Compressed sensing in Hilbert spaces

Yann Traonmilin, Gilles Puy, Rémi Gribonval and Mike E. Davies

Abstract In many linear inverse problems, we want to estimate an unknown vector belonging to a high-dimensional (or infinite-dimensional) space from few linear measurements. To overcome the ill-posed nature of such problems, we use a low-dimension assumption on the unknown vector: it belongs to a low-dimensional model set. The question of whether it is possible to recover such an unknown vector from few measurements then arises. If the answer is yes, it is also important to be able to describe a way to perform such a recovery. We describe a general framework where appropriately chosen random measurements guarantee that recovery is possible. We further describe a way to study the performance of recovery methods that consist in the minimization of a regularization function under a data-fit constraint.

1 Introduction

Many signal processing tasks aim at estimating a signal x from its observation y. The signal x can often be described by a continuous physical phenomenon and the observations y are made of a finite collection of scalar measurements. The most basic example of such observations is a sampled version of the signal x (e.g. for a sound recorded at a given sampling rate, the continuous x is the electrical signal produced by the microphone over time). More generally, we

Yann Traonmilin e-mail: yann.traonmilin@inria.fr, Rémi Gribonval
INRIA Rennes - Bretagne Atlantique, Campus de Beaulieu 35042 Rennes Cedex, France.

Gilles Puy
Technicolor, 975 Avenue des Champs Blancs, 35576 Cesson-Sévigné, France. Gilles Puy contributed to the results reported in this chapter when he was at INRIA Rennes.

Mike E. Davies
Institute for Digital Communications (IDCom), University of Edinburgh, The King's buildings, Edinburgh, EH9 3JL.

consider observations y modeled as

$$y = Ax + e \tag{1}$$

where $x \in \mathcal{H}$, $y \in \mathcal{F}$ and $\mathcal{H}, \mathcal{F}$ are Hilbert spaces of finite or infinite dimension. The operator A is a linear map and e is a noise whose energy $\|e\|_{\mathcal{F}}$ is bounded. In most cases, the operator A models a finite number of measurements m. This Hilbert space setting is a way to have a general view of signal recovery problems in classical finite or infinite-dimensional spaces where signals (in a wide sense: time series, images, videos, ...) are modelled, e.g. the space of continuous signals with finite energy $\mathcal{L}^2(\mathbb{R}^d)$, the space of bandlimited signals with finite energy or its equivalent after sampling, $\ell^2(\mathbb{R}^d)$, or the finite-dimensional vector space $\mathbb{R}^d$.

1.1 Observation model and low complexity signals

Observing a continuous signal with finitely many linear measurements induces an information loss. If no further prior information on the signal is available, recovering x from y is generally not possible. However, if an (approximate) hypothesis of "low complexity" on x is available, enforcing the hypothesis in the recovery process can ensure that we are able to estimate x with reasonable accuracy. Low complexity can be defined in several ways. It often means that the signal lives in a "low-dimensional model" or can be described by few parameters. Two classical examples where low complexity helps to recover the signal are:

- Sampling of periodic band-limited signals in $\mathcal{H} = \mathcal{L}^2(\mathbb{R})$: if the signal is known to be band-limited with cut-off frequency B, it is possible to recover it perfectly provided it is sampled at a rate at least $2B$.
- Compressed sensing in $\mathcal{H} = \mathbb{R}^n$: if the signal is known to have at most k non-zero samples in $\mathbb{R}^n$, it can be recovered with high probability from m random Gaussian (or Fourier) observations provided $m \gtrsim k \log(n)$ [11] (We use the symbol $\gtrsim$ to say that there is an absolute constant C such that if $m \geq Ck \log(n)$ recovery is possible with high probability). Similarly, if the signal is an $n \times n$ matrix with rank at most r, in the space $\mathcal{H} = \mathbb{R}^{n \times n}$, it can be recovered with high probability from $m \gtrsim rn$ random Gaussian observations [14].

In the following, the notion of low complexity is summarized by the fact that x is well approximated by an element of a so-called *model set* Σ, where $\Sigma \subset \mathcal{H}$ is low-dimensional according to a notion of dimension that will be specified. The considered notion of dimension will be defined in Section 3.1 and is related to the number of unknowns we need to estimate to characterize the signal. In the context of linear inverse problems with such low-dimensional

models, a first objective is to obtain conditions on the linear operator A and the model set Σ that guarantee a possible recovery. From this perspective, the analysis of "low complexity recovery" is an extension of classical analyses of sparse recovery or low rank matrix recovery. A second objective, related to the field of compressed sensing, is dimension reduction, where the goal is to design a linear operator A (often with randomness) so that low complexity recovery is possible, with an emphasis on allowing the dimension m of the observation to be small.

1.2 Decoders

As the ultimate task is to recover x from y, the analysis of the observation of x must be held together with the study of the methods used to recover x from y, which we call *decoders*. In this chapter, we consider the general class of decoders which consist in minimizing a regularizer under a data fit constraint. We study estimates x^* of x of the form

$$x^* \in \underset{z \in \mathcal{H}}{\operatorname{argmin}} f(z) \; s.t. \; \|Az - (Ax + e)\|_{\mathcal{F}} \leq \epsilon. \tag{2}$$

Formulation (2) covers many decoders proposed in the literature, even though other formulations exist (e.g., minimizing $\|Az - (Ax + e)\|_{\mathcal{F}}$ under a constraint on $f(z)$, or using a Lagrangian formulation). The study presented in this chapter does not require x^* to be the unique minimizer of (2). It must be noted that this formulation somehow emphasizes practical signal processing applications because an estimation ϵ of the observation noise energy $\|e\|_{\mathcal{F}}$ is often available (e.g. in photography the noise level can be estimated using aperture time and light conditions, in image processing more advanced techniques allow to estimate noise level from an image [37]).

The main parameter of the decoder is the regularizer f. Its role is to force the estimate belongs to the chosen model set. The form of the data fit constraint ($\|\cdot\|_{\mathcal{F}}$) influences the types of noise that the decoder can robustly manage. This raises interesting questions that are, however, out of the scope of this chapter. The main qualities required for a decoder are: 1) to provide *exact recovery of vectors $x \in \Sigma$ in the noiseless setting*; 2) to be *stable to observation noise and robust to modeling error*.

We emphasize the role of two classes of decoders: "ideal" decoders and convex decoders.

- Given a problem with a model set Σ, the *ideal decoder* corresponds to minimizing (2) using $f := \iota_\Sigma$ the characteristic function of Σ, i.e. $\iota_\Sigma(x) = 0$ if $x \in \Sigma$, $\iota_\Sigma(x) = \infty$ otherwise. This decoder is called ideal, as it enforces perfectly the fact that the solution must belong to Σ (the prior on the unknown). Unfortunately, it is generally hard to calculate efficiently as the

function to minimize is both non-convex and non-smooth. Consequently, we often use a heuristic for the minimization or turn to a convex proxy to this minimization.

- The decoder is said to be a *convex decoder* when f is convex. Such a decoder is often easier to compute as the minimization problem has no local minimum other than the global minima even if it this does not guarantee that the minimization can be efficiently performed, see e.g. tensor recovery problems [28]. State of the art shows that having some additional hypothesis on the linear operator A enables to guarantee stability and robustness of certain convex decoders for classical model sets Σ.

1.3 The RIP: a tool for the study of signal recovery

As we just saw, studying signal recovery amounts to studying the interactions between the model Σ, the regularization f and the measurement operator A. We propose here to use a tool that enables us to separate the study of A with respect to Σ from the study of f with respect to Σ: the restricted isometry property (RIP). It is generally defined in our setting for a linear observation operator A on the so-called *secant set* $\Sigma - \Sigma := \{x - x' : x \in \Sigma, x' \in \Sigma\}$

Definition 1 (RIP). The linear operator $A : \mathcal{H} \to \mathcal{F}$ satisfies the RIP on the secant set $\Sigma - \Sigma$ with constant δ if for all $x \in \Sigma - \Sigma$:

$$(1 - \delta)\|x\|_{\mathcal{H}}^2 \leq \|Ax\|_{\mathcal{F}}^2 \leq (1 + \delta)\|x\|_{\mathcal{H}}^2 \tag{3}$$

where $\|\cdot\|_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{F}}$ are Euclidean norms on $\mathcal{H}$ and $\mathcal{F}$.

