Mathematical Methods for Machine Learning and Signal Processing SS 2019

Lectures 8 – 10: Compressed Sensing – Deterministic Results

V. Morgenshtern (compiled from notes by H. Bölcskei)

Agenda:

1. Signal separation problem
2. Spark
3. P0 recovery algorithm
4. Coherence
5. Basis pursuit recovery algorithm
6. Uncertainty principles
7. Square-root bottleneck

1 Signal separation problem

We begin by studying the signal separation problem. Consider the following example in which we are looking at images of neurons and would like to automatically separate the point-like structures from the curve-like structures.

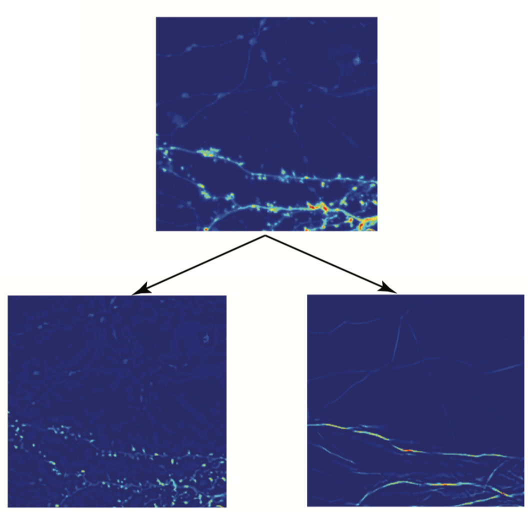


Figure 1: Separation of spines and dendrites, taken from work by Gitta Kutyniok.

There are many problems like this. For example here we are separating the image into cartoon-like part and texture.

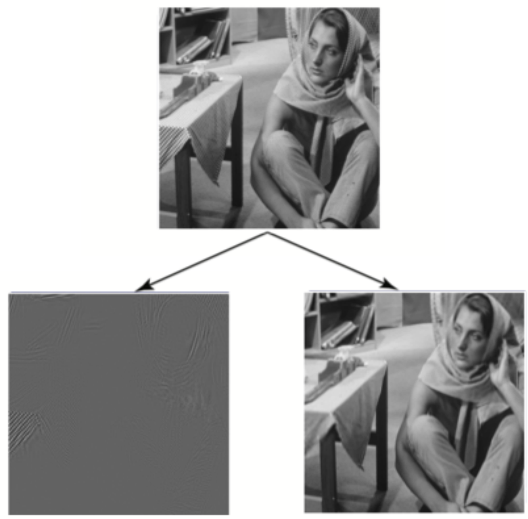


Figure 2: Separation of texture.

Here is one more example [1]:

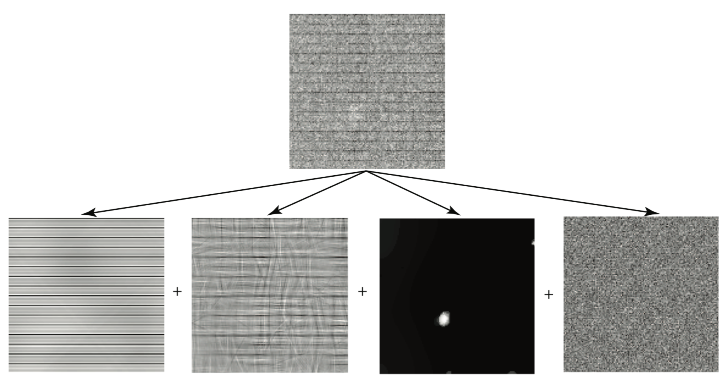


Figure 3: Separation of four structures.

How can we solve such problems?

Consider the example in Figure 1. Let's vectorize the image of point-like structures (bottom left) and collect the pixels into vector $\mathbf{z}_1 \in \mathbb{R}^m$. That point-like signals can be sparsely represented in wavelet domain. This means that we can write

$$\mathbf{y}_1 = \mathbf{A}\mathbf{w} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times m}$ is a square discrete wavelet transform matrix and $\mathbf{w} \in \mathbb{R}^m$ is the vector of wavelet coefficients. Importantly, $\mathbf{w}$ will be very sparse, meaning that most of its component will be zero or close to zero. We will say that $\mathbf{w}$ is s_1-sparse, and write $\|\mathbf{w}\|_0 \leq s_1$.

Definition 1. For any vector $\mathbf{x}$, the quasi-norm $\|\mathbf{x}\|_0$ denotes the number of nonzero entries in $\mathbf{x}$.

Let's vectorize the image of curve-like structures (bottom right) and collect the pixels into vector $\mathbf{z}_2 \in \mathbb{R}^m$. Curve-like structures also admit a sparse representation:

$$\mathbf{y}_2 = \mathbf{B}\mathbf{s} \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{m\times l},\ l > m$, is a frame of Shearlets and $\mathbf{s}$ contains shearlet coefficients of the signal $\mathbf{s}$. We will not study the details of Shearlet construction in this class. It is sufficient to say that Shearlets have been designed to reveal sparsity of curve-like images. Therefore, we assume that $\|\mathbf{s}\|_0 \le s_2$.

The signal separation problem can now be written as:

$$\mathbf{y} = \underbrace{\mathbf{A}\mathbf{x}_1}_{\mathbf{y}_1} + \underbrace{\mathbf{B}\mathbf{e}}_{\mathbf{y}_2} = \underbrace{[\mathbf{A}\ \mathbf{B}]}_{\mathbf{D}} \underbrace{\begin{bmatrix} \mathbf{w} \\ \mathbf{s} \end{bmatrix}}_{\mathbf{x}} = \mathbf{D}\mathbf{x}.$$

If we can recover $\mathbf{x}$ from $\mathbf{y}$, the we can easily obtain $\mathbf{y}_1$ and $\mathbf{y}_2$ via (1) and (2). We know that the sparsity of $\mathbf{x}$ satisfies $s = \|\mathbf{x}\|_0 = s_1 + s_2$.

