Let's not play dice in sampling

Baran Gözcü
LIONS, I&C, EPFL

***Abstract*—This report presents and investigates the problem of optimal subsampling given the sparsity and measurement bases in a compressive sensing setting. Starting with the restricted isometry property which is known to be a strong condition, we describe the coherence barrier and explain how this barrier is broken with variable density subsampling. For this approach we also present and interpret a recently proven theoretical guarantee. We then describe a deterministic sample selection technique using partial Hadamard matrices. Finally as a blend of some discussed approaches, we discuss a block diagonal measurement matrix design before giving a research proposal.**

***Index Terms*—compressive sensing, variable density sampling, RIP, deterministic sampling, measurement matrix design**

I. Introduction

COMPRESSIVE sensing has been a breakthrough in information acquisition in the past decade. Instead of first acquiring a signal and compressing it, this technique proposes to make a few projections onto a lower dimensional space. Using some prior about the signal, e.g. sparsity, the signal can then be reconstructed with high probability and with small or zero error. The advantage of compressive sensing is that when the measurements are expensive as in the case of medical imaging, this technique can greatly help in reducing number of necessary samples.

Proposal submitted to committee: August 31st, 2014; Candidacy exam date: September 5th, 2014; Candidacy exam committee: Emre Telatar, Volkan Cevher, Pierre Vandergheynst.

This research plan has been approved:

Date: _____

Doctoral candidate: _____
(name and signature)

Thesis director: _____
(name and signature)

Thesis co-director: _____
(if applicable) (name and signature)

Doct. prog. director: _____
(B. Falsafi) (signature)

EDIC-ru/05.05.2009

A. Problem Model

Let $\mathbf{x} = \Psi\boldsymbol{\alpha}$ be a $N \times 1$ vector that is s-sparse in an orthonormal basis $\Psi = [\boldsymbol{\psi_1}, \boldsymbol{\psi_2}, ...\boldsymbol{\psi_N}]$ with a sparse coefficient vector $\boldsymbol{\alpha}$. Denoting the $m \times N$ subsampling operator as P, the measurement basis as $\Phi = [\boldsymbol{\phi_1}, \boldsymbol{\phi_2}, ...\boldsymbol{\phi_N}]$, the $m \times N$ measurement matrix as $A = P\Phi^*$ and the orthogonal operator $U = \Phi^*\Psi$, we obtain the $m \times N$ measurement vector $\mathbf{y}$

$$\mathbf{y} = A\mathbf{x} = P\Phi^*\mathbf{x} = P\Phi^*\Psi\boldsymbol{\alpha} = PU\boldsymbol{\alpha} \tag{1}$$

by projecting the vector $\mathbf{x}$ onto the rowspace of A.

There are various algorithms to recover $\mathbf{x}$ from $\mathbf{y}$. In this write-up we will consider ℓ_1 norm minimization which substitutes ℓ_0 norm minimization which is not tractable. However the matrix A cannot be arbitrary and it needs to satisfy certain properties.

B. Restricted Isometry Property

Among others, a common approach to asses a measurement matrix A is so-called restricted isometry property (RIP) which is defined as follows. A matrix A has RIP-s property with RIP constant δ_s if

$$(1 - \delta_s)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 + \delta_s)\|\mathbf{x}\|_2^2 \tag{2}$$

for all s-sparse vectors $\mathbf{x}$. Similar to Johnson-Lindenstrauss Lemma (JL) which states that we can reduce the dimensionality of a group of vectors by preserving the pairwise distances using a Lipschitz mapping, RIP property requires a dimensionality reducing linear embedding to preserve the norms of all s sparse vectors. Once we have this property for the matrix $A\Psi$, we can solve the basis pursuit

$$\hat{\boldsymbol{\alpha}} = \underset{A\Psi\boldsymbol{\beta}=\mathbf{y}}{\arg\min}\|\boldsymbol{\beta}\|_1$$

and find an estimation of $\mathbf{x}$ as $\hat{\mathbf{x}} = \Psi\hat{\boldsymbol{\alpha}}$. The error guarantees are given in terms of the RIP constant. For example, a recent result states that if $A\Psi$ has RIP-$2s$ with $\delta_{2s} < 4/\sqrt{41}$ then we have the following error guarantee:

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \frac{C\sigma_s(\mathbf{x})_{\ell_1}}{\sqrt{s}} \tag{3}$$

where C is a constant that depends on only δ_{2s} and $\sigma_s(\mathbf{x})$ is ℓ_1 error of the best s-term approximation of $\mathbf{x}$ in the basis Ψ [1]. Since we assume that $\mathbf{x}$ is s-sparse in this basis, we have unique recovery. There are different algorithms and given the RIP constant and a desired error bound, we can choose a proper algorithm.

Certain classes of measurement matrices are known to satisfy RIP property. In Section II-A we are going to present

a simple proof that Gaussian matrices satisfy this property with high probability when $m = \mathcal{O}(s\log(N/s))$ [2]. In cases where a signal is not sparse in canonical basis, but sparse in e.g. wavelet basis, Gaussian matrices with this basis will still satisfy RIP property. This is called *universality*.

In many practical applications, we don't have the possibility to choose the matrix, the measurements are dictated by the physical structure of the device, e.g. in MRI, it is Fourier transform. In a discretized model, Φ is therefore the FFT matrix and signals are sparse in a wavelet basis Ψ. In these situations the matrix P needs to sample at least

$$m = \mathcal{O}(s\log^4(N)) \tag{4}$$

elements of the orthogonal matrix U for RIP to hold w.h.p. [3].