This property is a famous sufficient condition on A to guarantee the success of convex decoders (2) in the case of sparse and low rank signal recovery for appropriately chosen regularization f [21, 15, 34, 13, 17, 25]. Intuitively, the RIP requires the operator A to preserve the distance between any two elements of Σ (see Figure 1). Moreover, a lower RIP is a necessary condition for the existence of stable and robust decoders: given A and Σ, if a stable and robust decoder exists then, up to a global rescaling, A satisfies a lower RIP on the secant set $\Sigma - \Sigma$ [18, 7].

For example, in the case of sparse recovery, it is possible to show two facts.

- Fact 1: Random Gaussian matrices of size $m \times n$ satisfy the RIP on the set of $2k$-sparse vectors (the secant set of the set of k-sparse vectors) with constant $\delta < 1$ with high probability, provided $m \gtrsim \delta^{-2} k \log(n)$.
- Fact 2: As soon as A satisfies this RIP with constant $\delta < 1/\sqrt{2}$, it is guaranteed that minimization (2) with $f(\cdot) = \|\cdot\|_1$, the ℓ^1 norm, yields stable and robust recovery of all k-sparse vectors [9].

We see that the study of recovery guarantees in this case is separated in two steps: 1) a study of the behaviour of the linear operator A with respect to

the model set Σ (in terms of RIP property); and 2) a study of the behaviour of the regularizer f with respect to the model set Σ, that has consequences for all operators satisfying a RIP with a small enough constant.

The framework presented in the following generalizes these features in order to manage not only the classical sparse recovery/low-rank recovery and related compressed sensing theory, but much beyond to many sorts of low-dimensional model sets.

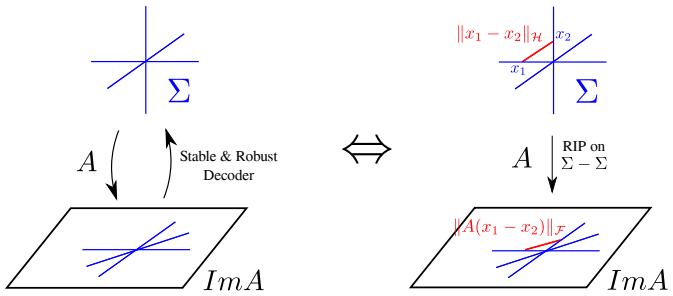


Fig. 1: A graphical representation of the equivalence between the existence of stable robust decoders and the RIP on the secant set. Operators satisfying the RIP approximately preserve distances between elements of Σ.

1.4 A general compressed sensing framework

The remaining part of this chapter shows how it is possible to generalize the steps we just mentioned. The proposed framework[1] consists in answering the following questions (summarized in Figure 2):

- Low-dimensional model: when is Σ "low-dimensional"? (Section 2)
- Dimension-reduction: given Σ, is there an operator A that satisfies the RIP on $\Sigma - \Sigma$? What level of dimension reduction can it achieve? (Section 3)
- What is a good regularizer? Given Σ and f, does a RIP of A on $\Sigma - \Sigma$ guarantee that f recovers the elements of Σ? (Section 4)

Section 5 mentions generalizations that were left out of the main of the chapter in order to keep the exposition accessible, and discusses what challenges we face to go beyond this general compressed sensing framework in Hilbert spaces.

[1] This chapter gives a unified view of the latest developments in the area found in [33] and [38].

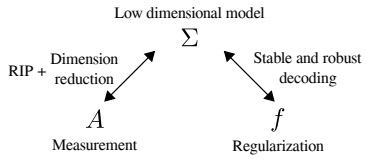


Fig. 2: Structure of the framework: the RIP framework allows to separate the study of dimension reduction and of decoding.

2 Low-dimensional models

We begin by precisely describing the low-dimensional models that will be considered in this chapter. We then focus on a model of *structured sparsity in levels*, which we will use as a running example to illustrate the different concepts used in this chapter.

2.1 Definition and examples

The results presented in [33] show that one can always construct a linear operator A that satisfies the RIP on $\Sigma - \Sigma$ if its normalized secant set $\mathcal{S}(\Sigma)$ has a finite intrinsic dimension. The *normalized secant set* of Σ is defined as

$$\mathcal{S}(\Sigma) := \left\{ z = \frac{y}{\|y\|_{\mathcal{H}}} : y \in (\Sigma - \Sigma) \setminus \{\mathbf{0}\} \right\}.$$

We substitute $\mathcal{S}$ for $\mathcal{S}(\Sigma)$ hereafter to simplify notations. We illustrate in Figure 3 the RIP on the normalized secant set which is equivalent to the RIP on the secant set.

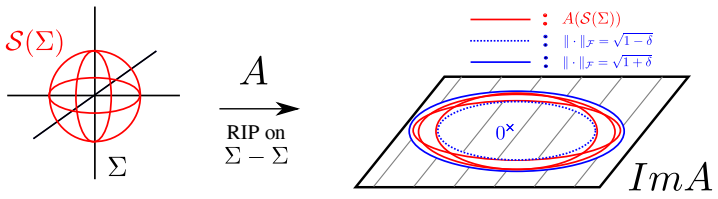


Fig. 3: A characterization of the RIP: the RIP on the normalized secant set. The image of the secant set must lie within a distance δ of the unit sphere.

In this chapter, we measure the intrinsic dimension of $\mathcal{S}$ using the upper box-counting dimension, which is linked to the notion of covering number.

Definition 2 (Covering number). Let $\alpha > 0$ and $\mathcal{S} \subset \mathcal{H}$. The covering number $N(\mathcal{S}, \alpha)$ of $\mathcal{S}$ is the minimum number of closed balls (with respect to the norm $\|\cdot\|_{\mathcal{H}}$) of radius α, with centers in $\mathcal{S}$, needed to cover $\mathcal{S}$.

The upper box-counting dimension is then defined as follows.

Definition 3 (Upper box-counting dimension). The upper box-counting dimension of $\mathcal{S}$ is

$$\mathrm{boxdim}(\mathcal{S}) := \limsup_{\alpha \to 0} \log[N(\mathcal{S}, \alpha)]/\log[1/\alpha].$$

Hence, as soon as $k > \mathrm{boxdim}(\mathcal{S})$, there exists a model-set dependent constant $\alpha_{\mathcal{S}} \in (0, 1/2)$ such that $N(\mathcal{S}, \alpha) \leq \alpha^{-k}$ for all $\alpha \leq \alpha_{\mathcal{S}}$. Further, if the covering number satisfies

$$N(\mathcal{S}, \alpha) \leq \left(\frac{C}{\alpha}\right)^k \tag{4}$$

then $\mathrm{boxdim}(\mathcal{S}) \leq k$.

We choose this definition of intrinsic dimension for two reasons. First, for many useful signal models – *e.g.*, sparse vectors, low-rank matrices, smooth manifolds – the upper box-counting dimension of the normalized secant set is known. The results presented in this chapter can thus be directly applied to these sets, without additional work. Second, one should be careful with the definition of intrinsic dimension used in an infinite-dimensional space. Indeed, for some definitions of dimension, there are examples where it is impossible to perform dimension reduction on vectors belonging to a set having a finite dimension (i.e. the set cannot be linearly and stably embedded in a finite-dimensional space [35, Chapter 6.1]). The upper box-counting dimension of the normalised secant set does not suffer from this issue.

In the following we will say informally that a model Σ is *low-dimensional* if $\mathrm{boxdim}(\mathcal{S}(\Sigma))$ is small compared to the ambient dimension of the Hilbert space $\mathcal{H}$ (which may be infinite). In many examples, the dimension $\mathrm{boxdim}(\mathcal{S})$ is of the order of the number of parameters needed to describe elements of the model, as in the case of classical sparsity or low rank matrices. For k-sparse vectors, the dimension of the normalized secant set $\mathcal{S}$ is of the order of k [25, Section C.2]. For $n \times n$ matrices of rank lower than r, the dimension of the normalized secant set $\mathcal{S}$ is of the order of rn.

2.2 Structured sparsity ...

As a running example, we use a refinement of the notion of sparsity as a way to introduce the general framework: we consider a model of *structured sparsity in levels*.

We start by describing *structured sparsity*, a now classical generalization of the plain sparsity model. In many applications, signals are not only sparse but also clustered in groups of significant coefficients in a transformed domain (Fourier domain, Radon domain,...). Structured sparsity (also called group-sparsity) is the assumption that the signal is supported on a few groups of coefficients [27, 5, 24].

