

Thesis for the degree: Doctor of Philosophy

Submitted to the Scientific Council of the
Weizmann Institute of Science, Rehovot, Israel

February 2024

Advisors:

Prof. Robert Krauthgamer
Prof. Boaz Nadler

Acknowledgments

As I reflect on my doctoral journey, I am humbled by the immense support I received. First and foremost, my deepest gratitude goes to my advisors, Prof. Robert Krauthgamer and Prof. Boaz Nadler, for their mentorship, advice, and patience. In addition to teaching me how to approach problems and convey my ideas, they have always encouraged me to find my own path. Their guidance during both challenging and joyous times has been invaluable.

I am also deeply grateful to my Ph.D. committee members, Prof. Uri Feige and Prof. Yaron Lipman, for their thoughtful insights and feedback.

A special acknowledgment goes to the administrative and support staff, whose efforts have been fundamental to my daily academic pursuits.

The Weizmann Institute of Science has been more than an academic home to me. The culture of perpetual learning, the informal discussions by the whiteboards on the grass, and the approachable nature of world-renowned researchers is what makes our department so unique. I am grateful to have been a part of this community of scholars and friends.

Finally, this thesis is lovingly dedicated to my family: to my late father, Ami z"l, and my mother, Shula, whose perseverance brought me into this world and got me to where I am today; to my brother, Orr, and his wife, Rinat, for their unwavering support; to Gili and Itay, for their continued generosity and assistance; to my husband, Omri, for his wholehearted belief in me throughout this journey; and to my children, Roni and Ofir, who will always be my favorite Ph.D. results.

Abstract

Modern data science is centered around learning from large-scale, high-dimensional, and noisy datasets. A key strategy for addressing this complexity involves identifying a simple model that captures the important features of the data, such as assuming that the parameter of interest is a *sparse* vector, meaning it has only a few non-zero values. Over the past few decades, sparse models have been used in various domains, ranging from signal and image processing to genomics analysis. However, even in the absence of noise, recovering the sparsity pattern of vectors (also known as their support set) is in general computationally challenging, and becomes even more difficult in extremely noisy settings.

An additional source of complexity is that the process of data collection and analysis is becoming more and more decentralized. From mobile phones, through autonomous vehicles, to geographically spread-out data centers – modern distributed networks gather and store enormous amounts of data. This shift poses new challenges for designing learning algorithms, that must process data from distributed sources while carefully navigating constraints related to communication, computation, and privacy.

In this dissertation, we study three sparse estimation problems in both centralized and distributed settings. In each problem, the goal is to recover a sparse signal, or its support set, based on measurements from a linear model corrupted by Gaussian (or sub-Gaussian) noise. We analyze sparse recovery algorithms, derive novel guarantees for their success based on the signal-to-noise ratio (SNR), and substantiate our results by simulation studies.

We first consider sparse linear regression in the presence of Gaussian noise. Orthogonal Matching pursuit (`OMP`) is a popular greedy algorithm for this problem, and we improve upon previous works studying this algorithm in two ways. First, we derive a slightly sharper sufficient condition for exact support recovery by `OMP` with high probability, and second, we show both theoretically and empirically that this sharper sufficient condition is tight. Thus, we uncover an SNR regime in which `OMP` succeeds.

Next, we focus on distributed settings with communication constraints. We study distributed sparse normal means estimation, where noisy observations of a high-dimensional sparse vector are distributed across multiple machines, with limited communication to a fusion center. Previous works showed that to achieve the centralized minimax rate for the ℓ_2 risk, the total communication must be high – at least linear in the dimension. This phenomenon occurs, however, at very weak signals. We show that at SNRs that are sufficiently high – but not enough for recovery by any individual machine – the support of the sparse vector can be correctly recovered with communication sublinear in its dimension. Furthermore, the communication decreases exponentially as a function of signal strength.

Finally, we leverage our insights from the first two studies and consider distributed `OMP`-based schemes for high-dimensional sparse linear regression. Such schemes are particularly suited for settings where end machines have both limited communication and computational resources. We prove that under suitable assumptions, distributed-`OMP` schemes recover the support of a sparse regression vector with communication per machine that is linear in its sparsity and logarithmic in the dimension. Again, this holds even at low SNRs, where individual machines are unable to recover the support on their own. Furthermore, our simulations show that distributed-`OMP` schemes are competitive with more computationally intensive methods, and in some cases even outperform them.

Collectively, the results in this thesis advance our statistical and algorithmic understanding of sparse estimation under noise in both centralized and distributed settings.

Declaration

I hereby declare that this thesis summarizes independent research I conducted as part of my doctoral studies. It is a comprehensive presentation of the findings derived in three research projects, performed in collaboration with my advisors Prof. Robert Krauthgamer and Prof. Boaz Nadler, who are the co-authors of the publications arising from this work.

Specifically, the results presented in Chapter 2 were published in the paper Amiraz et al. (2021), those in Chapter 3 in Amiraz et al. (2022), and the findings discussed in Chapter 4 have been accepted for publication as Amiraz et al. (2024).

Contents

Chapter 1

Introduction

A key challenge in statistical machine learning is to efficiently distill complex, high-dimensional data into understandable and manageable forms. One way of achieving this objective is through assuming that a simple sparse model underlies the data. Indeed, sparsity is a natural modeling assumption in high-dimensional settings, widely adopted across numerous studies over the past few decades (Foucart and Rauhut, 2013; Tibshirani et al., 2015).

Learning and estimation under sparsity assumptions has been applied in various fields. In fact, many natural signals such as acoustics, images, or video, have a sparse (or approximately sparse) representation in some appropriate basis, as demonstrated by transform-domain methods and technologies like JPEG compression (Wallace, 1991; Mallat, 1999). Moreover, redundant over-complete dictionaries enable even sparser representations of natural signals (Elad, 2010). This leads to a sparse linear regression problem, as discussed in Section 1.1.

Furthermore, in modern applications such as spectrum sensing and wireless sensor networks, machines constrained by limited computational resources and communication bandwidth are tasked with jointly estimating sparse quantities (Bazerque and Giannakis, 2010; Sharma et al., 2016; Ling and Tian, 2010). This requires distributed algorithms that balance accuracy, communication costs, computational efficiency, and privacy, as discussed in Section 1.2.

Despite the broad applications and theoretical foundations of sparsity, there are still open questions regarding the effect of the signal-to-noise ratio (SNR) on the success of sparse estimation algorithms. Such guarantees are essential to understanding the limits of existing algorithms and for guiding the development of new methodologies. In this thesis, we advance the theoretical foundations of sparse estimation by deriving SNR-based recovery guarantees for both centralized and distributed algorithms. Our contributions are summarized in Sections 1.1 and 1.2 below. We then propose in Section 1.3 several directions for future work.

1.1 Sparse Linear Regression

A fundamental inverse problem arising in machine learning, statistics and signal processing is to estimate an unknown sparse vector $\boldsymbol{\theta} \in \mathbb{R}^d$ from N linear measurements of it, often with $N < d$. Notable examples in signal processing include sparse recovery in a redundant representation and compressed sensing (Elad, 2010; Eldar and Kutyniok, 2012; Foucart and Rauhut, 2013). A notable example in statistics is linear regression with a sparse coefficient vector, in particular when there are more variables than observations (Tibshirani et al., 2015). 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 (Miller, 2002).

Assuming that the measurements are corrupted by additive Gaussian noise, the observed signal

$\mathbf{y} \in \mathbb{R}^N$ has the following form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \sigma\xi \tag{1.1}$$

where $\mathbf{X} \in \mathbb{R}^{N\times d}$ is a known matrix, $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown sparse vector, $\xi \in \mathbb{R}^N$ is a random Gaussian noise vector $\xi \sim N(\mathbf{0}, \mathbf{I}_N)$ and $\sigma > 0$ is the noise level. We say that $\boldsymbol{\theta}$ is *K-sparse* if $\|\boldsymbol{\theta}\|_0 = |\mathrm{supp}(\boldsymbol{\theta})| = K$ and denote its support by $\mathcal{S} = \mathrm{supp}(\boldsymbol{\theta})$. In statistics $\mathbf{X}$ is referred to as the design matrix, whereas in the signal processing literature it is often called the dictionary.

In sparse recovery, given the observed signal or response vector $\mathbf{y}$, the design matrix or dictionary $\mathbf{X}$ and the sparsity level K, the goal is to output an estimate $\hat{\boldsymbol{\theta}}$ that is close to the unknown vector of regression coefficients $\boldsymbol{\theta}$. Under the assumption that ξ is Gaussian and independent of $\boldsymbol{\theta}$, the maximum likelihood solution is

$$\hat{\boldsymbol{\theta}} = \arg\min \left\{ \|\mathbf{y} - \mathbf{X}\mathbf{z}\|_2 : \|\mathbf{z}\|_0 \le K \right\}. \tag{1.2}$$

In the noiseless case $\sigma = 0$, minimizing (1.2) is equivalent to finding a K-sparse vector $\hat{\boldsymbol{\theta}}$ such that $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\theta}}$. For $N < d$ this linear system is underdetermined and may have multiple solutions. Hence, for any $\sigma \ge 0$, Eq. (1.2) may in general also have multiple solutions. In certain regimes there exists a unique solution, for example when K is small compared to the size of the smallest linearly-dependent subset of design matrix columns (Donoho and Elad, 2003). Furthermore, even if a unique solution exists, finding it is in general NP-hard because the sparsity constraint is non-convex (Davis et al., 1997). In recent decades, several polynomial-time methods were developed for estimating $\hat{\boldsymbol{\theta}}$. Convex optimization-based methods such as Basis Pursuit use a relaxation of the l_0-norm of $\boldsymbol{\theta}$ to its l_1-norm (Tibshirani, 1996; Chen et al., 2001). Other recovery methods use non-convex penalty functions that promote sparsity (Figueiredo et al., 2007; Blumensath and Davies, 2008; Chartrand and Yin, 2008; Daubechies et al., 2010; Amir et al., 2021), or Bayesian approaches (Tipping, 2001; Wipf and Rao, 2004; Ji et al., 2008). Greedy methods estimate $\boldsymbol{\theta}$ by iteratively selecting columns that have high correlation with the residual part of the signal (Dai and Milenkovic, 2009; Needell and Tropp, 2009; Needell and Vershynin, 2010). For a review of sparse recovery algorithms, see (Marques et al., 2019) and the references therein.

In our first study, we focus on Orthogonal Matching Pursuit (`OMP`), which is one of the simplest and fastest greedy methods for sparse recovery (Chen et al., 1989; Pati et al., 1993; Mallat and Zhang, 1993). One key challenge in computing an estimate $\hat{\boldsymbol{\theta}}$ close to $\boldsymbol{\theta}$ is to accurately estimate its support. Hence, several authors studied conditions under which `OMP` exactly recovers the support of $\boldsymbol{\theta}$.

Several conditions for exact support recovery by `OMP` and by other methods have been studied. These include the Restricted Isometry Property (RIP) (Candes and Tao, 2005), the Exact Recovery Condition (ERC) (Tropp, 2004) and the Mutual Incoherence Property (MIP) (Donoho and Huo, 2001). For RIP and ERC based guarantees, see (Cai et al., 2018; Hashemi and Vikalo, 2016) and the references therein. While MIP is more restrictive than the other conditions, it is simple and tractable to compute for arbitrary dictionaries. We thus restrict our attention to coherence-based guarantees. Specifically, the coherence of the matrix $\mathbf{X}$ is defined as

$$\mu = \mu(\mathbf{X}) = \max_{i\neq j} \frac{|\langle \mathbf{x}_i, \mathbf{x}_j \rangle|}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}, \tag{1.3}$$

where $\mathbf{x}_i$ it the i-th column of $\mathbf{X}$. The sparsity K of a sparse vector $\boldsymbol{\theta}$ is said to satisfy the Mutual Incoherence Property (MIP) if

$$\mu < \frac{1}{2K-1}. \tag{1.4}$$

A fundamental result by Tropp (2004) is that the MIP condition is sufficient for exact support recovery by `OMP` in the noiseless case. Cai et al. (2010) proved that the MIP condition is sharp in the following setting: for each pair of positive integers (n, K), there exist a dictionary of size $2Kn \times (2K-1)n$ with coherence $\mu = \frac{1}{2K-1}$ and a K-sparse vector such that `OMP` fails to recover its support.

In the presence of additive Gaussian noise with noise level $\sigma > 0$, even if a K-sparse vector $\boldsymbol{\theta}$ satisfies the MIP condition (1.4), its exact support recovery will depend on the specific noise realization in the

observed signal $\mathbf{y}$. Hence, exact support recovery can only be guaranteed with a success probability $P_{\text{succ}} < 1$, which in general depends on the noise level σ, the sparsity level K, the magnitude of the non-zero coefficients of $\boldsymbol{\theta}$, the matrix dimensions N and d and the coherence μ. As we review in Chapter 2, Ben-Haim et al. (2010) developed a sufficient condition for OMP to recover the support of $\boldsymbol{\theta}$ in the presence of additive Gaussian noise with high probability. A similar result for a variant of OMP was proved by Cai and Wang (2011). Miandji et al. (2017) derive a similar sufficient condition in a different model where the nonzero elements of $\boldsymbol{\theta}$ are random variables. An interesting question is whether this sufficient condition is sharp, or can it be lowered further.

Our Contributions. Towards answering this question, we make two key contributions, presented in Chapter 2 and published in Amiraz et al. (2021). First, in Theorem 2.2 we derive a sharper sufficient condition for exact support recovery by OMP than that of Ben-Haim et al. (2010) and Cai and Wang (2011) by performing a tighter analysis of their proof. Our main result, stated formally in Theorem 2.3, shows that our sharper sufficient condition is tight, in the following sense. For a wide range of sparsity levels K, dictionary dimensions N, d and coherence values μ, there exist a dictionary $\mathbf{X}$ and a vector $\boldsymbol{\theta}$ with a signal-to-noise ratio that is slightly lower than that of our sufficient condition, for which with high probability OMP fails to recover its support. More specifically, we prove the following result.

Theorem 1.1 (Informal statement of Theorems 2.2 and 2.3.)**.** *Let* $\theta_{\min} = \min_{i \in \mathcal{S}} |\theta_i|$. *Under the MIP condition* (1.4)*, around*

$$\theta_{_{crit}} = \frac{\sigma\sqrt{2\log d}}{1-(2K-1)\mu(\mathbf{X})}$$

the following transition in the behavior of OMP *occurs:*

- *If the SNR is slightly higher, namely* $\theta_{\min} \gtrsim \theta_{_{crit}}\left(1+\sqrt{\frac{\log K}{\log d}}\right)$*, then w.h.p.* OMP *exactly recovers* $\mathcal{S}$*.*
- *If the SNR is slightly lower, namely* $\theta_{\min} \lesssim \theta_{_{crit}}\left(1-\sqrt{\frac{\log K}{\log d}}-\mu(\mathbf{X})\right)$*, then there exist* $\mathbf{X}$ *and* K*-sparse* $\boldsymbol{\theta}$ *for which w.h.p.* OMP *fails to recover* $\mathcal{S}$*.*

In Section 2.2 we present several simulations that support our theoretical analysis and show that this phase transition occurs for many common families of matrices $\mathbf{X}$ and sparse vectors $\boldsymbol{\theta}$. Thus, we identify the SNR regime in which OMP succeeds.

1.2 Distributed Statistical Estimation

In many modern applications, the collected datasets are too large to be stored or analyzed on a single machine. Moreover, due to communication or privacy constraints, the data at each machine cannot be sent to other machines in the network. This has sparked the development of distributed approaches for machine learning, statistical analysis, and data mining. A few examples of this vast body of work are McDonald et al. (2009); Bekkerman et al. (2011); Duchi et al. (2012); Guha et al. (2012). For a recent survey see Wimalajeewa and Varshney (2019).

In this thesis, we consider a popular distributed setting where M machines, each holding an independent set of n samples from some unknown distribution, are connected in a star topology to a central node, also called a fusion center or simply the center. The task of the fusion center is to estimate $\boldsymbol{\theta}$, a parameter of the distribution, using little communication with the M machines. In each round of communication, the fusion center may send a message to the machines (or a subset of them). Then, each contacted machine performs a local computation and sends its result back to the center. Finally, after the last communication round, the fusion center forms a global estimator $\hat{\boldsymbol{\theta}}$ based on these messages.

An interesting special case of this distributed setting is known as one-shot, embarrassingly parallel or split-and-merge schemes, where there is only a single round of two-way communication. A clear advantage of such one-shot schemes is their simplicity and ease of implementation.

Statistical inference in a distributed setting, in particular under communication constraints, raises several fundamental theoretical and practical questions. One question is what is the loss in statistical accuracy incurred by the scheme being distributed, i.e., compared to a centralized setting, whereby a single machine has access to all of the samples. Various works proposed multi-round communication-efficient schemes and analyzed their accuracy, see for example (Shamir et al., 2014; Zhang and Lin, 2015; Wang et al., 2017; Jordan et al., 2019). In the context of one-shot schemes, several works analyzed the case where the fusion center simply averages the estimators computed by the individual machines or for robustness, takes their median (Zhang et al., 2013b; Rosenblatt and Nadler, 2016; Minsker et al., 2019). In a high dimensional setting where the parameter of interest is a-priori known to be sparse, Lee et al. (2017) and Battey et al. (2018) considered a variant where the averaged estimator is further thresholded at the fusion center. A key finding in many of these papers is that in various scenarios and under suitable regularity assumptions, the ℓ_2 risk of the distributed estimate attains the same convergence rate as the centralized one, provided that the data is not split across too many machines.

Another important theoretical aspect in distributed learning is fundamental lower bounds on the achievable accuracy under communication as well as memory constraints, regardless of any specific inference scheme, see e.g. (Zhang et al., 2013a; Garg et al., 2014; Steinhardt et al., 2016; Cai and Wei, 2020; Zhu and Lafferty, 2018; Szabo and van Zanten, 2020; Acharya et al., 2020b), and similarly for the closely related problem of distributed detection (Acharya et al., 2020a; Szabo et al., 2020). Lower bounds on the estimation accuracy were also studied for problems involving a sparse quantity, including sparse linear regression, correlation detection and more (Steinhardt and Duchi, 2015; Braverman et al., 2016; Dagan and Shamir, 2018; Han et al., 2018). A central finding in these works is that to achieve the centralized minimax rate for the ℓ_2 risk, the communication must scale at least linearly in the ambient dimension.

However, when the task is to estimate a sparse quantity, then intuitively the communication should increase linearly with its sparsity level, and logarithmically with the ambient dimension, the latter being necessary for encoding the indices. Indeed, in the context of supervised learning, Acharya et al. (2019) showed that in various linear models with a sparse vector, optimal prediction error rates are achievable with total communication logarithmic in the dimension. However, they consider connectivity topology of a chain where each machine i sends a message only to machine $i+1$, and thus their algorithm is sequential and not compatible with one-shot inference schemes. An interesting question is the following: can problems that involve a sparsity prior admit one-shot algorithms with communication that is *sublinear* in the ambient dimension?

1.2.1 Distributed Sparse Normal Mean Estimation

In our second study, we consider sparse normal mean estimation, which is one of the simplest and most well-studied inference problems with sparsity priors, but in a distributed setting of M machines connected in a star topology to a fusion center. For simplicity we assume that each machine has the same number n of i.i.d. samples of the form

$$\mathbf{s}_i = \boldsymbol{\theta} + \sigma \boldsymbol{\xi}_i,$$

where the mean vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is exactly K-sparse, the noise is Gaussian, $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and $\sigma > 0$ is the noise level. Note that this problem can be viewed as a special case of distributed sparse linear regression where the design matrices are orthogonal.

We consider a one-shot communication scheme where the fusion center sends a setup message to each of the machines (or a subset of them), and then each contacted machine sends back its message to the center. We emphasize that in our setting the machines communicate only with the center and not with each other. Note that if the machines have prior knowledge of all problem parameters, then setup messages are not required. However, in any case the communication of this setup stage is often

negligible. The goal of the center is to recover the support of $\boldsymbol{\theta}$ under the constraint that the total communication between the fusion center and the machines (including the setup stage) is bounded by a budget of $B \ll d$ bits. As we discuss in Section 3.3, if $KM \ll \frac{d}{\log d}$ then achieving this goal implies that the vector $\boldsymbol{\theta}$ itself can be estimated with small ℓ_2 risk using communication sublinear in d.

For this sparse normal means problem, Braverman et al. (2016) and Han et al. (2018) derived communication lower bounds for the ℓ_2 risk of any estimator, and proved that to achieve the minimax rate, the total communication must be at least $\Omega(d)$. Shamir (2014) derived lower bounds for several other distributed problems involving M machines, each allowed to send a message of length at most b bits. His work implies that there exist d-dimensional distributions whose mean is a 1-sparse vector of sufficiently large magnitude, such that with $n = O(d \log d)$ samples per machine, any scheme with communication sublinear in d has only an $o(1)$ probability of exact support recovery. These works paint a pessimistic view, that to achieve the performance of the centralized solution, distributed inference must incur high communication costs.

Our contributions. Our main contribution, presented in Chapter 3 and published in Amiraz et al. (2022), is that in contrast to the previous works presented above, at SNRs that are sufficiently high, but not high enough for recovery by any individual machine, the support of $\boldsymbol{\theta}$ can be exactly recovered with total communication *sublinear* in the dimension d. Specifically, we present and analyze the performance of two distributed schemes. In the first scheme, which we denote `Top-L`, the center sends a parameter L to each contacted machine, which then sends back the indices of its top L coordinates in terms of magnitude. In the second scheme, which we denote `Thresholding`, the center sends a suitably chosen threshold t_m to each contacted machine, which then sends back the indices of coordinates that exceed the threshold. Our analysis is non-asymptotic, but the setting we have in mind is of a sparse vector in high dimension, namely $d \gg 1$ and $K \ll d$. Assuming that $\theta_{\min} = \min_{i \in \mathcal{S}} |\theta_i|$ is known to the center, we define the SNR scaling r such that $\theta_{\min} = \sigma\sqrt{\frac{2r \log d}{n}}$ and prove the following results.

Theorem 1.2 (Informal statement of Theorem 3.2.A.)**.** *Assume that $M \geq \Omega(\log d)$, that $\Omega\left(\frac{1}{\log d}\right) < r < 1$, and that d is sufficiently large. Then, with high probability, `Thresholding` with $t_m = \sqrt{2r \log d}$ recovers the support of the K-sparse vector $\boldsymbol{\theta}$ with $\tilde{O}\left(K + d^{1-r}\right)$ expected total communication bits.*

Note that since $r < 1$, the total communication is sublinear in d. As we show next, when the number of machines M is higher, the total communication can be lowered even further.

Theorem 1.3 (Informal statement of Theorem 3.2.B.)**.** *Assume that $M > \Omega\left(\log^{1.5} d\right)$, that $\Omega\left(\frac{\log M}{\log d}\right) < r < 1$, and that d is sufficiently large. Then, with high probability, `Thresholding` with an appropriately chosen threshold t_m recovers the support of the K-sparse vector $\boldsymbol{\theta}$ with expected communication of $\tilde{O}\left(KM + d^{1-r} \cdot e^{-f(M,r,d)}\right)$ bits, for an appropriately chosen function $f(M, r, d)$ that increases in all of its parameters.*

The function $f(M, r, d)$ increases in r in a way that the communication cost decreases *exponentially* as r increases towards 1, at which point the support of $\boldsymbol{\theta}$ may be found by a single machine using $O(K \log d)$ communication bits. In addition, Theorem 3.2.B implies the following counter-intuitive behavior of our `Thresholding` scheme: *more machines enable less communication.* Specifically, as discussed after Theorem 3.2.B, for some range of the problem parameters, as the number of machines is increased, exact support recovery is possible with *less* total communication. The reason is that a larger number of machines allows the center to set a higher threshold, which in turn implies that on average, each machine sends fewer indices back to the center. As we show, with a suitably chosen threshold, the total communication decreases, while the probability of exact recovery still remains close to one. In addition, if the number of machines M is sufficiently high, then the total communication cost can be further optimized by letting the center contact a subset of the machines, as presented in the next theorem.

Theorem 1.4 (Informal statement of Theorem 3.2.C.). *Assume that* $\Omega\left(\frac{1}{\log^2 d}\right) < r < 1$ *and let* $M_c = \tilde{O}\left(d^{\left(1-\sqrt{r}\right)^2}\right)$. *Further assume that* $M \geq M_c$ *and that* d *is sufficiently large. Then, with high probability, by contacting* M_c *machines,* `Thresholding` *with* $t_m = \sqrt{2\log d}$ *recovers the support of the* K*-sparse vector* $\boldsymbol{\theta}$ *with* $\tilde{O}\left(Kd^{\left(1-\sqrt{r}\right)^2}\right)$ *expected total communication bits.*

In Theorem 3.1 we prove a similar result for `Top-L` with $L = K$. In the reminder of Chapter 3 we further extend some of these results to the case of sub-Gaussian additive noise. Finally, we prove that if $KM \ll \frac{d}{\log d}$, then an additional single round of communication, also sublinear in d, results in an estimator for $\boldsymbol{\theta}$ that achieves the centralized rate for the ℓ_2 risk.

This idealized setting allows for a relatively simple analysis that showcases a tradeoff between the number of machines, SNR, and communication. Our work raises several issues for future research. First, it remains an open problem whether the SNR-communication tradeoff of our algorithms is optimal. Indeed, the derivation of tight SNR-dependent communication lower bounds for the sparse normal means problem is an interesting research topic. Second, we focus on the simple case where all machines have the same number of samples n and all samples have the same noise level σ. Another interesting research direction is to consider a more general setting where each machine i has a different number of samples n_i, or a different noise level σ_i. An additional interesting setting is where each machine observes different sparse vectors $\boldsymbol{\theta}_i$ with the same support $\mathcal{S}$ (or very similar supports $\mathcal{S}_i$). Note that there is no single SNR parameter in these cases since different machines have different effective SNRs. Lastly, building on the insights gained in this simple setting, we believe a similar behavior should hold for other popular statistical learning problems involving estimation of a sparse quantity in a high dimensional setting.

1.2.2 Distributed Sparse Linear Regression

In our third study, we revisited the sparse linear regression model with Gaussian noise, but in a distributed setting where M machines are connected in a star topology to a fusion center. Each machine $m \in [M]$ has for simplicity a vector of observations $\mathbf{y}^{(m)} \in \mathbb{R}^n$ of size $n = N/M$ and a design matrix $\mathbf{X}^{(m)} \in \mathbb{R}^{n\times d}$, related via

$$\mathbf{y}^{(m)} = \mathbf{X}^{(m)}\boldsymbol{\theta} + \sigma\boldsymbol{\xi}^{(m)}, \tag{1.5}$$

where the vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is the unknown K-sparse vector of regression coefficients, $\boldsymbol{\xi}^{(m)} \in \mathbb{R}^n$ is a Gaussian noise vector, i.e., $\boldsymbol{\xi}^{(m)} \sim \mathcal{N}(0, \mathbf{I}_n)$, independent across machines, and $\sigma > 0$ is the noise level. Recall that in a centralized setting, as described in (1.1), all entries of the response vector and design matrix are provided as input, and common tasks are to accurately estimate $\boldsymbol{\theta}$ as well as its support $\mathcal{S} = \mathrm{supp}(\boldsymbol{\theta}) = \{i \mid \theta_i \neq 0\}$.

Some distributed methods attempt to recover the centralized solution that would have been computed by the fusion center, if it had access to all $N = nM$ observations 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 single round divide-and-conquer scheme is for each machine to send its own dense estimate of $\boldsymbol{\theta}$, each of size d, to the fusion center, which then averages these estimates. For a wide range of problems, the resulting estimator 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 (1.5), 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.

In Chapter 4 we consider distributed estimation of the sparse vector $\boldsymbol{\theta}$ in the model (1.5), 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 this setting, 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). 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, where each vector $\mathbf{x}_i$ at each machine was independently sampled at random, and their proofs rely heavily on the symmetry between all non-support variables. Thus, they are inapplicable in our framework.

Our Contributions. In Chapter 4 we analyze several distributed-`OMP` schemes that estimate the support of $\boldsymbol{\theta}$. The first scheme, `D-OMP`, is a one-shot scheme, while the second scheme, `DJ-OMP`, requires K rounds of communication between the machines and the fusion center. Our results detailed below have been accepted for publication in the Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS) Amiraz et al. (2024).

Our key contribution is the derivation of the following recovery guarantee for `DJ-OMP`.

Theorem 1.5 (Informal statement of Theorem 4.1.)**.** *Let* $\theta_{\min} = \min_{i\in\mathcal{S}} |\theta_i|$*, and define the SNR scaling* r *such that* $\theta_{\min} = \frac{\sigma\sqrt{2r\log d}}{1-(2K-1)\mu_{\max}}$*, where* $\mu_{\max} = \mu_{\max}\left(\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(M)}\right) = \max_{m\in[M]}\mu(\mathbf{X}^{(m)})$*. Suppose that the MIP condition* (1.4) *holds for each of the* M *matrices* $\mathbf{X}^{(m)}$*, that* $r \gtrsim \mu_{\max}$*, that* $M \geq \tilde{\Omega}\left(Kd^{\frac{(1-\sqrt{r})^2}{(1-\mu_{\max})^3}}\right)$*, and that* d *is sufficiently large. Then, w.h.p.* `DJ-OMP` *with* K *communication rounds recovers the support of the* K*-sparse vector* $\boldsymbol{\theta}$*.*

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 matrices $\mathbf{X}^{(m)}$, 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.

To complement our theoretical analysis, we compare via simulations the support-recovery success of the distributed-OMP algorithms D-OMP and DJ-OMP with 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.

1.3 Directions for Future Research

In summary, this thesis advances the field of sparse estimation by establishing novel guarantees for recovering sparse vectors from linear measurements corrupted by Gaussian (and sub-Gaussian) noise in both centralized and distributed settings. For each of the three problems considered in this thesis, there are various questions for future research that are outlined in each chapter.

On a broader scale, we believe that our results could be extended to a wide array of sparse estimation problems. For instance, our methodologies could be adapted to distributed sparse logistic regression and other generalized linear models. Another intriguing direction for future research is the online or streaming setting, where the goal is to recover a support set shared between multiple measurement vectors that are observed sequentially, where the support set may vary over time. Thus, the insights from our work offer a foundation for developing algorithms and deriving SNR-based bounds for diverse sparse recovery problems.

Chapter 2

Tight Recovery Guarantees for OMP under Gaussian Noise[1]

Orthogonal Matching pursuit (OMP), described in Algorithm 1, is a popular algorithm to estimate an unknown sparse vector from multiple linear measurements of it. Assuming exact sparsity and that the measurements are corrupted by additive Gaussian noise, the success of OMP is often formulated as exactly recovering the support of the sparse vector. Several authors derived a sufficient condition for exact support recovery by OMP with high probability depending on the signal-to-noise ratio, defined as the magnitude of the smallest non-zero coefficient of the vector divided by the noise level. We make two contributions, presented in Section 2.1. First, we derive a slightly sharper sufficient condition for two variants of OMP, in which either the sparsity level or the noise level is known. Next, we show that this sharper sufficient condition is tight, in the following sense: for a wide range of problem parameters, there exist a dictionary of linear measurements and a sparse vector with a signal-to-noise ratio slightly below that of the sufficient condition, for which with high probability OMP fails to recover its support. In Section 2.2, we present simulations which illustrate that our condition is tight for a much broader range of dictionaries. All proofs appear in Section 2.3.

2.1 Main Results

We first introduce key notations and assumptions that will hold throughout the chapter. We denote $\theta_{\min} = \min_{i \in \mathcal{S}} |\boldsymbol{\theta}_i|$ and define the following effective noise factor

$$\sigma_{\text{eff}}(K, \mu) = \frac{\sigma}{1 - (2K - 1)\,\mu}.$$

We assume that the MIP condition (1.4) holds, so $\sigma_{\text{eff}}(K, \mu)$ is well defined and strictly positive. Moreover, we refer to the columns $\mathbf{x}_i$ of $\mathbf{X}$ as the *atoms* of the dictionary and assume for simplicity that they are normalized to have unit norm $\|\mathbf{x}_i\| = 1$.

For measurements that are corrupted by additive Gaussian noise, Ben-Haim et al. (2010) derived the following sufficient condition for OMP to recover the support of $\boldsymbol{\theta}$ with high probability.

Theorem 2.1 (Ben-Haim et al. (2010))**.** *Let $\boldsymbol{\theta}$ be an unknown vector with known sparsity K, and let $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \sigma\boldsymbol{\xi}$, where $\mathbf{X} \in \mathbb{R}^{N \times d}$ is a dictionary with normalized columns and coherence μ, and $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_N)$. Suppose that the MIP condition (1.4) holds and that for some $\alpha \geq 0$*

$$\theta_{\min} \geq 2\sigma_{\text{eff}}(K, \mu)\sqrt{2\,(1 + \alpha)\log d}. \tag{2.1}$$

[1] The results presented in this chapter have been published in Amiraz et al. (2021). The notations have been changed for consistency with the other chapters of the thesis.

Algorithm 1: Orthogonal Matching Pursuit (OMP)

input : dictionary $\mathbf{X} \in \mathbb{R}^{N \times d}$, signal $\mathbf{y} \in \mathbb{R}^N$, sparsity level K
output: estimated vector $\hat{\boldsymbol{\theta}}_K \in \mathbb{R}^d$

initialize the residual $\mathbf{r}_0 = \mathbf{y}$ and the estimated support $\hat{\mathcal{S}}_0 = \emptyset$
for $t = 1, \ldots, K$ **do**
  calculate $j = \arg\max \left\{ \frac{|\langle \mathbf{x}_i, \mathbf{r}_{t-1} \rangle|}{\|\mathbf{x}_i\|} : i \in [d] \right\}$
  add $\hat{\mathcal{S}}_t = \hat{\mathcal{S}}_{t-1} \cup \{j\}$
  calculate $\hat{\boldsymbol{\theta}}_t = \arg\min \left\{ \|\mathbf{y} - \mathbf{X}\mathbf{z}\|_2 : \mathbf{z} \in \mathbb{R}^d, \mathrm{supp}(\mathbf{z}) = \hat{\mathcal{S}}_t \right\}$
  update $\mathbf{r}_t = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}_t$
end

Then, OMP with K iterations successfully recovers the support of $\boldsymbol{\theta}$ with probability at least

$$1 - \frac{1}{d^{\alpha}\sqrt{\pi\,(1+\alpha)\log d}}. \tag{2.2}$$

In many practical cases K is unknown while the noise level σ is known. Denote by OMP* a variant of Algorithm 1 where instead of performing K iterations, the algorithm stops when the maximal correlation of the residual with any dictionary atom is smaller than a threshold τ, i.e., $\left\|\mathbf{X}^T \mathbf{r}_t\right\|_{\infty} \leq \tau$. Cai and Wang (2011, Thm. 8) proved the following analogue of Theorem 2.1 for this OMP* variant: under the MIP condition (1.4) and condition (2.1), OMP* with threshold $\tau = \sigma\sqrt{2(1+\alpha)\log d}$ recovers the support of $\boldsymbol{\theta}$ with probability at least $1 - K/d^{\alpha}\sqrt{2\log d}$.

2.1.1 Sharper sufficient condition

By performing a tighter analysis of the proofs of Ben-Haim et al. (2010) and Cai and Wang (2011), we derive a sharper sufficient condition than (2.1) for exact support recovery by both OMP and OMP*. However, this sharper sufficient condition comes at a price, whereby the success probability is a function not only of the vector length d, but also of its sparsity level K. The following theorem formalizes this statement and is proved in Section 2.3.1.