C. Coherence

Although RIP has been a practical tool and a strong machinery to prove many results in compressive sensing (CS), it is known to be a strong condition for unique recovery. Moreover it is not always simple to verify the RIP property of a given matrix. A so-called RIP*less* theory of compressive sensing based on the coherence of a matrix constitutes a major alternative to RIP [4]. For a matrix U, the coherence $\mu(U)$ is defined as

$$\mu(U) = \max_{i,j} |U_{ij}|^2 \in [1/N, 1]$$

The number of uniformly taken measurements for a unique recovery w.h.p. via basis pursuit is given as

$$m = \mathcal{O}\left(\mu(U) N s \log(N)\right) \tag{5}$$

which is a considerable reduction in number of measurements compared to (4) if the coherence is around its minimum value $1/N$ which is the case when for example U is the FFT matrix. Note that this result is non-uniform, i.e. for a fixed $\mathbf{x}$, m uniformly taken samples of U can recover that fixed $\mathbf{x}$ unlike the stronger guarantee given by RIP (4) which is uniform, i.e. valid for all $\mathbf{x}$. This will be further explained in Section II-A. However this is not a practical drawback. The bound (5) suggests that we are limited by the coherence of the isometry U. For example Fourier and wavelet bases are highly coherent, i.e. the multiplication of their matrices $U = \Phi\Psi$ has a high coherence, $\mu(U) = \mathcal{O}(1)$ as N goes to infinity. This would require $m = \mathcal{O}\left(sN\log(N)\right)$ samples taken with a uniform distribution. This is called coherence barrier and addressed in [5] which will be presented in Section II-B.

In practice it has been shown that a variable density subsampling reduces the number of required measurements and gives a good reconstruction quality which would not be possible if the same reduced number of samples were taken uniformly [6]. However the density functions used are empirical and even if they perform well, we don't know if they are optimum, i.e. whether or not they give theoretically the best reconstruction quality with a minimum number of samples. This is the general theme of this write-up which presents different approaches to this problem and gives a research proposal.

There are two important factors that affect optimal subsampling density. The first one is coherence and various approaches considered minimizing the coherence of U. In [7], the authors formulated the optimum density as the solution of a convex optimization problem. This approach can also incorporate support knowledge of a signal and closely matches the performance of empirical approaches existing in the literature.

Another important factor is the distribution of nonzero elements in the signal, i.e. the structure of the sparsity which will be explained in II-B where we present a so-called *flip test* to show the importance of this factor. The authors of [5] argue that both factors must be taken into account, minimizing the coherence is not sufficient alone. They make an extensive performance analysis in order to give robust theoretical guarantees in terms of coherences and sparsities at each level of the signal support. However despite providing recovery guarantees, they keep using empirical distributions without presenting a justification of them based on their theory.

In Section II-C we will give an information theoretic view of subsampling problem in compressive sensing. Inspired by polar coding, [8] establishes a good connection and analogy between the two fields and proposes a deterministic subsampling scheme.

In the last section we will present some preliminary work, give conclusions and future directions.

D. Notations

We show vectors by bold characters, random vectors with bold capital letters and matrices by capital letters. e denotes the Euler's number.

II. Survey of the Selected Papers

A. A simple proof of RIP property [2]

1) Background:

As explained above, RIP property of measurement matrices allows us to give error guarantees for ℓ_1 minimization. Gaussian matrices are known to have RIP property with smaller number of rows compared to any other type of matrices, they are therefore optimal in this sense.

In [2] a proof of RIP property based on a concentration inequality that Gaussian or other distributions such as Bernoulli distribution satisfy is given. The proof makes use of covering numbers in finite dimensional space to pass from the concentration inequality to the RIP condition. As the same concentration inequalities have also been used to provide new proofs of Johnson-Lindenstrauss Lemma, this brings out a connection between JL lemma and RIP property. It is also shown in [2] that the minimum number of rows that Gaussian matrices need to have in order to achieve the RIP property is indeed optimum.

2) Results:

A) RIP property for Gaussian matrices

Theorem 1. *(Thm 5.2. in [2]) Let the random matrix $A \in \mathbb{R}^{m\times N}$ satisfy the following concentration inequality. For each given vector $\mathbf{x} \in \mathbb{R}^N$,*

$$Pr\left(\left|\|A\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \geq \delta\|\mathbf{x}\|_2^2\right) \leq 2e^{-c_\delta m} \tag{6}$$

where c_δ is a positive constant that only depends on $\delta \in (0,1)$. Then for a given $\delta \in (0,1), \exists\ \tilde{c}_\delta, \hat{c}_\delta > 0$ such that the matrix A satisfies RIP of order-s with a failing probability less than $2e^{-\hat{c}m}$ whenever $m \geq \tilde{c}s\log(N/s)$.

Proof Overview:

Note that the concentration inequality (6) is equivalent to stating that for each given $\mathbf{x}$

$$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1+\delta)\|\mathbf{x}\|_2^2 \tag{7}$$

with failing probability less than or equal to $2e^{-c_\delta m}$. This is very similar in form to RIP condition (2). Now we are going to describe how to go from (7) to (2). They indeed share the same inequality, but the statements have a crucial difference. Inequality (7) holds *for each* fixed $\mathbf{x}$ whereas the definition of RIP requires the inequality (2) to hold *for all* s-sparse $\mathbf{x}$, which is less likely to happen compared to the former if it was not restricted to s-sparse vectors. The theorem says that this restriction of the statement to s-sparse vectors gives the same order of failing probability as the concentration inequality if m is large enough.

Usually, they way to go from (7) to (2) is simply to apply a union bound on failing probabilities. Indeed if we were to prove the JL lemma instead, using the concentration inequality (7), this would be straightforward. Given a set T of t vectors, applying a union bound over $\binom{t}{2}$ pairwise distance vectors would reveal that $\mathcal{O}(\log(t)/\delta^2)$ samples are sufficient to preserve the pairwise distances

$$(1-\delta)\|\mathbf{u}-\mathbf{v}\|_2^2 \leq \|A\mathbf{u}-A\mathbf{v}\|_2^2 \leq (1+\delta)\|\mathbf{u}-\mathbf{v}\|_2^2$$

for all $\mathbf{u}, \mathbf{v} \in T$. However (2) has to hold for all and *infinitely many* s-sparse $\mathbf{x}$ which prevents us to directly apply a union bound that would result in an infinite sum. But if we could cover all $s-$sparse vectors with finitely many elements up to a distortion parameter, we could still apply the union bound.

