An algorithm for variable density sampling with block-constrained acquisition

Claire Boyer$^{(1)}$, Pierre Weiss$^{(2)}$, and Jérémie Bigot$^{(3)}$

$^{(1)}$ Institut de Mathématiques de Toulouse, Université de Toulouse, France
claire.boyer@math.univ-toulouse.fr
$^{(2)}$ Institut des Technologies Avancées du Vivant, Toulouse, France
pierre.armand.weiss@gmail.com
$^{(3)}$ DMIA, Institut Supérieur de l'Aéronautique et de l'Espace, Toulouse, France
jeremie.bigot@isae.fr

January 24, 2014

Abstract

Reducing acquisition time is of fundamental importance in various imaging modalities. The concept of variable density sampling provides a nice framework to address this issue. It was justified recently from a theoretical point of view in the compressed sensing (CS) literature. Unfortunately, the sampling schemes suggested by current CS theories may not be relevant since they do not take the acquisition constraints into account (for example, continuity of the acquisition trajectory in Magnetic Resonance Imaging - MRI). In this paper, we propose a numerical method to perform variable density sampling with block constraints. Our main contribution is to propose a new way to draw the blocks in order to mimic CS strategies based on isolated measurements. The basic idea is to minimize a tailored dissimilarity measure between a probability distribution defined on the set of isolated measurements and a probability distribution defined on a set of blocks of measurements. This problem turns out to be convex and solvable in high dimension. Our second contribution is to define an efficient minimization algorithm based on Nesterov's accelerated gradient descent in metric spaces. We study carefully the choice of the metrics and of the prox function. We show that the optimal choice may depend on the type of blocks under consideration. Finally, we show that we can obtain better MRI reconstruction results using our sampling schemes than standard strategies such as equiangularly distributed radial lines.

Key-words: Compressed Sensing, blocks of measurements, blocks-constrained acquisition, dissimilarity measure between discrete probabilities, optimization on metric spaces.

1 Introduction

Compressive Sensing (CS) is a recently developed sampling theory that provides theoretical conditions to ensure the exact recovery of signals from a few number of linear measurements (below the Nyquist rate). CS is based on the assumption that the signal to reconstruct can be represented by a few number of atoms in a certain basis. We say that the signal $\boldsymbol{x} \in \mathbb{C}^n$ is s-sparse if

$$\|\boldsymbol{x}\|_{\ell^0} \leq s,$$

where $\|\cdot\|_{\ell^0}$ denotes the ℓ_0 pseudo-norm, counting the number of non-zero entries of $\boldsymbol{x}$. Original CS theorems [Don06, CRT06, CP11a] assert that a sparse signal $\boldsymbol{x}$ can be faithfully reconstructed via ℓ_1-minimization:

$$\min_{\boldsymbol{z} \in \mathbb{C}^n} \|\boldsymbol{z}\|_{\ell^1} \qquad \text{such that} \qquad \boldsymbol{A}_\Omega \boldsymbol{z} = \boldsymbol{y}, \tag{1}$$

where $\boldsymbol{A}_\Omega \in \mathbb{C}^{p \times n}$ $(p \leq n)$ is a sensing matrix, $\boldsymbol{y} = \boldsymbol{A}_\Omega \boldsymbol{x} \in \mathbb{C}^p$ represents the vector of linear projections, and $\|\boldsymbol{z}\|_{\ell^1} = \sum_{i=1}^n |z_i|$ for all $\boldsymbol{z} = (z_1, \ldots, z_n) \in \mathbb{C}^n$. More precisely CS results state that $p = O(s \ln(n))$ measurements are enough to guarantee exact reconstruction provided that $\boldsymbol{A}_\Omega$ satisfies some incoherence property.

One way to construct $\boldsymbol{A}_\Omega$ is by randomly extracting rows from a full sensing matrix $\boldsymbol{A} \in \mathbb{C}^{n \times n}$ that can be written as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1^* \\ \vdots \\ \boldsymbol{a}_n^* \end{pmatrix}, \tag{2}$$

where $\boldsymbol{a}_i^*$ denotes the i-th row of $\boldsymbol{A}$. In the context of Magnetic Resonance Imaging (MRI) for instance, the full sensing matrix $\boldsymbol{A}$ consists in the composition of a Fourier transform with an inverse wavelet transform. This choice is due to the fact that the acquisition is done in the Fourier domain, while the images to be reconstructed are assumed to be sparse in the wavelet domain. In this setting, a fundamental issue is constructing $\boldsymbol{A}_\Omega$ by extracting appropriate rows from the full sensing matrix $\boldsymbol{A}$. A theoretically founded approach to build $\boldsymbol{A}_\Omega$ (i.e. constructing of sampling schemes) consists in randomly extracting rows from $\boldsymbol{A}$ according to a given density. This approach requires to define a discrete probability distribution $\boldsymbol{p} = (\boldsymbol{p}_i)_{1 \leq i \leq n}$ on the set of integers $\{1, \ldots, n\}$ that represents the indexes of the rows of $\boldsymbol{A}$. We call this procedure variable density sampling. This term appeared in the early MRI paper [SPM95]. It was recently given a mathematical definition in [CCKW13]. One possibility to construct $\boldsymbol{p}$ is to choose its i-th component $\boldsymbol{p}_i$ to be proportional to $\|\boldsymbol{a}_i^*\|_{\ell^\infty}^2$ (see [Rau10, PVW11, BBW13, CCW13]) i.e.

$$\boldsymbol{p}_i = \frac{\|\boldsymbol{a}_i^*\|_{\ell^\infty}^2}{\sum_{k=1}^n \|\boldsymbol{a}_k^*\|_{\ell^\infty}^2}, \; i = 1, \ldots, n. \tag{3}$$

In the MRI setting, another strategy ensuring good reconstruction is to choose $\boldsymbol{p}$ according to a polynomial radial distribution [KW12] in the so-called *k-space* i.e. the 2D Fourier plane where low frequencies are centered. Other strategies are possible. For example, [AHPR13] propose a multilevel uniformly random subsampling approach.

All these strategies lead to sampling schemes that are made of a few but isolated measurements, see e.g. Figure 1 (a). However, in many applications, the number of measurements is not of primary importance relative to the path the sensor must take to collect the measurements. For instance, in MRI, sampling is done in the Fourier domain along continuous and smooth curves [Wri97, LKP08]. Another example of the need to sample continuous trajectories can be found in mobile robots monitoring where robots have to spatially sample their environment under kinematic and energy consumption constraints [HPH$^+$11].

This paper focuses on the acquisition of linear measurements in applications where the physics of the sensing device allows to sample a signal from pre-defined blocks of measurements. We define a block of measurements as an arbitrary set of isolated measurements, that could be contiguous in the Fourier plane for instance. As an illustrative example (that will be used throughout the paper), one may consider sampling patterns generated by randomly drawing a set of straight lines in the Fourier plane or *k-space* as displayed in Figure 1(b). This kind of sampling patterns is particularly relevant in the case of MRI acquisition with echo planar sampling strategies, see e.g. [LDSP08].

Acquiring data by blocks of measurements raises the issue of designing appropriate sampling schemes. In this paper, we propose to randomly extract blocks of measurements that are made of several rows from the full sensing matrix $\boldsymbol{A}$. The main question investigated is how to choose an

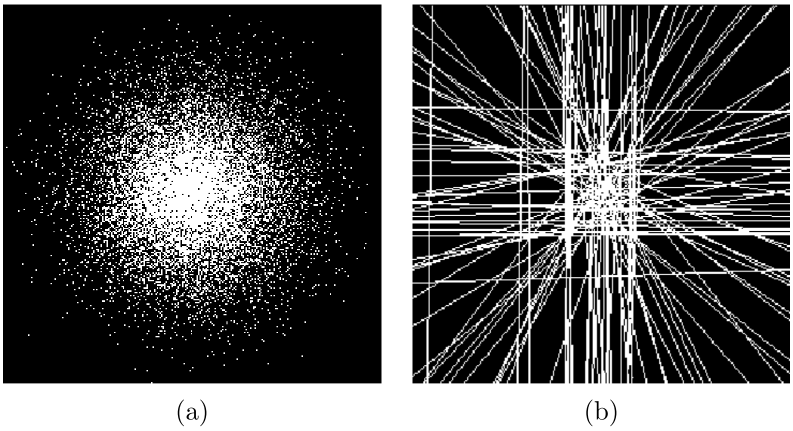


Figure 1: **An example of MRI sampling schemes in the *k-space* (the 2D Fourier plane where low frequencies are centered)** (a): Isolated measurements drawn from a probability measure $\boldsymbol{p}$ having a radial distribution. (b): Sampling scheme based on a dictionary of blocks of measurements: blocks consist of discrete lines of the same size.

appropriate probability distribution from which blocks of measurements will be drawn. A first step in this direction [BBW13, PDG12] was recently proposed. In [BBW13], we have derived a theoretical probability distribution in the case of blocks of measurements to design a sensing matrix $\boldsymbol{A}_\Omega$ that guarantees an exact reconstruction of s-sparse signals with high probability. Unfortunately, the probability distributions proposed in [BBW13] and [PDG12] are difficult to compute numerically and seem suboptimal in practice.

In this paper, we propose an alternative strategy which is based on the numerical resolution of an optimization problem. Our main idea is to construct a probability distribution $\boldsymbol{\pi}$ on a dictionary of blocks. The blocks are drawn independently at random according to this distribution. We propose to choose $\boldsymbol{\pi}$ in such a way that the resulting sampling patterns are similar to those based on isolated measurements, such as the ones proposed in the CS literature. For this purpose, we define a dissimilarity measure to compare a probability distribution $\boldsymbol{\pi}$ on a dictionary of blocks and a target probability distribution $\boldsymbol{p}$ defined on a set of isolated measurements. Then, we propose to choose an appropriate distribution $\boldsymbol{\pi}[\boldsymbol{p}]$ by minimizing its dissimilarity with a distribution $\boldsymbol{p}$ on isolated measurements that is known to lead to good sensing matrices.

This paper is organized as follows. In Section 2, we introduce the notation. In Section 3, we describe the problem setting. Then, we construct a dissimilarity measure between probability distributions lying in different, but spatially related domains. We then formulate the problem of finding a probability distribution $\boldsymbol{\pi}[\boldsymbol{p}]$ on blocks of measurements as a convex optimization problem. In Section 4, we present an original and efficient way to solve this minimization problem via a dual formulation and an algorithm based on the accelerated gradient descents in metric spaces [Nes05]. We study carefully how the theoretical rates of convergence are affected by the choice of norms and prox-functions on the primal and dual spaces. Finally, in Section 5, we propose a dictionary of blocks that is appropriate for MRI applications. Then, we compare the quality of MRI images reconstructions using the proposed sampling schemes and those currently used in the context of MRI acquisition, demonstrating the potential of the proposed approach on real scanners.

2 Notation

We consider d-dimensional signals for any $d \in \mathbb{N}^*$, of size $n_1 \times n_2 \times \ldots \times n_d = n$. Let E and F denote finite-dimensional vector spaces endowed with their respective norms $\|.\|_E$ and $\|.\|_F$. In the paper, we identify E to $\mathbb{R}^m$ and F to $\mathbb{R}^n$. We denote by E^* and F^*, respectively the dual spaces of E and F. For $s \in E^*$ and $x \in E$ we denote by $\langle s, x\rangle_{E^*\times E}$ the value of s at x. The notation $\langle \cdot, \cdot\rangle$ will denote the usual inner product in a Euclidean space. The norm of the dual space E^* is defined by:

$$\|s\|_{E^*} = \max_{\substack{x\in E \\ \|x\|_E=1}} \langle s, x\rangle_{E^*\times E}\,.$$

Let $\boldsymbol{M} : E \to F^*$ denote some operator. When M is linear, we denote its adjoint operator by $\boldsymbol{M}^* : F \to E^*$. The subordinate operator norm is defined by :

$$\|\boldsymbol{M}\|_{E\to F^*} = \sup_{\|x\|_E\leq 1} \|\boldsymbol{M}x\|_{F^*}$$

When the spaces E^* and F are endowed with ℓ^q and ℓ^p norms respectively, we will use the following notation for the operator norm of $\boldsymbol{M}^*$:

$$\|\boldsymbol{M}^*\|_{F\to E^*} = \|\boldsymbol{M}^*\|_{p\to q}.$$

We set $\Delta_m \subset E$ to be the simplex in $E = \mathbb{R}^m$, and $\Delta_n \subset F$ to be the simplex in $F = \mathbb{R}^n$. For $\boldsymbol{\pi} \in \Delta_m$ and an index $j \in \{1, \ldots, m\}$ we denote by $\boldsymbol{\pi}_j$ the j-th component of $\boldsymbol{\pi}$.

Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ denote a closed convex function. Its Fenchel conjugate is denoted g^*. The relative interior of a set $X \subseteq \mathbb{R}^n$ is denoted $\mathrm{ri}(X)$. Finally, the normal cone to X at a point x on the boundary of X is denoted $\mathcal{N}_X(x)$.

3 Variable density sampling with block constraints

3.1 Problem setting

In this paper, we assume that the acquisition system is capable of sensing a finite set $\{y_1, \ldots, y_n\}$ of linear measurements of a signal $\boldsymbol{x} \in \mathbb{R}^{n_s}$ such that

$$y_i = \langle \boldsymbol{a}_i^*, \boldsymbol{x}\rangle, \qquad \forall i = 1, \ldots, n,$$

where $\boldsymbol{a}_i^*$ denotes the i-th row of the full sensing matrix $\boldsymbol{A}$. Let us define a set $\mathcal{I} = \{I_1, \ldots, I_m\}$ where each $I_k \subseteq \{1, \ldots, n\}$ denotes a set of indexes. We assume that the acquisition system has physical constraints that impose sensing simultaneously the following sets of measurements

$$E_k = \{y_i, i \in I_k\}, \qquad \forall k = 1, \ldots, m.$$

In what follows, we refer to $\mathcal{I}$ as the blocks dictionary.

For example in MRI, $n = n_s$ is the number of pixels or voxels of a 2D or 3D image, and y_i represents the i-th discrete Fourier coefficient of this image. In this setting, the sets of indexes I_k may represent straight lines in the discrete Fourier domain as in Figure 1(b). In Section 5.1, we give further details on the construction of such a dictionary.