Theorem 2.2. *Let $\boldsymbol{\theta}$ be an unknown fixed vector with known sparsity K, and let $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \sigma\boldsymbol{\xi}$, where $\mathbf{X} \in \mathbb{R}^{N \times d}$ is a dictionary with normalized columns and coherence μ, and $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_N)$. Suppose that the MIP condition (1.4) holds, that $K \leq d^{\beta}$ for some $0 < \beta < 1$ and that for some $\alpha \geq 0$*

$$\theta_{\min} \geq \sigma_{\mathrm{eff}}(K, \mu)\left(1 + \sqrt{\beta}\right)\sqrt{2\,(1+\alpha)\log d}. \tag{2.3}$$

Then, OMP with K iterations successfully recovers the support of $\boldsymbol{\theta}$ with probability at least

$$1 - \frac{1}{\sqrt{\pi\,(1+\alpha)\log d}}\left(\frac{1}{d^{\alpha}} + \frac{1}{d^{\alpha\beta}\sqrt{\beta}}\right). \tag{2.4}$$

Moreover, under the same conditions OMP* with threshold $\tau = \sigma\sqrt{2(1+\alpha)\log d}$ successfully recovers the support of $\boldsymbol{\theta}$ with probability at least (2.4).

We emphasize that in both Theorem 2.1 and Theorem 2.2 the vector $\boldsymbol{\theta}$ is fixed and the probability of success is over the realizations of the noise vector $\boldsymbol{\xi}$. Eq. (2.3) is sharper than Eq. (2.1) since $\beta < 1$. Simulations in Section 2.2 illustrate the tightness of this result.

2.1.2 Near-tightness of the OMP recovery guarantee

According to either Eq. (2.2) or (2.4), the smallest α that still guarantees exact support recovery with probability tending to 1 as $d \to \infty$ is $\alpha = 0$. Therefore, the weakest sufficient condition for OMP to recover the exact support of $\boldsymbol{\theta}$ with high probability for $d \gg 1$ is

$$\theta_{\min} \geq \sigma_{\text{eff}}(K, \mu) \left(1 + \sqrt{\beta}\right) \sqrt{2 \log d}. \tag{2.5}$$

An interesting question is thus whether this sufficient condition is sharp, or could the right hand side in (2.5) be lowered further.

The main result of this chapter, formalized in Theorem 2.3 below, is that the above condition is quite tight. Informally, our result can be stated as follows: for a wide range of sparsity levels K, dictionary dimensions N, d and coherence values μ, there exist a dictionary $\mathbf{X} \in \mathbb{R}^{N \times d}$ and a K-sparse vector $\boldsymbol{\theta} \in \mathbb{R}^d$ with

$$\theta_{\min} \approx \sigma_{\text{eff}}(K, \mu) \left(1 - \mu - \sqrt{\beta}\right) \sqrt{2 \log d}, \tag{2.6}$$

for which OMP fails to recover its support with probability $1 - o(1)$. In particular, the failure probability for this specific $\mathbf{X}$ and $\boldsymbol{\theta}$ tends to 1 as $d \to \infty$. As shown by the simulations in Section 2.2, OMP fails with high probability under condition (2.6) in a much broader range of cases. These include a case where the dictionary atoms are drawn independently and uniformly at random from the unit sphere and a case where the dictionary is composed of two orthogonal matrices (the identity matrix and the Hadamard matrix with normalized columns).

If K is constant or polylogarithmic in d, then as $d \to \infty$ we can take $\beta > 0$ arbitrarily small. In this case, the bounds (2.5) and (2.6) match, up to a multiplicative factor of $1 - \mu$. Finally, for various dictionaries the coherence μ is itself small. For example, if each entry of the dictionary is drawn independently and uniformly at random from $\pm 1/\sqrt{N}$, then with probability exceeding $1 - \delta^2$ the coherence is $\mu \leq 2\sqrt{N^{-1} \log \frac{d}{\delta}}$ (Tropp and Gilbert, 2007). Hence, $\mu \to 0$ if d is sub-exponential in N.

To formally state our theorem, we introduce the following notations. First, let

$$\rho = \rho\,(K, \mu) = \sqrt{\frac{1 - (K-1)\mu}{K}} \tag{2.7}$$

and

$$\tilde{\mu} = \tilde{\mu}\,(K, \mu) = \frac{\mu^2}{\rho^2} = \frac{\mu^2 K}{1 - (K-1)\mu}. \tag{2.8}$$

Both quantities are well defined, since by the MIP condition (1.4), $1 - (K-1)\mu > 0$. It can be easily shown that $\sqrt{\mu} < \rho \leq 1/\sqrt{K}$ and $\tilde{\mu} < \mu$. Next, denote $\tilde{N} = N - K$ and $\tilde{d} = d - K$. Let $\mu_{\min}\,(a, b)$ be the smallest possible coherence of an $a \times b$ overcomplete dictionary with $a < b$. To prove our theorem we construct a dictionary that consists of several parts. One of these parts is a $\tilde{N} \times \tilde{d}$ dictionary with coherence $L = L\left(\tilde{N}, \tilde{d}\right) = \mu_{\min}\left(\tilde{N}, \tilde{d}\right)$. By the theory of Grassmannian frames,

$$L \geq \sqrt{\frac{\tilde{d} - \tilde{N}}{\tilde{N}(\tilde{d} - 1)}} \tag{2.9}$$

(see for example Strohmer and Heath, 2003). In fact, L may be strictly higher since Grassmannian frames do not exist for every pair $(\tilde{N}, \tilde{d})$. However it can not be much higher, since by Tropp and Gilbert (2007) $L \leq 2\sqrt{\tilde{N}^{-1} \log \tilde{d}}$.

We now give a rigorous statement of our result, whose proof appears in Section 2.3.2.

Theorem 2.3. *Let N, d be integers such that $N < d$. Let K be an integer and let μ be a number that satisfy the MIP condition* (1.4) *and the following set of inequalities:*

$$K \leq \min\left\{d^{\beta}, N\right\} \tag{2.10}$$

where $0 < \beta < 1$,

$$K \leq \frac{3 - L - \sqrt{8 - 8L}}{L}, \tag{2.11}$$

and

$$\frac{(1 + L(K-1))\left(1 - \sqrt{1 - \frac{4L(2K-1-LK)}{(1+L(K-1))^2}}\right)}{2(2K - 1 - LK)} \leq \mu \leq \frac{(1 + L(K-1))\left(1 + \sqrt{1 - \frac{4L(2K-1-LK)}{(1+L(K-1))^2}}\right)}{2(2K - 1 - LK)}. \tag{2.12}$$

Then, there exists a dictionary $\mathbf{X} \in \mathbb{R}^{N \times d}$ with coherence μ and a corresponding K-sparse vector $\boldsymbol{\theta} \in \mathbb{R}^d$ satisfying

$$\theta_{\min} = \sigma_{\text{eff}}(K, \mu) \cdot \left\{ \sqrt{2(1-\mu)(1-\tilde{\mu}) \log \tilde{d}} - \sqrt{2\beta(1-\rho^2) \log d} \right.$$
$$\left. - c_0 \sqrt{(1-\tilde{\mu}) \log\log \tilde{d}} - \left(\rho + \sqrt{\tilde{\mu}}\right) \sqrt{2 \log\log d} \right\} \tag{2.13}$$

where ρ is given in (2.7)*, $\tilde{\mu}$ is given in* (2.8) *and $c_0 > 0$ is a universal constant, such that with probability at least*

$$P_0 = 1 - 6e^{-C\sqrt{\log\log \tilde{d} \min\left\{\mu^{-1}, \log \tilde{d}\right\}}} - \left(\log d \sqrt{\pi \log\log d}\right)^{-1} - \left(\sqrt{\pi\beta \log d}\right)^{-1},$$

OMP fails to recover the support of $\boldsymbol{\theta}$ from $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \sigma\boldsymbol{\xi}$.

Remark 2.1. *Let us now illustrate that conditions* (2.11) *and* (2.12) *are not very restrictive. It is instructive to consider the over-complete case with $d = JN$ for $J > 1$, with $d, N \gg 1$ and sparsity K much smaller than N, such that $\tilde{N} \approx N$. By the theory of Grassmannian frames $L \approx \frac{C(J)}{\sqrt{N}}$ for an appropriate $C(J) > 0$ (Strohmer and Heath, 2003). Under the MIP condition* (1.4)*, $K \lesssim 0.5\frac{\sqrt{N}}{C(J)}$, while under condition* (2.11)*, $K \lesssim (3 - \sqrt{8})\frac{\sqrt{N}}{C(J)} \approx 0.17\frac{\sqrt{N}}{C(J)}$.*

For values of K such that LK is much smaller than 1, condition (2.12) *can be approximated by a binomial approximation $\sqrt{1-\varepsilon} \approx 1 - \varepsilon/2$ for small $\varepsilon = \frac{4L(2K-1-LK)}{(1+L(K-1))^2}$ as*

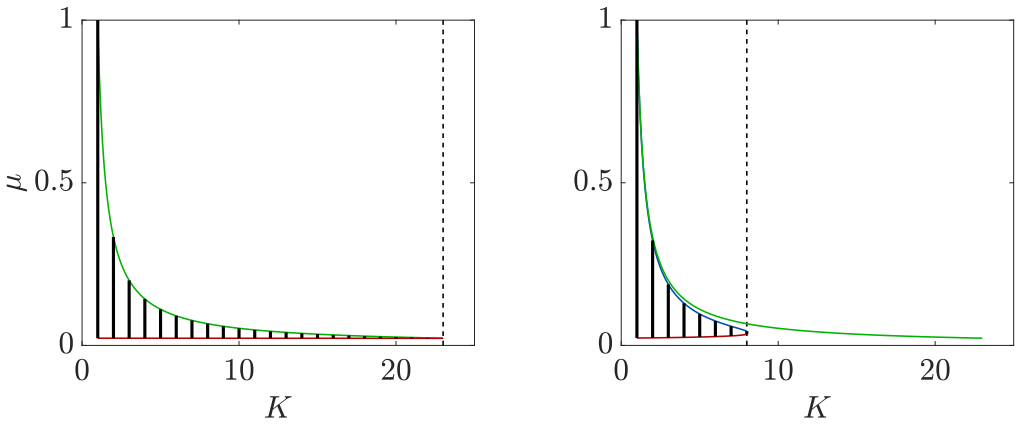
$$L \lesssim \mu \lesssim \frac{1}{2K-1} - L.$$

The inequality $L \lesssim \mu$ follows essentially from frame lower bounds whereas the other inequality is very close to the MIP condition (1.4)*. Hence, condition* (2.12) *is only slightly more restrictive than MIP. This comparison is visualized in Figure 2.1.*

Remark 2.2. *We now show how Eq.* (2.13) *may be approximated by Eq.* (2.6)*. First, for Theorem 2.3 to be meaningful, the right hand side of Eq.* (2.13) *must be positive. We now show that this is indeed the case for typical parameter values. If $K = d^{\beta}$ for $\beta < 1$, then $\log \tilde{d} = \log d + \log\left(1 - \frac{1}{d^{1-\beta}}\right) \approx \log d$. Recall that $\tilde{\mu} < \mu$ and that $\rho > 0$. Hence, the first two terms on the right hand side of* (2.13) *can be approximated as follows*

$$\begin{aligned} \sqrt{2(1-\mu)(1-\tilde{\mu}) \log \tilde{d}} - \sqrt{2\beta(1-\rho^2) \log d} &> (1-\mu)\sqrt{2 \log \tilde{d}} - \sqrt{2\beta \log d} \\ &\approx (1 - \mu - \sqrt{\beta})\sqrt{2 \log d}. \end{aligned}$$

In addition, the last two terms on the right hand side of equation (2.13) *are small compared to the first term, since they are of order $\sqrt{\log\log d}$. Hence,* (2.13) *may be approximated by* (2.6)*.*



(a) (K, μ) by the MIP condition (1.4)

(b) (K, μ) by condition (2.12)

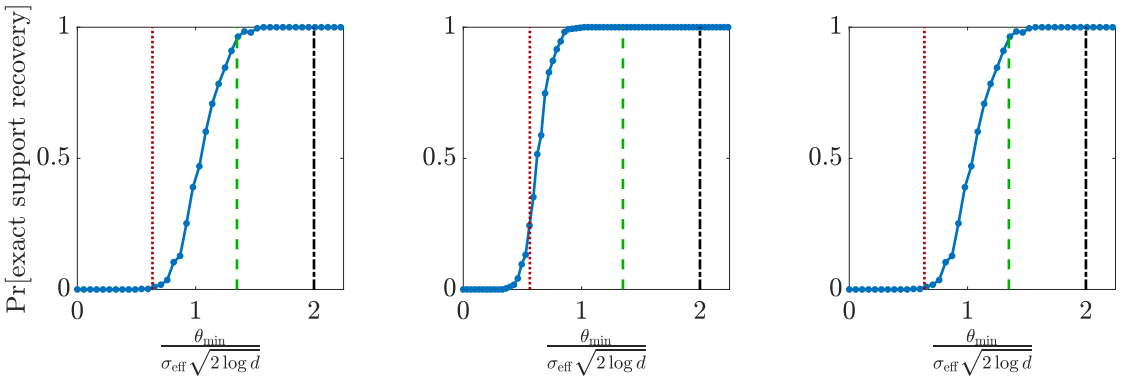
Figure 2.1: Comparison of various conditions on (K, μ) for a dictionary of size $\tilde{N} = 1020$ and $\tilde{d} = 2040$. In the left panel, the green curve is the MIP condition (1.4) and the red curve is the frame lower bound from Eq. (2.9). The vertical dashed line is the maximal sparsity K where the two curves intersect. The solid black lines between these two curves are values of (K, μ) for which if $\theta_{\min}$ is sufficiently large, then support recovery is guaranteed with high probability. In the right panel, the blue curve is condition (2.12) and the black lines are the values of (K, μ) where our Theorem 2.3 holds. Namely, if $\theta_{\min}$ is small then there are provable cases where `OMP` will fail. Comparing the two panels shows that Theorem 2.3 applies to most but not all values of (K, μ) where Theorems 2.1 and 2.2 hold.

2.2 Simulations

We present several simulations to illustrate our sharper sufficient condition in Theorem 2.2 and our near-tightness result in Theorem 2.3. We generated $N \times d$ dictionaries and K-sparse vectors with coefficients of equal magnitude ν. For each vector $\boldsymbol{\theta}$, we drew random noise $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_N)$ with noise level $\sigma = 1$ and computed the signal $\mathbf{y}$ as in Eq. (1.1).

In Setting 1, we considered the probability of exact support recovery of K-sparse vectors with sparsity $K = 3$ using three dictionaries of size $(N, d) = (4096, 8192)$. The first is a two-ortho dictionary $\mathbf{X}_1 = [\mathbf{I}\,\mathbf{H}]$ composed of two orthogonal matrices – the identity matrix and the Hadamard matrix with normalized columns. The second $\mathbf{X}_2$ is a dictionary whose atoms are drawn independently and uniformly at random from the unit sphere. For these two dictionaries the K-sparse vectors were drawn independently and uniformly at random from the $\binom{d}{K}$ possible vectors. The third dictionary $\mathbf{X}_3$ and its corresponding K-sparse vector are the ones used to construct the near-tightness example in the proof of Theorem 2.3 (see Eqs. (2.24) and (2.27)). Figure 2.2 depicts the empirical probability that `OMP` recovered the exact support of the unknown sparse vector in Setting 1, averaged over 500 realizations. It is interesting to note that our sufficient condition in Theorem 2.2 indeed improves over that of Ben-Haim et al. (2010). In addition, our sufficient condition is relatively sharp for small values of K. Another important observation is that even though condition (2.6) was derived considering $\mathbf{X}_3$ corresponding to the third panel, we see that the condition holds for different types of dictionaries as well.

In Setting 2, we study the probability of exact support recovery for different sparsity levels $K = 2, 4, 6$ for the specific dictionary and K-sparse vector used in the proof of Theorem 2.3. For each value of K, the corresponding dictionary of size $(N, d) = (4096, 8192)$ and coherence $\mu = 0.03$ was constructed using the alternating projection method of Tropp et al. (2005). Figure 2.3 depicts the empirical probability that `OMP` recovered the exact support of the unknown sparse vector in Setting 2, averaged over 500 realizations. Note that condition (2.6) is conservative since in our proof we analyze



(a) Two-ortho dictionary (b) Random dictionary (c) Our dictionary

Figure 2.2: The solid blue line in each panel is the empirical probability of exact support recovery of a sparse vector by OMP as a function of its normalized signal-to-noise ratio in Setting 1. The dash-dotted black line is the sufficient condition (2.1) by Ben-Haim et al. (2010). The dashed green line is the sharper sufficient condition (2.3). The dotted red line is the approximate condition (2.6) below which OMP fails with high probability by Theorem 2.3. In all panels, when fixing a high success probability (for example 0.95), our sharper sufficient condition (2.3) is indeed closer to the empirical probability than the sufficient condition (2.1) by Ben-Haim et al. (2010), and in the first panel it is in fact tight.

failure only in the first step of the algorithm. However, it cannot be increased much further, since the probability of recovery increases sharply at higher values of the normalized signal-to-noise ratio. Finally, we remark that similar results are obtained for other values of N, d, K and μ.

2.3 Proofs

The following auxiliary lemma will be useful in proving both Theorems 2.2 and 2.3. Its proof appears in Section 2.3.3.

Lemma 2.1. *Let* $(X_1, \ldots, X_{n_1}) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ *where* $\mathbf{\Sigma}_{ii} = 1$ *for all* $i \in [n_1]$. *For any* $\eta > 0$ *and* $n_2 \geq n_1$ *the following holds*

$$\Pr\left[\max_{i\in[n_1]} |X_i| < \sqrt{2\eta \log n_2}\right] \geq 1 - \frac{n_1}{n_2^{\eta}\sqrt{\pi\eta \log n_2}}. \tag{2.14}$$

2.3.1 Proof of Theorem 2.2

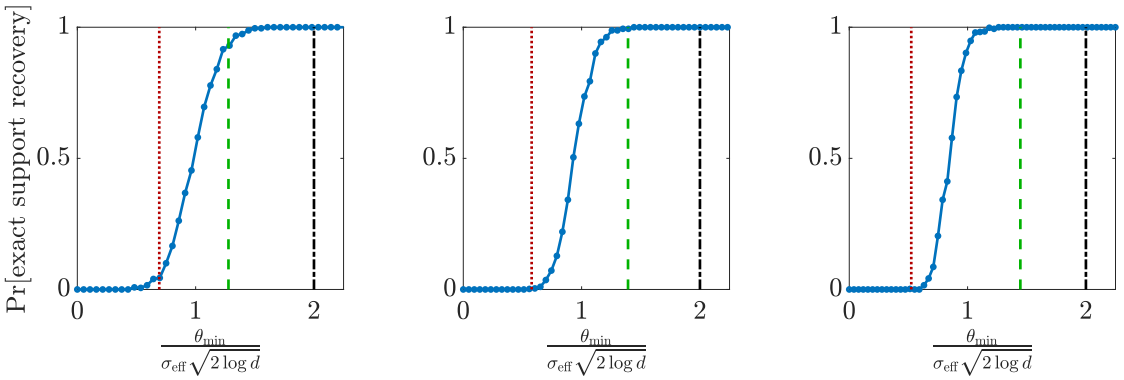
The proof is based on a tighter analysis of the proof of Ben-Haim et al. (2010, Thm 4). First, we define "bad" random events B_K and $B_{\tilde{d}}$ which indicate that the largest magnitude of inner products of the noise with support atoms and with non-support atoms is larger than their respective thresholds. We then define the "good" random event G that indicates that neither B_K nor $B_{\tilde{d}}$ occurs, and prove that the event G occurs with probability (2.4). Next, we show that under the event G, OMP with K iterations successfully recovers the support of $\boldsymbol{\theta}$. Finally, we prove that OMP* with threshold $\tau = \sigma\sqrt{2(1+\alpha)\log d}$ stops after exactly K iterations, and therefore also successfully recovers the support of $\boldsymbol{\theta}$.

In details, we define the following two random events

$$B_K = \left\{\max_{k\in\mathcal{S}} |\langle \mathbf{x}_k, \sigma\boldsymbol{\xi}\rangle| \geq \sqrt{\beta}\tau\right\}$$

and

$$B_{\tilde{d}} = \left\{\max_{i\notin\mathcal{S}} |\langle \mathbf{x}_i, \sigma\boldsymbol{\xi}\rangle| \geq \tau\right\},$$



(a) Sparsity $K = 2$ (b) Sparsity $K = 4$ (c) Sparsity $K = 6$

Figure 2.3: The solid blue line in each figure is the empirical probability of exact support recovery of a sparse vector by OMP as a function of its normalized signal-to-noise ratio in Setting 2. The dash-dotted black line is the sufficient condition (2.1) by Ben-Haim et al. (2010). The dashed green line is the sharper sufficient condition (2.3). The dotted red line is the approximate condition (2.6) below which OMP fails with high probability by Theorem 2.3. In all panels, when fixing a high success probability (for example 0.95), our sharper sufficient condition (2.3) is indeed closer to the empirical probability than the sufficient condition (2.1) by Ben-Haim et al. (2010), and in the first and second panel it is in fact tight.

and let the random event $G = \left(B_K \cup B_{\tilde{d}}\right)^C$ be the complement of their union. Note that while these definitions depend on the unknown support set $\mathcal{S}$, this is only for the sake of the analysis, and we do not assume that OMP receives the support $\mathcal{S}$ as input.

Next, we prove that the event G occurs with probability at least (2.4). Since the dictionary atoms are normalized, each random variable $\langle \mathbf{x}_i, \boldsymbol{\xi} \rangle$ is a standard Gaussian random variable. Therefore, applying Lemma 2.1 with $n_1 = K$, $n_2 = d$ and $\eta = (1+\alpha)\,\beta$ gives

$$\Pr\left[B_K\right] \leq \frac{K}{d^{(1+\alpha)\beta}\sqrt{\pi\,(1+\alpha)\,\beta\log d}}.$$

Since $K \leq d^{\beta}$, then

$$\Pr\left[B_K\right] \leq \frac{1}{d^{\alpha\beta}\sqrt{\pi\,(1+\alpha)\,\beta\log d}}.$$

Similarly, we can apply Lemma 2.1 again with $n_1 = \tilde{d} = d - K$, $n_2 = d$ and $\eta = 1+\alpha$ and get

$$\Pr\left[B_{\tilde{d}}\right] \leq \frac{\tilde{d}}{d^{(1+\alpha)}\sqrt{\pi\,(1+\alpha)\log d}}.$$

Since $\tilde{d} < d$, then

$$\Pr\left[B_{\tilde{d}}\right] \leq \frac{1}{d^{\alpha}\sqrt{\pi\,(1+\alpha)\log d}}.$$

By the definition of G and a union bound,

$$\Pr\left[G\right] = \Pr\left[\left(B_K \cup B_{\tilde{d}}\right)^c\right] \geq 1 - \Pr\left[B_K\right] - \Pr\left[B_{\tilde{d}}\right],$$

which proves that the event G occurs with probability at least (2.4).

The following lemma shows that under condition (2.3), one step of the OMP algorithm chooses an atom in the support $\mathcal{S}$.

Lemma 2.2. *Let* $\mathbf{z}$ *be an unknown vector with sparsity* K *and support* $\mathcal{S} = \operatorname{supp}\{\mathbf{z}\}$*, and let* $\mathbf{r} = \mathbf{X}\mathbf{z}+\sigma\boldsymbol{\xi}$ *where* $\mathbf{X} \in \mathbb{R}^{N\times d}$ *is a dictionary with normalized columns and coherence* μ*, and* $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_N)$*. Suppose that the MIP condition* (1.4) *holds, that* $K \leq d^\beta$ *for some* $0 < \beta < 1$ *and that for some* $\alpha \geq 0$

$$\max_{i\in\mathcal{S}} |\mathbf{z}_i| \geq \sigma_{\text{eff}}(K,\mu)\left(1+\sqrt{\beta}\right)\sqrt{2\,(1+\alpha)\log d}. \tag{2.15}$$

Then under the event G*,*

$$\max_{k\in\mathcal{S}} |\langle \mathbf{x}_k, \mathbf{r}\rangle| > \max_{i\notin\mathcal{S}} |\langle \mathbf{x}_i, \mathbf{r}\rangle|. \tag{2.16}$$

Proof of Lemma 2.2. Denote by $z_{\max} = \max_{i\in\mathcal{S}} |\mathbf{z}_i|$. Under the event G, the largest magnitude of an inner product of the observed signal $\mathbf{r}$ with a non-support atom $i \notin \mathcal{S}$ is

$$\begin{aligned}
\max_{i\notin\mathcal{S}} |\langle \mathbf{x}_i, \mathbf{r}\rangle| &= \max_{i\notin\mathcal{S}} \left| \langle \mathbf{x}_i, \sigma\boldsymbol{\xi}\rangle + \sum_{j\in\mathcal{S}} \mathbf{z}_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle \right| \\
&\leq \max_{i\notin\mathcal{S}} |\langle \mathbf{x}_i, \sigma\boldsymbol{\xi}\rangle| + \max_{i\notin\mathcal{S}} \sum_{j\in\mathcal{S}} |\mathbf{z}_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle| \\
&< \tau + K\mu z_{\max}.
\end{aligned} \tag{2.17}$$

Similarly,

$$\begin{aligned}
\max_{k\in\mathcal{S}} |\langle \mathbf{x}_k, \mathbf{r}\rangle| &= \max_{k\in\mathcal{S}} \left| \mathbf{z}_k + \langle \mathbf{x}_k, \sigma\boldsymbol{\xi}\rangle + \sum_{j\in\mathcal{S}\setminus\{k\}} \mathbf{z}_j \langle \mathbf{x}_k, \mathbf{x}_j\rangle \right| \\
&\geq z_{\max} - \max_{k\in\mathcal{S}} \left| \langle \mathbf{x}_k, \sigma\boldsymbol{\xi}\rangle + \sum_{j\in\mathcal{S}\setminus\{k\}} \mathbf{z}_j \langle \mathbf{x}_k, \mathbf{x}_j\rangle \right| \\
&> z_{\max} - \sqrt{\beta}\tau - (K-1)\,\mu z_{\max}.
\end{aligned} \tag{2.18}$$

Combining the last two equations gives

$$\max_{k\in\mathcal{S}} |\langle \mathbf{x}_k, \mathbf{r}\rangle| - \max_{i\notin\mathcal{S}} |\langle \mathbf{x}_i, \mathbf{r}\rangle| > z_{\max} - (2K-1)\,\mu z_{\max} - \sqrt{\beta}\tau - \tau.$$

Substituting for τ implies that Eq. (2.16) holds under condition (2.15). □

Next, assume that G occurs. We prove the first part of Theorem 2.2 by induction. Consider the first iteration of `OMP`, described in Algorithm 1. In line 3, `OMP` chooses an atom $\mathbf{x}_i$ whose inner product with $\mathbf{y}$ is maximal. In other words, condition (2.16) must hold for $\mathbf{r} = \mathbf{y}$ and $\mathbf{z} = \boldsymbol{\theta}$ for `OMP` to select an atom $i \in \mathcal{S}$ at the first iteration. When G occurs, then by condition (2.3) and by Lemma 2.2 `OMP` selects a support atom, i.e., $\hat{\mathcal{S}}_1 \subseteq \mathcal{S}$. Assume by induction that the set of atoms that were selected in all previous $1 \leq t < K$ iterations is a subset of the support set, i.e., $\operatorname{supp}\left\{\hat{\boldsymbol{\theta}}_t\right\} = \hat{\mathcal{S}}_t \subseteq \mathcal{S}$. Hence,

$$\mathbf{r}_t = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}_t = \mathbf{X}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\right) + \sigma\boldsymbol{\xi}, \tag{2.19}$$

where $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t$ is a sparse vector whose support is contained in $\mathcal{S}$. In addition, since `OMP` selects exactly one atom at each iteration,

$$\left|\operatorname{supp}\left\{\hat{\boldsymbol{\theta}}_t\right\}\right| = t < K = |\operatorname{supp}\{\boldsymbol{\theta}\}|.$$

Hence, at least one entry in $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t$ is equal to its corresponding entry in $\boldsymbol{\theta}$ and

$$\max_{i\in\mathcal{S}} \left|\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\right)_i\right| \geq \min_{i\in\mathcal{S}} |\boldsymbol{\theta}_i| = \theta_{\min}. \tag{2.20}$$

Since by Eq. (2.3) $\theta_{\min}$ is larger than the bound in Eq. (2.15), we can apply Lemma 2.2 with $\mathbf{r} = \mathbf{r}_t$ and $\mathbf{z} = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t$ to conclude that under the event G,

$$\max_{k \in \mathcal{S}} |\langle \mathbf{x}_k, \mathbf{r}_t \rangle| > \max_{i \notin \mathcal{S}} |\langle \mathbf{x}_i, \mathbf{r}_t \rangle| .$$

This implies that OMP chooses a support atom at iteration $t+1$. Therefore by induction the OMP algorithm recovers the unknown support of $\boldsymbol{\theta}$ under the event G, which concludes the proof of the first part of Theorem 2.2.

It remains to show that the OMP* algorithm with threshold $\tau = \sigma\sqrt{2(1+\alpha)\log d}$ does not stop early before the K-th iteration, and that it does stop after the K-th iteration. At iteration $1 \le t \le K$,

$$\left\|\mathbf{X}^T \mathbf{r}_t\right\|_\infty = \max_{i \in [d]} |\langle \mathbf{x}_i, \mathbf{r}_t \rangle| \ge \max_{k \in \mathcal{S}} |\langle \mathbf{x}_k, \mathbf{r}_t \rangle|.$$

Under the event G, by Eqs. (2.18) and (2.20),

$$\max_{k \in \mathcal{S}} |\langle \mathbf{x}_k, \mathbf{r}_t \rangle| > \theta_{\min} \left(1 - (K-1)\,\mu\right) - \sqrt{\beta}\tau.$$

Finally, by condition (2.3),

$$\left\|\mathbf{X}^T \mathbf{r}_t\right\|_\infty > \frac{\left(1+\sqrt{\beta}\right)\left(1-(K-1)\,\mu\right)}{1-(2K-1)\,\mu}\tau - \sqrt{\beta}\tau = \frac{1-\left(\left(1-\sqrt{\beta}\right)K-1\right)\mu}{1-(2K-1)\,\mu}\tau > \tau,$$

which proves that OMP* does not stop early.

At the end of iteration $t = K$ all support atoms have been selected. Let $\boldsymbol{\theta}_{\mathcal{S}} \in \mathbb{R}^K$ and $\mathbf{X}_{\mathcal{S}} \in \mathbb{R}^{N \times K}$ be the vector $\boldsymbol{\theta}$ and the dictionary $\mathbf{X}$ restricted to the support $\mathcal{S}$ (respectively), and let $P_{\mathcal{S}} = \mathbf{X}_{\mathcal{S}} \mathbf{X}_{\mathcal{S}}^\dagger = \mathbf{X}_{\mathcal{S}} \left(\mathbf{X}_{\mathcal{S}}^T \mathbf{X}_{\mathcal{S}}\right)^{-1} \mathbf{X}_{\mathcal{S}}^T$ be the projection of the observed signal onto the linear subspace spanned by the elements of $\mathcal{S}$. Then

$$\mathbf{r}_K = \mathbf{y} - \mathbf{X}_{\mathcal{S}} \mathbf{X}_{\mathcal{S}}^\dagger \mathbf{y} = (I - P_{\mathcal{S}})\,\mathbf{y} = (I - P_{\mathcal{S}})\,\mathbf{X}_{\mathcal{S}} \boldsymbol{\theta}_{\mathcal{S}} + (I - P_{\mathcal{S}})\,\sigma\boldsymbol{\xi}.$$

Since $I - P_{\mathcal{S}}$ is a projection to the linear space that is orthogonal to the subspace spanned by the elements of $\mathcal{S}$, the first term of the last equation above is zero. Hence, under the event G,

$$\left\|\mathbf{X}^T \mathbf{r}_K\right\|_\infty = \max_{i \in [d]} |\langle \mathbf{x}_i, \mathbf{r}_K \rangle| = \max_{i \in [d]} |\langle \mathbf{x}_i, (I - P_{\mathcal{S}})\,\sigma\boldsymbol{\xi} \rangle| \le \max_{i \in [d]} |\langle \mathbf{x}_i, \sigma\boldsymbol{\xi} \rangle| \le \tau.$$

Therefore, OMP* stops after exactly K iterations. This concludes the proof of Theorem 2.2. □

2.3.2 Proof of Theorem 2.3

First, we present an outline of the proof. Given parameters d, N, K, μ with $N < d$, and where K, μ satisfy conditions (2.10)-(2.12), we construct a dictionary $\mathbf{X} \in \mathbb{R}^{N \times d}$ with coherence μ and a sparse vector $\boldsymbol{\theta} \in \mathbb{R}^d$ with sparsity K. We show that when the smallest coefficient in $\boldsymbol{\theta}$ is sufficiently small as in condition (2.13), then with probability at least P_0, OMP fails to detect a support atom already at the first iteration, and therefore fails to recover the support of $\boldsymbol{\theta}$.

To prove the theorem we shall use the following auxiliary lemmas. The first lemma concerns the maximum of several correlated normal random variables.

Lemma 2.3. *Let $(X_1, \ldots, X_n) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{ii} = 1$ for all $i \in [n]$ and $0 < |\boldsymbol{\Sigma}_{ij}| \le \eta < 1$ for all $i \ne j \in [n]$. For $M_n = \max_{i \in [n]} X_i$, the following hold:*

1. *(Lopes (2018)). There exists a universal constant $c_0 > 1$ such that*

$$\mathrm{E}\left[M_n\right] \ge \sqrt{2(1-\eta)\log n} - (c_0 - 1)\sqrt{\log\log n}. \tag{2.21}$$

2. (Tanguy (2015)). There exists $C > 0$ such that for any $n \geq 2$ and $t > 0$,

$$\Pr\left[|M_n - \mathrm{E}\left[M_n\right]| > t\right] \leq 6e^{-Ct\sqrt{\min\left\{\frac{1}{\eta}, \log n\right\}}}. \tag{2.22}$$

In constructing our specific dictionary, we will use the following lemma whose proof appears in Section 2.3.3.

Lemma 2.4. *For any integer $K > 1$ and any $0 \leq \mu < \frac{1}{K-1}$, there exist vectors $\mathbf{a}_1, \ldots, \mathbf{a}_K \in \mathbb{R}^K$ such that for all $i, j \in [K]$*

$$\langle \mathbf{a}_i, \mathbf{a}_j \rangle = \begin{cases} 1 & i = j \\ -\mu & i \neq j. \end{cases} \tag{2.23}$$

Proof of Theorem 2.3. Recall the notations $\tilde{N} = N - K$ and $\tilde{d} = d - K$. Given the sparsity K and coherence μ, we first construct vectors $\mathbf{a}_1, \ldots, \mathbf{a}_K \in \mathbb{R}^K$ as in Lemma 2.4. Next, we construct our dictionary $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$ as follows

$$\mathbf{X} = \begin{bmatrix} \mathbf{a}_1 & \ldots & \mathbf{a}_K & \sqrt{\tilde{\mu}}\bar{\mathbf{a}} & \ldots & \sqrt{\tilde{\mu}}\bar{\mathbf{a}} \\ \mathbf{0} & \ldots & \mathbf{0} & \sqrt{1-\tilde{\mu}}\mathbf{w}_{K+1} & \ldots & \sqrt{1-\tilde{\mu}}\mathbf{w}_d \end{bmatrix}, \tag{2.24}$$

where $\bar{\mathbf{a}} = \frac{\sum_{i\in[K]} \mathbf{a}_i}{\left\|\sum_{i\in[K]} \mathbf{a}_i\right\|}$ and the constant $\tilde{\mu}$ is defined in Eq. (2.8). For future use, note that

$$\left\| \sum_{i\in[K]} \mathbf{a}_i \right\| = \sqrt{\sum_{i\in[K]} \sum_{i'\in[K]} \langle \mathbf{a}_i, \mathbf{a}_{i'} \rangle} = \sqrt{K\left(1 - (K-1)\,\mu\right)}. \tag{2.25}$$

The key requirements of the vectors $\mathbf{w}_{K+1}, \ldots, \mathbf{w}_N \in \mathbb{R}^{\tilde{N}}$ is that they have unit norm $\|\mathbf{w}_i\| = 1$ and that they satisfy the following condition

$$\max_{K+1 \leq i < j \leq d} |\langle \mathbf{w}_i, \mathbf{w}_j \rangle| \leq L. \tag{2.26}$$

As the following lemma shows, condition (2.26) implies that the coherence of $\mathbf{X}$ is μ. The proof appears in Section 2.3.3.