Without additional assumption on $\mathbf{x}$ we cannot recover $\mathbf{x}$ from $\mathbf{y} = \mathbf{D}\mathbf{x}$, because we have more unknowns than observations. Concretely, $\mathbf{D} \in \mathbb{R}^{m\times n}$ with $n = m + l > m$ is a fat matrix.

Perhaps sparsity of $\mathbf{x}$ could help. For example, if $s \le m$, and if we knew where the nonzero elements of $\mathbf{x}$ are located (knew the support of $\mathbf{x}$), we could just invert the corresponding matrix and recover $\mathbf{x}$. Can we also recover $\mathbf{x}$ without knowing the support of $\mathbf{x}$?

2 Spark

First let's discuss the conditions that are certainly necessary to be able to recover $\mathbf{x}$ from $\mathbf{y}$. When does the dictionary $\mathbf{D}$ admits successful recovery of all s-sparse vectors?

First observe that the set of all s-sparse vector is a union of subspaces in $\mathbb{R}^n$:

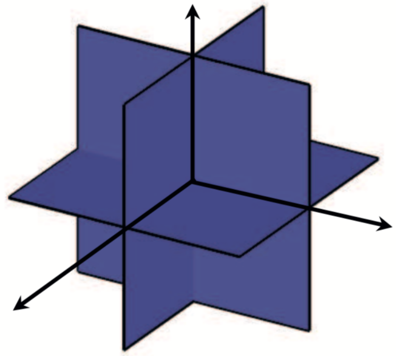


Figure 4: The set of s-sparse vectors

Each subspace corresponds to a distinct sparsity pattern.

The recovery is only possible if all s-sparse vectors are distinguishable:

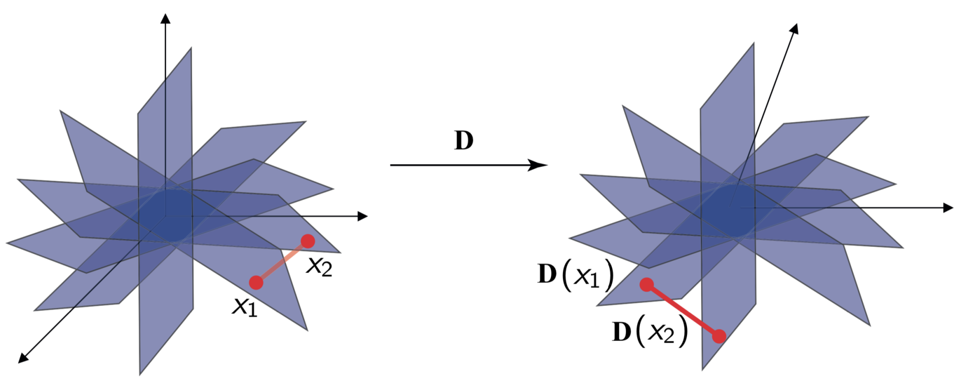


Figure 5: Measurement matrix is injective for s-sparse vectors

Mathematically, for all $\mathbf{x}_1, \mathbf{x}_2$ that are s-sparse with $\mathbf{x}_1 \neq \mathbf{x}_2$

$$\|\mathbf{D}\left(\mathbf{x}_1 - \mathbf{x}_2\right)\|_2^2 > 0.$$

Hence, all collections of $2s$ columns of $\mathbf{D}$ have to be linearly independent. Clearly, this is possible only if $m \geq 2s$.

Definition 2. The spark of a matrix $\mathbf{A}$ denoted by spark$(\mathbf{A})$ is defined as the cardinality of the smallest set of linearly dependent columns.

3 P0 recovery algorithm

For a given matrix $\mathbf{D}$ of dimension $m \times n$, uniqueness of recovery of s-sparse vectors $\mathbf{x}$ from the observation $\mathbf{y} = \mathbf{D}\mathbf{x}$ is guaranteed for

$$s < \frac{\text{spark}(\mathbf{D})}{2}.$$

This can be done via combinatorial search:

$$\text{(P0)} \quad \text{find } \arg\min \|\hat{\mathbf{x}}\|_0 \text{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}}.$$

Suppose that $\|\mathbf{x}\|_0 \leq s$ and $s < \frac{\text{spark}(\mathbf{D})}{2}$. Let $\tilde{\mathbf{x}} \neq \mathbf{x}$ with $\|\tilde{\mathbf{x}}\|_0 \leq s$ and $\mathbf{y} = \mathbf{D}\tilde{\mathbf{x}}$, then

$$0 = \mathbf{y} - \mathbf{y} = \mathbf{D}\mathbf{x} - \mathbf{D}\tilde{\mathbf{x}} = \mathbf{D}\left(\mathbf{x} - \tilde{\mathbf{x}}\right).$$

The sparsity of $\mathbf{x} - \tilde{\mathbf{x}}$ is bounded as $\|\mathbf{x} - \tilde{\mathbf{x}}\|_0 \leq 2s$.

Since $2s < \text{spark}(\mathbf{D})$, we know that

$$\|\mathbf{D}\left(\mathbf{x} - \tilde{\mathbf{x}}\right)\| > 0, \;\; \mathbf{x} - \tilde{\mathbf{x}} \neq 0$$

as every set of $2s$ columns of $\mathbf{D}$ is linearly independent. Therefore (P0) recovers $\mathbf{x}$ uniquely.