We propose to partially sense the signal using the following procedure:

(i) Construct a discrete probability distribution $\boldsymbol{\pi} \in \Delta_m$.

(ii) Draw i.i.d. indexes $k_1, \ldots, k_b$ from the probability distribution $\boldsymbol{\pi}$ on the set $\{1, \ldots, m\}$, with $1 \leq b \leq m$.

(iii) Sense randomly the signal $\boldsymbol{x}$ by considering the random set of blocks of measurements $\left(E_{k_j}\right)_{j\in\{1,\ldots,b\}}$, which leads to the construction of the following sensing matrix

$$\boldsymbol{A}_\Omega = \left(\boldsymbol{a}_i^*\right)_{i\,\in\,\cup_{j=1}^b I_{k_j}}\,.$$

The main objective of this paper is to provide an algorithm to construct the discrete probability distribution $\boldsymbol{\pi}$ based on the knowledge of a target discrete probability distribution $\boldsymbol{p} \in \Delta_n$ on the set $\{y_1, \ldots, y_n\}$ of isolated measurements. The problem of choosing a distribution $\boldsymbol{p}$ leading to good image reconstruction is not addressed in this paper, since there already exist various theoretical results and heuristic strategies in the CS literature on this topic [LKP08, CCW13, AHPR13, KW12].

3.2 A variational formulation

In order to define $\boldsymbol{\pi}$, we propose to minimize a dissimilarity measure between $\boldsymbol{\pi} \in \Delta_m$ and $\boldsymbol{p} \in \Delta_n$. The difficulty lies in the fact that these two probability distributions belong to different spaces. We propose to construct a dissimilarity measure $\mathcal{D}(\boldsymbol{\pi}, \boldsymbol{p}, \mathcal{I})$ that depends on the blocks dictionary $\mathcal{I}$. This dissimilarity measure will be minimized over $\boldsymbol{\pi} \in \Delta_m$ using numerical algorithms with m being relatively large (typically $10^4 \le m \le 10^{10}$). Therefore, it must have appropriate properties such as convexity, for the problem to be solvable in an efficient way.

Mapping the m-dimensional simplex to the n-dimensional one

In order to define a reasonable dissimilarity measure, we propose to construct an operator $\boldsymbol{M}$ that maps a probability distribution $\boldsymbol{\pi} \in \Delta_m$ to some $\boldsymbol{p}' \in \Delta_n$:

$$\begin{aligned} \boldsymbol{M} : \quad & E \longrightarrow F^* \\ & \boldsymbol{\pi} \longmapsto \boldsymbol{p}', \end{aligned}$$

where for $i \in \{1, \ldots, n\}$,

$$\boldsymbol{p}'_i = \frac{\sum_{k=1}^{m} \boldsymbol{\pi}_k \mathbb{1}_{i \in I_k}}{\sum_{j=1}^{n} \sum_{k'=1}^{m} \boldsymbol{\pi}_{k'} \mathbb{1}_{j \in I_{k'}}}, \tag{4}$$

where $\mathbb{1}_{i \in I_k}$ is equal to 1 if $i \in I_k$, 0 otherwise. The i-th element of $\boldsymbol{p}'$ represents the probability to draw the i-th measurement y_i by drawing blocks of measurements according to the probability distribution $\boldsymbol{\pi}$. The operator $\boldsymbol{M}$ satisfies the following property by construction :

$$\boldsymbol{M}\Delta_m \subseteq \Delta_n.$$

A sufficient condition for the mapping $\boldsymbol{M}$ to be a linear operator

Note that the operator $\boldsymbol{M}$ is generally non linear, due to the denominator in (4). This is usually an important drawback for the design of numerical algorithms involving the operator $\boldsymbol{M}$. However, if the sets $(I_k)_{k \in \{1,\ldots,m\}}$ all have the same cardinality (or length) equal to ℓ, the denominator in (4) is equal to ℓ. In this case, $\boldsymbol{M}$ becomes a linear operator. In this paper, we will focus on this setting, which is rich enough for many practical applications:

Assumption 3.1. *For $k \in \{1, \ldots, m\}$, Card $(I_k) = \ell$, where ℓ is some positive integer.*

Let us provide two important results for the sequel.

Proposition 3.2. *For $\ell > 1$, $\boldsymbol{M}\Delta_m \subsetneq \Delta_n$, i.e. $\boldsymbol{M}\Delta_m$ is a strict subset of Δ_n.*

Proof. By definition of the convex envelope, $\boldsymbol{M}\Delta_m = \operatorname{conv}(\{\boldsymbol{M}_{:,i}, i \in \{1, \ldots, m\}\})$, where $\boldsymbol{M}_{:,i}$ denotes the i-th column of $\boldsymbol{M}$. For $\ell > 1$, $\{\boldsymbol{M}_{:,i}, i \in \{1, \ldots, m\}\}$ is a subset of Δ_n that does not contain the extreme points of the simplex. ■

In practice, Proposition 3.2 means that it is impossible to reach exactly an arbitrary distribution $\boldsymbol{p} \in \Delta_n$, except for the trivial case of isolated measurements.

Proposition 3.3. *Suppose that Assumption 3.1 holds, then for $p \in [1, \infty]$,*

$$\|\boldsymbol{M}^*\|_{p \to \infty} = \ell^{-\frac{1}{p}}.$$

Proof. Under Assumpiton 3.1, all the columns of $\boldsymbol{M}$ have only ℓ non-zero coefficients equal to $1/\ell$. With $\|\cdot\|_F = \|\cdot\|_{\ell^p}$, we can thus derive that

$$\begin{aligned}\|\boldsymbol{M}^*\|_{p\to\infty} &= \max_{\|x\|_p=1} \|\boldsymbol{M}^* x\|_{\ell^\infty} = \max_{1\le i\le m} \max_{\|x\|_p=1} \langle \boldsymbol{M}_{:,i}, x\rangle \\ &= \max_{1\le i\le m} \|\boldsymbol{M}_{:,i}\|_{F^*} = \max_{1\le i\le m} \|\boldsymbol{M}_{:,i}\|_q \\ &= \ell^{-\frac{1}{p}},\end{aligned}$$

where $\boldsymbol{M}_{:,i}$ denotes the i-th column of $\boldsymbol{M}$, and q is the conjugate of p satisfying $1/p + 1/q = 1$. ■

Measuring the dissimilarity between $\boldsymbol{\pi}$ and $\boldsymbol{p}$ through the operator M

Now that we have introduced the mapping $\boldsymbol{M}$, we propose to define a dissimilarity measure between $\boldsymbol{\pi} \in \Delta_m$ and $\boldsymbol{p} \in \Delta_n$. To do so, we propose to compare $\boldsymbol{M}\boldsymbol{\pi}$ and $\boldsymbol{p}$ that are both vectors belonging to the simplex Δ_n. Owing to Proposition 3.2, it is hopeless to find some $\tilde{\boldsymbol{\pi}} \in \Delta_m$ satisfying $\boldsymbol{M}\tilde{\boldsymbol{\pi}} = \boldsymbol{p}$ for an arbitrary target density $\boldsymbol{p}$. Therefore, we can only expect to get an approximate solution by minimizing a dissimilarity measure $\mathcal{D}(\boldsymbol{M}\boldsymbol{\pi}, \boldsymbol{p})$. For obvious numerical reasons, $\mathcal{D}$ should be convex in $\boldsymbol{\pi}$. Among statistical distances, the most natural ones are the total variation distance, Kullback-Leibler of more generally f-divergences. Among this family, total variation presents the interest of having a dual of bounded support. We will exploit this property to design efficient numerical algorithms in Section 4. In the sequel, we will thus use $\mathcal{D}(\boldsymbol{M}\boldsymbol{\pi}, \boldsymbol{p}) = \|\boldsymbol{M}\boldsymbol{\pi} - \boldsymbol{p}\|_{\ell^1}$ to compare the distributions $\boldsymbol{M}\boldsymbol{\pi}$ and $\boldsymbol{p}$.

Entropic regularization

In applications such as MRI, the number m of columns of $\boldsymbol{M}$ is larger than the number n of its rows. Therefore, $\mathrm{Ker}(\boldsymbol{M}) \neq \emptyset$ and there exist multiple $\boldsymbol{\pi} \in \Delta_m$ with the same dissimilarity measure $\mathcal{D}(\boldsymbol{M}\boldsymbol{\pi}, \boldsymbol{p})$. In this case, we propose to take among all these solutions, the one minimizing the neg-entropy $\mathcal{E}$ defined by

$$\mathcal{E} : \boldsymbol{\pi} \in \Delta_m \longmapsto \sum_{j=1}^{m} \boldsymbol{\pi}_j \log(\boldsymbol{\pi}_j), \tag{5}$$

with the convention that $0\log(0) = 0$. We recall that the entropy $\mathcal{E}(\boldsymbol{\pi})$ is proportional to the Kullback-Leibler divergence between $\boldsymbol{\pi}$ and the uniform distribution $\boldsymbol{\pi}^c$ in Δ_m (i.e. such that $\boldsymbol{\pi}^c_j = \frac{1}{m}$ for all j). Therefore, among all the solutions minimizing $\mathcal{D}(\boldsymbol{M}\boldsymbol{\pi}, \boldsymbol{p})$, choosing the distribution $\boldsymbol{\pi}(\boldsymbol{p})$ minimizing $\mathcal{E}(\boldsymbol{\pi})$ gives priority to entropic solutions, i.e. probability distributions which maximize the covering of the sampling space if we proceed to several drawings of blocks of measurements. Therefore, we can finally write the following regularized problem defined by

$$\min_{\boldsymbol{\pi}\in\Delta_m} F_\alpha(\boldsymbol{\pi}), \tag{PP}$$

where

$$F_\alpha(\boldsymbol{\pi}) = \|\boldsymbol{M}\boldsymbol{\pi} - \boldsymbol{p}\|_{\ell^1} + \alpha\mathcal{E}(\boldsymbol{\pi}),$$

for some regularization parameter $\alpha > 0$. Adding the neg-entropy has the effect of spreading out the probability distribution $\boldsymbol{\pi}$, which is a desirable property. Moreover, the neg-entropy is strongly convex on the simplex Δ_m. This feature is of primary importance for the numerical resolution of the above optimization problem. Note that an appropriate choice of the regularization parameter α is also important, but this issue will not be addressed in this paper.

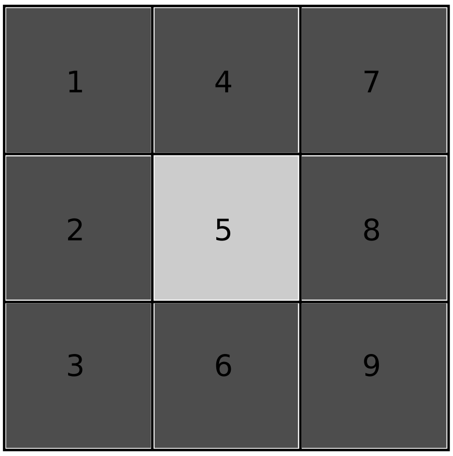


Figure 2: Illustration of a target distribution concentrated on the central pixel of a 3×3 images. The pixels are numbered, and this order is kept in the design of $\boldsymbol{M}$ and $\boldsymbol{\pi}$.

A toy example

To illustrate the interest of Problem (PP), we design a simple example. Consider a 3×3 image. Define the target distribution $\boldsymbol{p}$ as a dirac on the central pixel (numbered 5 in Figure 2). Consider a blocks dictionary composed of horizontal and vertical lines. In that setting, the operator $\boldsymbol{M}$ is given by

$$\boldsymbol{M} = \frac{1}{3}\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

For such a matrix, there are various distributions minimizing $\|\boldsymbol{M}\boldsymbol{\pi} - \boldsymbol{p}\|_{\ell^1}$. For example, one can choose $\boldsymbol{\pi}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}^*$ or $\boldsymbol{\pi}_2 = \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix}^*$. The solution maximizing the entropy is $\boldsymbol{\pi}_2$. In the case of image processing, this solution is preferable since it leads to better covering of the acquisition space. Note that, among all the ℓ^p-norms $(1 \leq p < +\infty)$, only the ℓ^1-norm is such that $\|\boldsymbol{M}\boldsymbol{\pi}_1 - \boldsymbol{p}\|_{\ell^1} = \|\boldsymbol{M}\boldsymbol{\pi}_2 - \boldsymbol{p}\|_{\ell^1}$. This property is once again desirable since we want the regularizing term (and not the fidelity term) to force choosing the proper solution.

4 Optimization

In this section, we propose a numerical algorithm to solve Problem (PP). Note that despite being convex, this optimization problem has some particularities that make it difficult to solve. Firstly, the parameter $\boldsymbol{\pi} \in \Delta_m$ lies in a very high dimensional space. In our experiments, n varies between 10^4 and 10^7 while m varies between 10^4 and 10^{10}. Moreover, the function $\mathcal{E}$ is differentiable but its gradient is not Lipschitz, and the total variation distance $\|\cdot\|_{\ell^1}$ is non-differentiable.

The numerical resolution of Problem (PP) is thus a delicate issue. Below, we propose an efficient strategy based on the numerical optimization of the dual problem of (PP), and on the use of Nesterov's ideas [Nes05]. Contrarily to most first order methods proposed recently in the literature [BC11, Nes13, CDV10] which are based on Hilbert space formalisms, Nesterov's algorithm is stated in a (finite dimensional) normed space. We thus perform the minimization of the dual problem on a metric space, and we carefully study the optimal choice of the norms in the primal and dual spaces. We show that depending on the blocks length ℓ, the optimal choice might well be different from the standard ℓ^2-norm. Such ideas stem back from (at least) [CT93], but were barely used in the domain of image processing.

4.1 Dualization of the problem

Our algorithm consists in solving the problem dual to (PP) in order to avoid the difficulties related to the non-differentiability of the ℓ^1-norm. Proposition 4.1 and 4.3 state that the dual of problem (PP) is differentiable. We will use this feature to design an efficient first-order algorithm and use the primal-dual relationships (Proposition 4.4) to retrieve the primal solution.

