Submatrices with non-uniformly selected random supports and insights into sparse approximation

Simon Ruetz SIMON.RUETZ@UIBK.AC.AT
Karin Schnass KARIN.SCHNASS@UIBK.AC.AT
University of Innsbruck
Technikerstraße 13
6020 Innsbruck, Austria

Abstract

In this paper we derive tail bounds on the norms of random submatrices with non-uniformly distributed supports. We apply these results to sparse approximation and conduct an analysis of the average case performance of thresholding, Orthogonal Matching Pursuit and Basis Pursuit. As an application of these results we characterise sensing dictionaries to improve average performance in the non-uniform case and test their performance numerically.

Keywords: random submatrices, non-uniform sampling, matrix Chernoff, sparse approximation

1. Introduction

Motivation: In sparse approximation, the goal is to find a sparse solution to an underdetermined system of linear equations. A signal $y \in \mathbb{R}^d$ is assumed to be a linear combination of a small number $S \ll d$ of elements ϕ_i, called atoms, out of a larger set, called the dictionary. Denoting the dictionary by $\Phi = (\phi_1, \ldots, \phi_K) \in \mathbb{R}^{d\times K}$ and by Φ_I the restriction to the columns indexed by the set I, called the support, one assumes that

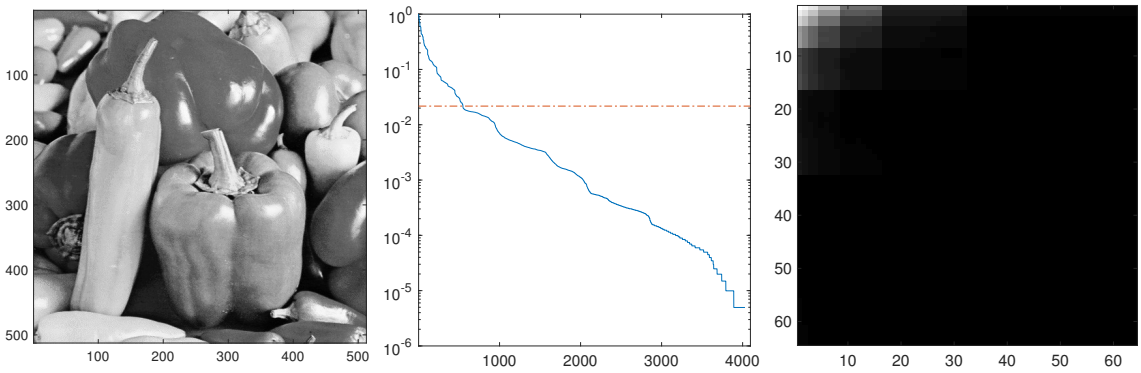
$$y \approx \sum_{k\in I} \phi_k x_k = \Phi_I x_I \quad \text{s.t.} \quad |I| = S.$$

The sparse approximation problem amounts to finding the vector x and its support I given the dictionary Φ and signal y. In general, this is a NP-hard optimisation problem, hence sparse approximation algorithms such as thresholding, Orthogonal Matching Pursuit (OMP) and Basis Pursuit (BP) were proposed. It turns out that in order to prove support recovery guarantees for these algorithms, information about the extreme singular values of Φ_I is needed.

Let $\|\cdot\|_{2,2}$ denote the operator norm and $\mathbb{I}$ the identity matrix. Deterministic methods to bound $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2}$ for *arbitrary* supports I are of limited use since the restrictions on the dictionary Φ are too stringent. This started the study of random collections of columns of the dictionary Φ. In [22] it was first shown that under rather mild conditions on the dictionary Φ, *most* subdictionaries Φ_I are close to an isometry - i.e. $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} \leq \vartheta_0 < 1$, with later improvements in [8]. So far, all available results on the conditioning of random subdictionaries rely on the supports I to be drawn from the uniform distribution. Unfortunately this assumption is rarely satisfied for practically relevant signal classes, where some atoms

of the underlying dictionary are usually more likely to appear in a sparse representation than others.

To demonstrate this non-homogeneity, we conduct the following small experiment. We take the 2D Haar-Wavelet decomposition of all normalised 64×64 patches from the image *Peppers* and apply a threshold[1] of $\sqrt{\log(d)/d}/6$ for $d = 64^2$ to the coefficients to get sparse approximations. We then count how often each atom has a non-zero coefficient to get a proxy for its inclusion probability in a sparse support I. Figure 1 shows the relative frequency of each element of the 2D Haar-Wavelet basis. It comes as no surprise that



(a) (b) (c)

Figure 1: (a) Original image from which the patches are extracted. (b) Relative frequency of wavelet coefficients above threshold (blue) - average frequency (red) on a log scale. (c) Locations of non-zeros coefficients in the 2D Haar-Wavelet basis - the higher the row or column index the smaller the corresponding wavelet

low frequency (large) wavelets are much more likely to appear in the sparse supports than high frequency (small) wavelets. So the supports of the sparse signals exhibit a non-uniform structure which previous results on the conditioning of random subdictionaries do not cover. We try to close this gap by defining two non-uniform support distributions and deriving tail bounds on the norms of the resulting random submatrices. This allows us to derive recovery guarantees of the sparse supports for a larger class of practically relevant signals.

Prior work: As mentioned above, Tropp [22] and Chrétien and Darses [8] derived concentration inequalities for the operator norm of random submatrices with uniformly distributed supports. These results were applied to BP showing that BP recovers the correct support and coefficients under rather mild conditions on the dictionary [23]. For OMP, similar results were developed in [18], whereas for thresholding average case results appeared in [19]. In [15] the dictionary D is assumed to be a concatenation of two dictionaries ϕ and ψ, i.e. $D = (\phi, \psi)$. There a concentration inequality on the extreme singular values of submatrices that consist of a *fixed* set of columns with cardinality n_a of the first dictionary and a *random* set of columns n_b of the second dictionary are derived. This allows to model signals where some atoms are known to be in the support while some others are picked uniformly at random.

1. The threshold is inspired by the expected size of the largest inner product of a wavelet with noise drawn uniformly at random from the unit sphere.

The idea of using the structure of sparse signals to improve recovery of the sparse coefficients can also be found in the field of compressed sensing. The aim in compressed sensing is to recover a sparse signal $y \in \mathbb{R}^d$ from an incomplete set of linear measurements $z = Ay$, where $A \in \mathbb{R}^{m \times d}$ and $m \ll d$, [6, 10]. The signal y is assumed to be sparse or compressible in some (orthonormal) basis or frame Φ, i.e. $y = \Phi x$ for a sparse coefficient vector x.
From a theoretical point of view the best measurement matrices A, achieving the smallest m for a given sparsity level S, are random matrices. Unfortunately in many practical applications it is not possible or efficient to use random matrices, since they cannot be realised by the underlying physical measurement process, such as in compressed magnet resonance imaging (MRI). Instead one is given an (often orthonormal) measurement matrix $\Psi \in \mathbb{R}^{d \times d}$ and has to find a *subsampling pattern* $\Omega \subseteq \{1, ..., K\}$ which selects m rows of Ψ, so that for $A = P_\Omega \Psi$ the signal y resp. the coefficients x can be reliably reconstructed from $z = Ay = A\Phi x = \bar{A}x$.
As in sparse approximation, rather strong assumptions on the matrix $A\Phi = \bar{A}$ are needed in order to guarantee recovery for all sparse x. In [5] the elements of Ω were assumed to be chosen uniformly at random in order to employ probabilistic arguments to derive sufficient conditions for recovery for relatively small m. Over the years, various different subsampling strategies - most of them highly non-uniform - were proposed (see for example [3, 7, 16, 1, 14]). Underlying the success of these variable density sampling strategies is the highly non-uniform structure of the sparse supports. So it was shown that previous lower bounds on the size of m are too pessimistic and performance can be improved if the subsampling pattern takes the support structure of the sparse signals into account [1, 14].
Contribution: We derive tail bounds on the operator norm of non-uniformly chosen submatrices. The supports are assumed to follow either a Poisson sampling model or a rejective sampling model thus allowing us to model a large class of non-uniform distributions. Our results rely on a generalisation of a Theorem by Chrétien and Darses [8]. The main tool to handle non-uniformly distributed S-sparse supports is a kind of Poissonisation argument where we provide a generalised version of Lemma 4.1 of [12]. We apply these results to derive sufficient conditions for sparse approximation to work with high probability for thresholding, OMP and BP. In the CS setup this analysis provides a criterion to decide between two possible measurement matrices A_1 and A_2 depending on the frequency of the basis elements. Further, if there is no design freedom for the dictionary or CS matrix, we show how to incorporate this prior information about the coefficient distribution into the algorithms using the ideas of preconditioning and sensing dictionaries.
Organisation: Section 2 collects our notations and defines the setting we work in. In Section 3 we state our results on norms of non-uniformly distributed random submatrices and apply those concentration inequalities to sparse approximation in Section 4. Finally we incorporate this knowledge in the construction of special sensing dictionaries in Section 5 and show how they improve performance.

2. Notation and setting

A quick note on the notation used throughout this text. Let $A \in \mathbb{R}^{d \times K}$ and $B \in \mathbb{R}^{K \times m}$. By A_k and A^k we denote the k-th column and k-th row of A respectively and by A^* the transpose of the matrix A. For $1 \leq p, q, r \leq \infty$ we set $\|A\|_{p,q} := \max_{\|x\|_q = 1} \|Ax\|_p$. Recall

that $\|AB\|_{p,q} \leq \|A\|_{q,r}\|B\|_{r,p}$ and $\|Ax\|_q \leq \|A\|_{q,p}\|x\|_p$. Frequently encountered quantities are

$$\|A\|_{\infty,2} = \max_{k\in\{1,\dots,d\}} \|A^k\|_2 \quad \text{and} \quad \|A\|_{2,1} = \max_{k\in\{1,\dots,K\}} \|A_k\|_2,$$

denoting the maximum ℓ_2-norm of a row and the maximum ℓ_2-norm of a column of A respectively. Note that $\|A\|_{\infty,2} = \|A^*\|_{2,1}$. Further note that $\|A\|_{\infty,1}$ simply is the maximum absolute entry of the matrix A. For ease of notation we sometimes write $\|A\| = \|A\|_{2,2}$ for the operator norm - corresponding to the largest absolute singular value of A. For a vector $v \in \mathbb{R}^d$, we denote by $\|v\|_{\min} := \min_i |v_i|$ the smallest absolute value of v and $\|v\|_{\max} := \|v\|_\infty$ the maximal absolute value of v. For a subset $I \subseteq \mathbb{K} := \{1,\dots,K\}$, called the support, we denote by $A_I \in \mathbb{R}^{d\times S}$ the submatrix with columns indexed by I and by $A_{I,I} \in \mathbb{R}^{S\times S}$ the submatrix with columns and rows indexed by I. We denote by $A_I^\dagger$ the Moore-Penrose pseudo inverse of the matrix A_I and by $P(A_J) := A_J A_J^\dagger$ the projection onto the column span of A_J. As was noted in the introduction we want the supports to follow a non-uniform distribution, allowing some columns, called atoms, to be picked more frequently than others. We are going to use the following two sampling models which define two probability measures on $\mathcal{P}(\mathbb{K})$ that allow us to model non-uniform distributions for our supports.

