Greedy-type sparse recovery from heavy-tailed measurements

***Abstract*—Recovering a s-sparse signal vector $x \in \mathbb{C}^n$ from a comparably small number of measurements $y := (Ax) \in \mathbb{C}^m$ is the underlying challenge of compressed sensing. By now, a variety of efficient greedy algorithms has been established and strong recovery guarantees have been proven for random measurement matrices $A \in \mathbb{C}^{m \times n}$.**
However, they require a strong concentration of A^*Ax around its mean x (in particular, the Restricted Isometry Property), which is generally not fulfilled for heavy-tailed matrices. In order to overcome this issue and even cover applications where only limited knowledge about the distribution of the measurements matrix is known, we suggest substituting A^*Ax by a median-of-means estimator.
In the following, we present an adapted greedy algorithm, based on median-of-means, and prove that it can recover any s-sparse unit vector $x \in \mathbb{C}^n$ up to a l_2-error $\|x - \hat{x}\|_2 < \epsilon$ with high probability, while only requiring a bound on the fourth moment of the entries of A. The sample complexity is of the order $\mathcal{O}(s \log(n \log(\frac{1}{\epsilon})) \log(\frac{1}{\epsilon}))$.

I. Introduction

Motivated by various applications in signal processing and the publications of Candès, Romberg, Tao, and Donoho [1], [2], a variety of research in the field of compressed sensing, targeting the recovery of a sparse signal from a small number of measurements, has been established.

In the following, it is assumed that the measurements are of the form

$$y_j = a^{(j)*}x \qquad \longleftrightarrow \qquad y = Ax,$$

where $x \in \mathbb{C}^n$ denotes the s-sparse signal, $y \in \mathbb{C}^m$ the measurement vector with $m \ll n$, and $A \in \mathbb{C}^{m \times n}$ a random measurement matrix with $E[A^*Ax] = x$.
Besides algorithms for solving the initially proposed basis pursuit ($\min \|\hat{x}\|_1$ s.t. $A\hat{x} = y$), more efficient greedy algorithms as the Orthogonal Matching Pursuit (OMP)[3], Compressive Sampling Matching Pursuit (CoSaMP) [4] or Iterative hard thresholding [5] have been established. However, those methods are based on a strong concentration of A^*Ax around x, namely, the Restricted Isometry Property (RIP).
Requiring a strong concentration of A^*Ax is equivalent to requiring a strong concentration of the sample mean of m iid. random variables $X^{(j)} := ma^{(j)}a^{(j)*}x$ around their mean $E[X^{(j)}] = x$:

$$AA^*x = \sum_{j=1}^{m} a^{(j)}a^{(j)*}x =: \frac{1}{m}\sum_{j=1}^{m} X^{(j)} =: \bar{X}$$

For Gaussian measurement matrices (and other well-concentrated distributions), comparably sharp tail bounds for $|\bar{X}_i - x_i|$ exist. However, this is a major challenge for heavy-tailed distributions or in scenarios with only limited knowledge about the underlying distribution.
In [6], the authors pointed out the difficulties for an RIP-based analysis for matrices with weak concentration and instead established a new, l_1-specific technique to obtain recovery guarantees for the basis pursuit covering heavy-tailed matrices. However, their theory is not applicable for greedy algorithms. To the best of our knowledge, there are no successful recovery guarantees for greedy algorithms for heavy-tailed matrices.