Proposition 4.1. *Let* $J_\alpha(\boldsymbol{q}) := \langle \boldsymbol{p}, \boldsymbol{q} \rangle_{F^* \times F} - \alpha \log \left(\sum_{\ell=1}^m \exp \left(-\frac{(\boldsymbol{M}^* \boldsymbol{q})_\ell}{\alpha} \right) \right)$, *for* $\boldsymbol{q} \in F$. *The dual problem to* (PP) *is:*

$$- \min_{\boldsymbol{q} \in B_\infty} J_\alpha(\boldsymbol{q}), \tag{DP}$$

in the sense that $\min_{\boldsymbol{\pi} \in \Delta_m} F_\alpha(\boldsymbol{\pi}) = \max_{\boldsymbol{q} \in B_\infty} -J_\alpha(\boldsymbol{q})$, *where* B_∞ *is the* ℓ^∞*-ball of unit radius in* F.

Proof. The proof is available in Appendix A. ■

In order to study the regularity properties of J_α, and so the solvability of (DP), we use the strong convexity of the neg-entropy $\mathcal{E}$ with respect to $\|\cdot\|_E$. First, let us recall one version of the definition of the strong convexity in Banach spaces.

Definition 4.1. *We say that* $f : F \to \mathbb{R}$ *is* σ*-strongly convex with respect to* $\|\cdot\|_F$ *on* $F' \subset F$ *if*

$$\forall x, y \in F', \quad \forall t \in [0,1], \quad f(tx + (1-t)y) \leq t f(x) + (1-t) f(y) - \frac{\sigma}{2} t(1-t) \|x - y\|_F^2. \tag{6}$$

We define the convexity modulus σ_f *of* f *as the largest positive real* σ *satisfying Equation* (6).

Proposition 4.2. *For* $\|\cdot\|_E = \|\cdot\|_{\ell^p}$, $p \in [1, +\infty]$, *the convexity modulus of the neg-entropy on the simplex* Δ_m *is* $\sigma_{\mathcal{E}} = 1$.

Proof. The proof is available in Appendix B. ■

Proposition 4.3. *The function* J_α *is convex and its gradient is Lipschitz continuous i.e.*

$$\|\nabla J_\alpha(\boldsymbol{q}_1) - \nabla J_\alpha(\boldsymbol{q}_2)\|_{F^*} \leq L_\alpha \|\boldsymbol{q}_1 - \boldsymbol{q}_2\|_F \qquad \forall (\boldsymbol{q}_1, \boldsymbol{q}_2) \in F^2.$$

with constant

$$L_\alpha = \frac{\|\boldsymbol{M}^*\|_{F \to E^*}^2}{\alpha \sigma_{\mathcal{E}}}. \tag{7}$$

Moreover, ∇J_α *is locally Lipschitz around* $\boldsymbol{q} \in F$ *with constant*

$$L_\alpha(\boldsymbol{q}) = \frac{\|\boldsymbol{M}^*\|_{F \to E^*}^2}{\alpha \sigma_{\mathcal{E}}(\boldsymbol{\pi}(\boldsymbol{q}))}, \tag{8}$$

where $\sigma_{\mathcal{E}}(\boldsymbol{\pi}) := \inf_{\|\boldsymbol{h}\|_E = 1} \left\langle \mathcal{E}''(\boldsymbol{\pi}) \boldsymbol{h}, \boldsymbol{h} \right\rangle$ *is the local convexity modulus of* $\mathcal{E}$ *around* $\boldsymbol{\pi}$, *and an explicit expression for* $\boldsymbol{\pi}(\boldsymbol{q})$ *is given in* (19).

Proof. The proof is available in Appendix C. ■

Note that a standard reasoning would rather lead to $L_\alpha = \frac{\|\boldsymbol{M}^*\|^2_{2\to 2}}{\alpha\sigma_{\mathcal{E}}}$, which is usually much larger than bound (7). Proposition 4.3 implies that Problem (DP) is efficiently solvable by Nesterov's algorithm [Nes05]. Therefore, we will first solve the dual problem (DP). Then, we use the relationships between the primal and dual solutions (as described in Proposition 4.4) to finally compute a primal solution $\boldsymbol{\pi}^\star$ for Problem (PP).

Proposition 4.4. *The relationships between the primal and dual solutions*

$$\boldsymbol{\pi}^\star = \operatorname*{arg\,min}_{\boldsymbol{\pi}\in\Delta_m} F_\alpha(\boldsymbol{\pi}) \qquad \textit{and} \qquad \boldsymbol{q}^\star = \operatorname*{arg\,min}_{\boldsymbol{q}\in B_\infty} J_\alpha(\boldsymbol{q})$$

are given by

$$\boldsymbol{\pi}^\star_j = \frac{\exp\left(-\frac{(\boldsymbol{M}^*\boldsymbol{q}^\star)_j}{\alpha}\right)}{\sum_{k=1}^m \exp\left(-\frac{(\boldsymbol{M}^*\boldsymbol{q}^\star)_k}{\alpha}\right)}, \qquad \forall j \in \{1,\dots,m\}\,. \tag{9}$$

Furthermore,

$$\operatorname{sign}\left(\boldsymbol{M}\boldsymbol{\pi}^\star - \boldsymbol{p}\right) = \operatorname{sign}\left(\boldsymbol{q}^\star\right). \tag{10}$$

Proof. Equation (9) is a direct consequence of (19). To derive the second equation (10), it suffices to write the optimality conditions of the problem $\max\limits_{\boldsymbol{q}\in B_\infty} \langle \boldsymbol{M}\boldsymbol{\pi}^\star - \boldsymbol{p}, \boldsymbol{q}\rangle_{F^*\times F} + \alpha\mathcal{E}(\boldsymbol{\pi}^\star)$. It leads to:

$$\boldsymbol{M}\boldsymbol{\pi}^\star - \boldsymbol{p} \in \mathcal{N}_{B_\infty}(\boldsymbol{q}^\star) \Leftrightarrow \operatorname{sign}\left(\boldsymbol{M}\boldsymbol{\pi}^\star - \boldsymbol{p}\right) = \operatorname{sign}\left(\boldsymbol{q}^\star\right).$$

■

4.2 Numerical optimization of the dual problem

Now that the dual problem (DP) is fully characterized, we propose to solve it using Nesterov's optimal accelerated projected gradient descent [Nes05] for smooth convex optimization.

4.2.1 The algorithm

Nesterov's algorithm is based on the choice of a prox-function d of the set B_∞, i.e. a continuous function that is strongly convex on B_∞ w.r.t. $\|\cdot\|_F$. Let σ_d denote the convexity modulus of d, we further assume that $d(\boldsymbol{q}_c) = 0$ so that

$$d(\boldsymbol{q}) \geq \frac{\sigma_d}{2}\|\boldsymbol{q} - \boldsymbol{q}_c\|^2_F \qquad \forall \boldsymbol{q} \in B_\infty,$$

where $\boldsymbol{q}_c = \operatorname*{arg\,min}\limits_{\boldsymbol{q}\in B_\infty} d(\boldsymbol{q})$. Nesterov's algorithm is described in Algorithm 1.

Theorem (4.5) summarizes the theoretical guarantees of Algorithm 1.

Theorem 4.5. *[Nes05, Theorem 2] Algorithm 1 ensures that*

$$\begin{aligned} J_\alpha(\boldsymbol{y}_k) - J_\alpha(\boldsymbol{q}^\star) &\leq \frac{4L_\alpha d(\boldsymbol{q}^\star)}{\sigma_d(k+1)(k+2)} \\ &\leq \frac{4\|\boldsymbol{M}^*\|^2_{F\to E^*} d(\boldsymbol{q}^\star)}{\alpha\sigma_{\mathcal{E}}\sigma_d(k+1)(k+2)}, \end{aligned} \tag{11}$$

where $\boldsymbol{q}^\star$ is an optimal solution of Problem (DP).

Algorithm 1 Resolution scheme for smooth optimization proposed by [Nes05]

```
Initialization: choose $\boldsymbol{q}_0 \in B_\infty$.
for $k = 0 \dots K$ do
    Compute $J_\alpha(\boldsymbol{q}_k)$ and $\nabla J_\alpha(\boldsymbol{q}_k)$
    Find $\boldsymbol{y}_k \in \operatorname*{arg\,min}_{\boldsymbol{y} \in B_\infty} \langle \nabla J_\alpha(\boldsymbol{q}_k), \boldsymbol{y} - \boldsymbol{q}_k \rangle + \frac{1}{2} L_\alpha \|\boldsymbol{y} - \boldsymbol{q}_k\|_F^2$
    Find $\boldsymbol{z}_k \in \operatorname*{arg\,min}_{\boldsymbol{q} \in B_\infty} \frac{L_\alpha}{\sigma_d} d(\boldsymbol{q}) + \sum_{i=0}^{k} \frac{i+1}{2} \left[ J_\alpha(\boldsymbol{q}_i) + \langle \nabla J_\alpha(\boldsymbol{q}_i), \boldsymbol{q} - \boldsymbol{q}_i \rangle \right]$
    Set $\boldsymbol{q}_{k+1} = \frac{2}{k+3} \boldsymbol{z}_k + \frac{k+1}{k+3} \boldsymbol{y}_k$.
end for
Set the primal solution to $\boldsymbol{\pi}_j = \frac{\exp\left(-\frac{(\boldsymbol{M}^* \boldsymbol{y}_K)_j}{\alpha}\right)}{\sum_{k=1}^m \exp\left(-\frac{(\boldsymbol{M}^* \boldsymbol{y}_K)_k}{\alpha}\right)}, \quad \forall j \in \{1, \dots, m\}$.
```

Since $d(\boldsymbol{q}^\star)$ is generally unknown, we can bound (11) by

$$\frac{4\|\boldsymbol{M}^*\|_{F \to E^*}^2 D}{\alpha \sigma_{\mathcal{E}} \sigma_d (k+1)(k+2)}. \tag{12}$$

where $D = \max_{\boldsymbol{q} \in B_\infty} d(\boldsymbol{q})$. Note that until now, we got theoretical guarantees in the dual space but not in the primal. What matters to us is rather to obtain guarantees on the primal iterates, which can be summarized by the following theorem.

Theorem 4.6. *Denote*

$$\boldsymbol{\pi}_k = \frac{\exp\left(-\frac{(\boldsymbol{M}^* \boldsymbol{y}_k)}{\alpha}\right)}{\left\|\exp\left(-\frac{(\boldsymbol{M}^* \boldsymbol{y}_k)}{\alpha}\right)\right\|_{\ell^1}}.$$

where $\boldsymbol{y}_k$ is defined in Algorithm 1. The following inequality holds:

$$\|\boldsymbol{\pi}_k - \boldsymbol{\pi}^\star\|_E^2 \leq \frac{8\|\boldsymbol{M}^*\|_{F \to E^*}^2 D}{\alpha^2 \sigma_{\mathcal{E}}^2 \sigma_d (k+1)(k+2)}.$$

The proof is given in Appendix D. It is a direct consequence of a more general result of independent interest.

4.2.2 Choosing the prox-function and the metrics

Algorithm 1 depends on the choice of $\|\cdot\|_E$, $\|\cdot\|_F$ and d. The usual accelerated projected gradient descents consist in setting $\|\cdot\|_E = \|\cdot\|_{\ell^2}$, $\|\cdot\|_F = \|\cdot\|_{\ell^2}$ and $d(\cdot) = \frac{1}{2}\|\cdot\|_{\ell^2}^2$. However, we will see that it is possible to change the algorithm's speed of convergence by making a different choice. In this paper we concentrate on the usual ℓ^p-norms, $p \in [1, +\infty]$.

Choosing a norm on E: The following proposition shows an optimal choice for $\|\cdot\|_{E^*}$.

Proposition 4.7. *The norm $\|\cdot\|_{E^*}$ that minimizes (12) among all ℓ^p-norms, $p \in [1, +\infty]$ is $\|\cdot\|_{\ell^\infty}$. Note however that the minimum local Lipschitz constant $L_\alpha(\boldsymbol{q})$ for $\boldsymbol{q} \in F$ might be reached for another choice of $\|\cdot\|_{E^*}$.*

Proof. From Proposition 4.2, we get that $\sigma_{\mathcal{E}}$ remains unchanged no matter how $\|\cdot\|_E$ is chosen among ℓ^p-norms. The choice of $\|\cdot\|_E$ is thus driven by the minimization of $\|\boldsymbol{M}^*\|_{F\to E^*}$. From the operator norm definition, it is clear that the best choice consists in setting $\|\cdot\|_{E^*} = \|\cdot\|_{\ell^\infty}$ since the ℓ^∞-norm is the smallest of all ℓ^p-norms. ■

According to Proposition 4.7, choosing $\|\cdot\|_{E^*}$ to be $\|\cdot\|_{\ell^\infty}$ leads to consider $\|\cdot\|_E$ to be $\|\cdot\|_{\ell^1}$. As shown by Proposition 3.3, it is clear that the norm $\|\boldsymbol{M}^*\|_{F\to E^*}$ may vary a lot with respect to $\|\cdot\|_F$ for the particular operator $\boldsymbol{M}$ considered in this paper.

Choosing a norm on F and a prox-function d: by Proposition 4.7 the norm $\|\cdot\|_F$ and the prox function d should be chosen in order to minimize $\frac{\|\boldsymbol{M}^*\|^2_{F\to\infty} D}{\sigma_d}$. We are unaware of a general theory to make an optimal choice despite recent progresses in that direction. The recent paper [dJ13] proposes a systematic way of selecting $\|\cdot\|_F$ and d in order to make the algorithm complexity invariant to change of coordinates for a general optimization problem. The general idea in [dJ13] is to choose $\|\cdot\|_F$ to be the Minkowski gauge of the constraints set (of the optimization problem), and d to be a strongly convex approximation of $\frac{1}{2}\|\cdot\|^2_F$. However, this strategy is not shown to be optimal. In our setting, since the constraints set is B_∞, this would lead to choose $\|\cdot\|_F = \|\cdot\|_{\ell^\infty}$. Unfortunately, there is no good strongly convex approximation of $\frac{1}{2}\|\cdot\|^2_{\ell^\infty}$.