Definition 1 (Poisson sampling) *Let δ_j denote a sequence of K independent Bernoulli 0-1 random variables with expectation p_j such that $\sum_{j=1}^K p_j = S$. We say the supports I follow the Poisson sampling model, if*

$$I := \{i \mid \delta_i = 1\}.$$

Each support $I \subseteq \mathbb{K}$ is chosen with probability

$$\mathbb{P}(I) = \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j). \tag{1}$$

Supports following a Poisson sampling model have (by definition of the Bernoulli r.v.) cardinality S *on average*. This comes with the big advantage that the probability of one atom appearing in the support is independent of the others, allowing us to make use of concentration inequalities for sums of independent random matrices later on. The drawback of this model is that the supports are not exactly S sparse. This can be achieved by keeping only those supports that have cardinality S and *throwing away* the rest. This amounts to simply conditioning the above Poisson sampling model on the event that exactly S of the Bernoulli r.v. are equal to 1, leading to our second support distribution model.

Definition 2 (Rejective sampling) *Let δ_j denote a sequence of K independent Bernoulli 0-1 random variables with expectation p_j such that $\sum_{j=1}^K p_j = S$ and denote by $\mathbb{P}$ the probability measure of the corresponding Poisson sampling model. We say our supports follow the rejective sampling model, if each support $I \subseteq \mathbb{K}$ is chosen with probability*

$$\mathbb{P}_S(I) := \mathbb{P}(I \mid |I| = S) = \begin{cases} c\prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) & \text{if} \quad |I| = S \\ 0 & \text{else} \end{cases}, \tag{2}$$

where c is a constant to ensure that $\mathbb{P}_S$ is a probability measure.

The distributions of the supports in the above two sampling models are uniquely defined by the expectations of the Bernoulli random variables. For more information on Poisson and rejective sampling, we refer the interested reader to [12]. We call the square diagonal matrix $W := \mathrm{diag}((\sqrt{p_k})_k)$ the weight matrix. Let R be the square diagonal *selector matrix* whose diagonal entries are the δ_j, i.e. $R = \mathrm{diag}((\delta_k)_k)$ and denote by R' an independent copy of R. Further let $\vec{A}_{jk} = A_{jk} e_j \otimes e_k$ be the matrix with only non-zero entry A_{jk}. This allows us to write

$$RAR = \sum_{i,j} \delta_i \delta_j \vec{A}_{ij}.$$

Note that by properties of the operator norm, the two random variables $\|A_I\|$ and $\|AR\|$ have the same distribution.

3. Main results

We now present our main results on submatrices whose support is sampled from a non-uniform distribution. We begin by stating the concentration inequality for the operator norm of non-uniformly picked random submatrices, before turning to some special cases arising in sparse approximation. Then we state a concentration inequality for the maximal row norm of random column-submatrices. Lastly we state and proof a kind of Poissonisation argument - of independent interest - which is key for our proofs. Note that we state our results only for the rejective sampling model, but they hold for the Poisson sampling model as well - see Remark 8.

3.1 Operator norm of random submatrices

The aim is to get a tail bound for the random variable $\|H_{I,I}\|_{2,2}$, where I is distributed according to the models introduced above and H is a matrix with zero diagonal. As expected, the result shows how the more frequently picked entries have a higher impact on the operator norm than less important ones.

Theorem 3 *Let $H \in \mathbb{R}^{K\times K}$ be a matrix with zero diagonal and assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $r \geq 2e^2 \|WHW\|_{2,2}$*

$$\mathbb{P}_S\bigg(\|H_{I,I}\|_{2,2} \geq r \bigg) \leq 216K \exp\left(- \min\left\{ \frac{r^2}{4e^2\|HW\|_{\infty,2}^2}, \frac{r^2}{4e^2\|WH\|_{2,1}^2}, \frac{r}{2\|H\|_{\infty,1}} \right\} \right).$$

Proof [Outline] We follow the proof that appeared in Chrétien and Darses [8] with some minor changes to account for the non-uniformly distributed supports and the extension to non-symmetric matrices. Their proof consists of roughly three steps. First they bound the failure probability of the rejective sampling model by the independent Poisson sampling model

$$\mathbb{P}_S\left(\|RHR\|_{2,2} \geq r\right) \leq 2\mathbb{P}\left(\|RHR\|_{2,2} \geq r\right).$$

Then they use a decoupling argument to make the selection of rows and columns independent, i.e.

$$\mathbb{P}\left(\|RHR\|_{2,2} \geq r\right) \leq 72\mathbb{P}\left(\|RHR'\|_{2,2} \geq r/2\right),$$

where R' is an independent copy of R. Then they apply the matrix Chernoff inequality three times to finish the proof. Our proof in the non-uniform, non-symmetric case follows the above outline very closely. The main difficulty lies in bounding the rejective model by the Poisson model, which is why we had to provide Lemma 7. The second and third steps are straightforward extensions of their argument. For the sake of completeness we provide a detailed proof in the appendix. ■

3.1.1 Special cases - hollow (cross)-Gram matrices

In this subsection we look at the special case $H = \Phi^*\Phi - \mathbb{I}$ that appears naturally in the sparse approximation framework. Previous results showed that success of recovery depends on the coherence $\mu := \max_{i \neq j} |\langle \phi_i, \phi_j \rangle|$ and the conditioning of the subdictionary Φ_I, i.e.

$$\vartheta_I := \|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} = \max\left\{\lambda_{\max}^2(\Phi_I) - 1, 1 - \lambda_{\min}^2(\Phi_I)\right\}.$$

Here $\lambda_{\max}^2$ and $\lambda_{\min}^2$ denote the biggest and smallest eigenvalue of $\Phi_I^*\Phi_I$ respectively. In this setting, the matrix $H := \Phi^*\Phi - \mathbb{I}$ is called the hollow Gram matrix and we call $\mu := \max_{i \neq j} |\langle \phi_i, \phi_j \rangle| = \|H\|_{\infty,1}$ the coherence. Applying Theorem 3 to this matrix, we get the following bound on ϑ_I.

Corollary 4 *Let $\Phi \in \mathbb{R}^{d \times K}$ be a dictionary with unit norm columns and assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $r \geq 2e^2\|WHW\|_{2,2}$*

$$\mathbb{P}_S\bigg(\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} \geq r\bigg) \leq 216K \exp\left(-\min\left\{\frac{r^2}{4e^2\|HW\|_{\infty,2}^2}, \frac{r}{2\mu}\right\}\right).$$

In this setting H is symmetric, hence $H^*W = HW$. The result can be used to bound

$$\mathbb{P}_S\left(\|\Phi_I\|_{2,2} \gtrless \sqrt{1 \pm r}\right) \quad \text{and} \quad \mathbb{P}_S\left(\|(\Phi_I^*\Phi_I)^{-1}\|_{2,2} \geq \frac{1}{1-r}\right).$$

This comes in handy when trying to prove recovery guarantees of sparse approximation algorithms later in this text.
Another frequently arising quantity is the cross-Gram matrix $H := \Psi^*\Phi - \mathrm{diag}(\Psi_I^*\Phi_I)$, where Φ and Ψ are dictionaries. In this setting, we call $\hat{\mu} := \max_{i \neq j} |\langle \phi_i, \psi_j \rangle|$ the cross-coherence. Applying Theorem 3 yields

Corollary 5 *Let $\Psi, \Phi \in \mathbb{R}^{d \times K}$ be dictionaries and assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $r \geq 2e^2\|WHW\|_{2,2}$*

$$\mathbb{P}_S\bigg(\|\Psi_I^*\Phi_I - \mathrm{diag}(\Psi_I^*\Phi_I)\| \geq r\bigg) \leq 216K \exp\bigg(-\min\bigg\{\frac{r^2}{4e^2\|HW\|_{\infty,2}^2}, \frac{r^2}{4e^2\|WH\|_{2,1}^2}, \frac{r}{2\hat{\mu}}\bigg\}\bigg).$$

Note that in contrast to Corollary 3.1.1 the matrix H is not symmetric any more, hence we need to control both $\|HW\|_{\infty,2}$ and $\|WH\|_{2,1}$. In contrast to previous works the above results are in terms of the maximal row norm of the weighted Gram matrix. By using the bounds

$$\begin{aligned}
&\|HW\|_{\infty,2} \leq \|\Psi^*\Phi W\|_{\infty,2} \leq \|\Psi^*\|_{\infty,2}\|\Phi W\|_{2,2} = \|\Phi W\|_{2,2},\\
&\|WH\|_{2,1} = \|H^*W\|_{\infty,2} \leq \|\Phi^*\|_{\infty,2}\|\Psi W\|_{2,2} = \|\Psi W\|_{2,2},\\
&\|WHW\|_{2,2} \leq \|\Psi W\|_{2,2}\|\Phi W\|_{2,2}
\end{aligned}$$

one would get bounds similar in spirit to the results of Chrétien and Darses [8] and Tropp [22].
We stick to the quantities $\|HW\|_{\infty,2}^2$ and $\|WH\|_{2,1}^2$ to see how the weights of the distribution interact with the structure of H. Intuitively the above results state that the more frequently an atom is picked, the less coherent it should be to all the other atoms in order for a random submatrix to be well-conditioned.
The generality of this result allows for $p_i \in [0,1]$, which thus includes models where some atoms are already known to be in the support and some to not appear at all. This allows for models where a dictionary D is a concatenation of two dictionaries ϕ and ψ, i.e. $D = (\phi, \psi)$ and the submatrix of interest consists of a *fixed* set of columns with cardinality n_a of the first dictionary and a *random* set of columns n_b of the second dictionary. Such a scenario can easily be modeled by setting the p_i and the weight matrix W accordingly and would yield similar results to [15].

3.2 Maximum row norm of a random restriction

Another frequently encountered random variable in sparse approximation is the maximal row norm $\|H_I\|_{\infty,2}$. Given a weight matrix W, the following Lemma states that one can expect this quantity to be approximately of size $\|HW\|_{\infty,2}$. This can be significantly smaller than the worst case $\max_{i,j}|H_{i,j}|\sqrt{S}$ for $|I| \leq S$, depending on the structure of H and W. Plugging in $H = \Psi^*\Phi - \mathrm{diag}(\Psi^*\Phi)$ we again see that the more frequently picked atoms should have smaller coherences in order for $\|HW\|_{\infty,2}$ to be small. This result is an integral part of the proof of Theorem 3 and hence we defer its proof to the appendix.

Lemma 6 *Let $H \in \mathbb{R}^{d\times K}$ be some matrix. Assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $v > 0$*

$$\mathbb{P}_S\left(\|H_I\|_{\infty,2} \geq v\right) \leq 2K\left(e\frac{\|HW\|_{\infty,2}^2}{v^2}\right)^{\frac{v^2}{\mu^2}}.$$

3.3 Poissonisation argument in the non-uniform case

As already mentioned, we have to bound the failure probability under the rejective sampling model by the failure probability under the Poisson sampling model in order to apply concentration inequalities for sums of independent random variables. In the uniform case the following lemma is not needed, as one can argue that the supports can also be sampled by

drawing one atom after the other to get a uniform support distribution - see Claim (3.29) p. 2173 in [4]. For the non-uniform case it is not that easy. Lemma 4.1 of [12] almost provides the result that we need, but has too restrictive assumptions on the expectations p_i. Therefore we prove[2] the following result which does not have any constraints on the expectations p_i.