II. Median-of-Means

By definition, heavy-tailed distributions have a significantly higher probability for outliers, which negatively affects the sample mean and, as a consequence, prevents successful recovery guarantees for greedy algorithms. For that reason, we suggest replacing the inadequately concentrating sample mean by a more robust median-of-means estimator $\hat{\mu}$.
For computing $\hat{\mu}$, the m measurements have to be split into K subsets of size J. In the next step, the sample mean $\bar{X}^{(k)}$ of every subset has to be computed.

$$\underbrace{\begin{bmatrix} X^{(1,1)} \\ \vdots \\ X^{(J,1)} \end{bmatrix}}_{\Rightarrow \bar{X}^{(1)}} \quad \underbrace{\begin{bmatrix} X^{(1,2)} \\ \vdots \\ X^{(J,2)} \end{bmatrix}}_{\Rightarrow \bar{X}^{(2)}} \quad \cdots \quad \underbrace{\begin{bmatrix} X^{(1,K)} \\ \vdots \\ X^{(J,K)} \end{bmatrix}}_{\Rightarrow \bar{X}^{(K)}}$$

By taking the entry-wise median (in the complex case separately for the real and imaginary part) over all sample means $\bar{X}^{(1)}, \ldots, \bar{X}^{(K)}$, the median-of-means estimator $\hat{\mu}$ is obtained. As the median is very robust against outliers, $\hat{\mu}$ even exhibits an exponential concentration in K:

Lemma 1. *Assume a random variable X_i has mean $E[X_i] = x_i$ and variance $Var[X_i] \leq \sigma^2 \|x\|_2^2 < \infty$. Then, the median-of-means estimator $\hat{\mu}_i$, defined as*

$$\hat{\mu}_i = \textit{median}\{\bar{X}_i^{(1)}, \ldots, \bar{X}_i^{(K)}\} \quad \textit{with } \bar{X}_i^{(k)} = \frac{1}{J}\sum_{j=1}^{J} X_i^{(j,k)},$$

fulfills

$$P(|\hat{\mu}_i - x_i| \geq \gamma) \leq 2e^{-K/2}$$

if $J \geq \frac{2e^2\sigma^2\|x\|_2^2}{\gamma^2}$.

Proof. The proof follows the proof idea of [7, Theorem 5] with appropriate adaptations.

– **Case I: x, y, A are real**:
By assumption,

$$Var[\bar{X}_i] = \frac{1}{J^2}\sum_{j=1}^{J} Var[X_i^{(j)}] \leq \frac{\sigma^2\|x\|_2^2}{J}$$

By applying Chebyshev's inequality, the following tail bound is obtained

$$p_J := P(|\bar{X}_i - x_i| \geq \gamma) \leq \frac{Var[\bar{X}_i]}{\gamma^2} \leq \frac{\sigma^2\|x\|_2^2}{J\gamma^2}.$$

For every $k \in [K]$, one can define the Bernoulli random variable $I^{(k)} := \mathbf{1}\{|\bar{X}_i^{(k)} - x| \geq \gamma\}$ with parameter p_J. By the definition of the median, $|\hat{\mu}_i - x_i| \geq \gamma$ can only be fulfilled if either at least half of the $\bar{X}_i^{(k)}$ are larger than $x_i + \gamma$ or at least half of them are smaller than $x_i - \gamma$. Therefore,

$$P\left(|\hat{\mu}_i - x_i| \geq \gamma\right) \leq P(\sum_{k=1}^{K} I^{(k)} \geq \frac{K}{2})$$

Applying the multiplicative Chernoff bound, yields

$$P\left(\sum_{k=1}^{K} I^{(k)} \geq (1+\lambda)Kp\right) \leq \left(\frac{e^{\lambda}}{(1+\lambda)^{1+\lambda}}\right)^{Kp}$$
$$= e^{-Kp}\left(\frac{e}{1+\lambda}\right)^{(1+\lambda)Kp}, \qquad \lambda > 0.$$

By assumption, $J \geq \frac{2e^2\sigma^2\|x\|_2^2}{\gamma^2}$ and therefore $p_J \leq \frac{1}{2e^2}$. Choosing $(1+\lambda)Kp = K/2$ concludes the proof for the real case:

$$P(|\hat{\mu}_i - x_i| \geq \gamma) \leq e^{-Kp_J}(2ep_J)^{K/2} \leq (2ep_J)^{K/2}$$
$$\leq \left(\frac{2e\sigma^2\|x\|_2^2}{J\gamma^2}\right)^{K/2} \leq e^{-K/2}.$$