Lemma 2.5. *Assume that K, μ satisfy conditions* (2.11) *and* (2.12)*. Then, under condition* (2.26)*, the coherence of the dictionary $\mathbf{X}$ of Eq.* (2.24) *is exactly μ.*

Before proceeding we remark that such a dictionary $\mathbf{X}$ indeed exists. Specifically, Lemma 2.6 in Section 2.3.3 shows that if $\mathbf{w}_{K+1}, \ldots, \mathbf{w}_N$ are drawn independently and uniformly at random from the unit sphere, and μ satisfies condition (2.12) with a (possibly) slightly higher value $L = 2\sqrt{\frac{\log \tilde{d}}{\tilde{N}}}$, then condition (2.26) holds with high probability.

Let us now analyze the inability of `OMP` to successfully recover the support of an underlying K-sparse vector $\boldsymbol{\theta}$, given $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \sigma\boldsymbol{\xi}$. Consider a dictionary $\mathbf{X}$ of the form (2.24) and the vector

$$\boldsymbol{\theta} = \nu \sum_{j=1}^{K} \mathbf{e}_j, \tag{2.27}$$

which implies that $\mathcal{S} = \{1, \ldots, K\}$ and $\theta_{\min} = \nu$. Note that $\mathbf{X}\boldsymbol{\theta} = \nu \sum_{i\in[K]} \mathbf{x}_i = \nu \begin{pmatrix} \sum_{i\in[K]} \mathbf{a}_i \\ \mathbf{0} \end{pmatrix}$. From this point on we view $\mathbf{X}$ and $\boldsymbol{\theta}$ as fixed and the randomness is only over realizations of the noise vector $\boldsymbol{\xi}$.

Our goal is to show that if ν is sufficiently small such that condition (2.13) holds, then with high probability OMP fails to recover the support $\mathcal{S}$. For future use, we introduce the following two random variables that depend on the noise $\boldsymbol{\xi}$,

$$A_{\text{out}} = \max_{i \notin \mathcal{S}} |\langle \mathbf{x}_i, \mathbf{y} \rangle| \tag{2.28}$$

and

$$A_{\text{in}} = \max_{k \in \mathcal{S}} |\langle \mathbf{x}_k, \mathbf{y} \rangle| \,. \tag{2.29}$$

A sufficient condition for the failure of OMP, as described in Algorithm 1, is that it would choose a non-support atom in the first step of the algorithm with probability $\geq P_0$, or equivalently if

$$\Pr\left[A_{\text{out}} > A_{\text{in}}\right] \geq P_0.$$

As we shall see below, due to dependencies between various inner products $\langle \mathbf{x}_i, \mathbf{y} \rangle$, the probability that $A_{\text{out}} > A_{\text{in}}$ is difficult to analyze. Instead, we will introduce two other random variables B_{out} and B_{in} which satisfy $A_{\text{out}} \geq B_{\text{out}}$, $B_{\text{in}} \geq A_{\text{in}}$ and for which $\Pr\left[B_{\text{out}} > B_{\text{in}}\right]$ is simpler to analyze.

First, we decompose the noise into its support elements $\boldsymbol{\xi}_K \in \mathbb{R}^K$ and non-support elements $\boldsymbol{\xi}_{\tilde{N}} \in \mathbb{R}^{\tilde{N}}$, such that $\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\xi}_K \\ \boldsymbol{\xi}_{\tilde{N}} \end{pmatrix}$. Next, we analyze the random variable A_{out} and define the random variable B_{out}. Using the value for $\tilde{\mu}$ in Eq. (2.8) and value of the norm in Eq. (2.25), the inner product of the observed signal $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \sigma\boldsymbol{\xi}$ with a non-support atom $i \notin \mathcal{S}$ is

$$\begin{aligned}
\langle \mathbf{x}_i, \mathbf{y} \rangle &= \left\langle \mathbf{x}_i, \nu \sum_{j=1}^{K} \mathbf{x}_j \right\rangle + \langle \mathbf{x}_i, \sigma\boldsymbol{\xi} \rangle \\
&= \left\langle \begin{pmatrix} \sqrt{\tilde{\mu}}\bar{\mathbf{a}} \\ \sqrt{1-\tilde{\mu}}\mathbf{w}_i \end{pmatrix}, \nu \sum_{j=1}^{K} \begin{pmatrix} \mathbf{a}_j \\ \mathbf{0} \end{pmatrix} \right\rangle + \langle \mathbf{x}_i, \sigma\boldsymbol{\xi} \rangle \\
&= \nu \sum_{j=1}^{K} \left\langle \sqrt{\tilde{\mu}}\bar{\mathbf{a}}, \mathbf{a}_j \right\rangle + \sigma \langle \mathbf{x}_i, \boldsymbol{\xi} \rangle = \frac{\nu\sqrt{\tilde{\mu}}}{\sqrt{K\left(1-(K-1)\,\mu\right)}} \sum_{j=1}^{K} \sum_{j'=1}^{K} \langle \mathbf{a}_{j'}, \mathbf{a}_j \rangle + \sigma \langle \mathbf{x}_i, \boldsymbol{\xi} \rangle \\
&= \frac{\nu\mu}{1-(K-1)\,\mu} \left(K\left(1-(K-1)\,\mu\right)\right) + \sigma \langle \mathbf{x}_i, \boldsymbol{\xi} \rangle = \nu K\mu + \sigma \langle \mathbf{x}_i, \boldsymbol{\xi} \rangle \\
&= \nu K\mu + \sigma\sqrt{\tilde{\mu}} \langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle + \sigma\sqrt{1-\tilde{\mu}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle \,.
\end{aligned} \tag{2.30}$$

We define B_{out} by

$$B_{\text{out}} = \nu K\mu - \sigma\sqrt{\tilde{\mu}} \left|\langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle\right| + \sigma\sqrt{1-\tilde{\mu}} \max_{i \notin \mathcal{S}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle \,. \tag{2.31}$$

Using Eq. (2.28) and (2.30),

$$\begin{aligned}
A_{\text{out}} &= \max_{i \notin \mathcal{S}} |\langle \mathbf{x}_i, \mathbf{y} \rangle| \geq \max_{i \notin \mathcal{S}} \langle \mathbf{x}_i, \mathbf{y} \rangle \\
&= \nu K\mu + \sigma\sqrt{\tilde{\mu}} \langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle + \sigma\sqrt{1-\tilde{\mu}} \max_{i \notin \mathcal{S}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle \\
&\geq \nu K\mu - \sigma\sqrt{\tilde{\mu}} \left|\langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle\right| + \sigma\sqrt{1-\tilde{\mu}} \max_{i \notin \mathcal{S}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle = B_{\text{out}}.
\end{aligned}$$

We now analyze the random variable A_{in} and define the random variable B_{in}. The inner product of the observed signal $\mathbf{y}$ with a support atom $k \in \mathcal{S}$ is

$$\begin{aligned}
\langle \mathbf{x}_k, \mathbf{y} \rangle &= \left\langle \mathbf{x}_k, \nu \sum_{j=1}^{K} \mathbf{x}_j \right\rangle + \langle \mathbf{x}_k, \sigma\boldsymbol{\xi} \rangle \\
&= \nu \sum_{j=1}^{K} \langle \mathbf{a}_k, \mathbf{a}_j \rangle + \sigma \langle \mathbf{x}_k, \boldsymbol{\xi} \rangle = \nu\left(1-(K-1)\,\mu\right) + \sigma \langle \mathbf{a}_k, \boldsymbol{\xi}_K \rangle \,.
\end{aligned} \tag{2.32}$$

To circumvent the dependence between the random variables $\langle \mathbf{a}_k, \boldsymbol{\xi}_K \rangle$ and $\langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle$, we decompose each $\mathbf{a}_k$ into two components, $\mathbf{a}_k^{\parallel}$ which is parallel to $\bar{\mathbf{a}}$ and $\mathbf{a}_k^{\perp}$ which is orthogonal to $\bar{\mathbf{a}}$. Using Eq. (2.7) and (2.25), for each $k \in \mathcal{S}$

$$\langle \mathbf{a}_k, \bar{\mathbf{a}} \rangle = \frac{1}{\left\| \sum_{i \in [K]} \mathbf{a}_i \right\|} \sum_{i \in [K]} \langle \mathbf{a}_k, \mathbf{a}_i \rangle = \frac{1 - (K-1)\,\mu}{\sqrt{K\,(1 - (K-1)\mu)}} = \rho. \tag{2.33}$$

Thus $\mathbf{a}_k^{\perp} = \mathbf{a}_k - \langle \mathbf{a}_k, \bar{\mathbf{a}} \rangle\, \bar{\mathbf{a}} = \mathbf{a}_k - \rho \bar{\mathbf{a}}$. Combining this relation with Eq. (2.32), we can rewrite A_{in} as

$$A_{\text{in}} = \max_{k \in \mathcal{S}} \left| \nu\,(1 - (K-1)\,\mu) + \sigma \left\langle \mathbf{a}_k^{\perp} + \rho \bar{\mathbf{a}}, \boldsymbol{\xi}_K \right\rangle \right|.$$

We define B_{in} by

$$B_{\text{in}} = \nu\,(1 - (K-1)\,\mu) + \sigma\rho\,|\langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle| + \sigma \max_{k \in \mathcal{S}} \left| \left\langle \mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K \right\rangle \right|. \tag{2.34}$$

By the triangle inequality $A_{\text{in}} \le B_{\text{in}}$.

Now that we defined B_{out} and B_{in}, we proceed to analyze the probability that $B_{\text{out}} > B_{\text{in}}$, or equivalently,

$$\sqrt{1-\tilde{\mu}} \max_{i \notin \mathcal{S}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle > \frac{\nu}{\sigma_{\text{eff}}(K,\mu)} + \left(\rho + \sqrt{\tilde{\mu}} \right) |\langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle| + \max_{k \in \mathcal{S}} \left| \left\langle \mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K \right\rangle \right|. \tag{2.35}$$

For constants b_1, b_2 that will be determined later, denote the following three probabilities

$$P_1 = \Pr\left[\max_{i \notin \mathcal{S}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle > b_1 \right],$$

$$P_2 = \Pr\left[|\langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle| < b_2 \right],$$

and

$$P_3 = \Pr\left[\max_{k \in \mathcal{S}} \left| \left\langle \mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K \right\rangle \right| < \sqrt{1-\tilde{\mu}} b_1 - \frac{\nu}{\sigma_{\text{eff}}(K,\mu)} - \left(\rho + \sqrt{\tilde{\mu}} \right) b_2 \right].$$

By the statistical independence of $\boldsymbol{\xi}_K$ and $\boldsymbol{\xi}_{\tilde{N}}$ and the linear independence of $\bar{\mathbf{a}}$ and $\mathbf{a}_k^{\perp}$ for all $k \in \mathcal{S}$,

$$\begin{aligned} \Pr\left[B_{\text{out}} > B_{\text{in}}\right] &\ge \Pr\left[\max_{i \notin \mathcal{S}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle > b_1 \right] \times \Pr\left[|\langle \bar{\mathbf{a}}, \boldsymbol{\xi}_K \rangle| < b_2 \right] \\ &\quad \times \Pr\left[\max_{k \in \mathcal{S}} \left| \left\langle \mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K \right\rangle \right| < \sqrt{1-\tilde{\mu}} b_1 - \frac{\nu}{\sigma_{\text{eff}}(K,\mu)} - \left(\rho + \sqrt{\tilde{\mu}} \right) b_2 \right] \\ &= P_1 \cdot P_2 \cdot P_3. \end{aligned}$$

Hence, instead of proving that (2.35) holds with probability at least P_0, it suffices to prove that $P_1 \cdot P_2 \cdot P_3 \ge P_0$.

We proceed by calculating each of these probabilities, beginning with P_1. Since $\mathbf{w}_{K+1}, \ldots, \mathbf{w}_N$ are fixed unit vectors, each inner product between $\mathbf{w}_i$ and the vector of standard normals $\boldsymbol{\xi}_{\tilde{N}}$ is a standard normal random variable. Let

$$b_1 = \sqrt{2\,(1-\mu) \log \tilde{d}} - c_0 \sqrt{\log \log \tilde{d}},$$

where c_0 is the constant from Lemma 2.3. Denote by $M_{\tilde{d}} = \max_{i \notin \mathcal{S}} \langle \mathbf{w}_i, \boldsymbol{\xi}_{\tilde{N}} \rangle$. By the first part of Lemma 2.3,

$$\begin{aligned} \mathrm{E}\left[M_{\tilde{d}}\right] &\ge \sqrt{2(1-\mu) \log \tilde{d}} - (c_0 - 1) \sqrt{\log \log \tilde{d}} \\ &= b_1 + \sqrt{\log \log \tilde{d}}. \end{aligned}$$

Therefore, by the triangle inequality,

$$\begin{aligned} P_1 &= \Pr\left[M_{\tilde{d}} > b_1\right] \geq \Pr\left[M_{\tilde{d}} > \mathrm{E}\left[M_{\tilde{d}}\right] - \sqrt{\log\log\tilde{d}}\right] \\ &\geq \Pr\left[\left|M_{\tilde{d}} - \mathrm{E}\left[M_{\tilde{d}}\right]\right| < \sqrt{\log\log\tilde{d}}\right]. \end{aligned}$$

By the second part of Lemma 2.3,

$$P_1 \geq 1 - 6e^{-C\sqrt{\log\log\tilde{d}\min\left\{\mu^{-1},\,\log\tilde{d}\right\}}}.$$

Next, we calculate P_2. Let $b_2 = \sqrt{2\log\log d}$. The term $\langle\bar{\mathbf{a}}, \boldsymbol{\xi}_K\rangle$ is simply a standard normal variable. By Lemma 2.1 with parameters $n_1 = 1$, $n_2 = \log d$ and $\eta = 1$, we obtain that

$$P_2 = \Pr\left[|\langle\bar{\mathbf{a}}, \boldsymbol{\xi}_K\rangle| < \sqrt{2\log\log d}\right] > 1 - \frac{1}{\log d\sqrt{\pi\log\log d}}.$$

Lastly, we calculate P_3. Recall that by construction $\theta_{\min} = \nu$. By Eq. (2.13),

$$\frac{\nu}{\sigma_{\mathrm{eff}}(K,\mu)} = \sqrt{1-\tilde{\mu}}b_1 - \left(\rho + \sqrt{\tilde{\mu}}\right)b_2 - \sqrt{2\beta\left(1-\rho^2\right)\log d}.$$

Therefore,

$$\begin{aligned} P_3 &= \Pr\left[\max_{k\in\mathcal{S}}\left|\left\langle\mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K\right\rangle\right| < \sqrt{1-\tilde{\mu}}b_1 - \frac{\nu}{\sigma_{\mathrm{eff}}(K,\mu)} - \left(\rho + \sqrt{\tilde{\mu}}\right)b_2\right] \\ &= \Pr\left[\max_{k\in\mathcal{S}}\left|\left\langle\mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K\right\rangle\right| < \sqrt{2\beta\left(1-\rho^2\right)\log d}\right] \\ &= \Pr\left[\max_{k\in\mathcal{S}}\left|\frac{\left\langle\mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K\right\rangle}{\sqrt{1-\rho^2}}\right| < \sqrt{2\beta\log d}\right]. \end{aligned}$$

Note that for all $k \in \mathcal{S}$, $\left\|\mathbf{a}_k^{\perp}\right\| = \sqrt{1-\rho^2}$. Hence, each random variable $\frac{\left\langle\mathbf{a}_k^{\perp}, \boldsymbol{\xi}_K\right\rangle}{\sqrt{1-\rho^2}}$ is Gaussian with zero mean and variance 1. We can apply Lemma 2.1 with $n_1 = K$, $n_2 = d$ and $\eta = \beta$, and use the inequality (2.10) to get

$$P_3 \geq 1 - \frac{K}{d^{\beta}\sqrt{\pi\beta\log d}} \geq 1 - \frac{1}{\sqrt{\pi\beta\log d}}.$$

By a union bound, for sufficiently large N and d the probability that `OMP` fails to recover the support $\mathcal{S}$ is at least P_0, which completes the proof of Theorem 2.3. □

2.3.3 Proofs of Lemmas

To conclude, we prove the auxiliary lemmas.

Proof of Lemma 2.1. The proof is similar to that of Ben-Haim et al. (2010, Lemma 2). By Šidák (1967, Thm. 1), since $X_1, \ldots, X_{n_1}$ are jointly Gaussian random variables, then

$$\Pr\left[\max_{i\in[n_1]}|X_i| < \sqrt{2\eta\log n_2}\right] \geq \prod_{i\in[n_1]}\Pr\left[|X_i| \leq \sqrt{2\eta\log n_2}\right] = \Pr\left[|X_1| \leq \sqrt{2\eta\log n_2}\right]^{n_1}. \quad (2.36)$$

Each X_i is a standard normal random variable. Therefore,

$$\Pr\left[|X_1| \leq x\right] = 1 - 2Q\left(x\right),$$

where $Q(x)$ is the Gaussian tail probability function. Applying the inequality

$$Q\left(x\right) \leq \frac{1}{x\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

with $x = \sqrt{2\eta \log n_2}$ gives

$$\Pr\left[|X_1| \leq \sqrt{2\eta \log n_2}\right] \geq 1 - \frac{e^{-\eta \log n_2}}{\sqrt{\pi \eta \log n_2}} = 1 - \frac{1}{n_2^{\eta}\sqrt{\pi \eta \log n_2}}. \tag{2.37}$$

Inserting Eq. (2.37) into (2.36) and using the inequality $(1-a)^n \geq 1 - an$ completes the proof. □

Proof of Lemma 2.4. Let $G \in \mathbb{R}^{K \times K}$ be the following symmetric matrix with entries

$$G_{ij} = \begin{cases} 1 & i = j \\ -\mu & i \neq j. \end{cases}$$

Hence, G can be rewritten as a rank-one perturbation of the identity matrix

$$G = -\mu \mathbf{1}\mathbf{1}^T + (1+\mu)\,\mathbf{I}.$$

If $\mu < \frac{1}{K-1}$, then G is positive definite. Therefore, it is the Gram matrix of a set of linearly independent vectors, i.e., there exist $\mathbf{a}_1, \ldots, \mathbf{a}_K$ such that condition (2.23) holds, which completes the proof (Horn and Johnson, 2012, p. 441).

For completeness, we describe an explicit construction. Let $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_K]$ be an orthogonal matrix where $\mathbf{v}_1 = \frac{1}{\sqrt{K}}\mathbf{1}$ and

$$\mathbf{A} = \begin{pmatrix} \sqrt{\frac{1-(K-1)\mu}{K}}\mathbf{1}^T \\ \sqrt{1+\mu}\mathbf{v}_2^T \\ \vdots \\ \sqrt{1+\mu}\mathbf{v}_K^T \end{pmatrix}.$$

Let us now prove that $\mathbf{A}$ indeed satisfies condition (2.23). Since $\mathbf{V}$ is orthogonal, its rows also form an orthonormal basis of $\mathbb{R}^K$. First, consider the diagonal entries of the Gram matrix $\mathbf{A}^T\mathbf{A}$. For all $i \in [K]$, $\frac{1-(K-1)\mu}{K} = -\mu + (1+\mu)\,\frac{1}{K} = -\mu + (1+\mu)\,\mathbf{V}_{1i}^2$, and therefore

$$(\mathbf{A}^T\mathbf{A})_{ii} = \frac{1-(K-1)\,\mu}{K} + (1+\mu)\sum_{k=2}^{K}\mathbf{V}_{ki}^2 = -\mu + (1+\mu)\sum_{k=1}^{K}\mathbf{V}_{ki}^2 = 1.$$

Similarly, for all $i \neq j \in [K]$, $\frac{1-(K-1)\mu}{K} = -\mu + (1+\mu)\,\frac{1}{K} = -\mu + (1+\mu)\,\mathbf{V}_{1i}\mathbf{V}_{1j}$, and therefore

$$(\mathbf{A}^T\mathbf{A})_{ij} = \frac{1-(K-1)\,\mu}{K} + (1+\mu)\sum_{k=2}^{K}\mathbf{V}_{ki}\mathbf{V}_{kj} = -\mu + (1+\mu)\sum_{k=1}^{K}\mathbf{V}_{ki}\mathbf{V}_{kj} = -\mu.$$

□

Proof of Lemma 2.5. To prove that the coherence of $\mathbf{X}$ is μ we need to analyze three types of dot products $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. The first type is $1 \leq i < j \leq K$, the second type is $1 \leq i \leq K < j \leq d$, and the third is $K+1 \leq i < j \leq d$.

Beginning with the first type, by construction, for any $1 \leq i < j \leq K$,

$$|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| = |\langle \mathbf{a}_i, \mathbf{a}_j \rangle| = \mu.$$

For the second type, by Eq. (2.8), for any $1 \le i \le K < j \le d$,

$$|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| = \left| \left\langle \mathbf{a}_i, \sqrt{\tilde{\mu}} \bar{\mathbf{a}} \right\rangle \right| = \frac{\mu}{\rho} \frac{1}{\left\| \sum_{i' \in [K]} \mathbf{a}_{i'} \right\|} \left| \sum_{i' \in [K]} \langle \mathbf{a}_i, \mathbf{a}_{i'} \rangle \right|.$$

Inserting Eq. (2.7) and (2.25),

$$|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| = \mu \sqrt{\frac{K}{1-(K-1)\mu}} \frac{1-(K-1)\mu}{\sqrt{K\left(1-(K-1)\mu\right)}} = \mu.$$

Finally, we address the third type. By the triangle inequality and condition (2.26),

$$|\langle \mathbf{x}_i, \mathbf{x}_j \rangle| = |\tilde{\mu} + (1-\tilde{\mu}) \langle \mathbf{w}_i, \mathbf{w}_j \rangle| \le \tilde{\mu} + (1-\tilde{\mu}) |\langle \mathbf{w}_i, \mathbf{w}_j \rangle| \le \tilde{\mu} + (1-\tilde{\mu}) L.$$

It remains to show that for values of μ in the range of Eq. (2.12),

$$\tilde{\mu} + (1-\tilde{\mu}) L \le \mu. \tag{2.38}$$

Using the definition (2.8) of $\tilde{\mu}$, condition (2.38) is

$$\frac{K\mu^2}{1-(K-1)\mu} + \left(1 - \frac{K\mu^2}{1-(K-1)\mu}\right) L \le \mu.$$

In turn, this inequality can be rewritten as the following quadratic equation

$$\mu^2 \left(2K - 1 - LK\right) - \mu \left(L\left(K-1\right)+1\right) + L \le 0.$$

Notice that since $L < 1$, the term $2K - 1 - LK > K - 1 \ge 0$. The above inequality is thus satisfied by values of μ in Eq. (2.12). This range is not empty if $\frac{4L(2K-1-LK)}{(L(K-1)+1)^2} \le 1$. It is easy to verify that this condition holds for K values in (2.11). Note that the above condition also holds for $K \ge \frac{3-L+\sqrt{8-8L}}{L}$, however this range is often not possible due to the MIP condition (1.4). □

Lemma 2.6. *Let $\mathbf{w}_{K+1}, \ldots, \mathbf{w}_N$ be $\tilde{d}$ vectors drawn independently and uniformly at random from the $\tilde{N}$-dimensional unit sphere. Suppose $\tilde{d} = \tilde{d}_{\tilde{N}} \to \infty$ satisfies $\frac{\log \tilde{d}}{\tilde{N}} \to 0$ as $\tilde{N} \to \infty$. Then as $\tilde{N} \to \infty$, condition (2.26) with $L = 2\sqrt{\frac{\log \tilde{d}}{\tilde{N}}}$ holds with probability $e^{-1/\sqrt{8\pi \log \tilde{d}}}$.*

To prove Lemma 2.6, we need the following auxiliary lemma which bounds the largest magnitude of an inner product between random unit vectors.

Lemma 2.7 (Cai and Jiang (2012))**.** *Let $\mathbf{a}_1, \ldots \mathbf{a}_N$ be i.i.d. vectors drawn uniformly at random from the N-dimensional unit sphere and let*

$$L_N = \max_{1 \le i < j \le d} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|.$$

Suppose $d = d_N \to \infty$ satisfies $\frac{\log d}{N} \to 0$ as $N \to \infty$. Then as $N \to \infty$, the random variable

$$N \log(1 - L_N^2) + 4 \log d - \log \log d$$

converges weakly to an extreme value distribution with the distribution function $F(y) = 1 - e^{-Ke^{y/2}}$ for $y \in \mathbb{R}$ and $K = \frac{1}{\sqrt{8\pi}}$.

Proof of Lemma 2.6. Note that in the regime stated in the lemma, $L_{\tilde{N}} \to 0$ as $\tilde{N} \to \infty$. Hence $\log(1 - L_{\tilde{N}}^2) \approx -L_{\tilde{N}}^2$. By Lemma 2.7 for $y = -\log\log(\tilde{d})$ and $L = 2\sqrt{\frac{\log \tilde{d}}{\tilde{N}}}$, in the limit

$$\begin{aligned} \Pr\left[\tilde{N}\log(1 - L_{\tilde{N}}^2) + 4\log\tilde{d} - \log\log\tilde{d} \geq -\log\log\tilde{d}\right] &= \Pr\left[L_{\tilde{N}} \leq L\right] \\ &= e^{-Ke^{-\log\log\tilde{d}/2}} \\ &= e^{-1/\sqrt{8\pi\log\tilde{d}}}. \end{aligned}$$

Therefore as $\tilde{N} \to \infty$, $\max_{K+1 \leq i < j \leq d} |\langle \mathbf{w}_i, \mathbf{w}_j \rangle| = L_{\tilde{N}} \leq L$ and condition (2.26) is satisfied with probability $e^{-1/\sqrt{8\pi\log(\tilde{d})}}$. □

Chapter 3

Distributed Sparse Normal Means Estimation with Sublinear Communication [1]

We consider the problem of sparse normal means estimation in a distributed setting with communication constraints. We assume there are M machines, each holding d-dimensional observations of a K-sparse vector $\boldsymbol{\theta}$ corrupted by additive Gaussian noise. The M machines are connected in a star topology to a fusion center, whose goal is to estimate the vector $\boldsymbol{\theta}$ with a low communication budget. In Section 3.1 we formalize the distributed sparse normal means problem and characterize the relevant SNR regime.

Previous works have shown that to achieve the centralized minimax rate for the ℓ_2 risk, the total communication must be high – at least linear in the dimension d. This phenomenon occurs, however, at very weak signals. We show that at signal-to-noise ratios (SNRs) that are sufficiently high – but not enough for recovery by any individual machine – the support of $\boldsymbol{\theta}$ can be correctly recovered with significantly less communication.

In Section 3.2, we present two algorithms for distributed estimation of a sparse mean vector corrupted by either Gaussian or sub-Gaussian noise. We then prove that above certain SNR thresholds, with high probability, these algorithms recover the correct support with total communication that is *sublinear* in the dimension d. Furthermore, the communication decreases *exponentially* as a function of signal strength. Section 3.3 discusses the relation between exactly recovering the support of a vector and estimating it with small ℓ_2 risk, and shows a reduction from the latter to the former with one additional round of sublinear communication.

Section 3.4 elaborates on how our results relate to the lower bounds of Braverman et al. (2016), Han et al. (2018) and Shamir (2014). In Section 3.5, we present simulations that illustrate the performance of our algorithms in different parameter regimes. All proofs appear in Section 3.6 and simulation details in Section 3.7.

Notation. We use the standard $O(\cdot), \Omega(\cdot), \Theta(\cdot)$ notation to hide constants independent of the problem parameters and the notation $\tilde{O}(\cdot)$ to hide terms that are at most polylogarithmic in d. For functions f, g the notations $f = o(g)$ and $f \ll g$ imply that $f/g \to 0$ as $d \to \infty$. The term exact recovery of the support $\mathcal{S}$ with high probability means that an estimator $\hat{\mathcal{S}}$ correctly estimates the support, i.e., $\Pr\left[\hat{\mathcal{S}} = \mathcal{S}\right] \to 1$ as $d \to \infty$ and the number of machines $M = M(d)$ tends to infinity at a suitable rate, as detailed in each theorem. We use the notation $\lceil x \rceil$ for the smallest integer larger than or equal to

[1] The results presented in this chapter have been published in Amiraz et al. (2022). The notations have been changed for consistency with the other chapters of the thesis.

x.

3.1 Problem Setup

Each machine has n independent samples corrupted by additive Gaussian noise of a known level σ, implying that their empirical mean is a sufficient statistic. Thus, we simplify by assuming that each machine i stores a single observation of $\boldsymbol{\theta}$ under an effective noise level of $\sigma_{\text{eff}} = \frac{\sigma}{\sqrt{n}}$. Since σ is known, without loss of generality, we assume that the noise level across all machines is standardized to $\sigma_{\text{eff}} = 1$. This leads to an equivalent model where each machine has a single observation

$$\mathbf{s}_i = \boldsymbol{\theta} + \boldsymbol{\xi}_i, \tag{3.1}$$

where $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\boldsymbol{\theta}$ is exactly K-sparse.

For simplicity we assume that the sparsity level K is known to the fusion center and that $\theta_j \geq 0$ for all $j \in [d]$. However, with slight variations our methods can work when K is unknown or for vectors $\boldsymbol{\theta}$ that have both positive and negative entries. We further assume a lower bound $\theta_{\min}$ on its smallest non-zero coordinate, namely $\theta_j \geq \theta_{\min}$ for all $j \in \mathcal{S} = \{i \,|\theta_i > 0\}$. It will be convenient to use the natural scaling

$$\theta_{\min} = \sqrt{2r \log(d-K)}. \tag{3.2}$$

We focus on the following question: Given a lower bound on the signal-to-noise ratio (SNR) r, how much communication is sufficient for *exact recovery of the support* $\mathcal{S}$ of a K-sparse vector $\boldsymbol{\theta}$ with high probability?

Let us first discuss what is the interesting regime for the SNR parameter r. Recall that for $d - K \gg 1$, the maximum of $d - K$ i.i.d. standard Gaussian random variables is tightly concentrated around $\sqrt{2\log(d-K)}$. At a high SNR $r > 1$, each individual machine can thus exactly recover the support set $\mathcal{S}$ with high probability. Hence, it suffices that only one machine sends $O(K \log d)$ bits to the fusion center. At the other extreme, let $r < \frac{c}{M}$ for a fixed $0 < c < 1$. Here, even in a centralized setting, exact support recovery with high probability is not possible. To see this, note that the empirical mean of all samples is a sufficient statistic, and its effective SNR is $c < 1$. Therefore, with probability tending to 1 as $d \to \infty$, its smallest support entry is smaller than its largest non-support entry. If the index of $\theta_{\min}$ is chosen uniformly at random, then any algorithm would fail to recover the support. Hence, the relevant SNR values are

$$\frac{1}{M} < r < 1. \tag{3.3}$$

In this range, a single machine cannot individually recover the support with high probability. Yet, as we show next, for a large subrange of the SNR values given in Eq. (3.3), exact support recovery by the fusion center is possible with total communication $o(d)$ bits. Furthermore, as r increases towards 1, the total communication decays exponentially fast to $O\left(K \log^{1+c} d\right)$ for an appropriate constant $c > 0$.

3.2 Distributed Algorithms for Sparse Normal Means

We present two one-shot algorithms for the distributed sparse normal means problem and derive non-asymptotic bounds on their performance, namely, their probability of exact recovery and their total communication. For both algorithms, the lower bound r on the SNR is assumed to be known to the center and is used to decide how many machines to communicate with and what messages to send them. We use the notation M_c for the number of contacted machines, which is different in each theorem. For our analysis below, we assume the total number of machines is sufficiently large, in particular $M \geq M_c$, which is a stronger condition than the centralized lower bound $M > 1/r$.

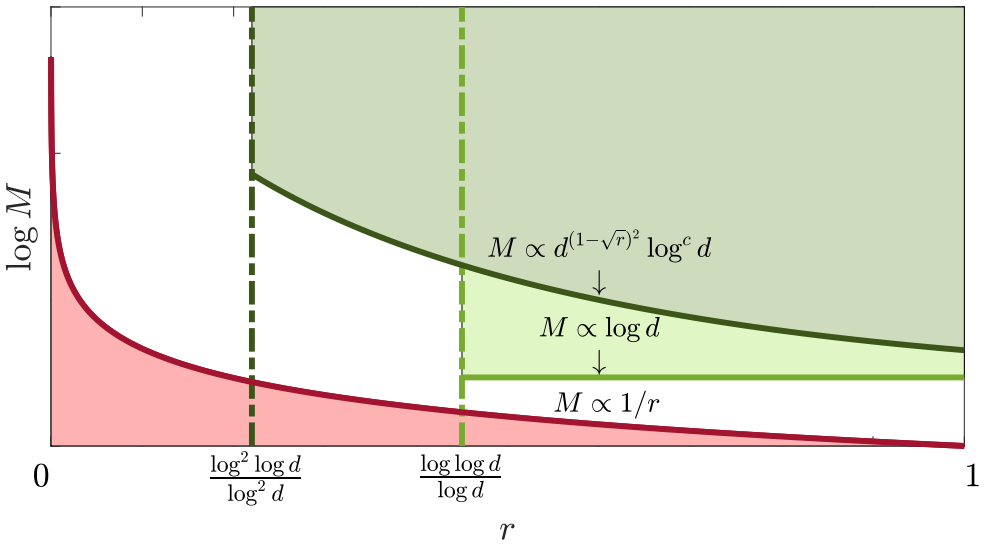


Figure 3.1: Illustration of communication regimes as a function of (r, M), where M is on a logarithmic scale. In the pink area, exact support recovery is information-theoretically impossible even in the centralized setting. In the green areas, our distributed algorithms achieve exact support recovery with communication that is sublinear in d.