Lemma 7 (Poissonisation) *Denote by $\mathbb{P}$ the probability measure corresponding to the Poisson sampling model* (1) *and by $\mathbb{P}_S$ the probability measure corresponding to the rejective sampling model* (2) *- both with the same weight matrix W. Let $f : \mathcal{P}(\mathbb{K}) \mapsto \{0, 1\}$ be such that for all $I, J \in \mathcal{P}(\mathbb{K})$*

$$f(I) \leq f(J) \quad \textit{if} \quad I \subseteq J.$$

Then for all $I \subseteq \mathbb{K}$

$$\mathbb{P}_S\left(f(I) = 1\right) \leq 2\,\mathbb{P}\left(f(I) = 1\right).$$

Proof Note that the conditions on f imply that if $f(J) = 0$ for some J, then $f(I) = 0$ for all $I \subset J$. We start by showing that for $0 \leq T \leq K - 1$ we have

$$\mathbb{P}\big(f(I) = 1 \;\big|\; |I| = T\big) \leq \mathbb{P}\big(f(I) = 1 \;\big|\; |I| = T + 1\big).$$

Expanding the conditional probability we get

$$\frac{\sum_{I:|I|=T} f(I)\mathbb{P}(I)}{\sum_{I:|I|=T} \mathbb{P}(I)} \leq \frac{\sum_{J:|J|=T+1} f(J)\mathbb{P}(J)}{\sum_{J:|J|=T+1} \mathbb{P}(J)},$$

which is equivalent to

$$\sum_{I:|I|=T} f(I)\mathbb{P}(I) \sum_{J:|J|=T+1} \mathbb{P}(J) \leq \sum_{J:|J|=T+1} f(J)\mathbb{P}(J) \sum_{I:|I|=T} \mathbb{P}(I). \tag{3}$$

By combining the sums on both sides and subtracting

$$\sum_{J:|J|=T+1} \sum_{I:|I|=T} \mathbb{P}(J)\mathbb{P}(I)f(I)f(J)$$

on both sides we see that (3) is equivalent to

$$\sum_{J:|J|=T+1} \sum_{I:|I|=T} \mathbb{P}(J)\mathbb{P}(I)f(I)[1 - f(J)] \leq \sum_{J:|J|=T+1} \sum_{I:|I|=T} \mathbb{P}(J)\mathbb{P}(I)f(J)[1 - f(I)] \tag{4}$$

Now the crucial step is to see that we can partition these sums in a very special way. For a pair (I, J), by definition of the Poisson sampling model, we can write $\mathbb{P}(I)\mathbb{P}(J)$ in the following way

$$\mathbb{P}(I)\mathbb{P}(J) = \prod_{i\in I} p_i \prod_{j\notin I}(1 - p_j) \prod_{i\in J} p_i \prod_{j\notin J}(1 - p_j) = \prod_{i\in I\cap J} p_i^2 \prod_{i\in I\triangle J} p_i(1 - p_i) \prod_{j\notin I\cup J} (1 - p_j)^2.$$

2. The result might be known but extremely well hidden, thus forcing us to prove it.

This implies that for two pairs (I, J), (I', J') with

$$I \cap J = I' \cap J' \quad \text{and} \quad I \triangle J = I' \triangle J' \quad \text{we have} \quad \mathbb{P}(I)\mathbb{P}(J) = \mathbb{P}(I')\mathbb{P}(J').$$

This allows us to define natural partitions on the set of pairs (I, J) such that the probability $\mathbb{P}(I)\mathbb{P}(J)$ is constant on each partition: Let $k \in \mathbb{T}$, $A \subseteq \mathbb{K}$ with $|A| = k$ and $B \subseteq \mathbb{K} \setminus A$ with $|B| = 2(T-k)+1$. A will be the intersection and B will model the symmetric difference of the sets I and J respectively. For such a combination of A, B we define

$$\mathcal{Q}_{A,B} := \{(I, J) : I, J \subseteq \mathbb{K}, |I| = T, |J| = T+1, I \cap J = A, I \triangle J = B\}.$$

Note that each pair (I, J) with $|I| = T$, $|J| = T+1$ can be *uniquely* assigned to one $\mathcal{Q}_{A,B}$. So if

$$\sum_{(I,J)\in\mathcal{Q}_{A,B}} f(I)[1-f(J)] \leq \sum_{(I,J)\in\mathcal{Q}_{A,B}} f(J)[1-f(I)] \tag{5}$$

for all possible choices of A, B then (4) follows and we are done.
We start with the special case $|A| = 0$ and fix $B \subseteq \mathbb{K}$ with $|B| = 2T+1$. With slight abuse of notation we write $I^c := B \setminus I$ for the complement in B. With this notation (5) becomes

$$\sum_{\substack{I\subseteq B\\|I|=T}} f(I)(1-f(I^c)) \leq \sum_{\substack{J\subseteq B\\|J|=T+1}} f(J)(1-f(J^c)).$$

Remembering that $f(I) \leq f(I \cup \{i\})$ and $f(J) \geq f(J \setminus \{i\})$ we get

$$\begin{aligned}
\sum_{\substack{I\subseteq B\\|I|=T}} f(I)(1-f(I^c)) &= \sum_{\substack{I\subseteq B\\|I|=T}} f(I)(1-f(I^c))\frac{1}{T+1}\sum_{i\in I^c} 1 \\
&= \frac{1}{T+1}\sum_{\substack{I\subseteq B\\|I|=T}} f(I)(1-f(I^c))\sum_{i\in I^c} f(I\cup\{i\})(1-f(I^c\setminus\{i\})) \\
&\leq \frac{1}{T+1}\sum_{\substack{I\subseteq B\\|I|=T}}\sum_{i\in I^c} f(I\cup\{i\})(1-f(I^c\setminus\{i\})) \\
&= \frac{1}{T+1}(T+1)\sum_{\substack{J\subseteq B\\|J|=T+1}} f(J)(1-f(J^c)).
\end{aligned}$$

If $|A| > 0$ then the same argument as above replacing $f(\cdot)$ with $f(A \cup \cdot)$ and T with $T - |A|$ yields (5) for all possible choices of A and B. Thus we get

$$\mathbb{P}\big(f(I) = 1 \;\big|\; |I| = T\big) \leq \mathbb{P}\big(f(I) = 1 \;\big|\; |I| = T+1\big).$$

Now we are finally in a position to prove our result. Note that

$$\begin{aligned}\mathbb{P}\left(f(I)=1\right) &= \sum_{k=1}^{K} \mathbb{P}\left(f(I)=1 \mid |I|=k\right)\mathbb{P}\left(|I|=k\right) \\ &\geq \mathbb{P}\left(f(I)=1 \mid |I|=S\right)\sum_{k=S}^{K}\mathbb{P}\left(|I|=k\right) \\ &\geq \mathbb{P}_S\left(f(I)=1\right)\cdot\frac{1}{2},\end{aligned}$$

where the last inequality follows from Theorem 3.2 of [13] which says that if the mean number of successes of K independent trials is an integer S, the median is also S. ■

Remark 8 *Applying the above result on the functions* $f_1(I) := \mathbb{1}_{\left\{\|H_{I,I}\|_{2,2}\geq t\right\}}$ *and* $f_2(I) := \mathbb{1}_{\{\|H_I\|_{\infty,2}\geq t\}}$ *we get*

$$\mathbb{P}_S\left(\|H_{I,I}\|_{2,2}\geq r\right) \leq 2\mathbb{P}\left(\|H_{I,I}\|_{2,2}\geq r\right)$$

and

$$\mathbb{P}_S\left(\|H_I\|_{\infty,2}\geq v\right) \leq 2\mathbb{P}\left(\|H_I\|_{\infty,2}\geq v\right).$$

Even though we stated our results only for the rejective sampling model, all of our proofs consist of first bounding the failure probability under the rejective sampling model by the failure probability under Poisson sampling model. Hence all of our results hold for the Poisson sampling model as well, with the failure bound actually improved by a factor $1/2$.

4. Application to sparse approximation

In this section we apply the derived result to sparse approximation. The starting point of sparse approximation is an underdetermined system of linear equations for which one tries to find the sparsest solution. Assuming that the signal y is a linear combination of S columns of a dictionary Φ, we show under which conditions sparse approximation algorithms are successful. To that end we define the following statistical model of our signals.

Definition 9 (Signal model) *We model our signals as*

$$y = \Phi_I x_I = \sum_{k=1}^{S}\phi_{i_k}x_{i_k}, \quad x_{i_k} = c_k\sigma_k, \ \ \forall k\in\{1,\dots,S\},$$

where $\Phi\in\mathbb{R}^{d\times K}$ *is a dictionary of* K *normalised atoms,* $I=\{i_1,\dots i_S\}$ *is the random support and* $c=\{c_1,\dots c_S\}$ *is an arbitrary sequence of strictly positive coefficients. We assume* $I\subseteq\mathbb{K}$ *is chosen according to the rejective sampling model with probabilities* $p_1,\dots,p_K$ *such that* $\sum_{i=1}^{K}p_i = S$ *and denote by* W *the corresponding weight matrix. Further we assume that the signs* σ_i *form an independent Rademacher sequence, i.e.* $\sigma_i=\pm1$ *with equal probability.*

This definition allows us to use probabilistic arguments to show that in the majority of cases, sparse approximation algorithms are able to recover the support under mild conditions on the dictionary Φ and on the coefficients x. We denote by $\mathbb{P}_y := \mathbb{P}_{\sigma,S}$ the product measure of the signs and the support and by $\mu := \max_{i\neq j}|\langle\phi_i,\phi_j\rangle|$ the coherence of the dictionary Φ.

4.1 Thresholding

We start by considering the fastest and conceptually easiest sparse approximation algorithm. Thresholding works by finding the indices corresponding to the S largest values of $|\langle y, \phi_i \rangle|$, i.e.

$$\begin{aligned} &\text{find} \quad J = \operatorname{argmax}_{|I|=S} \|\Phi_I^* y\|_1 \quad \text{and} \\ &\text{reconstruct} \quad x_J = P(\Phi_J) y. \end{aligned}$$

In slight abuse of notation, let $\|c\|_{\min} := \min_i |c_i|$. In [19], average case results for thresholding were derived for the uniform case. There, a sufficient condition for thresholding to work with high probability was $S\mu^2 \log(K) \lesssim \|c\|_{\min}^2 / \|c\|_{\max}^2$. We extend these results to the non-uniform case and show how the structure of the dictionary interacts with the distribution of coefficients.

Theorem 10 (Thresholding) *Assume that the signals follow the model in* (9)*, where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix and denote by $H = \Phi^*\Phi - \mathbb{I}$ the hollow Gram-matrix. If*

$$\mu^2 \leq \frac{\|c\|_{\min}^2}{8\|c\|_{\max}^2 \log(4K/\varepsilon)}, \qquad \text{and} \qquad \|HW\|_{\infty,2}^2 \leq \frac{\|c\|_{\min}^2}{8e^2\|c\|_{\max}^2 \log(4K/\varepsilon)},$$

then thresholding recovers the support with probability at least $1 - \varepsilon$.