This is done in two steps. First we fix a support set S of size s and consider only unit vectors $\mathbf{x}$ with the support S, since A is linear. It can be shown that there exists a finite set of maximum $(12/\delta)^s$ unit vectors with same support S such that for every unit vector $\mathbf{x}$ having the support S, there exists a unit vector in the finite set at a distance less than $\delta/4$ to $\mathbf{x}$ [9]. Applying a union bound on this finite set and using this upper bound on the distance, $\delta/4$, an inequality similar to (7) is proven to hold for all $\mathbf{x}$ with the fixed support S with failing probability less than $2(12/\delta)^s e^{-c_{\delta/2}m}$.

This was for all vectors $\mathbf{x}$ in the fixed support S, so the problem of having infinitely many vectors is resolved. In the second step, we apply again a union bound, this time over $\binom{n}{s} \leq (eN/s)^s$ different possible supports of size s and after some algebra we obtain the condition $m > \tilde{c}s\log(N/s)$ for (7) to hold for all s-sparse $\mathbf{x}$.

B) Optimality of Gaussian matrices

The first part of [2] gives the following bound after a discussion on Gelfand width of a unit ℓ_1 ball, $U(\ell_1^N)$: $\exists c_0 > 0$ such that $\forall m \in \{1,2,...N\}$

$$\frac{1}{c_0}\sqrt{\frac{\log(eN/m)}{m}} \leq E_{m,N}(U(\ell_1^N))_{\ell_2} \leq 2c_0\sqrt{\frac{\log(eN/m)}{m}} \tag{8}$$

where

$$E_{m,N}(U(\ell_1^N))_{\ell_2} = \inf_{(A,\Delta)} \sup_{x\in U(\ell_1^N)} \|x - \Delta(A\mathbf{x})\|_{\ell_2}$$

is the worst case error of the best possible encoder-decoder pair (A,Δ) which first reduces the dimensionality of a signal $\mathbf{x} \in U(\ell_1^N)$ to m and then maps it back to the dimension N. From this we can see the optimality of Gaussian measurement matrices as follows. Taking the supremum of (3) over $U(\ell_1^N)$ and considering the lower bound in (8) we have,

$$\begin{aligned}\frac{1}{c_0}\sqrt{\frac{\log(eN/m)}{m}} &\leq \sup_{x\in U(\ell_1^N)} \|x - \Delta(A\mathbf{x})\|_{\ell_2} \\ &\leq \sup_{x\in U(\ell_1^N)} \frac{C}{\sqrt{s}}\sigma_s(x)_{\ell_1} \leq \frac{C}{\sqrt{s}}\end{aligned}$$

which implies that the number of measurements m is bounded below by

$$m \geq c_2 s\log(eN/m) \tag{9}$$

It has been stated in Theorem 1 that a measurement matrix $A \in \mathbb{R}^{m\times N}$ satisfying the concentration inequality (6) has RIP of order-s w.h.p. if $m \geq \tilde{c}s\log(N/s)$. This condition can be rephrased using the following lemma.

Lemma 1. *(Lemma C6-c in [3]) Let the integers $m, N, s \geq 1$ with $N \geq s$ be given and let $c, d > 0$ be constants. Then we have*

$$m \geq cs\log\left(\frac{dN}{m}\right) \implies m \geq c's\log\left(\frac{dN}{s}\right), c' = \frac{ec}{e+c} \tag{10}$$

Now choosing $d = e$, we can conclude that Gaussian matrices satisfy RIP-s property w.h.p. if $m \geq c''s\log(N/m)$ with $c'' = (\tilde{c}e)/(\tilde{c}-e)$. Note that this is a stronger condition due to direction of the implication in the lemma, but we needed this in order to be able to compare with (9). We now see that the number of measurements that Gaussian matrices require for RIP to hold is optimum in the sense that it achieves the lower bound (9) up to a constant.

3) Discussion:

RIP condition is a strong requirement for signal reconstruction using ℓ_1 minimization. As it has been mentioned above, it requires far more than necessary number of measurements and it does not take the sparsity structure into account. It cannot help us alone for a solution to the optimal subsampling problem, however the basic tools and analysis used in the proof could constitute a good base in the future research.

B. Breaking the coherence barrier: A new theory for compressed sensing [5]

1) Background:

In [5] a CS setting where sparsity and measurement bases are imposed by the physical setup is considered. It is known that due to the coherence problem between two orthogonal bases, a considerable reduction in number of measurements is not possible with uniform sampling (Figure-1c,d). If the number of measurements are reduced, then we are not anymore in guaranteed region and the quality of reconstruction deteriorates (Figure-1e,f). However if we apply the variable density sampling used in [5] with the same number of measurements, we obtain a reasonable reconstruction. (Figure-1g,h).

In order to explain the success of variable density sampling, the authors propose new concepts. *Asymptotic incoherence* refers to the phenomena of matrix entries approaching to zero in some regions (Figure-1a) even if some regions preserve a high amplitude. This structure of matrix U can be exploited using a so called *multilevel sampling* to adapt different sampling rates to different coherence regions. (Definition-2 below).