In this paper, we thus study the influence of $\|\cdot\|_F$ and d both theoretically and experimentally, with $\|\cdot\|_F \in \{\|\cdot\|_{\ell^1}, \|\cdot\|_{\ell^2}, \|\cdot\|_{\ell^\infty}\}$. Propositions 4.8, 4.9 and 4.10 summarize the theoretical algorithm complexity in different regimes.

Proposition 4.8. *Let $p' \in]1,2]$. Define $d_{p'}(x) = \frac{1}{2}\|x\|^2_{p'}$. Then*

- *For $p \in [p',\infty]$, $d_{p'}$ is $(p'-1)$-strongly convex w.r.t. $\|\cdot\|_p$.*
- *For $p \in [1,p']$, $d_{p'}$ is $(p'-1)n^{(1/p'-1/p)}$-strongly convex w.r.t. $\|\cdot\|_p$.*

Proof. The proof is a direct consequence of [JN08, Proposition 3.6] and of the fact that for $p' \geq p$,

$$\|x\|_{p'} \leq \|x\|_p \leq n^{(1/p-1/p')}\|x\|_{p'}.$$

■

Proposition 4.9. *Suppose that Assumption 3.1 holds. Set $\|\cdot\|_F = \|\cdot\|_p$ and $d = d_{p'}$ with $p \in [1,\infty]$ and $p' \in]1,2]$. For all this family of norms and prox-functions, the one minimizing the complexity bound (12) is*

- *$p' = 2$ and $p \in [1,2]$, if $\ell^2 = n$. For this choice, we get*

$$J_\alpha(\boldsymbol{y}_k) - J_\alpha(\boldsymbol{q}^\star) \leq \frac{2\sqrt{n}}{\alpha(k+1)(k+2)}. \tag{13}$$

- *$p = p' = 2$, if $\ell^2 < n$. For this choice, we get*

$$J_\alpha(\boldsymbol{y}_k) - J_\alpha(\boldsymbol{q}^\star) \leq \frac{2n}{\alpha\ell(k+1)(k+2)}. \tag{14}$$

- *$p = 1$ and $p' = 2$, if $\ell^2 > n$. For this choice, we get*

$$J_\alpha(\boldsymbol{y}_k) - J_\alpha(\boldsymbol{q}^\star) \leq \frac{2n^{3/2}}{\alpha\ell^2(k+1)(k+2)}. \tag{15}$$

Proof. The result is a direct consequence of Proposition 4.8. ■

Unfortunately, the bounds in (13), (14) and (15) are dimension dependent. Moreover, the optimal choice suggested by Proposition 4.9 is different from the Minkowski gauge approach suggested in [dJ13]. Indeed, in all the cases described in Proposition 4.5, the optimal choice $\|\cdot\|_F$ differs from $\|\cdot\|_{\ell^\infty}$. The difficulty to apply this approach is to find a function $d \simeq 1/2\|\cdot\|_{\ell^\infty}^2$ strongly convex w.r.t. $\|\cdot\|_{\ell^\infty}$. A simple choice consists in setting $d_\varepsilon = \frac{1}{2}\|\cdot\|_{\ell^\infty}^2 + \frac{\varepsilon}{2}\|\cdot\|_{\ell^2}^2$. This function is ε-strongly convex w.r.t. $\|\cdot\|_{\ell^\infty}$. We thus get the following proposition:

Proposition 4.10. *Suppose that Assumption 3.1 holds, with $\ell = \sqrt{n}$. Set $\|\cdot\|_F = \|\cdot\|_{\ell^\infty}$, $d_\varepsilon(\cdot) = \frac{1}{2}\|\cdot\|_{\ell^\infty}^2 + \frac{\varepsilon}{2}\|\cdot\|_{\ell^2}^2$.*

$$J_\alpha(\boldsymbol{y}_k) - J_\alpha(\boldsymbol{q}^\star) \leq \frac{2\,(1/\varepsilon + n)}{\alpha(k+1)(k+2)}.$$

In particular, for $\varepsilon \propto \frac{1}{n}$, $J_\alpha(\boldsymbol{y}_k) - J_\alpha(\boldsymbol{q}^\star) = O\left(\frac{n}{\alpha k^2}\right)$.

Note that this complexity bound is worse than that of Proposition 4.9 in the case where $\ell = \sqrt{n}$. In the next section, we intend to illustrate and to confirm in practice the different rates of convergence, predicted by the theoretical results in Proposition 4.9.

4.3 Numerical experiments on convergence

In this section, we are willing to emphasize the improvement achieved by appropriately choosing the norms $\|.\|_E$, $\|.\|_F$, and the prox-function d. To do so, we run experiments on a dictionary of blocks of measurements having all the same size $\ell = 256$, described in Section 5.1, for 2D images of size 256×256. At first, we choose $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^2}$ and $d = \frac{1}{2}\|.\|_{\ell^2}^2$ and we perform Algorithm 1 for this dictionary. In fact, this first case (the norm on E differs from the usual $\|\cdot\|_{\ell^2}$) nearly corresponds to a standard accelerated gradient descent [NN04]. In a second time, we set $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^\infty}$ $d = \frac{1}{2}\|.\|_{\ell^2}^2$. In Figure 3, we display the decrease of the objective function in both settings. Figure 3 points out that a judicious selection of norms on E and F can significantly speed up the convergence: for 29 000 iterations, the standard accelerated projected gradient descent reaches a precision of 10^{-5} whereas Algorithm 1 with $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^\infty}$, i.e. a "modified" gradient descent, reaches a precision of 10^{-3}. The conclusions for this numerical experiment appear to be faithful to what was predicted by the theory, see Proposition 4.5. For the sake of completeness, we add in Figure 3 (in green) the case where $\|.\|_E = \|.\|_{\ell^2}$, $\|.\|_F = \|.\|_{\ell^2}$ and $d = \frac{1}{2}\|.\|_{\ell^2}^2$, which is an usual choice in practice. Clearly, this is the slowest rate of convergence observed.

Finally, we perform the algorithm for $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^2}$, and $d = \frac{1}{2}\|.\|_{\ell^2}^2$ by changing the value of L_α. The value of L_α provided by Proposition 4.3 is tight uniformly on B_∞. However, the local Lipschitz constant of ∇J_α varies rapidly inside the domain. In practice, the Lipschitz constant around the minimizer may be much smaller than L_α (note that $\pi^\star \in \mathrm{ri}(\Delta_m)$ for all $\alpha > 0$). In this last heuristic approach, we will decrease L_α by substantial factors without losing practical convergence. This result is presented in Figure 3 where the black curve denotes convergence result when the Lipschitz constant L_α has been divided by 100. We can observe that in this case, it suffices 1500 iterations to reach the precision obtained by the case $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^2}$ and $d = \frac{1}{2}\|.\|_{\ell^2}^2$ (in red) after 29000 iterations. Let us give an intuitive explanation to this positive behaviour. To simplify the reasoning, let us assume that $\boldsymbol{\pi}^\star$ is the uniform probability distribution. First notice that the choice of $\|\cdot\|_E$ only influences the Lipschitz constant of ∇J_α but does not change the algorithm, so that we can play with the norm on E to decrease the local Lipschitz constant. Furthermore, the choice of $\|\cdot\|_E$ minimizing the global Lipschitz constant may be different from the one minimizing the local Lipschitz constant. Considering that $\|\cdot\|_E = \|\cdot\|_{\ell^2}$, from Equation (8), we get that $L_\alpha(\boldsymbol{q}^\star) = \frac{\|\boldsymbol{M}^*\|_{2\to 2}^2}{\alpha\sigma_E(\boldsymbol{\pi}^\star)}$. Using Perron-Frobenius theorem, it can be shown that $\|\boldsymbol{M}^*\|_{2\to 2}^2 = O(1)$ for our choice of dictionary, and $\sigma_E(\boldsymbol{\pi}^\star) = m$ for $\|\cdot\|_E = \|\cdot\|_{\ell^2}$. From this simple reasoning, we can conclude that the local Lipschitz constant around $\boldsymbol{\pi}^\star$ is no greater than $O(1/m)$. This means that if the minimizer

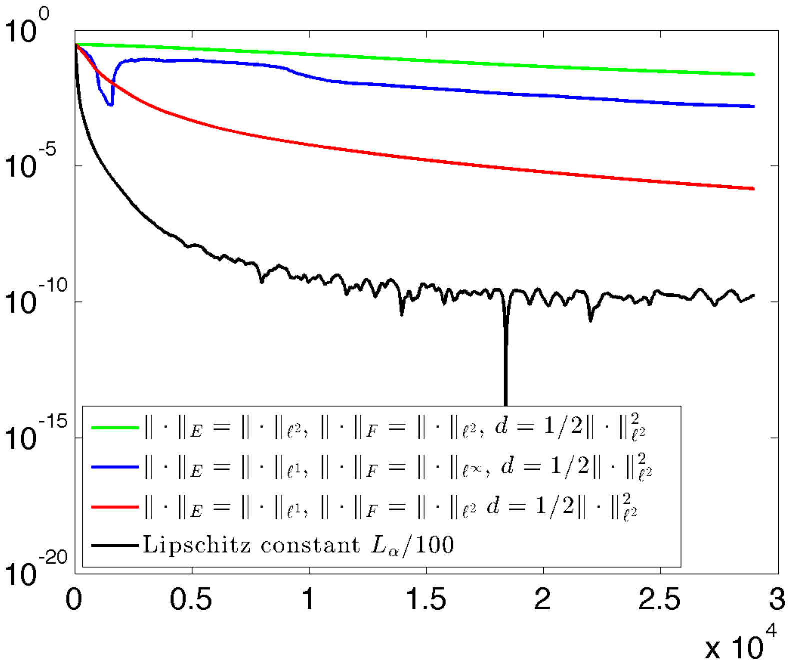


Figure 3: Convergence curves in a semi-logarithmic scale for Algorithm 1 ($\alpha = 10^{-2}$) (number of iterations on the x-axis) in green the case where $\|.\|_E = \|.\|_{\ell^2}$, $\|.\|_F = \|.\|_{\ell^2}$, $d = \frac{1}{2}\|.\|_{\ell^2}^2$, in red the case where $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^2}$, $d = \frac{1}{2}\|.\|_{\ell^2}^2$, in blue the case where $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^\infty}$ $d = \frac{1}{2}\|.\|_{\ell^2}^2$, and in black the case where $\|.\|_E = \|.\|_{\ell^1}$, $\|.\|_F = \|.\|_{\ell^2}$, $d = \frac{1}{2}\|.\|_{\ell^2}^2$ with a restricted Lipschitz constant $L'_\alpha = L_\alpha/100$.

is sufficiently far away from the simplex boundary, we can decrease L_α by a significant factor without loosing convergence.

5 Numerical results

In this section, we assess the reconstruction performance of the sampling patterns using the approach described in Section 4.2 with $\alpha = 10^{-2}$. We compare it to standard approaches used in the context of MRI. We call $\boldsymbol{\pi}[\boldsymbol{p}]$ the probability distribution $\boldsymbol{\pi}^\star$ resulting from the minimization problem (PP) for a given target distribution $\boldsymbol{p}$ on isolated measurements.

5.1 The choice of a particular dictionary of blocks

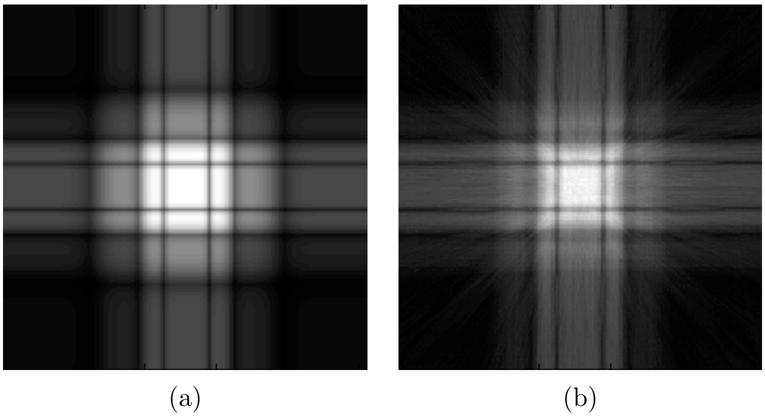
From a numerical point of view, we study a particular system of blocks of measurements. The dictionary used in all numerical experiments of this article is composed of discrete lines of length ℓ, joining any pixel on the edge of the image to any pixel on the opposite edge, as in Figure 1(b). Note that the number of blocks in this dictionary is $n_1^2+n_2^2$ for an image of size $n_1 \times n_2$. The choice of such a dictionary is particularly relevant in MRI, since the gradient waveforms that define the acquisition paths is subject to bounded-gradient and slew-rate constraints, see e.g. [LKP08]. Moreover the practical implementation on the scanner of straight lines is straightforward since it is already in use in standard echo-planar imaging strategies.

Remark that, in such a setting, the mapping $\boldsymbol{M}$, defined in (4), is a linear mapping that can be represented by a matrix of size $n \times m$ with $\boldsymbol{M}_{i,j} = 1/\ell$ when the i-th pixel belongs to the j-th block, for $i = 1, \dots, n$ and $j = 1, \dots, m$.

One may argue that in MRI, dealing with samples lying on continuous lines (and not discrete grids) is more realistic in the design of the MR sequences. To deal with this issue, one could resort to the use of the Non-Uniform Fast Fourier Transform. This technique is however much more computationally intensive. In this paper we thus stick to values of the Fourier transform located on the Euclidean grid. This is commonly used in MRI with regridding techniques.

5.2 The reconstructed probability distribution

We are willing to illustrate the fidelity of $\boldsymbol{\pi}[\boldsymbol{p}]$, the solution of Problem (PP), to a given target $\boldsymbol{p}$. In the setting of 2D MR sensing, with the dictionary of lines in dimension $n_1 = n_2 = 256$ described in the previous subsection. We set the target probability distribution $\boldsymbol{p} = \boldsymbol{p}_{\text{opt}}$ the one suggested by current CS theories on the set of isolated measurements. It is proportional to $\|\boldsymbol{a}_k^*\|_{\ell^\infty}^2$, see [PVW11, CCW13, BBW13]. To give an idea of what the resulting probability distribution $\boldsymbol{\pi}[\boldsymbol{p}_{\text{opt}}]$ looks like, we draw 100 000 independent blocks of measurements according to $\boldsymbol{\pi}[\boldsymbol{p}_{\text{opt}}]$ and count the number of measurement for each discrete Fourier coefficient. The result is displayed on Figure 4. This experiment underlines that our strategy manages to catch the overall structure of the target probability distibution.


