5 Appendix A: Notations and definitions

The full list of notations that are used in this paper is in Table 5.

Table 5: Notation of Main Variables

Notation	Description
m	total number of measurements
n	signal dimension
s	sparsity level
L	size of each bootstrap sample
L/m	bootstrap sampling ratio
K	number of bootstrap samples / the number of estimates
$\boldsymbol{A}$	the original sensing matrix of size $m \times n$
$\boldsymbol{y}$	the original measurements vector of size $m \times 1$
$\mathcal{I}$	a multi-set or set
$\mathcal{I}_j$	the j-th Bootstrap sample, $j = 1, 2, ..., K$, length of $\mathcal{I}_j = L$
$(\cdot)[\mathcal{I}]$	takes rows supported on $\mathcal{I}$ and throws away elements in $\mathcal{I}^c$
$\boldsymbol{A}[\mathcal{I}_j]$	bootstrapped sampling matrix for bootstrap sample $\mathcal{I}_j$
$\boldsymbol{y}[\mathcal{I}_j]$	measurement vector corresponds to bootstrap sample $\mathcal{I}_j$
$\boldsymbol{x}_j$	the j-th column of matrix $\boldsymbol{X}$;
	a feasible solution corresponds to $(\boldsymbol{A}[\mathcal{I}_j], \boldsymbol{y}[\mathcal{I}_j])$
$\widehat{\boldsymbol{x}}_j$	the optimal solution corresponds to $(\boldsymbol{A}[\mathcal{I}_j], \boldsymbol{y}[\mathcal{I}_j])$
$(\cdot)[i]$	the i-th row of a matrix/ vector.
$\boldsymbol{x}[i]$	the i-th row of matrix $\boldsymbol{X}$
$\lVert\boldsymbol{X}\rVert_{p,q}$	takes ℓ_q norms on rows of $\boldsymbol{X}$; stacks those as a vector
	and then computes ℓ_p norm. The precise form is in (8).
$\lVert\boldsymbol{X}\rVert_{1,2}$	row sparsity norm
$\lVert\boldsymbol{X}\rVert_{1,1}$	the ℓ_1 norm on vectorized $\boldsymbol{X}$

5.1 Mixed $\ell_{p,q}$ norm of a matrix

The mixed $\ell_{p,q}$ norm on matrix $\boldsymbol{X}$ is defined as:

$$\begin{aligned}\|\boldsymbol{X}\|_{p,q} &= (\sum_{i=1}^{n} \|\boldsymbol{x}[i]^{\boldsymbol{T}}\|_q^p)^{1/p} \\ &= \|(\|\boldsymbol{x}[1]^{\boldsymbol{T}}\|_q, \|\boldsymbol{x}[2]^{\boldsymbol{T}}\|_q, ..., \|\boldsymbol{x}[n]^{\boldsymbol{T}}\|_q)^{\boldsymbol{T}}\|_p,\end{aligned} \tag{8}$$

where $\boldsymbol{x}[i]$ denotes the $i-$th row of matrix $\boldsymbol{X}$. Intuitively, the mixed $\ell_{p,q}$ norm essentially takes ℓ_q norms on rows of $\boldsymbol{X}$ first; then stacks those as a vector and then computes its ℓ_p norm. Note when $p = q$, the $\ell_{p,p}$ norm of $\|\boldsymbol{X}\|$ is simply the ℓ_p vector norm of the vectorized $\boldsymbol{X}$. The row sparsity penalty that we employed $\ell_{1,2}$ norm in JOBS is essentially a special case of (8) taking $p = 1, q = 2$.

5.2 Mixed $\ell_{p,q}$ norm over block partition of a vector

Similarly to the $\ell_{p,q}$ norm on matrix in (8), we introduce a more general form: the mixed $\ell_{p,q}$ norm over a block partition of a vector. The definition for $\ell_{p,q}$ norm over block partition $\mathcal{B} =$

$\{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_b\}$ for a vector $\|\boldsymbol{x}\|_{p,q|\mathcal{B}}$:

$$\begin{aligned}\|\boldsymbol{x}\|_{p,q|\mathcal{B}} =& (\sum_{i=1}^{b} \|\boldsymbol{x}[\mathcal{B}_i]^{\boldsymbol{T}}\|_q^p)^{1/p} \\ &= \|(\|\boldsymbol{x}[\mathcal{B}_1]^{\boldsymbol{T}}\|_q, ..., \|\boldsymbol{x}[\mathcal{B}_b]^{\boldsymbol{T}}\|_q)\|_p.\end{aligned} \tag{9}$$

It is not difficult to see that the $\ell_{p,q}$ norm of a matrix is a special case of $\ell_{p,q}$ norm over block of the vectorized version of that matrix. In fact, the mixed $\ell_{1,2}$ norm on matrix $\boldsymbol{X}$ can also be expressed as a mixed $\ell_{1,2|\mathcal{B}}$ norm on the vectorized $\boldsymbol{X}$ given $\mathcal{B}$, where the block partition is row-wise.

6 Appendix B: Preliminaries

We summarize the theoretical results that are needed for understanding and analyzing our algorithm mathematically. We offer a quick review of several concepts including block sparsity, Null Space Property (NSP) (Cohen et al., 2009), Restricted Isometry Property (RIP) (Candes, 2008) for classical sparse signal recovery as well as Block Null Space Property (BNSP) (Gao et al., 2015), Block Restricted Isometry Property (BRIP) (Eldar & Mishali, 2009) for block sparse signal recovery.

6.1 Block Sparsity

Since row sparsity is a special case of block sparsity (or more precisely, the non-overlapping group sparsity) (Eldar & Mishali, 2009), we therefore can employ the tools from block sparsity to analyze our problem. Block sparsity is a generalization of the standard ℓ_1 sparsity. To start, we recall its definition.

Definition 8 (Block Sparsity, from (Eldar & Mishali, 2009)) *$\boldsymbol{x} \in \mathbb{R}^n$ is $s-$block sparse with respect to a partition $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_b\}$ of $\{1, 2, ..., n\}$ if for $\boldsymbol{x} = (\boldsymbol{x}[\mathcal{B}_1], \boldsymbol{x}[\mathcal{B}_2], ..., \boldsymbol{x}[\mathcal{B}_b])$, the block sparsity level is $\|\boldsymbol{x}\|_{0,2|\mathcal{B}} := \sum_{i=1}^{b} 1\{\|\boldsymbol{x}[\mathcal{B}_i]\|_2 > 0\} \leq s$ and the relaxation $\ell_{1,2}$ norm is $\|\boldsymbol{x}\|_{1,2|\mathcal{B}} := \sum_{i=1}^{b} \|\boldsymbol{x}[\mathcal{B}_i]\|_2$.*

The block sparsity level $\|\boldsymbol{x}\|_{0,2|\mathcal{B}}$ counts the number of non-zero blocks of the given a block partition $\mathcal{B}$. The $\ell_{1,2}$ norm $\|\boldsymbol{x}\|_{1,2|\mathcal{B}} := \sum_{i=1}^{b} \|\boldsymbol{x}[\mathcal{B}_i]\|_2$ is one of its convex relaxations. For the same sparse vector $\boldsymbol{x}$, the block sparsity level is in general smaller than the sparsity level given a non-overlapping block partition.

The $\ell_{1,2}$ minimization is a special case of block sparse minimization, with each element in the block partition containing all indices of a row. The results of block sparsity such as BNSP, BRIP can be useful tools to analyze our algorithm.

6.2 Null Space Property (NSP) and Block-NSP (BNSP)

The NSP for standard sparse recovery and block sparse signal recovery are summarized below. BNSP is obtained from a more general result of BNSP of $\ell_{p,2}$ block norm stated in (9) from (Gao et al., 2015) taking $p = 1$.

Theorem 9 (NSP, from (Cohen et al., 2009)) *Every $s-$sparse signal $\boldsymbol{x} \in \mathbb{R}^n$ is a unique solution to $\mathbf{P_1}$: $\min \|\boldsymbol{x}\|_1$ s.t. $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ if and only if $\boldsymbol{A}$ satisfies NSP of order s: for any set $\mathcal{S} \subset \{1, 2, .., n\}$, card$(\mathcal{S}) \leq s$,*

$$\|\boldsymbol{v}[\mathcal{S}]\|_1 < \|\boldsymbol{v}[\mathcal{S}^c]\|_1,$$

for all $\boldsymbol{v} \in \mathrm{Null}(\boldsymbol{A})\backslash\{\mathbf{0}\}$, where $\boldsymbol{v}[\mathcal{S}]$ denotes the vector equals to $\boldsymbol{v}$ on a index set $\mathcal{S}$ and zero elsewhere.

Definition 10 (BNSP, from (Gao et al., 2015)) *Every $s-$block sparse signal $\boldsymbol{x}$ with respect to block assignment $\mathcal{B}$, is a unique solution to $\min \|\boldsymbol{x}\|_{1,2|\mathcal{B}}$ s.t. $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ if and only if matrix $\boldsymbol{A}$ satisfies block null space property over $\mathcal{B}$ of order s: for any set $\mathcal{S} \subset \{1, 2, ..., n\}$ with card$(\mathcal{S}) \leq s$,*

$$\|\boldsymbol{v}[\mathcal{S}]\|_{1,2|\mathcal{B}} < \|\boldsymbol{v}[\mathcal{S}^c]\|_{1,2|\mathcal{B}},$$

for all $\boldsymbol{v} \in \mathrm{Null}(\boldsymbol{A})\backslash\{\boldsymbol{0}\}$, where $\boldsymbol{v}[\mathcal{S}]$ denotes the vector equal to $\boldsymbol{v}$ on a block index set $\mathcal{S}$ and zero elsewhere.

6.3 Restricted Isometry Property (RIP) and Block-RIP (BRIP)

Although NSP directly characterizes the ability of success for sparse recovery, verifying the BNSP condition is computationally intractable and it is also not suitable for quantifying performance in noisy cases since it is a binary (True or False) metric instead of a continuous one. Restricted Isometry Properties: RIP (Candes, 2008) and BRIP (Eldar & Mishali, 2009) are introduced for those purposes.