However the authors argue that minimizing the overall coherence is not sufficient and the sparsity structure, i.e. where the nonzero elements are located, also makes a difference. This is shown via a so-called Flip Test. The authors take a signal $\mathbf{x}$ and by flipping its coefficients in its sparsity basis they obtain a new signal, $\mathbf{x}_{fp}$, which is measured using the same subsampling mapping as $\mathbf{x}$ (Figure–1i), reconstructed with basis pursuit and flipped again (Figure–1j). It is seen that a subsampling map that is good for a particular sparsity structure is not working for a structure in which the nonzero elements are relocated. They conclude that the optimal subsampling map depends on the sparsity structure as well and cannot be explained by only coherence and total sparsity which both signals share. It should be emphasized that this is not a contradiction with RIP, since in both cases (Figure–1g,j) we are out of guaranteed regions given by RIP but in the first one we choose a subsampling map that allows us to operate under the allowed regime of RIP or the universal coherence.

The problem that [5] adresses is precisely to provide a theoretical guarantee for variable density subsampling given a measurement scheme and sparsity levels of the signal. Note that this doesn't mean to find the optimum theoretical subsampling map, indeed the authors use also empirical distributions [5], [10].

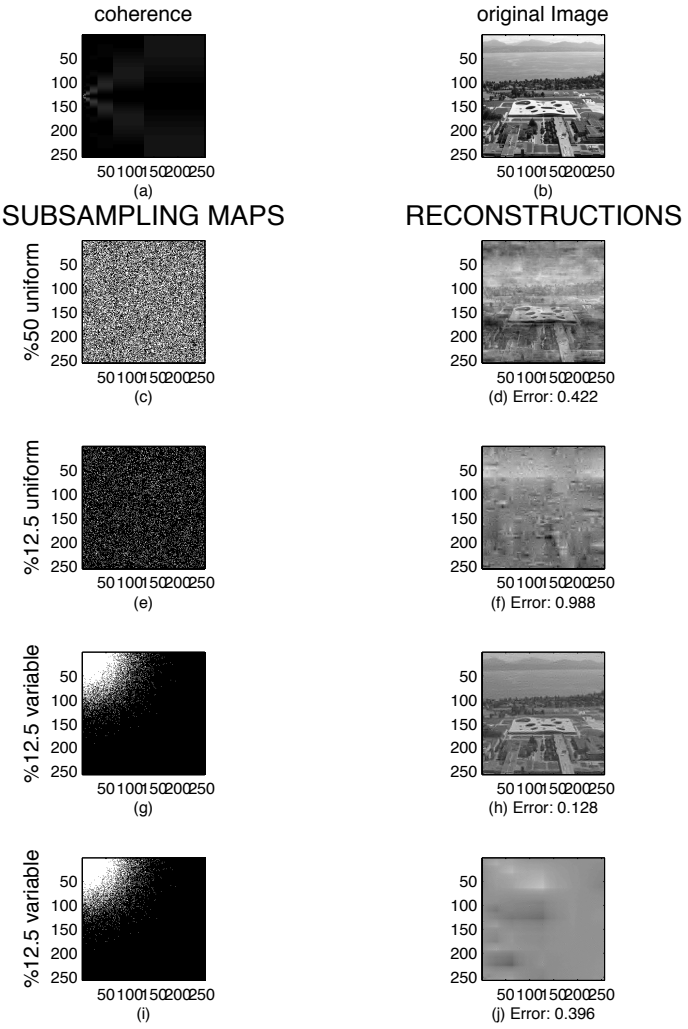

Fig. 1. (a) Coherence of Daubechies 4 wavelet and Fourier basis in 1D (b) Original image to be measured with a partial Hadamard matrix (c-f) Subsampling mappings and reconstruction errors (i-j) Flip Test

2) Results:

Before stating the main result, we need to give the following definitions.

Definition 1. *(Multilevel sparsity) Let* $\mathbf{N} = (N_1, N_2, ...N_r)$ *be a partition of* N *into* r *positive integers, i.e.* $N = \sum_{l=1}^{r} N_l$. *This corresponds to dividing the length of signal* $\boldsymbol{\alpha} \in \mathbb{C}^N$ *into* r *consecutive levels of size* N_l. *Assume that we know the sparsity of the signal* $\boldsymbol{\alpha}$ *at each level* $l \in \{1...r\}$, $s_l \leq N_l$, *in addition to our knowledge of total sparsity* $s = \sum_{l=1}^{r} s_l$. *We say that* $\boldsymbol{\alpha}$ *is* $(\mathbf{s}, \mathbf{N})$ *sparse,* $\mathbf{s} = (s_1, s_2....s_r)$, *if the sparsity of the signal at each level* l *is equal to* s_l. *We denote the set of* $(\mathbf{s}, \mathbf{N})$ *signals by* $\Sigma_{\mathbf{s},\mathbf{N}}$.

Definition 2. *(Multilevel sampling) Let* $\tilde{\mathbf{N}} = (\tilde{N}_1, \tilde{N}_2, ...\tilde{N}_r)$ *be another partition of* N *into* r *positive integers, i.e.* $N = \sum_{k=1}^{r} \tilde{N}_k$. *Let also* $\mathbf{m} = (m_1, m_2, ...m_r)$ *with* $m_k \leq \tilde{N}_k \; \forall k \in \{1...r\}$ *be a partition of total number of measurements,* m, *into numbers of measurements at each level*