Determining the spark of a dictionary is a combinatorial problem and leads to huge computational complexity even for small problem size. Specifically, every set of a columns out of the $\binom{n}{a}$ possible sets has to be checked for linear independence and the parameter a has to be increased starting from two.

4 Coherence

In the following, we assume that every column of a dictionary $\mathbf{D}$ is normalized to unit 2-norm.

We next derive a lower bound on the spark in terms of the coherence of the dictionary, $\mu(\mathbf{D})$, defined according to

Definition 3. For $\mathbf{D} = [\mathbf{d}_1 \ldots \mathbf{d}_n] \in \mathbb{C}^{m \times n}$ with $\|\mathbf{d}_i\|_2 = 1$ for all i, the coherence is defined as $\mu(\mathbf{D}) = \max_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|$.

Theorem 1. [2], [3] (P0) applied to $\mathbf{y} = \mathbf{D}\mathbf{x}$ recovers $\mathbf{x}$ if

$$\|\mathbf{x}\|_0 = s < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right).$$

Proof. We will show that $\text{spark}(\mathbf{D}) \geq 1 + 1/\mu(\mathbf{D})$. Consider $\mathbf{h} \in \mathbb{C}^n$ with $\|\mathbf{h}\|_0 = \text{spark}(\mathbf{D})$ and $\mathbf{D}\mathbf{h} = \mathbf{0}$, i.e. $\mathbf{h} \in \mathcal{N}(\mathbf{D})$. Then, we have

$$\mathbf{d}_l h_l = -\sum_{r \neq l} \mathbf{d}_r h_r, \quad \text{for all } l \in \{1, \ldots, n\}.$$

Left-multiplying both sides by $\mathbf{d}_l^{\mathsf{H}}$ and using $\|\mathbf{d}_l\|_2 = 1$ yields

$$h_l = -\sum_{r \neq l} \mathbf{d}_l^{\mathsf{H}} \mathbf{d}_r h_r,$$

which implies

$$|h_l| = \left| \sum_{r \neq l} \mathbf{d}_l^{\mathsf{H}} \mathbf{d}_r h_r \right| \leq \sum_{r \neq l} \left| \mathbf{d}_l^{\mathsf{H}} \mathbf{d}_r \right| |h_r| \leq \mu(\mathbf{D}) \sum_{r \neq l} |h_r| \quad \text{for } l \in \{1, \ldots, n\}.$$

Adding $\mu(\mathbf{D})|h_l|$ on both sides results in

$$(1 + \mu(\mathbf{D}))\,|h_l| \leq \mu(\mathbf{D})\|\mathbf{h}\|_1 \quad \text{for } l \in \{1, \ldots, n\}. \tag{3}$$

Summing over all l for which $h_l \neq 0$ finally leads to

$$\begin{aligned} &(1 + \mu(\mathbf{D}))\,\|\mathbf{h}\|_1 \leq \mu(\mathbf{D})\|\mathbf{h}\|_1 \operatorname{spark}(\mathbf{D}) \\ &\Rightarrow \operatorname{spark}(\mathbf{D}) \geq 1 + \frac{1}{\mu(\mathbf{D})}. \end{aligned}$$

□

Notice that determining $\mu(\mathbf{D})$ has the complexity of doing the first step in the computation of $\operatorname{spark}(\mathbf{D})$, i.e., checking whether any two columns are linearly independent. Therefore coherence can we quickly calculated for every dictionary.

5 Basis pursuit recovery algorithm

(P0) recovery algorithm is a combinatorial problem and leads to huge computational complexity even for small problem size. Specifically, for every set of s locations out of the $\binom{n}{s}$ possible sets we need to solve a system of linear equation. Is there a more efficient algorithm to recover the signal?

In this section, we consider the recovery algorithm

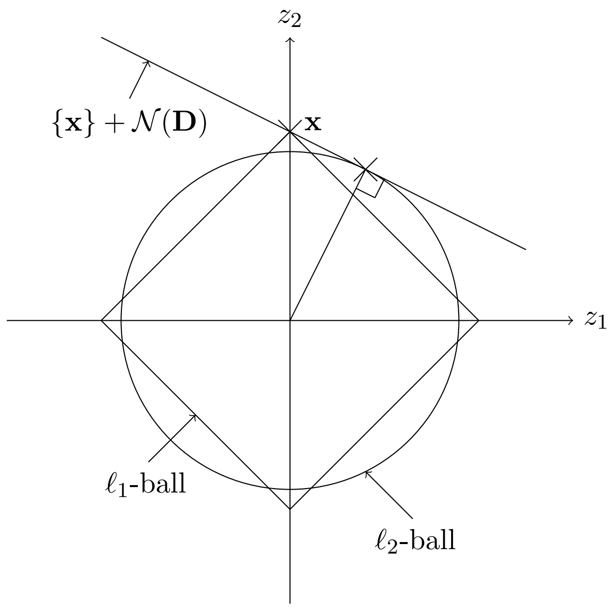
$$\text{(P1)} \quad \text{find } \arg\min \|\hat{\mathbf{x}}\|_1 \text{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}}$$

(P1) is often referred to as basis pursuit (BP). This is a linear program and is therefore efficiently solvable even for huge problem sizes.