– **Case II: x, y, A are complex**:
Denote by $\Re(x_i)$ the real part and by $\Im(x_i)$ the imaginary part of x_i. By our definition, the median over a complex set has to be taken separately for the real part and imaginary part of its elements. Therefore, $\hat{\mu}_i =: \Re(\hat{\mu}_i) + i\Im(\hat{\mu}_i)$, where $\Re(\hat{\mu}_i) = \text{median}\{\Re(\bar{X}_i^{(1)}), \ldots, \Re(\bar{X}_i^{(K)})\}$ (resp. for $\Im(\hat{\mu}_i)$).
By triangle inequality and union bound,

$$\begin{aligned} &P(|\hat{\mu}_i - x_i| \geq \gamma) \\ &\leq P(|\Re(\hat{\mu}_i) - \Re(x_i)| + |\Im(\hat{\mu}_i) - \Im(x_i)| \geq \gamma) \\ &\leq P(|\Re(\hat{\mu}_i) - \Re(x_i)| \geq \gamma) + P(|\Im(\hat{\mu}_i) - \Im(x_i)| \geq \gamma) \\ &\leq 2e^{-K/2} \end{aligned}$$

As $P(|\Re(\bar{X}_i - x_i)| \geq \gamma) \leq P(|\bar{X}_i - x_i| \geq \gamma)$ (resp. for $\Im$), the bound for the real case above holds for both summands in the second line separately, which concludes the proof. □

Corollary 2. *Assume $X_1, \ldots, X_n$ are random variables with mean $E[X_i] = x_i$ and variance $Var[X_i] \leq \sigma^2 < \infty$ for all $i \in [n]$. Then,*

$$P(\|\hat{\mu} - x\|_\infty \geq \gamma) \leq \eta$$

for

$$J \geq \frac{2e^2\sigma^2\|x\|_2^2}{\gamma^2} \qquad K \geq 2\log(\frac{2n}{\eta})$$

Proof. The theorem follows directly from Lemma 1 by choosing K such $2e^{-K/2} \leq \frac{\eta}{n}$ and applying a union bound over all $i \in [n]$. □

Algorithm 1 Approximation from random measurements via median-of-means

Input: Measurement matrix $A \in \mathbb{C}^{m\times n}$ and vector of measurements $y \in \mathbb{C}^m$ with $m = JK$;
Output: An approximation $\hat{\mu}$ of the s-sparse signal $x \in \mathbb{C}^n$ fulfilling $\|\hat{\mu} - x\|_\infty < \gamma$ with high probability.

function MoM($y; A; J; K$)
  Split A in matrices $A^{(k)} \in \mathbb{C}^{J\times n}$ and y in corresponding vectors $y^{(k)} \in \mathbb{C}^J$ $\forall k \in [K]$.
  Set $x^{(0)} = 0$
  for $k = 1$ to K **do**
    Compute $\bar{X}^{(k)} = \frac{m}{J}A^{(k,l)*}(y^{(k,l)} - A^{(k,l)}x^{(l-1)})$
  return median$\{\bar{X}^{(1)}, \ldots, \bar{X}^{(K)}\}$

Theorem 3. *Let $x \in \mathbb{C}^n$ be a signal, $A \in \mathbb{C}^{m\times n}$ a random measurement matrix with centered iid. entries with moments*

$$E[|a_i^{(j)}|^2] = \frac{1}{m} \quad \text{and} \quad E[|a_i^{(j)}|^4] \leq \frac{\sigma^2}{m^2} < \infty,$$

and $y := (Ax) \in \mathbb{C}^m$ the corresponding measurement vector. Then, the output $\hat{\mu}$ of Algorithm 1 with J and K as in Corollary 2 and $m \geq JK$ fulfills

$$\|\hat{x} - x\|_\infty < \gamma$$

with probability $1 - \eta$.