(a)

(b)

Figure 4: Illustration of the fidelity of $\boldsymbol{\pi}[\boldsymbol{p}_{\text{opt}}]$ to $\boldsymbol{p}_{\text{opt}}$. (a): on the left hand side, we present the target probability distribution $\boldsymbol{p}_{\text{opt}}$ (b): on the right hand side, we perform 100000 i.i.d. drawings according to $\boldsymbol{\pi}[\boldsymbol{p}_{\text{opt}}]$ of blocks from the blocks dictionary and count the number of times that a point is sampled at each location.

5.3 Reconstruction results

In this section, we compare the reconstruction quality of MR images for different acquisition schemes. The comparison is always performed for schemes with an equivalent number of isolated measurements. We recall that in the case of MR images, the acquisition is done in the Fourier domain, and MR images are supposed to be sparse in the wavelet domain. Therefore, the full sensing matrix $\boldsymbol{A} = (\boldsymbol{a}_1|\boldsymbol{a}_2|\ldots|\boldsymbol{a}_n)^*$, which models the acquisition process, is the composition of a Fourier transform with an inverse Wavelet transform. The reconstruction is done via ℓ^1-minimization as presented in (1), using Douglas-Rachford algorithm [CP11b]. It was proven in various papers [CCW13, CCKW13, AHPR13] that MRI image quality can be strongly improved by fully acquiring the center of the Fourier domain via a mask defined by the support of the mother wavelet, see Figure 5. Therefore, for every type of schemes used in our reconstruction test, we first fully acquire this mask.

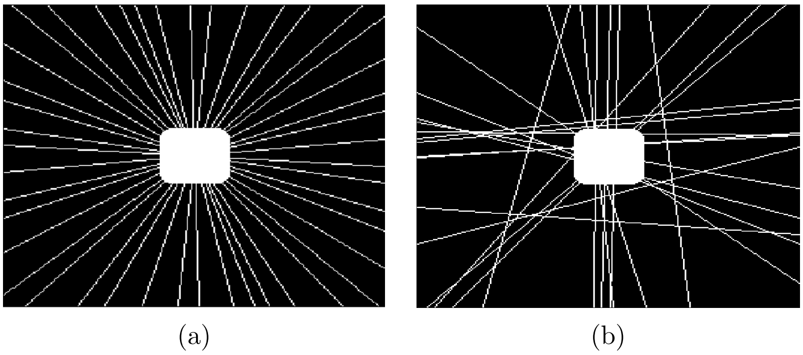


Figure 5: Different schemes based (a) on the golden angle pattern, and (b) on the dictionary proposed in Section 5.1. Both schemes are combined with a mask which fully samples the center of the Fourier domain. In both cases, the proportion of total measurements represents 10% of the full image, while the mask defined by the support of the mother wavelet represents 3% of the full image.

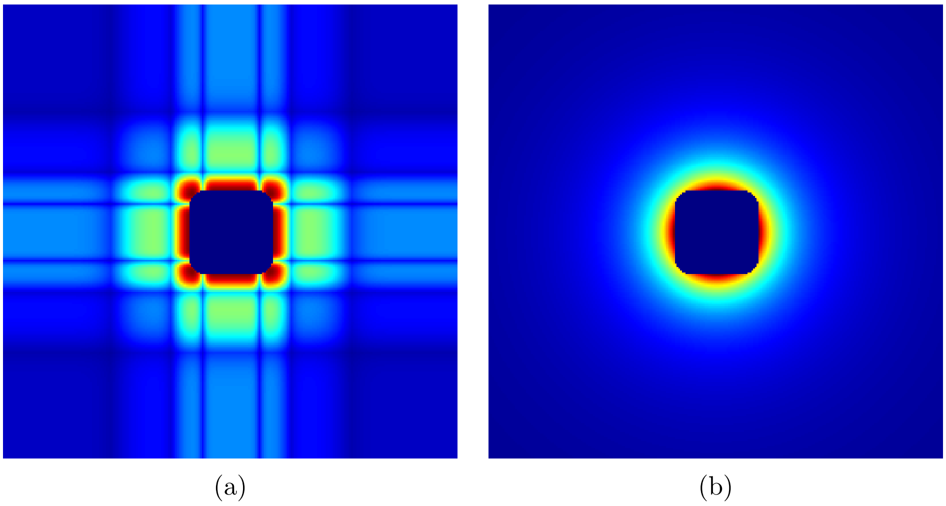
The various schemes considered in this paper are based on blocks of measurements and on heuristic schemes that are widely used in the context of MRI. They will consist in:

- Equiangularly distributed radial lines: the scheme is made of lines always intersecting the center of the acquisition domain, and that are distributed uniformly [LDSP08].
- Golden angle scheme: the scheme is made of radial lines separated from the golden angle, i.e 111.246°. This technique is used often in MRI sequences, and it gives good reconstruction results in practice [WSK+07].
- Random radial scheme: radial lines are drawn uniformly at random [CRCP12].
- Scheme based on the dictionary described in Section 5.1
  - Blocks are drawn according to $\boldsymbol{\pi}\left[\boldsymbol{p}_{\mathrm{rad}}\right]$ which is the resulting probability distribution obtained by minimizing Problem (PP) for $\boldsymbol{p} = \boldsymbol{p}_{\mathrm{rad}}$. The distribution $\boldsymbol{p}_{\mathrm{rad}}$ a radial distribution that decreases as $\mathcal{O}\left(\frac{1}{k_x^2+k_y^2}\right)$. This choice was justified recently in [KW12] and used extensively in practice. Note that $\boldsymbol{p}_{\mathrm{rad}}$ is set to 0 on the k-space center since it is already sampled deterministically, see Figure 6 (b).
  - Blocks are drawn according to $\boldsymbol{\pi}\left[\boldsymbol{p}_{\mathrm{opt}}\right]$, where $\boldsymbol{p}_{\mathrm{opt}}$ is defined by (3), which is the resulting probability distribution obtained by minimizing Problem (PP) for $\boldsymbol{p} = \boldsymbol{p}_{\mathrm{opt}}$ defined in [CCW13, BBW13]. Once again, $\boldsymbol{p}_{\mathrm{opt}}$ is set to 0 on the k-space center, see Figure 6 (a).

Setting 256×256

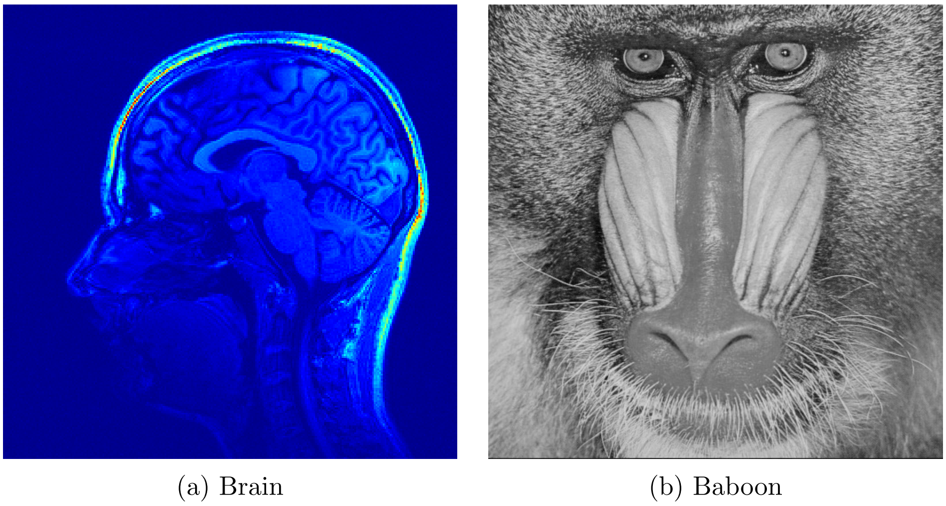
The numerical experiment is run for images of size $n_0 \times n_0$ with $n_0 = 256$. The full dictionary described in Section 5.1 contains lines of length $\ell = n_0$ pixels connecting every point on the edge of the image to every point on the opposite side. For each proportion of measurements $(10\%, 15\%, 20\%, 25\%, 30\%, 40\%, 50\%)$, we proceed to 100 drawings of schemes when the considered scheme is random. Reconstruction results, for the reference images showed in Figure 7 and for various sampling schemes, are displayed in the form of boxplots of PSNR in Figure 8 (a)(c).

Figure 8 shows that the schemes based on the approach presented in this article give better results than random radial schemes, for any proportion of measurements. The improvement in terms of PSNR is generally between 1 and 2 dB. The schemes based on $\boldsymbol{\pi}\left[\boldsymbol{p}_{\mathrm{opt}}\right]$ and $\boldsymbol{\pi}\left[\boldsymbol{p}_{\mathrm{rad}}\right]$ are competitive with those based on the golden angle or equiangularly distributed schemes in the



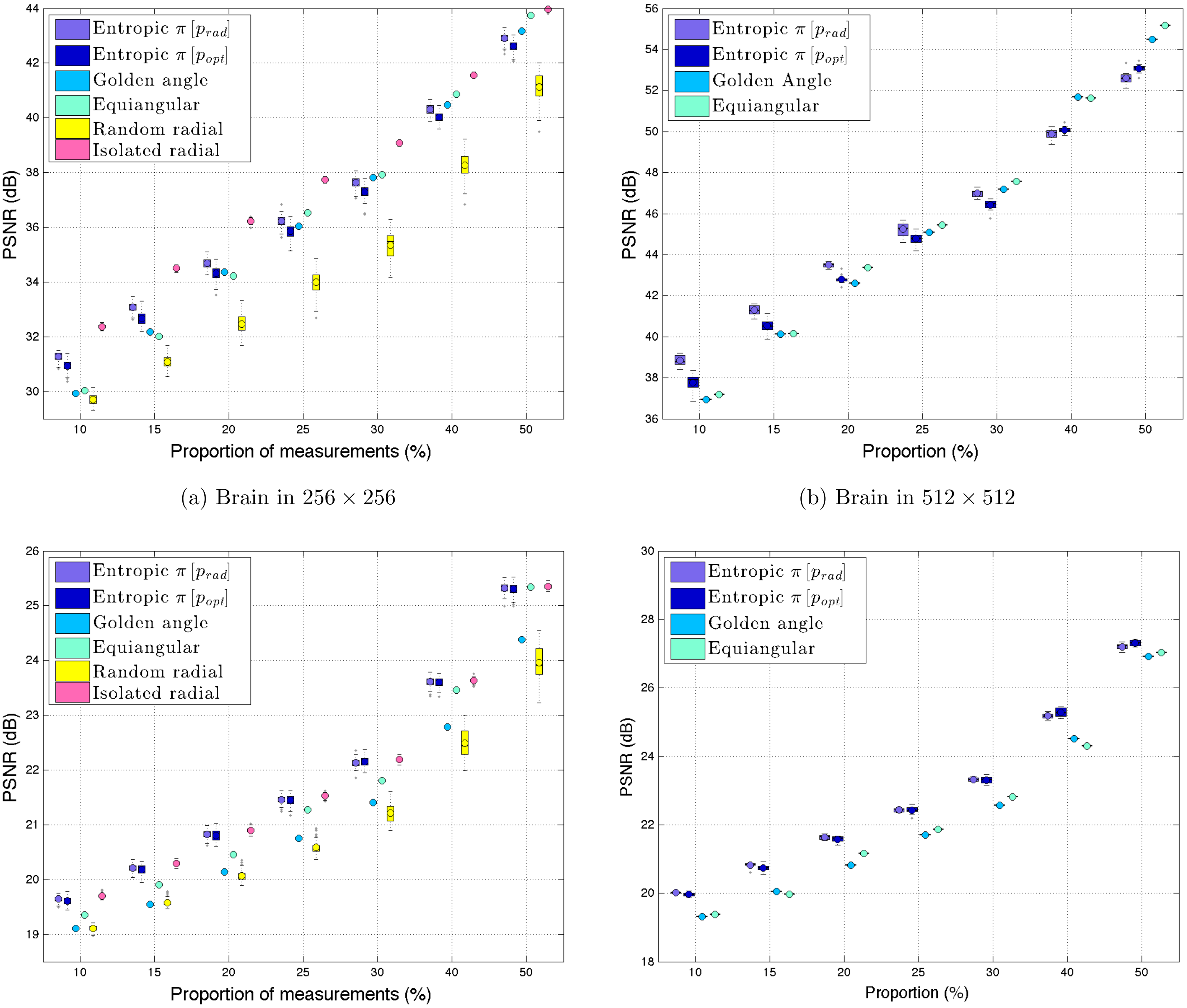
(a) (b)

Figure 6: Target probabilities on pixels (in red, high values, and in dark blue, values close to 0). (a) displays the distribution proportional to $\|\boldsymbol{a}_i^*\|_{\ell^\infty}^2$ defined in [CCW13], (b) displays a radial distribution as presented in [KW12]. The center has been set to zero, since it will be sampled by the mask in a deterministic way.



(a) Brain (b) Baboon

Figure 7: Reference images to reconstruct for the settings 256 × 256 and 512 × 512 .