Proof By definition of the algorithm, thresholding recovers the full support if

$$\|\Phi_{I^c}^* y\|_{\max} < \|\Phi_I^* y\|_{\min}.$$

Note that the signals have two sources of randomness, σ and I. Plugging in the definition of y we derive a bound on the failure probability

$$\begin{aligned} \mathbb{P}_y(\|\Phi_I^* y\|_{\min} < \|\Phi_{I^c}^* y\|_\infty) &= \mathbb{P}_y\left(\|\Phi_I^*\Phi_I x_I\|_{\min} < \|\Phi_{I^c}^*\Phi_I x_I\|_\infty\right) \\ &\leq \mathbb{P}_y\left(\|c\|_{\min} - \|(\Phi_I^*\Phi_I - \mathbb{I})x_I\|_\infty < \|\Phi_{I^c}^*\Phi_I x_I\|_\infty\right) \\ &\leq \mathbb{P}_y\left(\|c\|_{\min} < 2\|H_I x_I\|_\infty\right). \end{aligned}$$

Where we used that $x_{i_k} = \sigma_k c_k$, where $\sigma \in \mathbb{R}^S$ is an independent Rademacher sequence. Now as the signs σ are independent from the support I, we can apply Hoeffding's inequality to each entry of $H_I\sigma$ (Lemma 23) and use Lemma 6 to get

$$\begin{aligned} \mathbb{P}_y(\|\Phi_I^* y\|_{\min} < \|\Phi_{I^x}^* y\|_{\max}) &\leq \mathbb{P}_y\left(\|H_I x_I\|_\infty \geq \frac{\|c\|_{\min}}{2} \;\middle|\; \|H_I\|_{\infty,2} < \gamma\right) + \mathbb{P}_S\left(\|H_I\|_{\infty,2} \geq \gamma\right) \\ &\leq 2K \exp\left(-\frac{\|c\|_{\min}^2}{8\|c\|_{\max}^2\gamma^2}\right) + 2K\left(e\frac{\|HW\|_{\infty,2}^2}{\gamma^2}\right)^{\frac{\gamma^2}{\mu^2}}. \end{aligned}$$

Setting $\gamma^2 = \frac{\|c\|_{\min}^2}{8\|c\|_{\max}^2 \log(4K/\varepsilon)}$ we see that the conditions of the Theorem imply that the failure probability does not exceed ε. ∎

4.2 OMP

One of the most popular sparse approximation algorithms is the Orthogonal Matching Pursuit (OMP). This greedy algorithm finds the support iteratively, adding one index at a time to the current support. In every step, it picks the index of the atom which has the largest absolute inner product with the residual and then updates the residual. Initialising $r_0 = y$ and $J_0 = \emptyset$, it

$$\begin{aligned} &\text{finds} \quad j = \operatorname{argmax}_k |\langle \phi_k, r_i \rangle| \quad \text{and} \\ &\text{updates} \quad J_{i+1} = J_i \cup \{j\} \quad \text{resp.} \quad r_{J_{i+1}} = y - P(\Phi_{J_{i+1}})y, \end{aligned}$$

until a stopping criterion is met. Hence to prove that OMP recovers the correct support, one needs to ensure that it picks an atom from the support in each step. So assume OMP has successfully found $J \subseteq I$ in the i-th step, it will find another correct atom if

$$\|\Phi_{I^c}^* r_J\|_\infty < \|\Phi_L^* r_J\|_\infty,$$

where $L := I \setminus J$. Based on this observation we prove the following Theorem.

Theorem 11 (OMP) *Assume that the signals follow the model in* (9)*, where the support* $I \subseteq \mathbb{K}$ *is chosen according to the rejective sampling model with probabilities* $p_1, \ldots, p_K$ *such that* $\sum_{i=1}^K p_i = S$*. Further let* W *denote the corresponding weight matrix. Assume that the hollow Gram-matrix* $H = \Phi^*\Phi - \mathbb{I}$ *satisfies* $\|WHW\|_{2,2} \leq \frac{1}{4e^2}$*. If*

$$\begin{aligned} \|HW\|_{\infty,2}^2 &\leq \min\left\{ \min_{L \subseteq \{1,\ldots,S\}} \frac{\|c_L\|_\infty^2}{16e^2 \|c_L\|_2^2}, \frac{1}{16e^2 \log(216K/\varepsilon)} \right\} \quad \textit{and} \\ \mu &\leq \min\left\{ \min_{L \subseteq \{1,\ldots,S\}} \frac{\|c_L\|_\infty}{4\|c_L\|_2 \sqrt{\log(218K/\varepsilon)}}, \frac{1}{4\log(218K/\varepsilon)} \right\}, \end{aligned}$$

then OMP recovers the correct support with probability at least $1 - \varepsilon$*.*

Proof Set $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} =: \vartheta_I$ and assume that $\vartheta_I < 1/2$. We start by expanding the residual in step i

$$r_J = y - P(\Phi_J)y = \Phi_I x_I - P(\Phi_J)\Phi_I x_I = \Phi_{I\setminus J} x_{I\setminus J} - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_{I\setminus J} x_{I\setminus J}$$

Set $L := I \setminus J$. By definition, OMP finds another correct atom in the next step if

$$\|\Phi_{I^c}^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty < \|\Phi_L^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty, \quad (6)$$

i.e. the inner products with the residual of the remaining atoms in the support are bigger than the inner products with the residual of atoms outside the support. Writing this differently, we get the sufficient condition

$$\begin{aligned} &\|\Phi_{I^c}^*\Phi_L x_L\|_\infty + \|\Phi_{I^c}^*\Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L\|_\infty \\ &\qquad < \|x_L\|_\infty - \|(\Phi_L^*\Phi_L - \mathbb{I})x_L\|_\infty - \|\Phi_L^*\Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L\|_\infty, \end{aligned}$$

Note that

$$\max\left\{\|\Phi_{I^c}^*\Phi_L\|_{\infty,2}, \|\Phi_{I^c}^*\Phi_J\|_{\infty,2}, \|\Phi_L^*\Phi_L - \mathbb{I}\|_{\infty,2}, \|\Phi_L^*\Phi_J\|_{\infty,2}\right\} \leq \|H_I\|_{\infty,2}.$$

So OMP works if

$$2\|H_I\|_{\infty,2}\|x_L\|_2 + 2\|H_I\|_{\infty,2}\|(\Phi_J^*\Phi_J)^{-1}\|_{2,2}\|\Phi_J^*\Phi_L\|_{2,2}\|x_L\|_2 < \|x_L\|_\infty, \tag{7}$$

By properties of the operator norm we have $\|\Phi_J^*\Phi_L\|_{2,2} \leq \vartheta_I$ and $\|(\Phi_J^*\Phi_J)^{-1}\|_{2,2} \leq \frac{1}{1-\vartheta_I}$. Plugging this into (7) we see that OMP will pick a correct atom in the next step, if

$$\|H_I\|_{\infty,2}\left(2 + 2\frac{\vartheta_I}{1-\vartheta_I}\right) < \frac{\|x_L\|_\infty}{\|x_L\|_2}.$$

So on the set $\{\vartheta_I < 1/2\}$ the columns of Φ_I are linearly independent and we need to have $\|H_I\|_{\infty,2} < \min\limits_{L\subseteq\{1,\dots,S\}} \dfrac{\|c_L\|_\infty}{4\|c_L\|_2} =: \gamma$ for OMP to find the correct support. So by Corollary 4 and Lemma 6 we get

$$\begin{aligned}\mathbb{P}_S(\|\Phi_{I^c}^* r_J\|_\infty \geq \|\Phi_L^* r_J\|_\infty) &\leq \mathbb{P}_S(\vartheta_I \geq 1/2) + \mathbb{P}_S(\|H_I\|_{\infty,2} \geq \gamma) \\ &\leq 216K\exp\left(-\min\left\{\frac{1}{16e^2\|HW\|_{\infty,2}^2}, \frac{1}{4\mu}\right\}\right) + 2K\left(e\frac{\|HW\|_{\infty,2}^2}{\gamma^2}\right)^{\frac{\gamma^2}{\mu^2}}.\end{aligned}$$

Owing to the conditions on μ and $\|HW\|_{\infty,2}$ in the theorem, the right hand side does not exceed ε. ∎

Remark 12 *Note that for coefficients $c_k \sim \alpha^k$ we can always lower bound $\|c_L\|_\infty/\|c_L\|_2 > \sqrt{1-\alpha^2}$. So in the case of uniformly distributed supports ($p_i = S/K$) and a very incoherent dictionary the conditions above reduce to*

$$S\mu^2 \lesssim 1-\alpha^2 \quad \textit{and} \quad S\mu^2\log K \lesssim 1,$$

which are essentially the same conditions recently derived in [18] for exactly sparse signals. This is quite surprising, since this new proof is not only shorter but more importantly does not assume random signs of the coefficients but only a random support.

4.3 BP

A very popular alternative to the above algorithms is the Basis Pursuit principle. Instead of tackling the NP-hard problem of finding the sparsest solution with greedy methods, it instead aims to solve the convex relaxation

$$\hat{x} = \operatorname{argmin}\|x\|_1 \quad \text{s.t.} \quad y = \Phi x. \tag{8}$$

The average case performance in the uniform case of this optimisation problem has been extensively studied [22, 17, 4]. We give a short proof how these results can be transferred to the non-uniform case.

Theorem 13 *Assume that the signals follow the model in* (9)*, where the support* $I \subseteq \mathbb{K}$ *is chosen according to the rejective sampling model with probabilities* $p_1, \dots, p_K$ *such that* $\sum_{i=1}^{K} p_i = S$*. Further let* W *denote the corresponding weight matrix. Assume that the hollow Gram-matrix* $H = \Phi^*\Phi - \mathbb{I}$ *satisfies* $\|WHW\|_{2,2} \leq \frac{1}{4e^2}$*. If*

$$\mu \leq \frac{1}{4\log(220K/\varepsilon)}, \qquad \text{and} \qquad \|HW\|_{\infty,2}^2 \leq \frac{1}{16e^2\log(220K/\varepsilon)},$$

then BP recovers the correct coefficients with probability at least $1-\varepsilon$*.*

Proof We use results for fixed supports such that ℓ_1 minimisation yields the exact solution [21, 11]. Then we show that under the assumptions of the theorem these conditions are satisfied with high probability.