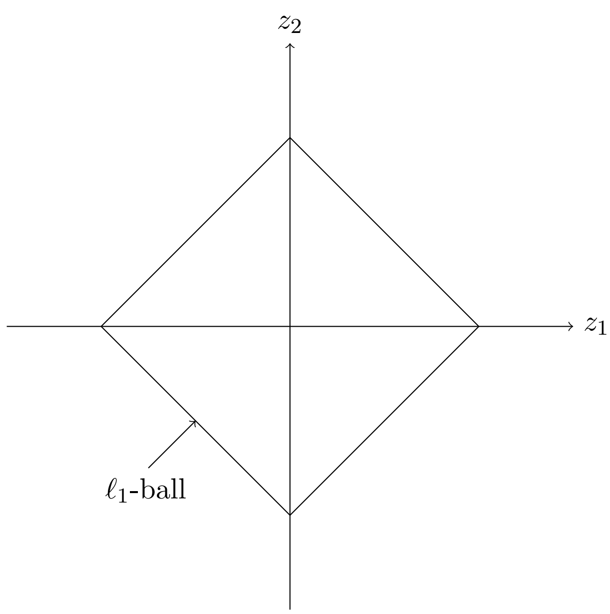
For early results on $l1$-reconstruction see [4] and [5].

Why does ℓ_1-reconstruction work?

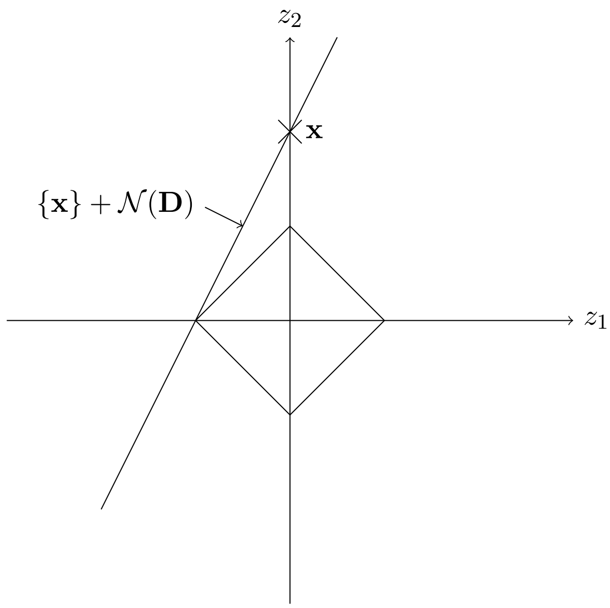
$$\begin{gathered} \arg\min \|\hat{\mathbf{x}}\|_1 \text{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}} \\ \Updownarrow \\ \arg\min \|\hat{\mathbf{x}}\|_1 \text{ subject to } \hat{\mathbf{x}} \in (\{\mathbf{x}\} + \mathcal{N}(\mathbf{D})) \end{gathered}$$



Let us draw the scaled ℓ_1-ball, i.e., $\{(z_1, z_2) : |z_1| + |z_2| = \text{ const.}\}$. Consider the case $z_1, z_2 > 0$. Then, $z_1 + z_2 = \text{ const.}$ and, therefore, $z_2 = \text{ const. } - z_1$. By symmetry, the ℓ_1-ball must look as depicted below.



Clearly, (P1) cannot always recover the correct solutions, e.g., consider the scenario in the figure below.



Can we characterize analytically under which conditions (P1) finds the correct solution?

5.1 Null space property

The following property of $\mathbf{D}$ is central to the success of (P1).

Definition 4. A matrix $\mathbf{D}$ is said to satisfy the *null space property* relative to a set $\mathcal{S}$ if

$$P_1(\mathcal{S}, \mathbf{D}) \triangleq \max_{\mathbf{h} \in \mathcal{N}(\mathbf{D}),\ \mathbf{h} \neq 0} \frac{\sum_{k \in \mathcal{S}} |h_k|}{\sum_k |h_k|} < \frac{1}{2}.$$

It is said to satisfy the null space property of order s if it satisfies the null space property relative to any set $\mathcal{S}$ with $|\mathcal{S}| \leq s$.

5.2 Recovery guarantee

Theorem 2. Fix $\mathbf{x} = [x_1, \ldots, x_n]^\mathsf{T}$ with support set $\mathcal{S}$ and let $\mathbf{y} = \mathbf{D}\mathbf{x}$. If $P_1(\mathcal{S}, \mathbf{D}) < 1/2$, then $\mathbf{x}$ is the unique solution to

$$\text{(P1)} \quad \text{find } \arg\min \|\hat{\mathbf{x}}\|_1 \text{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}}.$$

Proof. We need to prove that for all $\mathbf{h} = [h_1, \ldots, h_n]^\mathsf{T} \in \mathcal{N}(\mathbf{D})$

$$\sum_k |x_k + h_k| > \sum_k |x_k|.$$

Application of the reverse triangle inequality

$$|a + b| \geq |a| - |b|$$

to the LHS of the above equation yields

$$\begin{aligned}\sum_k |x_k + h_k| &= \sum_{k \notin \mathcal{S}} |x_k + h_k| + \sum_{k \in \mathcal{S}} |x_k + h_k| \\ &= \sum_{k \notin \mathcal{S}} |h_k| + \sum_{k \in \mathcal{S}} |x_k + h_k| \\ &\geq \sum_{k \notin \mathcal{S}} |h_k| + \sum_{k \in \mathcal{S}} |x_k| - \sum_{k \in \mathcal{S}} |h_k| .\end{aligned}$$