Proof. This follows directly from Corollary 2 as

$$E[X_i^{(j)}] = mE[e_i^* a^{(j)} a^{(j)*} x]$$
$$= \underbrace{mE[|a_i^{(j)}|^2]x_i}_{=x_i} + m\underbrace{E[a_i^{(j)}]}_{=0}E[\textstyle\sum_{l\neq i}\bar{a}_l^{(j)}x_l] = x_i$$

and

$$Var[X_i^{(j)}] = E[|X_i^{(j)}|^2]$$
$$= \underbrace{m^2E[|a_i^{(j)}|^4]|x_i|^2}_{\leq\sigma^2|x_i|^2} + \underbrace{mE[|a_i^{(j)}|^2]}_{=1}\underbrace{mE[|\textstyle\sum_{l\neq i}\bar{a}_l^{(j)}x_l|^2]}_{=m\sum_{l\neq i}E[|\bar{a}_l^{(j)}|^2]|x_l|^2}$$
$$\leq \sigma^2|x_i|^2 + \textstyle\sum_{l\neq i}|x_l|^2 \leq \sigma^2\|x\|_2^2$$

Applying Jensen's inequality, $E[|a_i^{(j)}|^4] \geq E[|a_i^{(j)}|^2]^2 \Rightarrow \sigma^2 \geq 1$ which explains the last inequality. □

Remark 4. *Even if the original measurement matrix does not fulfill $E[|a_i^{(j)}|^2] = \frac{1}{m}$, matrix and measurements can be scaled to fulfill the corresponding requirement of Theorem 3 as long as the fourth moment is bounded and the second moment is known.*

III. Iterative Median-of-Means Algorithm

As demonstrated in Theorem 3, the approximation $\hat{\mu}(x)$ obtained by Algorithm 1 fulfills $\|\hat{\mu}(x) - x\|_\infty < \gamma$ with high probability. While the l_∞-bound can be used to identify large entries of x, the naive l_2-bound exhibits an undesirable scaling in n:

$$\|\hat{\mu}(x) - x\|_2 = \sqrt{\sum_{i=1}^{n} |\hat{\mu}_i(x) - x_i|^2} \leq \sqrt{n}\|\hat{\mu}(x) - x\|_\infty \leq \sqrt{n}\gamma$$

The scaling in n can be reduced to a scaling in s by applying an entry-wise hard-thresholding operator

$$h_\gamma(\hat{\mu})_i := h_\gamma(\hat{\mu}_i) := \begin{cases} \hat{\mu}_i & \text{for } |\hat{\mu}_i| \geq \gamma \\ 0 & \text{for } |\hat{\mu}_i| < \gamma \end{cases} \qquad \forall i \in [n].$$

In Figure 1, the possible intervals of the deviations of $\hat{\mu}(x)$ from x are visualized. As $\|\hat{\mu}(x) - x\|_\infty < \gamma$, $x_i = 0$ implies $h_\gamma(\hat{\mu}_i(x)) = 0$, and, further,

$$\text{supp}(h_\gamma(\hat{\mu}(x))) \subseteq \text{supp}(x). \quad (1)$$

No effect can be seen for $|x_i| > 2\gamma$ which implies $|\hat{\mu}_i(x)| > \gamma$, and therefore, $h_\gamma(\hat{\mu}_i(x)) = \hat{\mu}_i(x)$.

While the last two properties are beneficial, applying the thresholding operator can increase the l_∞-error for $|x_i| \in [\gamma, 2\gamma)$ with $|\hat{\mu}_i(x)| < \gamma$, and therefore, $h_\gamma(\hat{\mu}_i(x)) = 0$ which doubles the l_∞ bound $|h_\gamma(\hat{\mu}_i(x)) - x_i| < 2\gamma$.

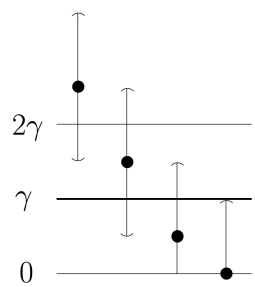

Figure 1. Visualization of the possible deviation of $|\hat{\mu}|$ from $|x|$ for a precision of γ. The black dots symbolize the entries of x, while the entries of $\hat{\mu}_i(x)$ lie within the open intervals.