Proposition 14 ([21, 11]) *Assume* $y = \sum_{i\in I} \phi_i c_i \sigma_i$*, for some* $I \subset \{1, .., K\}$ *with* $|I| = S$*. If*

$$\|\Phi_{I^c}^* \Phi_I (\Phi_I^* \Phi_I)^{-1} \sigma_I\|_\infty < 1,$$

then x *is the unique solution to the* l_1*-minimisation problem* (8)*.*

Now set $M := \Phi_{I^c}^* \Phi_I (\Phi_I^* \Phi_I)^{-1}$ and $\vartheta_I := \|\Phi_I^* \Phi_I - \mathbb{I}\|$. As usual we note that

$$\|M\|_{\infty,2} = \|\Phi_{I^c}^* \Phi_I (\Phi_I^* \Phi_I)^{-1}\|_{\infty,2} \leq \|\Phi_{I^c}^* \Phi_I\|_{\infty,2} \|(\Phi_I^* \Phi_I)^{-1}\|_{2,2} \leq \|H_I\|_{\infty,2} \frac{1}{1-\vartheta_I}.$$

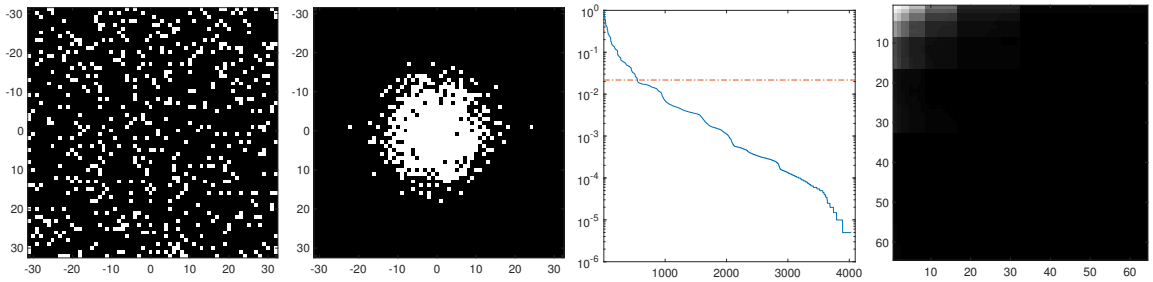
Now Corollary 4 together with applying Hoeffding's inequality to each entry of $M\sigma$ (Lemma 23) and Lemma 6 yield

$$\begin{aligned}
\mathbb{P}_y\left(\|M\sigma\|_\infty \geq 1\right) &\leq \mathbb{P}_y\left(\|M\sigma\|_\infty \geq 1 \mid \|M\|_{\infty,2} \leq 2\gamma\right) + \mathbb{P}_S\left(\vartheta_I \geq 1/2\right) + \mathbb{P}_S(\|H_I\|_{\infty,2} \geq \gamma) \\
&\leq 2K \exp\left(-\frac{1}{8\gamma^2}\right) + 216K \exp\left(-\min\left\{\frac{1}{16e^2\|HW\|_{\infty,2}^2}, \frac{1}{4\mu}\right\}\right) \\
&\qquad + 2K\left(e\frac{\|HW\|_{\infty,2}^2}{\gamma^2}\right)^{\frac{\gamma^2}{\mu^2}}.
\end{aligned}$$

Setting $\gamma^2 = \frac{1}{8\log(220K/\varepsilon)}$ we see that under the conditions of the Theorem, the failure probability is bounded by ε. ∎

To illustrate our results we conduct the following small experiment. We take the 2D Haar-Wavelet decomposition of 1000 randomly chosen normalised patches y_n of size 64×64 from the image *Peppers* before applying a threshold of $\sqrt{\log(d)/d}/6$ for $d = 64^2$ on the coefficients to get a sparse approximation. Counting how often each atom is used we get a proxy for the probability of any atom being in the sparse support I Figure 2 (c-d). We denote by W the corresponding weight matrix and by D the vectorised 2D Haar-Wavelet basis. Now we are given two measurement matrices derived from subsampled vectorised 2D-DCT matrices which we denote by $A_1 \in \mathbb{R}^{m\times d}$ and $A_2 \in \mathbb{R}^{m\times d}$. The subsampling pattern is generated by two different subsampling strategies - see Figure 2 (a-b). For our experiment we set $m = 512$. We are tasked with solving the minimisation problem

$$\hat{x} = \operatorname{argmin} \|x\|_1 \quad \text{s.t.} \quad A_i y = A_i D x \tag{9}$$



(a) (b) (c) (d)

Figure 2: (a) The K-space $\{(k_1, k_2) : -\sqrt{K}/2 + 1 \leq k_1, k_2 \leq \sqrt{K}/2\}$ with the frequencies used for the measurement matrix A_1 (a) and the measurement matrix A_2 (b). (c) Expectation of each atom to be in the support (blue) and average expectations for comparison (red) on a log scale. (d) Locations of non-zero coefficients of patches in the 2D-Haar Wavelet Basis.

and are given the choice between the two measurement matrices A_1 and A_2. Our results tell us that as long as the sparse supports of our signals follow the distribution described by the weight matrix W, we should pick the sensing dictionary A_i that minimises the quantities μ, $\|HW\|_{\infty,2}$ and $\|WHW\|_{2,2}$ (where H is the hollow Gram matrix $A_i DD^* A_i^* - \mathrm{diag}(A_i DD^* A_i^*)$). Looking at Table 1 columns 1-3 we see that for signals following the distribution specified by W, our results suggest A_2 yields better performance. To test the actual performance, we used BP to recover the coefficients x_n from the set of

	μ	$\|HW\|_{\infty,2}$	$\|WHW\|_{2,2}$	MSE
A_1	0.89	3.80	2.80	0.18
A_2	0.98	0.84	0.87	0.06

Table 1: The first line corresponds to the uniform subsampling strategy, the second line to the variable density subsampling strategy.

incomplete measurements $A_i y_n = A_i D x_n$. Note that the coefficients x_n are not sparse, but compressible. Looking at the mean squared error (MSE) $\frac{1}{N}\sum_{n=1}^{N} \|y_n - D\hat{x}_n\|_2^2$ in Table 1, we see that even though strictly speaking our theory does not apply here (as these signals are not perfectly sparse) the quantities $\|HW\|_{\infty,2}$ and $\|WHW\|_{2,2}$ seem to be good predictors of average performance for signals where the sparse support (in this case of the biggest entries) follows a distribution specified by a weight matrix W.

5. Sensing dictionaries and preconditioning

As an application of our results we construct a sensing dictionary to improve the average performance of a dictionary for thresholding and OMP, given that we know the distribution of supports. We then extend these ideas to BP via preconditioning.

In the most general sense a sensing dictionary[3] Ψ for a given dictionary Φ is a matrix of the same size as Φ, whose columns satisfy $\langle \psi_k, \phi_k \rangle = 1$ for all $k \in \mathbb{K}$. It can be used in greedy algorithms to replace the original dictionary in the atom selection step. Sensing dictionaries improving the worst case performance of OMP and thresholding were first characterised and constructed in [20]. In [19] those ideas were generalised to construct sensing dictionaries that improve the average performance. We extend these average case results to non-uniformly distributed supports to see how the distribution interacts with the structure of the sensing dictionary.

The main idea in thresholding and OMP is to determine which atoms to include in the support by looking at the absolute inner products between the signal and the atoms. Using a sensing dictionary changes this step in the thresholding algorithm to

$$\begin{aligned} &\text{find} \quad J = \operatorname{argmax}_{|I|=S} \|\Psi_I^* y\|_1 \quad \text{and} \\ &\text{reconstruct} \quad x_J = P(\Phi_J) y. \end{aligned}$$

For OMP, similarly, the sensing dictionary comes into play when choosing the next atom to add to the support while the residual update step stays the same. Initialising $r_0 = y$ and $J_0 = \emptyset$, for OMP with sensing dictionary Ψ one has to

$$\begin{aligned} &\text{find} \quad j = \operatorname{argmax}_k |\langle \psi_k, r_i \rangle| \quad \text{and} \\ &\text{update} \quad J_{i+1} = J_i \cup j \quad \text{resp.} \quad r_{J_{i+1}} = y - P(\Phi_{J_{i+1}}) y, \end{aligned}$$

until a stopping criterion is met. Now we will show how to construct a sensing dictionary given knowledge about the distribution of the supports.

Assuming that the distribution of our supports follows a Poisson or rejective sampling model with known weight matrix W, Theorems 24 and 25 in the appendix show that a sensing dictionary with good average case performance should ideally minimise $\|(\Psi^*\Phi - \mathbb{I})W\|_{\infty,2}$. We now try to find Ψ such that this quantity is minimised under the constraint that $\operatorname{diag}(\Psi^*\Phi) = \mathbb{I}$. First note that the quantity $\|(\Psi^*\Phi - \mathbb{I})W\|_{\infty,2}^2$ is bounded from above by $\|(\Psi^*\Phi - \mathbb{I})W\|_F^2$. Minimising the Frobenius norm instead of the maximum row norm has the big advantage that there exists an easy to find analytic solution. For ease of notation let $P := W^2$. Following [19] we use Lagrangian multipliers and derive both the objective and the constraint function along ψ_j to get

$$\frac{d}{d\psi_j} \|\Psi^*\Phi W\|_F^2 = \sum_i 2\langle \phi_i, \psi_j \rangle \phi_i p_i = 2\Phi P \Phi^* \psi_j$$

$$\frac{d}{d\psi_j} \langle \phi_j, \psi_j \rangle = \phi_j.$$

So we see that for thresholding and OMP, the sensing dictionary should be set to

$$\Psi := (\Phi P \Phi^*)^{-1} \Phi D,$$

where D is a diagonal matrix s.t. $\langle \phi_i, \psi_i \rangle = 1$ for all $i \in \mathbb{K}$. This compares nicely to the result in [19], where they arrived at $\Psi = (\Phi\Phi^*)^{-1}\Phi D$ for the special case $p_i = S/K$. This

3. Note that strictly speaking $\Psi \neq \Phi$ is not actually a dictionary, as the columns are not normalised.

shows how the distribution of coefficients changes the optimal sensing dictionary via the diagonal matrix P. Figures 2 and 3 show how the performance of thresholding and OMP improves when using sensing dictionaries for various dictionaries and distributions.
For BP it is not that simple to use a different sensing dictionary. Instead we use preconditioning, multiplying the original dictionary by an invertible matrix from the left and by a diagonal matrix from the right. Inspired by the heuristic argument above, we set

$$\Psi = (\Phi P \Phi^*)^{-1/2} \Phi D^{1/2},$$

where D is a diagonal matrix s.t. $\langle \psi_i, \psi_i \rangle = 1$ for all $i \in \{1, \dots, K\}$. We then change the BP minimisation problem to

$$\min \|z\|_1 \quad \text{such that} \quad \tilde{y} = \Psi z,$$

where $\tilde{y} = (\Phi P \Phi^*)^{-1/2} y$. This is equivalent to the original optimisation problem, as D is a diagonal matrix with positive entries on its diagonal and $(\Phi P \Phi^*)^{-1/2}$ is invertible.

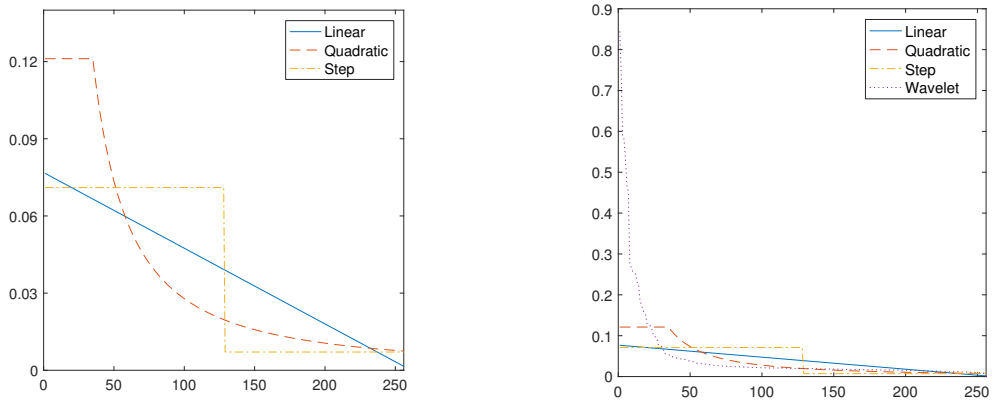


Figure 3: (a) Expectations of the Bernoulli random variables employed in our distribution models. (b) The same plot with the relative frequency of the wavelet coefficients from Figure 1 for comparison.