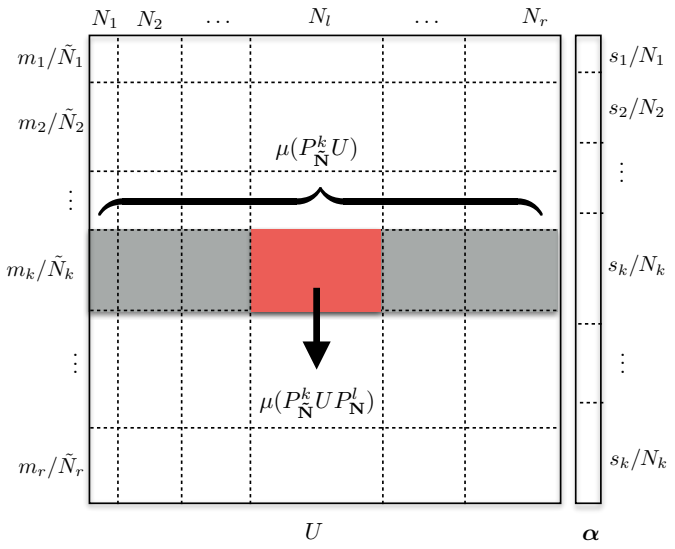

Fig. 2. Multilevel sparsity, multilevel sampling and local coherence

k. We refer to this scheme of uniformly sampling m_k rows of $\tilde{N}_k$ rows at each level-k a $(\mathbf{m}, \tilde{\mathbf{N}})$ multilevel sampling scheme.

Definition 3. *(Local Coherence) For given partitions $\mathbf{N}$ and $\tilde{\mathbf{N}}$ of N and an isometry $U \in \mathbb{C}^{N\times N}$ the local coherence of U is given by*

$$\mu_{\tilde{\mathbf{N}},\mathbf{N}}(k,l) = \sqrt{\mu(P^k_{\tilde{\mathbf{N}}}UP^l_{\mathbf{N}}) \cdot \mu(P^k_{\tilde{\mathbf{N}}}U)}, \quad k,l = 1,...,r$$

where $P^k_{\tilde{\mathbf{N}}}$ denotes the projection matrix corresponding to indices $1 + \sum_{i=1}^{k-1} \tilde{N}_i ..... \sum_{i=1}^{k} \tilde{N}_i$.

These first three definitions are shown in Figure-2 for illustrative purposes as they are central to the understanding of the main result. The following two definitions are also necessary for the statement of Theorem-2.

Definition 4. *$((\mathbf{s},\mathbf{N}) - $ term approximation error$)$ Let $\mathbf{x} = \sum_i \alpha_j \psi_j$ be a signal in $\mathbb{C}^N$. The $(\mathbf{s},\mathbf{N})$-term approximation error of $\mathbf{x}$ in the orthonormal basis Ψ is defined as*

$$\sigma_{\mathbf{s},\mathbf{N}}(x) = \min_{\boldsymbol{\beta}\in\Sigma_{\mathbf{s},\mathbf{N}}} \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1.$$

Definition 5. *(Relative Sparsity) For a given multilevel sparsity $(\mathbf{s},\mathbf{N})$, an isometry $U \in \mathbb{C}^{N\times N}$ and a partition of N, $\tilde{\mathbf{N}} = (\tilde{N}_1, \tilde{N}_2, ...\tilde{N}_r)$, the k^{th} relative sparsity is given by $\tilde{s}_k = \tilde{s}_k(\tilde{\mathbf{N}},\mathbf{N},\mathbf{s}) = \max_{\beta\in\Theta} \|P^k_{\tilde{\mathbf{N}}}U\beta\|^2$ and Θ is the set given by*

$$\Theta = \{\beta : \|\beta\|_\infty \leq 1, |supp(P^l_{\mathbf{N}}\beta)| = s_l, l = 1,...,r\}.$$

The main theorem is then as follows.

Theorem 2. *(Thm 4.4. in [5])*

Let a signal $\mathbf{x} = \Psi\boldsymbol{\alpha}$ be measured as $\mathbf{y} = PU\boldsymbol{\alpha}$ where U is an isometry. Let P sample the rows of U according to $(\mathbf{s},\mathbf{N})$. Also let $(\mathbf{m},\tilde{\mathbf{N}})$ satisfy the following: $\forall\epsilon \in (0, 1/e]$ and $\forall k \in 1,...r$

$$m_k \geq \tilde{N}_k \cdot \log(1/e) \left(\sum_{l=1}^{r} \mu_{\tilde{\mathbf{N}},\mathbf{N}}(k,l) \cdot s_l \right) \cdot \log(N) \quad (11)$$

with $m_k \geq \hat{m}_k \cdot \log(1/e) \cdot \log(N)$ and $\hat{m}_k$ satisfies

$$1 \geq \sum_{l=1}^{r} \left(\frac{\tilde{N}_k}{\hat{m}_k} - 1 \right) \cdot \mu_{\tilde{\mathbf{N}},\mathbf{N}}(k,l) \cdot \hat{s}_l \quad (12)$$

for all $l = 1,...,r$ and for all $\hat{s}_1, \hat{s}_2, ..., \hat{s}_r \in (0,\infty)$ with

$$\hat{s}_1 + \hat{s}_2 + ... + \hat{s}_r \leq s_1 + s_2 + ... + s_r, \quad \hat{s}_k \leq \tilde{s}_k(\tilde{\mathbf{N}},\mathbf{N},\mathbf{s}).$$

Suppose now that $\hat{\boldsymbol{\alpha}}$ is a minimizer of

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \ s.t. \ PU\boldsymbol{\beta} = \mathbf{y}$$

Then with probability at least $1 - s\epsilon$, we have the following error bound for some constant C:

$$\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2 \leq C\sigma_{\mathbf{s},\mathbf{N}}(\mathbf{x}).$$

The theorem seems to be complicated but we can interpret it as follows. In contrast with (5) where an error guarantee is given in terms of global coherence and global sparsity which results in a high number of measurements, the number of measurements m_k at each level k now depends mostly on the local sparsity s_k and local coherences $\mu_{\tilde{\mathbf{N}},\mathbf{N}}(k,l)$ as seen in condition (11). Note that the local coherences of other levels do not completely disappear, i.e. U is not completely block diagonal, that is why there is an interdependency between different levels. In the case of Fourier measurements and wavelet sparsity, the effect of other levels to level-k is shown to exponentially decrease as we are further away from that level, therefore U becomes nearly block diagonal and this shows that Fourier and wavelet bases are asymptotically incoherent even if they are globally coherent. This gives a theoretical explanation of using variable density sampling that breaks the coherence barrier.