Formally, we consider an orthonormal Hilbert basis $(e_i)_{i\in\mathbb{N}}$ of $\mathcal{H}$ and a finite collection G of non-overlapping finite *groups* of indices, i.e. subsets $g \subset \mathbb{N}$ with $|g| < \infty$ and $g \cap g' = \emptyset$ whenever $g \neq g'$. The restriction of the vector $x \in \mathcal{H}$ to the group g is $x_g := \sum_{i\in g}\langle x, e_i\rangle e_i$. A *group support* is a subset $T \subset G$ and the restriction of x to the group support T is $x_T := \sum_{g\in T} x_g$. The group support of $x \in \mathcal{H}$, denoted $\mathrm{gsupp}(x)$, is the smallest $T \subset G$ such that $x_T = x$. The size of the group support of x, denoted $|\mathrm{gsupp}(x)|$, is the cardinality of $\mathrm{gsupp}(x)$ (to be distinguished from the number of non zero coordinates in x).

Given an integer k, the *k-group-sparse model* is defined as

$$\Sigma_k := \{x \in \mathcal{H},\ |\mathrm{gsupp}(x)| \leq k\}. \tag{5}$$

Let d be the size of the biggest group. We have the following covering of $\mathcal{S}(\Sigma_k)$:

$$N(\mathcal{S}(\Sigma_k), \alpha) \leq \left(\tfrac{C}{\alpha}\right)^{dk} \tag{6}$$

where C is a constant depending on d.

2.3 ... in levels

Consider a collection of J orthogonal spaces $\mathcal{H}_j \subset \mathcal{H}$ each equipped with a k_j-group-sparse model Σ_j as defined in (5) (each with its Hilbert basis and its set G_j of groups). Since the subspaces are orthogonal, there is a natural isomorphism between their direct sum and their Cartesian product. It is simpler to work with the latter, and *structured sparsity in levels* is associated to the model (see Figure 4)

$$\Sigma := \left\{ x \in \mathcal{H}, x = \sum_{j=1}^{J} x_j, x_j \in \Sigma_{k_j} \right\}, \tag{7}$$

which is identified to the Cartesian product of the models $\Sigma_{k_1} \times \Sigma_{k_2} \times \ldots \times \Sigma_{k_j}$.

Two examples were this model is useful are: medical imaging (MRI) and simultaneous signal and noise sparse modeling [1, 36, 39]:

- In MRI, the different levels where the signal is sparse are wavelet scales. MRI images are generally sparser at fine wavelet scales than large wavelet scales. This allows for more flexibility in the modeling of the signal than the simple sparsity model.



Fig. 4: A representation of structured sparsity in levels in $\mathcal{H}$. A structured sparsity in level model is formed by different structured sparsity models in orthogonal subspaces.

- Simultaneous signal and noise sparse modeling is a convenient setting for the separation of a signal sparse in some domain from noise that is sparse in another domain. An observed signal y is modeled as the super-imposition of two components, $y = A_1x_1+A_2x_2$ where A_1x_1 is the signal of interest, x_1 lives in the (structured) sparse model Σ_{k_1}, A_2x_2 is noise, and x_2 lives in the (structured) sparse model Σ_{k_2}. This model is also related to the separation of transients from stationary parts in audio, or for the decomposition of images into cartoon and texture [32]. As $y = [A_1\ A_2]x$ with $x = [x_1^T,\ x_2^T]^T$, this corresponds to a two-level (structured) sparse model for x.

For structured sparsity in levels, we have [38]:

$$\begin{aligned} N(\mathcal{S}, \alpha) &\leq N(\mathcal{S}(\Sigma_{k_1}, \alpha) \times \ldots \times N(\mathcal{S}(\Sigma_{k_j}), \alpha) \\ &\leq \left(\tfrac{C_1}{\alpha}\right)^{d_1 k_1} \times \ldots \times \left(\tfrac{C_J}{\alpha}\right)^{d_J k_J} \end{aligned} \tag{8}$$

where C_j are constants that are of the order of the dimension of each level times the maximum size of groups d_j in level j. Hence up to log factors, the upper box-counting dimension of $\mathcal{S}$ in this case is of the order of $\sum d_j k_j$.

3 Dimension reduction with random linear operators

Now that we have defined the notion of dimension of a model Σ that we work with, and the desirable RIP property of a linear operator A, the remaining question is: how to construct a *dimension-reducing* linear operator $A : \mathcal{H} \to \mathbb{R}^m$ that satisfies the RIP on $\Sigma - \Sigma$?

Consider an MRI-like scenario with a sparsity in levels signal model $\Sigma_{k_1} \times \ldots \times \Sigma_{k_J}$ in a wavelet basis. The fact that the signals in Σ have a support restricted to the first J wavelet scales implies that their energy decreases at high frequencies. Intuitively, it thus seems unnecessary to probe very high frequencies in the measurement process for this type of signals [2]. A good approximation of the signals can be obtained by probing all frequencies up to a certain bandlimit B. This process corresponds to a projection from the infinite-dimensional space $\mathcal{H}$ to a finite-dimensional space of size B. However, the dimension B, though finite, might still be reduced. Indeed, the signals are not just concentrated in the first J wavelet scales, they are also sparse in levels. A dimension-reducing step can thus be envisioned after the projection onto the first B Fourier coefficients with, *e.g.*, a random Gaus-

sian matrix. Ideally the final dimension m should satisfy, up to log factors, $m = O\left(\sum_{j=1}^{J} k_j\right)$ (of the order of the number of parameters describing the model). Intuition thus suggests to build the operator A in two steps: a projection onto a finite (but high) dimensional space followed by a multiplication with a random matrix.

In fact, the authors of [33] present such a construction in the general setting that first projects the signal onto a subspace $H \subset \mathcal{H}$ of finite (but potentially large) dimension, then reduces the dimension using a random linear operator on H (see Figure 5).

	Projection		Reduction	
$\Sigma \subset \mathcal{H}$	$\rightarrow$	Linear space $H \subset \mathcal{H}$	$\rightarrow$	Linear space $= \mathbb{R}^{\mathsf{m}}$
Low-dimension		Finite dimension		Low dimension
$\mathrm{boxdim}(\mathcal{S}(\Sigma)) < \infty$		$\dim(H) = d < \infty$		$m \ll d$

Fig. 5: Strategy for dimension reduction: we aim at reducing the dimension of vectors belonging to Σ leaving an infinite-dimensional space $\mathcal{H}$.

3.1 Projection on a finite-dimensional subspace

Assuming that $\mathrm{boxdim}(\mathcal{S})$ is finite we will see that, given $0 < \alpha < 1$, there always exists a finite-dimensional subspace $H \subset \mathcal{H}$ such that

$$(1-\alpha)\|x\|_{\mathcal{H}} \leq \|P_H x\|_{\mathcal{H}} \leq \|x\|_{\mathcal{H}} \tag{9}$$

for all $x \in \Sigma - \Sigma$, where P_H denotes the orthogonal projection onto H.

In the example of Fourier sampling of signals sparse in a Haar basis, it is possible to directly exhibit such a projection P_H by sampling low Fourier frequencies. However, one can generally construct H as follows. First, build an α-cover of the normalized secant set $\mathcal{S}$. As $\mathrm{boxdim}(\mathcal{S})$ is finite, $N(\alpha, \mathcal{S}) < +\infty$ balls are sufficient to build this cover. Let now $\mathcal{C}$ be the set containing the center of these balls. It is then sufficient to take $H = \mathrm{span}\, \mathcal{C}$, see Figure 6 and [33]. We remark that in the worst case the cardinality of $\mathcal{C}$ is exponential in the dimension of $\mathcal{S}$, hence H can have a dimension of the order of $e^{c(\alpha)\times \mathrm{boxdim}(\mathcal{S})}$. Yet the important message to take away at this stage is that:

If the normalized secant set $\mathcal{S}(\Sigma)$ has a finite upper box-counting dimension, then there exists a finite-dimensional subspace $H \subset \mathcal{H}$ that approximates all vectors in $\mathcal{S}$ with precision α.

In the next section, we describe how to further reduce the dimension to $m = O(\text{boxdim}(\mathcal{S}))$ after this first projection.