In our first algorithm, denoted `Top-L`, the center sends a parameter L to M_c machines. Each contacted machine i sends back a message S_i with the indices of the L highest coordinates of its sample $\mathbf{s}_i$. Our second algorithm is threshold-based; the center sends a threshold t_m to M_c machines, and each contacted machine i sends back all indices j with $s_{i,j} > t_m$. In both two algorithms, the center then estimates the support of $\boldsymbol{\theta}$ by a voting procedure. We prove in Theorems 3.1 and 3.2 that under suitable assumptions, and in particular for a sufficiently high SNR, both algorithms achieve exact support recovery with high probability using sublinear communication. In particular, we show in Theorem 3.2.A that if $r = \Omega\left(\log^{-1}(d-K)\right)$, then with high probability `Thresholding` with $M_c = O(\log d)$ machines and $t_m = \sqrt{2r\log(d-K)}$ recovers the support of the K-sparse vector $\boldsymbol{\theta}$ using $\tilde{O}\left((d-K)^{1-r} + K\right)$ communication bits in expectation. The total communication cost is sublinear in d provided that $K \ll d$ and $r = \Omega\left(\frac{\log\log d}{\log d}\right)$. Moreover, increasing the threshold allows for a tradeoff between M_c and the expected message length per machine. As we show in Theorems 3.2.B and 3.2.C, perhaps counter-intuitively, given more than $O(\log d)$ machines, the fusion center can recover the support using *less* total communication, by setting a higher threshold. Specifically, if $r = \Omega\left(\log^{-2}(d-K)\right)$, then with high probability `Thresholding` with $M_c = \tilde{O}\left((d-K)^{\left(1-\sqrt{r}\right)^2}\right)$ machines and $t_m = \sqrt{2\log(d-K)}$ recovers the support of the K-sparse vector $\boldsymbol{\theta}$ using $\tilde{O}\left(K(d-K)^{\left(1-\sqrt{r}\right)^2}\right)$ communication bits. Note that the resulting total communication cost is sublinear in d, provided that K is at most polylogarithmic in d and $r = \Omega\left(\frac{\log^2\log d}{\log^2 d}\right)$, or equivalently $\theta_{\min} = \Omega\left(\frac{\log\log d}{\sqrt{\log d}}\right)$ We also prove a similar result for the `Top-L` algorithm with $L = K$ in Theorem 3.1. Finally, in Section 3.2.3 we extend some of these results to the case of additive sub-Gaussian noise.

To put our results in context, we illustrate in Figure 3.1 the different communication regimes as a function of the SNR r and the number of machines M for $K = 1$. As discussed above, if $r < \frac{1}{M}$, then even with infinite communication, exact support recovery with high probability is information-

Algorithm 2: `Top-L`

At the fusion center
- **input :** dimension d, number of machines M, SNR r, sparsity level K, parameter L
- **output:** setup message $\mathbf{q}$
- if $L = K = 1$, then M_c is given by Eq. (3.4), otherwise it is given by Eq. (3.7)
- send message $\mathbf{q}$ that contains the value L to each of the first M_c machines

At each machine $i = 1, \ldots, M_c$
- **input :** setup message $\mathbf{q}$, sample $\mathbf{s}_i$
- **output:** message S_i to center
- compute permutation α_i that sorts the vector $\mathbf{s}_i$ in descending order, $s_{i,\alpha_i(1)} \geq \cdots \geq s_{i,\alpha_i(d)}$
- send to the center the L indices with the largest coordinates, $S_i = \{\alpha_i(1), \ldots, \alpha_i(L)\}$

At the fusion center
- **input :** messages $S_1, \ldots, S_{M_c}$, sparsity level K
- **output:** estimated support $\hat{\mathcal{S}}$
- for each coordinate $j \in [d]$, let the set of votes it received be $\mathbf{V}_j = \{i \in [M_c] : j \in S_i\}$ and let their number be $\mathbf{v}_j = |\mathbf{V}_j|$
- compute permutation π that sorts the indices by descending number of votes, $\mathbf{v}_{\pi(1)} \geq \cdots \geq \mathbf{v}_{\pi(d)}$
- return $\hat{\mathcal{S}} = \{\pi(1), \ldots, \pi(K)\}$

theoretically impossible. The corresponding (r, M) values are in the pink area below the red curve which delineate the relation $r \cdot M = 1$. By our Theorems 3.1 and 3.2, exact recovery with sublinear communication is possible in the light green and dark green areas. In the white area, distributed exact support recovery is possible using communication that is at least linear in d. An example of a recovery scheme in this range is to send the entire sample (up to a quantization error). It remains an open question whether exact support recovery with sublinear communication is possible for (r, M) values in the white area.

3.2.1 `Top-L` Algorithm

In the `Top-L` algorithm, the center uses its knowledge of the parameters d, M, r, K to determine the number of machines M_c to contact, and sends them a parameter $L \in \mathbb{N}$. The i-th contacted machine then sends a message S_i consisting of the L indices with the largest coordinates of its vector $\mathbf{s}_i$. Given the messages $S_1, \ldots, S_{M_c}$, the fusion center counts how many votes each index received and estimates the support to be the K indices with the highest number of votes. Voting ties can be broken arbitrarily. This scheme is outlined in Algorithm 2. Its total communication cost is $B = O(LM_c \log d)$ bits.

Remark 3.1. *The above description assumes that the fusion center knows the sparsity level K. However the following simple variant can handle a case where only an upper bound $K_{\max} \geq K$ is known. In this case, the number of contacted machines M_c is determined using $K_{\max}$ instead of K, and each contacted machine sends its top $L \geq K_{\max}$ indices to the fusion center. The center then estimates the support as the set of indices that received more votes than a suitable threshold $t_c(d)$.*

We prove that for sufficiently high SNR, the `Top-L` algorithm recovers the exact support of $\boldsymbol{\theta}$ with high probability. To ease the presentation and highlight the main ideas of the proof, we first analyze the case $L = K = 1$ and then extend the analysis to general $L \geq K \geq 1$. The proofs of the theorems stated below appear in Appendix 3.6.1.

Motivated by the required number of machines for proving Theorem 3.1.A, we define the quantity

$$M_0(d,r) = \left\lceil \max\left\{1, \frac{\sqrt{2\pi}e\left(2\left(1-\sqrt{r}\right)^2\log d + 1\right)}{\left(1-\sqrt{r}\right)\sqrt{2\log d}} \cdot d^{\left(1-\sqrt{r}\right)^2}\right\} \cdot 8\log d \right\rceil. \tag{3.4}$$

Notice that for any fixed SNR $r < 1$, $M_0(d,r)$ is *sublinear* in d, and up to polylogarithmic terms it is proportional to $d^{\left(1-\sqrt{r}\right)^2} < d$. The following theorem provides a support recovery guarantee in the setting $K = L = 1$.

Theorem 3.1.A. *Assume $r < 1$ and that $M_0(d,r) \le \min\{M, d\}$. Then, if the center contacts $M_c = M_0$ machines, the `Top-L` algorithm with $L = 1$ recovers the support of a 1-sparse vector $\boldsymbol{\theta}$ with probability at least $1 - d^{-1} - e^3 d^{-3}$. Its total communication is $O(M_0 \log d)$ bits.*

Several insights follow from Theorem 3.1.A. First, recall that for any $r < 1$ no machine can successfully recover the support of $\boldsymbol{\theta}$ on its own. Yet, for $d \gg 1$ and for any fixed $r < 1$, as implied by the theorem, the fusion center can recover the support of $\boldsymbol{\theta}$ by communicating with only $M_0(d,r)$ machines, receiving from each machine its own mostly inaccurate estimate of the support. Second, as the SNR lower bound r increases towards 1, the algorithm needs to contact fewer machines and thus less communication to succeed with high probability. Moreover, by Eq. (3.4), $M_0(d,r)$ decreases *exponentially* fast with r. Lastly, for a fixed r the required number of machines $M_0(d,r)$ and thus the total communication cost both increase sublinearly with d.

Next, we consider the more general case where the unknown vector $\boldsymbol{\theta}$ is exactly sparse with sparsity level at most K, and its support is estimated by the `Top-L` algorithm with parameter $L \ge K$. To this end, we define the auxiliary quantities

$$a = a(K,L,d) = \sqrt{2\log\frac{d-K}{L-K+1}}, \tag{3.5}$$

$$b = b(K,L,d,r) = a - \sqrt{2r\log(d-K)}, \tag{3.6}$$

and the quantity

$$M_{K,L}(d,r) = \left\lceil \max\left\{1, \frac{4\sqrt{2\pi}\left(b^2+1\right)}{b} \cdot (d-K)^{\left(\sqrt{1-\frac{\log(L-K+1)}{\log(d-K)}}-\sqrt{r}\right)^2}\right\} \cdot 8\log d \right\rceil. \tag{3.7}$$

The following theorem provides a support recovery guarantee in this setting.

Theorem 3.1.B. *Assume $r < 1$ and that $M_{K,L}(d,r) \le \min\left\{M, \frac{d-K}{L}\right\}$. Then, if the center contacts $M_c = M_{K,L}$ machines, the `Top-L` algorithm with $K \le L < (d-K)/2$ recovers the support of a K-sparse vector $\boldsymbol{\theta}$ with probability at least $1 - Kd^{-1} - e^3 d^{-3}$ using $O(LM_{K,L}\log d)$ communication bits.*

While the expressions in Theorem 3.1.B are more involved than those of Theorem 3.1.A, similar insights to those mentioned above continue to hold. In addition, it is easy to check that the total communication is monotonically increasing in L. Hence, if K is a-priori known, it is best to choose $L = K$. In this case, the `Top-L` algorithm incurs a total communication cost of $O\left(K \cdot (d-K)^{\left(1-\sqrt{r}\right)^2}\log^{2.5} d\right)$, which is sublinear in d provided that K is at most polylogarithmic in d and $r = \Omega\left(\frac{\log^2\log d}{\log^2 d}\right)$.

Remark 3.2. *One can consider a variant of the algorithm that sends $L < K$ randomly selected indices out of the top K. In such a variant, the message length per machine is shorter, and hence in general, the center would need to contact more machines for exact support recovery with high probability. A theoretical analysis of this scheme is beyond the scope of the current manuscript.*

Algorithm 3: `Thresholding`

At the fusion center
- **input :** dimension d, number of machines M, SNR r, sparsity level K
- **output:** setup message $\mathbf{q}$
- depending on M, calculate M_c and t_m by their expressions in Theorem 3.2.A or 3.2.B or 3.2.C
- send message $\mathbf{q} = \text{Trunc}(t_m, \lfloor \log_2 t_m \rfloor, \lceil \log_2 d \rceil)$ to each of the first M_c machines

At each machine $i = 1, \ldots, M_c$
- **input :** setup message $\mathbf{q}$, sample $\mathbf{s}_i$
- **output:** message S_i to center
- construct threshold $\hat{t}_m = \text{Approx}(\mathbf{q}, \lfloor \log_2 t_m \rfloor, \lceil \log_2 d \rceil)$
- let $S_i = \left\{ j \in [d] : s_{i,j} > \hat{t}_m \right\}$
- send S_i to center

At the fusion center
- **input :** messages $S_1, \ldots, S_{M_c}$, sparsity level K
- **output:** estimated support $\hat{\mathcal{S}}$
- for each coordinate $j \in [d]$, let the set of votes it received be $\mathbf{V}_j = \{i \in [M_c] : j \in S_i\}$ and let their number be $\mathbf{v}_j = |\mathbf{V}_j|$
- compute permutation π that sorts the indices by descending number of votes, $\mathbf{v}_{\pi(1)} \geq \cdots \geq \mathbf{v}_{\pi(d)}$
- return $\hat{\mathcal{S}} = \{\pi(1), \ldots, \pi(K)\}$

3.2.2 `Thresholding` Algorithm

In our second algorithm, the fusion center chooses a threshold $t_m = t_m(d, r, M, K)$ and sends (a truncated binary representation of) it to a subset of the machines $M_c = M_c(d, r, M, K) \leq M$. Each contacted machine i sends back all indices j such that $s_{i,j} > t_m$. Similarly to the `Top-L` algorithm, given the messages $S_1, \ldots, S_{M_c}$ and the sparsity level K, the fusion center estimates the support as the K indices with the highest number of votes. Voting ties can be broken arbitrarily. The scheme is outlined in Algorithm 3. If instead of the sparsity level K only an upper bound on it $K_{\max} \geq K$ is known, and $K_{\max} \ll d$, then the fusion center can set t_m and M_c by approximating $d - K \approx d$. In addition, the center estimates the support as outlined in Remark 3.1.

The `Thresholding` algorithm has several desirable properties. First, it is simple to implement in a distributed setting. Second, in the centralized setting, thresholding algorithms were shown to be optimal in various aspects (see Section 3.3 for further details). Third, adjusting the threshold allows for a tradeoff between the number of contacted machines and the expected message length per machine. Notice that if the SNR is sufficiently high, but still $r < 1$, i.e., not high enough for recovery by any individual machine, there may not even be a need to contact all machines to recover the support. By the same logic, when the SNR is lower, one can lower the threshold. Of course, this would incur a higher communication cost. Hence, since the fusion center knows both r and M, it can set an optimal threshold t_m and send it only to $M_c \leq M$ machines, which ensures exact support recovery with high probability at minimal communication cost (among all possible thresholds).

To complete the description of the algorithm, we now describe our approximation of a real number by a finite amount of bits. Recall that the scientific binary representation of a number $x \in \mathbb{R}$ consists of a bit representing its sign and bits $\{b_j\}_{j \in \mathbb{Z}}$, such that $|x| = \sum_{j=-\infty}^{\lfloor \log_2 |x| \rfloor} b_j 2^j$. One can approximate x by truncating its binary representation at a predetermined precision level. Specifically, given two parameters $U, P \in \mathbb{N}$, let the procedure $\mathbf{q} = \text{Trunc}(x, U, P)$ output a truncated binary representation of x of length $U + P + 2$ such that $\mathbf{q} = (\mathbb{1}\{x \geq 0\}, b_{-P}, \ldots, b_U)$. Given $\mathbf{q}$, let the procedure $\hat{x} = \text{Approx}(\mathbf{q}, U, P)$ construct an approximation for x, given by $\hat{x} = \text{sign}(x) \cdot \sum_{j=-P}^{U} b_j 2^j$. If $U \geq \lfloor \log_2 |x| \rfloor$, then $\hat{x}$ and x consist of the same bits up to the P-th bit after the binary dot, and thus the resulting

approximation error is bounded by $|\hat{x} - x| < 2^{-P}$. This scheme is a variant of Szabo and van Zanten (2020, Algorithm 1).

In our analysis we assume that $\theta_{\max} = \max_{j \in \mathcal{S}} \theta_j$ is at most polynomial in d. Thus taking $U, P = O(\log d)$ ensures that with high probability all quantities of interest are approximated up to $d^{-O(1)}$ error. In addition, since P, U only depend on d and on the bound $\theta_{\max}$, they can be set in advance without communication.

We analyze the performance of the `Thresholding` algorithm in three regimes, in terms of the number of contacted machines M_c: small, intermediate, and large (clearly under the constraint that $M \geq M_c$). For each regime, we derive a different threshold $t_m = t_m(d, r, M)$, where the SNR parameter r and sparsity level K are assumed to be known. In the small M_c regime, considered in Theorem 3.2.A, the number of contacted machines is logarithmic in d. The corresponding threshold t_m given by (3.8) is relatively small. In the intermediate regime, considered in Theorem 3.2.B, all M machines are contacted and the threshold t_m, given by Eq. (3.11), increases as a function of M. Finally, when the number of available machine is sufficiently large, as described in Theorem 3.2.C, the center contacts only a subset M_c of all machines, where the value of M_c is chosen to minimize the total communication, while still achieving exact support recovery with high probability. The proofs appear in Appendix 3.6.2.

Theorem 3.2.A. *Assume that $d \geq 16$ and $M \geq 16 \log d$. Further assume $\frac{\log 5}{\log(d-K)} < r < 1$. Then, with probability at least $1 - (K+1)/d$, `Thresholding` with $M_c = \lceil 16 \log d \rceil$ and*

$$t_m = \sqrt{2r \log(d - K)} \tag{3.8}$$

recovers the support of the K-sparse vector $\boldsymbol{\theta}$ using

$$O\left((d-K)^{1-r} r^{-0.5} \log^{1.5} d + K \log^2 d\right) \tag{3.9}$$

communication bits in expectation.

The communication cost (3.9) is sublinear in d for all $r > \frac{2 \log \log d}{\log(d-K)}$ and $K \ll d/\log^2 d$. Note that in the above theorem, the number of contacted machines is fixed at $16 \log d$ and correspondingly, the threshold does not depend on the total number of machines M. The next theorem shows that contacting all machines with a higher threshold that depends on the total number of machines, can lead to exact support recovery with even less communication than (3.9).

Theorem 3.2.B. *Let $d \geq 15$ and assume that $32\sqrt{e\pi} \log^{1.5} d \leq M \leq d$. Further assume $r < 1$ and that*

$$r > \frac{\left(\sqrt{2 \log \frac{5M}{\sqrt{2\pi} 4 \log d}} - \sqrt{2 \log \frac{M}{32\sqrt{\pi} \log^{1.5} d}} + \frac{1}{d}\right)^2}{2 \log(d-K)}. \tag{3.10}$$

Then, with probability at least $1 - (K+1)/d$, `Thresholding` with $M_c = M$ and

$$t_m = \sqrt{2r \log(d-K)} + \sqrt{2 \log \frac{M}{32\sqrt{\pi} \log^{1.5} d}} \tag{3.11}$$

recovers the support of the K-sparse vector $\boldsymbol{\theta}$ using

$$O\left(KM \log d + (d-K)^{1-r} e^{-2\sqrt{r \log(d-K) \log \frac{M}{32\sqrt{\pi} \log^{1.5} d}}} \log^{2.5} d\right) \tag{3.12}$$

communication bits in expectation.

It is interesting to study the behavior of the total communication cost in Eq. (3.12). The first term increases with M, whereas the second term decreases with M. It is easy to show that the total

communication cost is minimized at $M_{\text{opt}} = \tilde{O}\left((d-K)^{\left(1-\sqrt{r}\right)^2}\right)$. This leads to a perhaps counter-intuitive result, that in the range $O\left(\log^{1.5} d\right) < M < M_{\text{opt}}$, as the number of machines increases exact recovery is possible with *less* total communication. Once the number of available machines is larger than M_{opt}, there is no benefit in contacting all machines. In terms of total communication, it is best to simply contact $M_c = M_{\text{opt}}$ of them, as stated in the following theorem.

Theorem 3.2.C. *Assume that* $d - K \geq 20$ *and* $\left(\frac{\log 10}{\log(d-K)}\right)^2 < r < 1$. *Let*

$$M_c = \left\lceil \frac{8\sqrt{2\pi}\left(\left(1-\sqrt{r}\right)^2 2\log\left(d-K\right)+1\right)}{\left(1-\sqrt{r}\right)\sqrt{2\log\left(d-K\right)}} \left(d-K\right)^{\left(1-\sqrt{r}\right)^2} \log d \right\rceil, \tag{3.13}$$

and assume that $M \geq M_c$. *Then, with probability at least* $1-(K+1)/d$, `Thresholding` *with*

$$t_m = \sqrt{2\log\left(d-K\right)} \tag{3.14}$$

and M_c *machines recovers the support of the* K*-sparse vector* $\boldsymbol{\theta}$ *using*

$$O\left(K\left(d-K\right)^{\left(1-\sqrt{r}\right)^2} \log^{2.5} d\right) \tag{3.15}$$

communication bits in expectation.

Let us now compare the `Top-L` and `Thresholding` algorithms, in terms of communication cost and recovery guarantees. By Theorems 3.1.B and 3.2.C, with appropriately set parameters the algorithms exhibit qualitatively similar performances for high SNR and large number of machines M. The main differences between the two algorithms occur when M is small, for example logarithmic in d. If the SNR is low, for example $r = O\left(\log^{-2} d\right)$, then the `Top-L` algorithm with $L = K$ might fail to recover the support, whereas, by Theorem 3.2.A, the `Thresholding` algorithm succeeds to recover it. However, substituting $r = O\left(\log^{-2} d\right)$ in Eq. (3.15) results in total communication cost superlinear in d. In contrast, if the SNR is slightly higher, namely $r = O\left(\frac{\log^2 \log d}{\log^2 d}\right)$, then by Theorems 3.1.B and 3.2.A, with high probability both algorithms succeed, and the `Top-L` algorithm with $L = K$ incurs less total communication cost than the `Thresholding` algorithm. However, the `Thresholding` algorithm is more robust in the following sense. If the sparsity level K is fixed and the center only knows an upper bound on it $K_{\max} = cK$ for $c > 1$, then the `Top-L` algorithm with $L = K_{\max}$ incurs a communication cost that is linear in c, while the `Thresholding` algorithm incurs a communication cost that is roughly the same as when $c = 1$.

3.2.3 Extension to sub-Gaussian noise

Let us outline in this section how some of our results above can be extended to the case of additive sub-Gaussian noise. Specifically, we assume that each machine i has n i.i.d. samples of the form $\mathbf{s}_i^j = \boldsymbol{\theta} + \boldsymbol{\xi}_i^j$ for $j = 1, \ldots, n$, where the mean vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is exactly K-sparse and all noise coordinates $\xi_{i,k}^j$ are i.i.d. sub-Gaussian random variables with parameter σ^2 (also known as the variance proxy). We assume all noise coordinates have the same variance $\tilde{\sigma}^2 = E\left[\left(\xi_{i,k}^j\right)^2\right]$ and finite third absolute moment $\eta = E\left[|\xi_{i,k}^j|^3\right]$. It is easy to show that $\sigma^2 \geq \tilde{\sigma}^2$ (Rigollet, 2015, Lemma 1.4). In our analysis, we shall assume that for some fixed $0 < \lambda \leq 1$,

$$\tilde{\sigma}^2 \geq \lambda^2 \cdot \sigma^2. \tag{3.16}$$

To account for having n samples per machine, we generalize the definition of the scaling parameter r as follows

$$\theta_{\min} = \frac{\tilde{\sigma}}{\sqrt{n}}\sqrt{2r\log(d-K)}. \tag{3.17}$$

Denote by `Thresholding`* the following scheme, where each contacted machine i computes the normalized empirical mean vector

$$\tilde{\mathbf{s}}_i = \frac{1}{\tilde{\sigma}\sqrt{n}}\sum_{j=1}^{n}\mathbf{s}_i^j. \tag{3.18}$$

Accordingly, each machine i computes its message as

$$S_i = \left\{j \in [d] : \tilde{s}_{i,j} > \hat{t}_m\right\}. \tag{3.19}$$

Note that the effective signal strength in each machine, corresponding to its sample $\tilde{\mathbf{s}}_i$, is $\frac{\sqrt{n}\theta_{\min}}{\tilde{\sigma}} = \sqrt{2r\log(d-K)}$, which matches Eq. (3.2) above.

Given sufficiently many samples per machine, results similar to those we proved for Gaussian noise hold for the case of sub-Gaussian noise. As an example, the following theorem is a variant of Theorem 3.2.C for the `Thresholding` algorithm. Its proof appears in Appendix 3.6.3. A similar result can be derived for the `Top-L` algorithm.

Theorem 3.3. *Consider exact support recovery with n samples per machine, corrupted by additive sub-Gaussian noise as described above. Assume that the SNR r satisfies the following, for a suitable universal constant $C > 0$*

$$\Omega\left(\frac{1}{\log(d-K)}\right) < \left(1-\sqrt{r}\right)^2 < C\lambda^2, \tag{3.20}$$

that the sample size in each machine satisfies

$$n = \Omega\left(\frac{\eta^2}{\tilde{\sigma}^6}(1-\sqrt{r})^6\log^3(d-K)\right). \tag{3.21}$$

and that $d-K$ is sufficiently large. Let $M_c = O\left((d-K)^{\left(1-\sqrt{r}\right)^2}\log^{1.5} d\right)$ and assume that $M \geq M_c$. Then, with probability at least $1 - O\left(\frac{K}{d}\right)$, the `Thresholding`* algorithm with

$$t_m = \sqrt{2\log(d-K)} \tag{3.22}$$

and M_c machines recovers the support of the K-sparse vector $\boldsymbol{\theta}$ using

$$O\left(K\,(d-K)^{\left(1-\sqrt{r}\right)^2}\log^{2.5} d\right) \tag{3.23}$$

communication bits in expectation.

The proof of Theorem 3.3 uses both lower bounds and upper bounds on the tail probability of the noise. For the tail lower bound, we use a result of Nagaev (2002), which requires a minimal number of samples per machine, as stated in Eq. (3.21). Note that this requirement is rather mild. For r bounded away from one, only a polylogarithmic in d number of samples per machine suffices. For the lower bound to hold, we also require in (3.20) that the SNR parameter r cannot be arbitrarily close to 1, as otherwise n could tend to zero in Eq. (3.21). In contrast, such an upper bound on r does not appear in Theorem 3.2.C.

Another key difference from Theorem 3.2.C is a strict lower bound on the SNR r, as stated in Eq. (3.20), which implies $r > \left(1-\sqrt{C}\lambda\right)^2$. The reason for this is a rather crude upper tail probability

approximation we apply in our proof, which uses the sub-Gaussian property of the noise. We remark that by the central limit theorem, if we require a much larger number of samples per machine with d and K fixed, then results similar to Theorem 3.2.C may be derived, even without assuming sub-Gaussianity of the noise. In particular, with sufficient number of samples per machine, the lower bound on the SNR r will be the same as in Theorem 3.2.C.

3.3 Sublinear Distributed Algorithms with Small ℓ_2 Risk

In the previous section we considered distributed estimation of the support of $\boldsymbol{\theta}$. Another common task is to estimate the vector $\boldsymbol{\theta}$ itself, with both small ℓ_2 risk and low total communication. We show that this can be achieved with only a single additional round of communication. Furthermore, under certain parameter regimes, specifically $KM \ll \frac{d}{\log d}$, the resulting estimate achieves the centralized ℓ_2 risk, with sublinear total communication. The proof of this result is based on the fact that both of our algorithms achieve exact support recovery with high probability. We thus first discuss the relation between support recovery and ℓ_2 risk, as well as lower bounds for the centralized minimax risk.

3.3.1 On Exact Support Recovery and ℓ_2 Risk

Let us first briefly discuss estimation of $\boldsymbol{\theta}$ in a centralized setting with M samples and noise level σ. Without any assumptions on the vector $\boldsymbol{\theta}$, the empirical mean $\bar{\mathbf{s}} = \frac{1}{M}\sum_i \mathbf{s}_i$ is a rate-optimal estimator. When $\boldsymbol{\theta}$ is assumed to be sparse, various works suggested and theoretically analyzed the set of diagonal estimators $\mathcal{O}_{\text{diag}}$. An estimator $\hat{\boldsymbol{\theta}} \in \mathcal{O}_{\text{diag}}$ has the form $\hat{\theta}_j = a_j(\bar{s}_j)\,\bar{s}_j$ for all $j \in [d]$, where each $a_j(\cdot)$ is a scalar function. For further details see for example Mallat (1999, Chapter 11).

Projection oracle risk. In analyzing the lowest risk achievable in the set $\mathcal{O}_{\text{diag}}$, a key notion is the *projection oracle risk*, defined as the smallest expected ℓ_2 error of a diagonal projection estimator $\hat{\boldsymbol{\theta}}^{\text{oracle}}$ but with additional prior knowledge of $\boldsymbol{\theta}$, such that $\hat{\theta}_j^{\text{oracle}} = a_j(\theta_j)\,\bar{s}_j$ and $a_j \in \{0,1\}$. It is easy to show that $\hat{\theta}_j^{\text{oracle}} = \bar{s}_j \cdot \mathbf{1}(|\theta_j| > \sigma/\sqrt{M})$. Its corresponding risk is

$$R_{\text{oracle}}(\boldsymbol{\theta}) = \mathrm{E}\left[\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{\text{oracle}}\right\|^2\right] = \sum_{j=1}^{d} \min\left\{\frac{\sigma^2}{M}, \theta_j^2\right\} \le \frac{K\sigma^2}{M}. \tag{3.24}$$

Note that the projection oracle is not a realizable estimator, as it relies on knowledge of the underlying $\boldsymbol{\theta}$ for support recovery. However, the oracle risk provides a lower bound for the risk of any diagonal estimator. Also note that given a lower bound on the SNR, of the form $\min_{j\in\mathcal{S}}|\theta_j| > \sigma/\sqrt{M}$, the oracle risk is $R_{\text{oracle}}(\boldsymbol{\theta}) = K\sigma^2/M$.

Centralized lower bound. Donoho and Johnstone (1994, Theorem 3) proved the following lower bound on the asymptotic minimax rate among all diagonal estimators,

$$\lim_{d\to\infty} \inf_{\hat{\boldsymbol{\theta}}\in\mathcal{O}_{\text{diag}}} \sup_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{\mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2]}{\frac{\sigma^2}{M} + R_{\text{oracle}}(\boldsymbol{\theta})} \frac{1}{2\log d} = 1. \tag{3.25}$$

Moreover, they proved that thresholding at a suitable level achieves this minimax rate.

In the result above, no assumptions are made neither regarding the sparsity of $\boldsymbol{\theta}$, nor on its SNR or equivalently on $\theta_{\min}$. Indeed, the proof of (3.25) relies on a construction of vectors $\boldsymbol{\theta}$ with $\log d$ coordinates having values slightly smaller than $\frac{\sigma}{\sqrt{M}}\sqrt{2\log d}$, namely with a low SNR. Thus, it cannot be used as a lower bound for the centralized minimax rate in our setting. In fact, if $\boldsymbol{\theta}$ is K-sparse and $\theta_{\min}$ is sufficiently high, then asymptotically as $d \to \infty$ with $\frac{KM\log d}{d} \to 0$, the risk of a suitable thresholding estimator is equal to $R_{\text{oracle}}(\boldsymbol{\theta})\,(1 + o(1))$. The reason is that in this case one can achieve exact support recovery with high probability. We now prove a similar result for the distributed setting.

Algorithm 4: Protocol Π

At the fusion center
  input : estimated support set $\hat{\mathcal{S}}$
  output: setup message $\mathbf{q}$
  send message $\mathbf{q}$ which contains the set $\hat{\mathcal{S}}$ to each of the M machines
At each machine $m = 1, \ldots, M$
  input : setup message $\mathbf{q}$, sample $\mathbf{s}_i$, precision parameters U, P
  output: message $\mathbf{w}_i$ to center
  for each $k \in \hat{\mathcal{S}}$, calculate $\mathbf{w}_{i,k} = \mathrm{Trunc}(s_{i,k}, U, P)$
  send to center $\mathbf{w}_i = \left\{\mathbf{w}_{i,k} : k \in \hat{\mathcal{S}}\right\}$
At the fusion center
  input : messages $\mathbf{w}_1, \ldots, \mathbf{w}_M$
  output: estimated vector $\hat{\boldsymbol{\theta}}$
  for each $i \in [M]$ and each $k \in \hat{\mathcal{S}}$, reconstruct $z_{i,k} = \mathrm{Approx}(\mathbf{w}_{i,k}, U, P)$
  for each $k \in \hat{\mathcal{S}}$, calculate the mean $\bar{z}_k = \frac{1}{M}\sum_{i\in[M]} z_{i,k}$
  return $\hat{\boldsymbol{\theta}}^{\Pi}$ where $\hat{\theta}^{\Pi}_j = \bar{z}_j \cdot \mathbb{1}\left\{j \in \hat{\mathcal{S}}\right\}$

3.3.2 The ℓ_2 Risk of `Top-L` and `Thresholding`

The `Top-L` and `Thresholding` algorithms described in Section 3.2, output an estimated support set $\hat{\mathcal{S}}$. As we describe now, using an additional round of communication, the center can also estimate the vector $\boldsymbol{\theta}$ itself. In particular, we consider the following protocol, denoted Π: First, the center sends the indices of $\hat{\mathcal{S}}$ to all M machines. Then, each machine i replies with the binary representation $\mathbf{w}_{i,k} = \mathrm{Trunc}(s_{i,k}, U, P)$ for the estimated support coordinates $k \in \hat{\mathcal{S}}$, for appropriately chosen $U, P = O(\log d)$. The center computes $z_{i,k} = \mathrm{Approx}(\mathbf{w}_{i,k}, U, P)$ and calculates the empirical mean $\bar{z}_k = \frac{1}{M}\sum_{i\in[M]} z_{i,k}$. Finally, the center estimates $\boldsymbol{\theta}$ as follows

$$\hat{\theta}^{\Pi}_j = \bar{z}_j \cdot \mathbb{1}\left\{j \in \hat{\mathcal{S}}\right\}.$$

The scheme is outlined in Algorithm 4.

The following corollary shows that applying Π to the set $\hat{\mathcal{S}}$ computed by one of our algorithms yields an estimator $\hat{\boldsymbol{\theta}}^{\Pi}$ with ℓ_2 risk $R_{\Pi} = \mathrm{E}\left[\left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{\Pi}\right\|^2\right]$ which is near-oracle. Its proof appears in Appendix 3.6.4.

Corollary 3.1. *Let $d \geq 5$. Assume that the conditions of Theorem 3.1.B hold and let $\hat{\mathcal{S}} \subset [d]$ be the estimate computed by the `Top-L` algorithm. In addition, assume that $\theta_{\max} < d^{\gamma}$ for $\gamma > 0$. Then, the ℓ_2 risk of $\hat{\boldsymbol{\theta}}^{\Pi}$ with precision parameters $P = \lceil \log_2 d \rceil$ and $U = \left\lfloor \log_2(d^{\gamma} + \sqrt{4(\gamma+1)\log d}) \right\rfloor$ is bounded as follows*

$$R_{\Pi} \leq \frac{K}{M}\left(1 + d^{-1} + d^{-2}\right) + \frac{2K\theta^2_{\min}}{d}. \tag{3.26}$$

The expected total communication cost of Π is $O(KM \log d)$. Thus, in an asymptotic setting where $K, M, d \to \infty$ with $\frac{KM\log d}{d} \to 0$, the protocol Π has sublinear expected communication cost and its ℓ_2 risk is $R_{oracle}(\boldsymbol{\theta})(1 + o(1))$.

If we assume that the conditions of either Theorem 3.2.A, Theorem 3.2.B or Theorem 3.2.C hold, then essentially the same proof shows that a two-round algorithm that first estimates the support of $\boldsymbol{\theta}$ by the respective `Thresholding` algorithm and then applies protocol Π as a second round to estimate the vector $\boldsymbol{\theta}$ itself can achieve near-oracle ℓ_2 risk as well. Similarly, the expected total communication cost is sublinear in d if $KM \ll \frac{d}{\log d}$.

Remark 3.3. *An interesting question is whether one round of sublinear communication suffices to estimate $\boldsymbol{\theta}$ with near-oracle ℓ_2 risk. A natural candidate solution is a variant of the* `Thresholding` *algorithm where each machine sends its indices that pass the threshold t_m and their corresponding coordinate values truncated to $O(\log d)$ precision. If the number of machines is large, then our analysis suggests that only a small fraction of the machines would send messages to the center, which would result in high risk compared to the centralized risk. However, if $M = O(\log d)$, then by our analysis of Theorem 2.A, at least half of the machines would send to the center each of the support elements, which should be sufficient information for estimating $\boldsymbol{\theta}$ with near-centralized rate. Note that the sent coordinate values are biased, and thus simply computing their mean would result in an over-estimate of each θ_j. Therefore, the analysis of Theorem 3.2.A and Corollary 3.1 cannot be applied directly to this one-round variant. We believe that a more delicate fusion technique should result in estimating $\boldsymbol{\theta}$ with small ℓ_2 risk, but we do not investigate this further due to our focus on support recovery.*

3.4 Relation to Previous Works

In the context of the distributed sparse normal means problem, several works derived communication lower bounds for exact support recovery and for the ℓ_2 risk of any distributed scheme with total communication budget B. We now describe in further detail three closely related previous works and their relation to our results.

3.4.1 Lower Bounds on the ℓ_2 Risk in Distributed Settings

Braverman et al. (2016, Theorem 4.5) and Han et al. (2018, Theorem 7) derived communication lower bounds for the distributed minimax ℓ_2 risk of estimating a K-sparse vector $\boldsymbol{\theta}$. Their results imply that to achieve the centralized minimax rate, the required total communication by any distributed algorithm must be at least linear in d. However, their proof relies on sparse vectors with a very low signal-to-noise ratio. In contrast, in scenarios where the SNR is sufficiently high these bounds do not apply, and as our theoretical analysis reveals, both exact support recovery and rate-optimal ℓ_2 risk are achievable with sublinear communication, provided that $KM \ll \frac{d}{\log d}$.