In [5] the theorem is also given for infinite dimensional setting in Hilbert space and the finite dimensional case is derived as a special case of this. The proof is extensive and requires many tools from probability theory. The infinite dimensional setting allows an analog compressed sensing in which one can overcome the errors arising due to discretization.

3) Discussion:

The theoretical guarantee that [5] provided in attempt to explain the success of variable density subsampling is an important contribution. However [5] does not asses the empirical density schemes that are present in the literature using Theorem-2 nor proposes an optimum subsampling scheme as done in [7]. This would necessitate the formulation of an optimization problem in order to minimize the quantities on the left side of (11) and (12). This approach can potentially give an ultimate answer to the problem of optimum subsampling if we maintain a good statistics of sparsity levels for certain signals. Another shortcoming of [5] is the complexity of the conditions in the theorem which naturally raises the question whether or not they could be made more user-friendly. Finally one wonders if it is possible to make the analysis without imposing

the same number r of multivariate sampling and sparsity levels given that this is not the case in empirical subsampling schemes in use. This would facilitate the evaluation of the empirical subsampling schemes.

C. Adaptive sensing using deterministic partial Hadamard matrices [8]

1) Background:

In [8], the authors consider the problem of finding deterministic measurement matrices that can well preserve the entropy of random vectors. Therefore a probability distribution over the input signal is assumed. The main result is that a family of deterministically truncated Hadamard matrices, preserves entropy of a random vector of i.i.d discrete random variables with measurement rate approaching to zero as the problem dimension increases.

2) Results :

Theorem 3. *Let* $J_N = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}^{\otimes n}$, $N = 2^n$, *n=1,2... be a family of matrices. Let* $\mathbf{Y} = J_N\mathbf{X}$ *be a linear transformation of the input random vector* $\mathbf{X} = [X_1 X_2 .... X_N]$ *of* N *i.i.d. discrete random variables with probability distribution* p_x *supported over a subset of* $\mathbb{Z}$. *Define* $H_i = H\left(Y_i | Y_0, Y_1 ... Y_{i-1}\right)$ *where* H *is the entropy function. For a given* ϵ, *also define* A_N *to be matrix whose rows are the rows of* J_N *with indices* i *for which* $H_i > \epsilon$. *Then the measurement vector* $A_N\mathbf{X}_N$ *and measurement rate satisfy*

$$\frac{H(\mathbf{X}_N | A_N\mathbf{X}_N)}{N} < \epsilon \quad \text{and} \quad \limsup_{N\to\infty} \frac{m_N}{N} = 0 \tag{13}$$

In parallel with RIP property regarding the preservation of the norm of a deterministic vector, the family of matrices that satisfies (13) are said to be satisfying a property called restricted iso-entropy property (REP) with a certain measurement rate which is, in the case of Theorem 3, asymptotically zero.

Proof Overview:

The proof is based on source polarization and uses a martingale convergence theorem [11]. The main difference is that in compressive sensing, signals are in real field therefore the binary addition is replaced by addition in real field. The result is also different as stated in Theorem 3: the rate of significant indices, $H_i > \epsilon$, goes to zero unlike the source polarization where the rate goes to the entropy of the source.

We will briefly describe the martingale construction. Consider at time n, a Hadamard transformation of size $N = 2^n, \mathbf{Y} = J_N\mathbf{X}$. For each n we label the measurements with binary labels $[\omega]_n$ of size n, e.g. when $n = 2$ we denote the possible outputs Y_1, Y_2, Y_3, Y_4 with the binary subindices starting from 0: $Y_{00}, Y_{01}, Y_{10}, Y_{11}$. The subindex $[\omega]_n$ denotes first n-bit truncation of an infinite length sequence $\omega \in \Omega = \{0,1\}^\infty$ i.e. $[\omega]_n = \omega_1\omega_2, .., \omega_n$. We assign a uniform probability measure on the space of all infinite binary

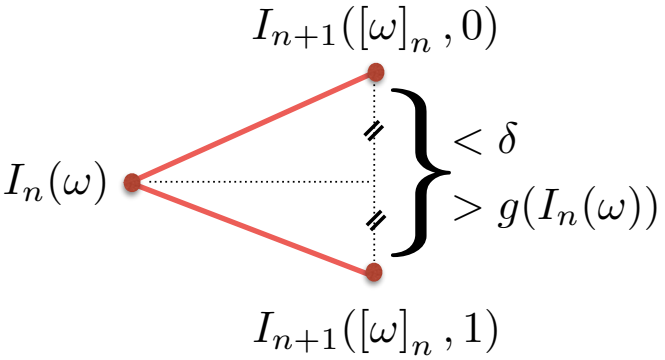

Fig. 3. A branch of the martingale at time n and $n+1$

labelings. On the other hand, a superindex denotes the set of all prior elements, e.g. when $n = 4$ and $\omega = 0.11...$, $Y^{[\omega]_n} = \{Y_1, Y_2, Y_3\}$ .We now define a stochastic process I_n as $I_n = H(Y_{[\omega]_n} | Y^{[\omega]_n})$ and the σ-algebra F_n as the collection of sets

$$\{\omega \in \Omega : \omega_1 = b_1, \omega_2 = b_2, ..., \omega_i = b_i\}$$

for $1 \le i \le n$ and $b_i \in \{0,1\}$. Also we define $F_0 = \{0, \Omega\}$ and $I_0 = H(X)$.