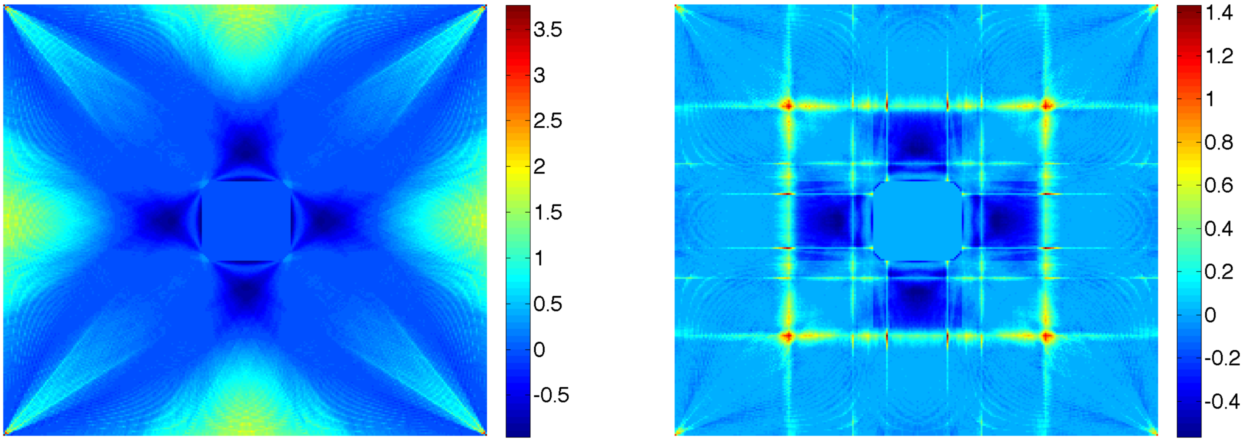
(a) Brain in 256 × 256

(b) Brain in 512 × 512

(c) Baboon in 256 × 256

(d) Baboon in 512 × 512

Figure 8: Box plots for PSNR of the reconstructed images (brain, baboon) with respect to the proportion of measurements chosen in the scheme $(10\%, 15\%, 20\%, 25\%, 30\%, 40\%, 50\%)$. The undersampling ratio for all schemes is the ratio between the number of sampled distinct frequencies and the total number of possible measurements. This means that duplicated frequencies are accounted for only once.



(a) (b)

Figure 9: Difference between the target probabilities $\boldsymbol{p}$ and $\boldsymbol{M}\boldsymbol{\pi}(\boldsymbol{p})$ relatively to the magnitude of $\boldsymbol{p}$, i.e. we show the following quantity $\frac{(\boldsymbol{M}(\boldsymbol{\pi}(\boldsymbol{p})))_i-\boldsymbol{p}_i}{\boldsymbol{p}_i}$ for the i-th sampling location, (a) for the radial distribution $\boldsymbol{p}_{\text{rad}}$, we see that we "sub-draw" by a factor of 50 % around the mask, and we "over-draw" by a factor of 150 % at the center of the edges. (b) for the CS optimal distribution $\boldsymbol{p}_{\text{opt}}$, we see that we "sub-draw" by a factor of 40 % around the mask. Note that the sub-drawing effect cannot be avoided: indeed, we cannot reach any target probability distribution via $\boldsymbol{M}$, see Proposition 3.2.

case where the proportion of measurements is low (less than 20% of measurements). We observe that for 10% measurements, schemes based on our dictionary and drawn according to $\boldsymbol{\pi}\left[\boldsymbol{p}_{\text{rad}}\right]$ outperform by more than 1 dB the standard sampling strategies. Increasing the PSNR of 1dB is significant and can be qualitatively observed in the reconstructed image.

Figures 8(a) and (c) allow to compare the quality of the reconstructions using different sampling schemes and different undersampling ratios. In this experiment, it can be seen that block-constrained acquisition never outperforms acquisitions based on indepenedent measurements. This was to be expected since adding constraints reduces the space of possible sampling patterns. Once again, note that independent drawings are however not conceivable in many contexts such as MRI. In this Figure it can also be seen that the proposed sampling approach always produces results comparable to the standard sampling schemes and tend to produce better results for low sampling ratios.

Finally, in Figure 9, we illustrate that block-constrained acquisition does not allow to reach an arbitrary target distribution by showing the difference between $\boldsymbol{p}_{\text{rad}}$ and the probability distribution $\boldsymbol{M}\left(\boldsymbol{\pi}\left[\boldsymbol{p}_{\text{rad}}\right]\right)$ which is defined on the set of isolated measurements. This confirms Proposition 3.2 experimentally.

Setting 512×512

Given that in CS the quality of the reconstruction can be resolution dependent, as described in [AHPR13], we have decided to run the same numerical experiment on 512×512 images. The numerical experiment is run for images of size $n_0 \times n_0$ with $n_0 = 512$. The full dictionary described in Section 5.1 contains lines of length $\ell = n_0$ connecting every point on the edge of the image to every point on the opposite side. For each proportion of measurements $(10\%, 15\%, 20\%, 25\%, 30\%, 40\%, 50\%)$, we proceed to 10 drawings of sampling schemes when the considered scheme is random. The images of reference to reconstruct are the same as in the setting 256 × 256, see Figure 7.

Quality of reconstructions are compared in Figure 8(b) and (d) for the golden or equiangularly distributed lines and our proposed method based on $\boldsymbol{\pi}(\boldsymbol{p}_{\text{opt}})$ and $\boldsymbol{\pi}(\boldsymbol{p}_{\text{rad}})$. This experiment shows that the PSNR of the reconstructed images is significantly improved by using the proposed method until 30% of measurements for the brain image and until 40% of measurements for the baboon one. We can remark that for both images, for a same proportion of measurements, the PSNR of the reconstructed images increases while the resolution increases. This numerical experiment suggests that the proposed sampling approach might be significantly better than traditional ones in the MRI context for high resolution images. In Figure 10 (a), we present the reconstructed image of the brain from 15% of measurements in the case of a golden angle scheme. In Figure 10 (b), we present the reconstructed image of the brain from 15% of measurements in the case of a realization of schemes based on $\boldsymbol{\pi}(\boldsymbol{p}_{\text{rad}})$. The latter's PSNR is 41.88 dB whereas in the golden scheme case, the PSNR only reaches 40.13 dB. In Figure 10 (c) and (d), we display the corresponding difference images to the reference image, which underlines the improvement of 1.7 dB in our method.

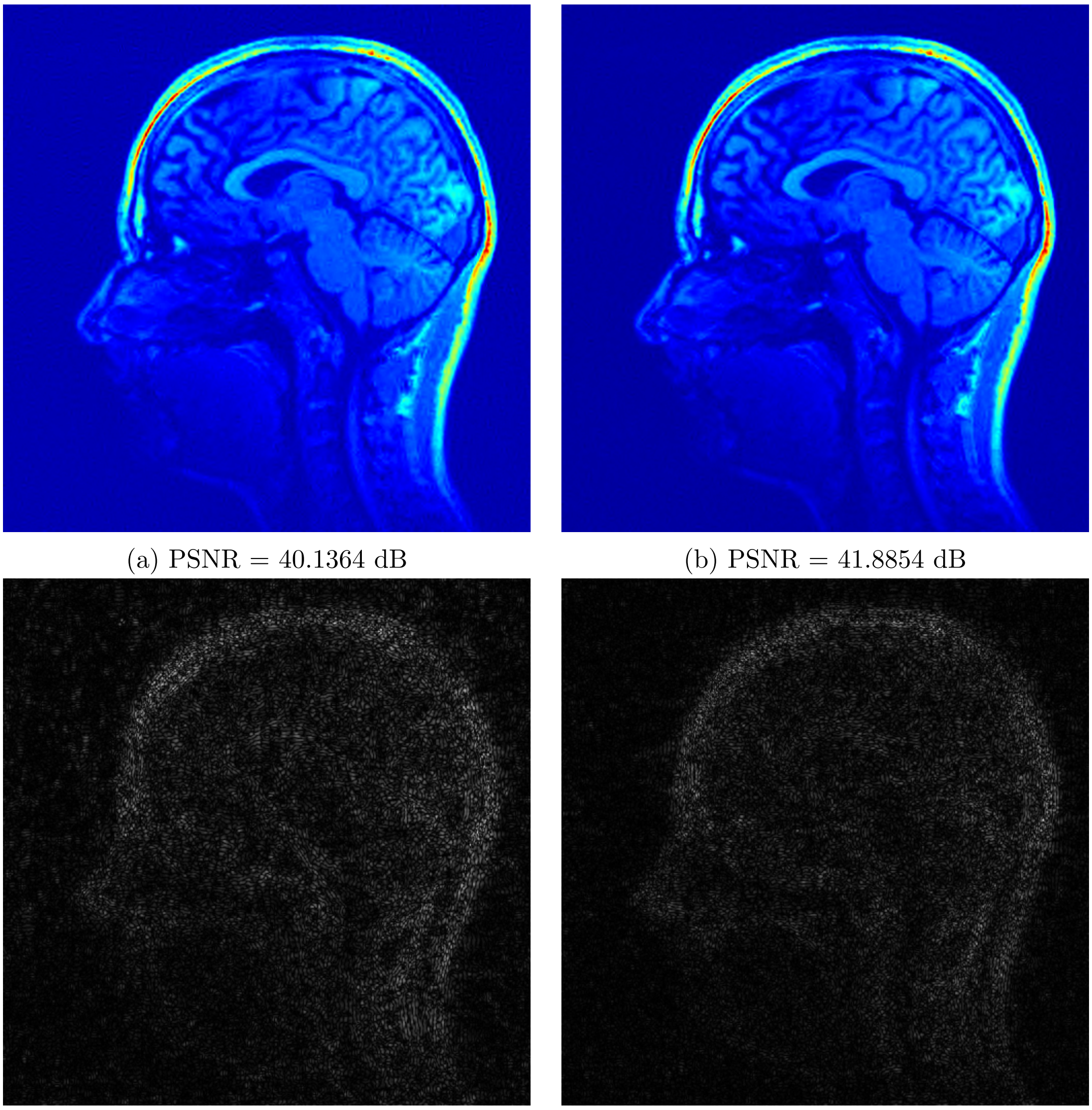
As a side remark, let us mention that in MRI, sampling diagonal or horizontal lines actually takes the same scanning time (even though the diagonals are longer), since gradient coils work independently in each direction. In the MRI example, the length of the path is thus less meaningful that the number of scanned lines. In Figure 11, we show different sampling schemes based on the golden angle pattern or on our method with the same number of lines, and we show the corresponding reconstructions of brain images.

Remark. In both settings, for the brain image, schemes based on $\boldsymbol{\pi}\left[\boldsymbol{p}_{\text{rad}}\right]$ lead to better reconstruction results in terms of PSNR than schemes from $\boldsymbol{\pi}\left[\boldsymbol{p}_{\text{opt}}\right]$. This can be explained by the fact that $\boldsymbol{p}_{\text{opt}}$ is the probability density given by CS theory which provides guarantees for any s-sparse image to reconstruct. However, brain images or natural images have a structured sparsity in the wavelet domain: indeed, their wavelet transform is not uniformly s-sparse, the approximation part contains more non-zero coefficients than the rest of the details parts. We can infer that $\boldsymbol{p}_{\text{rad}}$ manages to catch the sparsity structure of the wavelet coefficients of the considered images.

6 Conclusion

In this paper, we have focused on constrained acquisition by blocks of measurements. Sampling schemes are constructed by drawing blocks of measurements from a given dictionary of blocks according to a probability distribution $\boldsymbol{\pi}$ that needs to be chosen in an appropriate way. We have presented a novel approach to compute $\boldsymbol{\pi}$ in order to imitate existing sampling schemes in CS that are based on the drawing of isolated measurements. For this purpose, we have defined a notion of dissimilarity measure between a probability distribution on a dictionary of blocks and a probability distribution on a set of isolated measurements. This setting leads to a convex minimization problem in high dimension. In order to compute a solution to this optimization problem, we proposed an efficient numerical approach based on the work of [Nes05]. Our numerical study highlights the fact that performing minimization on a metric space rather than a Hilbert space might lead to significant acceleration. Finally, we have presented reconstruction results using this new approach in the case of MRI acquisition. Our method seems to provide better reconstruction results than standard strategies used in this domain. We believe that this last point brings interesting perspectives for 3D MRI reconstruction.

As an outlook, we plan to extend the proposed numerical method to a wider setting and to provide better theoretical guarantees for cases where the Lipschitz constant of the gradient may vary across the domain. A first step in this direction was proposed recently in [GK13]. We also plan to accelerate the matrix-vectors product involving $\boldsymbol{M}$ by using fast Radon transforms. This step is unavoidable to apply our algorithm in 3D or 3D+t problems for which we expect