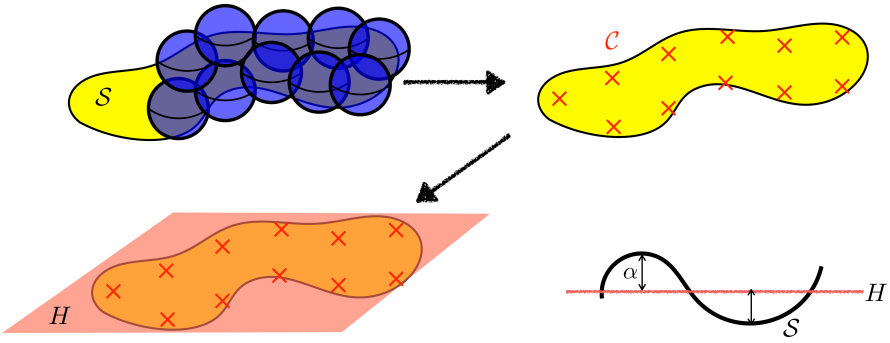


Fig. 6: Construction of H. *Top left*: cover of $\mathcal{S}$ with $N(\mathcal{S}, \alpha)$ balls of radius α. *Top right*: the centers of the balls, indicated by the red crosses, form an α-cover, denoted by $\mathcal{C}$, for $\mathcal{S}$. *Bottom left*: H is defined as the linear span of the vectors in $\mathcal{C}$. *Bottom right*: H approximates $\mathcal{S}$ with precision α.

3.2 Dimension reduction step

After the projection onto the finite-dimensional space H of the previous section, the goal is now to reduce the dimension down to $O(\text{boxdim}(\mathcal{S}))$. As most compressive sensing techniques use random observations to reduce the dimension, it seems natural to follow this route.

Denote by d the dimension of the subspace H and $(e_1, \ldots, e_d)$ an arbitrary orthonormal basis of H. By abuse of notation, identify the projection onto H with the linear operator $P_H : \mathcal{H} \to \mathbb{R}^d$ that returns the coordinates of the orthogonal projection onto H in the basis $(e_1, \ldots, e_d)$. The idea is now to compose P_H with a random matrix $M \in \mathbb{R}^{m \times d}$ to build $A : \mathcal{H} \to \mathbb{R}^m$, *i.e.*, $A = MP_H$. Ideally, we would like A to satisfy the RIP, and $m \approx O(\text{boxdim}(\mathcal{S}))$: a number of measurements of the order of the dimension of the model. In this case, we would be assured that the ideal decoder is stable and robust and that the reduction of dimension is close to optimal.

3.2.1 Randomized dimension reduction

To exhibit a linear operator A satisfying the RIP with constant δ, one can first identify a finite-dimensional subspace $H \subset \mathcal{H}$ such that (9) holds with α small enough, then *build* a random $M : \mathbb{R}^d \to \mathbb{R}^m$ satisfying a RIP with small

enough constant δ'. Sometimes one is directly provided with a random linear operator from $\mathcal{H} \to \mathbb{R}^m$ and needs to *check* whether the RIP holds with high probability. The approach described in [33] makes it possible to handle both cases. With a slight abuse of notation, in this subsection $\mathcal{H}$ stands either for the original Hilbert space (case of a given random operator) or for $\mathbb{R}^d$ (two-step construction considered above).

Consider M a *random* linear operator from $\mathcal{H}$ to $\mathbb{R}^m$. An example results from the independent draw of m identically distributed random vectors $a_i \in \mathcal{H}$, so that for $x \in \mathcal{H}$, $Mx := (\langle a_i, x\rangle)_{i=1}^m$. A convenient way to help M satisfy the RIP is to choose its probability distribution so that, for any vector $x \in \mathcal{H}$,

$$\mathbb{E}_M \|Mx\|_2^2 = \|x\|_{\mathcal{H}}^2. \tag{10}$$

With the above *isotropy assumption*, a draw M of the random linear operator satisfies the RIP on $\Sigma - \Sigma$ if, and only if,

$$\left| \|Mx\|_2^2 - \mathbb{E}_{\tilde{M}} \|\tilde{M}x\|_2^2 \right| \leq \delta \|x\|_{\mathcal{H}}^2, \tag{11}$$

for all $x \in \Sigma - \Sigma$, where we emphasize with the $\tilde{M}$ notation that the expectation $\mathbb{E}$ is with respect to a linear operator with the same *distribution* as the one from which the particular M is drawn. As discussed in Section 5, even without the isotropy assumption (10), one can establish dimension reduction results using (11) as a generalized definition of the RIP [33].

To prove that M satisfies the RIP, the authors in [33] require it to satisfy two concentration inequalities. Define

$$\begin{aligned} h_M \colon \mathcal{H} &\longrightarrow \mathbb{R} \\ x &\longmapsto \|Mx\|_2^2 - \|x\|_{\mathcal{H}}^2. \end{aligned}$$

The assumption is that there exists two constants $c_1, c_2 \in (0, \infty]$ such that for any fixed $y, z \in \mathcal{S}(\Sigma) \cup \{0\}$,

$$\mathbb{P}_M \left\{ |h_M(y) - h_M(z)| \geq \lambda \|y - z\|_{\mathcal{H}} \right\} \leq 2e^{-c_1 m \lambda^2}, \quad \text{for } 0 \leq \lambda \leq c_2/c_1 \tag{12}$$

$$\mathbb{P}_M \left\{ |h_M(y) - h_M(z)| \geq \lambda \|y - z\|_{\mathcal{H}} \right\} \leq 2e^{-c_2 m \lambda}, \quad \text{for } \lambda \geq c_2/c_1. \tag{13}$$

By taking $z = 0$ in (12) and (13), we see that the above properties imply that, for any fixed vector in the normalized secant set, $y \in \mathcal{S}$, $\|My\|_2^2$ stays close to its expected value $\mathbb{E}_M \|My\|_2^2 = \|y\|_{\mathcal{H}}^2 = 1$ with high probability. Proving that the RIP holds consists in showing that, with high probability on the draw of M, this property actually holds uniformly for *all* vectors in $\mathcal{S}$, not just for any fixed vector $y \in \mathcal{S}$. Among other properties, this generalisation to the entire set $\mathcal{S}$ is proved by using the fact that for any fixed $y, z \in \mathcal{S}$, if $\|y - z\|_{\mathcal{H}}$ is small then the difference between $\|My\|_2^2 - \|Mz\|_2^2$ and $\mathbb{E}_{\tilde{M}} \|\tilde{M}y\|_2^2 - \mathbb{E}_{\tilde{M}} \|\tilde{M}z\|_2^2$ is also small with high probability.

These concentration inequalities together with the finite dimension of $\mathcal{S}$ suffice to conclude on a sufficient number of measurements for M to satisfy the RIP [33, Theorem II.2].

Theorem 1. *Let $M : \mathcal{H} \to \mathbb{R}^m$ be a random linear map that satisfies (12) and (13). Assume that $\operatorname{boxdim}(\mathcal{S}) < s$ (there exists $0 < \alpha_{\mathcal{S}} < \frac{1}{2}$ such that $N(\mathcal{S}, \alpha) \leq \alpha^{-s}$ for all $0 < \alpha < \alpha_{\mathcal{S}}$).*

Then for any $\xi, \delta_0 \in (0,1)$, M satisfies the RIP on $\Sigma - \Sigma$ with constant $\delta \leq \delta_0$ with probability at least $1 - \xi$ provided that

$$m \geq \frac{1}{\delta_0^2} \frac{C}{\min(c_1, c_2)} \max\left\{ s \log\left(\frac{1}{\alpha_{\mathcal{S}}}\right), \log\left(\frac{6}{\xi}\right) \right\}, \tag{14}$$

where $C > 0$ is an absolute constant.

This theorem states that if the random operator M satisfies appropriate concentration inequalities and the set $\mathcal{S}$ has finite upper box-counting dimension, then reducing the dimension of the vectors in Σ is possible (recall that these vectors possibly live in an infinite-dimensional space). A number of measurements m of the order of the dimension of the secant set $\mathcal{S}$ (only) is sufficient to be able to recover elements of Σ, whatever the ambient dimension of $\mathcal{H}$ (which can be infinite). We remark that the sufficient number of measurements grows as the RIP constant decreases (the closer A is to an isometry for elements in $\mathcal{S}$). In particular, the typical $\log n$ factor appearing in standard results for compressed sensing of k-sparse vectors in $\mathcal{H} = \mathbb{R}^n$ is in fact related to $\alpha_{\mathcal{S}}$ rather than the ambient dimension. Related results, independent of the ambient dimension, have been achieved for manifold embedding [23, 20].

For a fixed dimension of $\mathcal{S}$, if we wish to ensure an arbitrarily small probability that the RIP fails to hold, $\xi \leq 6\,(\alpha_{\mathcal{S}})^s$, then the number of measurements m also grows as ξ approaches zero. Vice-versa, as the ratio between m and its minimum value $m_0 = \frac{1}{\delta^2} \frac{C}{\min(c_1,c_2)}\, s \log\left(\frac{1}{\alpha_{\mathcal{S}}}\right)$ grows, the RIP holds with probability exponentially close to 1.