Combined, this leads to the following l_2-bound

$$\|h_\gamma(\hat{\mu}(x)) - x\|_2 = \sqrt{\sum_{i \in \text{supp}(h_\gamma(\hat{\mu}(x)))} |h_\gamma(\hat{\mu}_i(x)) - x_i|^2} \leq \sqrt{s}\|h_\gamma(\hat{\mu}(x)) - x\|_\infty \leq \sqrt{s}2\gamma \quad (2)$$

Due to the strong scaling of $m \in \mathcal{O}(\frac{1}{\gamma^2})$ in γ, a small l_2-norm can only be achieved by a large increase of the number of measurements.

Instead, an iterative procedure will be defined which allows for an increasing precision while keeping J constant. For simplicity, assume that x has unit norm (i.e., $\|x\|_2 = 1$). Setting $x^{(1)} = h_\gamma(\hat{\mu}(x))$, in the second iteration not x but $x - x^{(1)}$ has to be recovered. By Eq. 1, the sparsity is still bounded by s, while, by Eq. 2, the l_2-norm is bounded by $\sqrt{s}2\gamma$.

So, in order to obtain an approximation $\hat{\mu}(x - x^{(1)})$ with a precision of $\alpha\gamma$ (for an $\alpha \in (0,1)$) while keeping J constant, the following inequality has to be fulfilled:

$$J \geq \frac{\overbrace{\|x\|_2^2}^{=1}}{\gamma^2} \overset{!}{\geq} \frac{\overbrace{\|x - x^{(1)}\|_2^2}^{s(2\gamma)^2 \geq}}{(\alpha\gamma)^2} \qquad \Rightarrow \gamma \leq \alpha \frac{1}{2\sqrt{s}}$$

Therefore, set $\gamma := \alpha\frac{1}{2\sqrt{s}}$ (the choice of α will be discussed in Remark 6).

A last issue has to be addressed: Theorem 3 assumes a fixed x which is independent of A. $\hat{\mu}(x)$, and consequently, $x - x^{(1)}$ do not fulfill the independence on A. Therefore, A and y have to be partitioned into L blocks, where L denotes the number of iterations. Due to this, the current approximation and the next block will always be independent. Defining $x^{(l)} = x^{(l-1)} + h_{\alpha^{l-1}\gamma}(\hat{\mu}(x - x^{(l-1)}))$ recursively, leads to the iterative Algorithm 2 and the main result, Theorem 5.

Algorithm 2 Approximation from random measurements via iterative median-of-means

Input: Measurement matrix $A \in \mathbb{C}^{m \times n}$ and vector of measurements $y \in \mathbb{C}^m$ with $m = JKL$; $\alpha \in (0,1)$.
Output: An approximation $\hat{x}$ of the s-sparse signal $x \in \mathbb{C}^n$ fulfilling $\|x - \hat{x}\|_2 \leq \alpha^L \|x\|_2$.

function Iterative-MoM($y; A; N; K; L; \alpha$)
  Split A in matrices $A^{(k,l)} \in \mathbb{C}^{J \times n}$ and y in corresponding vectors $y^{(k,l)} \in \mathbb{C}^J$ $\forall k \in [K], l \in [L]$.
  Set $x^{(0)} = 0$
  for $l = 1$ to L **do**
    for $k = 1$ to K **do**
      Compute $\bar{X}^{(k)} = A^{(k,l)*}(y^{(k,l)} - A^{(k,l)}x^{(l-1)})$
    $\hat{\mu} = \text{median}\{\bar{X}^{(1)}, \ldots, \bar{X}^{(K)}\}$
    $x^{(l)} = x^{(l-1)} + h_{\alpha^l \frac{\|x\|_2}{2\sqrt{s}}}(\hat{\mu})$
  return $x^{(L)}$