In more detail, Braverman et al. (2016) considered blackboard communication protocols, where all machines communicate via a public blackboard and the total number of bits that they can write in the transcript is bounded by B. Denote the set of estimators whose inputs are blackboard communication protocols by $\mathcal{O}_{\text{BB}}$ and the set of all K-sparse d dimensional vectors as $\mathcal{S}_{d,K}$. Their Theorem 4.5 states that if $d > 2K$, then the ℓ_2 risk of any distributed estimator in this model is lower bounded by

$$R_{\text{BB}} = \inf_{\hat{\boldsymbol{\theta}} \in \mathcal{O}_{\text{BB}}} \sup_{\boldsymbol{\theta} \in \mathcal{S}_{d,K}} \mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2] \geq \Omega\left(\min\left\{\sigma^2 K, \max\left\{\sigma^2 K \frac{d}{B}, \frac{\sigma^2 K}{M}\right\}\right\}\right). \tag{3.27}$$

Note that if the total communication B is sublinear in d, then the above simplifies to $\Omega(\sigma^2 K)$, which is significantly larger than the centralized minimax rate, Eq. (3.25). The reason is that R_{BB} involves a supremum over all K-sparse vectors, without any assumptions on their SNR. Indeed, in their analysis a vector $\boldsymbol{\theta}$ with extremely low SNR is used to prove the bound.

Han et al. (2018) considered a more restricted case of one-shot protocols where each of the M machines has a budget of at most b bits that are sent simultaneously to the center, i.e. $B = Mb$. Denote by $\mathcal{O}_{\text{sim}}$ the set of estimators based on such protocols. Their Theorem 7 states that if $d \geq 2K$ and $M \geq \frac{Kd^2 \log(d/K)}{\min\{b^2, d^2\}}$, then the risk is lower bounded by

$$R_{\text{sim}} = \inf_{\hat{\boldsymbol{\theta}} \in \mathcal{O}_{\text{sim}}} \sup_{\boldsymbol{\theta} \in \mathcal{S}_{d,K}} \mathbb{E}[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2] \geq \Omega\left(\frac{\sigma^2 K}{M} \cdot \log(d/K) \cdot \max\left\{\frac{d}{b}, 1\right\}\right). \tag{3.28}$$

Two remarks are in place here. First, our protocol Π described in Section 3.3 requires two rounds of two-way communication between the center and the machines instead of one-round of one-way

communication from the machines to the center. In addition, during the first round a subset of the machines may not be contacted and thus remain idle. For these reasons our estimator $\hat{\boldsymbol{\theta}}^{\Pi}$ is not in $\mathcal{O}_{\text{sim}}$, and thus the above lower bound does not apply to it.

Second, the lower bound (3.28) does not apply for estimators in $\mathcal{O}_{\text{sim}}$ with sublinear communication, since the condition on M translates to requiring $B \geq d$. To show this, notice that if $B < d$ then in particular each machine has a sublinear communication budget, i.e., $b = d^{\beta}$ for $0 < \beta < 1$. The requirement on the number of machines then translates to $M \geq Kd^{2-2\beta} \log(d/K)$, and thus the total communication budget is $B = Mb \geq Kd^{2-\beta} \log(d/K)$, which is superlinear in d for all $\beta < 1$.

3.4.2 Lower Bound on Exact Support Recovery in a Distributed Setting

Shamir (2014) proved lower bounds for several distributed estimation problems under communication constraints. The considered setting is of distributed (b, n, M) protocols whereby each machine $i \in [M]$ constructs a message S_i of length at most b bits based on its own n i.i.d. samples and the messages $S_1, \ldots S_{i-1}$ sent by the previous $i-1$ machines. Shamir analyzed a specific problem of distributed detection of a special coordinate $j \in [d]$, whose mean is $\tau > 0$, whereas the mean of all other coordinates $i \neq j$ is zero. The following corollary of Shamir's Theorem 6 upper bounds the success probability of detecting j by any distributed (b, n, M) protocol. For completeness, its proof appears in the appendix.

Corollary 3.2. *Consider the class of exact support recovery problems in $d \geq 21$ dimensions, and all possible distributions of a d-dimensional random vector $\mathbf{u}$ such that:*

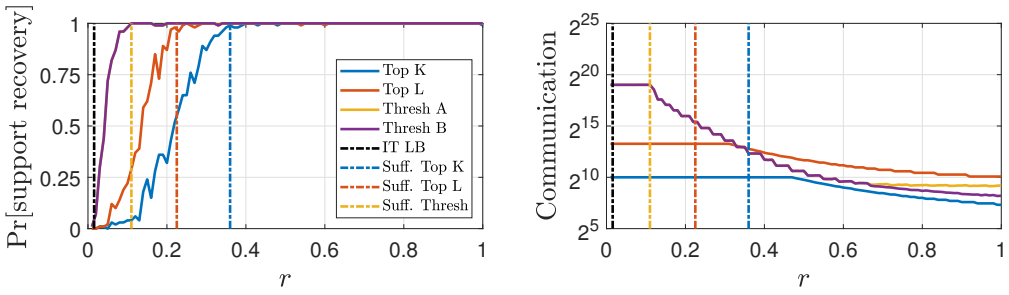
1. *There exists one coordinate j for which $\mathrm{E}[u_j] = \tau > 0$ with $\tau = O(\frac{1}{d \log d})$, whereas $\mathrm{E}[u_i] = 0$ for all other coordinates $i \neq j$.*
2. *All coordinates $i \in [d]$ have the same second moment $\tilde{\sigma}^2 = \mathrm{E}[u_i^2] = \frac{1}{d}$.*
3. *For all coordinates $i \in [d]$, the random variable $(u_i - \mathrm{E}[u_i]) \sim \mathrm{subG}(1)$.*

Assume that $n \leq cd \log d$ for a suitable constant $c > 0$. Then for any estimate $\hat{J}$ of j returned by a (b, n, M) protocol, there exists a distribution as above such that

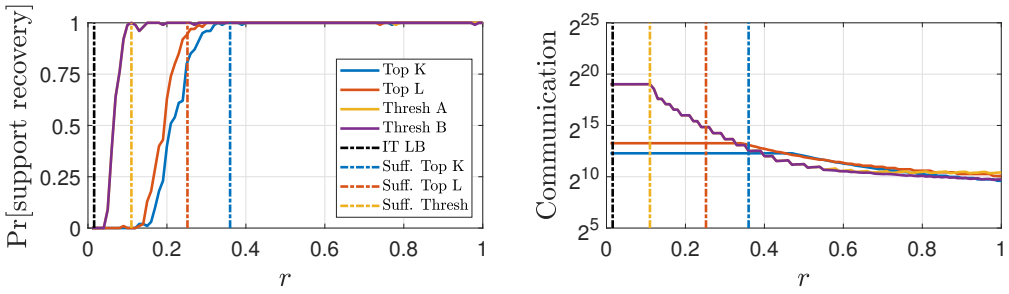
$$\Pr\left[\hat{J} = j\right] \leq O\left(\frac{1}{d} + \sqrt{\frac{Mb}{d}}\right). \tag{3.29}$$

Let us discuss what is the effective signal to noise ratio that corresponds to the above result. Assume that each of M machines has n i.i.d. samples of a vector $\mathbf{s}$ with a distribution as in Corollary 3.2. Similar to (3.17), we define the effective SNR parameter r via the relation $\tau = \frac{\tilde{\sigma}}{\sqrt{n}}\sqrt{2r \log d}$. Taking $\tau = \frac{C}{d \log d}$ and $n = cd \log d$ gives an effective SNR $r = O(\frac{1}{\log^2 d})$. Suppose that each machine sends a message of length b bits, such that the total communication is sublinear in d, namely $Mb \ll d$. Then by Corollary 3.2 the probability of exact support recovery by any (b, n, M) distributed scheme with $n = cd \log d$ samples per machine is $o(1)$.

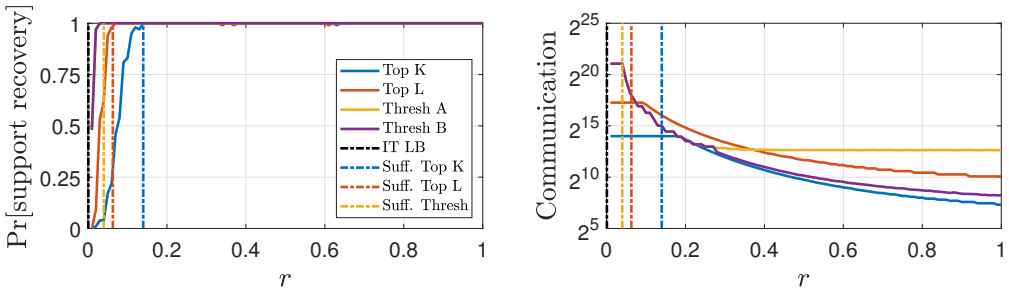
It is important to note that the problem considered in our work and that in Corollary 3.2 are *different*. Specifically, our setting involves measurements of the form sparse signal plus additive noise, where the noise is assumed independent of the signal. Hence, in our setting the second moment of the vector u depends on the signal, namely $\mathbb{E}[u_j^2] = \theta_j^2 + \sigma^2$. In contrast, the distributions in Corollary 3.2 have a constant second order moment (see condition 2). Furthermore, the distribution constructed to prove Corollary 3.2 is not of the form signal plus noise, with noise that is independent of the signal. Hence, the above lower bound is not applicable to our setting. As we proved in Section 3.2, for SNR parameters only slightly higher than $O\left(\frac{1}{\log^2 d}\right)$, namely $r = \Omega\left(\frac{\log^2 \log d}{\log^2 d}\right)$, exact support recovery for signal plus Gaussian noise type observations is possible using sublinear communication. It would be interesting to study if any distributed scheme can recover the support using sublinear communication for SNR values below our aforementioned bound, and to derive tight lower bounds for signal plus noise type distributions.



(a) Setting 1: $d = 2^{15}$, $M = 2^6$, $K = 1$, $L = 10$



(b) Setting 2: $d = 2^{15}$, $M = 2^6$, $K = 5$, $L = 10$



(c) Setting 3: $d = 2^{15}$, $M = 2^{10}$, $K = 1$, $L = 10$

Figure 3.2: The plots on the left depict the success probability of the algorithms from Section 3.2 as a function of r in Settings 1-3. The plots on the right depict the communication cost of the algorithms on a logarithmic scale as a function of r in these settings. The blue curve corresponds to the `Top-L` algorithm with $L = K$, the red curve corresponds to the `Top-L` algorithm with $L > K$, the orange and purple curves correspond to variants A and B of the `Thresholding` algorithm, respectively. The vertical black line is a lower bound on the performance of all algorithms. The colored vertical lines are sufficient SNR bounds for the corresponding algorithms, as described in the main text.

3.5 Simulations

We present several simulations that illustrate the ability of our algorithms to detect the support of a K-sparse d-dimensional vector $\boldsymbol{\theta}$ with sublinear communication. We compare the performance of the `Top-L` algorithm with $L = K$ (blue), the `Top-L` algorithm with $L > K$ (red), variant A of the `Thresholding` algorithm which contacts all machines, i.e., $M_c = M$ (orange), and variant B of the `Thresholding` algorithm which limits the number of contacted machines, i.e., $M_c < M$ (purple). See Appendix 3.7 for details on optimizing simulation parameters.

Figure 3.2 depicts the success probabilities and communication costs (on a logarithmic scale) of the aforementioned algorithms as a function of r, averaged over 100 noise realizations. We consider three

different settings of parameters M and K. In all settings the dimension is $d = 2^{15}$ and in the `Top-L` algorithm with $L > K$ we set $L = 10$. In Setting 1, $M = 2^6$ and $K = 1$; in Setting 2, $M = 2^6$ and $K = 5$; and in Setting 3 $M = 2^{10}$ and $K = 1$.

The vertical black dashed line is the centralized information theoretic lower bound of $1/M$. This line represents the necessary SNR, below which even centralized algorithms fail with high probability. In addition, we define a sufficient SNR bound for each algorithm, above which it exactly recovers the support with high probability $1 - O(K/d)$. The vertical blue and red dashed lines correspond to sufficient SNR bounds for the `Top-L` algorithm with $L = K$ and $L > K$, respectively. The vertical orange dashed line corresponds to the sufficient SNR bounds for the `Thresholding` algorithms. Note that these bounds are conservative, and while they are quite tight in the presented settings, the actual range of SNRs where the algorithms are successful is often larger.

The simulation results reveal several interesting behaviors. First, when the SNR is extremely low, i.e., to the left of the dashed black line, none of our algorithms succeeds with high probability. Second, no algorithm uniformly outperforms the others for all parameter regimes. At low SNR values, the `Thresholding` algorithms have a higher success probability compared to the `Top-L` algorithms, but require higher communication costs. Similarly, at low SNR values the success probability of the `Top-L` algorithm increases with L at the expense of higher communication. At high SNR values, all algorithms succeed with high probability, but the communication costs of the algorithms depend on the parameter settings. For example, the `Top-L` algorithm with $L = K$ can either incur a lower communication cost compared to the `Thresholding` B algorithm (Setting 1), or a higher one (Setting 2), or they can be comparable (Setting 3). In addition, there is a wide range of SNR values for which the communication costs of all algorithms decrease exponentially with r and their total communication costs are sublinear in d.

To understand how a higher sparsity level K affects the performance of the algorithms, we compare between Setting 1 and Setting 2. The communication cost of the `Top-L` algorithm with $L = K$ increases linearly with K. In contrast, dependence of the communication costs of the `Thresholding` algorithms on K varies with the SNR. Specifically, at low SNR values they are comparable for different values of K, but for high SNR values they increase linearly with K. This phenomenon is consistent with the higher number of messages containing support indices.

Finally we compare between Setting 1 and Setting 3 to understand how the availability of more machines affects the performance of the algorithms. With more machines, the `Top-L` algorithms succeed at much lower SNR values, at the expense of higher communication costs. Variant A of the `Thresholding` algorithm has a higher communication cost in Setting 3 compared to Setting 1 since it uses all machines. However, there is still a large range of SNR values where it is smaller than d, due to its adaptive threshold. As shown by our proofs, when M is large, variant B of the `Thresholding` algorithm performs similarly to the `Top-L` algorithm with $L = K$, and they outperform the other algorithms.

3.6 Proofs

Denote the complement of the standard normal cumulative distribution function by $\Phi^c(t) = \Pr[Z > t]$ where $Z \sim \mathcal{N}(0, 1)$. In our proofs we shall use the following well known auxiliary lemmas.

Lemma 3.1 (Gaussian tail bounds)**.** *For $t > 0$,*

$$\frac{t}{\sqrt{2\pi}(t^2+1)} e^{-t^2/2} \leq \Phi^c(t) \leq \frac{1}{\sqrt{2\pi}t} e^{-t^2/2}. \tag{3.30}$$

If in addition $t \geq 1$,

$$\Phi^c(t) \geq \frac{1}{2\sqrt{2\pi}t} e^{-t^2/2}. \tag{3.31}$$

A consequence of Eq. (3.30) is that the maximum of $n-1$ i.i.d. standard normal random variables $Z_1, \ldots, Z_{n-1} \sim \mathcal{N}(0,1)$ is highly concentrated around $\sqrt{2\log n}$. In particular, by the well known identity $\left(1-\frac{1}{n}\right)^{n-1} \geq \frac{1}{e}$, for all $n \geq 2$

$$\begin{aligned}
\Pr\left[\max_{i\in[n-1]} Z_i > \sqrt{2\log n}\right] &= 1-\left(1-\Phi^c\left(\sqrt{2\log n}\right)\right)^{n-1} \\
&\leq 1-\left(1-\frac{1}{n\sqrt{4\pi\log n}}\right)^{n-1} \leq 1-\left(1-\frac{1}{n}\right)^{n-1} \\
&\leq 1-e^{-1}, \qquad (3.32)
\end{aligned}$$

where the third step follows from $4\pi\log n > 1$.

Lemma 3.2 (Chernoff (1952))**.** *Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli random variables and let X denote their sum. Then, for any $\delta \geq 0$,*

$$\Pr\left[X \geq (1+\delta)\,\mathrm{E}\left[X\right]\right] \leq e^{-\frac{\delta^2 \mathrm{E}[X]}{2+\delta}}, \qquad (3.33)$$

and for any $0 \leq \delta \leq 1$,

$$\Pr\left[X \leq (1-\delta)\,\mathrm{E}\left[X\right]\right] \leq e^{-\frac{\delta^2 \mathrm{E}[X]}{2}}. \qquad (3.34)$$

Towards proving the main theorems, we introduce a few definitions. Denote by $I_{i,k}$ the indicator that machine i sends the index k to the fusion center. Note that for each k, the random variables $I_{1,k}, \ldots, I_{M,k}$ are independent and identically distributed. Further denote $p_k = \Pr\left[I_{i,k}=1\right]$ and notice that it is the same for all machines i. Our proofs use a stochastic dominance argument for lower bounding the number of votes $\mathbf{v}_k$ received by each support index $k \in \mathcal{S}$. Towards this goal, we define a Binomial random variable $\mathbf{v}_{\min} \sim Bin\left(M_c, p_{\min}\right)$, where $p_{\min}$ is the probability that machine i sends a support index whose nonzero coordinate is $\theta_{\min}$. By definition of $\theta_{\min}$, the random variable $\mathbf{v}_{\min}$ is stochastically dominated by $\mathbf{v}_k \sim Bin\left(M_c, p_k\right)$ for each $k \in S$. For exact support recovery, it suffices that for some threshold t_c, each support index receives more than t_c votes, and each non-support index receives less than t_c votes. For our proof, we set the threshold

$$t_c = 4\log d. \qquad (3.35)$$

We conclude this subsection with two useful lemmas. First, we show that if $p_{\min}$ is sufficiently high, then any support index receives a number of votes exceeding t_c with high probability.

Lemma 3.3. *Let $p_{\min}$ be the probability defined above, let M_c be the number of contacted machines, and let t_c be the threshold in Eq.* (3.35)*. If $p_{\min} \geq \frac{2t_c}{M_c}$, then*

$$\Pr[\min_{k\in\mathcal{S}} \mathbf{v}_k < t_c] \leq \frac{K}{d}. \qquad (3.36)$$

Proof. Let $\delta = 1 - \frac{t_c}{M_c p_{\min}}$. By the Chernoff bound (3.34),

$$\Pr[\mathbf{v}_{\min} < t_c] = \Pr\left[\mathbf{v}_{\min} < M_c p_{\min}(1-\delta)\right] \leq \exp\left(-M_c p_{\min}\delta^2/2\right). \qquad (3.37)$$

The assumption $p_{\min} \geq \frac{2t_c}{M_c}$ implies that $\delta > 1/2$ and $\delta M_c p_{\min} > t_c$. Thus

$$\Pr[\mathbf{v}_{\min} < t_c] \leq \exp\left(-t_c/4\right) \leq 1/d, \qquad (3.38)$$

where the last inequality follows from Eq. (3.35).

Now, fix $k \in \mathcal{S}$. By the independence of the noises in different machines $\mathbf{v}_k \sim Bin\left(M_c, p_k\right)$. By definition of $\theta_{\min}$, the coordinate $\theta_k \geq \theta_{\min}$ and thus $p_k \geq p_{\min}$. Therefore,

$$\Pr[\mathbf{v}_k < t_c] \leq \Pr[\mathbf{v}_{\min} < t_c].$$

By (3.38), applying a union bound over $k \in S$ proves (3.36). □

Next, let us consider the non-support coordinates. The following lemma shows that if p_j is sufficiently low for each non-support index $j \notin \mathcal{S}$, then no non-support index receives more than t_c votes with high probability.

Lemma 3.4. *Let M_c be the number of contacted machines and let t_c be the threshold in Eq. (3.35). If for each $j \notin \mathcal{S}$, the probability $p_j \leq \frac{t_c}{5M_c}$, then*

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

Proof. The average number of messages at the fusion center containing index j is $\mathbb{E}[\mathbf{v}_j] = M_c p_j$. Let

$$\delta = \frac{1}{M_c p_j}\left(t_c - M_c p_j\right) = \frac{t_c}{M_c p_j} - 1.$$

The assumption $p_j \leq \frac{t_c}{5M_c}$ implies that $\delta M_c p_j = t_c - M_c p_j \geq \frac{4t_c}{5}$ and $\delta \geq 4$, which in turn implies that $\delta/(2+\delta) \geq 2/3$. Note that for each $j \in [d]$ the random variables $I_{1,j}, \ldots, I_{M,j}$ are independent. By a Chernoff bound (3.33),

$$\Pr[\mathbf{v}_j > t_c] = \Pr\left[\mathbf{v}_j > M_c p_j(1+\delta)\right] \leq \exp\left(-\frac{\delta}{2+\delta}\delta M_c p_j\right) \leq \exp\left(-\frac{8t_c}{15}\right). \tag{3.39}$$

By Eq. (3.35), the above probability is at most d^{-2}. We conclude by applying a union bound, $\Pr[\max_{j \notin \mathcal{S}} \mathbf{v}_j > t_c] \leq (d-k)\Pr[\mathbf{v}_j > t_c] \leq 1/d$. □

3.6.1 Proof of Theorem 3.1

We begin by proving Theorem 3.1.A where $L = K = 1$ and then outline the necessary changes in order to prove Theorem 3.1.B for $L \geq K \geq 1$.

For future use, note that by definition of the `Top-L` algorithm, the probability that machine i sends a coordinate $k \in [d]$ is

$$p_k = \Pr\left[\exists j_1, \ldots, j_{d-L} \in [d] \setminus \{k\} : \; s_{i,k} > s_{i,j_1}, \ldots, s_{i,j_{d-L}}\right]. \tag{3.40}$$

The communication of the `Top-L` algorithm is $B = O\left(LM_c \log d\right)$ since the center sends one message to each participating machine indicating $L < d$, and each of these machines sends back exactly L indices.

Proof of Theorem 3.1.A. Without loss of generality, let the support index be $\mathcal{S} = \{1\}$. Thus,

$$p_1 = \Pr\left[s_{i,1} > \max_{j>1} s_{i,j}\right].$$

We show that w.h.p. both $\mathbf{v}_1 > t_c$ and $\mathbf{v}_j < t_c$ for all $j > 1$.

By the law of total probability and the independence of the random variables $\xi_{i,j}$,

$$\begin{aligned}
p_{\min} &= \Pr\left[\sqrt{2r\log d} + \xi_{i,1} > \max_{j>1}\xi_{i,j}\right] \\
&\geq \Pr\left[\sqrt{2r\log d} + \xi_{i,1} > \sqrt{2\log d} \;\middle|\; \max_{j>1}\xi_{i,j} < \sqrt{2\log d}\right] \cdot \Pr\left[\max_{j>1}\xi_{i,j} < \sqrt{2\log d}\right] \\
&= \Pr\left[\xi_{i,1} > \left(1-\sqrt{r}\right)\sqrt{2\log d}\right] \cdot \Pr\left[\max_{j>1}\xi_{i,j} < \sqrt{2\log d}\right].
\end{aligned}$$

Recall that the random variables $\xi_{i,j}$ are i.i.d. standard Gaussians. By Eq. (3.32),

$$\Pr\left[\max_{j>1}\xi_{i,j} < \sqrt{2\log d}\right] \geq e^{-1}.$$

Therefore, by the Gaussian tail bound (3.30),

$$p_{\min} \geq e^{-1}\cdot\frac{\left(1-\sqrt{r}\right)\sqrt{2\log d}}{\sqrt{2\pi}\left(2\left(1-\sqrt{r}\right)^2\log d+1\right)}d^{-\left(1-\sqrt{r}\right)^2}. \tag{3.41}$$

Combining Eq. (3.41) with the bound (3.4) implies that $p_{\min} \geq \frac{2t_c}{M_c}$, and thus we can apply Lemma 3.3 and get that $\Pr[\mathbf{v}_1 < t_c] \leq d^{-1}$.

Now consider a non-support index $j > 1$. By symmetry considerations, the probability that machine i sends j to the center is

$$p_j = \frac{1-p_1}{d-1}.$$

Recall that by definition of $\theta_{\min}$, the coordinate $\theta_k \geq \theta_{\min}$ and thus $p_1 \geq p_{\min}$. Since for any strictly positive SNR $p_1 > p_{\min} > \frac{1}{d}$, it follows that $p_j < \frac{1}{d}$ for each $j > 1$. Hence, the expected number of votes for index j is $\mathrm{E}[\mathbf{v}_j] = M_c p_j < \frac{M_c}{d}$. Let $\delta = \frac{t_c}{M_c p_j} - 1$ and note that the assumption $M \leq d$ implies that $M_c p_j \leq 1$ and hence $\delta \geq 4\log d - 1 > 0$. By the Chernoff bound (3.33),

$$\begin{aligned}
\Pr[\mathbf{v}_j > t_c] &= \Pr\left[\sum_{i=1}^{M_c} I_{i,j} > (1+\delta)M_c p_j\right] \\
&\leq e^{-\frac{\delta^2 M_c p_j}{\delta+2}} = e^{-\frac{\left(4\log d - M_c p_j\right)^2}{4\log d + M_c p_j}} \\
&= e^{-4\log d - M_c p_j + 4M_c p_j \frac{4\log d}{4\log d + M_c p_j}} \leq e^3 d^{-4}.
\end{aligned}$$

By a union bound over all $d-1$ non-support coordinates,

$$\Pr\left[\max_{j>1}\mathbf{v}_j > t_c\right] \leq (d-1)\cdot e^3 d^{-4} \leq e^3 d^{-3}.$$

By an additional union bound on the two events, the algorithm outputs the correct support index with probability at least $1 - d^{-1} - e^3 d^{-3}$. □

Proof of Theorem 3.1.B. The proof is similar to that of Theorem 3.1.A, with the following changes. For any threshold $a \in \mathbb{R}$, the probability p_k that $k \in \mathcal{S}$ is sent to the fusion center is lower bounded by

$$\begin{aligned}
p_k &\geq \Pr\left[s_{i,k} > a\ ,\ \sum_{j\notin\mathcal{S}} \mathbb{1}\left\{s_{i,j} > a\right\} \leq L-K\right] \\
&= \Pr\left[\xi_{i,k} > a-\theta_k\ ,\ \sum_{j\notin\mathcal{S}} \mathbb{1}\left\{\xi_{i,j} > a\right\} \leq L-K\right].
\end{aligned}$$

Set $a = a(K,L,d)$ and $b = b(K,L,d,r)$ by Eqs. (3.5) and (3.6) respectively. Recall that $\xi_{i,j}$ are i.i.d. for all $i \in [M_c]$ and $j \in [d]$, i.e., the two events in the probability above are independent of each other. Combining this with the definition of $\theta_{\min}$ yields

$$p_{\min} \geq \Phi^c(b)\cdot\Pr\left[\sum_{j\notin\mathcal{S}} \mathbb{1}\left\{Z_j > a\right\} \leq L-K\right], \tag{3.42}$$

where $Z_j \sim \mathcal{N}(0,1)$.

We begin by bounding the first term of Eq. (3.42). If $b \leq 0$ then $\Phi^c(b) \geq 1/2$. Otherwise, by the Gaussian tail bound (3.30),

$$\Phi^c(b) \geq \frac{b}{\sqrt{2\pi}\,(b^2+1)}\,(d-K)^{-\left(\sqrt{1-\frac{\log(L-K+1)}{\log(d-K)}}-\sqrt{r}\right)^2}.$$

Next, we show that with probability $\geq \frac{1}{4}$ the number of non-support indices that pass the threshold a is upper bounded by $L-K$. Denote by p_a the probability that a standard normal random variable passes the threshold a, i.e., $p_a \equiv \Phi^c(a)$. By Eq. (3.30), p_a is upper bounded by

$$p_a \leq \frac{1}{\sqrt{2\pi}a} \cdot \frac{L-K+1}{d-K}. \tag{3.43}$$

Next, let $\delta = \frac{L-K+1}{p_a(d-K)} - 1$. Note that the assumption $K \leq L < (d-K)/2$ implies that $\sqrt{2\pi}a \geq \sqrt{4\pi\log 2} > 1$, and thus $\delta > 0$. By the Chernoff bound (3.33),

$$\begin{aligned}
\Pr\left[\sum_{j\notin\mathcal{S}} \mathbb{1}\left\{Z_j > a\right\} \geq L-K+1\right] &= \Pr\left[\sum_{j\notin\mathcal{S}} \mathbb{1}\left\{Z_j > a\right\} \geq (1+\delta)\,p_a\,(d-K)\right] \\
&\leq e^{-\delta^2 p_a(d-K)/(2+\delta)} \\
&= e^{-\left(\frac{L-K+1}{p_a(d-K)}-1\right)^2 \frac{p_a(d-K)}{1+\frac{L-K+1}{p_a(d-K)}}} \\
&= e^{-\frac{(L-K+1-p_a(d-K))^2}{L-K+1+p_a(d-K)}}.
\end{aligned}$$

For $A_1, A_2 > 0$ the function $e^{-\frac{(A_1-A_2)^2}{A_1+A_2}}$ is monotonically increasing in A_2. Letting $A_1 = L-K+1$ and $A_2 = p_a\,(d-K)$, we can now apply the upper bound on A_2 in Eq. (3.43) to the equation above. Thus the complementary probability, i.e., the second term in Eq. (3.42), can be lower bounded as follows

$$\begin{aligned}
\Pr\left[\sum_{j\notin\mathcal{S}} \mathbb{1}\left\{Z_j > a\right\} \leq L-K\right] &\geq 1 - e^{-(L-K+1)\frac{\left(1-1/\sqrt{2\pi}a\right)^2}{1+1/\sqrt{2\pi}a}} \\
&\geq 1 - e^{-\frac{\left(1-\sqrt{4\pi\log 2}\right)^2}{1+\sqrt{4\pi\log 2}}} \geq \frac{1}{4},
\end{aligned}$$

where the second inequality follows from the assumption $K \leq L < (d-K)/2$.

By Eq. (3.7) the probability $p_{\min} \geq \frac{2t_c}{M_c}$ and thus $\Pr\left[\min_{k\in\mathcal{S}} \mathbf{v}_k < t_c\right] \leq Kd^{-1}$ by Lemma 3.3. Let $W_i \sim Bin\,(K, p_{\min})$ be a binomial random variable that serves as a lower bound for how many of the support coordinates machine i sends to the center. By the law of total probability and symmetry of the non-support indices, the probability that machine i sends to the center a non-support index $j \notin \mathcal{S}$ is

$$\begin{aligned}
p_j &\leq \sum_{n=0}^{K} \Pr\left[i \in \mathbf{V}_j | W_i = n\right] \cdot \Pr\left[W_i = n\right] \\
&= \sum_{n=0}^{K} \frac{L-n}{d-K} \Pr\left[W_i = n\right] = \frac{L-Kp_{\min}}{d-K} \leq \frac{L}{d-K}.
\end{aligned} \tag{3.44}$$

Using the requirement $M_{K,L} \leq \frac{d-K}{L}$, the rest of the proof continues in the same manner. □

3.6.2 Proof of Theorem 3.2

Note that we set the precision parameters P, U such that $t_m - 1/d \leq \hat{t}_m \leq t_m$. By definition of the `Thresholding` algorithm, the probability that machine i sends a support coordinate $k \in \mathcal{S}$ is

$$p_k = \Pr\left[s_{i,k} > \hat{t}_m\right] \geq \Pr\left[\xi_{i,k} > t_m - \theta_k\right]. \tag{3.45}$$

Thus, for the extreme case $\theta_k = \theta_{\min}$,

$$p_{\min} \geq \Phi^c\left(t_m - \theta_{\min}\right). \tag{3.46}$$

For a non-support coordinate $j \notin \mathcal{S}$, the Gaussian tail bound (3.30) implies that

$$p_j = \Pr\left[\xi_{i,j} > \hat{t}_m\right] \leq \frac{e^{-\hat{t}_m^2/2}}{\sqrt{2\pi}\hat{t}_m} \leq e^{t_m/d}\frac{e^{-t_m^2/2}}{\sqrt{2\pi}(t_m - 1/d)}. \tag{3.47}$$

In terms of communication, each coordinate $j \in [d]$ appears in $M_c p_j$ messages on average. In addition, in the setup stage the fusion center sends M_c messages with the truncated threshold $\hat{t}_m$, whose binary representation is $O(\log d)$ bits long. Hence the average total communication is

$$\mathrm{E}\left[B\right] = O\left(M_c \log d + \left(\sum_{k\in\mathcal{S}} p_k + \sum_{j\notin\mathcal{S}} p_j\right) M_c \log d\right) = O\left(\left(K + \sum_{j\notin\mathcal{S}} p_j\right) M_c \log d\right), \tag{3.48}$$

where the last step follows from the trivial bound $p_k \leq 1$ for each $k \in \mathcal{S}$.

We now proceed to proving the sub-theorems.

Proof of Theorem 3.2.A. By Eqs. (3.8) and (3.46),

$$p_{\min} \geq \Phi^c\left(0\right) = \frac{1}{2}. \tag{3.49}$$

Since $M_c = \lceil 16 \log d \rceil$ and by Eq. (3.35), we have that $p_{\min} \geq 2t_c/M_c$, and thus $\Pr[\min_{k\in\mathcal{S}} \mathbf{v}_k < t_c] \leq \frac{K}{d}$ by Lemma 3.3. Now fix $j \notin \mathcal{S}$. Due to the assumptions $d \geq 16$ and $r > \frac{\log 5}{\log(d-K)}$, by Eq. (3.47) we have that

$$p_j \leq \frac{e^{\sqrt{2r\log(d-K)}/d}\left(d-K\right)^{-r}}{\sqrt{2\pi}\left(\sqrt{2r\log\left(d-K\right)} - 1/d\right)} \leq \frac{e^{\sqrt{2\log 5}/16}}{10\left(\sqrt{\pi\log 5} - 1/16\right)} \leq \frac{t_c}{5M_c}.$$

Applying Lemma 3.4 yields $\Pr[\max_{j\notin\mathcal{S}} |\mathbf{v}_j > t_c] \leq 1/d$.

Finally, the average total communication follows from inserting the expressions for p_j and M_c into Eq. (3.48).

Proof of Theorem 3.2.B. Note that the bound $M > \sqrt{e}\cdot 32\sqrt{\pi}\log^{1.5} d$ implies that $2\log\frac{M}{32\sqrt{\pi}\log^{1.5} d} \geq 1$. By the expression (3.11) for t_m and the Gaussian tail bound (3.31),

$$\begin{aligned} p_{\min} &\geq \Phi^c\left(\sqrt{2\log\frac{M}{32\sqrt{\pi}\log^{1.5} d}}\right) \geq \frac{1}{2\sqrt{2\pi}\sqrt{2\log\frac{M}{32\sqrt{\pi}\log^{1.5} d}}} e^{-\log\frac{M}{32\sqrt{\pi}\log^{1.5} d}} \\ &= \frac{1}{2}\cdot\frac{16\log d}{M}\sqrt{\frac{\log d}{\log\frac{M}{32\sqrt{\pi}\log^{1.5} d}}} \geq \frac{2t_c}{M}, \end{aligned} \tag{3.50}$$

where the last inequality follows from the upper bound on M. Thus, by Lemma 3.3, $\min_{k\in\mathcal{S}} \mathbf{v}_k < t_c$ with probability at most K/d. Due to Assumption (3.10), $\hat{t}_m \geq \sqrt{2\log\frac{5M}{\sqrt{2\pi}4\log d}}$, and thus by the first inequality of Eq. (3.47),

$$p_j \leq \frac{1}{\sqrt{2\pi}\sqrt{2\log\frac{5M}{\sqrt{2\pi}4\log d}}} e^{-\log\frac{5M}{\sqrt{2\pi}4\log d}} = \frac{4\log d}{5M\sqrt{2\log\frac{5M}{\sqrt{2\pi}4\log d}}} < \frac{t_c}{5M}, \tag{3.51}$$

where the last inequality follows from Eq. (3.35) and the condition on M. Thus by Lemma 3.4, $\Pr[\max_{j\notin\mathcal{S}} \mathbf{v}_j > t_c] \le 1/d$.