Remark 1. The sufficient condition $m \geq m_0$ is *not* necessary. There are actually pathological sets Σ whose normalized secant set has an *infinite upper box counting dimension* and for which some operators $M : \mathcal{H} \to \mathbb{R}^m$ *with only* $m = 1$ measurement satisfy the RIP [33].

3.2.2 Some examples

When given a random linear operator $A : \mathcal{H} \to \mathbb{R}^m$, one can leverage the above result to check whether A satisfies the RIP with high probability. Alternatively, one can *construct* such an operator by pursuing the strategy

described at the beginning of this section. We now need to choose the matrix $M \in \mathbb{R}^{m\times d}$. Examples of matrices $M \in \mathbb{R}^{m\times d}$ such that the operator $A = MP_H$ satisfies (12) and (13) are:

- matrices with independent random Gaussian entries with mean 0 and variance $1/m$;
- matrices whose entries are independent random Bernoulli variables $\pm 1/\sqrt{m}$;
- matrices whose rows are independently drawn from the Euclidean sphere of radius $\sqrt{d/m}$ in $\mathbb{R}^d$ using the uniform distribution.

If M is one of the above matrices (or more generally a matrix with independent subgaussian rows), considering the orthogonally projected model set $\Sigma' = P_H\Sigma$, its normalized secant set $\mathcal{S}' = \mathcal{S}(\Sigma')$, and $s > \text{boxdim}(\mathcal{S}') = \text{boxdim}(\mathcal{S})$, we have [33]: M satisfies the RIP on $\Sigma' - \Sigma'$ with constant $\delta' < \delta_0$ with high probability provided

$$m \geq \frac{C'}{\delta_0^2} \max\left\{ s \log\left(\frac{1}{\alpha_{\mathcal{S}'}}\right), \log\left(\frac{6}{\xi}\right)\right\}, \tag{15}$$

where C' is a constant that depends on the distribution of M.

3.3 Summary

To summarize, a generic strategy (a way to implement the strategy in Figure 5) to build a compressive sensing measurement operator for a set Σ that has a normalized secant set $\mathcal{S}$ of finite upper box-counting dimension is :

1. Find a (potentially high-dimensional) finite-dimensional space H whose orthogonal projection operator satisfies (9). A generic construction ofsuch a space is presented in Section 3.1.
2. Compose this projection operator with a random projection operator M (a random matrix) such that (12) and (13) holds.

Now that we have described how a we can build operators preserving low complexity models, we can turn to the study of the performance of methods used to recover x from y.

4 Performance of regularizers for the recovery of low-dimensional models

As they satisfy the RIP, the linear operators A built with the technique just described in Section 3 preserve the low-dimensional model Σ in the

sense that stable reconstruction of vectors from Σ is possible with the so-called "ideal decoder". Yet, this decoder is often intractable in practice as it involves possibly non-convex and/or non-smooth optimization. We now turn to general decoders, with an emphasis on convex decoders: minimization algorithms are well known for such decoders and they are often possible to implement with off-the-shelf algorithms, as in the classical cases of basis pursuit (ℓ^1 norm minimization) or nuclear norm minimization.

4.1 Convex decoders and atomic norms

In the framework of minimization (2), it is interesting to consider a particular class of convex functions: atomic norms with atoms included in the model set Σ [17]. Considering a set $\mathcal{A} \subset \mathcal{H}$, commonly called the set of *atoms*, the corresponding *atomic "norm"* is built using the convex hull of $\mathcal{A}$.

Definition 4 (Convex hull). The convex hull of a set $\mathcal{A}$ is:

$$\operatorname{conv}(\mathcal{A}) := \left\{ x = \sum c_i a_i : a_i \in \mathcal{A}, c_i \in \mathbb{R}_+, \sum c_i = 1 \right\} \tag{16}$$

Definition 5 (Atomic norm). The atomic "norm" induced by the set $\mathcal{A}$ is defined as:

$$\|x\|_{\mathcal{A}} := \inf \{ t \in \mathbb{R}_+ : x \in t \cdot \overline{\operatorname{conv}}(\mathcal{A}) \} \tag{17}$$

where $\overline{\operatorname{conv}}(\mathcal{A})$ is the closure of $\operatorname{conv}(\mathcal{A})$ in $\mathcal{H}$. The function $\|x\|_{\mathcal{A}}$ is a convex gauge that is not always a norm. It is a norm if $\mathcal{A}$ is symmetrical and bounded. We will keep the term atomic norm in the general case as an abuse of notation. This norm is finite only on the set

$$\mathcal{E}(\mathcal{A}) := \mathbb{R}_+ \cdot \overline{\operatorname{conv}}(\mathcal{A}) = \{ x = t \cdot y, t \in \mathbb{R}_+, y \in \overline{\operatorname{conv}}(\mathcal{A}) \} \subset \mathcal{H}. \tag{18}$$

It can be extended to $\mathcal{H}$ by setting $\|x\|_{\mathcal{A}} := +\infty$ if $x \notin \mathcal{E}(\mathcal{A})$.

Atoms are often *normalized*: a vector u is *normalized* if $\|u\|_{\mathcal{H}} = 1$.

Remark 2. Atomic norms are interesting because given any convex regularization function it is always possible to find an atomic norm that performs noiseless recovery better (in the sense that it permits recovery for more measurement operators A [38]).

4.1.1 Classical examples of atomic norms

As pointed out in [17], many well know norms used for low complexity recovery are atomic norms:

- ℓ^1 norm in $\mathbb{R}^n$: $\mathcal{A}$ is the set of canonical orthonormal basis vectors multiplied by a real scalar with modulus 1, i.e. the normalized 1-sparse vectors.
- Nuclear norm: $\mathcal{A}$ is the set of normalized rank one matrices.
- Gauge generated by a finite polytope: $\mathcal{A}$ is composed of the vertices of a polytope.
- Spectral norm: $\mathcal{A}$ is the set of normalized orthogonal matrices.

4.1.2 Group norms in levels

For our running example, the model $\Sigma = \Sigma_1 \times \dots \Sigma_J$ associated to structured sparsity in levels, we consider a similar class of atomic norms: the group norms in levels.

Given the subspace $\mathcal{H}_j$ associated to the j-th level, $\mathcal{S}_j(1) \subset \mathcal{H}_j$ its unit sphere, G_j its set of groups, and $\Sigma_{1,j}$ the associated 1-group sparse model, consider the collection of atoms of the j-the level:

$$\mathcal{A}_j := \Sigma_{1,j} \cap \mathcal{S}_j(1). \tag{19}$$

The corresponding atomic norm is associated to the finite-dimensional space

$$\mathcal{E}(\mathcal{A}_j) = \operatorname{span}(\{e_i\}_{i \in \cup_{g \in G_j}})$$

and simply given by

$$\|x\|_{\mathcal{A}_j} = \begin{cases} \sum_{g \in G} \|x_g\|_{\mathcal{H}}, & x \in \mathcal{E}(\mathcal{A}); \\ +\infty, & x \notin \mathcal{E}(\mathcal{A}) \end{cases} \tag{20}$$

The norm $\|x\|_{\mathcal{A}_j}$ is called a *group norm*, a *structured norm* or a *mixed* $\ell^1 - \ell^2$ *norm* [41].

A natural regularizer for the *structured sparsity in levels* model is defined as follows in $\mathcal{H}_1 \times \dots \times \mathcal{H}_J$:

$$f_w : (x_1, \dots x_J) \mapsto w_1 \|x_1\|_{\mathcal{A}_1} + \dots + w_J \|x_J\|_{\mathcal{A}_J} \tag{21}$$

with weights $w_j > 0$. We will show in the next sections that setting appropriately the weights in each level can yield recovery guarantees of various strengths.