Definition 11 (RIP, from (Candes, 2008)) *A matrix $\boldsymbol{A}$ with ℓ_2-normalized columns satisfies RIP of order s if there exists a constant $\delta_s(\boldsymbol{A}) \in [0,1)$ such that for every $s-$sparse $\boldsymbol{v} \in \mathbb{R}^n$,*

$$(1-\delta_s(\boldsymbol{A}))\|\boldsymbol{v}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{v}\|_2^2 \leq (1+\delta_s(\boldsymbol{A}))\|\boldsymbol{v}\|_2^2. \tag{10}$$

Definition 12 (BRIP, from (Eldar & Mishali, 2009)) *A matrix $\boldsymbol{A}$ with ℓ_2-normalized columns satisfies Block RIP with respect to block partition $\mathcal{B}$ of order s if there exists a constant $\delta_{s|\mathcal{B}}(\boldsymbol{A}) \in [0,1)$ such that for every $s-$block sparse $\boldsymbol{v} \in \mathbb{R}^n$ over $\mathcal{B}$,*

$$(1-\delta_{s|\mathcal{B}}(\boldsymbol{A}))\|\boldsymbol{v}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{v}\|_2^2 \leq (1+\delta_{s|\mathcal{B}}(\boldsymbol{A}))\|\boldsymbol{v}\|_2^2. \tag{11}$$

If we take the location of each entry as one block, the block sparsity RIP reduces to the standard RIP condition. Therefore, BRIP is a generalization of RIP.

6.4 Noisy Recovery bounds based on RIP constants

It is well-known that certain RIP conditions imply NSP conditions for both classical sparse recovery and block sparse recovery. More specifically, if the RIP constant in the order $2s$ is strictly less than $\sqrt{2}-1$, then it implies that NSP is satisfied in the order of s. This applies to sparse recovery (Candes, 2008) and block sparse recovery (Eldar & Mishali, 2009).

Stated below are the error bound for conventional sparse recovery based on ℓ_1 minimization and the RIP constant as well as for block sparse recovery based on BRIP constant.

Theorem 13 (Sparse recovery error bound, from (Candes, 2008)) *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{z}$, $\|\boldsymbol{z}\|_2 \leq \epsilon$; $\boldsymbol{x}_0$ is $s-$sparse and minimizes $\|\boldsymbol{x}-\boldsymbol{x}^\star\|_2$ over all $s-$ sparse signals, and the vector $\boldsymbol{e}$ represents the s-sparse approximation error vector $\boldsymbol{e} = \boldsymbol{x}^\star - \boldsymbol{x}_0$. If $\delta_{2s}(\boldsymbol{A}) \leq \delta < \sqrt{2}-1$ and $\boldsymbol{x}^{\boldsymbol{\ell_1}}$ is the solution of ℓ_1 minimization, then*

$$\|\boldsymbol{x}^{\boldsymbol{\ell_1}} - \boldsymbol{x}^\star\|_2 \leq \mathcal{C}_0(\delta)s^{-1/2}\|\boldsymbol{e}\|_1 + \mathcal{C}_1(\delta)\epsilon,$$

where $\mathcal{C}_0(\cdot), \mathcal{C}_1(\cdot)$ are certain constants, depending on the RIP constant $\delta_{2s}(\boldsymbol{A})$. These two constants are in the form of non-decreasing functions of δ: $\mathcal{C}_0(\delta) = \frac{2(1-(1-\sqrt{2})\delta)}{1-(1+\sqrt{2})\delta}$ and $\mathcal{C}_1(\delta) = \frac{4\sqrt{1+\delta}}{1-(1+\sqrt{2})\delta}$.

Theorem 14 (Block sparse recovery error bound, from (Eldar & Mishali, 2009)) *Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{z}$, $\|\boldsymbol{z}\|_2 \leq \epsilon$; $\boldsymbol{x}_{0|\mathcal{B}}$ is $s-$block sparse and minimizes $\|\boldsymbol{x}-\boldsymbol{x}^\star\|_2$ over all $s-$block sparse signals, and the vector $\boldsymbol{e}_\mathcal{B}$ represents the s-sparse approximation error vector $\boldsymbol{e}_\mathcal{B} = \boldsymbol{x}^\star - \boldsymbol{x}_{0|\mathcal{B}}$. If $\delta_{2s|\mathcal{B}}(\boldsymbol{A}) \leq \delta < \sqrt{2}-1$, $\boldsymbol{x}^{\boldsymbol{\ell_{1,2|\mathcal{B}}}}$ is the solution of block sparse minimization, then*

$$\|\boldsymbol{x}^{\boldsymbol{\ell_{1,2|\mathcal{B}}}} - \boldsymbol{x}^\star\|_2 \leq \mathcal{C}_0(\delta)s^{-1/2}\|\boldsymbol{e}_\mathcal{B}\|_{1,2|\mathcal{B}} + \mathcal{C}_1(\delta)\epsilon,$$

where $\mathcal{C}_0(\cdot), \mathcal{C}_1(\cdot)$ are the same non-decreasing functions of δ as in Theorem 13.

6.5 Sample Complexity for i.i.d. Gaussian or Bernoulli Random Matrices

With $\boldsymbol{A}$ being a random matrix in which entries are identically and independently distributed (i.i.d.), previous work in (Baraniuk et al., 2008) builds a relationship between the sample complexity for random matrices to a desired RIP constant as a direct implication from Johnson-Lindenstrauss lemma as stated below.

Theorem 15 (Sample Complexity, from (Baraniuk et al., 2008)) *Let entries of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ from Gaussian $\mathcal{N}(0, 1/m)$ or Bernoulli $1/\sqrt{m}$ Bern(0.5). Let $\xi, \delta \in (0,1)$ and assume $m \geq \beta\delta^{-2}(s\ln(n/s) + \ln(\xi^{-1}))$ for a universal constant $\beta > 0$, then $\mathbb{P}(\delta_s(\boldsymbol{A}) \leq \delta) \geq 1 - \xi$.*

By rearranging the terms in this theorem, the sample complexity can be derived: when m is sufficiently large, which is in the order of $\mathcal{O}(2s\ln(n/2s))$, there is a high probability that the RIP constant of order $2s$ is sufficiently small.

7 Appendix C: JOBS-Noiseless, a two step relaxation of ℓ_1 minimization

JOBS recovers the true sparse solution because it is essentially a relaxation of the original ℓ_1 minimization problem in a multiple vectors fashion. Therefore, it is not so surprising that JOBS relaxation can recover the true solution: exactly in the noiseless case and within some neighbourhood of the ground truth in noisy case.

We demonstrate that JOBS is a two-step relaxation procedure of ℓ_1 minimization. For a ℓ_1 minimization with a unique solution $\boldsymbol{x}^\star$, the multiple measurement vectors (MMV) equivalence is: for $j = 1, 2, .., K$

$$\mathbf{P_1}(K) : \min \|\boldsymbol{X}\|_{1,1} \text{ s.t. } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_j, \tag{12}$$

where $\|\boldsymbol{X}\|_{1,1} = \sum_i \|\boldsymbol{x}[i]^{\boldsymbol{T}}\|_1$ as mentioned in Table 5. We show that this MMV form (12) is equivalent to the original ℓ_1 minimization problem. If the original problem $\mathbf{P_1}$ has a unique solution $\boldsymbol{x}^\star$, then the solution to the MMV problem $\mathbf{P_1}(K)$ in (12) yields a row sparse solution $\boldsymbol{X}^\star = (\boldsymbol{x}^\star, \boldsymbol{x}^\star, ..., \boldsymbol{x}^\star)$. This result can be derived via contradiction. The reverse direction is also true: if the MMV problem $\mathbf{P_1}(K)$ has a unique solution, it implies that the $\mathbf{P_1}$ must also have a unique solution. Details are stated in Lemma 18 in Appendix 10.1.

Since the $\ell_{1,1}$ norm of $\boldsymbol{X}$ essentially takes ℓ_1 norm of its vectorized version, it only enforces the sparsity for all elements in $\boldsymbol{X}$ without any structure such as the support consistency across its columns. To obtain the JOBS form, We first relax the $\ell_{1,1}$ norm in (12) to the $\ell_{1,2}$ norm. For all $j = 1, 2, .., K$

$$\mathbf{P_{12}}(K) : \ \min \|\boldsymbol{X}\|_{1,2} \text{ s.t. } \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_j. \tag{13}$$

From here, to obtain Noiseless JOBS version, we further drop all constraints that are not in $\mathcal{I}_j$ from (13) for estimator $\boldsymbol{x}_j, j = 1, 2, ..., K$. Then we obtain the noiseless version of JOBS:

$$\mathbf{J_{12}} : \min \|\boldsymbol{X}\|_{1,2} \text{ s.t. } \boldsymbol{y}[\mathcal{I}_j] = \boldsymbol{A}[\mathcal{I}_j]\boldsymbol{x}_j, \tag{14}$$

This two-step relaxation process is illustrated in Figure 5.

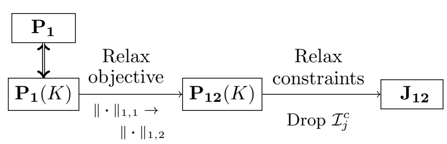


Figure 5: JOBS framework is a two-step relaxation of ℓ_1 minimization

The noisy version can be obtained similarly. We formulate the MMV version of the original ℓ_1 problem; relax the regularizer from $\ell_{1,1}$ norm to $\ell_{1,2}$ norm, and then further relax the objective function by dropping the constraints that are not on the selected subset $\mathcal{I}_j$ for the $j-$th estimate $\boldsymbol{x}_j$ to obtain the proposed form $\mathbf{J_{12}^{\boldsymbol{\lambda}}}$.

Because JOBS procedure is a two-step relaxation of the ℓ_1 minimization, it gives some insight of why JOBS algorithm can correctly recover sparse solution, which is important for analyzing the algorithm. In Section 2, we will establish the correctness of JOBS algorithm rigorously.