Towards computing the expected communication of the algorithm, we bound p_j more carefully using the second inequality of Eq. (3.47),

$$\begin{aligned} p_j &\le \frac{e^{\left(\sqrt{2r\log(d-K)}+\sqrt{2\log\frac{M}{32\sqrt{\pi}\log^{1.5}d}}\right)/d}(d-K)^{-r}\,e^{-2\sqrt{r\log(d-K)\log\frac{M}{32\sqrt{\pi}\log^{1.5}d}}}}{\sqrt{2}\left(\sqrt{2r\log(d-K)}+\sqrt{2\log\frac{M}{32\sqrt{\pi}\log^{1.5}d}}-\frac{1}{d}\right)}\cdot\frac{32\log^{1.5}d}{M} \\ &\le (d-K)^{-r}\,e^{-2\sqrt{r\log(d-K)\log\frac{M}{32\sqrt{\pi}\log^{1.5}d}}}\cdot\frac{32\log^{1.5}d}{M}, \end{aligned} \tag{3.52}$$

where the second inequality follows from bounding $\sqrt{2r\log(d-K)}+\sqrt{2\log\frac{M}{32\sqrt{\pi}\log^{1.5}d}}-\frac{1}{d}>1$ and from $d\ge 15$, which implies that $e^{\left(\sqrt{2r\log(d-K)}+\sqrt{2\log\frac{M}{32\sqrt{\pi}\log^{1.5}d}}\right)/d}<\sqrt{2}$. By inserting Eq. (3.52) into Eq. (3.48), the expected communication of the algorithm is

$$\mathrm{E}[B] = O\left(KM\log d + (d-K)\cdot(d-K)^{-r}\,e^{-2\sqrt{r\log(d-K)\log\frac{M}{32\sqrt{\pi}\log^{1.5}d}}}\frac{\log^{1.5}d}{M}\cdot M\log d\right).$$

Rearranging completes the proof. □

Proof of Theorem 3.2.C. Recall that M_c and t_m are given by Eqs. (3.13) and (3.14), respectively. By the Gaussian tail bound (3.30),

$$\begin{aligned} p_{\min} &\ge \Phi^c\left(\left(1-\sqrt{r}\right)\sqrt{2\log(d-K)}\right) \\ &\ge \frac{\left(1-\sqrt{r}\right)\sqrt{2\log(d-K)}}{\sqrt{2\pi}\left(\left(1-\sqrt{r}\right)^2 2\log(d-K)+1\right)}(d-K)^{-\left(1-\sqrt{r}\right)^2} = \frac{2t_c}{M_c}. \end{aligned} \tag{3.53}$$

Thus by Lemma 3.3, $\min_{k\in\mathcal{S}}\mathbf{v}_k < t_c$ with probability at most K/d.

Fix a non-support index $j\notin\mathcal{S}$. Note that the assumption $d-K\ge 20$ implies that $e^{\sqrt{2\log(d-K)}/d}<\sqrt{2}$. Thus, by Eq. (3.47),

$$p_j \le \frac{e^{\sqrt{2\log(d-K)}/d}}{\sqrt{2\pi}\left(\sqrt{2\log(d-K)}-1/d\right)}(d-K)^{-1} \le \frac{1}{\sqrt{\pi}\left(\sqrt{2\log(d-K)}-1/d\right)}(d-K)^{-1}. \tag{3.54}$$

It is easy to verify that the aforementioned assumption and the condition $r>\left(\frac{\log 10}{\log(d-K)}\right)^2$, or equivalently, $d^{2\sqrt{r}}\ge 100$, imply that $p_j<\frac{t_c}{5M_c}$. Thus the desired bound $\Pr[\max_{j\notin\mathcal{S}}\mathbf{v}_j>t_c]\le 1/d$ follows from Lemma 3.4.

By inserting Eq. (3.54) into Eq. (3.48), the expected communication of the algorithm is

$$\mathrm{E}[B] = O\left(KM_c\log d + (d-K)\cdot\frac{1}{\sqrt{2\pi}\sqrt{2\log(d-K)}}(d-K)^{-1}\cdot M_c\log d\right).$$

Using Eq. (3.13) concludes the proof. □

3.6.3 Proof of Theorem 3.3

Towards proving Theorem 3.3, we first recall the definition of sub-Gaussian random variables and cite a couple of useful results.

Definition 3.1 ((Rigollet, 2015, Definition 1.2)). *A random variable $X \in \mathbb{R}$ is said to be sub-Gaussian with sub-Gaussian parameter (variance proxy) $\sigma^2 > 0$ if $\mathrm{E}[X] = 0$ and its moment generating function satisfies $\mathrm{E}[\exp(sX)] \le \exp\left(\sigma^2 s^2/2\right)$ for all $s \in \mathbb{R}$. We write $X \sim \mathrm{subG}(\sigma^2)$.*

Let $X_1, \ldots, X_n \sim \mathrm{subG}\left(\sigma^2\right)$ be i.i.d. sub-Gaussian random variables. We denote the variance of each r.v. by $\tilde{\sigma}^2 = E\left[X_j^2\right]$ and their third absolute moment by $\eta = E|X_j|^3$, and assume that it is finite $\eta < \infty$. In our proofs for the Gaussian noise case, we used the upper and lower bounds on tail probabilities described in Lemma 3.1. We now show that similar bounds hold for sub-Gaussian noise.

To prove tail lower bounds, we need the following result.

Theorem 3.4 ((Nagaev, 2002, Corollary 3)). *Let $X_1, \ldots, X_n \sim \mathrm{subG}\left(\sigma^2\right)$ be i.i.d. sub-Gaussian random variables as defined above and let $\tilde{X} = \frac{1}{\tilde{\sigma}\sqrt{n}} \sum_{i=1}^n X_i$. If $1.7 < t \le \frac{1}{25}\sqrt{n}\frac{\tilde{\sigma}^3}{\eta}$, then*

$$\Pr\left[\tilde{X} > t\right] > \Phi^c(t) \exp\left(-\frac{(2.35\eta/\tilde{\sigma}^3 + 0.2)t^3}{\sqrt{n}}\right)\left(1 - \frac{(16.88\eta/\tilde{\sigma}^3 + 6.58)t}{\sqrt{n}}\right). \tag{3.55}$$

To prove tail upper bounds, we use the following lemma, which is an easy corollary of Lemma 1.3 and Corollary 1.7 of Rigollet (2015).

Lemma 3.5. *Let $X_1, \ldots, X_n \sim \mathrm{subG}\left(\sigma^2\right)$ be i.i.d. sub-Gaussian random variables as defined above and let $\tilde{X} = \frac{1}{\tilde{\sigma}\sqrt{n}} \sum_{i=1}^n X_i$. Then, $\tilde{X} \sim \mathrm{subG}(\sigma^2/\tilde{\sigma}^2)$, and there exists a constant $c > 0$ such that for any $t > 0$,*

$$\Pr\left[\tilde{X} > t\right] \le \exp\left(-c\tilde{\sigma}^2 t^2/\sigma^2\right). \tag{3.56}$$

We now proceed to prove the theorem.

Proof of Theorem 3.3. The proof is similar to that of Theorem 3.2.C, with the following changes, pertaining to the probabilities of sending the support and non-support indices, i.e., Eqs. (3.53) and (3.54). Let $t = (1 - \sqrt{r})\sqrt{2\log(d-K)}$ and let the number of samples in each machine satisfy Eq. (3.21) as follows,

$$n \ge \frac{\left(2.35\eta/\tilde{\sigma}^3 + 0.2\right)^2}{\log^2 2} t^6 = \frac{8\left(2.35\eta/\tilde{\sigma}^3 + 0.2\right)^2}{\log^2 2}(1 - \sqrt{r})^6 \log^3(d-K). \tag{3.57}$$

Further assume that $1.7 < (1 - \sqrt{r})\sqrt{2\log(d-K)} \le \frac{1}{25}\sqrt{n}\frac{\tilde{\sigma}^3}{\eta}$. It is easy to verify that for the values of n in Eq. (3.57), the SNR r satisfies the left inequality in (3.20). We begin with applying Theorem 3.4 to lower bound the probability that machine i sends a support index $k \in \mathcal{S}$. The first term of Eq. (3.55), i.e. $\Phi^c(t)$, is identical to that of Eq. (3.53). By Eq. (3.57), the second term is at least $1/2$. By Jensen's inequality, $\eta/\tilde{\sigma}^3 \ge 1$. Thus, the third term is

$$1 - (16.88\eta/\tilde{\sigma}^3 + 6.58)\frac{t}{\sqrt{n}} > 1 - \frac{4\log 2}{(1 - \sqrt{r})^2 \log(d-K)},$$

which is also greater than $\frac{1}{2}$ for sufficiently large $d - K$. Hence, Theorem 3.4 implies that

$$\Pr\left[\tilde{\xi}_{i,k} > (1 - \sqrt{r})\sqrt{2\log(d-K)}\right] > \frac{1}{4}\Phi^c\left((1 - \sqrt{r})\sqrt{2\log(d-K)}\right). \tag{3.58}$$

Let

$$M_c = \left\lceil 4 \cdot \frac{8\sqrt{2\pi}\left((1 - \sqrt{r})^2 2\log(d-K) + 1\right)}{(1 - \sqrt{r})\sqrt{2\log(d-K)}} (d-K)^{(1-\sqrt{r})^2} \log d \right\rceil. \tag{3.59}$$

Then, Eq. (3.53) can be replaced with

$$p_{\min} > \frac{1}{4} \cdot \frac{(1-\sqrt{r})\sqrt{2\log(d-K)}}{\sqrt{2\pi}\left((1-\sqrt{r})^2 2\log(d-K)+1\right)} (d-K)^{-(1-\sqrt{r})^2} = \frac{2t_c}{M_c}. \tag{3.60}$$

The bound $\Pr[\min_{k\in\mathcal{S}} \mathbf{v}_k < t_c] \le K/d$ follows from Lemma 3.3.

As for the non-support indices, by Lemma 3.5 the r.v. $\tilde{\xi}_{i,j} \sim \mathrm{subG}(\sigma^2/\tilde{\sigma}^2)$. Thus, by the tail bound (3.56) and by Condition (3.16),

$$p_j = \Pr\left[\tilde{\xi}_{i,j} > \hat{t}_m\right] \le \exp\left(-c\tilde{\sigma}^2\hat{t}_m^2/\sigma^2\right) \le \exp\left(-c\lambda^2\hat{t}_m^2\right).$$

As in the proof of Theorem 3.2.C, recall that the truncated threshold satisfies that $\hat{t}_m > t_m - 1/d$. Hence,

$$p_j \le \exp\left(-c\lambda^2 t_m^2\right)\exp\left(2c\lambda^2 t_m/d\right).$$

Next, we insert into the right hand side above the value of t_m, Eq. (3.22). For sufficiently large $d-K$, the second term above is bounded by say 2. Hence,

$$p_j \le O\left((d-K)^{-2c\lambda^2}\right). \tag{3.61}$$

Condition (3.20) with a constant $C > 2c$ implies that for sufficiently large $d-K$, as in the original proof, $p_j \le \frac{t_c}{5M_c}$ where t_c is given in Eq. (3.35). Thus the desired bound $\Pr[\max_{j\notin\mathcal{S}} \mathbf{v}_j > t_c] \le 1/d$ follows from Lemma 3.4. The communication bound (3.23) follows directly from Eqs. (3.61) and (3.59).

3.6.4 Proof of Corollary 3.1

We first analyze the total communication cost of Π. Each machine i sends a message $\mathbf{w}_i$ consisting of the truncated binary representations of $s_{i,k}$ for $k \in \hat{\mathcal{S}}$. Recall that the length of each $\mathbf{w}_{i,k}$ is $P+U+2$ bits. Since $P, U = O(\log d)$, the expected total communication cost of Π is $O(KM\log d)$.

Let $\hat{\boldsymbol{\theta}}$ be the output of protocol Π, and recall that $\hat{\theta}_j = \bar{z}_j \cdot \mathbb{1}\left\{j \in \hat{\mathcal{S}}\right\}$. By linearity of expectation and the law of total probability,

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right\|_2^2\right] = \sum_{j\in[d]} \mathrm{E}\left[\left(\theta_j - \hat{\theta}_j\right)^2\right] = \sum_{j\in[d]} \left(\mathrm{E}\left[(\theta_j-\bar{z}_j)^2\right]\Pr\left[j\in\hat{\mathcal{S}}\right] + \theta_j^2 \Pr\left[j\notin\hat{\mathcal{S}}\right]\right). \tag{3.62}$$

We now bound each of the terms in the RHS.

Fix $j \in [d]$. Since $\mathrm{E}[\bar{s}_j] = \theta_j$, it follows that

$$\mathrm{E}\left[(\theta_j-\bar{z}_j)^2\right] = \mathrm{E}\left[(\theta_j-\bar{s}_j+\bar{s}_j-\bar{z}_j)^2\right] = \mathrm{E}\left[(\theta_j-\bar{s}_j)^2\right] + \mathrm{E}\left[(\bar{s}_j-\bar{z}_j)^2\right].$$

Furthermore, since the noise in different machines is i.i.d.,

$$\mathrm{E}\left[(\theta_j-\bar{s}_j)^2\right] = \frac{1}{M}\mathrm{E}\left[(\theta_j-s_{i,j})^2\right] = \frac{1}{M},$$

and

$$\mathrm{E}\left[(\bar{s}_j-\bar{z}_j)^2\right] = \frac{1}{M}\mathrm{E}\left[(s_{i,j}-z_{i,j})^2\right]$$

for any fixed $i \in [M]$.

We now bound $\mathrm{E}\left[\left(s_{i,j}-z_{i,j}\right)^2\right]$ for any fixed $j\in[d]$ and $i\in[M]$. Since $s_{i,j}\sim\mathcal{N}(\theta_j,1)$ and $z_{i,j}$ is a deterministic function of it $z_{i,j}(s_{i,j})$, then

$$\mathrm{E}\left[\left(s_{i,j}-z_{i,j}\right)^2\right]=\int_{-\infty}^{\infty}\left(s-z\left(s\right)\right)^2\frac{\exp\left(-\left(s-\theta_j\right)^2/2\right)}{\sqrt{2\pi}}ds.$$

If $s<2^{U+1}$, then the truncation step of the protocol implies that the remainder is bounded such that $|s-z|\leq 2^{-P}$. Otherwise, the value s is higher than the range that is representable using $U+1$ bits before the binary dot, and thus the magnitude $|s-z|$ can be as large as $|s|$ itself. Therefore,

$$\mathrm{E}\left[\left(s_{i,j}-z_{i,j}\right)^2\right]\leq 2^{-P}+2\int_{2^{U+1}}^{\infty}s^2\frac{\exp\left(-\left(s-\theta_j\right)^2/2\right)}{\sqrt{2\pi}}ds.$$

Using integration by parts,

$$\int_{2^{U+1}}^{\infty}s^2\frac{e^{-(s-\theta_j)^2/2}}{\sqrt{2\pi}}ds\leq\frac{1}{\sqrt{2\pi}}\left(2^{U+1}+\theta_j\right)e^{-\left(2^{U+1}-\theta_j\right)^2/2}+\left(1+\theta_j^2\right)\int_{2^{U+1}}^{\infty}\frac{e^{-(s-\theta_j)^2/2}}{\sqrt{2\pi}}ds.$$

By the Gaussian tail bound 3.30,

$$\begin{aligned}\int_{2^{U+1}}^{\infty}s^2\frac{e^{-(s-\theta_j)^2/2}}{\sqrt{2\pi}}ds &\leq \frac{1}{\sqrt{2\pi}}\left(2^{U+1}+\theta_j+\frac{1+\theta_j^2}{2^{U+1}-\theta_j}\right)e^{-\left(2^{U+1}-\theta_j\right)^2/2}\\ &\leq \frac{1}{\sqrt{2\pi}}\left(\sqrt{4\left(\gamma+1\right)\log d}+2d^{\gamma}+\frac{1+d^{2\gamma}}{\sqrt{4\left(\gamma+1\right)\log d}}\right)d^{-2(\gamma+1)}\leq\frac{1}{2d},\end{aligned}$$

where the second inequality follows from the bound $\theta_{\max}<d^{\gamma}$ and the selection U, and the last inequality holds for all $d\geq 5$ and $\gamma\geq 0$. Finally, Since $P=\lceil\log_2 d\rceil$,

$$\mathrm{E}\left[\left(s_{i,j}-z_{i,j}\right)^2\right]\leq d^{-2}+d^{-1}.$$

In addition, since $\mathrm{E}\left[|\hat{\mathcal{S}}|\right]=K$, the sum $\sum_{j\in[d]}\Pr\left[j\in\hat{\mathcal{S}}\right]=K$, and thus the first term in the RHS of Eq. (3.62) is bounded by $\frac{K}{M}\left(1+d^{-2}+d^{-1}\right)$.

It remains to prove that for each support index $k\in\mathcal{S}$,

$$\theta_k^2\Pr\left[k\notin\hat{\mathcal{S}}\right]\leq 2\theta_{\min}^2/d.$$

Denote by G the "good" event that each non-support index $j\notin\mathcal{S}$ receives less than $t_c=4\log d$ votes. Fix $k\in\mathcal{S}$. By the law of total probability,

$$\Pr\left[k\notin\hat{\mathcal{S}}\right]\leq\left(\Pr\left[k\notin\hat{\mathcal{S}}|G\right]\Pr\left[G\right]+\left(1-\Pr\left[G\right]\right)\right).$$

Conditioned on G, the index $k\in\hat{\mathcal{S}}$ if $\mathbf{v}_k>t_c$. The complementary probability can be bounded by Chernoff (3.34),

$$\Pr\left[k\notin\hat{\mathcal{S}}|G\right]\leq\Pr\left[\mathbf{v}_k<t_c\right]\leq e^{-\frac{1}{2}(M_cp_k-t_c)\left(1-\frac{t_c}{M_cp_k}\right)}.$$

Recall that under the conditions of Theorem 3.1.B, $p_j\leq\frac{t_c}{5M_c}$ for each non-support index $j\notin\mathcal{S}$. Therefore $\Pr\left[G\right]\geq 1-1/d$ by Lemma 3.4. Thus,

$$\theta_k^2\Pr\left[k\notin\hat{\mathcal{S}}\right]\leq\theta_k^2\left(e^{-\frac{1}{2}(M_cp_k-t_c)\left(1-\frac{t_c}{M_cp_k}\right)}\left(1-1/d\right)+1/d\right).\tag{3.63}$$

In addition, recall that p_k is defined by Eq. (3.40) for the `Top-L` algorithm (or by Eq. (3.45) for the `Thresholding` algorithm), and decays exponentially with θ_k. Therefore, the right hand side of Eq. (3.63) is monotonically decreasing in θ_k, and thus upper bounded by

$$\theta_{\min}^2 \left(e^{-\frac{1}{2}(M_c p_{\min} - t_c)\left(1 - \frac{t_c}{M_c p_{\min}}\right)} (1 - 1/d) + 1/d \right).$$

Recall that the assumption $p_{\min} \geq \frac{2t_c}{M_c}$ also holds under the conditions of Theorem 3.1.B. Thus we can apply Eq. (3.38) and get the desired bound

$$\theta_k^2 \Pr\left[k \notin \hat{\mathcal{S}}\right] \leq \theta_{\min}^2 \left(\frac{1}{d} (1 - 1/d) + 1/d \right) \leq \frac{2\theta_{\min}^2}{d}.$$

We now turn to proving the last part of the corollary. The lower bound in Condition (3.3) implies that $\theta_{\min} > 1/\sqrt{M}$. Thus, by Eq. (3.24), the oracle risk is $R_{\text{oracle}}(\boldsymbol{\theta}) = K/M$. In addition, the upper bound in Condition (3.3) implies that $\theta_{\min}^2 \leq 2 \log d$. Taking $K, M, d \to \infty$ with $\frac{KM \log d}{d} \to 0$ yields the desired result.

3.6.5 Proof of Corollary 3.2

The proof is similar to that of Theorem 5 of Shamir (2014). The main step of the proof, which is proved in Shamir's Theorem 6, is deriving an upper bound on the probability of detecting the special coordinate j for some "hard" distribution. It remains to prove that this distribution indeed satisfies the conditions specified in Corollary 3.2.

Within the proof of his Theorem 5, Shamir (2014) defined the following problem, which he referred to as hide-and-seek 2.

Definition 3.2 (Hide-and-seek Problem 2)**.** *Let $0 < \rho < \frac{1}{2}$. Consider the set of distributions $\{\Pr_j(\cdot)\}_{j=1}^d$ over $\{-\mathbf{e}_i, +\mathbf{e}_i\}_{i=1}^d$, defined as*

$$\Pr_j(\mathbf{e}_i) = \begin{cases} \frac{1}{2d} & i \neq j \\ \frac{1}{2d} + \frac{\rho}{d} & i = j \end{cases} \qquad \Pr_j(-\mathbf{e}_i) = \begin{cases} \frac{1}{2d} & i \neq j \\ \frac{1}{2d} - \frac{\rho}{d} & i = j \end{cases}.$$

Given an i.i.d. sample of Mn instances generated from $\Pr_j(\cdot)$, where j is unknown, detect j.

Let $\mathbf{u} \in \mathbb{R}^d$ be a random vector sampled from $\Pr_j(\cdot)$. We now verify that it satisfies the conditions specified in Corollary 3.2.

1. By construction, there exists j for which $\mathrm{E}[u_j] = 2\rho/d$, whereas $\mathrm{E}[u_i] = 0$ for all other coordinates $i \neq j$. Thus, the first condition holds for $\tau = 2\rho/d > 0$.
2. For each coordinate $i \in [d]$, the value $u_i^2 = 1$ with probability $1/d$ and 0 otherwise, and thus $\mathrm{E}[u_i^2] = 1/d$.
3. For each coordinate $i \in [d]$, the random variable $(u_i - \mathrm{E}[u_i])$ equals $+1$ w.p. $1/2d$, -1 w.p. $1/2d$, and 0 w.p. $1 - 1/d$. Its absolute values are bounded by 1, and thus it is sub-Gaussian with parameter 1.

Shamir proved the following theorem, which bounds the success probability of detecting j.

Theorem 3.5 ((Shamir, 2014, Theorem 6))**.** *Consider the hide-and-seek problem 2 on $d > 1$ coordinates, with some bias $\rho \leq \min\left\{\frac{1}{27}, \frac{1}{9 \log d}, \frac{d}{14n}\right\}$ and sample size Mn. Then for any estimate $\hat{J}$ of the biased coordinate returned by any (b, n, M) protocol, there exists some coordinate j such that*

$$\Pr_j\left[\hat{J} = j\right] \leq \frac{3}{d} + 11\sqrt{\frac{Mb}{d}}. \tag{3.64}$$

To complete the proof of Corollary 3.2, note that if $d \geq 21$ and $n \leq \frac{9}{14} d \log d$, then $\frac{1}{9 \log d} \leq \min\left\{\frac{1}{27}, \frac{d}{14n}\right\}$. Thus, Shamir's Theorem 6 with $\tau = \frac{2\rho}{d} \leq \frac{2}{9d \log d}$ proves the corollary.

□

3.7 Simulation Parameter Settings

For simplicity of the proofs we did not fully optimize the choices of M_c and thresholds. We outline below the choices used for our simulations in Section 3.5. In terms of setup message length, in all simulations L is represented by $\log_2 L$ bits and $\hat{t}_m$ is represented with $U = 3$ bits before the binary dot and $P = 3$ bits after the binary dot.

`Top-L` **Algorithm.** We define the following random variables that represent bounds on the number of votes that a support coordinate $k \in \mathcal{S}$ receives $Y_s^{top}(d, r, K, L) = Bin\left(M_c, p_s^{top}\right)$ and on the number of votes that a non-support coordinate $j \notin \mathcal{S}$ receives $Y_n^{top}(d, r, K, L) = Bin\left(M_c, p_n^{top}\right)$, where $p_s^{top} = p_s(d, r, K, L)$ is the probability that $k \in \mathcal{S}$ is sent by machine i, defined in Eq. (3.42), and $p_n^{top} = p_n(d, r, K, L)$ is the probability that $j \notin \mathcal{S}$ is sent by machine i, defined in Eq. (3.44).

With high probability Y_n^{th} does not deviate from its expectation by more than a $\frac{\log(d-K)}{\log\log(d-K)}$ multiplicative bound. Thus, we set the number of contacted machines as

$$M_c = \max\left\{\left\lceil \frac{1}{p_s^{top}(d, r, K, L)} \cdot \frac{\log(d-K)}{\log\log(d-K)} \right\rceil, 1\right\}. \tag{3.65}$$

Intuitively, this selection ensures that the expected number of votes for a fixed support index is equal to the maximal expected number of votes for any non-support index.

In all of our simulations $\frac{\log(d-K)}{\log\log(d-K)} \cdot \mathrm{E}Y_n^{top} < 1$. Hence, the sufficient SNR bound for the `Top-L` algorithm (vertical blue/red line) is the minimal r for which $\mathrm{E}Y_s^{top} \geq 2$, i.e., the support indices have at least 2 votes in expectation while the non-support indices have at most 1.

`Thresholding` **Algorithm.** Similarly to the calculation for the `Top-L` algorithm, we define $Y_s^{th}(d, r, K, t_m) = Bin\left(M_c, p_s^{th}\right)$ and $Y_n^{th}(d, r, K, t_m) = Bin\left(M_c, p_n^{th}\right)$ as the number of votes for a support coordinate and non-support coordinate respectively, where p_s^{th} is by Eq. (3.46) and p_n^{th} is by Eq. (3.47).

For variant A, given r and M, we set the number of contacted machines $M_c = M$ and the threshold t_m as the highest t s.t.

$$\Pr\left[Y_s^{th}(d, r, K, t) < \mathrm{E}Y_n^{th}(d, r, K, t) \frac{\log(d-K)}{\log\log(d-K)}\right] < \frac{1}{d}. \tag{3.66}$$

Intuitively, Eq. (3.66) requires that the probability that the number of votes for a fixed support index is higher than the maximal expected number of votes for any non-support index is lower than d^{-1}.

In variant B, the parameters t_m and M_c are set in the following manner. If for $M_c = M$ the threshold $t < \sqrt{2 \log \frac{d-K}{K}}$, then $t_m = t$ as in variant A. Otherwise, we set $t_m = \sqrt{2 \log \frac{d-K}{K}}$ and take the lowest M_c for which Eq. (3.66) with $t = t_m$ holds.

Let $r_{\min}$ and $t_{\min}$ denote the minimal r value and the corresponding t value for which Eq. (3.66) holds, respectively. $r_{\min}$ is the sufficient SNR bound for the `Thresholding` algorithm (vertical orange line). Note that when $r < r_{\min}$, there is no value of t for which this Eq. (3.66) holds. For completeness of the simulations, in this case we set $t = t_{\min}$.

Chapter 4

Recovery Guarantees for Distributed-OMP[1]

This Chapter studies high-dimensional sparse linear regression, a problem we formalize in Section 4.1. In Section 4.2, we present distributed schemes for this problem, 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.

Section 4.3 details our theoretical contributions. 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. In Section 4.4, we introduce simulations showing that distributed-OMP schemes are competitive with more computationally intensive methods, and in some cases even outperform them. In Section 4.5, we contextualize our work by discussing the gap in communication required for support recovery between high-SNR and low-SNR settings.

Detailed proofs are provided in Section 4.6. In Section 4.7 we show both theoretically and via simulations that, in some cases, a distributed-OMP scheme can recover the support of $\boldsymbol{\theta}$ even when the sparsity level K is unknown. Additional simulation results and implementation details appear in Sections 4.8 and 4.9, respectively.

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 $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$.

4.1 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

[1]The results presented in this chapter have been accepted for publication in the Proceedings of the Conference on Artificial Intelligence and Statistics (AISTATS) Amiraz et al. (2024).

Algorithm 5: `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 $\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\mathbf{z} \in \mathbb{R}^d, \operatorname{supp}(\mathbf{z}) = S} \|\mathbf{y} - \mathbf{X}\mathbf{z}\|_2$
compute residual $\mathbf{r} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}$
output index $j = \arg\max \left\{ \frac{|\langle \mathbf{x}_i, \mathbf{r} \rangle|}{\|\mathbf{x}_i\|} : i \in [d] \right\}$

sparse regression model (1.5), 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

$$\mathbf{y}^{(m)} = \mathbf{X}^{(m)}\boldsymbol{\theta} + \sigma\xi^{(m)}, \tag{4.1}$$

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

4.2 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 5, 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 6.

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, $S_t = S_{t-1} \cup \{j_t\}$. 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 7.

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,

Algorithm 6: `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 7: `Distributed Joint OMP (DJ-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 $j_t = \operatorname{argmax}_j \mathbf{v}_j^{(t)}$

    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

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.

4.3 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 4.6.

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

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

$$\mu\left(\mathbf{A}\right) = \max_{i \neq j} \frac{\left|\langle \mathbf{a}_i, \mathbf{a}_j \rangle\right|}{\left\|\mathbf{a}_i\right\|_2 \left\|\mathbf{a}_j\right\|_2}. \tag{4.2}$$

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

$$\mu(\mathbf{A}) < \frac{1}{2K-1}. \tag{4.3}$$

A fundamental result by Tropp (2004) is that in an ideal noise-less setting ($\sigma = 0$), the MIP condition (4.3) 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

$$\mu_{\max} = \mu_{\max}(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}) = \max_{m \in [M]} \mu^{(m)}. \tag{4.4}$$

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

$$\mu_{\max} < \frac{1}{2K-1}. \tag{4.5}$$

Eq. (4.5) implies that all machines satisfy the MIP condition (4.3). 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* (4.5) *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* (4.5) *instead of MIP* (4.3) *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 μ whose columns have unit norm, and a noise level σ, define

$$\theta_{\text{crit}}(\mu, d, K, \sigma) = \frac{\sigma\sqrt{2\log d}}{1-(2K-1)\mu}. \tag{4.6}$$

Notice that $\theta_{\text{crit}}(\mu, d, K, \sigma)$ is well defined under the MIP condition (4.3).

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 μ and K-sparse vectors $\boldsymbol{\theta} \in \mathbb{R}^d$ for which given $\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \sigma\xi$, 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, (4.1) is equivalent to

$$\mathbf{y}^{(m)} = \tilde{\mathbf{X}}^{(m)}\tilde{\boldsymbol{\theta}}^{(m)} + \sigma\xi^{(m)}, \tag{4.7}$$

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

$$\min_m \|\mathbf{x}_k^{(m)}\|\,|\theta_k| \geq \tilde{\theta}_{\min}, \quad \forall k \in \mathcal{S}. \tag{4.8}$$

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

$$r = \left(\frac{\tilde{\theta}_{\min}}{\theta_{\text{crit}}(\mu_{\max}, d, K, \sigma)}\right)^2. \tag{4.9}$$

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 (4.9), 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 $\tilde{\theta}_{\min}^{(c)}$ is a value such that for all support indices $k \in \mathcal{S}$, $|\theta_k| \geq \tilde{\theta}_{\min}^{(c)}/\|\mathbf{X}_k\|$, 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 (4.6) for θ_{crit}, an SNR regime that is interesting to study in the distributed setting is

$$\frac{1}{M}\left(\frac{1-(2K-1)\mu_{\max}}{1-(2K-1)\mu(\mathbf{X})}\right)^2 < r < 1. \tag{4.10}$$

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. (4.10), the DJ-OMP scheme can still achieve exact support recovery.

4.3.1 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,

$$\delta = \delta\left(K, \mu_{\max}\right) = \tfrac{(K-1)\mu_{\max}^2}{1-(K-2)\mu_{\max}}. \tag{4.11}$$

As we show in Section 4.6.5, under the max-MIP condition (4.5), $\delta \leq \mu_{\max}$. Our first assumption is that the number of machines is sufficiently large, with the dependence on K encoded in the quantity δ.

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

$$M_c\left(d, K, \mu_{\max}, r\right) = K\left\lceil \frac{16\log d}{\Phi^c\left(\frac{(1-\sqrt{r})\sqrt{2\log d}}{\sqrt{1-\delta}(1-\mu_{\max})}\right)} \right\rceil. \tag{4.12}$$

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

$$M_c\left(d, K, \mu_{\max}, r\right) \approx K d^{\left(\frac{1-\sqrt{r}}{\sqrt{1-\delta}(1-\mu_{\max})}\right)^2}, \tag{4.13}$$

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 ϵ leading to a lower SNR. However, for our proof to work, ϵ cannot be arbitrarily small, and we set it as follows.

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

$$\frac{\sqrt{\mu_{\max}+\delta}}{1+\sqrt{\mu_{\max}+\delta}} < \epsilon < 1. \tag{4.14}$$

Importantly, for $\mu_{\max}$ small, ϵ 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

$$Q_0\left(d, K\right) = \tfrac{\log\left(88\sqrt{2}K\right)}{\log d}, \tag{4.15}$$

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

$$Q_1 = \tfrac{1-(1-\mu_{\max})\sqrt{1-\delta}\left((1-\epsilon)\sqrt{1-\mu_{\max}}-\sqrt{Q_0}\right)}{1-2\mu_{\max}K\sqrt{1-\delta}\frac{1-\mu_{\max}}{1-(2K-1)\mu_{\max}}}, \tag{4.16}$$

$$Q_2 = \tfrac{\sqrt{2+2(\mu_{\max}+\delta)}\left(1+\sqrt{1-\delta}(1-\mu_{\max})\sqrt{Q_0}\right)}{\sqrt{1-\delta}(1-\mu_{\max})+\sqrt{2+2(\mu_{\max}+\delta)}}. \tag{4.17}$$

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

$$\sqrt{r} \geq \begin{cases} Q_2 & (4K-1)\mu_{\max} - 2K\mu_{\max}^2 \geq 1 \\ \min(Q_1, Q_2) & \text{otherwise} \end{cases} \tag{4.18}$$

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* (4.5) *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* (4.8)*,*

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

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 4.1 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 4.7.*

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.

4.4 Simulation Results

We compare experimentally the following algorithms, which have different computation and communication costs (see Table 4.1): (i) Deb-Lasso where each machine computes a debiased-Lasso estimate

Table 4.1: Communication and Computation Costs

Algorithm	**Communication cost**	**Computational cost,** $K \ll d^{1/3}$
Single `OMP`	$\tilde{O}(K)$	$O(ndK)$
`Deb-Lasso`	$\tilde{O}(d)^2$	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`		

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 σ.