4.1.3 Atomic norm associated to a union of subspace model

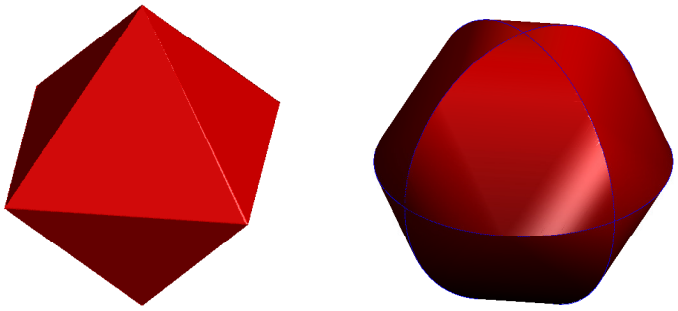
Many classical model sets Σ (the set of sparse vectors, the set of low-rank matrices, etc.) are homogeneous: if $x \in \Sigma$ then $\alpha x \in \Sigma$ for any scalar. As such they are (finite or infinite) unions of subspaces. Given any union of subspaces $\Sigma \subset \mathcal{H}$, the norm associated to its *normalized atoms*

$$\mathcal{A}(\Sigma) := \Sigma \cap \mathcal{S}(1) \tag{22}$$

will be of particular interest for the RIP analysis described in the next sections. As a shorthand notation, we define

$$\|\cdot\|_{\Sigma} := \|\cdot\|_{\Sigma \cap \mathcal{S}(1)}. \tag{23}$$

This norm is sometimes useful as a regularizer to perform recovery (i.e. by choosing $f(z) = \|z\|_{\Sigma}$ in minimization (2)). For the particular case where Σ is the set of k-sparse vectors, $\|\cdot\|_{\Sigma}$ is know as the k-support norm [3]. It is known to yield stable recovery guarantees for certain k-sparse vectors [3], however it has been shown that these results cannot be made uniform for all k-sparse vectors (and consequently similar negative results hold for structured sparsity in levels) [38]. We show in Figure 7 a representation of the ℓ^1-norm and of the k-support norm $\|\cdot\|_{\Sigma}$ for $k = 2$ in 3D ($\mathcal{H} = \mathbb{R}^3$), which are two atomic norms induced by normalized atoms included in the model set Σ.



(a) $\|\cdot\|_1$ (b) $\|\cdot\|_{\Sigma}$

Fig. 7: The unit ball of $\|\cdot\|_1$ (left) and the unit ball of $\|\cdot\|_{\Sigma}$ (k-support norm) for $\Sigma = \Sigma_2$ the set of 2-sparse vectors in 3D (right).

4.2 Stable and robust recovery of unions of subspaces

The main result from [38] states that the stability of *any* decoder of the form (2) is guaranteed provided the linear operator A satisfies a RIP on the secant set $\Sigma - \Sigma$ with a constant $\delta < \delta_{\Sigma}(f)$ holds, where $\delta_{\Sigma}(f)$ is a constant that depends only on the regularizer f and the model set Σ (we give and discuss the definition of $\delta_{\Sigma}(f)$ in Section 4.3 below).

4.2.1 Stable recovery in the presence of noise

Elements of the model can be stably recovered [38, Theorem 1.2]:

Theorem 2 (RIP condition for stable recovery of a union of subspaces). *Assume that Σ is a union of subspaces. Then, for any continuous linear operator A on $\mathcal{H}$ that satisfies the RIP on the secant set $\Sigma - \Sigma$ with constant $\delta < \delta_\Sigma(f)$ we have: for all $x \in \Sigma$, $e \in \mathcal{F}$ such that $\|e\|_\mathcal{F} \leq \epsilon$ (recall that ϵ is an* estimation *of the noise level used as a parameter of the decoder), with x^* the result of minimization (2),*

$$\|x^* - x\|_\mathcal{H} \leq C_\Sigma(f, \delta) \cdot (\|e\|_\mathcal{F} + \epsilon) \tag{24}$$

where $C_\Sigma(f, \delta) < +\infty$.

We refer the reader to [38, Theorem 1.2] for an explicit expression of $C_\Sigma(f, \delta)$. It is increasing with respect to the RIP constant δ : the worse the RIP constant is, the worse the stability constant is (see for example its expression for structured sparsity in levels in Theorem 4).

4.2.2 Robustness to modeling error

Regarding robustness to modeling error, generic results often use the so-called A-norm [7] (not to be confused with the atomic norm: here the A refers to the measurement operator) as an intermediate tool to measure the distance from a vector x to the model set Σ. Given a constant C, the A-norm is defined by

$$\|\cdot\|_{A,C} := C \cdot \|A \cdot\|_\mathcal{F} + \|\cdot\|_\mathcal{H}. \tag{25}$$

It is more convenient to express robustness results with respect to a norm that does not depend on the measurement operator A. We provide here a robustness result where the modeling error with respect to the regularizer f is used (this is more in line with the classical literature for ℓ^1 minimization of nuclear norm minimization). Consider the (symmetrized) distance with respect to f:

$$d_f(x, \Sigma) = \inf_{\tilde{x} \in \Sigma} \frac{f(x - \tilde{x}) + f(\tilde{x} - x)}{2}. \tag{26}$$

When f is a positively homogeneous, non-negative and convex regularizer that bounds the A-norm, robustness with respect to d_f also generally holds [38, Theorem 3.2]:

Theorem 3. *Let Σ be union of subspaces. Let f be positively homogeneous, non-negative and convex with $f(x) < +\infty$ for $x \in \Sigma$. Consider a continuous linear operator A satisfying the RIP on $\Sigma - \Sigma$ with constant $\delta < \delta_\Sigma(f)$, and a noise level $\|e\|_\mathcal{F} \leq \epsilon$. Denote C_Σ the constant from Theorem 2, and assume that for all $u \in \mathcal{H}$, $\|u\|_{A,C_\Sigma} \leq C_{f,A,\Sigma} \cdot f(u)$ for some $C_{f,A,\Sigma} < \infty$. Then, for*

all $x \in \mathcal{H}$, $e \in \mathcal{F}$, such that $\|e\|_{\mathcal{H}} \leq \eta \leq \epsilon$, any minimizer x^* of (2) satisfies

$$\|x^* - x\|_{\mathcal{H}} \leq C_\Sigma \cdot (\|e\|_{\mathcal{F}} + \epsilon) + 2C_{f,A,\Sigma} \cdot d_f(x_0, \Sigma). \tag{27}$$

Remark 3. To apply this theorem we need $C_{f,A,\Sigma} < \infty$. This is the case for most classical examples (sparse recovery with ℓ^1-norm, low-rank matrix recovery with the nuclear norm). It is also true for the case where f being a convex gauge induced by a bounded closed convex set containing 0 and $\mathcal{H}$ is of finite dimension.

Remark 4. Both Theorem 2 and Theorem 3 can be extended to the case where Σ is a *cone* instead of a union of subspaces, with a definition of $\delta_\Sigma(f)$ adapted compared to the one given later in Section 4.3 (See Section 5).

4.2.3 Example: the case of sparsity in levels

Consider the model set Σ corresponding to our running example of structured sparsity in levels, and choose as a regularizer the weighted atomic norm $f_w(\cdot)$ defined in (21). One can show [38, Theorem 4.1] that $\delta_\Sigma(f_w) \geq \frac{1}{\sqrt{2}}$ for $J = 1$ and

$$\delta_\Sigma(f_w) \geq \frac{1}{\sqrt{2 + J\kappa_w^2}}$$

for $J \geq 2$, where $\kappa_w := \max(w_j\sqrt{k_j})/\min(w_j\sqrt{k_j})$. In particular, for the particular weights $w_j = 1/\sqrt{k_j}$, we have $\delta_\Sigma(f_w) \geq \frac{1}{\sqrt{2+J}}$ for $J \geq 2$.