8 Appendix D: Proofs of JOBS theorems

8.1 Proof of Theorem 2: correctness of JOBS

The first part of Theorem 2 can be directly shown from the BNSP for block sparse minimization problems as in (Eldar & Mishali, 2009). We only need to show the procedure to prove the latter part. If BNSP of order s is satisfied for $\{\boldsymbol{A}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_2], ..., \boldsymbol{A}[\mathcal{I}_K]\}$, then each bootstrap matrix $\boldsymbol{A}[\mathcal{I}_j]$ satisfies the Null Space Property (NSP) of order s, which is proven in Appendix 10.2. Consequently, for all $j = 1, 2, ..., K$, $\boldsymbol{x}^\star$ also turns out to be the optimal solution to all sub-problems: $\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}_j} \|\boldsymbol{x}_j\|_1$ s.t. $\boldsymbol{y}[\mathcal{I}_j] = \boldsymbol{A}[\mathcal{I}_j]\boldsymbol{x}_j$.

For $\boldsymbol{X}$ to be a feasible solution, consider its $\ell_{1,2}$ norm, we have:

$$\|\boldsymbol{X}\|_{1,2} = \sum_{i=1}^{n}(\sum_{j=1}^{K}(x_{ij}^2))^{1/2} = \sqrt{K}\sum_{i=1}^{n}(\frac{1}{K}\sum_{j=1}^{K}(x_{ij}^2))^{1/2}.$$

By concavity of the square root, we have

$$\begin{aligned}
&\|\boldsymbol{X}\|_{1,2} \\
&\geq \sqrt{K}\sum_{i=1}^{n}\frac{1}{K}\sum_{j=1}^{K}\sqrt{x_{ij}^2} = \sqrt{K}\frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{n}|x_{ij}| \\
&\geq \sqrt{K}\frac{1}{K}\sum_{j=1}^{K}\min_{\substack{\boldsymbol{x}_j: x_{1j},...,x_{nj} \\ \boldsymbol{A}[\mathcal{I}_j]\boldsymbol{x}_j = \boldsymbol{y}[\mathcal{I}_j]}}\sum_{i=1}^{n}|x_{ij}| \\
&= \sqrt{K}\frac{1}{K}\sum_{j=1}^{K}\min_{\boldsymbol{x}_j:\boldsymbol{A}[\mathcal{I}_j]\boldsymbol{x}_j=\boldsymbol{y}[\mathcal{I}_j]}\|\boldsymbol{x}_j\|_1 \\
&= \sqrt{K}\|\boldsymbol{x}^\star\|_1.
\end{aligned}$$

Since $\boldsymbol{X}^\star = (\boldsymbol{x}^\star, \boldsymbol{x}^\star, ..., \boldsymbol{x}^\star)$ is a feasible solution and $\|\boldsymbol{X}^\star\|_{1,2} = \|(\boldsymbol{x}^\star, \boldsymbol{x}^\star, ..., \boldsymbol{x}^\star)\|_{1,2} = \sqrt{K}\|\boldsymbol{x}^\star\|_1$, it achieves the lower bound. By the uniqueness part of the theorem, we can concluded that $\boldsymbol{X}^\star$ is the unique solution. Since the JOBS solution takes the average over columns of multiple estimates, we can easily deduce that JOBS returns the correct answer.

8.2 Proof of Theorem 4: JOBS performance bound of for exactly $s-$sparse signals

If the true solution is exactly $s-$sparse, the sparse approximation error is zero. Then the noise level of performance only relates to measurements noise. For ℓ_1 minimization, $\boldsymbol{z}$ is the noise vector and we use matrix $\boldsymbol{Z} = (\boldsymbol{z}[\mathcal{I}_1], \boldsymbol{z}[\mathcal{I}_2], ..., \boldsymbol{z}[\mathcal{I}_K])$ to denote the noise matrix in JOBS. We bound the distance of $\|\boldsymbol{Z}\|_{2,2}$ to its expected value using Hoeffding's inequalities stated in Hoeffding (1963).

Theorem 16 (Hoeffding's Inequalities) *Let $X_1, ..., X_n$ be independent bounded random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one. Denote their sum by $S_n = \sum_{i=1}^n X_i$. Then for any $\zeta > 0$, we have:*

$$\mathbb{P}\Big\{S_n - \mathbb{E}S_n \geq \zeta\Big\} \leq \exp\frac{-2\zeta^2}{\sum_{i=1}^n(b_i - a_i)^2} \;\; \textit{and} \tag{15}$$

530

$$\mathbb{P}\Big\{S_n - \mathbb{E}S_n \leq -\zeta\Big\} \leq \exp\frac{-2\zeta^2}{\sum_{i=1}^n(b_i - a_i)^2}. \tag{16}$$

Here, the entire noise vector is $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{y} = (z[1], z[2], ..., z[m])^T$, $\|\boldsymbol{z}\|_\infty = \max_{i=1,2,...,m}|z[i]| < \infty$. We consider the matrix $\boldsymbol{Z}\circ\boldsymbol{Z} = (\xi_{ji})$, where $\circ$ is the entry-wise product. The quantity that we are interested in $\|\boldsymbol{Z}\|_{2,2}$ is the sum of all entries in $\boldsymbol{Z}\circ\boldsymbol{Z}$. Each element in this matrix $\boldsymbol{Z}\circ\boldsymbol{Z}$ is drawn i.i.d from the squares of entries in $\boldsymbol{z}$: $\{z[1], z[2], ..., z[m]\}$ with equal probability. Let Ξ be the underlining random variable and Ξ obeys a discrete uniform distribution:

$$\mathbb{P}(\Xi = z^2[i]) = \frac{1}{m}, i = 1, 2, ..., m. \tag{17}$$

The lower and upper bound of Ξ is then

$$0 \leq \min_i z^2[i] \leq \Xi \leq \|\boldsymbol{z}\|_\infty^2. \tag{18}$$

We use zero as lower bound for Ξ instead of the minimum value to simplify the terms. The expected power of $\boldsymbol{Z}$ is

$$\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 = \frac{KL}{m}\|\boldsymbol{z}\|_2^2. \tag{19}$$

Applying Hoeffding's inequality for any $\tau > 0$ leads to

$$\mathbb{P}\{\|\boldsymbol{Z}\|_{2,2}^2 - \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 - \tau \leq 0\} \geq 1 - \exp\frac{-2\tau^2}{KL\|\boldsymbol{z}\|_\infty^4}. \tag{20}$$

Next, let $\widehat{\boldsymbol{X}}$ be the solution of $\mathbf{J}_{\mathbf{12}}^{\boldsymbol{\lambda}}$. Theorem 14 yields

$$\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_{2,2}^2 - \mathcal{C}_1^2(\delta)\|\boldsymbol{Z}\|_{2,2}^2 \leq 0\} = 1. \tag{21}$$

Let Δ denote the difference between the solution to the truth solution scaled by the $\mathcal{C}_1$ constant. Hence, $\Delta = \frac{1}{\mathcal{C}_1(\delta)}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_{2,2}$ and (21) becomes

$$\mathbb{P}\{\Delta - \|\boldsymbol{Z}\|_{2,2} \leq 0\} = 1. \tag{22}$$

Since $\boldsymbol{Z}$ depends on the choice of $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K$, we derive the typical performance by studying the distance of the solution to the expected noise level of JOBS.

$$\begin{aligned}
&\mathbb{P}\{\Delta^2 - \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 - \tau^2 \leq 0\} \\
&= \mathbb{P}\{\Delta^2 - \|\boldsymbol{Z}\|_{2,2}^2 + \|\boldsymbol{Z}\|_{2,2}^2 - \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 - \tau^2 \leq 0\} \\
&\geq \mathbb{P}\{\Delta^2 - \|\boldsymbol{Z}\|_{2,2}^2 \leq 0, \|\boldsymbol{Z}\|_{2,2}^2 - \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 - \tau^2 \leq 0\} \\
&\quad \text{(The first and the second parts are independent)} \\
&= \mathbb{P}\{\Delta^2 - \|\boldsymbol{Z}\|_{2,2}^2 \leq 0\}\mathbb{P}\{\|\boldsymbol{Z}\|_{2,2}^2 - \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 - \tau^2 \leq 0\} \\
&\quad \text{(using (22) and (20))} \\
&\geq 1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4}.
\end{aligned}$$

In summary, this procedure results in

$$\mathbb{P}\{\Delta^2 \leq \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 + \tau^2\} \geq 1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4}. \tag{23}$$

We can bound the squared error as follows:

$$\begin{aligned}
&\mathbb{P}\{\Delta \leq (\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2)^{1/2} + \tau\} \\
&= \mathbb{P}\{\Delta^2 \leq \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 + \tau^2 + 2\tau(\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2)^{1/2}\} \\
&\geq \mathbb{P}\{\Delta^2 \leq \mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2 + \tau^2\}.
\end{aligned} \tag{24}$$

Combining (23) and (24), we arrive at

$$\mathbb{P}\{\Delta \leq (\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2)^{1/2} + \tau\} \geq 1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4}. \tag{25}$$

Since $f(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2$ is convex, we can apply Jensens' inequality to establish:

$$\|\frac{1}{K}\sum_{j=1}^K \widehat{\boldsymbol{x}}_j - \boldsymbol{x}^\star\|_2^2 \leq \frac{1}{K}\sum_{j=1}^K \|\widehat{\boldsymbol{x}}_j - \boldsymbol{x}^\star\|_2^2. \tag{26}$$

The JOBS estimate is averaged column-wise over all estimates: $\boldsymbol{x}^{\boldsymbol{J}} = \frac{1}{K}\sum_{j=1}^K \widehat{\boldsymbol{x}}_j$. Therefore, equation (26) is essentially

$$\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2^2 - \frac{1}{K}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_{2,2}^2 \leq 0\} = 1. \tag{27}$$