Therefore, the theorem would follow if we can show that

$$\sum_{k \notin \mathcal{S}} |h_k| > \sum_{k \in \mathcal{S}} |h_k| .$$

Adding $\sum_{k \in \mathcal{S}} |h_k|$ to both sides of the above equation results in the equivalent requirement:

$$\sum_k |h_k| > 2 \sum_{k \in \mathcal{S}} |h_k|$$

which is satisfied iff

$$\underbrace{\frac{\sum_{k \in \mathcal{S}} |h_k|}{\sum_k |h_k|}}_{P_1(\mathcal{S}, \mathbf{D})} < \frac{1}{2}.$$

The last requirement is satisfied for all $\mathbf{h} \in \mathcal{N}(\mathbf{D})$ since $P_1(\mathcal{S}, \mathbf{D}) < 1/2$ by assumption. $\square$

The theorem above guarantees that the basis pursuit algorithm will successfully recover every signal, if $\mathbf{D}$ satisfies the null space property relative to the support of the signal. Therefore, if $\mathbf{D}$ satisfies the null space property of order s, the algorithm will recover successfully every signal with sparsity level no larger than s. How large can s possibly be? Next we will answer this question in terms of coherence $\mu(\mathbf{D})$ of the dictionary $\mathbf{D}$.

5.3 Low coherence implies null space property

Consider $\mathbf{h} \in \mathcal{N}(\mathbf{D})$ and let $\mathcal{S}$ denote the support of $\mathbf{x}$. Due to the proof of Theorem 1, (3), we know that

$$(1 + \mu(\mathbf{D}))\, |h_l| \leq \mu(\mathbf{D}) \|\mathbf{h}\|_1, \quad \text{for all } l = 1, \ldots, n.$$

Summing over all $l \in \mathcal{S}$, we get

$$\begin{aligned}(1 + \mu(\mathbf{D})) \sum_{l \in \mathcal{S}} |h_l| &\leq \mu(\mathbf{D}) \|\mathbf{h}\|_1 |\mathcal{S}| \\ (1 + \mu(\mathbf{D}))\, \frac{\sum_{l \in \mathcal{S}} |h_l|}{\sum_l |h_l|} &\leq \mu(\mathbf{D}) |\mathcal{S}| \\ (1 + \mu(\mathbf{D}))\, P_1(\mathcal{S}, \mathbf{D}) &\leq \mu(\mathbf{D}) |\mathcal{S}| \\ P_1(\mathcal{S}, \mathbf{D}) &\leq \frac{1}{1 + 1/\mu(\mathbf{D})} |\mathcal{S}|.\end{aligned}$$

Therefore, if $\frac{1}{1+1/\mu(\mathbf{D})}|\mathcal{S}| < 1/2$, then $P_1(\mathcal{S}, \mathbf{D}) < 1/2$, i.e., the desired upper bound on the cardinality of the support set is given by

$$|\mathcal{S}| < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$$

just as in the case of recovery via (P0).

We therefore proved the following theorem.

Theorem 3. [2], [3] (P1) applied to $\mathbf{y} = \mathbf{D}\mathbf{x}$ recovers $\mathbf{x}$ if

$$\|\mathbf{x}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right).$$

6 Other signal recovery problems

Our motivation so far has been the signal separation recovery problem. We now list other related problems that can be addressed in a similar way.

6.1 Inpainting

Consider the following setting:

- the signal $\mathbf{x}$ is sparse with unknown support set
- we observe $\mathbf{y} = \mathbf{A}\mathbf{x}$, $\mathbf{A}$ is, for example, a discrete wavelet transform matrix
- only a subset of the entries of $\mathbf{y} = \mathbf{A}\mathbf{x}$ is available
- inpainting amounts to filling in the missing entries
- we account for the missing entries by taking the observation to be

$$\mathbf{z} = \underbrace{\mathbf{A}\mathbf{x}}_{\mathbf{y}} + \mathbf{e} = \mathbf{A}\mathbf{x} + \mathbf{I}\mathbf{e}$$

  and choose $\mathbf{e}$ such that the entries of $\mathbf{z} = \mathbf{y} + \mathbf{e}$ corresponding to the missing entries in $\mathbf{y}$ are set to some arbitrary value, e.g., zero.

If there are not too many entries missing or the area to be inpainted is not too big, $\mathbf{e}$ will be sparse.

To summarize, we observe

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{I}\mathbf{e}$$

and know that $\mathbf{x}, \mathbf{e}$ are sparse. Based on the observation $\mathbf{z}$, we want to recover $\mathbf{x}$ and $\mathbf{e}$. Here is an example of in-painting in action. The image on the right is reconstructed from the image on the left:

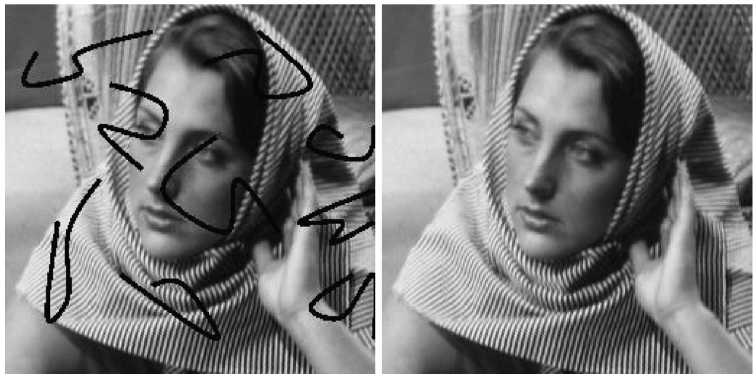


Figure 6: Inpainting.

6.2 Clipping

We observe $\mathbf{z} = g_a(\mathbf{y})$, where the function $g_a(\mathbf{y})$ realizes entry-wise signal clipping to the interval $[-a, a]$.