At time $n+1$, in order to obtain I_{n+1}, we can use the recursive structure of the Hadamard matrices. More precisely, let $\mathbf{Y} = J_N\mathbf{X}$, $\hat{\mathbf{Y}} = J_N\hat{\mathbf{X}}$. From the construction of the process I, it immediately results that

$$I_{n+1}([\omega]_n, 0) = H(Y_{[\omega]_n} + \hat{Y}_{[\omega]_n} | Y^{[\omega]_n}, \hat{Y}^{[\omega]_n}) \tag{14}$$

It can be shown that (I_n, F_n) is a martingale, i.e.

$$I_n(w) = E\{I_{n+1} | F_n\} = \frac{1}{2}\left[I_{n+1}([\omega]_n, 0) + I_{n+1}([\omega]_n, 1)\right] \tag{15}$$

Using martingale convergence theorem, we can conclude that I_n converges to a random variable I_∞ almost surely. Therefore for a fixed $\delta > 0$ and for every ω in convergence set, there exists a $n_0(\delta, \omega)$ such that whenever $n \ge n_0$, we have $|I_{n+1}(\omega) - I_n(\omega)| < \delta$, written in Cauchy sequence form. By using the martingale property (15), this is equivalent to

$$|I_{n+1}([\omega]_n, 0) - I_n(\omega)| < \delta \tag{16}$$

whenever $n \ge n_0$ as any of two possible level differences between I_{n+1} with regard to I_n are the same (Figure-3). Now comes to play an entropy power inequality proved in [8] which will give a lower bound on the level difference as a function of I_n.

Theorem 4. *Let* p *be a probability distribution over* $\mathbb{Z}$. *Then*

$$H(p * p) - H(p) \ge g(H(p))$$

where $*$ *denotes convolution,* $g : \mathbb{R}_+ \to \mathbb{R}_+$ *is a strictly increasing function with* $\lim_{x\to\infty} g(x) = \frac{1}{8\log 2}$ *and* $g(x) = 0$ *iff* $x = 0$.

Considering the relation (14), Theorem 4 implies that $|I_{n+1}([\omega]_n, 0) - I_n(\omega)| \geq g(I_n(\omega))$. Using (16), whenever $n \geq n_0$,

$$0 \leq I_n(\omega) < g^{-1}(\delta)$$

which means that $I_n \to I_\infty = 0$ almost surely. This implies that $I_n \to 0$ in probability, therefore

$$\limsup_{N\to\infty} \frac{m_N}{N} = \limsup_{N\to\infty} Pr(I_n > \epsilon) = 0$$

where the first equality is due to the rule according to which the rows are subsampled and the second is due to definition of convergence in probability. This shows that the measurement rate of the family of matrices, $\{A_N\}$, is zero. Showing that it preserves the entropy is relatively simpler and uses a few inequalities involving conditional expectations [8].

An important question to ask is what happens when we do not restrict the signals to be discrete. This is also answered in [8]: For continuous distributions the measurement rate goes to 1 as ϵ tends to 0. The formal statement is as follows.

Lemma 2. *Let p_u be the uniform distribution over $[-1, 1]$ and $Q : [-1, 1] \to \{1, 2, ..q\}$ be a uniform quantizer for the RV X with MMSE less than γ. Assume that $\{A_N\}$ be a family of full rank measurement matrices in $\mathbb{R}^{m_N \times N}$ satisfying*

$$\frac{H(Q(\mathbf{X}_N)|A_N\mathbf{X}_N)}{N} < \epsilon \text{ and } \limsup_{N\to\infty} \frac{m_N}{N} = \rho$$

where Q, when applied to a vector, quantizes each element of the vector. Then $\rho \to 1$ as $\epsilon \to 0$.

This suggests that we cannot do an analog-to-analog compression with an asymptotically zero measurement rate unlike the discrete case above, but there is a trade-off between quantization error and entropy preservation bound.

3) Discussion :

The technique has two attractive features. First, as it is also common in the compressive sensing, the Hadamard structure allows a fast computation like in FFT. Second, the indices corresponding to high entropies in the case of lower nonzero probability are contained in the indices of high entropies in the case of higher nonzero probability. This is called *nested* property and allows one to adaptively refine the reconstruction by adding extra measurements from lower entropy indices if the sparsity level is not known a priori.

The results of [8] could be of high relevance for the problem of optimum subsampling in compressive sensing as it offers an alternative view of this problem. However there are some questions to be answered before making a fair comparison with the existing techniques applied to real signals. For example it should be clarified how entropy preservation could be connected to ℓ_p-norm error bounds in order to give some conventional error guarantees. Another important question is whether or not this method could work with a sparsity basis other than the canonical one.

III. Conclusion

A. Preliminary work

Other than being an important factor for optimal subsampling of an orthogonal matrix, structured sparsity might also help us to design purpose-built measurement matrices. Assuming that we know the sparsity s_l of a signal $\mathbf{x} = \Psi\boldsymbol{\alpha}$ at each wavelet level l in addition to our knowledge of total sparsity $s = \sum_l s_l$, we can design a measurement matrix $A \in \mathbb{R}^{m\times N}$ composed of a block diagonal matrix and the inverse wavelet transform matrix:

$$A = \begin{bmatrix} A_s & 0 & \dots & \dots & \dots & 0 \\ 0 & A_0 & & & & \vdots \\ \vdots & & A_1 & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & A_{L-1} \end{bmatrix} \times \Psi^*$$

where $L = \log_2 N$ is the number of wavelet levels. Each Gaussian submatrix A_l has dimensions $m_l \times N_l$ with $N_l = 2^l$ and A_s is a Gaussian random variable. This matrix is similar to the aforementioned asymptotically incoherent matrix, but instead we don't need to increase the dimension for incoherence, we simply design it block diagonal. At each level l, we then uniformly select $m_l = \tilde{c}_l s_l log(N_l/s_l)$ rows so that RIP is satisfied with probability at least $1 - 2e^{-\hat{c}m_l}$. We have observed and we conjecture that the inequality

$$m = \sum_l m_l = \sum_l \tilde{c}_l s_l \log(N_l/s_l) \leq \tilde{c}s \log(N/s)$$

holds and gives us a substantial gain. We don't have a rigorous proof yet, but the proof is likely to use Jensen's inequality and a union bound over the failing probability of each reconstruction block.