Now, we consider the typical performance of the JOBS solution and recall that Δ denotes the difference between the solution to the truth solution scaled by the $\mathcal{C}_1$ constant: $\Delta = \frac{1}{\mathcal{C}_1(\delta)}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_{2,2}$. We can then bound the probability of error.

$$\begin{aligned}
&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 - \frac{\mathcal{C}_1(\delta)}{\sqrt{K}}((\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2)^{1/2} + \tau) \leq 0\} \\
=&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 - \frac{1}{\sqrt{K}}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2 \\
&+ \frac{1}{\sqrt{K}}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2 - \frac{\mathcal{C}_1(\delta)}{\sqrt{K}}((\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2)^{1/2} + \tau) \leq 0\} \\
\geq&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 - \frac{1}{\sqrt{K}}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2 \leq 0, \\
&\Delta \leq (\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2)^{1/2} + \tau\} \\
=&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 - \frac{1}{\sqrt{K}}\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2 \leq 0\} \\
&\mathbb{P}\{\Delta \leq (\mathbb{E}\|\boldsymbol{Z}\|_{2,2}^2)^{1/2} + \tau\} \qquad \text{(by (27) and (25))} \\
\geq&1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4}.
\end{aligned} \tag{28}$$

Substituting the expected noise level derived in (19) yields

$$\begin{aligned}
&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 \leq \mathcal{C}_1(\delta)(\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2 + \frac{\tau}{\sqrt{K}})\} \\
&\qquad\qquad \geq 1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4}.
\end{aligned}$$

By replacing $\tau/\sqrt{K}$ with τ, the quantity on the right hand side of the equation then becomes $1 - \exp\frac{-2K\tau^4}{L\|\boldsymbol{z}\|_\infty^4}$ and we have proved the theorem.

8.3 Proof of JOBS performance bound for general sparse signals

Similarly to prove the JOBS performance in the exact $s-$ sparse scenario, we here establish the proofs for the general sparse signals recovery for JOBS algorithm.

First, according to the general block sparse recovery Theorem 14, we consider the distance from the recovered solution $\widehat{\boldsymbol{X}}$ to the truth solution $\boldsymbol{X}^\star$.

$$\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_{2,2} - \mathcal{C}_0(\delta)s^{-1/2}\sqrt{K}|\boldsymbol{e}\|_1 - \mathcal{C}_1(\delta)\|\widetilde{\boldsymbol{Z}}\|_{2,2} \leq 0\} = 1. \tag{29}$$

To simplify our notation, we use $\epsilon(\boldsymbol{e})$ and $\epsilon(\boldsymbol{Z})$ for noise associated with $s-$ sparse approximation error and measurement error.

$$\begin{aligned}
\epsilon(\boldsymbol{e}) &= \mathcal{C}_0(\delta)s^{-1/2}\sqrt{K}|\boldsymbol{e}\|_1 \\
\epsilon(\boldsymbol{Z}) &= \mathcal{C}_1(\delta)\|\widetilde{\boldsymbol{Z}}\|_{2,2}.
\end{aligned} \tag{30}$$

We start the analysis:

$$\begin{aligned}
&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon^2(\boldsymbol{e}) + \mathbb{E}\epsilon^2(\boldsymbol{Z}) + \tau^2 + 2\epsilon(\boldsymbol{e})\mathbb{E}\epsilon(\boldsymbol{Z}) + 2\tau\epsilon(\boldsymbol{e})) \leq 0\} \\
=&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon^2(\boldsymbol{e}) + \mathbb{E}\epsilon^2(\boldsymbol{Z}) + \tau^2 + 2\epsilon(\boldsymbol{e})\mathbb{E}\epsilon(\boldsymbol{Z}) + 2\tau\epsilon(\boldsymbol{e})) \\
&- \epsilon^2(\boldsymbol{Z}) + \epsilon^2(\boldsymbol{Z}) - 2\epsilon(\boldsymbol{e})\epsilon(\boldsymbol{Z}) + 2\epsilon(\boldsymbol{e})\epsilon(\boldsymbol{Z}) \leq 0\} \\
\geq&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon^2(\boldsymbol{e}) + \epsilon^2(\boldsymbol{Z}) + 2\epsilon(\boldsymbol{e})\epsilon(\boldsymbol{Z})) \leq 0\} \\
&\mathbb{P}\{\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0, \\
&2\epsilon(\boldsymbol{e})(\epsilon(\boldsymbol{Z}) - \mathbb{E}\epsilon(\boldsymbol{Z}) - \tau)) \leq 0\}
\end{aligned} \tag{31}$$

Next, we establish that if the first condition is true, it implies the second condition: cond1 == True: $\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0 \Longrightarrow$ cond2 == True: $\epsilon(\boldsymbol{Z}) - \mathbb{E}\epsilon(\boldsymbol{Z}) - \tau \leq 0$.

$$\begin{aligned}
&\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0 \\
&\quad\Longleftrightarrow \epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0 \\
&\Longrightarrow \epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 - 2\tau\mathbb{E}\epsilon^2(\boldsymbol{Z}) \leq 0 \\
&\quad\Longleftrightarrow \epsilon^2(\boldsymbol{Z}) - (\mathbb{E}\epsilon^2(\boldsymbol{Z}) + \tau)^2 \leq 0 \\
&\quad\Longleftrightarrow \epsilon(\boldsymbol{Z}) - \mathbb{E}\epsilon(\boldsymbol{Z}) - \tau \leq 0
\end{aligned} \tag{32}$$

Therefore, the probability of cond2 conditional on cond1 is one. By Bayes Rule, we conclude the joint probably of cond1 and cond2 are both True equals to the probability of cond1 being True.

$$\begin{aligned}
&\mathbb{P}\{\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0, 2\epsilon(\boldsymbol{e})(\epsilon(\boldsymbol{Z}) - \mathbb{E}\epsilon(\boldsymbol{Z}) - \tau)) \leq 0\} \\
=&\mathbb{P}\{\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0\} \\
&\mathbb{P}\{(\epsilon(\boldsymbol{Z}) - \mathbb{E}\epsilon(\boldsymbol{Z}) - \tau)) \leq 0 | \epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0\} \\
=&\mathbb{P}\{\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0\}
\end{aligned} \tag{33}$$

We continue from (31),

$$\begin{aligned}
&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon^2(\boldsymbol{e}) + \mathbb{E}\epsilon^2(\boldsymbol{Z}) + \tau^2 + 2\epsilon(\boldsymbol{e})\mathbb{E}\epsilon(\boldsymbol{Z}) + 2\tau\epsilon(\boldsymbol{e})) \leq 0\} \\
\geq&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon(\boldsymbol{e}) + \epsilon^2(\boldsymbol{Z}))^2 \leq 0\} \\
&\mathbb{P}\{\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0\} \\
=&\mathbb{P}\{\epsilon^2(\boldsymbol{Z}) - \mathbb{E}\epsilon^2(\boldsymbol{Z}) - \tau^2 \leq 0\} \qquad \text{(by (20))} \\
&\geq 1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4 \mathcal{C}_1(\delta)^4}.
\end{aligned} \tag{34}$$

Next, we studied the error bound and relax the error bound by dropping one non-negative cross term $2\tau\mathbb{E}\epsilon(\boldsymbol{Z})$:

$$\begin{aligned}
&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon(\boldsymbol{e}) + \mathbb{E}\epsilon(\boldsymbol{Z}) + \tau)^2 \leq 0\} \\
=&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon^2(\boldsymbol{e}) + \mathbb{E}\epsilon^2(\boldsymbol{Z}) + \tau^2 + 2\epsilon(\boldsymbol{e})\mathbb{E}\epsilon(\boldsymbol{Z}) + 2\tau\epsilon(\boldsymbol{e}) + 2\tau\mathbb{E}\epsilon(\boldsymbol{Z})) \leq 0\} \\
\geq&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon^2(\boldsymbol{e}) + \mathbb{E}\epsilon^2(\boldsymbol{Z}) + \tau^2 + 2\epsilon(\boldsymbol{e})\mathbb{E}\epsilon(\boldsymbol{Z}) + 2\tau\epsilon(\boldsymbol{e})) \leq 0\}
\end{aligned} \tag{35}$$

Here, we started to link the JOBS solution error bound to total the error of all the joint sparsity program using the result in in equation (27) derived from Jensen's inequality of convex function:

$$\begin{aligned}
&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 - \frac{1}{\sqrt{K}}(\epsilon(\boldsymbol{e}) + \mathbb{E}\epsilon(\boldsymbol{Z}) + \tau) \leq 0\} \\
=&\mathbb{P}\{K\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2^2 - (\epsilon(\boldsymbol{e}) + \mathbb{E}\epsilon(\boldsymbol{Z}) + \tau)^2 \leq 0\} \\
=&\mathbb{P}\{K\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2^2 - (\epsilon(\boldsymbol{e}) + \mathbb{E}\epsilon(\boldsymbol{Z}) + \tau)^2 \\
&\quad + \|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - \|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2\} \leq 0\} \\
\geq&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon(\boldsymbol{e}) + \mathbb{E}\epsilon(\boldsymbol{Z}) + \tau)^2 \leq 0\} \\
&\mathbb{P}\{K\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2^2 - \|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2\} \leq 0\} \\
=&\mathbb{P}\{\|\widehat{\boldsymbol{X}} - \boldsymbol{X}^\star\|_2^2 - (\epsilon(\boldsymbol{e}) + \mathbb{E}\epsilon(\boldsymbol{Z}) + \tau)^2 \leq 0\} \qquad \text{(by (27))} \\
&\geq 1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4 \mathcal{C}_1(\delta)^4}.
\end{aligned} \tag{36}$$

From here, we plug in results from previous equations (19),(30):

$$\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}} - \boldsymbol{x}^\star\|_2 - (\mathcal{C}_0(\delta)s^{-1/2}\|\boldsymbol{e}\|_1 + \mathcal{C}_1(\delta)\sqrt{\frac{L}{m}}\epsilon + \frac{\tau}{\sqrt{K}}) \leq 0\} \geq 1 - \exp\frac{-2\tau^4}{KL\|\boldsymbol{z}\|_\infty^4 \mathcal{C}_1(\delta)^4}. \tag{37}$$