Clipping can equivalently be modeled as

$$\mathbf{z} = \mathbf{y} + \mathbf{e}$$

with $\mathbf{e} = g_a(\mathbf{y}) - \mathbf{y}$. Notice that the error locations can be determined by comparing the entries of $\mathbf{z}$ to the clipping threshold a.

Consequently, we observe

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{I}\mathbf{e}$$

where $\mathbf{e}$ can depend on $\mathbf{A}$ and $\mathbf{x}$.

6.3 Recovery of signals subject to impulse noise

In this scenario, a spectrally sparse signal with unknown spectrum is (sparsely) corrupted by impulses with unknown locations, i.e., we observe

$$\mathbf{z} = \mathbf{F}\mathbf{x} + \mathbf{I}\mathbf{e}.$$

where $\mathbf{F}$ is the DFT matrix and $\mathbf{I}$ is the identity matrix.

6.4 Recovery of signals subject to narrowband interference

We observe a sparse signal corrupted by spectrally sparse noise, i.e.,

$$\mathbf{z} = \mathbf{I}\mathbf{x} + \mathbf{F}\mathbf{e}.$$

where $\mathbf{F}$ is the DFT matrix and $\mathbf{I}$ is the identity matrix.

7 Uncertainty principles and signal recovery

The main reference for this section is [6].

In all examples in Section 6.1 – Section 6.4 the dictionary $\mathbf{D}$ is the concatenation of two ONBs $\mathbf{A}$ and $\mathbf{B}$. In this case, refined bounds on $\text{spark}(\mathbf{D}) = \text{spark}([\mathbf{A}\ \mathbf{B}])$ exist. We derive them next. In the derivation, we will use an unexpected connection between the uncertainty principles, similar to the ones used in physics, and signal recovery guarantees.

For a vector in the null-space of $\mathbf{D} = [\mathbf{A}\ \mathbf{B}]$, we have

$$[\mathbf{A}\ \mathbf{B}] \underbrace{\begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix}}_{\mathbf{v}} = 0.$$

Hence,

$$\mathbf{A}\mathbf{p} + \mathbf{B}\mathbf{q} = \mathbf{0} \Rightarrow \mathbf{A}\mathbf{p} = \mathbf{B}\,(-\mathbf{q}) \triangleq \mathbf{s}.$$

The signal $\mathbf{s}$ is represented in two different ways, namely as an expansion in the dictionary $\mathbf{A}$ and as an expansion in the dictionary $\mathbf{B}$.

Finding the vector $\mathbf{v}$ with minimum 0-norm among all vectors that satisfy

$$[\mathbf{A}\ \mathbf{B}]\mathbf{v} = \mathbf{0}$$

amounts to answering the question: How sparse can $\mathbf{p}$ and $\mathbf{q}$, the expansion coefficients of the *same* vector $\mathbf{s}$ in two different ONBs, concurrently be? This is a question directly related to uncertainty principle, as we explain next.

7.1 Uncertainty principles

Theorem 4. Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times m}$ be two unitary matrices. Let $\mathbf{A}\mathbf{p} = -\mathbf{B}\mathbf{q}$. Then $\|\mathbf{p}\|_0\|\mathbf{q}\|_0 \geq 1/\mu^2([\mathbf{A}\ \mathbf{B}])$.

Proof. Since $\mathbf{A}$ is unitary, left multiplication of the identity $\mathbf{A}\mathbf{p} = -\mathbf{B}\mathbf{q}$ by $\mathbf{A}^{\mathsf{H}}$ yields $\mathbf{p} = -\mathbf{A}^{\mathsf{H}}\mathbf{B}\mathbf{q}$. An entry of $\mathbf{p}$ satisfies

$$\begin{aligned} |p_k| &= \left|[\mathbf{A}^{\mathsf{H}}\mathbf{B}\mathbf{q}]_k\right| = \left|\sum_l [\mathbf{A}^{\mathsf{H}}\mathbf{B}]_{k,l} q_l\right| \leq \sum_l |\langle \mathbf{a}_k, \mathbf{b}_l\rangle|\, |q_l| \\ &\leq \max_l |\langle \mathbf{a}_k, \mathbf{b}_l\rangle|\, \|\mathbf{q}\|_1 \leq \mu([\mathbf{A}\ \mathbf{B}])\|\mathbf{q}\|_1. \end{aligned}$$

Summation of $k \in \text{supp}(\mathbf{p})$ yields

$$\|\mathbf{p}\|_1 \leq \|\mathbf{p}\|_0 \mu([\mathbf{A}\ \mathbf{B}])\|\mathbf{q}\|_1.$$

Left multiplication of the identity $\mathbf{A}\mathbf{p} = -\mathbf{B}\mathbf{q}$ by $\mathbf{B}^{\mathsf{H}}$ similarly gives

$$\|\mathbf{q}\|_1 \leq \|\mathbf{q}\|_0 \mu([\mathbf{A}\ \mathbf{B}])\|\mathbf{p}\|_1.$$

Multiplying both inequalities together and rearranging the terms we obtain the result. $\square$

Comparison to the classical uncertainty principle. In signal analysis, it is widely know that the signal and its Fourier transform cannot simultaneously be tightly concentrated. This fact can be expressed in many different ways. The classical way is in terms of concentration of the second moment.