Theorem 5. *Let $x \in \mathbb{C}^n$ be s-sparse signal with unit norm, $A \in \mathbb{C}^{m \times n}$ a random measurement matrix with centered iid. entries with moments*

$$E[|a_i^{(j)}|^2] = \frac{1}{m} \quad \text{and} \quad E[|a_i^{(j)}|^4] \leq \frac{\sigma^2}{m^2} < \infty,$$

and $y := (Ax) \in \mathbb{C}^m$ the corresponding measurement vector. Then, the output $\hat{x}$ of Algorithm 2

$$\|\hat{x} - x\|_2 < \epsilon$$

with probability $1 - \eta$ if $m \geq JKL$ and

$$J \geq \mathbf{s}\frac{8e^2\sigma^2}{\alpha^2} \quad K \geq 2\log(\mathbf{n}\frac{2L}{\eta}) \quad L \geq \frac{\log(\epsilon)}{\log(\alpha)} \quad \alpha \in (0,1)$$

Proof. By induction,

$$\text{supp}(x - x^{(l)}) \subseteq \cdots \subseteq \text{supp}(x - x^{(1)}) \subseteq \text{supp}(x)$$
$$\|x - x^{(l)}\|_2 \leq \sqrt{s}2(\alpha^{l-1}\gamma) = \alpha^l.$$

The choice of L guarantees $\alpha^L \leq \epsilon$, while the slight adaptation of K is the result of a union bound over all L iterations. The proof follows now directly from Theorem 3 as $x - x^{(l-1)}$ is independent of $A^{(l)}$ and $y^{(l)}$ for all $l \in [L]$, due to the partitioning of A and y. □

Remark 6. *As mentioned before, a decrease of γ strongly increases J, and consequently, the number of measurements. As $\gamma := \alpha\frac{1}{2\sqrt{s}}$, α can not be chosen too small. On the other hand, an α close to 1 increases L, the number of iterations. Minimizing over the product KL, $\alpha = \frac{1}{\sqrt{e}}$ is obtained.*

IV. Numerical Analysis

In the following, the numerical performance of Algorithm 2 will be analyzed.
The entries of A are chosen to be iid. student-t distributed with 5 degrees of freedom and, then, scaled to fulfill the requirements of Theorem 2, which leads to $\sigma^2 = 9$.

For a dimension of $n = 2000$ and sparsity $s = 10$, the constants in Theorem 2, still dominate the required number of measurements. Instead, we chose $J = 160$ and $K = 7$. The parameter α is set to $\frac{1}{\sqrt{e}}$ as suggested by Remark 6. The sparse vector x is chosen to have unit norm with s linearly increasing entries (from ca. 0.05 to 0.5) on random positions.

Then, Algorithm 2 has been performed 10 times for different matrices A and the worst result for every step has been plotted in Figure IV. Despite the significantly lower values for J and K, the l_2-error $\|x^{(l)} - x\|_2$ of the iterates of our algorithm (blue) stayed always well below the theoretical bound α^l (red).

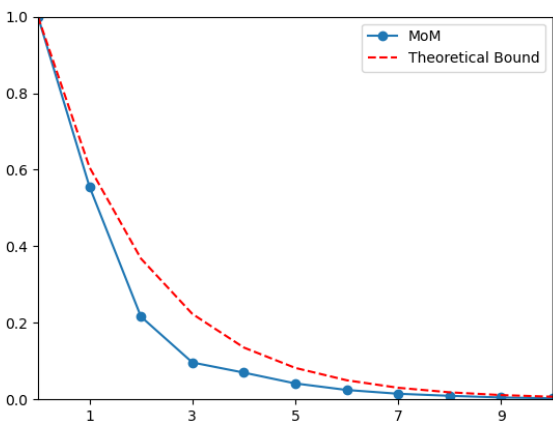

Figure 2. Comparison between $\|x^{(l)} - x\|_2$ and its theoretical bound α^l for $x^{(l)}$ obtained by Algorithm 2.
(max. l_2-error of every iteration for 10 different matrices A)

In Figure IV, the performance of two modifications of Algorithm 2 can be seen. For a good comparison, each of the three methods is applied to the same A and y. As explained in the last chapter, A and y have to be partitioned to guarantee the necessary independence between underlying signal and measurement matrix. If only one of those samples is used for every iteration, the algorithm appears way more unstable and often fails (orange), which indicates that the required independence is not only a proof aritfact. Nevertheless, without partitioning, a larger number of measurements could be used for every iteration - a trade-off which remains subject of further research.