By replacing $\tau/\sqrt{K}$ with $\tau/\mathcal{C}_1(\delta)$, the quantity on the right hand side of the equation then becomes $1-\exp\frac{-2K\tau^4}{L\|\boldsymbol{z}\|_\infty^4}$ and we have proved the theorem.

$$\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{J}}-\boldsymbol{x}^\star\|_2 \le (\mathcal{C}_0(\delta)s^{-1/2}\|\boldsymbol{e}\|_1+\mathcal{C}_1(\delta)(\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2+\tau))\} \ge 1-\exp\frac{-2K\tau^4}{L\|\boldsymbol{z}\|_\infty^4}. \tag{38}$$

9 Appendix E: Proofs of Bagging theorems

9.1 Proof of Bagging Performance bound for exactly s-sparse signals

Let $\boldsymbol{x}_1^{\boldsymbol{B}}, \boldsymbol{x}_2^{\boldsymbol{B}}, ..., \boldsymbol{x}_K^{\boldsymbol{B}}$ be the solutions of individually solved problems and the solution of the bagging scheme $\boldsymbol{x}^{\boldsymbol{B}}$ is obtained from their average: $\boldsymbol{x}^{\boldsymbol{B}}=\frac{1}{K}\sum_{j=1}^K \boldsymbol{x}_j^{\boldsymbol{B}}$. We consider the distance to the true solution $\boldsymbol{x}^\star$ from each estimate separately. Here, the desired upper bound is the square root of the expected power of each noise vector: $(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2}=\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2$, where $\mathcal{I}$ is a multi-set of size L with each element randomly sampled from $\{1,2,...,m\}$. For any $\tau>0$, we have:

$$\begin{aligned}
&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2-\mathcal{C}_1(\delta)((\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2}+\tau)\le 0\}\\
=&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2-\mathcal{C}_1(\delta)(((\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2}+\tau)^2)^{1/2}\le 0\}\\
\ge&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2-\mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2+\tau^2)^{1/2}\le 0\}\\
=&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2+\tau^2)\le 0\}.
\end{aligned}$$

Consider using the average of errors for each estimate $\frac{1}{K}\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2$, we can establish

$$\begin{aligned}
&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2-\mathcal{C}_1(\delta)((\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2}+\tau)\le 0\}\\
&=\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-\frac{1}{K}\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2\\
&+\frac{1}{K}\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2+\tau^2)\le 0\}\\
&\ge\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-\frac{1}{K}\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2\le 0,\\
&\frac{1}{K}\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2+\tau^2)\le 0\}
\end{aligned}$$

(from the independence of two terms)

$$\begin{aligned}
&=\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-\frac{1}{K}\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2\le 0\}\\
&\times\mathbb{P}\{\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-K{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2+\tau^2)\le 0\}.
\end{aligned}$$

By Jensen's inequality, the bagging error is smaller than the averaged error of each individual estimator as in (26) and the first term holds with probability 1. Therefore, we have:

$$\begin{aligned}
&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2-\mathcal{C}_1(\delta)((\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2}+\tau)\le 0\}\\
\ge&\mathbb{P}\{\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2-K{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2+\tau^2)\le 0\}\\
=&1-\mathbb{P}\{\sum_{j=1}^K\|\boldsymbol{x}_j^{\boldsymbol{B}}-\boldsymbol{x}^\star\|_2^2\ge K{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2+\tau^2)\}.
\end{aligned} \tag{39}$$

From this procedure, we can reduce the error bound for the bagging algorithm to bound the sum of individual errors.

Let the random variable of error for each bagged estimator be $\boldsymbol{x}(\mathcal{I})$: $\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2$, where $\mathcal{I}$ denotes a bootstrap sample of size L and $\boldsymbol{x}(\mathcal{I})$ is the bagged solution from ℓ_1 minimization on the bootstrap sample $\mathcal{I}$: $\boldsymbol{x}(\mathcal{I}) = \arg\min \|\boldsymbol{x}\|_1$ s.t. $\|\boldsymbol{y}[\mathcal{I}] - \boldsymbol{A}[\mathcal{I}]\|_2^2 \leq \epsilon(\mathcal{I})$. The power of all errors for each bagged estimators $\|\boldsymbol{x}_j^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2$ are realizations generated i.i.d. from the distribution of $\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2$. We proceed with the proof using the following lemma that establishes the tail bound of the sum of i.i.d. bounded random variables. It is a generalization of Hoeffding's inequality and the details of its proof can be found in Appendix 10.4.

Lemma 17 (Tail bound of the sum of i.i.d. bounded Random variables) *Let $Y_1, Y_2, ..., Y_n$ be i.i.d. observations of bounded random variable Y: $a \leq Y \leq b$ and the expectation $\mathbb{E}Y$ exists. Then, for any $\zeta > 0$,*

$$\mathbb{P}\{\sum_{i=1}^n Y_i \geq n\zeta\} \leq \exp\{-\frac{2n(\zeta - \mathbb{E}Y)^2}{(b-a)^2}\}. \tag{40}$$

In this case, we consider the lower bound a and the upper bound b of the error $\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2$. Clearly this term is non-negative, hence, we can set $a = 0$. The upper bound is obtained from the error bound of ℓ_1-minimization in Theorem 13. For all $\mathcal{I}$:

$$\mathbb{P}\{\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 - {\mathcal{C}_1}^2(\delta)\|\boldsymbol{z}[\mathcal{I}]\|_2^2 \leq 0\} = 1. \tag{41}$$

According to the norm equivalence inequality, we have

$$\|\boldsymbol{z}[\mathcal{I}]\|_2^2 \leq (\sqrt{L}\|\boldsymbol{z}[\mathcal{I}]\|_\infty)^2 \leq (\sqrt{L}\|\boldsymbol{z}\|_\infty)^2 = L\|\boldsymbol{z}\|_\infty^2. \tag{42}$$

From this, we can set $b = {\mathcal{C}_1}^2(\delta)L\|\boldsymbol{z}\|_\infty^2$.

We can now apply the sum of i.i.d. bounded random variable in Theorem 17 to analyze our problem. By (39), the parameter ζ in (40) turns out to be: $\zeta = {\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2 + \tau^2)$. Hence,

$$\begin{aligned}\mathbb{P}\{\sum_{j=1}^K \|\boldsymbol{x}_j - \boldsymbol{x}^\star\|_2^2 - K\zeta \geq 0\} \leq \\ \exp\{-\frac{2K(\zeta - \mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|)^2}{{\mathcal{C}_1}^4(\delta)L^2\|\boldsymbol{z}\|_\infty^4}\}.\end{aligned} \tag{43}$$

To simplify the right hand side, let us consider the expected bagged error: $\mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 = \frac{1}{|m^L|}\sum_{\mathcal{I}} \|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2$. Our bound in (41) implies that

$$\mathbb{P}\{\frac{1}{|m^L|}\sum_{\mathcal{I}} \|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 \leq \frac{1}{|m^L|}\sum_{\mathcal{I}} {\mathcal{C}_1}^2(\delta)\|\boldsymbol{z}_{\mathcal{I}}\|_2^2\} = 1,$$

which is equivalent to

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 &\leq \frac{1}{|m^L|}\sum_{\mathcal{I}} {\mathcal{C}_1}^2(\delta)\|\boldsymbol{z}[\mathcal{I}]\|_2^2 \\ &= \mathbb{E}\,{\mathcal{C}_1}^2(\delta)\|\boldsymbol{z}_{\mathcal{I}}\|_2^2 = {\mathcal{C}_1}^2(\delta)\mathbb{E}\|\boldsymbol{z}_{\mathcal{I}}\|_2^2.\end{aligned} \tag{44}$$

From here, it is easy to see that

$$\begin{aligned}&\zeta - \mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 \\ =&{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2 + \tau^2) - \mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 \\ \geq&{\mathcal{C}_1}^2(\delta)(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2 + \tau^2) - {\mathcal{C}_1}^2(\delta)\mathbb{E}\|\boldsymbol{z}_{\mathcal{I}}\|_2^2 = {\mathcal{C}_1}^2(\delta)\tau^2.\end{aligned} \tag{45}$$

The right hand side of (43) is upper bounded by $\exp\{-\frac{2K\tau^4}{L^2\|\boldsymbol{z}\|_\infty^4}\}$. Substituting this result into (39), we can obtain the result in our main bagging theorem.

9.2 Proof of Bagging performance bound of bagging for approximately sparse signals

In this section, we are working with the case when the true solution $\boldsymbol{x}^\star$ is only approximately sparse. In other words, its sparsity level may exceed s and the $s-$sparse approximation error is no longer necessarily zero. Let ϵ_s denote the sparse approximation error $\epsilon_s = \mathcal{C}_0(\delta)s^{-1/2}\|\boldsymbol{e}\|_1$. The square root of the expected power of each noise vector is $(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2)^{1/2} = \sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2$. We consider the following bound:

$$
\begin{aligned}
&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2 - (\epsilon_s + \mathcal{C}_1(\delta)((\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2)^{1/2} + \tau)) \leq 0\}\\
=&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2 - (\epsilon_s + \mathcal{C}_1(\delta)((\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2)^{1/2} + \tau))^2 \leq 0\}\\
\geq&\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2 -\\
&((\epsilon_s + \mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2)^{1/2})^2 + {\mathcal{C}_1}^2(\delta)\tau^2) \leq 0\}.
\end{aligned}
$$

Set $\zeta' = (\epsilon_s + \mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2)^{1/2})^2 + {\mathcal{C}_1}^2(\delta)\tau^2$ and consider using the averages of the errors $\frac{1}{K}\sum_{j=1}^K \|\boldsymbol{x}_j^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2$ as an intermediate term. Repeating the same proving technique as in (39) yields

$$
\begin{aligned}
\mathbb{P}\{\|\boldsymbol{x}^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2 - \zeta'\} &\geq \mathbb{P}\{\sum_{j=1}^K \|\boldsymbol{x}_j^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2 - K\zeta' \leq 0\}\\
&= 1 - \mathbb{P}\{\sum_{j=1}^K \|\boldsymbol{x}_j^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2 \geq K\zeta'\}.
\end{aligned}
$$

According to Lemma 17 , we have:

$$
\begin{aligned}
&\mathbb{P}\{\sum_{j=1}^K \|\boldsymbol{x}_j^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2 \geq K\zeta'\} \leq\\
&\exp\{-\frac{2K(\zeta' - \mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2)^2}{(b' - a')^2}\}.
\end{aligned}
\tag{46}
$$

Here, $a' = 0$ and $b' = (\epsilon_s + \mathcal{C}_1(\delta)\sqrt{L}\|\boldsymbol{z}\|_\infty)^2$. The lower bound a' is set to zero since the error for any bagged estimator $\|\boldsymbol{x}_j^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2$ is non-negative. The upper bound b' can be obtained using Theorem 13 and substituting in the upper bound of the noise power as derived in (42).