In comparison, Ayaz et al. [4] gave a uniform recovery result with the mixed $\ell^1 - \ell^2$-norm for structured compressed sensing under a RIP hypothesis. They showed that a RIP constant $\delta < \sqrt{2}-1$ for vectors in the secant set guarantees the recovery of vectors from the model. The above result shows that the RIP constant of Ayaz et al. can be improved to $\frac{1}{\sqrt{2}}$. In [1], a model of sparsity in levels was introduced: it is in fact a structured sparsity in levels model with classical sparsity (each group is reduced to a single coordinate) in each level. In [6], Bastounis et al. showed that when the model Σ is sparsity in levels and $f(\cdot) = \sum_j \|\cdot\|_{\mathcal{A}_j} = \|\cdot\|_1$ (i.e., with weights $w_j = 1$, in this case, $\kappa_w^2 = \kappa_1^2$ is the maximum ratio of sparsity between levels), the RIP with constant $\delta = 1/\sqrt{J(\kappa_1 + 0.25)^2 + 1)}$ on $\Sigma - \Sigma$ guarantees recovery. This constant is improved to the constant $\delta_\Sigma(f_w) \geq 1/\sqrt{2+J}$ when weighting the norm of each level with $w_j = 1/\sqrt{k_j}$. The above result further extends the work of Bastounis et al. to general structured sparsity. The following theorem [38, Theorem 4.3] summarizes the result with this optimal weighting:

Theorem 4. *Let Σ be the model set associated to structured sparsity in levels, and consider $f = f_w$ as a regularizer, with the adapted weights $w_j = 1/\sqrt{k_j}$. Suppose the continuous linear operator A satisfies the RIP with constant $\delta < \delta_\Sigma(f)$ on the secant set $\Sigma - \Sigma$. Then for all $x \in \mathcal{H}$, $e \in \mathcal{F}$ such that $\|e\|_{\mathcal{F}} \leq \epsilon$, and x^* the result of minimization (2), we have*

$$\|x^* - x\|_{\mathcal{H}} \leq C_\Sigma(f,\delta)(\|e\|_{\mathcal{F}} + \epsilon) + D_\Sigma(f,\delta) \cdot d_f(x,\Sigma) \tag{28}$$

where :

- *For* $J = 1$, $\delta_0 = \frac{1}{\sqrt{2}}$, $C_\Sigma(f,\delta) \leq \frac{2\sqrt{1+\delta}}{1-\delta\sqrt{2}}$ *and* $D_\Sigma(f,\delta) = 2(1+\sqrt{1+\delta}C_\Sigma(f,\delta))/\sqrt{k}$.
- *For* $J \geq 2$, $\delta_0 = \sqrt{\frac{1}{2+J}}$, $C_\Sigma(f,\delta) \leq \frac{(1+\sqrt{1+J})\sqrt{1+\delta}}{1-\delta\sqrt{2+J}}$ *and* $D_\Sigma(f,\delta) = 2\sqrt{2}(1+\sqrt{1+\delta}C_\Sigma(f,\delta))$.

This result recovers classical guarantees with ℓ^1 minimization for sparse recovery. Since $\delta_\Sigma(f) \geq 1/\sqrt{2+J}$ for $J \geq 2$, combining Theorem 4 for $\delta < 1/\sqrt{2+J}$ with results from Section 3 yields [38] that

$$m \geq O\left(J \sum_{j=1}^{J} \left(k_j d_j + k_j \log\left(\tfrac{3e|G_j|}{k_j}\right)\right)\right).$$

subgaussian measurements are sufficient to guarantee stable and robust recovery with f_w, where $w_j = 1/\sqrt{k_j}$.

Remark 5. The factor J might seem pessimistic, and we attribute its presence to the generality of the result. Should the structure of the observation matrix A be taken into account, better results can be achieved. In fact, if A is a block diagonal matrix where each block A_j has size $m_j \times n_j$, uniform recovery guarantees with the ℓ^1-norm hold if and only if uniform recovery holds on each block: this is possible as soon as each block A_j of A satisfies the RIP with some constant $\delta_j < \frac{1}{\sqrt{2}}$ on $\Sigma_j - \Sigma_j$, which is in turn exactly equivalent to the RIP with constant $\delta < \frac{1}{\sqrt{2}}$ on $\Sigma - \Sigma$.

Remark 6. To make sense of Theorem 4 in the *infinite-dimensional setting*, the domain where the regularizer f is finite must be extended outside of $\mathcal{E}(\Sigma)$ while keeping a finite constant $D_\Sigma(f,\delta)$. This can be done on a case-by-case basis when properties of A and f allow to conclude. For example, as Adcock and Hansen in [1], consider the following setting: $\mathcal{H} = \ell^2(\mathbb{N})$ with Hilbert basis $(e_i)_{i=1,+\infty}$. Consider Σ a sparsity in levels model in $(e_1,..,e_N)$. Let $f = \|\cdot\|_1$. Then f is an extension of the definition of f_w in $\mathcal{E}(\Sigma)$ to the whole space $\mathcal{H}$ (with $w_j = 1$ for all j). In [1], the measurement operator A is a collection of (Fourier) measurements that have *a strong balancing property*. The important fact here is that this property requires $\|A^H A\|_\infty \leq C'$ where $\|\cdot\|_\infty$ is the maximum of the ℓ^∞-norms of the coefficients of $A^H A$ (where A^H is the Hermitian conjugate of A). With such an hypothesis, for any $u \in \mathcal{H}$, we have: $\|Au\|_2^2 = |\langle u, A^H Au\rangle| \leq \|A^H Au\|_\infty \|u\|_1 \leq \|A^H A\|_\infty \|u\|_1 \|u\|_1 \leq C'\|u\|_1^2$. Thus in this case the A-norm is bounded by the ℓ^1-norm: $\|\cdot\|_{A,C} \leq (1 + C\sqrt{C'})\|u\|_1$.

4.3 Definition and calculation of $\delta_\Sigma(f)$

When Σ is a union of subspaces, the sufficient RIP constant for recovery of elements of Σ with f is defined as

$$\delta_\Sigma(f) := \inf_{z \in \mathcal{T}_f(\Sigma)\setminus\{0\}} \sup_{x \in \Sigma} \delta_\Sigma(x, z). \tag{29}$$

where

$$\delta_\Sigma(x, z) := \frac{-\mathcal{R}e\langle x, z\rangle}{\|x\|_{\mathcal{H}} \sqrt{\|x + z\|_\Sigma^2 - \|x\|_{\mathcal{H}}^2 - 2\mathcal{R}e\langle x, z\rangle}}. \tag{30}$$

$$\tag{31}$$

and $\mathcal{T}_f(\Sigma)$ is the set of descent vectors of f at points of Σ:

$$\mathcal{T}_f(\Sigma) := \{z \in \mathcal{H} : \exists x \in \Sigma \ / \ f(x + z) \leq f(x)\} \tag{32}$$

It is important to note that the constant $\delta_\Sigma(f)$ only depends on the geometry of Σ and f. This constant measures the quality of f as regularizer to recover elements of Σ under a RIP assumption: the larger $\delta_\Sigma(f)$, the weaker the assumption on the linear operator A to ensure stable and robust recovery in Theorem 3.

To obtain concrete results, one needs to lower bound the above expression. As the supremum in the expression of $\delta_\Sigma(f)$ is a priori hard to compute explicitly, for $z \in \mathcal{T}_f(\Sigma)$ one can intuitively seek an element $x \in \Sigma$ that maximizes the correlation with $-z$ (i.e., such that $-\mathcal{R}e\langle x, z\rangle$ is maximized). For the model associated to structured sparsity in levels, with the regularizer $f = f_w$ and weights $w_j = 1/\sqrt{k_j}$, this consists in taking $x = -z_T$ where T is the support such that $z_T \in \Sigma$ and z_T concentrates the most energy of z. With such an x, denoting $z_{T_c} = z - z_T$, one can show that

$$\delta(-z_T, z) = \frac{1}{\sqrt{\frac{\|z_{T_c}\|_\Sigma^2}{\|z_T\|_{\mathcal{H}}^2} + 1}} \tag{33}$$

Since $z \in \mathcal{T}_f(\Sigma)$, one shows that the fact that z_T concentrates the most energy implies that $f(z_{T_c}) \leq f(z_T)$, which in turns allows one to conclude that $\|z_{T_c}\|_\Sigma^2 / \|z_T\|_{\mathcal{H}}^2 \leq 1 + J$ is bounded using a control of $\|z_{T_c}\|_\Sigma^2$ obtained by extending Cai's sparse decomposition of polytopes [9]. This leads to the bound $\delta_\Sigma(f_w) \geq 1/\sqrt{2 + J}$ mentioned in Section 4.2.3.

5 Generality of the whole framework

The proof of the existence of random linear maps that reduce dimension while satisfying the RIP is valid for any finite-dimensional model set Σ in any Hilbert space. The guarantees for convex decoders from Section 4 allow to define a critical RIP value for any union of subspaces Σ and any regularizer f (for some pairs this may yield $\delta_\Sigma(f) = 0$, e,g, the right hand side of Figure 7). Overall, the compressive sensing framework described in this chapter is thus very general and we give here an overview of examples where it applies.

5.1 A flexible way to guarantee recovery

The following list summarizes the results of the combined framework of Section 3 and 4 for classical pairs of model Σ and regularizer f. It states the model Σ, the considered regularizer f, a lower bound on our sufficient RIP constant $\delta_\Sigma(f)$, and a sufficient number of (random subgaussian) measurements m to guarantee recovery using the construction from Section 3.