5.1 Numerical results

To test the performance of our sensing dictionaries and preconditioning, we conduct the following experiment. We build 2 dictionaries, each with 256 atoms of dimension 128. The columns of the first dictionary are drawn uniformly at random from the unit sphere and the second dictionary is a uniformly subsampled Discrete Cosine Basis with subsequent normalisation. We consider three different distribution models: quadratic, linear and step - see Figure 3. For each distribution model and each support size between 1 and 80 we construct 1000 signals by choosing the support according to the rejective sampling model specified in Section 2. The sparse coefficients of x have absolute value one with random signs, i.e. $x_i = \pm 1$ with equal probability. We then compare how often thresholding,

OMP and BP can recover the full support when using the original dictionary, the uniform average case sensing dictionary ($P = \mathbb{I}\frac{S}{K}$), and the distribution specific average case sensing dictionary (or the preconditioned matrix for BP). The results for thresholding and OMP are displayed in Table 2 and Table 3 respectively. Table 4 shows how the preconditioning changes the recovery rates for BP. As can be seen, incorporating prior knowledge about the distribution of supports into the algorithms improves performance quite significantly for all 3 algorithms.

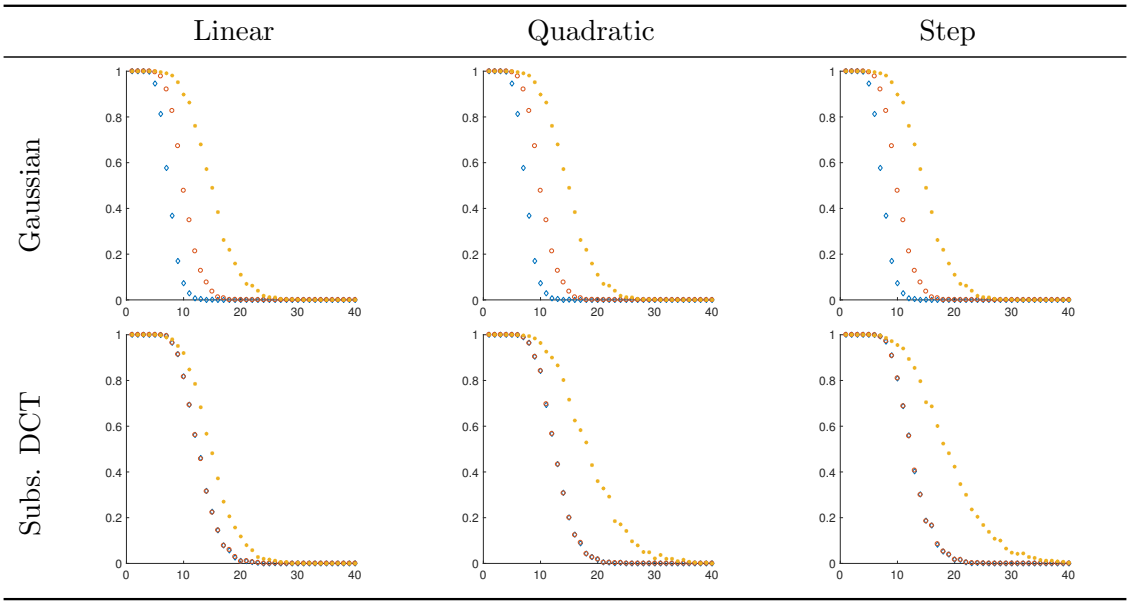


Table 2: The recovery rates for thresholding with different sensing dictionaries are plotted on the y-axis and the size of sparse supports on the x-axis. Blue corresponds to no sensing dictionary, red to the uniform average case sensing dictionary and orange to the distribution specific average case sensing dictionary.

6. Discussion

In this paper we have derived concentration inequalities for norms of random subdictionaries with non-uniformly distributed supports. This has allowed us to derive sufficient conditions for sparse approximation algorithms to recover the correct support given that the support of coefficients follows a rejective sampling or Poisson sampling model. We have shown that recovery of signals depends on the structure of the cross-Gram matrix and the distribution of supports, proving that more frequently used atoms should be more incoherent than less frequently used ones. The generalisation from uniformly to non-uniformly distributed supports gives valuable insight into how, in a compressed sensing setup, measurement matrices should be chosen or constructed. For both thresholding and OMP it was shown that using sensing dictionaries that take the distribution of supports into account improves performance. Using precondition to extend this argument to BP, it was also shown that prior knowledge about the distribution leads to improved performance for BP as well.

Our next goal is to use these results to prove convergence of dictionary learning algorithms for signals where the atoms of the generating dictionary are not equally used - as seems to happen in practice. Not only should this show that the more frequently used atoms converge faster, but it should also give insights how to best estimate the size of the generating dictionary.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) under Grant no. Y760. Finally, many thanks go to Elisabeth Schneckenreiter for proof-reading the manuscript.

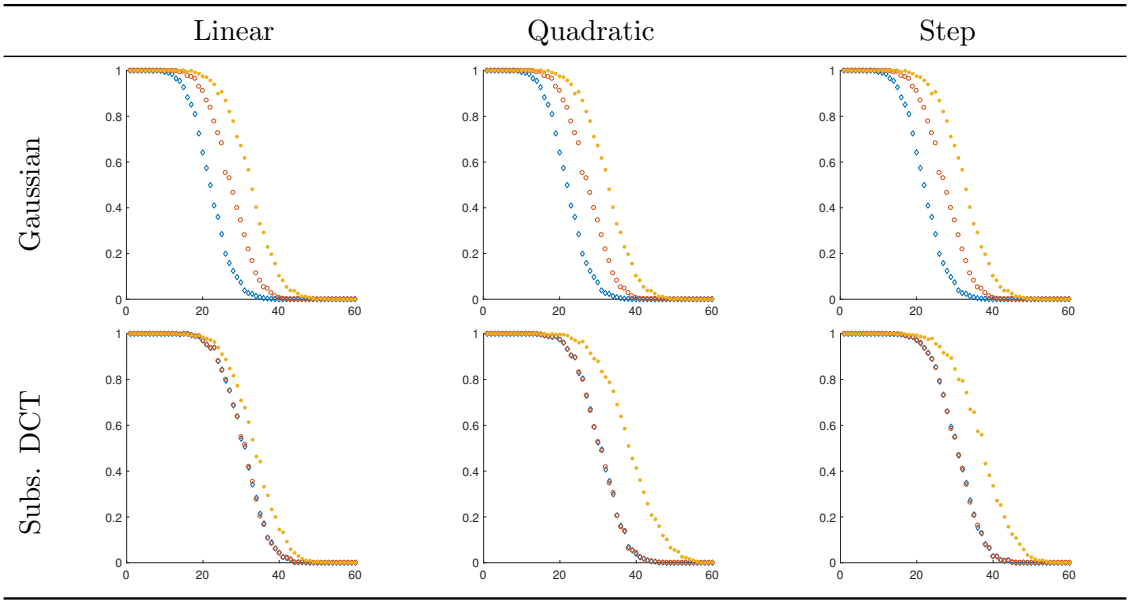


Table 3: The recovery rates for OMP with different sensing dictionaries are plotted on the y-axis and the size of sparse supports on the x-axis. Blue corresponds to no sensing dictionary, red to the uniform average case sensing dictionary and orange to the distribution specific average case sensing dictionary.

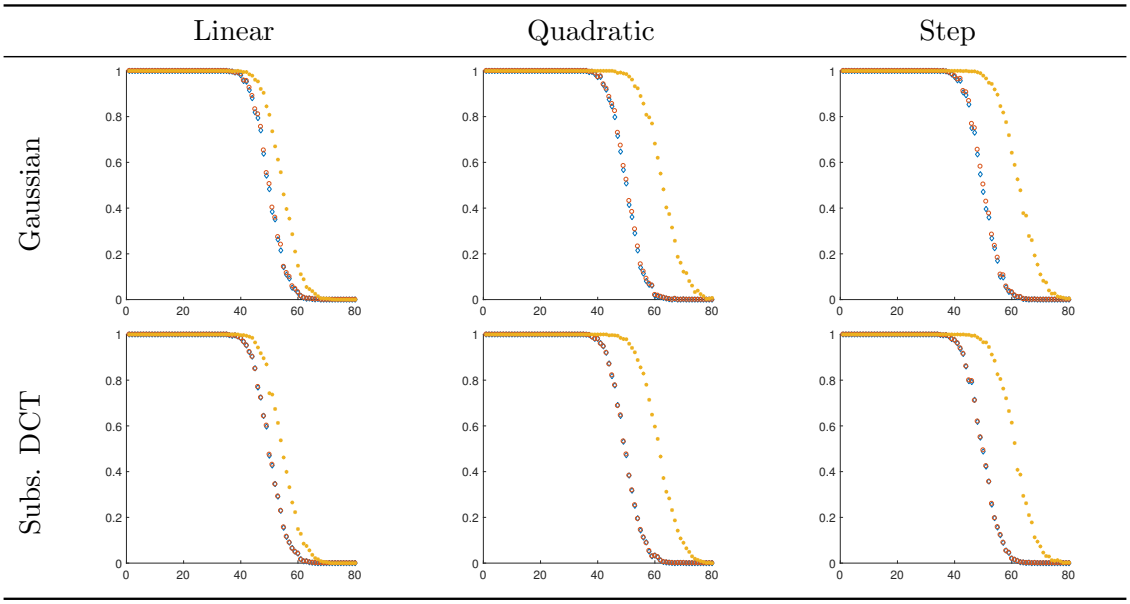


Table 4: The recovery rates for BP with different preconditioning strategies are plotted on the y-axis and the size of sparse supports on the x-axis. Blue corresponds to the original ℓ_1 minimisation problem, red corresponds to preconditioning in the uniform case and orange corresponds to preconditioning with the correct weights.

Appendix A. Proof of Theorem 3

The proof follows the one that appeared in Chrétien and Darses [8] with some minor changes to account for the non-uniformly distributed supports and the extension to non-symmetric matrices. We start with an argument that lets us decouple the random variables selecting the rows and columns. This is crucial for the application of concentration inequalities for sums of independent random matrices later in the proof.

Proposition 15 *Let $H \in \mathbb{R}^{d\times K}$ be some matrix. Assume $I \subseteq \mathbb{K}$ is chosen according to the Poisson sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^{K} p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $r \geq 0$*

$$\mathbb{P}\left(\|RHR\| \geq r\right) \leq 36\, \mathbb{P}\left(\|RHR'\| \geq r/2\right),$$

where R' is an independent copy of R.