Next, consider the term $\zeta' - \mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 = (\mathcal{C}_0(\delta)s^{-1/2}\|\boldsymbol{e}\|_1 + \mathcal{C}_1(\delta)\sqrt{\frac{L}{m}}\|\boldsymbol{z}\|_2)^2 + {\mathcal{C}_1}^2(\delta)\tau^2 - \mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2$. We can upper bound the expected value of the error of bagged estimator with same approach in (44). From Theorem 13, for all $\mathcal{I}$:

$$
\mathbb{P}\{\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 \leq (\epsilon_s + \mathcal{C}_1(\delta)\|\boldsymbol{z}[\mathcal{I}]\|_2)^2\} = 1. \tag{47}
$$

Since $\mathcal{I}$ takes value of all m^L choices with equal probability, the following result is implied from (47):

$$
\mathbb{P}\{\mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 \leq \mathbb{E}(\epsilon_s + \mathcal{C}_1(\delta)\|\boldsymbol{z}[\mathcal{I}]\|_2)^2\} = 1. \tag{48}
$$

Since $f(x) = x^2$ is a convex function, applying Jensen's inequality results in

$$
(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2)^2 \leq \mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2.
$$

Since the square root $x^{1/2}$ is a increasing function of x, taking square root preserves the sign of the inequality:

$$
\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2 \leq (\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2)^{1/2}. \tag{49}
$$

Then, from (48), we have:

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{x}(\mathcal{I}) - \boldsymbol{x}^\star\|_2^2 &\leq \mathbb{E}(\epsilon_s + \mathcal{C}_1(\delta)\|\boldsymbol{z}[\mathcal{I}]\|_2)^2\\
&= \epsilon_s^2 + {\mathcal{C}_1}^2(\delta)\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2 + 2\epsilon_s\mathcal{C}_1(\delta)\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2\\
&\text{(by (49))}\\
&\leq \epsilon_s^2 + {\mathcal{C}_1}^2(\delta)\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2 + 2\epsilon_s\mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}_{\mathcal{I}}\|_2^2)^{1/2}\\
&= (\epsilon_s + \mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}[\mathcal{I}]\|_2^2)^{1/2})^2.
\end{aligned}
$$

Finally, we can bound the term $\zeta' - \mathbb{E}\|\boldsymbol{x}_{(\mathcal{I})} - \boldsymbol{x}^\star\|_2^2$:

$$\begin{aligned}
&\zeta' - \mathbb{E}\|\boldsymbol{x}_{(\mathcal{I})} - \boldsymbol{x}^\star\|_2^2 \\
=&(\epsilon_s + \mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2})^2 + {\mathcal{C}_1}^2(\delta)\tau^2 - \mathbb{E}\|\boldsymbol{x}_{(\mathcal{I})} - \boldsymbol{x}^\star\|_2^2 \\
\geq&((\epsilon_s + \mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2})^2 + {\mathcal{C}_1}^2(\delta)\tau^2 \\
&- (\epsilon_s + \mathcal{C}_1(\delta)(\mathbb{E}\|\boldsymbol{z}_{[\mathcal{I}]}\|_2^2)^{1/2})^2 = {\mathcal{C}_1}^2(\delta)\tau^2.
\end{aligned}$$

One can observe that the upper bound of this difference is ${\mathcal{C}_1}^2(\delta)\tau^2$, which is the same as in the case of the exact $s-$sparse signal in (45). The bound for (46) can be upper bounded by

$$\mathbb{P}\{\sum_{j=1}^{K}\|\boldsymbol{x}_j^{\boldsymbol{B}} - \boldsymbol{x}^\star\|_2^2 - K\zeta' \geq 0\} \leq \exp\{-\frac{2K{\mathcal{C}_1}^4(\delta)\tau^4}{(b')^2}\},$$

where $b' = (\mathcal{C}_0(\delta)s^{-1/2}\|\boldsymbol{e}\|_1 + \mathcal{C}_1(\delta)\sqrt{L}\|\boldsymbol{z}\|_\infty)^2$.

10 Appendix F: Theory for NSP and RIP

10.1 Proof of the reverse direction for noiseless recovery

Lemma 18 *If the MMV problem $\mathbf{P_1}(K)$, $K > 1$, in (12) has a unique solution, it will be of form $\boldsymbol{X}^\star = (\boldsymbol{x}^\star, \boldsymbol{x}^\star, ..., \boldsymbol{x}^\star)$. Then, there is a unique solution to $\mathbf{P_1}$: $\boldsymbol{x}^\star$.*

Let us prove the other direction. If $\mathbf{P_1}(K)$ has a unique solution, the solution must be in the form of $\boldsymbol{X}^\star = (\boldsymbol{x}^\star, \boldsymbol{x}^\star, ..., \boldsymbol{x}^\star)$, and it implies that $\mathbf{P_1}$ has a unique solution $\boldsymbol{x}^\star$.

If $\mathbf{P_1}(K)$ has a unique solution, then it is equivalent to say that $\boldsymbol{A}$ satisfied BNSP of order s. For all $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_K) \neq \boldsymbol{O}, \boldsymbol{v}_j \in \text{Null}(\boldsymbol{A})$, we have $\forall\ \mathcal{S}, |\mathcal{S}| \leq s, \|\boldsymbol{V}[\mathcal{S}]\|_{1,2} < \|\boldsymbol{V}[\mathcal{S}^c]\|_{1,2}$. This implies that $\forall\ \boldsymbol{V} = (\boldsymbol{v}, \mathbf{0}, \mathbf{0}, ..., \mathbf{0}), \boldsymbol{v} \in \text{Null}(\boldsymbol{A})\backslash\{\mathbf{0}\}$, BNSP is satisfied. Since in this case, except the first column, all others are zero and therefore do not contribute any to the group norm. Mathematically, for all $\mathcal{S}$, $\|\boldsymbol{V}[\mathcal{S}]\|_{1,2} = \|\boldsymbol{v}[\mathcal{S}]\|_1$. We, therefore, will have the BNSP of order s, implying the NSP for $\boldsymbol{A}$ of order s.

10.2 JOBS matrix satisfies BNSP implies that each block matrix satisfies NSP

Using a similar analysis as in previous subsection 10.1, we conclude that a block diagonal matrix satisfies BNSP of order s implies that each submatrix satisfies NSP of order s. The block diagonal JOBS matrix $\boldsymbol{A}^{\boldsymbol{J}} = \text{block\_diag}(\boldsymbol{A}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_2], ..., \boldsymbol{A}[\mathcal{I}_K])$ satisfies BNSP of order s. Then, for all $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_K) \neq \boldsymbol{O}, \boldsymbol{v}_j \in \text{Null}(\boldsymbol{A}[\mathcal{I}_j]), j = 1, 2, .., K$, we have $\forall\ \mathcal{S}, |\mathcal{S}| \leq s, \|\boldsymbol{V}[\mathcal{S}]\|_{1,2} < \|\boldsymbol{V}[\mathcal{S}^c]\|_{1,2}$. This implies that $\forall\ \boldsymbol{V} = (\mathbf{0}, ..., \boldsymbol{v}_j, ..., \mathbf{0}), \boldsymbol{v}_j \in \text{Null}(\boldsymbol{A}[\mathcal{I}_j])\backslash\{\mathbf{0}\}$, BNSP is satisfied, which essentially states that NSP is statisfied for $\boldsymbol{A}[\mathcal{I}_j]$.

10.3 Proof of Proposition 3

To prove this proposition, we give an alternative form of RIP and BRIP which are stated in the following two propositions. Alternative form of RIP as a function of matrix induced norm is given as follows.

Proposition 19 (Alternative form of RIP) *Matrix $\boldsymbol{A}$ has ℓ_2-normalized columns, and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\mathcal{S} \subset \{1, 2, ..., n\}$ with size smaller or equal to s and $\boldsymbol{A}_\mathcal{S}$ takes columns of $\boldsymbol{A}$ with indices in $\mathcal{S}$. The RIP constant of order s of $\boldsymbol{A}$, $\delta_s(\boldsymbol{A})$ is:*

$$\delta_s(\boldsymbol{A}) = \max_{\mathcal{S}\subseteq\{1,2,...,n\},|\mathcal{S}|\leq s} \|\boldsymbol{A}_\mathcal{S}^{\boldsymbol{T}}\boldsymbol{A}_\mathcal{S} - \boldsymbol{I}\|_{2\to 2}, \tag{50}$$

where $\boldsymbol{I}$ is an identity matrix of size $s \times s$ and $\|\cdot\|_{2\to2}$ is the induced $2-$norm defined as for any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{2\to2} = \sup_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$.

This proposition can be directly derived from the definition of RIP constant. Similarly, we can derive the alternative form of BRIP constant as a function of matrix induced norm.