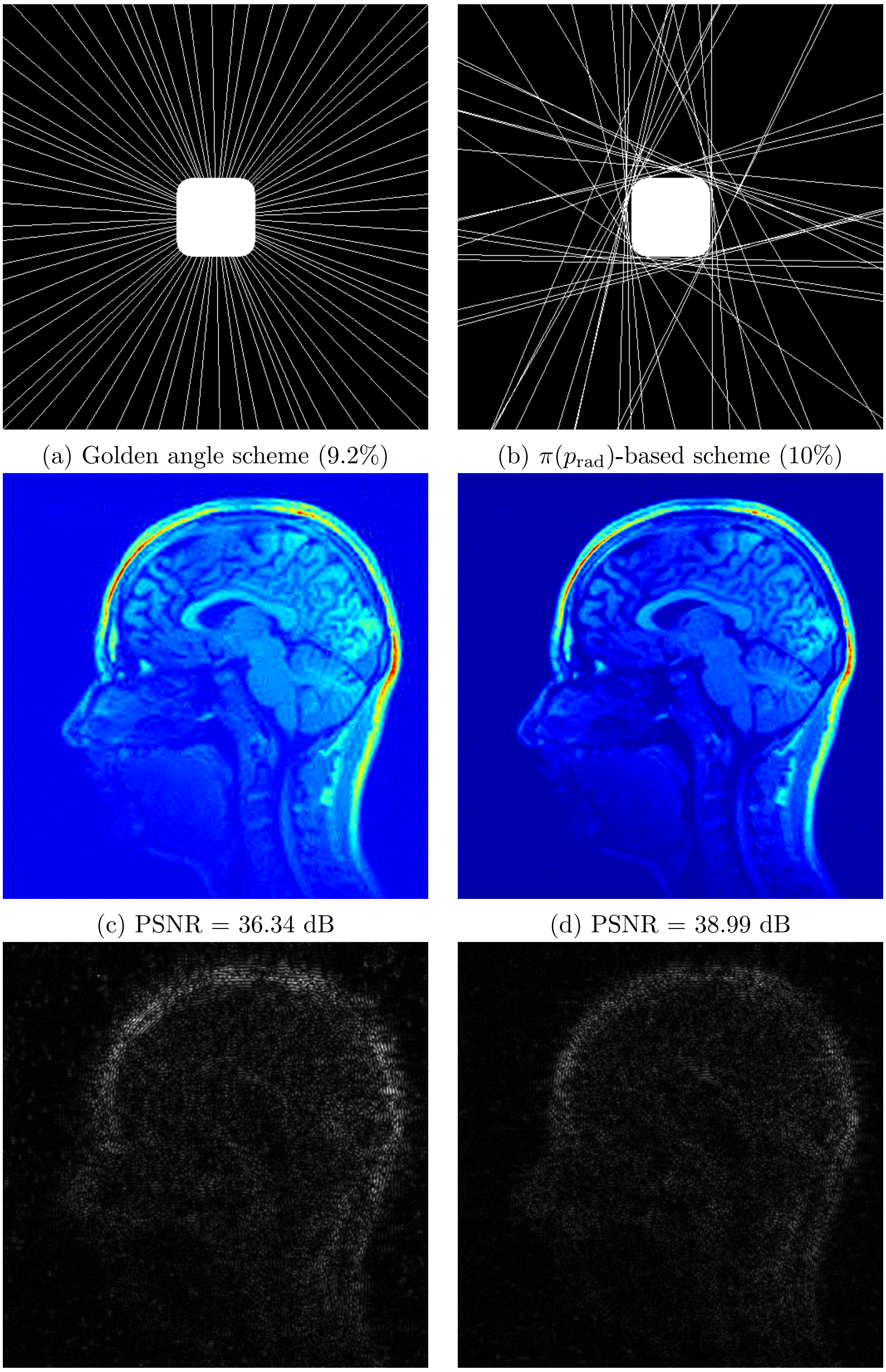
(a) PSNR = 40.1364 dB

(b) PSNR = 41.8854 dB

(c)

(d)

Figure 10: Comparison of the reconstructed images from 15% of measurements for a 512×512 image for a golden angle scheme (a), and a scheme based on our dictionary and $\boldsymbol{\pi}(\boldsymbol{p}_{\text{rad}})$ (b). We respectively plot the absolute difference to the reference image for the reconstruction using a golden angle scheme in (c) and for the reconstruction using a scheme based on $\boldsymbol{\pi}(\boldsymbol{p}_{\text{rad}})$ in (d). Note that in (c) and (d), the gray levels are in the same scale.



(a) Golden angle scheme (9.2%)

(b) $\pi(p_{\text{rad}})$-based scheme (10%)

(c) PSNR = 36.34 dB

(d) PSNR = 38.99 dB

(e) Golden angle

(f) $\boldsymbol{\pi}(\boldsymbol{p}_{\text{rad}})$-based scheme

Figure 11: Reconstruction examples. We plot schemes made of 37 lines based on the golden angle pattern (a), or based on our method with $\boldsymbol{\pi}(\boldsymbol{p}_{\text{rad}})$ (b). Drawing 37 lines in both cases leads to a cover of the sampling space by 9.2% in the case of the golden angle scheme, and by 10% for the $\boldsymbol{\pi}(\boldsymbol{p}_{\text{rad}})$-based scheme. Note that despite a difference of 0.8% in the covering the k-space, the scanning time is the same for both schemes. In (c) and (d) we display the corresponding reconstructions via ℓ^1-minimization. We can see that we improve the reconstruction of more than 2 dB with our method. At the bottom, we show the corresponding absolute difference with the reference image. Note that the gray levels have the same scaling in (e) and (f).

important benefits compared to the small images we tested until now. Finally we are currently collaborating with physicists at Neurospin, CEA and plan to implement the proposed sampling schemes on real MRI scanners.

Acknowledgement

The authors wish to thank Jean-Baptiste Hiriart-Urruty, Fabrice Gamboa and Alexandre Vignaud for fruitful discussion. They also thank the referees for their remarks which helped clarifying the paper. This work was partially supported by the CIMI (Centre International de Mathématiques et d'Informatique) Excellence program, by ANR SPH-IM-3D (ANR-12-BSV5-0008), by the FMJH Program Gaspard Monge in optimization and operation research, and by the support to this program from EDF.

A Proof of Proposition 4.1

First, we express the Fenchel-Rockafellar dual problem [Roc97]:

$$\begin{aligned}
&\min_{\boldsymbol{\pi}\in\Delta_m} \|\boldsymbol{p}-\boldsymbol{M}\boldsymbol{\pi}\|_{\ell^1} + \alpha\mathcal{E}(\boldsymbol{\pi}) \\
&\qquad = \min_{\boldsymbol{\pi}\in\Delta_m} \max_{\boldsymbol{q}\in B_\infty} \langle \boldsymbol{M}\boldsymbol{\pi}-\boldsymbol{p},\boldsymbol{q}\rangle_{F^*\times F} + \alpha\mathcal{E}(\boldsymbol{\pi}) \\
&\qquad = \max_{\boldsymbol{q}\in B_\infty} \min_{\boldsymbol{\pi}\in\Delta_m} \langle \boldsymbol{M}^*\boldsymbol{q},\boldsymbol{\pi}\rangle_{E^*\times E} - \langle \boldsymbol{p},\boldsymbol{q}\rangle_{F^*\times F} + \alpha\mathcal{E}(\boldsymbol{\pi}) \\
&\qquad = \max_{\boldsymbol{q}\in B_\infty} -J_\alpha(\boldsymbol{q})
\end{aligned}$$

where B_∞ stands for the ℓ^∞-ball of unit radius and

$$J_\alpha(\boldsymbol{q}) = -\min_{\boldsymbol{\pi}\in\Delta_m} \langle \boldsymbol{M}^*\boldsymbol{q},\boldsymbol{\pi}\rangle_{E^*\times E} - \langle \boldsymbol{p},\boldsymbol{q}\rangle_{F^*\times F} + \alpha\mathcal{E}(\boldsymbol{\pi}). \tag{16}$$

The solution $\boldsymbol{\pi}(\boldsymbol{q})$ of the minimization problem (16) satisfies

$$\boldsymbol{M}^*\boldsymbol{q} + \alpha\left(\log(\boldsymbol{\pi}(\boldsymbol{q}))+1\right) \in -\mathcal{N}_{\Delta_m}(\boldsymbol{\pi}(\boldsymbol{q})) \quad \text{if} \quad \boldsymbol{\pi}(\boldsymbol{q}) \in \operatorname{ri}(\Delta_m), \tag{17}$$

where $\mathcal{N}_{\Delta_m}(\boldsymbol{\pi}(\boldsymbol{q}))$ denotes the normal cone to the set Δ_m at the point $\boldsymbol{\pi}(\boldsymbol{q})$, and $\operatorname{ri}(\Delta_m)$ denotes the relative interior of Δ_m. Equation (17) can be rewritten in the following way

$$\boldsymbol{M}^*\boldsymbol{q} + \alpha\log(\boldsymbol{\pi}(\boldsymbol{q})) = (-\lambda-\alpha)\mathbb{1}, \text{with} \quad \lambda\in\mathbb{R}^+ \quad \text{and} \quad \boldsymbol{\pi}(\boldsymbol{q})\in\Delta_m. \tag{18}$$

By choosing $\lambda = \alpha\log\left(\sum_{k=1}^m \exp\left(-\frac{(\boldsymbol{M}^*\boldsymbol{q})_k}{\alpha}\right)\right) - \alpha$ and plugging it into (18) we get that

$$(\boldsymbol{\pi}(\boldsymbol{q}))_j = \frac{\exp\left(-\frac{(\boldsymbol{M}^*\boldsymbol{q})_j}{\alpha}\right)}{\sum_{k=1}^m \exp\left(-\frac{(\boldsymbol{M}^*\boldsymbol{q})_k}{\alpha}\right)}, \qquad \forall j\in\{1,\dots,m\}. \tag{19}$$

It remains to plug (19) in (16) to obtain (DP) with

$$J_\alpha(\boldsymbol{q}) = \langle \boldsymbol{p},\boldsymbol{q}\rangle_{F^*\times F} - \alpha\log\left(\sum_{k=1}^m \exp\left(-\frac{(\boldsymbol{M}^*\boldsymbol{q})_k}{\alpha}\right)\right).$$

B Proof of Proposition 4.2

The neg-entropy is continuous, and twice continuously differentiable on $\operatorname{ri}(\Delta_m)$. Then, in order to prove its strong convexity, it is sufficient to bound from below its positive diagonal Hessian with respect to $\|\cdot\|_E$. We have

$$\left\langle \mathcal{E}''(\boldsymbol{\pi})\boldsymbol{h}, \boldsymbol{h} \right\rangle = \sum_{i=1}^{m} \frac{(h_i)^2}{\pi_i}, \qquad \text{for} \quad \boldsymbol{\pi} \in \operatorname{ri}(\Delta_m), \quad \text{and} \quad \boldsymbol{h} \in \mathbb{R}^m. \tag{20}$$

Using Cauchy-Schwartz's inequality,

$$\begin{aligned} \|\boldsymbol{h}\|_{\ell^1}^2 &= \left(\sum_{i=1}^{m} \frac{|h_i|}{\sqrt{\pi_i}}\sqrt{\pi_i}\right)^2 \leq \left(\sum_{i=1}^{m} \frac{h_i^2}{\pi_i}\right)\left(\sum_{i=1}^{m} \pi_i\right) \\ &\leq \underbrace{\|\boldsymbol{\pi}\|_{\ell^1}}_{=1} \left\langle \mathcal{E}''(\boldsymbol{\pi})\boldsymbol{h}, \boldsymbol{h} \right\rangle. \end{aligned}$$

Therefore, $\mathcal{E}$ is 1-strongly convex on the simplex with respect to $\|.\|_{\ell^1}$. Since for all $p \in [1, \infty]$, $\|.\|_{\ell^1} \geq \|.\|_p$, we get:

$$\|\boldsymbol{h}\|_{\ell^p}^2 \leq \left\langle \mathcal{E}''(\boldsymbol{\pi})\boldsymbol{h}, \boldsymbol{h} \right\rangle, \quad \forall \boldsymbol{h} \in \mathbb{R}^m, \boldsymbol{\pi} \in \operatorname{ri}(\Delta_m).$$

Moreover if $(\boldsymbol{\pi}_n)_{n\in\mathbb{N}}$ denotes a sequence of $\operatorname{ri}(\Delta_m)$ pointwise converging to the first element of the canonical basis e_1 and $h = e_1$, then

$$\lim_{n\to+\infty} \langle \mathcal{E}''(\boldsymbol{\pi}_n)h, h\rangle = \|h\|_{\ell^p}^2 = 1$$

so that the inequality is tight.

C Proof of Proposition 4.3

The proof is based on similar arguments as [Nes05, Theorem 1]. First, notice that

$$\begin{aligned} &\langle \nabla\mathcal{E}(\boldsymbol{\pi}(\boldsymbol{q}_2)) - \nabla\mathcal{E}(\boldsymbol{\pi}(\boldsymbol{q}_1)), \boldsymbol{\pi}(\boldsymbol{q}_2) - \boldsymbol{\pi}(\boldsymbol{q}_1)\rangle \\ &= \left\langle \int_{t=0}^{1} \mathcal{E}''(\boldsymbol{\pi}_1 + t(\boldsymbol{\pi}_2 - \boldsymbol{\pi}_1))(\boldsymbol{\pi}_2 - \boldsymbol{\pi}_1)dt, \boldsymbol{\pi}(\boldsymbol{q}_2) - \boldsymbol{\pi}(\boldsymbol{q}_1) \right\rangle \\ &\geq \sigma_{\mathcal{E}}[\boldsymbol{\pi}_1, \boldsymbol{\pi}_2]\, \|\boldsymbol{\pi}_2 - \boldsymbol{\pi}_1\|_E^2, \end{aligned} \tag{21}$$

where

$$\sigma_{\mathcal{E}}[\boldsymbol{\pi}_1, \boldsymbol{\pi}_2] = \inf_{t\in[0,1]} \sigma_{\mathcal{E}}(t\boldsymbol{\pi}_1 + (1-t)\boldsymbol{\pi}_2)$$

is the local convexity modulus of $\mathcal{E}$ on the segment $[\boldsymbol{\pi}(\boldsymbol{q}_1), \boldsymbol{\pi}(\boldsymbol{q}_2)]$.