Proof Let η_i for $1 \leq i \leq K$ be a series of i.i.d. Rademacher random variables. We follow the approach of Chrétien/Darses [8] and Tropp [23] who refer to Bourgain/Tzafriri [2] and de la Peña/Giné [9]. We define

$$Z = Z(\eta, \delta) := \sum_{i\neq j} (1 - \eta_i\eta_j)\delta_i\delta_j\vec{H}_{ij}.$$

Setting $Y = \sum_{i\neq j} \delta_i\delta_j\vec{H}_{ij}\eta_i\eta_j$, we can write

$$Z = RHR - Y.$$

Recall the Hahn-Banach Theorem.

Theorem 16 (Hahn-Banach) *Let X be a real vector space and p a sublinear functional on X. Let f be a linear functional defined on a subspace $A \subset X$, and satisfying $f(a) \leq p(a)$ for all $a \in A$. Then there exists a linear functional $\tilde{f}$ on X satisfying*

$$\begin{aligned} \tilde{f}(a) &= f(a) \quad \textit{for all} \quad a \in A \quad \textit{and} \\ \tilde{f}(x) &\leq p(x) \quad \textit{for all} \quad x \in X. \end{aligned}$$

From now on we work conditional on a choice of I (i.e. we fix our sequence δ_i, therefore the support set I and the entries of R are fixed as well). Denote by $A = \{\lambda RHR \mid \lambda \in \mathbb{R}\}$ the subspace generated by RHR and define a linear form $f(\lambda RHR) = \lambda \|RHR\|$ on this subspace. By definition we have $f(a) \leq \|a\| =: p(a)$ for all $a \in A$, where the properties of the operator norm imply that p is a sublinear functional. Thus the Hahn-Banach Theorem gives us the existence of a linear functional $\tilde{f}$ satisfying

$$\tilde{f}(RHR) = f(RHR) = \|RHR\|$$

and

$$\tilde{f}(Z) \leq \|Z\|.$$

Using the linearity of $\tilde{f}$ and that Y is symmetric around 0 we get

$$\begin{aligned} \mathbb{P}_\eta\left(\|Z\| \geq \|RHR\|\right) &= \mathbb{P}_\eta\left(\|Z\| \geq \tilde{f}(RHR)\right) \\ &\geq \mathbb{P}_\eta\left(\tilde{f}(Z) \geq \tilde{f}(RHR)\right) \\ &= \mathbb{P}_\eta\left(\tilde{f}(-Y) + \tilde{f}(RHR) \geq \tilde{f}(RHR)\right) = \mathbb{P}_\eta\left(\tilde{f}(Y) \geq 0\right), \end{aligned}$$

where again by linearity of $\tilde{f}$ we have

$$\tilde{f}(Y) = \sum_{\substack{i \neq j \\ i,j \in I}} \tilde{f}(\vec{H}_{ij})\eta_i\eta_j = \sum_{\substack{i > j \\ i,j \in I}} \left[\tilde{f}(\vec{H}_{ij}) + \tilde{f}(\vec{H}_{ji})\right]\eta_i\eta_j.$$

So we see that $\tilde{f}(Y)$ is a homogeneous Rademacher chaos of order 2. For ease of notation write $\xi := \tilde{f}(Y)$. As ξ is a centered real random variable we can write $\mathbb{E}[|\xi|] = 2\mathbb{E}[\xi \mathbb{I}_{\xi>0}]$ and a simple application of Hölders inequality yields

$$\mathbb{E}[|\xi|]^2 = 4\mathbb{E}[\xi \mathbb{I}_{\xi>0}]^2 \leq 4\mathbb{P}\left(\xi > 0\right)\mathbb{E}[\xi^2].$$

Write $\mathbb{E}[\xi^2] = \mathbb{E}[\xi^{2/3}\xi^{4/3}]$ and apply Hölders inequality again with $p = \frac{3}{2}$ and $q = 3$ to get

$$\mathbb{E}[\xi^2] \leq \mathbb{E}[|\xi|]^{\frac{2}{3}}\mathbb{E}[\xi^4]^{\frac{1}{3}}.$$

Putting the above together we arrive at

$$\mathbb{P}(\xi > 0) \geq \frac{1}{4}\frac{\mathbb{E}[|\xi|]^2}{\mathbb{E}[\xi^2]} \geq \frac{1}{4}\frac{\mathbb{E}[\xi^2]^2}{\mathbb{E}[\xi^4]}.$$

Since ξ is a homogeneous Rademacher chaos of order 2 we can apply Lemma 2.1 of Chrétien and Darses [8], which states

$$\frac{\mathbb{E}[\xi^2]^2}{\mathbb{E}[\xi^4]} \geq \frac{1}{9}.$$

So

$$\mathbb{P}_\eta \left(\|Z\| \geq \|RHR\| \right) \geq \frac{1}{36}.$$

Multiplying both sides with $\mathbb{I}_{\{\|RHR\| \geq r\}}$ and taking the expectation w.r.t. to I we get

$$\mathbb{P}(\|RHR\| \geq r) \leq 36 \, \mathbb{P}(\|Z\| \geq r).$$

Now by the same argument as in Tropp [23], Proposition 2.1 there exists a $\bar{\eta} \in \{-1, 1\}^K$ s.t.

$$\mathbb{P}\left(\|Z\| \geq r \right) = \mathbb{E}\left[\mathbb{E}\left(\mathbb{I}_{\{\|Z(\eta,\delta)\| \geq r\}} \mid \eta \right) \right] \leq \mathbb{E}\left(\mathbb{I}_{\{\|Z(\bar{\eta},\delta)\| \geq r\}} \right) = \mathbb{P}\left(\|Z(\bar{\eta}, \delta)\| \geq r \right).$$

Setting $T = \{i : \bar{\eta}_i = 1\}$, we see by the definition of Z

$$Z(\bar{\eta}, \delta) = 2 \sum_{j \in T, k \in T^c} \delta_j \delta_k \vec{H}_{jk} + 2 \sum_{j \in T^c, k \in T} \delta_j \delta_k \vec{H}_{jk} = 2 \sum_{j \in T, k \in T^c} \delta_j \delta_k (\vec{H}_{jk} + \vec{H}_{kj}).$$

Now we can do the decoupling. As δ_i for $i \in T$ are independent from δ_j for $j \in T^c$ we can replace δ_j for $j \in T^c$ with δ' which is an independent copy of δ. Thus

$$\mathbb{P}\left(\|Z\| \geq r \right) \leq \mathbb{P}\left(\| \sum_{j \in T, k \in T^c} \delta_j \delta'_k (\vec{H}_{jk} + \vec{H}_{kj}) \| \geq r/2 \right),$$

Note that (after reordering) this matrix is of the form $\begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix}$, where A corresponds to $\sum_{j \in T, k \in T^c} \delta_j \delta'_k \vec{H}_{jk}$ and B to $\sum_{j \in T, k \in T^c} \delta_j \delta'_k \vec{H}_{kj}$. The operator norm of this reordered matrix satisfies

$$\left\| \begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} B^*B & 0 \\ 0 & A^*A \end{pmatrix} \right\| = \max\{\|A\|^2, \|B\|^2\}.$$

As the spectral norm of a submatrix is always less than or equal to the spectral norm of the whole matrix we get by reintroducing the missing entries

$$\mathbb{P}(\|Z\| \geq r) \leq \mathbb{P}(\|RHR'\| \geq r/2).$$

Putting everything together yields the desired result. ■

Now we are in a position to apply concentration inequalities for sums of independent random matrices. For that recall the Matrix Chernoff inequality, which can be found in [24].

Theorem 17 (Matrix Chernoff inequality [24]) *Let $X_1, ..., X_K$ be independent, symmetric and positive semi-definite random matrices taking values in $\mathbb{R}^{d\times d}$. Now let $B, m > 0$ and assume that for all $1 \le k \le K$*

$$\|X_k\| \le B \quad \text{and} \quad \|\sum_{k=1}^{K} \mathbb{E}X_k\| \le m.$$

Then, for all $t > 0$

$$\mathbb{P}\left(\|\sum_{k=1}^{K} X_k\| \ge t\right) \le d\left(\frac{em}{t}\right)^{t/B}.$$

Now we are going to derive a bound on $\mathbb{P}\left(\|RHR'\| \ge r\right)$ by applying the Matrix Chernoff inequality 3 times. We first use the randomness of R' while holding R fixed, then we bound the two resulting terms involving R. This leads to the following result

Lemma 18 *Let $H \in \mathbb{R}^{d\times K}$ be some matrix. Assume $I, I' \subseteq \mathbb{K}$ - leading to the selector matrices R, R' - are chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^{K} p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $r > 0$*

$$\mathbb{P}\left(\|RHR'\| \ge r\right) \le K\left(e\frac{u^2}{r^2}\right)^{\frac{r^2}{v^2}} + K\left(e\frac{\|WHW\|^2}{u^2}\right)^{\frac{u^2}{\|HW\|^2_{\infty,2}}} + K\left(e\frac{\|WH\|^2_{2,1}}{v^2}\right)^{\frac{v^2}{\mu^2}}. \quad (10)$$

We begin by bounding $\mathbb{P}\left(\|RHR'\| \ge r\right)$.

Lemma 19 *Let $H \in \mathbb{R}^{d\times K}$ be some matrix. Assume $I' \subseteq \mathbb{K}$ - leading to the selector matrix R' - is chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^{K} p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $r > 0$*

$$\mathbb{P}\left(\|RHR'\| \ge r\right) \le K\left(e\frac{\|RHW\|^2}{r^2}\right)^{\frac{r^2}{\|RH\|^2_{2,1}}}.$$

Proof Using that for any matrix A, $\|AA^*\| = \|A^*A\| = \|A\|^2$ we see

$$\mathbb{P}\left(\|RHR'\| > r\right) = \mathbb{P}\left(\|RHR'\|^2 > r^2\right) = \mathbb{P}\left(\|RHR'H^*R\| > r^2\right).$$

Denoting by Z_j the j-th column of RH, we get

$$RHR'H^*R = \sum_{j=1}^{K} \delta'_j Z_j Z_j^*. \quad (11)$$

Then we have $\|Z_j Z_j^*\| = \|Z_j\|_2^2 \le \|RH\|^2_{2,1}$ and

$$\|\sum_{j=1}^{K} \mathbb{E}[\delta'_j Z_j Z_j^*]\| = \|\sum_{j=1}^{K} p_j Z_j Z_j^*\| = \|RHWWH^*R\| = \|RHW\|^2.$$

As the right hand side of (11) is a sum of independent random variables, an application of the Matrix Chernoff inequality yields the result. ■

Now we turn to bounding the two quantities $\|RHW\|$ and $\|RH\|_{2,1}$ by the same argument as above.