Proposition 20 (Alternative form of BRIP) *Let matrix $\boldsymbol{A}$ have ℓ_2-normalized columns and let $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_n\}$ be the group sparsity pattern that defines the row sparsity pattern, with $\mathcal{B}_i$ contains all indices corresponding to all elements of the $i-$th row. For $\mathcal{S} \subseteq \{1, 2, ..., n\}$, denote $\mathcal{B}(\mathcal{S}) = \{\mathcal{B}_i, i \in \mathcal{S}\}$ as the subsets that takes several groups with group indices in $\mathcal{S}$. For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with Block-RIP constant of order s, $\delta_{s|\mathcal{B}}(\boldsymbol{A})$ is*

$$\delta_{s|\mathcal{B}} = \max_{\mathcal{S}\subseteq\{1,2,...,n\}, |\mathcal{S}|\leq s} \|\boldsymbol{A}^{\boldsymbol{T}}_{\mathcal{B}(\mathcal{S})}\boldsymbol{A}_{\mathcal{B}(\mathcal{S})} - \boldsymbol{I}\|_{2\to 2}. \tag{51}$$

Without loss of generality, let us assume that all columns of $\boldsymbol{A}$ in the original ℓ_1 minimization have unit ℓ_2 norms. Therefore, $\boldsymbol{A}$ does not have any zero column. Before we calculate the RIP constant of the bootstrapped sensing matrices, we need to perform two operations: remove the duplicate rows from bootstrapped sensing matrices and then normalize the columns.

First, we remove the duplicated rows using the weighted scheme. In the noisy recovery problem, for a multi-set $\mathcal{I}$ that may contain duplicate, the set $\mathcal{U}$ denotes the set of all unique elements. In the constraint optimization, we can express the sum using occurrence times in $\mathcal{I}$ for each element using c_i. $\|\boldsymbol{A}[\mathcal{I}]\boldsymbol{x} - \boldsymbol{y}[\mathcal{I}]\|_2^2 = \sum_{i\in\mathcal{I}} \|\boldsymbol{a}[i]\boldsymbol{x} - \boldsymbol{y}[i]\|_2^2 = \sum_{i\in\mathcal{U}} \|\sqrt{c_i}\boldsymbol{a}[i]\boldsymbol{x} - \boldsymbol{y}[i]\|_2^2$. Therefore, the original program is equivalent to reducing the duplicated rows in the bootstrap sample using $\sqrt{c_i}$ as weights. Because sampling with replacement is uniform, therefore the expected values of occurrence times for each sample are the same. To denote this operation, we have $R \in \mathbb{R}^{u\in L}$, $\boldsymbol{R} = \text{diag}(\sqrt{c_1}, \sqrt{c_2}, ..., \sqrt{c_u})\boldsymbol{I}[\mathcal{U}]$, each row of $\boldsymbol{I}[\mathcal{U}]$ correponds to the unique vector of a row and this operation deletes the duplicated rows.

Second, we normalize the columns of these matrices using the following normalization procedure. For $\boldsymbol{M} \in \mathbb{R}^{u\times n}$, since the original matrix $\boldsymbol{A}$ does not have any zero column, $\boldsymbol{Q}(\boldsymbol{M}) \in \mathbb{R}^{n\times n}$ is a normalization matrix of $\boldsymbol{M}$ such that $\boldsymbol{M}\boldsymbol{Q}(\boldsymbol{M})$ has ℓ_2-normalized columns. Clearly, the normalization matrix $\boldsymbol{Q}$ of $\boldsymbol{M}$ is obtained by:

$$\boldsymbol{Q}(\boldsymbol{M}) = \text{diag}(\|\boldsymbol{m}_1\|_2^{-1}, \|\boldsymbol{m}_2\|_2^{-1}, ..., \|\boldsymbol{m}_n\|_2^{-1}), \tag{52}$$

where $\boldsymbol{m}_j$ denotes $j-$th column of $\boldsymbol{M}$.

Similary, we can construct $\boldsymbol{Q}_j$s using (52) to normalize the columns. Let the original JOBS matrix be $\boldsymbol{A^J} = \text{block\_diag}(\boldsymbol{A}[\mathcal{I}_1], \boldsymbol{A}[\mathcal{I}_2], ..., \boldsymbol{A}[\mathcal{I}_K])$. We first normalize each block and then obtain the normalized bootstrapped sensing matrix as: $\widetilde{\boldsymbol{A}[\mathcal{I}_j]} = \boldsymbol{R}_j\boldsymbol{A}[\mathcal{I}_j]\boldsymbol{Q}_j$. The original JOBS matrix can be transferred into the normalized version $\widetilde{\boldsymbol{A^J}} = \text{block\_diag}(\widetilde{\boldsymbol{A}[\mathcal{I}_1]}, \widetilde{\boldsymbol{A}[\mathcal{I}_2]}, ..., \widetilde{\boldsymbol{A}[\mathcal{I}_K]})$.

Now, we consider the BRIP constant for $\boldsymbol{A^J}$. In this derivation, column selection of a matrix is written as a right multiplication of the matrix $\boldsymbol{I}_{\mathcal{S}}(\cdot)$.

$$\begin{aligned}
\delta_{s|\mathcal{B}}(\boldsymbol{A^J}) &= \delta s|\mathcal{B}(\widetilde{\boldsymbol{A^J}}) \\
&= \max_{\mathcal{S}\subseteq\{1,2,..,n\},|\mathcal{S}|\leq s} \|(\widetilde{\boldsymbol{A^J}}\boldsymbol{I}_{\mathcal{B}(\mathcal{S})})^{\boldsymbol{T}}\widetilde{\boldsymbol{A^J}}\boldsymbol{I}_{\mathcal{B}(\mathcal{S})} - \boldsymbol{I}\|_{2\to 2} \\
&= \max_{\substack{\mathcal{S}\subseteq\{1,2,..,n\},\\ |\mathcal{S}|\leq s}} \max_j \|(\widetilde{\boldsymbol{A}[\mathcal{I}_j]}\boldsymbol{I}_{\mathcal{S}})^{\boldsymbol{T}}\,\widetilde{\boldsymbol{A}[\mathcal{I}_j]}\boldsymbol{I}_{\mathcal{S}} - \boldsymbol{I}\|_{2\to 2} \\
&= \max_{\substack{\mathcal{S}\subseteq\{1,2,..,n\},\\ |\mathcal{S}|\leq s}} \|\text{block_diag}((\widetilde{\boldsymbol{A}[\mathcal{I}_1]}\boldsymbol{I}_{\mathcal{S}})^{\boldsymbol{T}}\widetilde{\boldsymbol{A}[\mathcal{I}_1]}\boldsymbol{I}_{\mathcal{S}} - \boldsymbol{I}, \\
&\qquad ..., (\widetilde{\boldsymbol{A}[\mathcal{I}_K]}\boldsymbol{I}_{\mathcal{S}})^{\boldsymbol{T}}\widetilde{\boldsymbol{A}[\mathcal{I}_K]}\boldsymbol{I}_{\mathcal{S}} - \boldsymbol{I})\|_{2\to 2}.
\end{aligned}$$

The induced $2-$norm of a matrix equals to the max singular value of $\|\boldsymbol{D}\|_{2\to 2} = \sigma_{\max}(\boldsymbol{D})$ and if $\boldsymbol{D}$ is a block diagonal matrix $\boldsymbol{D} = \text{diag}(\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_K)$, then $\sigma_{\max}(\boldsymbol{D}) = \max_{j=1,2,...,K} \sigma_{\max}(\boldsymbol{D}_j)$.

Applying this property leads to

$$\begin{aligned}
&\delta_{s|\mathcal{B}}(\boldsymbol{A}^{\boldsymbol{J}}) \\
&= \max_{\substack{\mathcal{S}\subseteq\{1,2,..,n\},\\ |\mathcal{S}|\leq s}} \max_{j} \|(\widetilde{\boldsymbol{A}[\mathcal{I}_j]\boldsymbol{I}_{\mathcal{S}}})^{\boldsymbol{T}}\, \widetilde{\boldsymbol{A}[\mathcal{I}_j]}\boldsymbol{I}_{\mathcal{S}} - \boldsymbol{I}\|_{2\to 2} \\
&= \max_{j} \max_{\substack{\mathcal{S}\subseteq\{1,2,..,n\},\\ |\mathcal{S}|\leq s}} \|(\widetilde{\boldsymbol{A}[\mathcal{I}_j]\boldsymbol{I}_{\mathcal{S}}})^{\boldsymbol{T}}\, \widetilde{\boldsymbol{A}[\mathcal{I}_j]}\boldsymbol{I}_{\mathcal{S}} - \boldsymbol{I}\|_{2\to 2} \\
&= \max_{j} \delta_s(\boldsymbol{A}[\mathcal{I}_j]).
\end{aligned}$$

10.4 Proof of Lemma 17

To prove of this lemma, We would need the Markov's inequality for non-negative random variables here. Let X be a non-negative random variable and suppose that $\mathbb{E}X$ exists. For any $t > 0$, we have:

$$\mathbb{P}\{X > t\} \leq \frac{\mathbb{E}X}{t}. \tag{53}$$

We also need the upper bound of the moment generating function (MGF) of the random variable Y. Suppose that $a \leq Y \leq b$, then for all $t \in \mathbb{R}$,

$$\mathbb{E}\exp\{tY\} \leq \exp\{t\mathbb{E}Y + \frac{t^2(b-a)^2}{8}\}. \tag{54}$$