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 Σ 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 ℓ_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. Further implementation details appear in Appendix 4.9. Code that reproduces the results presented in this chapter is publicly available on GitHub.[3]

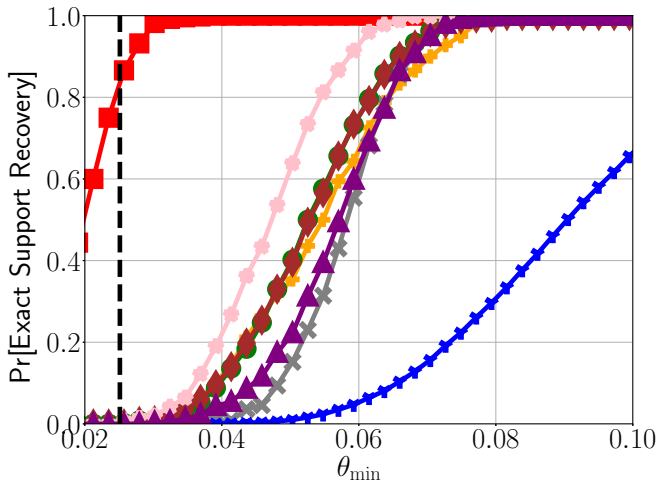
Figure 4.1 illustrates the empirical success probability of the various algorithms as a function of $\theta_{\min}$ in the two settings outlined above. Formally, for an algorithm A,

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

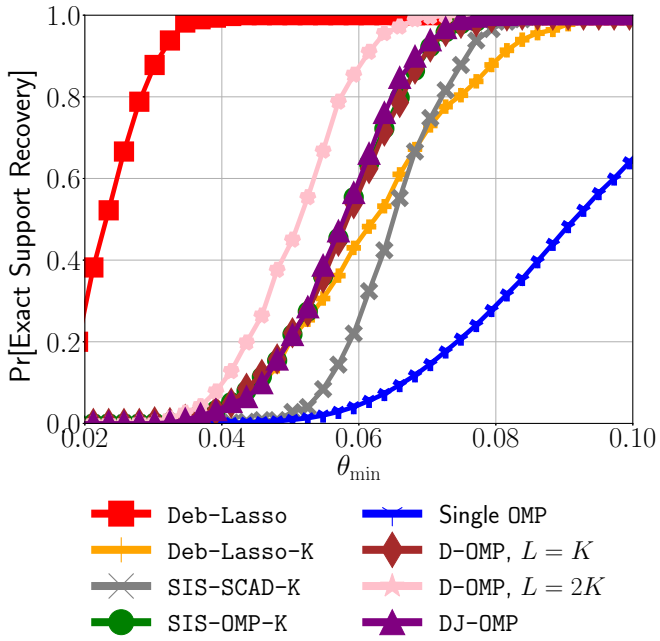
where $S_j^A(\theta_{\min})$ is the support set computed by algorithm A, for noise realization j and lower bound $\theta_{\min}$ on the non-zero coefficients of $\boldsymbol{\theta}$, and J is the total number of noise realizations, set to $J = 500$. The dashed vertical line in panel (a) is the lower bound $\theta_{\text{crit}}(\mu(\mathbf{X}), d, K, \sigma)$ of Eq. (4.6), 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.

[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.

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



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



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

Figure 4.1: Support Recovery as a Function of $\theta_{\min}$.

Figure 4.1 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

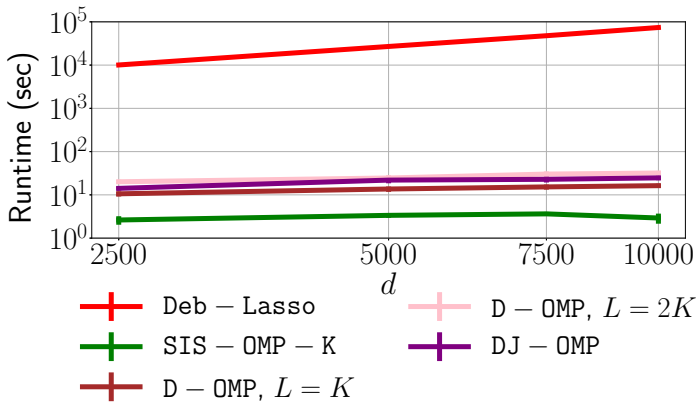


Figure 4.2: Runtime as a Function of d.

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 4.2 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, $\alpha = 0$ 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 4.8 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 (4.13).

4.5 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 ℓ_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.

4.6 Proofs

In this section we prove Theorem 4.1. For ease of presentation, in Section 4.6.1 we state and prove Theorem 4.2 which addresses the simpler case $K = 1$. The proof of Theorem 4.1 for the general case $K \geq 1$ appears in Section 4.6.2. The proofs of various auxiliary lemmas appear in Sections 4.6.3-4.6.6.

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.,

$$\hat{\boldsymbol{\theta}}^{(m)} = \arg\min_{\mathbf{z}\in\mathbb{R}^d} \left\| \mathbf{y}^{(m)} - \mathbf{X}^{(m)}\mathbf{z} \right\|_2 \text{ s.t. } \operatorname{supp}(\mathbf{z}) = s. \tag{4.19}$$

Then, it calculates the corresponding residual vector

$$\mathbf{r}^{(m)} = \mathbf{y}^{(m)} - \mathbf{X}^{(m)}\hat{\boldsymbol{\theta}}^{(m)}. \tag{4.20}$$

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

$$j^{(m)} = \arg\max_{i\in[d]} |\langle \tilde{\mathbf{x}}_i^{(m)}, \mathbf{r}^{(m)} \rangle|, \tag{4.21}$$

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 7, 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_{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

$$\tilde{\theta}_{\max}^{(m)} = \tilde{\theta}_{\max}^{(m)}(s) = \max_{k\in\mathcal{S}\setminus s} \left\{ \left\| \mathbf{x}_k^{(m)} \right\| |\theta_k| \right\}. \tag{4.22}$$

Similar to the definition of r in Eq. (4.9), we define

$$\rho^{(m)} = \rho^{(m)}(s) = \left(\frac{\tilde{\theta}_{\max}^{(m)}}{\theta_{\text{crit}}(\mu_{\max}, d, K, \sigma)} \right)^2, \tag{4.23}$$

where θ_{crit} is defined in Eq. (4.6). 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 4.3, 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. (4.12). 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,

$$\mathbf{v}_k > \max_{j \notin \mathcal{S}} \mathbf{v}_j. \tag{4.24}$$

A sufficient condition for (4.24) to occur is that for some suitable threshold $t_c = t_c(s) > 0$, both

$$\mathbf{v}_k > t_c, \tag{4.25}$$

and

$$\max_{j \notin \mathcal{S}} \mathbf{v}_j < t_c. \tag{4.26}$$

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,

$$F\left(d, K, \mu_{\max}, r\right) = \frac{1}{2}\Phi^c\left(\frac{(1-\sqrt{r})\sqrt{2\log d}}{\sqrt{1-\delta}(1-\mu_{\max})}\right). \tag{4.27}$$

Note that by this definition, Eq. (4.12) can be rewritten as

$$M_c\left(d, K, \mu_{\max}, r\right) = K\left\lceil \frac{8\log d}{F\left(d, K, \mu_{\max}, r\right)} \right\rceil. \tag{4.28}$$

We will show that Eqs. (4.25) and (4.26) indeed hold with high probability with the following threshold

$$t_c = t_c(s) = \frac{\sum_{m\in[M]} F\left(d, K, \mu_{\max}, \rho^{(m)}(s)\right)}{M \cdot F\left(d, K, \mu_{\max}, r\right)} 4\log d, \tag{4.29}$$

where r, $\rho^{(m)}$ and F are defined in Eqs. (4.9), (4.23), and (4.27) 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 4.1 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. (4.27) and appears in Section 4.6.3.

Lemma 4.1. *Under the max-MIP condition* (4.5)*, for any fixed $s \subset \mathcal{S}$, the threshold $t_c = t_c(s)$ defined in Eq.* (4.29) *satisfies*

$$t_c \geq 4\log d. \tag{4.30}$$

The following Lemma 4.2 states that if the expected number of votes for an index $k \in \mathcal{S} \setminus s$ is sufficiently high, then event (4.25) occurs with high probability. The next Lemma 4.3 shows that if the expected number of votes for each non-support index $j \notin \mathcal{S}$ is sufficiently low, then event (4.26) occurs with high probability. These lemmas follow from Chernoff bounds and are proved in Section 4.6.3 as well.

Lemma 4.2. *Assume the max-MIP condition* (4.5) *holds. Fix $s \subset \mathcal{S}$, and let $t_c = t_c(s)$ be given by Eq.* (4.29)*. If $\mathrm{E}[\mathbf{v}_k] \geq 2t_c$ for some $k \in \mathcal{S} \setminus s$, then*

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

Lemma 4.3. *Assume the max-MIP condition* (4.5) *holds. Fix $s \subset \mathcal{S}$, and let $t_c = t_c(s)$ be given by Eq.* (4.29)*. If for all non-support indices $j \notin \mathcal{S}$ it holds that $\mathrm{E}[\mathbf{v}_j] \leq \frac{t_c}{5}$ then*

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

It remains to bound $\mathrm{E}[\mathbf{v}_j]$ 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

$$p_j^{(m)} = \Pr\left[j^{(m)} = j\right], \tag{4.31}$$

where $j^{(m)}$ is defined in Eq. (4.21).

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 4.6.1, and then extend them to the general case $K \geq 1$ in Section 4.6.2.

4.6.1 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. (4.11), $\delta(1, \mu_{\max}) = 0$. Hence, the quantity F simplifies to

$$F(d, 1, \mu_{\max}, r) = \frac{1}{2}\Phi^c\left(\frac{1-\sqrt{r}}{1-\mu_{\max}}\sqrt{2\log d}\right), \tag{4.32}$$

and the quantity M_c from Eq. (4.28) reduces to

$$M_c(d, 1, \mu_{\max}, r) = \left\lceil \frac{8\log d}{F(d, 1, \mu_{\max}, r)} \right\rceil. \tag{4.33}$$

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

Assumption 4.4. $M \geq M_c(d, 1, \mu_{\max}, r)$.

Assumption 4.5. *The parameter $\epsilon = \epsilon(\mu_{\max})$ satisfies*

$$\frac{\sqrt{\mu_{\max}}}{1+\sqrt{\mu_{\max}}} < \epsilon < 1. \tag{4.34}$$

The quantity Q_0 reduces to

$$Q_0(d, 1) = \frac{\log\left(88\sqrt{2}\right)}{\log d}. \tag{4.35}$$

In addition, the expressions for Q_1 and Q_2 simplify to

$$Q_1(d, 1, \mu_{\max}, \epsilon) = \frac{1-(1-\mu_{\max})\left((1-\epsilon)\sqrt{1-\mu_{\max}} - \sqrt{Q_0}\right)}{1-2\mu_{\max}}, \tag{4.36}$$

$$Q_2(d, 1, \mu_{\max}) = \frac{\sqrt{2+2\mu_{\max}}\left(1+(1-\mu_{\max})\sqrt{Q_0}\right)}{1-\mu_{\max}+\sqrt{2+2\mu_{\max}}}. \tag{4.37}$$

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

Assumption 4.6 (SNR Condition)**.** *The SNR is sufficiently high,*

$$\sqrt{r} \geq \begin{cases} Q_2 & \mu_{\max} \geq 1/2 \\ \min(Q_1, Q_2) & \textit{otherwise} \end{cases} \tag{4.38}$$

Theorem 4.2. *Under Assumptions 4.4-4.6 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 4.3, when $\mu_{\max} \ll 1$ condition (4.38) 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.

Proof of Theorem 4.2

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

$$j^{(m)} = \arg\max_{i \in [d]} |\langle \tilde{\mathbf{x}}_i^{(m)}, \mathbf{y}^{(m)} \rangle|. \tag{4.39}$$

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 4.2, we derive a lower bound on the probability $p_k^{(m)}$ for the support index k in the following Lemma 4.4 and an upper bound on the probability $p_j^{(m)}$ for each non-support index $j \notin \mathcal{S}$ in the following Lemma 4.5. Their proofs appear in Section 4.6.4 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 4.4. *Assume that $\|\boldsymbol{\theta}\|_0 = K = 1$ and let $\mathcal{S} = \{k\} = \operatorname{supp}\{\boldsymbol{\theta}\}$. Further assume that the max-MIP condition* (4.5) *holds. For sufficiently large d, for each machine m,*

$$p_k^{(m)} \geq F\left(d, 1, \mu_{\max}, \rho^{(m)}\right), \tag{4.40}$$

where $p_k^{(m)}$ and F are defined in Eqs. (4.31) *and* (4.32) *respectively.*

Lemma 4.5. *Assume that $\|\boldsymbol{\theta}\|_0 = K = 1$ and let $\mathcal{S} = \operatorname{supp}\{\boldsymbol{\theta}\}$. Further assume that $\rho^{(m)}$ of Eq.* (4.23) *satisfies $\rho^{(m)} \leq 1$ for each machine m and that the max-MIP condition* (4.5) *holds. If ϵ satisfies Assumption 4.5, the SNR parameter r satisfies Assumption 4.6, and the dimension $d = d(\epsilon)$ is sufficiently large, then for each machine m and each non-support index $j \notin \mathcal{S}$,*

$$p_j^{(m)} \leq \frac{F\left(d, 1, \mu_{\max}, \rho^{(m)}\right)}{11}. \tag{4.41}$$

We now formally prove Theorem 4.2 by combining the above lemmas.

Proof of Theorem 4.2. For simplicity, we assume that the number of machines is $M = M_c\left(d, 1, \mu_{\max}, r\right)$, since a larger number of machines would only increase the probability of successful support recovery. We first analyze the probability that event (4.25) occurs. By Lemma 4.4, for the support index $k \in \mathcal{S}$, its expected number of votes is $\mathrm{E}\left[\mathbf{v}_k\right] = \sum_{m \in [M]} p_k^{(m)} \geq \sum_{m \in [M]} F\left(d, 1, \mu_{\max}, \rho^{(m)}\right)$. By the definitions of t_c in Eq. (4.29) and M_c in Eq. (4.33),

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

By Lemma 4.2, the event (4.25) occurs with probability at least $1 - 1/d$.

Next, we analyze the probability that event (4.26) occurs. Fix a non-support index $j \notin \mathcal{S}$. Since $\rho^{(m)} \leq 1$, then by Lemma 4.5, 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. (4.29) and M_c in Eq. (4.33),

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

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. (4.32), 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 4.3, the event (4.26) occurs with probability at least $1 - 1/d$. A union bound completes the proof. □

4.6.2 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 4.6. *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 (4.5), 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,*

$$\Pr[j_{center}(s) \notin \mathcal{S}] \leq 2d^{-1}. \tag{4.42}$$

First, let us show how Theorem 4.1 follows directly from Lemma 4.6.

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,

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

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

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

Since now the set s_{t-1} is fixed, we can bound each term via Lemma 4.6. This gives

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

which completes the proof. □

Next, we prove Lemma 4.6. Since $s \subset \mathcal{S}$, we need to bound the probability $p_j^{(m)}$ of Eq. (4.31) 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 4.7, lower bounds a different quantity $q^{(m)}$ defined as the probability that the index sent by machine m belongs to the support $\mathcal{S} \setminus s$,

$$q^{(m)} = q^{(m)}(s) = \Pr\left[j^{(m)} \in \mathcal{S} \setminus s\right]. \tag{4.43}$$

Lemma 4.8 upper bounds $p_j^{(m)}$ for each $j \notin \mathcal{S}$. Their proofs appear in Section 4.6.5.

Lemma 4.7. *Assume that the max-MIP condition* (4.5) *holds. For each machine m, for sufficiently large d,*

$$q^{(m)} \geq F\left(d, K, \mu_{\max}, \rho^{(m)}\right), \tag{4.44}$$

where $q^{(m)}$ and F are defined in Eqs. (4.43) *and* (4.27) *respectively.*

Lemma 4.8. *Assume that $\rho^{(m)}$ of Eq.* (4.23) *satisfies $\rho^{(m)} \leq 1$ for each machine m and that the max-MIP condition* (4.5) *holds. If ϵ 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 $j \notin \mathcal{S}$,*

$$p_j^{(m)} \leq \frac{F\left(d, K, \mu_{\max}, \rho^{(m)}\right)}{11K}, \tag{4.45}$$

where $p_j^{(m)}$ and F are defined in Eqs. (4.31) *and* (4.27) *respectively.*

We now formally prove Lemma 4.6 by combining the above lemmas.

Proof of Lemma 4.6. 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 (4.25) 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 4.7, 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. (4.29),

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

By definition of M_c in Eq. (4.28),

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

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 4.2, the event (4.25) occurs with probability at least $1 - 1/d$.

Similarly to the proof of Theorem 4.2, Lemmas 4.3 and 4.8 imply that the event (4.26) 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 4.8. A union bound completes the proof. □

4.6.3 Proofs of Lemmas 4.1, 4.2 and 4.3

We first prove Lemma 4.1 and then use it to prove Lemmas 4.2 and 4.3.

Proof of Lemma 4.1. By its definition in Eq. (4.27), the function F is monotonic increasing in its fourth argument. Next, by Eq. (4.22), $\tilde{\theta}_{\min} \leq \tilde{\theta}_{\max}^{(m)}$, 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. (4.29) concludes the proof. □

In the proofs below we use the following Chernoff bounds.

Lemma 4.9 (Chernoff (1952))**.** *Suppose $X_1, \ldots, X_d$ are independent Bernoulli random variables and let X denote their sum. Then, for any $\phi \geq 0$,*

$$\Pr\left[X \geq (1+\phi)\,\mathrm{E}\left[X\right]\right] \leq e^{-\frac{\phi^2 \mathrm{E}[X]}{2+\phi}}, \tag{4.46}$$

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

$$\Pr\left[X \leq (1-\phi)\,\mathrm{E}\left[X\right]\right] \leq e^{-\frac{\phi^2 \mathrm{E}[X]}{2}}. \tag{4.47}$$

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 $\left\{\xi^{(m)}\right\}_{m\in[M]}$ 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 4.9 and 4.1 to prove Lemmas 4.2 and 4.3.

Proof of Lemma 4.2. By the discussion above, we may apply the Chernoff bound (4.47) to the sum $\mathbf{v}_k$. Using the assumption $E_k \geq 2t_c$ and Lemma 4.1, 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 4.3. 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 (4.46) 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 4.1, the above probability is smaller than d^{-2}, and by applying a union bound we conclude that

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

□

4.6.4 Proofs of Lemmas 4.4 and 4.5

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.,

$$\mathbf{P}_{\mathcal{I}}^{(m)} = \tilde{\mathbf{X}}_{|\mathcal{I}}^{(m)}\left(\tilde{\mathbf{X}}_{|\mathcal{I}}^{(m)}\right)^\dagger. \tag{4.48}$$

For simplicity of notation, in Sections 4.6.4-4.6.6 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 4.10), a technical lemma regarding the Gaussian distribution, Lemma 4.11, whose proof appears in Section 4.6.6, and Lemma 4.12, which bounds the left tail probability of the maximum of correlated Gaussian random variables (Lopes and Yao, 2022).

Lemma 4.10 (Gaussian tail bounds (Gordon, 1941))**.** *For any $t > 0$,*

$$\frac{t}{\sqrt{2\pi}(t^2+1)} e^{-t^2/2} \leq \Phi^c(t) \leq \frac{1}{\sqrt{2\pi}t} e^{-t^2/2}. \tag{4.49}$$

Lemma 4.11. *For any $a, b \geq 0$,*

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

Lemma 4.12 ((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 (η, ζ) such that*

$$\Pr\left[\max_{i\in[d]} Z_i < \zeta\sqrt{2(1-\eta)\log d}\right] \leq C d^{-\frac{(1-\eta)(1-\zeta)^2}{\eta}} (\log d)^{\frac{1-\eta(2-\zeta)-\zeta}{2\eta}}. \tag{4.50}$$

To put Lemma 4.12 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 η is an upper bound on the correlation. We use this result with $\eta = \mu_{\max}$ and $\zeta = 1 - \epsilon$, where ϵ satisfies Assumption 4.2, in order to bound the probability that a non-support index is sent to the center and prove Lemma 4.5.

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. (4.7) its response vector $\mathbf{y} = \mathbf{y}^{(m)}$ admits the following form

$$\mathbf{y} = \tilde{\theta}_k \tilde{\mathbf{x}}_k + \sigma\xi. \tag{4.51}$$

Recall that by its definition in Eq. (4.22), $\tilde{\theta}_{\max} = \|\mathbf{x}_k\| \, |\theta_k| = |\tilde{\theta}_k|$. By Eq. (4.23) for ρ and Eq. (4.6) for θ_{crit} with $K = 1$,

$$\tilde{\theta}_{\max} = \frac{\sigma\sqrt{2\rho\log d}}{1-\mu_{\max}}. \tag{4.52}$$

We now prove the lemmas.

Proof of Lemma 4.4. Recall that p_k, defined in Eq. (4.31), 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,

$$|\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle| \geq \max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y}\rangle|. \tag{4.53}$$

where $\mathbf{y}$ is given by (4.51). To this end, we decompose the noise ξ in Eq. (4.51) as the sum of two components, the first $\xi_\parallel = \mathbf{P}_k\xi = \langle \tilde{\mathbf{x}}_k, \xi\rangle \tilde{\mathbf{x}}_k$ is parallel to $\tilde{\mathbf{x}}_k$, namely $\langle \tilde{\mathbf{x}}_k, \xi_\parallel\rangle = \langle \tilde{\mathbf{x}}_k, \xi\rangle$, and the second $\xi_\perp = \xi - \xi_\parallel = (\mathbf{I} - \mathbf{P}_k)\xi$, is orthogonal to $\tilde{\mathbf{x}}_k$, i.e., $\langle \tilde{\mathbf{x}}_k, \xi_\perp\rangle = 0$.

Next, we use this decomposition to bound the two terms in (4.53). Combining the expression (4.51) for $\mathbf{y}$, the decomposition of ξ and the fact that $\tilde{\theta}_{\max} = |\tilde{\theta}_k|$, the LHS of (4.53) can be bounded by

$$\begin{aligned} |\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle| &\geq \operatorname{sign}\left(\tilde{\theta}_k\right) \langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle = \operatorname{sign}\left(\tilde{\theta}_k\right) \left(\tilde{\theta}_k \langle \tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_k\rangle + \sigma \langle \tilde{\mathbf{x}}_k, \xi\rangle\right) \\ &= \tilde{\theta}_{\max} + \sigma \operatorname{sign}\left(\tilde{\theta}_k\right) \langle \tilde{\mathbf{x}}_k, \xi_\parallel\rangle. \end{aligned} \tag{4.54}$$

Similarly, the RHS of (4.53) can be bounded by

$$\begin{aligned} \max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y}\rangle| &= \max_{i\notin\mathcal{S}} \left|\tilde{\theta}_k \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k\rangle + \sigma \langle \tilde{\mathbf{x}}_i, \xi_\parallel + \xi_\perp\rangle\right| \\ &\leq \left(\tilde{\theta}_{\max} + \sigma |\langle \tilde{\mathbf{x}}_k, \xi\rangle|\right) \max_{i\notin\mathcal{S}} \{|\langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k\rangle|\} + \sigma \max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \xi_\perp\rangle| \\ &\leq \left(\tilde{\theta}_{\max} + \sigma \left|\langle \tilde{\mathbf{x}}_k, \xi_\parallel\rangle\right|\right) \mu_{\max} + \sigma \max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \xi_\perp\rangle|. \end{aligned} \tag{4.55}$$

where the second step follows from the triangle inequality and the definitions of $\tilde{\theta}_{\max}$ and $\xi_{\|}$, and the last step follows from the definition of $\mu_{\max}$. Combining Eq. (4.54) with Eq. (4.55) implies that a sufficient condition for (4.53) to hold is that

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

By Eq. (4.52), the above event may be written as

$$\max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \xi_\perp\rangle| \leq \operatorname{sign}\left(\tilde{\theta}_k\right) \langle \tilde{\mathbf{x}}_k, \xi_{\|}\rangle - \mu_{\max} \left|\langle \tilde{\mathbf{x}}_k, \xi_{\|}\rangle\right| + \sqrt{2\rho\log d}. \tag{4.56}$$

A key property is that $\xi_{\|}$ and $\xi_\perp$ 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,

$$\Pr\left[A\leq B\right] \geq \Pr\left[A\leq T\cap B\geq T\right] = \Pr\left[A\leq T\right]\cdot\Pr\left[B\geq T\right]. \tag{4.57}$$

Thus,

$$p_k \geq \Pr[A\leq T]\cdot\Pr]B\geq T] \tag{4.58}$$

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 $\langle\tilde{\mathbf{x}}_i,\xi_\perp\rangle$. We may write $\tilde{\mathbf{x}}_i = \mathbf{P}_k\tilde{\mathbf{x}}_i + (\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i$ Since $\xi_\perp = (\mathbf{I}-\mathbf{P}_k)\,\xi$, then $\langle\mathbf{P}_k\tilde{\mathbf{x}}_i,\xi_\perp\rangle = 0$, and $\langle\tilde{\mathbf{x}}_i,\xi_\perp\rangle = \langle(\mathbf{I}-\mathbf{P}_k)\,\tilde{\mathbf{x}}_i,\xi_\perp\rangle$. Normalizing the inner product by the norm of $(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i$ yields a standard normal random variable $Z_i = \frac{\langle\tilde{\mathbf{x}}_i,\xi_\perp\rangle}{\|(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i\|_2}\sim\mathcal{N}(0,1)$. By the definition of $\mu_{\max}$,

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

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

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

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 (4.49) 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

$$\Pr\left[A\leq\sqrt{2\log d}\right] \geq \left(1-\frac{\gamma_1}{\sqrt{\pi\log d}}d^{-1/\gamma_1^2}\right)^{d-1} \geq 1-\frac{\gamma_1}{\sqrt{\pi\log d}}d^{1-1/\gamma_1^2} \geq \frac{1}{2}, \tag{4.59}$$

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

We now bound $\Pr\left[B\geq T\right]$, where B is the RHS of (4.56). Since $\tilde{\mathbf{x}}_k$ has unit norm, by the definition of $\xi_{\|}$, then $Z=\langle\tilde{\mathbf{x}}_k,\xi_{\|}\rangle = \langle\tilde{\mathbf{x}}_k,\xi\rangle\sim\mathcal{N}(0,1)$. By the law of total probability,

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

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

$$\Pr\left[B \geq \sqrt{2\log d}\right] \geq \frac{1}{2}\Pr\left[\gamma_2 |Z| \geq \sqrt{2\log d} - \sqrt{2\rho\log d}\right] \geq \Phi^c\left(\frac{1-\sqrt{\rho}}{\gamma_2}\sqrt{2\log d}\right). \tag{4.60}$$

Inserting (4.59) and (4.60) with $\gamma_2 = 1 - \mu_{\max}$ into (4.58) and recalling the definition of F in (4.32) completes the proof of Lemma 4.4. □

Proof of Lemma 4.5. Fix a non-support index $j \notin \mathcal{S}$. Recall that p_j, defined in Eq. (4.31), 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.,

$$p_j = \Pr\left[|\langle\tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > \max_{i\neq j} |\langle\tilde{\mathbf{x}}_i, \mathbf{y}\rangle|\right]. \tag{4.61}$$

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

$$\Pr\left[|\langle\tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > \max_{i\notin\{k,j\}} |\langle\tilde{\mathbf{x}}_i, \mathbf{y}\rangle|\right] \tag{4.62}$$

and

$$\Pr\left[|\langle\tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > |\langle\tilde{\mathbf{x}}_k, \mathbf{y}\rangle|\right], \tag{4.63}$$

and then use the following inequality to upper bound (4.61) 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

$$\Pr\left[A > \max\{B, C\}\right] \leq \min\left\{\Pr\left[A > B\right], \Pr\left[A > C\right]\right\}. \tag{4.64}$$

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

$$|\langle\tilde{\mathbf{x}}_j, \mathbf{y}\rangle| = \left|\tilde{\theta}_k \langle\tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_k\rangle + \sigma\langle\tilde{\mathbf{x}}_j, \xi\rangle\right| \leq \tilde{\theta}_{\max}\mu_{\max} + \sigma|\langle\tilde{\mathbf{x}}_j, \xi\rangle|. \tag{4.65}$$

We first bound (4.62). For each non-support index $i \notin \mathcal{S}$ such that $i \neq j$,

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

Combining this with Eq. (4.65), rearranging terms, and recalling the relation between $\theta_{\max}$ and ρ in (4.52) yields

$$\Pr\left[|\langle\tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > \max_{i\notin\{k,j\}} |\langle\tilde{\mathbf{x}}_i, \mathbf{y}\rangle|\right] \leq \Pr\left[|\langle\tilde{\mathbf{x}}_j, \xi\rangle| + 2\mu_{\max}\frac{\sqrt{2\rho\log d}}{1-\mu_{\max}} > \max_{i\notin\{k,j\}}\langle\tilde{\mathbf{x}}_i, \xi\rangle\right]. \tag{4.66}$$

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

$$\Pr\left[D > E\right] \leq \Pr\left[D \geq T\right] + \Pr\left[E < T\right]. \tag{4.67}$$

Applying this inequality with $T = (1-\epsilon)\sqrt{2\left(1-\mu_{\max}\right)\log d}$ and $\epsilon \in (0, 1)$ as in Eq. (4.34), we can upper bound (4.66) by

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

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

$$2\Phi^c\left(a\sqrt{2\log d}\right). \tag{4.68}$$

We now bound the second term. It involves the maximum of $d-2$ correlated Gaussians, whose covariance matrix Σ has $\Sigma_{ii}=1$ for all i, and $\Sigma_{ij} = \mathrm{Cov}\left(\langle \tilde{\mathbf{x}}_i, \xi\rangle, \langle \tilde{\mathbf{x}}_l, \xi\rangle\right) = \langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_l\rangle \le \mu_{\max}$. Hence, we can apply Lemma 4.12 with $\eta = \mu_{\max}$ and $\zeta = 1-\epsilon$, which gives the following bound

$$C\,(d-2)^{-\frac{1-\mu_{\max}}{\mu_{\max}}\epsilon^2}\left(\log(d-2)\right)^{\frac{\epsilon-\mu_{\max}(1+\epsilon)}{2\mu_{\max}}}. \tag{4.69}$$

We now show that (4.68) is larger than (4.69), and thus

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > \max_{i\notin\{k,j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{y}\rangle|\right] \le 4\Phi^c\left(a\sqrt{2\log d}\right). \tag{4.70}$$

First note that if ρ is sufficiently large such that $a \le 0$, then (4.68) is larger than 1, and thus larger than (4.69). Otherwise, $a>0$ and using the lower bound for the Gaussian tail of (4.49), we may lower bound (4.68) by $d^{-a^2-o(1)}$, where $o(1)$ hides factors that are asymptotically smaller than 1. The term (4.69) 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 $d = d(\epsilon)$, (4.68) is larger than (4.69). Indeed, under condition (4.34), $\epsilon = \frac{\sqrt{\mu_{\max}}}{1+\sqrt{\mu_{\max}}} + \epsilon_0$ 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 (4.34) implies that (4.62) can be bounded by (4.70).

We now bound (4.63). For the support index k, by (4.51),

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

Combining this with (4.65) and plugging $\tilde{\theta}_{\max}$ in Eq. (4.52), the probability (4.63) is upper bounded by

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \xi\rangle| - \mathrm{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \xi\rangle > \sqrt{2\rho\log d}\right]. \tag{4.71}$$

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

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

$$\Pr\left[|H| - G > c\right] \le \Pr\left[H - G > c\right] + \Pr\left[-H - G > c\right]. \tag{4.72}$$

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 $\sigma_{HG} = \mathrm{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \tilde{\mathbf{x}}_j\rangle$. Since $|\sigma_{HG}| \le \mu_{\max}$, then by (4.72),

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

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

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{y}\rangle| > |\langle \tilde{\mathbf{x}}_k, \mathbf{y}\rangle|\right] \le 2\Phi^c\left(\sqrt{\frac{2\rho\log d}{2+2\mu_{\max}}}\right). \tag{4.73}$$

By Eq. (4.64), the probability (4.61) is at most the minimum between (4.70) and (4.73). By the monotonicity of the Gaussian CDF, it is upper bounded by

$$4\Phi^c\left(\max\left\{\left((1-\epsilon)\sqrt{1-\mu_{\max}}-\frac{2\mu_{\max}\sqrt{\rho}}{1-\mu_{\max}}\right),\sqrt{\frac{\rho}{2+2\mu_{\max}}}\right\}\sqrt{2\log d}\right). \tag{4.74}$$

Finally, to prove (4.41) of the lemma, we note that with Q_1 and Q_2 defined in Eqs. (4.36) and (4.37) respectively, by splitting to cases and applying some algebraic manipulations[4], condition (4.38) implies that

$$\frac{1-\sqrt{r}}{1-\mu_{\max}}+\sqrt{Q_0}<\max\left\{\left((1-\epsilon)\sqrt{1-\mu_{\max}}-\frac{2\mu_{\max}\sqrt{r}}{1-\mu_{\max}}\right),\sqrt{\frac{r}{2+2\mu_{\max}}}\right\}. \tag{4.75}$$

The definitions of r and ρ in Eqs. (4.9) and (4.23) imply that $\rho \geq r$. Thus, ρ satisfies condition (4.38) and hence condition (4.75). The RHS of (4.75) is the same as the maximum in (4.74) above. Thus, (4.74) is upper bounded by

$$4\Phi^c\left(\left(\frac{1-\sqrt{\rho}}{1-\mu_{\max}}+\sqrt{Q_0}\right)\sqrt{2\log d}\right). \tag{4.76}$$

Since $\rho \leq 1$, we can apply Lemma 4.11. Hence, by the definition of Q_0 in Eq. (4.35), and by the definition of F in Eq. (4.32),

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

which completes the proof of Lemma 4.5. □

4.6.5 Proof of Lemmas 4.7 and 4.8

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 4.6.4. Recall that given an input support set s, each machine m estimates its vector $\hat{\boldsymbol{\theta}}^{(m)}$ by solving the least squares problem (4.19). Thus, $\text{supp}\left(\hat{\boldsymbol{\theta}}^{(m)}\right)=s$ and

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

Denote by $\tilde{\boldsymbol{\xi}}^{(m)}$ the projection of the noise $\xi^{(m)}$ to the subspace orthogonal to the span of the columns of $\mathbf{X}^{(m)}_{|s}$, i.e., $\tilde{\boldsymbol{\xi}}^{(m)}=\left(\mathbf{I}-\mathbf{P}^{(m)}_{s}\right)\xi^{(m)}$. Given that $s\subset\mathcal{S}$, the residual $\mathbf{r}^{(m)}$ defined in Eq. (4.20) can

[4]First, consider the case $\mu_{\max}\geq 1/2$. By the max-MIP condition (4.5), $\mu_{\max}<1$, 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 (4.75) is smaller than $\sqrt{\frac{r}{2+2\mu_{\max}}}$ and thus smaller than the RHS of (4.75). Now consider the case $\mu_{\max}<1/2$. By (4.38), $\sqrt{r}>Q_1$ or $\sqrt{r}>Q_2$. By the same reasoning, the latter implies that the LHS of (4.75) 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 (4.75) 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 (4.75) is smaller than the maximum between the aforementioned terms.

be written in the following form

$$\begin{aligned}\mathbf{r}^{(m)} &= \mathbf{y}^{(m)} - \mathbf{X}^{(m)}\hat{\boldsymbol{\theta}}^{(m)} = \mathbf{y}^{(m)} - \mathbf{X}^{(m)}_{|s}\hat{\boldsymbol{\theta}}^{(m)}_{|s} = \left(\mathbf{I} - \mathbf{X}^{(m)}_{|s}\left(\mathbf{X}^{(m)}_{|s}\right)^{\dagger}\right)\mathbf{y}^{(m)} \\ &= \left(\mathbf{I} - \mathbf{P}^{(m)}_{s}\right)\mathbf{y}^{(m)} = \left(\mathbf{I} - \mathbf{P}^{(m)}_{s}\right)\left(\tilde{\mathbf{X}}^{(m)}\tilde{\boldsymbol{\theta}}^{(m)} + \sigma\xi^{(m)}\right) \\ &= \left(\mathbf{I} - \mathbf{P}^{(m)}_{s}\right)\sum_{l\in\mathcal{S}\setminus s}\tilde{\theta}^{(m)}_{l}\tilde{\mathbf{x}}^{(m)}_{l} + \sigma\tilde{\boldsymbol{\xi}}^{(m)}, \end{aligned} \tag{4.77}$$

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. (4.7), 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

$$\mu_s = \mu_s\left(K_d, \mu_{\max}\right) = \frac{K_d\mu^2_{\max}}{1 - (K_d - 1)\mu_{\max}}. \tag{4.78}$$

The following Lemma 4.13 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 4.6.6.