Assume that the function $g(t)$ has unit norm: $\int_{-\infty}^{\infty} |g(t)|^2 \, dt = 1$. The uncertainty principle states that

$$\underbrace{\int_{-\infty}^{\infty} t^2 \, |g(t)|^2 \, dt}_{\text{concentration in time}} \times \underbrace{\int_{-\infty}^{\infty} f^2 \, |\hat{g}(f)|^2 \, df}_{\text{concentration in frequency}} \geq \frac{1}{16\pi^2}, \tag{4}$$

where $\hat{g}(f)$ is the Fourier transform of $g(t)$.

If $\mathbf{A} = \mathbf{I}$ (the identity matrix) and $\mathbf{B} = \mathbf{F}$ (the DFT matrix), then $\mathbf{p}$ corresponds to $g(\cdot)$ and $-\mathbf{q}$ corresponds to $\hat{g}(\cdot)$. The bound we derived before is of the same type as (4):

$$\underbrace{\|\mathbf{p}\|_0}_{\text{concentration in time}} \quad \times \quad \underbrace{\|\mathbf{q}\|_0}_{\text{concentration in frequency}} \quad \geq \quad 1/\mu^2([\mathbf{A} \ \mathbf{B}]).$$

7.2 Signal recovery guarantees

The uncertainty relation in Theorem 4 states that $\mathbf{A}\mathbf{p} + \mathbf{B}\mathbf{q} = \mathbf{0}$ is only possible if $\|\mathbf{p}\|_0 \|\mathbf{q}\|_0 \geq 1/\mu^2([\mathbf{A} \ \mathbf{B}])$.

In the case where $\mathbf{D} = [\mathbf{A} \ \mathbf{B}]$ is a concatenation of two ONBs, the dictionary coherence evaluates to

$$\mu(\mathbf{D}) = \max_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle| = \max_{i,j} |\langle \mathbf{a}_i, \mathbf{b}_j \rangle| \, .$$

How can these uncertainty relations be used to obtain recovery thresholds?
By assumption, we have

$$\operatorname{spark}(\mathbf{D}) \geq \|\mathbf{p}\|_0 + \|\mathbf{q}\|_0 = \left\| \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix} \right\|_0 .$$

The arithmetic-mean – geometric-mean (AM-GM) inequality implies the following lower bound on the spark:

$$\operatorname{spark}(\mathbf{D}) \geq \|\mathbf{p}\|_0 + \|\mathbf{q}\|_0 \geq 2\sqrt{\|\mathbf{p}\|_0 \|\mathbf{q}\|_0} \geq \frac{2}{\mu(\mathbf{D})}. \tag{5}$$

Recall that the solution to (P0) is unique if the sparsity of the signal s satisfies

$$s < \frac{\operatorname{spark}(\mathbf{D})}{2}. \tag{6}$$

If sparsity is upper-bounded by

$$s < \frac{1}{\mu(\mathbf{D})}$$

then (5) guarantees that (6) is satisfied and, therefore, (P0) will recover the signal successfully. Compare this to the old threshold

$$s < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) \approx \frac{1}{2\mu(\mathbf{D})}$$

to observe that for dictionaries which are the concatenation of two ONBs we get an improvement in the recovery threshold by a factor of two.

8 Square-root bottleneck

All sparsity thresholds we obtained so far are proportional to $1/\mu(\mathbf{D})$. What is the largest sparsity threshold we can obtain? How small can the dictionary coherence $\mu(\mathbf{D})$ possibly be?

Theorem 5 (Welch bound, [7])**.** Let $\mathbf{D} \in \mathbb{C}^{m \times n}$ be a dictionary with coherence $\mu(\mathbf{D})$. Then,

$$\mu(\mathbf{D}) \geq \sqrt{\frac{n-m}{m\,(n-1)}},$$

where $m \leq n$.

Proof. Set $\mathbf{G} = \mathbf{D}^{\mathsf{H}}\mathbf{D} \in \mathbb{C}^{n \times n}$. Then, $\mathbf{G}$ has the following properties:

1. $\mathbf{G}$ has ones along its diagonal (since all dictionary columns have unit ℓ_2 norm);
2. $\mathbf{G}$ is positive semi-definite with rank (at most) m.

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathsf{T}}$ denote the vector of nonzero eigenvalues λ_i of $\mathbf{G}$. Then, we have

$$\operatorname{tr}(\mathbf{G}) = \sum_{i=1}^{m} \lambda_i = \|\boldsymbol{\lambda}\|_1 = n$$

$$\|\mathbf{G}\|_F^2 = \operatorname{tr}\left(\mathbf{G}^{\mathsf{T}}\mathbf{G}\right) = \sum_{i,j} \mathbf{G}_{ij}^{\mathsf{T}}\mathbf{G}_{ij} = \sum_{i=1}^{m} \lambda_i^2 = \|\boldsymbol{\lambda}\|_2^2.$$

Since

$$\left(\frac{1}{m}\sum_{i=1}^{m} \lambda_i\right)^2 \leq \frac{1}{m}\sum_{i=1}^{m} \lambda_i^2$$

by Cauchy-Schwarz inequality, it follow that $\|\boldsymbol{\lambda}\|_1^2 \leq m\|\boldsymbol{\lambda}\|_2^2$, which implies in turn that

$$\|\mathbf{G}\|_F^2 \geq \frac{n^2}{m}.$$

We thus have

$$\begin{aligned}\|\mathbf{G}\|_F^2 &= n + \sum_{i=1}^{n}\sum_{j \neq i} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|^2 \\ &\geq \frac{n^2}{m},\end{aligned}$$

which finally yields

$$\begin{aligned}\mu(\mathbf{D})^2 &\geq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|^2 \\ &\geq \frac{1}{n(n-1)} \left(\frac{n^2}{m} - n \right) \\ &= \frac{n-m}{m\,(n-1)}.\end{aligned}$$

□

For $n \gg m$ the Welch lower bound implies

$$\mu(\mathbf{D}) \geq \sqrt{\frac{n-m}{m\,(n-1)}} \approx \frac{1}{\sqrt{m}}$$

and hence all sparsity thresholds obtained so far obey the fundamental upper bound

$$s \lesssim \sqrt{m} \quad \Rightarrow \quad m \gtrsim s^2.$$

We see that there is a limit on all deterministic sparsity thresholds we have derived. We say that the thresholds are bounded by the *square-root bottleneck* meaning that recovery is only guaranteed if we take $m \approx s^2$ samples for an s-sparse signal.