This suggests that the number of measurements required for the block diagonal matrix is less than that of a full Gaussian matrix. This also suggests that with the same amount of measurements we can obtain a better performance.

As shown in the Figure-4, with less number of measurements we can recover the signal with a good performance using our method which is not the case for full Gaussian matrix. Figure-4a,b show the test signal and best-s term approximation ($s = 15$) that is used as an input to the full and block Gaussian measurements. The length of the signal is $N = 2^{15}$. When solved with ℓ_1 minimization, our approach takes about the same time as the full Gaussian matrix does, however it gives a much better reconstruction performance (4c,d). Finally solving for the block systems successively brings a speed-up factor of approximately four and doing this on a parallel architecture further improves this (4-e,f).

B. Research proposal

As discussed above, finding a joint formulation that considers both coherence and sparsity structures to find an optimum

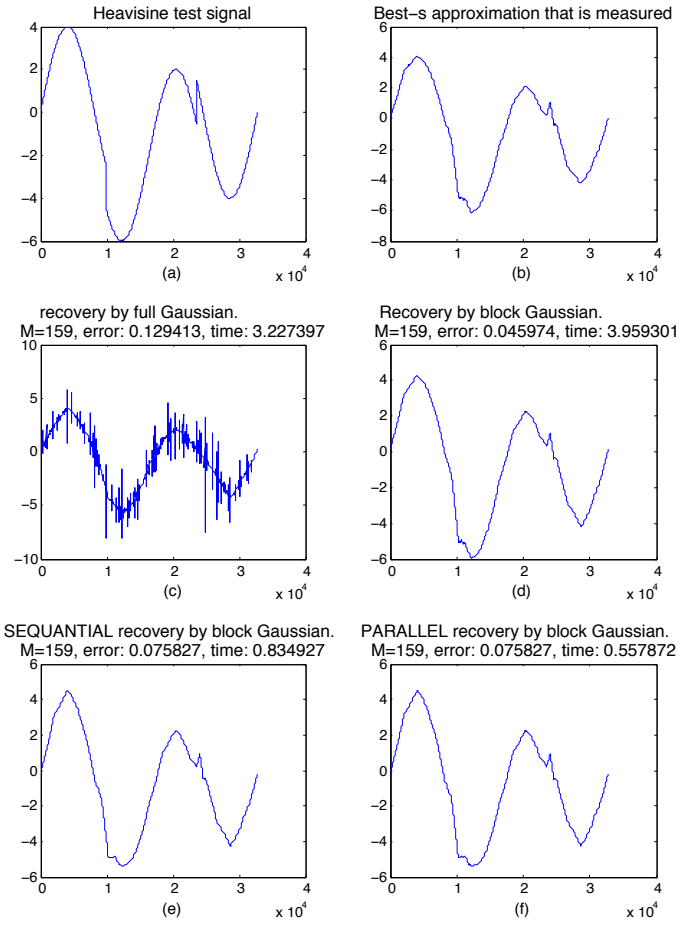

Fig. 4. A block diagonal matrix can be used for a better performance if the sparsity structure is known

variable density subsampling scheme and theoretically analyzing the performance of existing good empirical schemes is an important research problem and constitutes the focus of this research proposal. An answer to this problem could further reduce number of necessary measurements in different signal acquisition settings and give a better understanding of empirical subsampling schemes.

A promising direction is to improve the preliminary work done on designing block diagonal matrices. As a future work, we would like to work on 2D signals. It would be important to understand the limitations of this approach on a signal that is not exactly sparse. A performance investigation of other measurement systems such as expanders or matrices of Bernouilli random variables would be very relevant in terms of applicability. Merged with $e.g.$ Haar wavelet basis, this method using an expander matrix could potentially result in a physical hardware design as it would allow a limited number of different matrix entries.

Another research direction is to investigate whether it is possible to reconcile the deterministic method presented above with the optimal subsampling problem of the compressive sensing applications as an attempt to propose an alternative solution to this problem. The challenging obstacle is that the problems have different settings: the first one assumes a probabilistic distribution on the signal values whereas the conventional compressive sensing works with deterministic signals and uses random matrices for recovery. However the connections provided in [8] motivates a further investigation in this direction.

REFERENCES

[1] J. Andersson and J. Stromberg, "On the theorem of uniform recovery of random sampling matrices," 2014.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[3] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, 2013.

[4] E. J. Candes and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *Information Theory, IEEE Transactions on*, vol. 57, no. 11, pp. 7235–7254, 2011.

[5] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: A new theory for compressed sensing," *arXiv preprint arXiv:1302.0561v3*, 2013.

[6] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic resonance in medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[7] G. Puy, P. Vandergheynst, and Y. Wiaux, "On variable density compressive sampling," *Signal Processing Letters, IEEE*, vol. 18, no. 10, pp. 595–598, 2011.

[8] S. Haghighatshoar, E. Abbe, and E. Telatar, "Adaptive sensing using deterministic partial hadamard matrices," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. Ieee, 2012, pp. 1842–1846.

[9] G. G. Lorentz, M. von Golitschek, and Y. Makovoz, *Constructive approximation: advanced problems*. Springer Berlin, 1996, vol. 304.

[10] B. Roman, A. Hansen, and B. Adcock, "On asymptotic structure in compressed sensing," *arXiv preprint arXiv:1406.4178*, 2014.

[11] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *Information Theory, IEEE Transactions on*, vol. 55, no. 7, pp. 3051–3073, 2009.