Lemma 4.13. *Assume that the max-MIP condition* (4.5) *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 μ_s of Eq.* (4.78) *satisfies*

$$\mu_s \le \mu_{\max}, \tag{4.79}$$

and

$$K_u\left(\mu_{\max} + \mu_s\right) < K\mu_{\max}. \tag{4.80}$$

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

$$1 - \mu_s \le \left\|\left(\mathbf{I} - \mathbf{P}^{(m)}_{s}\right)\tilde{\mathbf{x}}^{(m)}_{i}\right\|^2_2 \le 1. \tag{4.81}$$

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

$$\left|\left\langle\tilde{\mathbf{x}}^{(m)}_{k}, \left(\mathbf{I} - \mathbf{P}^{(m)}_{s}\right)\tilde{\mathbf{x}}^{(m)}_{i}\right\rangle\right| \le \mu_{\max} + \mu_s, \tag{4.82}$$

and

$$\left\|\left(\mathbf{I} - \mathbf{P}^{(m)}_{s}\right)\left(\mathbf{I} - \mathbf{P}^{(m)}_{k}\right)\tilde{\mathbf{x}}^{(m)}_{i}\right\|^2_2 \ge 1 - \mu^2_{\max} - \mu_s\left(1 + \mu_{\max}\right)^2. \tag{4.83}$$

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

For future use, notice that by its definition in Eq. (4.78), μ_s is an increasing function of K_d. Since $K_d \le K - 1$, then the quantity δ of Eq. (4.11) satisfies

$$\delta\left(K, \mu_{\max}\right) = \mu_s\left(K - 1, \mu_{\max}\right) \ge \mu_s\left(K_d, \mu_{\max}\right). \tag{4.84}$$

In addition, by Eq. (4.79), under max-MIP condition (4.5), $\delta \le \mu_{\max} < 1$, and hence the quantities in Section 4.3 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. (4.22) and let $k \in \mathcal{S} \setminus s$ be an index for which

$$\|\tilde{\mathbf{x}}_k\| \cdot |\theta_k| = \tilde{\theta}_{\max} \tag{4.85}$$

(chosen arbitrarily in case of ties). By Eq. (4.23) for ρ and Eq. (4.6) for θ_{crit},

$$\tilde{\theta}_{\max} = \frac{\sigma\sqrt{2\rho\log d}}{1-(2K-1)\mu_{\max}}. \tag{4.86}$$

We now prove Lemmas 4.7 and 4.8.

Proof of Lemma 4.7. Recall that q, defined in Eq. (4.43), 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. (4.85) 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

$$|\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle| \geq \max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|. \tag{4.87}$$

Thus, to prove the lemma it suffices to lower bound the probability of event (4.87). Similarly to the proof of Lemma 4.4, we decompose the noise $\tilde{\boldsymbol{\xi}}$ in Eq. (4.77) 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 (4.87). By Eq. (4.77), 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. (4.81). 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 (4.82) and $\left|\tilde{\theta}_l\right| \leq \tilde{\theta}_{\max}$ by its definition in Eq. (4.22). Combining these bounds with the definition of $\tilde{\boldsymbol{\xi}}_{\|}$ implies that the LHS of (4.87) can be bounded by

$$\begin{aligned}
|\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle| &\geq \operatorname{sign}\left(\tilde{\theta}_k\right)\langle \tilde{\mathbf{x}}_k, \mathbf{r}\rangle \\
&\geq \tilde{\theta}_{\max}\left(\langle \tilde{\mathbf{x}}_k, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_k\rangle - \sum_{l\in\mathcal{S}\setminus(s\cup\{k\})} |\langle \tilde{\mathbf{x}}_k, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle|\right) + \operatorname{sign}\left(\tilde{\theta}_k\right)\sigma\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}\right\rangle \\
&\geq \tilde{\theta}_{\max}\left(1-\mu_s-(K_u-1)\left(\mu_{\max}+\mu_s\right)\right) + \sigma\operatorname{sign}\left(\tilde{\theta}_k\right)\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle.
\end{aligned} \tag{4.88}$$

The RHS of (4.87) can be bounded by

$$\begin{aligned}
\max_{i\notin\mathcal{S}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle| &= \max_{i\notin\mathcal{S}}\left|\sum_{l\in\mathcal{S}\setminus s}\tilde{\theta}_l\langle \tilde{\mathbf{x}}_i, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle + \sigma\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}+\tilde{\boldsymbol{\xi}}_{\|}\right\rangle\right| \\
&\leq \tilde{\theta}_{\max}\max_{i\notin\mathcal{S}}\sum_{l\in\mathcal{S}\setminus s}|\langle \tilde{\mathbf{x}}_i, (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_l\rangle| + \sigma\max_{i\notin\mathcal{S}}\left|\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}\right\rangle\right| + \sigma\left|\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}\right\rangle\right|\max_{i\notin\mathcal{S}}|\langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_k\rangle| \\
&\leq K_u\tilde{\theta}_{\max}\left(\mu_{\max}+\mu_s\right) + \sigma\max_{i\notin\mathcal{S}}\left|\left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}_{\perp}\right\rangle\right| + \sigma\mu_{\max}\left|\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}_{\|}\right\rangle\right|.
\end{aligned} \tag{4.89}$$

where the first step follows from Eq. (4.77) 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. (4.82) and the definitions of $\mu_{\max}$ in Eq. (4.4). Combining Eq. (4.88) with Eq. (4.89) implies that a sufficient condition for (4.87) 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. (4.86) and by the inequality (4.80), a sufficient condition for the previous event to occur is

$$\max_{i\notin\mathcal{S}}\left|\left\langle\tilde{\mathbf{x}}_i,\tilde{\boldsymbol{\xi}}_\perp\right\rangle\right| \le \operatorname{sign}\left(\tilde{\theta}_k\right)\left\langle\tilde{\mathbf{x}}_k,\tilde{\boldsymbol{\xi}}_\parallel\right\rangle - \mu_{\max}\left|\left\langle\tilde{\mathbf{x}}_k,\tilde{\boldsymbol{\xi}}_\parallel\right\rangle\right| + \sqrt{2\rho\log d}. \tag{4.90}$$

As in the proof of Lemma 4.4, denote the LHS of (4.90) by A, its RHS by B and let $T=\sqrt{2\log d}$. By Eq. (4.57), 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 $\tilde{\boldsymbol{\xi}}_\perp=(\mathbf{I}-\mathbf{P}_k)\tilde{\boldsymbol{\xi}}=(\mathbf{I}-\mathbf{P}_k)(\mathbf{I}-\mathbf{P}_s)\xi$. By the symmetry of projections, $\left\langle\tilde{\mathbf{x}}_i,\tilde{\boldsymbol{\xi}}_\perp\right\rangle=\langle\tilde{\mathbf{x}}_i,(\mathbf{I}-\mathbf{P}_k)(\mathbf{I}-\mathbf{P}_s)\xi\rangle=\langle(\mathbf{I}-\mathbf{P}_s)(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i,\xi\rangle$. 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. (4.83), $\|(\mathbf{I}-\mathbf{P}_s)(\mathbf{I}-\mathbf{P}_k)\tilde{\mathbf{x}}_i\|_2\ge\gamma_1$, where $\gamma_1=\sqrt{1-\mu_{\max}^2-\mu_s(1+\mu_{\max})^2}$. As in the proof of Lemma 4.4, it follows that

$$\Pr[A\le T]\ge 1-\frac{\gamma_1}{\sqrt{\pi\log d}}d^{-\frac{1}{\gamma_1^2}+1}\ge\frac{1}{2}, \tag{4.91}$$

where the last inequality holds for sufficiently large d and follows from noting that $\gamma_1\le 1$ by the max-MIP condition (4.5).

We now bound $\Pr[B\ge T]$, where B is the RHS of Eq. (4.90). By definition of $\tilde{\boldsymbol{\xi}}_\parallel$, the inner product $\left\langle\tilde{\mathbf{x}}_k,\tilde{\boldsymbol{\xi}}_\parallel\right\rangle=\left\langle\tilde{\mathbf{x}}_k,\tilde{\boldsymbol{\xi}}\right\rangle=\langle\tilde{\mathbf{x}}_k,(\mathbf{I}-\mathbf{P}_s)\xi\rangle$. 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. (4.81), $\|(\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_k\|_2\ge\sqrt{1-\mu_s}$. As in the proof of Lemma 4.4,

$$\Pr[B\ge T]\ge\Phi^c\left(\frac{1-\sqrt{\rho}}{\gamma_2}\sqrt{2\log d}\right), \tag{4.92}$$

where $\gamma_2=\sqrt{1-\mu_s}(1-\mu_{\max})$. Recall the definition of δ in Eq. (4.11). By Eq. (4.84), $\gamma_2\ge\sqrt{1-\delta}(1-\mu_{\max})$. Combining this with the bounds (4.91) and (4.92) completes the proof of Lemma 4.7. □

Proof of Lemma 4.8. Fix a non-support index $j\notin\mathcal{S}$. Recall that p_j, defined in Eq. (4.31), 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,

$$p_j\ge\Pr\left[|\langle\tilde{\mathbf{x}}_j,\mathbf{r}\rangle|\ge\max_{i\in\mathcal{A}}|\langle\tilde{\mathbf{x}}_i,\mathbf{r}\rangle|\right]. \tag{4.93}$$

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. (4.85). Next, we separately upper bound

$$\Pr\left[|\langle\tilde{\mathbf{x}}_j,\mathbf{r}\rangle|>\max_{i\notin\mathcal{S}\cup\{j\}}|\langle\tilde{\mathbf{x}}_i,\mathbf{r}\rangle|\right] \tag{4.94}$$

and

$$\Pr\left[|\langle\tilde{\mathbf{x}}_j,\mathbf{r}\rangle|>|\langle\tilde{\mathbf{x}}_k,\mathbf{r}\rangle|\right] \tag{4.95}$$

and then upper bound p_j using (4.64) 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

$$\begin{aligned} |\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| &= \left| \sum_{l\in\mathcal{S}\setminus s} \tilde{\theta}_l \langle \tilde{\mathbf{x}}_j, (\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_l\rangle + \sigma \left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}}\right\rangle \right| \\ &\leq \tilde{\theta}_{\max} \sum_{l\in\mathcal{S}\setminus s} |\langle \tilde{\mathbf{x}}_j, (\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_l\rangle| + \sigma \left|\left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}}\right\rangle\right| \leq \tilde{\theta}_{\max} K_u (\mu_{\max}+\mu_s) + \sigma \left|\left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}}\right\rangle\right| \end{aligned} \quad (4.96)$$

where the first equality follows from Eq. (4.77), the next inequality follows from the triangle inequality and the definition of $\tilde{\theta}_{\max}$ in Eq. (4.22), and the last inequality follow from (4.82). We now begin with event (4.94). By Eqs. (4.82) and (4.80), 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. (4.96), rearranging the terms, recalling the relation between $\tilde{\theta}_{\max}$ and ρ in (4.86) and applying inequality (4.80) yields

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| > \max_{i\notin\mathcal{S}\cup\{j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|\right] \leq \Pr\left[\left|\left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}}\right\rangle\right| + \frac{2K\mu_{\max}\sqrt{2\rho\log d}}{1-(2K-1)\mu_{\max}} > \max_{i\notin\mathcal{S}\cup\{j\}} \left\langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}\right\rangle\right]. \quad (4.97)$$

As in the proof of Lemma 4.5, applying (4.67) with $T = (1-\epsilon)\sqrt{2(1-\mu_{\max})\log d}$ and $\epsilon \in (0,1)$ as in (4.14), we can upper bound (4.97) by

$$\Pr\left[\left|\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}}\rangle\right| \geq a\sqrt{2\log d}\right] + \Pr\left[\max_{i\notin\mathcal{S}\cup\{j\}} \langle \tilde{\mathbf{x}}_i, \tilde{\boldsymbol{\xi}}\rangle < (1-\epsilon)\sqrt{2(1-\mu_{\max})\log d}\right], \quad (4.98)$$

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, $\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}}\rangle = \langle (\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_j, \xi\rangle$. By Eq. (4.81), the norm $\|(\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_j\|_2 \leq 1$ and thus the first term in (4.98) is bounded by

$$2\Phi^c\left(a\sqrt{2\log d}\right). \quad (4.99)$$

We now bound the second term in (4.98) using Lemma 4.12 with $Z_i = \left\langle \frac{(\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_i}{\|(\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_i\|_2}, \xi\right\rangle$. Towards this goal, notice that by Eq. (4.81), $\|(\mathbf{I}-\mathbf{P}_s)\tilde{\mathbf{x}}_i\|_2 \geq \sqrt{1-\mu_s}$. Thus, the second term in (4.98) is upper bounded by

$$\Pr\left[\max_{i\notin\mathcal{S}\cup\{j\}} Z_i < \frac{(1-\epsilon)\sqrt{2(1-\mu_{\max})\log d}}{\sqrt{1-\mu_s}}\right]. \quad (4.100)$$

Furthermore, by Eqs. (4.81) and (4.82), 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 4.12 with $\eta = \frac{\mu_{\max}+\mu_s}{1-\mu_s}$ and $\zeta = 1-\epsilon$ to obtain that (4.100) is bounded by

$$C(d-K-1)^{-\frac{1-\mu_{\max}}{\mu_{\max}+\mu_s}\epsilon^2} \log^{\left(\frac{1-\mu_{\max}}{\mu_{\max}+\mu_s}\epsilon-1\right)/2}(d-K-1). \quad (4.101)$$

Similarly to the proof of Lemma 4.5, under condition (4.14) on ϵ and for sufficiently large $d = d(\epsilon)$, (4.99) is larger than (4.101), and thus

$$\Pr\left[|\langle \tilde{\mathbf{x}}_j, \mathbf{r}\rangle| > \max_{i\notin\mathcal{S}\cup\{j\}} |\langle \tilde{\mathbf{x}}_i, \mathbf{r}\rangle|\right] \leq 4\Phi^c\left(a\sqrt{2\log d}\right). \quad (4.102)$$

We now turn to bounding (4.95). For the support index k, similarly to (4.88),

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

Combining the bound above with Eq. (4.96), recalling the relation between $\tilde{\theta}_{\max}$ and ρ in (4.86) and applying the inequality (4.80) 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 $H = \left\langle \tilde{\mathbf{x}}_j, \tilde{\boldsymbol{\xi}}\right\rangle = \langle (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_j, \xi\rangle$ and $G = \operatorname{sign}\left(\tilde{\theta}_k\right)\left\langle \tilde{\mathbf{x}}_k, \tilde{\boldsymbol{\xi}}\right\rangle = \operatorname{sign}\left(\tilde{\theta}_k\right)\langle (\mathbf{I}-\mathbf{P}_s)\,\tilde{\mathbf{x}}_k, \xi\rangle$. Notice that by Eqs. (4.81), (4.82), and (4.79), $\sigma_H^2, \sigma_G^2 \leq 1$. Combining Eqs. (4.11) and (4.84) implies that $\mu_s \leq \delta$. Hence, $|\sigma_{HG}| \leq \mu_{\max}+\mu_s \leq \mu_{\max}+\delta$. Thus, as in the proof of Lemma 4.5,

$$\Pr\left[|H| - G > c\right] \leq 2\Phi^c\left(\sqrt{\frac{2\rho\log d}{2+2\left(\mu_{\max}+\delta\right)}}\right). \tag{4.103}$$

By Eq. (4.64), the probability (4.93) is at most the minimum between (4.102) and (4.103). By the monotonicity of the Gaussian CDF, (4.93) is upper bounded by

$$4\Phi^c\left(\max\left\{\sqrt{\frac{\rho}{2+2\left(\mu_{\max}+\delta\right)}}, \left((1-\epsilon)\sqrt{1-\mu_{\max}} - \frac{2K\mu_{\max}\sqrt{\rho}}{1-(2K-1)\mu_{\max}}\right)\right\}\sqrt{2\log d}\right). \tag{4.104}$$

Similarly to the proof of Lemma 4.5, inserting the definitions of Q_1 and Q_2 in Eqs. (4.16) and (4.17) respectively, into Eq. (4.18) and rearranging various terms yields

$$\frac{1-\sqrt{r}}{\sqrt{1-\delta}\left(1-\mu_{\max}\right)} + \sqrt{Q_0} < \max\left\{\sqrt{\frac{r}{2+2\left(\mu_{\max}+\delta\right)}}, \left((1-\epsilon)\sqrt{1-\mu_{\max}} - \frac{2K\mu_{\max}\sqrt{r}}{1-(2K-1)\mu_{\max}}\right)\right\}. \tag{4.105}$$

The definitions of r and ρ in Eqs. (4.9) and (4.23) imply that $\rho \geq r$. Thus, ρ satisfies Eq. (4.18) and hence Eq. (4.105). The RHS of Eq. (4.105) is the same as the maximum in Eq. (4.104) above. Thus, Eq. (4.104) is upper bounded by

$$4\Phi^c\left(\left(\frac{1-\sqrt{\rho}}{\sqrt{1-\delta}\left(1-\mu_{\max}\right)} + \sqrt{Q_0}\right)\sqrt{2\log d}\right). \tag{4.106}$$

By the assumption $\rho \leq 1$, we can apply Lemma 4.11. Hence, by the definition of Q_0 in Eq. (4.15), and by the definition of F in Eq. (4.27),

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

which completes the proof of Lemma 4.8. □

4.6.6 Proofs of technical lemmas

Proof of Lemma 4.11. Classical results by Birnbaum (1942) and Komatu (1955) are that for all $x \geq 0$, the following inequalities hold

$$\frac{2e^{-x^2/2}}{\sqrt{2\pi}\left(\sqrt{x^2+4}+x\right)} < \Phi^c(x) < \frac{2e^{-x^2/2}}{\sqrt{2\pi}\left(\sqrt{x^2+2}+x\right)}. \tag{4.107}$$

Hence,

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

and

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

Combining the two yields the following

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

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

Towards proving Lemma 4.13, we prove the following Lemma 4.14, 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 4.14. *Let $s \subset \mathcal{S}$ and denote $K_d = |s|$. Assume that $(K_d - 1)\mu_{\max} < 1$. Then, for any pair of vectors $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^n$*

$$\langle \mathbf{a}_1, \mathbf{a}_2 \rangle - \frac{\left|\sum_{j\in s}\langle \tilde{\mathbf{x}}_j, \mathbf{a}_1\rangle\langle \tilde{\mathbf{x}}_j, \mathbf{a}_2\rangle\right|}{1-(K_d-1)\mu_{\max}} \leq \langle \mathbf{a}_1, (\mathbf{I}-\mathbf{P}_s)\mathbf{a}_2\rangle \leq \langle \mathbf{a}_1, \mathbf{a}_2\rangle + \frac{\left|\sum_{j\in s}\langle \tilde{\mathbf{x}}_j, \mathbf{a}_1\rangle\langle \tilde{\mathbf{x}}_j, \mathbf{a}_2\rangle\right|}{1-(K_d-1)\mu_{\max}}. \quad (4.108)$$

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

$$\|\mathbf{a}\|_2^2 - \frac{\sum_{j\in s}\langle \tilde{\mathbf{x}}_j, \mathbf{a}\rangle^2}{1-(K_d-1)\mu_{\max}} \leq \|(\mathbf{I}-\mathbf{P}_s)\mathbf{a}\|_2^2 \leq \|\mathbf{a}\|_2^2 - \frac{\sum_{j\in s}\langle \tilde{\mathbf{x}}_j, \mathbf{a}\rangle^2}{1+(K_d-1)\mu_{\max}}. \quad (4.109)$$

Proof of Lemma 4.14. 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 (4.108) and (4.109) trivially hold. Therefore, assume that $K_d \geq 1$. In this case

$$\langle \mathbf{a}_1, (\mathbf{I}-\mathbf{P}_s)\mathbf{a}_2\rangle = \langle \mathbf{a}_1, \mathbf{a}_2\rangle - \langle \mathbf{a}_1, \mathbf{P}_s\mathbf{a}_2\rangle. \quad (4.110)$$

By definition of $\mathbf{P}_s$ in Eq. (4.48),

$$\langle \mathbf{a}_1, \mathbf{P}_s\mathbf{a}_2\rangle = \left\langle \mathbf{a}_1, \tilde{\mathbf{X}}_{|s}\left(\tilde{\mathbf{X}}_{|s}^T\tilde{\mathbf{X}}_{|s}\right)^{-1}\tilde{\mathbf{X}}_{|s}^T\mathbf{a}_2\right\rangle = \left\langle \tilde{\mathbf{X}}_{|s}^T\mathbf{a}_1, \left(\tilde{\mathbf{X}}_{|s}^T\tilde{\mathbf{X}}_{|s}\right)^{-1}\tilde{\mathbf{X}}_{|s}^T\mathbf{a}_2\right\rangle. \quad (4.111)$$

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

$$\frac{1}{1+(K_d-1)\mu_{\max}} \leq \lambda_{\min}\left(\mathbf{A}^{-1}\right) \leq \lambda_{\max}\left(\mathbf{A}^{-1}\right) \leq \frac{1}{1-(K_d-1)\mu_{\max}}. \quad (4.112)$$

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. (4.111) yields

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

Combining the bounds in Eq. (4.112) 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

$$\frac{\left|\sum_{j \in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a}_1 \rangle \langle \tilde{\mathbf{x}}_j, \mathbf{a}_2 \rangle\right|}{1 + (K_d - 1)\,\mu_{\max}} \leq \left|\langle \mathbf{a}_1, \mathbf{P}_s \mathbf{a}_2 \rangle\right| \leq \frac{\left|\sum_{j \in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a}_1 \rangle \langle \tilde{\mathbf{x}}_j, \mathbf{a}_2 \rangle\right|}{1 - (K_d - 1)\,\mu_{\max}}. \tag{4.113}$$

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. (4.110) proves the inequality (4.108).

We now prove the inequality (4.109). Let $\mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a}$. Since each of the terms in the two sums in Eq. (4.113) is positive, we may remove the absolute values, i.e.,

$$\frac{\sum_{j \in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a} \rangle^2}{1 + (K_d - 1)\,\mu_{\max}} \leq \langle \mathbf{a}, \mathbf{P}_s \mathbf{a} \rangle \leq \frac{\sum_{j \in s} \langle \tilde{\mathbf{x}}_j, \mathbf{a} \rangle^2}{1 - (K_d - 1)\,\mu_{\max}}. \tag{4.114}$$

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

$$(\mathbf{I} - \mathbf{P}_s)^T (\mathbf{I} - \mathbf{P}_s) = (\mathbf{I} - \mathbf{P}_s)(\mathbf{I} - \mathbf{P}_s) = (\mathbf{I} - \mathbf{P}_s). \tag{4.115}$$

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 (4.114) completes the proof of (4.109) and of Lemma 4.14. □

Proof of Lemma 4.13. We begin with proving inequalities (4.79) and (4.80). By the max-MIP condition (4.5), $1 - (2K - 1)\,\mu_{\max} > 0$. Rearranging implies that

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

Combining this with the bound on μ_s in (4.84) gives

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

which proves (4.79). The max-MIP condition (4.5) 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 μ_s in (4.78) 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 (4.80).

We now prove the remaining inequalities using Lemma 4.14, beginning with (4.81). Since $\mathbf{P}_s$ is a projection matrix, for any index $i \notin s$,

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

Recall that for any distinct pair of indices $i \neq j$, it holds that $0 \leq \langle \tilde{\mathbf{x}}_j, \tilde{\mathbf{x}}_i \rangle^2 \leq \mu_{\max}^2$. By Eq. (4.109) with $\mathbf{a} = \tilde{\mathbf{x}}_i$,

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

which concludes the proof of (4.81).

Next, we prove inequality (4.82). By the right inequality in (4.108) with $\mathbf{a}_1 = \tilde{\mathbf{x}}_k$ and $\mathbf{a}_2 = \tilde{\mathbf{x}}_i$,

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

Thus, by the triangle inequality and by the definitions of $\mu_{\max}$ and μ_s in Eqs. (4.4) and (4.78) respectively,

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

Finally, we prove inequality (4.83). Recall that by the max-MIP condition (4.5), $\mu_{\max} < \frac{1}{K-1}$. For any distinct pair of indices $i \neq k$ such that $i, k \notin s$, Eq. (4.109) with $\mathbf{a} = (\mathbf{I} - \mathbf{P}_k)\, \tilde{\mathbf{x}}_i$ gives

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

which concludes the proof of Eq. (4.83). 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 (4.5). Now let $K > 1$. By the max-MIP condition (4.5), $\mu_{\max} < 1$ 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 (4.5). □

4.7 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 4.1, we prove that this variant,

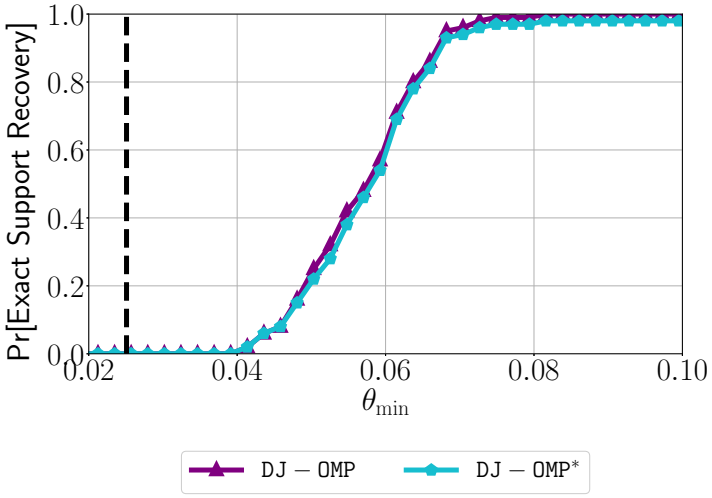


Figure 4.3: Support Recovery by DJ-OMP and DJ-OMP* as a Function of $\theta_{\min}$

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 4.1 completes this section.

Corollary 4.1. *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 4.3 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 4.1(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 4.1. Denote by B an event where the max-MIP condition (4.5) 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 ℓ_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 (4.5) 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

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

We now show that when event B does not occur, i.e., when the max-MIP condition (4.5) is satisfied, then DJ-OMP* recovers the support of $\boldsymbol{\theta}$ with probability at least $1-\frac{2^{K+1}-1}{d}$. By Lemma 4.1, 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. (4.77), the residual $\mathbf{r}^{(m)}$ is

$$\begin{aligned}\mathbf{r}^{(m)} &= \left(\mathbf{I}-\mathbf{P}_{\mathcal{S}}^{(m)}\right)\mathbf{y}^{(m)} = \left(\mathbf{I}-\mathbf{P}_{\mathcal{S}}^{(m)}\right)\left(\mathbf{X}^{(m)}\boldsymbol{\theta}^{(m)}+\sigma\xi^{(m)}\right)\\ &= \sigma\left(\mathbf{I}-\mathbf{P}_{\mathcal{S}}^{(m)}\right)\xi^{(m)}. \end{aligned} \tag{4.116}$$

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

$$p_j^{(m)} = \Pr\left[j = \arg\max_{i\notin\mathcal{S}}\left|\left\langle \mathbf{x}_i^{(m)}, \mathbf{r}^{(m)}\right\rangle\right|\right] = \Pr\left[j = \arg\max_{i\notin\mathcal{S}}\left|\left\langle \mathbf{x}_i^{(m)}, \left(\mathbf{I}-\mathbf{P}_{\mathcal{S}}^{(m)}\right)\xi^{(m)}\right\rangle\right|\right]. \tag{4.117}$$

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

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

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. □

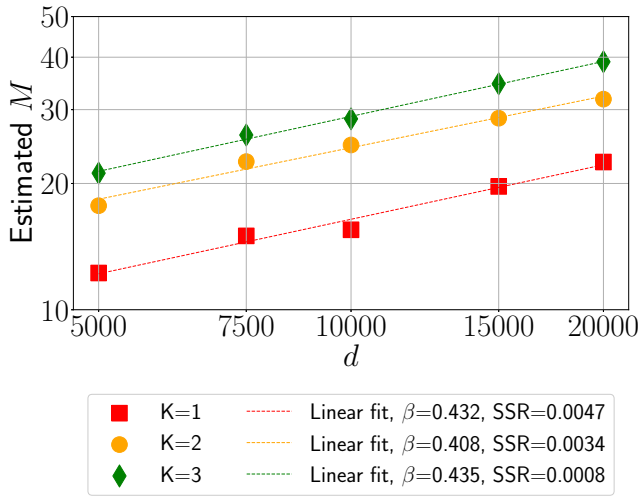


Figure 4.4: Number of Machines for Support Recovery by `DJ-OMP` vs. Dimension d

4.8 Additional Simulation Results

Theorem 4.1 holds under the max-MIP condition (4.5) 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 4.1(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 4.4. As described in Section 4.4, 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 β is empirically smaller than the exponent derived in Eq. (4.12). In addition, the estimated number of machines increases with K, which is also in accordance with Eq. (4.12). We obtained similar results when the matrices were slightly correlated, with slightly higher estimated number of machines.

4.9 Implementation Details

The code used to generate the simulations in Section 4.4 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).

Bibliography

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.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. Distributed signal detection under communication constraints. In *Conference on Learning Theory*, pages 41–63. PMLR, 2020a.

Jayadev Acharya, Clément L Canonne, and Himanshu Tyagi. General lower bounds for interactive high-dimensional estimation under information constraints. *arXiv preprint arXiv:2010.06562*, 2020b.

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.

Chen Amiraz, Robert Krauthgamer, and Boaz Nadler. Recovery guarantees for distributed-OMP. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024. To appear.

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.

Juan Andrés Bazerque and Georgios B. Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, 58(3):1847–1862, 2010.

Ron Bekkerman, Mikhail Bilenko, and John Langford. *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

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.

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.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020, 2016.

T Tony Cai and Hongji Wei. Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *CoRR abs/2001.08877*, 2020.

Tony Cai and Tiefeng Jiang. Phase transition in limiting distributions of coherence of high-dimensional random matrices. *Journal of Multivariate Analysis*, 107:24–39, 2012.

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

Tony Cai, Lie Wang, and Guangwu Xu. Stable recovery of sparse signals and an oracle inequality. *IEEE Transactions on Information Theory*, 56(7):3516–3522, 2010.

Xiaolun Cai, Zhengchun Zhou, Yang Yang, and Yong Wang. Improved sufficient conditions for support recovery of sparse signals via orthogonal matching pursuit. *IEEE Access*, 6:30437–30443, 2018.

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

Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3869–3872. IEEE, 2008.

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.

Yuval Dagan and Ohad Shamir. Detecting correlations with little memory and communication. In *Conference On Learning Theory*, pages 1145–1198, 2018.

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

Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010.

Geoff Davis, Stephane Mallat, and Marco Avellaneda. Adaptive greedy approximations. *Constructive approximation*, 13(1):57–98, 1997.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81 (3):425–455, 1994.

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.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2012.

Michael Elad. *Sparse and redundant representations: from theory to applications in signal and image processing*. Springer Science & Business Media, 2010.

Yonina C Eldar and Gitta Kutyniok, editors. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

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.

Mário AT Figueiredo, José M Bioucas-Dias, and Robert D Nowak. Majorization–minimization algorithms for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 16(12): 2980–2991, 2007.

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.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.

Ankit Garg, Tengyu Ma, and Huy Nguyen. On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2014.

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.

Saptarshi Guha, Ryan Hafen, Jeremiah Rounds, Jin Xia, Jianfu Li, Bowei Xi, and William S Cleveland. Large complex data: divide and recombine (d&r) with rhipe. *Stat*, 1(1):53–67, 2012.

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.

Yanjun Han, Ayfer Özgür, and Tsachy Weissman. Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory*, pages 3163–3188, 2018.

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.

Abolfazl Hashemi and Haris Vikalo. Sparse recovery via orthogonal least-squares under presence of noise. *arXiv preprint arXiv:1608.02554*, 2016.

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

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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

Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on signal processing*, 56(6):2346–2356, 2008.

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 sparse signal recovery for compressive sleeping wireless sensor networks. *IEEE Transactions on Signal Processing*, 58(7):3816–3827, 2010.

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 ℓ_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. On the maximum of dependent gaussian random variables: A sharp bound for the lower tail. *arXiv preprint arXiv:1809.08539*, 2018.

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 Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

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

Elaine Crespo Marques, Nilson Maciel, Lírida Naviner, Hao Cai, and Jun Yang. A review of sparse recovery algorithms. *IEEE Access*, 7:1300–1322, 2019.

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

Ryan McDonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S Mann. Efficient large-scale distributed training of conditional maximum entropy models. In *Advances in neural information processing systems*, pages 1231–1239, 2009.

Ehsan Miandji, Mohammad Emadi, Jonas Unger, and Ehsan Afshari. On probability of support recovery for orthogonal matching pursuit using mutual coherence. *IEEE Signal Processing Letters*, 24(11):1646–1650, 2017.

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

Stanislav Minsker et al. Distributed statistical estimation and rates of convergence in normal approximation. *Electronic Journal of Statistics*, 13(2):5213–5252, 2019.

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.

Sergey V Nagaev. Lower bounds on large deviation probabilities for sums of independent random variables. *Theory of Probability & Its Applications*, 46(1):79–102, 2002.

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.

Deanna Needell and Roman Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):310–316, 2010.

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/`.

Philippe Rigollet. 18.s997: High dimensional statistics. *Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu*, 2015.

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.

Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *International conference on machine learning*, pages 1000–1008. PMLR, 2014.

Shree Krishna Sharma, Eva Lagunas, Symeon Chatzinotas, and Björn Ottersten. Application of compressive sensing in cognitive radio communications: A survey. *IEEE Communications Surveys & Tutorials*, 18(3):1838–1860, 2016.

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.

Jacob Steinhardt, Gregory Valiant, and Stefan Wager. Memory, communication, and statistical queries. In *Conference on Learning Theory*, pages 1490–1516. PMLR, 2016.

Thomas Strohmer and Robert W Heath. Grassmannian frames with applications to coding and communication. *Applied and computational harmonic analysis*, 14(3):257–275, 2003.

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.

Botond Szabo and Harry van Zanten. Adaptive distributed methods under communication constraints. *Annals of Statistics*, 48(4):2347–2380, 2020.

Botond Szabo, Lasse Vuursteen, and Harry van Zanten. Optimal distributed testing in high-dimensional gaussian models. *arXiv preprint arXiv:2012.04957*, 2020.

Kevin Tanguy. Some superconcentration inequalities for extrema of stationary gaussian processes. *Statistics & Probability Letters*, 106:239–246, 2015.

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

Robert Tibshirani, Martin Wainwright, and Trevor Hastie. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

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.

Joel A Tropp, Inderjit S Dhillon, Robert W Heath, and Thomas Strohmer. Designing structured tight frames via an alternating projection method. *IEEE Transactions on Information Theory*, 51(1): 188–209, 2005.

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.

Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4): 30–44, 1991.

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*, 2019.

David P Wipf and Bhaskar D Rao. Sparse bayesian learning for basis selection. *IEEE Transactions on Signal processing*, 52(8):2153–2164, 2004.

Yuchen Zhang and Xiao Lin. Disco: Distributed optimization for self-concordant empirical loss. In *International conference on machine learning*, pages 362–370. PMLR, 2015.

Yuchen Zhang, John Duchi, Michael I Jordan, and Martin J Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pages 2328–2336, 2013a.

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013b.

Yuancheng Zhu and John Lafferty. Distributed nonparametric regression under communication constraints. In *International Conference on Machine Learning*, pages 6009–6017, 2018.