Take, e.g., $s = 30$ and $n = 1000$. The square-root bottleneck implies that we would need ≈ 900 samples to get recovery through (P1) or OMP. This is very disappointing.

Hence, the question: Can we improve upon the scaling behavior $s \lesssim m^2$?

Square-root bottleneck is tight. The answer is that the bound of the order $s \lesssim \sqrt{m}$ is the best we can obtain for general dictionaries $\mathbf{D}$ if recovery is to be guaranteed uniformly for all s-sparse signals.

Consider the case $\mathbf{D} = [\mathbf{I} \quad -\mathbf{F}]$, where $\mathbf{I}$ is the $m \times m$ identity matrix and $\mathbf{F}$ is the $m \times m$ DFT matrix. Assume that m is a perfect square: $m = \eta^2$, where $\eta \in \mathbb{N}$. Let $\mathbf{p} = [p_1, \ldots, p_m]^\mathsf{T}$ be defined as

$$p_l = \begin{cases} 1, & \text{if } l = 1 \mod \eta \\ 0, & \text{otherwise.} \end{cases}$$

It turns out that the DFT of $\mathbf{p}$, $\hat{\mathbf{p}} = [\hat{p}_1, \ldots, \hat{p}_m]^\mathsf{T} = \mathbf{F}\mathbf{p}$ is equal to $\mathbf{p}$ itself. This can be seen as follows:

$$\hat{p}_j = \frac{1}{\eta} \sum_{l=1}^{\eta^2} p_l e^{2\pi \mathrm{i}(l-1)(j-1)/\eta^2} = \frac{1}{\eta} \sum_{k=1}^{\eta} e^{2\pi \mathrm{i}(k-1)(j-1)/\eta} = \begin{cases} 1, & \text{if } j = 1 \mod \eta \\ 0, & \text{otherwise.} \end{cases}$$

This shows that $\hat{\mathbf{p}} = \mathbf{p}$. Therefore, $\mathbf{I}\mathbf{p} = \mathbf{F}\mathbf{p}$ and $\mathbf{D}\mathbf{x} = \mathbf{0}$ with $\mathbf{x} = [\mathbf{p}^\mathsf{T}\ \mathbf{p}^\mathsf{T}]^\mathsf{T} \neq \mathbf{0}$. Any reasonable reconstruction algorithm, having observed $\mathbf{y} = \mathbf{D}\mathbf{x} = \mathbf{0}$, will return $\mathbf{x} = \mathbf{0}$. This is not a correct answer, and, therefore, $\mathbf{x}$ cannot be recovered from $\mathbf{y}$ by any method whatsoever.

At the same time $\|\mathbf{x}\|_0 = 2\|\mathbf{p}\|_0 = 2\eta = 2\sqrt{m}$. We cannot recover arbitrary signals of sparsity higher than $2\sqrt{m}$ when $\mathbf{D} = [\mathbf{I}\ \mathbf{F}]$.

It is easy to see (homework) that $\mu(\mathbf{D}) = \frac{1}{\sqrt{m}}$. Therefore, this example demonstrates that if the sparsity bound is formulated in terms of coherence we cannot hope to get a bound better than

$$s \leq \frac{2}{\mu(\mathbf{D})}. \tag{7}$$

We also conclude that the bound found in (5) is tight.

This means that the square-root bottleneck type bound cannot be improved for general dictionaries without additional assumptions on the signal.

Breaking the square-root bottleneck. It turns out that much more optimistic results can be obtained in many cases of practical relevance.

First, we will see that nonnegativity assumption on the elements of the signal will allow us to prove a recovery result when $s < m/2$ in the case of super-resolution problem.

Second, we will see that s can scale linearly with m (up to logarithmic factors) if we allow for randomized sampling, i.e. we will inject randomness in the construction of the sampling matrix.

Also, under mild assumptions on $\mathbf{D}$ it is possible to prove recovery results with s linear in m (up to logarithmic factors) if we don't insist on perfect recovery for all signals, but only for most signals with random support.

References

[1] J. L. Starck, D. Donoho, and E. J. Candès, "Astronomical image representation by the curvelet transform," *A&A*, vol. 398, pp. 785–800, Nov. 2003.

[2] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, Mar. 2003.

[3] R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," *Applied and Computational Harmonic Analysis*, vol. 22, no. 3, pp. 335–355, 2007.

[4] B. F. Logan, *Properties of High-Pass Signals.* PhD thesis, Columbia University, New York, 1965.

[5] D. L. Donoho and B. F. Logan, "Signal recovery and the large sieve," *SIAM J. Appl. Math.*, vol. 52, pp. 577–591, Apr. 1992.

[6] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, pp. 906–931, June 1989.

[7] L. Welch, "Lower bounds on the maximum cross correlation of signals (corresp.)," *IEEE Transactions on Information Theory*, vol. 20, pp. 397–399, May 1974.