Lemma 20 *Let $H \in \mathbb{R}^{d\times K}$ be some matrix. Assume $I \subseteq \mathbb{K}$ is chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $u > 0$*

$$\mathbb{P}\left(\|RHW\| > u\right) \leq K\left(e\frac{\|WHW\|^2}{u^2}\right)^{\frac{u^2}{\|HW\|_{\infty,2}^2}}.$$

Proof Again using that for any matrix A, $\|AA^*\| = \|A^*A\| = \|A\|^2$ we see

$$\mathbb{P}\left(\|RHW\| > u\right) = \mathbb{P}\left(\|RHW\|^2 > u^2\right) = \mathbb{P}\left(\|WH^*RHW\| > u^2\right).$$

Now denote by Y_j the j-th row of HW then we get

$$WH^*RHW = \sum_{j=1}^K \delta_j Y_j^* Y_j. \tag{12}$$

We have $\|Y_j^*Y_j\| = \|Y_j\|_2^2 \leq \|HW\|_{\infty,2}^2$ and

$$\|\sum_{j=1}^K \mathbb{E}[\delta_j Y_j^* Y_j]\| = \|\sum_{j=1}^K p_j Y_j^* Y_j\| = \|WH^* \operatorname{diag}((p_k)_k) HW\| = \|WHW\|^2.$$

As the right hand side of (12) is a sum of independent random variables, an application of the Matrix Chernoff inequality yields the result. ■

We now restate and prove Lemma 6 for the Poisson sampling model. Note that by definition $\|RH^*\|_{2,1} = \|HR\|_{\infty,2} = \|H_I\|_{\infty,2}$. Recall that by Lemma 7

$$\mathbb{P}_S\left(\|H_I\|_{\infty,2} \geq v\right) \leq 2\,\mathbb{P}\left(\|H_I\|_{\infty,2} \geq v\right),$$

so this result translates immediately to the rejective sampling model.

Lemma 21 *Let $H \in \mathbb{R}^{d\times K}$ be some matrix. Assume $I \subseteq \mathbb{K}$ is chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix. Then, for all $v > 0$*

$$\mathbb{P}\left(\|H_I\|_{\infty,2} \geq v\right) \leq K\left(e\frac{\|HW\|_{\infty,2}^2}{v^2}\right)^{\frac{v^2}{\mu^2}}.$$

Proof We begin by writing $\|H_I\|$ as the maximum of a sum of independent random variables

$$\|H_I\|_{\infty,2}^2 = \max_{i\in\{1,\ldots,K\}} \sum_{j=1}^K \delta_j H_{ij}^2.$$

Now we fix $i \in \{1, .., K\}$ and apply the standard Chernoff inequality

$$\mathbb{P}\left(\sum_{j=1}^{K} \delta_j H_{ji}^2 \geq v^2\right) \leq \left(e\frac{\|HW\|_{\infty,2}^2}{v^2}\right)^{\frac{v^2}{\mu^2}}.$$

Taking a union bound yields the result. ■

Finally we can put everything together and prove our main result. The main difficulty lies in picking v and u such as to minimise the probability bound in (10).

Proof [Theorem 3] Set

$$\alpha := \min\left\{\frac{r^2}{4e^2\|WH\|_{2,1}^2}, \frac{r^2}{4e^2\|HW\|_{\infty,2}^2}, \frac{r}{2\mu}\right\} \qquad v^2 := \frac{r^2}{4\alpha} \qquad u^2 := \frac{r^2}{4e^2}.$$

Now these definitions and the assumption $r^2 \geq 4e^4\|WHW\|^2$ imply the following 6 inequalities

$$\frac{u^2}{\|HW\|_{\infty,2}^2} = \frac{r^2}{4e^2\|HW\|_{\infty,2}^2} \geq \alpha \qquad e\frac{\|WHW\|^2}{u^2} = \frac{4e^3\|WHW\|^2}{r^2} \leq e^{-1}$$

$$\frac{v^2}{\mu^2} = \frac{r^2}{4\alpha\mu^2} \geq \alpha \qquad e\frac{\|WH\|_{2,1}^2}{v^2} = \frac{4e\|WH\|_{2,1}^2\alpha}{r^2} \leq e^{-1}$$

$$\frac{r^2}{4v^2} = \frac{4r^2\alpha}{4r^2} = \alpha \qquad e\frac{4u^2}{r^2} = \frac{4er^2}{4e^2r^2} = e^{-1}.$$

So

$$\mathbb{P}_S\left(\|RHR\| \geq r\right) \leq 2\mathbb{P}\left(\|RHR\| \geq r\right) \leq 72\mathbb{P}\left(\|RHR'\| \geq r/2\right),$$

together with

$$\mathbb{P}\left(\|RHR'\| \geq r\right) \leq K\left(\left(e\frac{4u^2}{r^2}\right)^{\frac{r^2}{4v^2}} + \left(e\frac{\|WHW\|^2}{u^2}\right)^{\frac{u^2}{\|HW\|_{\infty,2}^2}} + \left(e\frac{\|WH\|_{2,1}^2}{v^2}\right)^{\frac{v^2}{\mu^2}}\right)$$

shows that

$$\mathbb{P}_S\left(\|RHR\| \geq r\right) \leq 216Ke^{-\alpha}.$$

■

Remark 22 *In the published version of Chrétien and Darses [8] there is a tiny bug in the proof of Proposition 4.2 in the way the variables u and v are balanced. In particular, for very small μ, inequality 4.17 may be violated. v^2 should instead be defined via an equality in 4.15, whereas 4.14 should be an inequality.*

For convenience we restate an easy consequence of Hoeffding's inequality.

Lemma 23 (Hoeffding) *Let $M \in \mathbb{R}^{K\times S}$ be a matrix and $x \in \mathbb{R}^S$ such that $\operatorname{sign}(x) \in \mathbb{R}^S$ is an independent Rademacher sequence. Then, for all $t \geq 0$*

$$\mathbb{P}\left(\|Mx\|_\infty \geq t\right) \leq 2K \exp\left(-\frac{t^2}{2\|M\|_{\infty,2}^2\|x\|_\infty^2}\right).$$

Proof We apply Hoeffding's inequality to the k-th entry of Mx, which yields

$$\mathbb{P}\left(|\textstyle\sum_j M_{kj}x_j| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_j M_{kj}^2x_j^2}\right) \leq 2\exp\left(-\frac{t^2}{2\|x\|_\infty^2\|M^k\|_2^2}\right).$$

The statement follows using a union bound and the identity $\|M\|_{\infty,2} = \max_k \|M^k\|_2$. ∎

Appendix B. Sensing matrices

Lemma 24 (Thresholding with sensing matrix) *Assume that the signals follow the model in (9), where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix and denote by $H := \Psi^*\Phi - \mathbb{I}$ the hollow cross-Gram matrix. If*

$$\|H\|_{\infty,1}^2 \leq \frac{\|c\|_{\min}^2}{8\|c\|_{\max}^2\log(4K/\varepsilon)}, \qquad \text{and} \qquad \|HW\|_{\infty,2}^2 \leq \frac{\|c\|_{\min}^2}{8e^2\|c\|_{\max}^2\log(4K/\varepsilon)},$$

then thresholding with sensing dictionary Ψ recovers the support with probability at least $1-\varepsilon$.

Proof Now by definition of the algorithm, thresholding recovers the full support if

$$\|\Psi_{I^c}^* y\|_{\max} < \|\Psi_I^* y\|_{\min}.$$

Repeating the steps from the proof of Theorem 10 with the obvious changes we obtain the result. ∎

Lemma 25 (OMP with sensing matrix) *Assume that the signals follow the model in (9), where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let W denote the corresponding weight matrix. Let Ψ be a sensing matrix and assume the hollow Gram-matrix $H = \Phi^*\Phi - \mathbb{I}$ satisfies $\|WHW\|_{2,2} \leq \frac{1}{4e^2}$. If*

$$\begin{aligned}
\|HW\|_{\infty,2}^2 &\leq \frac{1}{16e^2\log(216K/\varepsilon)} \\
\|H\|_{\infty,1} &\leq \frac{1}{4\log(218K/\varepsilon)} \\
\|(\Psi^*\Phi - \mathbb{I})W\|_{\infty,2}^2 &\leq \min_{L\subseteq\{1,\ldots,S\}} \frac{\|c_L\|_\infty^2}{16e^2\|c_L\|_2^2} \\
\|\Psi^*\Phi - \mathbb{I}\|_{\infty,1} &\leq \min_{L\subseteq\{1,\ldots,S\}} \frac{\|c_L\|_\infty}{4\|c_L\|_2\sqrt{\log(218K/\varepsilon)}},
\end{aligned}$$

then OMP with sensing matrix Ψ recovers the correct support with probability at least $1-\varepsilon$.

Proof Set $L := I \setminus J$. By definition, OMP finds another correct atom in the next step if

$$\|\Psi_{I^c}^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty < \|\Psi_L^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty.$$

Repeating the steps from the proof of Theorem 11 with the obvious changes we obtain the result. ■

References

[1] B. Adcock, A.C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. *Forum of Mathematics, Sigma*, 5:e4, 2017. doi: 10.1017/fms.2016.32.

[2] J. Bourgain and L. Tzafriri. Invertibility of "large" submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel J. Math*, 57(2):137–224, 1987.

[3] C. Boyer, P. Weiss, and J. Bigot. An algorithm for variable density sampling with block-constrained acquisition. *SIAM Journal on Imaging Sciences [electronic only]*, 7, 10 2013. doi: 10.1137/130941560.

[4] E. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37:2145–2177, 2009.

[5] E. Candès and Y. Plan. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011. doi: 10.1109/TIT.2011.2161794.

[6] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[7] N. Chauffert, P. Ciuciu, J. Kahn, and P. Weiss. Variable density sampling with continuous trajectories. *SIAM Journal on Imaging Sciences*, 7, 11 2013. doi: 10.1137/130946642.

[8] S. Chrétien and S. Darses. Invertibility of random submatrices via tail-decoupling and matrix Chernoff inequality. *Statistics and Probability Letters*, 82:1479–1487, 2012.

[9] V.H. De la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer New York, 1999.

[10] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[11] J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50:1341–1344, 2004.

[12] J. Hajek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

[13] K. Jogdeo and S. M. Samuels. Monotone convergence of binomial probabilities and a generalization of ramanujan's equation. *Ann. Math. Statist.*, 39:1191–1195, 1968.

[14] F. Krahmer and R. Ward. Stable and robust sampling strategies for compressive imaging. *arXiv:1210.2380*, 2012.

[15] P. Kuppinger, G. Durisi, and H. Bolcskei. Uncertainty relations and sparse signal recovery for pairs of general signal sets. *IEEE Transactions on Information Theory*, 58:263–277, 2012.

[16] G. Puy, P. Vandergheynst, and Y. Wiaux. On variable density compressive sampling. *IEEE Signal Processing Letters*, 18(10):595–598, 2011. doi: 10.1109/LSP.2011.2163712.

[17] P. Randall. *Sparse recovery via convex optimization.* Ph.d. thesis, n.4349, California Institute of Technology, 2009.

[18] K. Schnass. Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation. *IEEE Signal Processing Letters*, 25(12):1865–1869, 2018.

[19] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, 2007.

[20] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Transactions on Signal Processing*, 56(5):1994–2002, 2008.

[21] J. Tropp. Recovery of short, complex linear combinations via l1 minimization. *IEEE Transactions on Information Theory*, 51:1568–1570, 2005.

[22] J. Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1-24), 2008.

[23] J. Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus Mathematique*, 346:1271–1274, 2008.

[24] J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.