Next, recall that

$$J_\alpha(\boldsymbol{q}) = \max_{\boldsymbol{\pi}\in\Delta_m} \langle -\boldsymbol{M}^*\boldsymbol{q}, \boldsymbol{\pi}\rangle_{E^*\times E} + \langle \boldsymbol{p}, \boldsymbol{q}\rangle_{F^*\times F} - \alpha\mathcal{E}(\boldsymbol{\pi}).$$

The optimality conditions of the previous maximization problem for $J_\alpha(\boldsymbol{q}_1)$ and $J_\alpha(\boldsymbol{q}_1)$, $\boldsymbol{q}_1, \boldsymbol{q}_2 \in F$, lead to

$$\begin{aligned} \langle -\boldsymbol{M}^*\boldsymbol{q}_1 - \alpha\nabla\mathcal{E}(\boldsymbol{\pi}(\boldsymbol{q}_1)), \boldsymbol{\pi}(\boldsymbol{q}_2) - \boldsymbol{\pi}(\boldsymbol{q}_1)\rangle \leq 0, \\ \langle -\boldsymbol{M}^*\boldsymbol{q}_2 - \alpha\nabla\mathcal{E}(\boldsymbol{\pi}(\boldsymbol{q}_2)), \boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\rangle \leq 0. \end{aligned}$$

Combining the two previous inequalities, we can write that for $\boldsymbol{q}_1, \boldsymbol{q}_2 \in F$:

$$\begin{aligned}
&\langle \boldsymbol{M}^*(\boldsymbol{q}_1 - \boldsymbol{q}_2), \boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\rangle \geq \alpha \langle \nabla\mathcal{E}(\boldsymbol{\pi}(\boldsymbol{q}_2)) - \nabla\mathcal{E}(\boldsymbol{\pi}(\boldsymbol{q}_1)), \boldsymbol{\pi}(\boldsymbol{q}_2) - \boldsymbol{\pi}(\boldsymbol{q}_1)\rangle, \\
&\implies \|\boldsymbol{M}^*(\boldsymbol{q}_1 - \boldsymbol{q}_2)\|_{E^*} \|\boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\|_E \overset{(21)}{\geq} \alpha\sigma_{\mathcal{E}}[\boldsymbol{\pi}(\boldsymbol{q}_1), \boldsymbol{\pi}(\boldsymbol{q}_2)] \|\boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\|_E^2, \\
&\implies \|\boldsymbol{M}^*\|_{F\to E^*} \|\boldsymbol{q}_1 - \boldsymbol{q}_2\|_F \|\boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\|_E \geq \alpha\sigma_{\mathcal{E}}[\boldsymbol{\pi}(\boldsymbol{q}_1), \boldsymbol{\pi}(\boldsymbol{q}_2)] \|\boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\|_E^2.
\end{aligned}$$

Therefore, we can write that

$$\|\boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\|_E \leq \frac{\|\boldsymbol{M}^*\|_{F\to E^*} \|\boldsymbol{q}_1 - \boldsymbol{q}_2\|_F}{\alpha\sigma_{\mathcal{E}}[\boldsymbol{\pi}(\boldsymbol{q}_1), \boldsymbol{\pi}(\boldsymbol{q}_2)]}.$$

Noting that $\nabla J_\alpha(\boldsymbol{q}) = -\boldsymbol{M}\boldsymbol{\pi}(\boldsymbol{q}) + \boldsymbol{p}$, we can conclude that

$$\begin{aligned}
\|\nabla J_\alpha(\boldsymbol{q}_1) - \nabla J_\alpha(\boldsymbol{q}_2)\|_{F^*} &= \|\boldsymbol{M}(\boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2))\|_{F^*} \\
&\leq \|\boldsymbol{M}^*\|_{F\to E^*} \|\boldsymbol{\pi}(\boldsymbol{q}_1) - \boldsymbol{\pi}(\boldsymbol{q}_2)\|_E \\
&\leq \frac{\|\boldsymbol{M}^*\|_{F\to E^*}^2}{\alpha\sigma_{\mathcal{E}}[\boldsymbol{\pi}(\boldsymbol{q}_1), \boldsymbol{\pi}(\boldsymbol{q}_2)]} \|\boldsymbol{q}_1 - \boldsymbol{q}_2\|_F.
\end{aligned}$$

Let us consider a sequence $(\boldsymbol{q}_n)_{n\in\mathbb{N}}$ converging uniformly towards $\boldsymbol{q} \in B_\infty \subset F$. Since $\boldsymbol{\pi} : \boldsymbol{q} \in B_\infty \longmapsto \boldsymbol{\pi}(\boldsymbol{q}) \in \Delta_m$ is a continuous mapping, $\boldsymbol{\pi}_n := \boldsymbol{\pi}(\boldsymbol{q}_n)$ converges uniformly towards $\boldsymbol{\pi}(\boldsymbol{q})$. Thus,

$$\sigma_{\mathcal{E}}[\boldsymbol{\pi}(\boldsymbol{q}_n), \boldsymbol{\pi}(\boldsymbol{q})] \underset{n\to+\infty}{\longrightarrow} \sigma_{\mathcal{E}}(\boldsymbol{\pi}(\boldsymbol{q})).$$

D Proof of Theorem 4.6

Theorem 4.6 is a direct consequence of Lemma (D.1) below. A similar proof was proposed in [FP11], but not extended to a general setting.

Lemma D.1. *Let $f : \mathbb{R}^m \to \mathbb{R}\cup\{+\infty\}$ and $g : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ denote closed convex functions such that $A \cdot ri(dom(f)) \cap ri(dom(g)) \neq \emptyset$. Assume further that g is σ-strongly convex w.r.t. an arbitrary norm $\|\cdot\|$. Then*

(i) *Function g^* satisfies $dom(g^*) = \mathbb{R}^n$ and it is differentiable on $\mathbb{R}^n$.*

(ii) *Denote*

$$p(x) = f(Ax) + g(x) \tag{22}$$

$$d(y) = -g^*(-A^*y) - f^*(y) \tag{23}$$

and

$$x(y) = \nabla g^*(-A^*y).$$

Let x^* denote the minimizer of (22) and y^* denote any minimizer of (23). Then for any $y \in \mathbb{R}^m$ we have

$$\|x(y) - x^*\|^2 \leq \frac{2}{\sigma}(d(y) - d(y^*)). \tag{24}$$

Proof. Point (i) is a standard result in convex analysis. See e.g. [HUL96]. We did not find the result (ii) in standard textbooks and to our knowledge it is new. We assume for simplicity that g, g^*, f and f^* are differentiable. This hypothesis is not necessary and can be avoided at the expense of longer proofs. First note that

$$\inf_{x\in\mathbb{R}^n} p(x) = \sup_{y\in\mathbb{R}^m} -g^*(-A^*y) - f^*(y)$$

by Fenchel-Rockafellar duality. Since g is strongly convex ∇g is a one-to-one mapping and

$$\nabla g(\nabla g^*(x)) = x, \ \forall x \in \mathbb{R}^n. \tag{25}$$

The primal-dual relationships read

$$\begin{cases} A^* y^* + \nabla g(x^*) & = 0 \\ Ax^* - \nabla f^*(y^*) & = 0 \end{cases}$$

So that

$$\begin{aligned} x^* &= (\nabla g)^{-1}(-A^* y^*) \\ &= (\nabla g^*)(-A^* y^*). \end{aligned}$$

Let us define the following Bregman divergences quantities:

$$\begin{aligned} D_1(y) &:= f^*(y) - f^*(y^*) - \langle A \nabla g^*(-A^* y^*), y - y^* \rangle \\ D_2(y) &:= g^*(-A^* y) - g^*(-A^* y^*) + \langle A \nabla g^*(-A^* y^*), y - y^* \rangle. \end{aligned}$$

By construction

$$D_1(y) + D_2(y) = d(y) - d(y^*).$$

Moreover since y^* is the minimizer of d it satisfies $A\nabla g^*(-A^* y^*) = \nabla f^*(y^*)$. By replacing this expression in D_1 and using the fact that f^* is convex we get that

$$D_1(y) \geq 0, \ \forall y \in \mathbb{R}^n.$$

Using identity (25) we get:

$$D_2(y) = g^*(\nabla g(x(y))) - g^*(\nabla g(x^*)) + \langle x^*, \nabla g(x^*) - \nabla g(x(y)) \rangle. \tag{26}$$

Moroever, since (see e.g. [HUL96])

$$g(x) + g^*(x^*) = \langle x, x^* \rangle \Leftrightarrow x^* = \nabla g(x),$$

we get that

$$g^*(\nabla g(x(y))) = \langle \nabla g(x(y)), x(y) \rangle - g(x(y)),$$

and

$$g^*(\nabla g(x^*)) = \langle \nabla g(x^*), x^* \rangle - g(x^*).$$

Replacing these expressions in (26) we obtain

$$\begin{aligned} D_2(y) &= g(x^*) - g(x(y)) + \langle \nabla g(x(y)), x(y) - x^* \rangle \\ &\geq \frac{\sigma}{2} \|x(y) - x^*\|^2 \end{aligned}$$

since g is σ strongly convex w.r.t $\|\cdot\|$. To sum up we have:

$$\begin{aligned} d(y) - d(y^*) &= D_1(y) + D_2(y) \\ &\geq D_2(y) \\ &\geq \frac{\sigma}{2} \|x(y) - x^*\|^2 \end{aligned}$$

which is the desired inequality. ■

We now have all the ingredients to prove Theorem 4.6.

Proof of Theorem 4.6. The proof is a direct consequence of Lemma D.1. It can be obtained by setting $A \equiv \boldsymbol{M}$, $f(y) \equiv \|y - \boldsymbol{p}\|_1$ and $g(x) \equiv \alpha\mathcal{E}(x) + \chi_{\Delta_m}(\boldsymbol{\pi})$, with χ_{Δ_m} the indicator function of the set Δ_m. Thus $p(x) = f(Ax) + g(x) = F_\alpha(x)$ and $d(y) = J_\alpha(y)$. Then remark that $\boldsymbol{\pi}_k$ defined in Theorem 4.6 satisfies $\boldsymbol{\pi}_k = \nabla g^*(-A^* \boldsymbol{y}_k)$. By Proposition 4.2, we get that g is $\alpha\sigma_{\mathcal{E}}$-strongly convex w.r.t. $\|\cdot\|_{\ell^p}$, for all $p \in [1; \infty]$. It then suffices to use bound (12) together with Lemma D.1 to conclude. ■

References

[AHPR13] Ben Adcock, Anders C Hansen, Clarice Poon, and Bogdan Roman. Breaking the coherence barrier: asymptotic incoherence and asymptotic sparsity in compressed sensing. arXiv preprint arXiv:1302.0561, 2013.

[BBW13] Jérémie Bigot, Claire Boyer, and Pierre Weiss. An analysis of block sampling strategies in compressed sensing. Preprint, 2013.

[BC11] Heinz H Bauschke and Patrick L Combettes. Convex analysis and monotone operator theory in Hilbert spaces. Springer, 2011.

[CCKW13] Nicolas Chauffert, Philippe Ciuciu, Jonas Kahn, and Pierre. Weiss. Variable density sampling with continuous sampling trajectories. preprint, 2013.

[CCW13] Nicolas Chauffert, Philippe Ciuciu, and Pierre Weiss. Variable density compressed sensing in MRI. theoretical vs heuristic sampling strategies. In proceedings of IEEE ISBI, 2013.

[CDV10] Patrick L Combettes, Đinh Dũng, and Bng Công Vũ. Dualization of signal recovery problems. Set-Valued and Variational Analysis, 18(3-4):373–404, 2010.

[CP11a] Emmanuel J. Candes and Yaniv Plan. A probabilistic and ripless theory of compressed sensing. Information Theory, IEEE Transactions on, 57(11):7235–7254, 2011.

[CP11b] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pages 185–212. Springer, 2011.

[CRCP12] Rachel W Chan, Elizabeth A Ramsay, Edward Y Cheung, and Donald B Plewes. The influence of radial undersampling schemes on compressed sensing reconstruction in breast mri. Magnetic Resonance in Medicine, 67(2):363–377, 2012.

[CRT06] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. Information Theory, IEEE Transactions on, 52(2):489–509, 2006.

[CT93] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using bregman functions. SIAM Journal on Optimization, 3(3):538–543, 1993.

[dJ13] Alexandre d'Aspremont and Martin Jaggi. An optimal affine invariant smooth minimization algorithm. arXiv preprint arXiv:1301.0465, 2013.

[Don06] David Donoho. Compressed sensing. Information Theory, IEEE Transactions on, 52(4):1289–1306, 2006.

[FP11] Jalal M Fadili and Gabriel Peyré. Total variation projection with first order schemes. Image Processing, IEEE Transactions on, 20(3):657–669, 2011.

[GK13] Clóvis C Gonzaga and Elizabeth W Karas. Fine tuning nesterovs steepest descent algorithm for differentiable convex programming. Mathematical Programming, pages 1–26, 2013.

[HPH$^+$11] Robert Hummel, Sameera Poduri, Franz Hover, Urbashi Mitra, and Guarav Sukhatme. Mission design for compressive sensing with mobile robots. In Robotics and Automation (ICRA), 2011 IEEE International Conference on, pages 2362–2367. IEEE, 2011.

[HUL96] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. Convex Analysis and Minimization Algorithms: Part 1: Fundamentals, volume 1. Springer, 1996.

[JN08] Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. arXiv preprint arXiv:0809.0813, 2008.

[KW12] Felix Krahmer and Rachel Ward. Beyond incoherence: stable and robust sampling strategies for compressive imaging. arXiv preprint arXiv:1210.2380, 2012.

[LDSP08] Michael Lustig, David L. Donoho, Juan M. Santos, and John M. Pauly. Compressed sensing mri. Signal Processing Magazine, IEEE, 25(2):72–82, 2008.

[LKP08] Michael Lustig, Seung-Jean Kim, and John M Pauly. A fast method for designing time-optimal gradient waveforms for arbitrary k-space trajectories. Medical Imaging, IEEE Transactions on, 27(6):866–873, 2008.

[Nes05] Yurii Nesterov. Smooth minimization of non-smooth functions. Mathematical Programming, 103(1):127–152, 2005.

[Nes13] Yu. Nesterov. Gradient methods for minimizing composite functions. Mathematical Programming, 140(1):125–161, 2013.

[NN04] Yurii Nesterov and IU E Nesterov. Introductory lectures on convex optimization: A basic course, volume 87. Springer, 2004.

[PDG12] Adam C Polak, Marco F Duarte, and Dennis L Goeckel. Performance bounds for grouped incoherent measurements in compressive sensing. arXiv preprint arXiv:1205.2118, 2012.

[PVW11] Gilles Puy, Pierre Vandergheynst, and Yves Wiaux. On variable density compressive sampling. Signal Processing Letters, IEEE, 18(10):595–598, 2011.

[Rau10] Holger Rauhut. Compressive sensing and structured random matrices. Theoretical foundations and numerical methods for sparse recovery, 9:1–92, 2010.

[Roc97] R Tyrell Rockafellar. Convex analysis, volume 28. Princeton university press, 1997.

[SPM95] D. M. Spielman, J. M. Pauly, and C. H. Meyer. Magnetic resonance fluoroscopy using spirals with variable sampling densities. Magnetic resonance in medicine, 34(3):388–394, 1995.

[Wri97] Graham A. Wright. Magnetic resonance imaging. Signal Processing Magazine, IEEE, 14(1):56–66, 1997.

[WSK$^+$07] Stefanie Winkelmann, Tobias Schaeffter, Thomas Koehler, Holger Eggers, and Olaf Doessel. An optimal radial profile order based on the golden ratio for time-resolved mri. Medical Imaging, IEEE Transactions on, 26(1):68–76, 2007.