Back to Lemma 17, for $t > 0$,

$$\begin{aligned}
\mathbb{P}\{\sum_{i=1}^{n} Y_i \geq n\zeta\} &= \mathbb{P}\{\exp\{\sum_{i=1}^{n} Y_i\} \geq \exp\{n\zeta\}\} \\
&= \mathbb{P}\{\exp\{t\sum_{i=1}^{n} Y_i\} \geq \exp\{tn\zeta\}\}
\end{aligned}$$

using the Markov inequality in (53)

$$\begin{aligned}
&\leq \exp\{-tn\zeta\}\mathbb{E}\{\exp\{t\sum_{i=1}^{n} Y_i\}\} \\
&= \exp\{-tn\zeta\}\mathbb{E}\{\Pi_{i=1}^{n}\exp\{tY_i\}\} \\
&= \exp\{-tn\zeta\}\Pi_{i=1}^{n}\mathbb{E}\{\exp\{tY_i\}\}
\end{aligned}$$

by upper bound for MGF in (54)

$$\begin{aligned}
&\leq \exp\{-tn\zeta\}(\exp\{t\mathbb{E}Y + \frac{t^2(b-a)^2}{8}\})^n \\
&= \exp\{-tn\zeta + tn\mathbb{E}Y + \frac{t^2(b-a)^2 n}{8}\}.
\end{aligned}$$

The right hand side is a convex function with respect to t. Taking the derivative with respect to t and set it zero, we obtain the optimal t, $t^\star = \frac{4\zeta - 4EY}{(b-a)^2}$. The right hand side is minimized at value:

$$\exp\{-t^\star n\zeta + t^\star n\mathbb{E}Y + \frac{{t^\star}^2(b-a)^2 n}{8}\} = \exp\{\frac{-2n(\zeta-\mathbb{E}Y)^2}{(b-a)^2}\}.$$

11 Appendix G: Pseudo-code of JOBS implementation via ADMM

We present the pseudo-code for solving JOBS optimization problem via ADMM updates. The key difference to Bagging and the baseline ℓ_1 minimization here is that we employ the soft-thresholding operation on each row in JOBS (described in line 6 of Algorithm 1), rather than the common entry-wise thresholding operation on each individual sparse-code element in Bagging.

Algorithm 1 ADMM for solving JOBS

Require: Sensing matrix and measurements vector $(\boldsymbol{A}, \boldsymbol{y})$, bootstrap ratio and number of estimates $(L/m, K)$, sparse balancing ratio λ, learning rate ρ, maximum number of iterations MaxIter.
Initialization: $\widehat{\boldsymbol{X}}_0, \boldsymbol{W}_0, \boldsymbol{U}_0 \leftarrow \boldsymbol{O}$ (zero matrix of size $n \times K$)

generate K bootstrap samples of length L:
$\{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K\}$, and its corresponding $\{\boldsymbol{A}[\mathcal{I}_j], \boldsymbol{y}[\mathcal{I}_j]\}$
for $t = 1$: MaxIter **do**
  $\widehat{\boldsymbol{X}}$ update: $\widehat{\boldsymbol{x}}_j \leftarrow$
    $(\boldsymbol{A}[\mathcal{I}_j]^* \boldsymbol{A}[\mathcal{I}_j] + \rho \boldsymbol{I})^{-1}(\boldsymbol{A}[\mathcal{I}_j]^* \boldsymbol{y}[\mathcal{I}_j] + \rho(\boldsymbol{w} - \boldsymbol{u}))$
  $\widehat{\boldsymbol{X}} \leftarrow \alpha \widehat{\boldsymbol{X}} + (1-\alpha)\boldsymbol{W}$
  $\boldsymbol{W}$ update: applying shrinkage operations on each row. For $i = 1, 2, .., n$,
  $\boldsymbol{w}[i] \leftarrow \text{Shrinkage}_{\lambda/\rho}(\widehat{\boldsymbol{x}}[i] - \boldsymbol{u}[i])$,
  $\text{Shrinkage}_\kappa(\boldsymbol{x}) = \max(1 - \kappa/\|\boldsymbol{x}\|_2, 0)\boldsymbol{x}$
  $\boldsymbol{U}$ update: $\boldsymbol{U} = \boldsymbol{U} + \boldsymbol{X} - \boldsymbol{W}$
end for
JOBS solution is the average columns of solution matrix $\widehat{\boldsymbol{X}}$: $\boldsymbol{x}^{\boldsymbol{J}} = 1/K \sum \widehat{\boldsymbol{x}}_j$

12 Appendix H: Distribution of the unique number of elements for bootstrapping

H The bootstrap is essentially sampling with replacement, which is likely to create duplicate information. The performance of sampling with replacement and sampling without replacement (sub-sampling) can be linked by studying the quantity of the number of unique elements. In this section, we give the analytic form of the number of unique samples when there are finite number of measurements m and bootstrap sample L, as well as the form for asymptotic case as $m \to \infty$. The finite case is studied in a well-known statistics problem – the Birthday Problem (bir). We also show empirically that the finite m case is close in the asymptotic sense.

12.0.1 Unique number of Bootstrap Samples with Finite Sample m

We generate L samples from m samples uniformly at random with replacement ($L \leq m$). Let U denote the number of distinct samples among L samples. Clearly we have the number of distinct samples is between $[1, L]$ and the probability mass function is given by (bir), same as the famous Birthday problem in statistics:

$$\begin{aligned} \mathbb{P}(U = u) &= \binom{m}{u} \sum_{j=0}^{u} (-1)^j \binom{u}{j} (\frac{u-j}{m})^L, \\ u &= 1, 2, ..., L. \end{aligned} \tag{55}$$

In our problem, we are interested in finding the lower bound of U with certainty $1 - \alpha$

$$\mathbb{P}(U \geq d) = \sum_{U=d}^{L} \binom{m}{u} \sum_{j=0}^{u} (-1)^j \binom{u}{j} (\frac{u-j}{m})^L \geq 1 - \alpha. \tag{56}$$

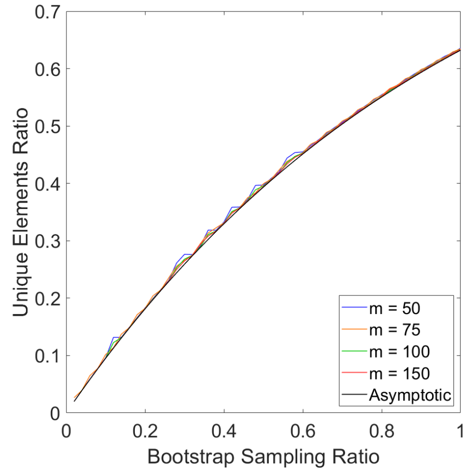
Therefore for

$$1 \geq \alpha \geq \sum_{u=0}^{d-1} \binom{m}{u} \sum_{j=0}^{u} (-1)^j \binom{u}{j} (\frac{u-j}{m})^L, \tag{57}$$

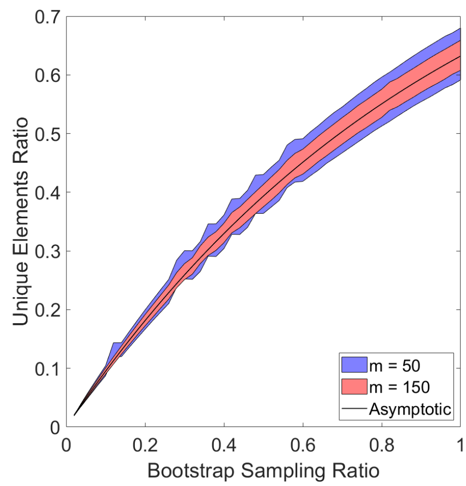
equation (56) is satisfied.

12.0.2 Asymptotic unique ratios of bootstrap samples

The theoretically unique percentage for asymptotic case when the number of total number of measurements goes to infinity $m \to \infty$ has been studied in the literature (Weiss, 1958; Mendelson et al.,



(a) The mean of unique element ratios under various sampling with replacement/ bootstrapping ratios with various number of measurements: $m = 50, 75, 100, 150$ and theoretical value when $m = \infty$.



(b) The area between of empirical mean plus and minus one empirical standard deviation. The blue area and the red area corresponds to the number of total measurements $m = 50$ and 150 respectively. The black line is the asymptotic mean and the asymptotic variance converges to zero

Figure 6: Unique element ratios with various bootstrapping ratios.

2016). In the limit case, the limiting distribution of the number of unique elements U is normal. The asymptotic mean for the unique number of elements over total number of measurements m is $\mathbb{E}\frac{U}{m} = 1 - \exp\{-r\}$, where r is the bootstrap sampling rate. The asymptotic variance of the unique ratio is then $\mathrm{Var}\frac{U}{m} = \frac{1}{m}(\exp\{-r\} - (1+r)\exp\{-2r\})$, which converges to zero when m is large.

12.0.3 FINITE NUMBER OF MEASUREMENTS m CASES ARE EMPRICIALLY CLOSE TO THE ASYMPTOTIC CASE

We generate 10000 trials of random sampling with replacement and then calculate the empirical unique percentage by counting the ratio of the number of unique elements over the total number of measurements m. The theoretical mean is consistently lower than the mean for a finite m. From the plot, the average unique elements in finite m cases $m = 50, 75, 100, 150$ are not so different from the theoretical value of the infinite sample size.

The empirical mean and the asymptotic value are plotted in Figure 6a, indicating that the numeric unique percentage is not that far from the asymptotic value even when the number of estimates is finite and small. Figure 6b illustrates the region between the mean plus and minus one standard of deviation. As the asymptotic case, the theoretical standard deviation converges to zero. We plotted the cases $m = 150$ and $m = 50$ compared to the asymptotic case. For both, the variance is tight and gets smaller when m becomes larger. For the same m, the variance of the unique number of elements become larger when the bootstrap ratio L/m is large.

12.1 THE SUB-SAMPLING VARIATION: SUB-JOBS

Bootstrapping (random sampling with replacement) creates duplicates within a bootstrap sample. Although it simplifies the analysis, in practice, duplicate information does not add value. One natural extension of the proposed framework is to use sub-sampling: sampling without replacement. The sub-sampling variation of Bagging is known as Subagging estimator in the literature (Bühlmann & Yu, 2000; Bühlmann, 2003). We adopt a similar name for the sub-sampling variation of the proposed method: Sub-JOBS. The only difference to the original scheme is that for each bootstrap sample $\mathcal{I}_j$, L distinct samples are generated by random sampling without replacement from m measurements.

In this paper, all the theoretical results are for the bootstrapping version for simplicity of presentation.