The second modification targets an underlying weakness of the median-of-means estimator. Different to mean or median, the median-of-means of a data set potentially changes when changing the order of the samples. If all outliers end up in only few of the K subsets, the median-of-means will most likely not be affected by them. If they are distributed over all K subsets and affect all means, the median-of-means should be affected as well. To compensate for this, the median-of-means is not only computed for one single ordering of the measurements. Instead, the median-of-means has to be computed again for multiple different permutations of the measurements. The 'improved' estimate is then obtained by taking the median over those median-of-means. While even further improvements can be expected for a larger number of permutations, we restricted ourselves to only 20 random permutations for performance reasons.
As indicated by the graph, this leads to a significant increase in performance in the first iteration. The performance of further steps might be restricted by the slow decrease of the threshold (i.e., x might already be recovered with high precision, but smaller values of the support of x are still set to 0 by the high threshold). This idea is part of ongoing research.

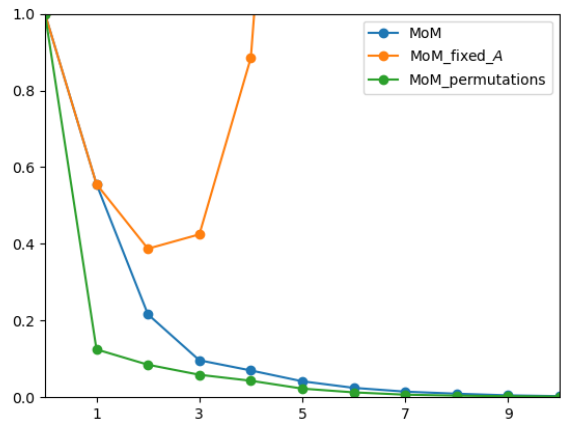

Figure 3. Comparison between the l_2-error of the iterates of Algorithm 2 and two modifications of the algorithm. For the orange results, $A^{(l)}$ is fixed for all iterations. For the green line, in every iteration, the median over 20 median-of-means estimates for different permutations of the measurements is taken.
(max. l_2-error of every iteration for 10 different matrices A)

V. Conclusion

The greedy algorithm presented in this work, reliably approximates a s-sparse vector from random measurements while requiring a comparably small number of measurements. The big advantage of the presented method – besides the efficient implementation – is the lack of strong concentration requirements on the measurement matrix. As long as the fourth moment can be bounded, our algorithm will provably work for any centered measurement matrix A.

Furthermore, an additional performance increase of a modified median-of-means estimator has been demonstrated empirically in the last chapter.

As listed in the introduction, there is a variety of greedy algorithms for recovering sparse signals which are based on the concentration of A^*Ax. We are convinced that the algorithm presented here is only one example where replacing the sample mean by the median-of-means is beneficial and suggest further research for different, more involved recovery algorithms.

REFERENCES

[1] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 8, pp. 1207–1223, 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[4] D. Needell and J. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[5] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[6] S. Dirksen, G. Lecué, and H. Rauhut, "On the gap between restricted isometry properties and sparse recovery conditions," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5478–5487, 2016.

[7] T. Fuchs, D. Gross, F. Krahmer, R. Kueng, and D. Mixon, "Sketching with kerdock's crayons: Fast sparsifying transforms for arbitrary linear maps," *SIAM Journal on Matrix Analysis and Applications*, vol. 43, no. 2, pp. 939–952, 2022.