- $\Sigma =$ Linear subspace of dimension n, $f =$ indicator function ι_Σ or $\|\cdot\|_\Sigma$.
  - $\delta_\Sigma(f) = 1$: this sufficient RIP constant was already known, e.g. [7].
  - Sufficient number of measurements: $m \gtrsim n$.
- $\Sigma = k$-sparse vectors in dimension n, $f = \ell^1$-norm.
  - $\delta_\Sigma(f) \geq 1/\sqrt{2}$. This is the sharp RIP constant of Cai et al. [9] (sharpness will be discussed in the next section).
  - Sufficient number of measurements: $m \gtrsim klog(n)$.
- $\Sigma =$ Matrices of rank lower than r in dimension $n \times n$, $f =$ nuclear norm.
  - $\delta_\Sigma(f) \geq 1/\sqrt{2}$. This is also the sharp RIP constant of Cai et al. [9].
  - Sufficient number of measurements: $m \gtrsim rn$.
- $\Sigma =$ Finite union of k 1D-half-spaces with coherence $\mu(\Sigma)$, $f = \|\cdot\|_\Sigma$.
  - $\delta_\Sigma(f) \geq \frac{2(1-\mu(\Sigma))}{3+2\mu(\Sigma)}$.
  - Sufficient number of measurements: $m \gtrsim \log(k)/\delta_\Sigma(f)^2$.
- $\Sigma =$ Permutation matrices of dimension $n \times n$, $f = \|\cdot\|_\Sigma$.
  - $\delta_\Sigma(f) \geq \frac{2}{3}$.
  - Sufficient number of measurements: $m \gtrsim n\log(n)$.

Combining the formalism of [33] with that of [38], these models can also be considered in an infinite-dimensional space $\mathcal{H}$, which is convenient to handle analog compressive sensing scenarios. Stable recovery guarantees are still valid in this infinite-dimensional setting. For robustness, one must make sure that the constant $C_{f,A,\Sigma}$ is finite. For convex f, this might need some further assumptions on the behaviour of f and A outside of the space $\mathcal{E}(\Sigma)$ (the subspace spanned by Σ) as mentionned in Section 4.2.3.

5.2 Uniform vs non-uniform recovery guarantees

The framework described in this chapter focuses on *uniform* recovery guarantees for arbitrary linear operators. Another trend of general framework for compressive sensing focuses on non-uniform guarantees for Gaussian observations. In particular, Chandrasekaran et al [17] studied the general non-uniform recovery from Gaussian observations with atomic norms. In this case, the goal is to show that, for any element x of the model, atomic norm minimization will recover x from Ax with high probability on the draw of A. In contrast, in the framework presented in this chapter, we established conditions so that (with high probability) *the same* linear operator A (i.e., a particular draw of a random operator) allows to stably and robustly recover *all elements of the model* with arbitrary regularizers. Moreover, these results are proved for general random matrices (typically, subgaussian matrices).

5.3 Extensions

The guarantees for convex decoders for unions of subspaces from Section 4 have further been extended to the case where the model set is a cone (a positively homogeneous sets) [38, Theorem 3.1]. This covers models such as (subsets of) the cone of positive semi-definite matrices, or that of non-negative matrices.

Beyond the pure Hilbert norm setting described in this chapter, the generalized definition of the RIP from equation (11) or its further generalizations to arbitrary norms in $\mathcal{H}$ can be used [33] to establish dimension reduction results for structured acquisition. An example is the use of random rank one projections (which are a subset of sub-exponential random matrices) [10], which offer a computationally efficient way to gather linear observations of a matrix, thus making them an interesting observation method for algorithmic purposes in the low rank matrix recovery problem. While the RIP constant $\delta_\Sigma(f)$ has not be extended to such settings yet, such developments seem accessible.

5.4 Sharpness of results?

In [33] the finite dimension of the normalized secant set allows one to conclude on the possibilities in terms of dimension reduction. Only a number of measurements of the order of the dimension is sufficient. However, this hypothesis is not necessary. It is possible to find a model Σ whose normalized secant set $\mathcal{S}$ has infinite upper box counting dimension such that there exists a measurement operator with the RIP on $\mathcal{S}$. Hence a weaker necessary and sufficient

condition on the "dimension" of $\mathcal{S}$ could exist to guarantee the existence of measurements operators with stable dimension reduction capabilities.

In terms of recovery for arbitrary regularizers, it has been shown that a sufficient RIP constant can be provided. For classical families of models and regularizers (sparse recovery with the ℓ^1 norm and low-rank matrix recovery with the nuclear norm), as well as for structured sparsity and the associated group norm the constant $\delta_\Sigma(f)$ is sharp in the following way: we know that there exist RIP matrices with constant arbitrarily close to $\delta_\Sigma(f)$ (here $1/\sqrt{2}$) which do not permit uniform recovery [19, 10] for some dimension of $\mathcal{H}$ and some sparsity k (or rank r). Considering sparsity in levels, we observe that $\delta_\Sigma(f)$ complies with the necessary dependency on the ratios of sparsity between levels and the number of levels J [6]. These sharpness results all consider *families* of models and regularizers: it is a worst case sharpness among these families of regularizer. However, one can consider the question of strong sharpness: for a given model Σ and regularizer f, what is the biggest RIP constant sufficient to guarantee recovery?

5.5 New Frontiers: Super-resolution and compressive learning

Much of the algorithmic and mathematical techniques revolving around the notion of sparsity in the context of inverse problems and compressive sensing have been developed with finite-dimensional models, involving e.g. a discretization of the time domain, or of the frequency domain. However, the physical phenomena underlying the acquisition of modern data from the analogue world are rather intrinsically continuous [40]. The generic framework for inverse problems and dimension reduction presented in this chapter is directly set up in an arbitrary Hilbert space setting, and as such it opens new perspectives for handling the analogue nature of many problems.

Super-resolution is one such problem. In super-resolution, one aims at recovering spikes combinations of few spikes from their low pass observation. While spikes are usually modeled with Dirac measures, which can be considered as belonging to certain Banach spaces of measures (e.g., equiped with the total variation norm), one way of bringing super-resolution close to the content of this chapter is to consider a kernel metric, which will bring a Hilbert structure to such Banach spaces [26]. Intuitively, this amounts to choosing a high resolution at which we will measure energy in the signal space. In this context, all the results on recovery guarantees and dimension reduction hold. Several questions remain standing: is it possible to find a sufficient RIP constant $\delta_\Sigma(f)$ that also holds in this context? Do usual models in Banach spaces have a normalized secant set with finite dimension? With the work of [12, 16, 22], we already know that low pass filtering allows to recover spikes up to some resolution with a convex decoder.

Another related problem is compressive learning. In [8, 29] it is shown empirically that Gaussian mixtures can be recovered from a so-called *sketch* of the data, which can be considered as random Fourier measurements of their probability density. Recent works suggests that for an an appropriately chosen kernel metric, the secant set of sufficiently separated mixtures of Diracs is of finite dimension for appropriately chosen kernel metric [30]. It is then possible to guarantee the success of the ideal decoder with random observations. Practical results have been obtained using a greedy heuristic approach to the problem [31]. These results seem to indicate a possible generalization of the theory of dimension reduction and convex recovery to these problems.

Acknowledgment

This work was supported in part by the European Research Council PLEASE project (ERC-StG-2011-277906) and the European Research Council C-SENSE project (ERC-ADG-2015-694888). M. E. Davies would like to acknowledge the support of EPSRC grant EP/J015180/1.

References

1. B. Adcock, A. Hansen, B. Roman, and G. Teschke. Generalized sampling: stable reconstructions, inverse problems and compressed sensing over the continuum. *Adv. in Imag. and Electr. Phys.*, 182(1):187–279, 2014.
2. B. Adcock and A. C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. *Foundations of Computational Mathematics*, 16(5):1263–1323, 2016.
3. A. Argyriou, R. Foygel, and N. Srebro. Sparse Prediction with the k-Support Norm. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1457–1465. Curran Associates, Inc., 2012.
4. U. Ayaz, S. Dirksen, and H. Rauhut. Uniform recovery of fusion frame structured sparse signals. *Applied and Computational Harmonic Analysis*, 2016.
5. R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *Information Theory, IEEE Transactions on*, 56(4):1982–2001, 2010.
6. A. Bastounis and A. C. Hansen. On random and deterministic compressed sensing and the restricted isometry property in levels. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 297–301. IEEE, 2015.
7. A. Bourrier, M. Davies, T. Peleg, P. Perez, and R. Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. *Information Theory, IEEE Transactions on*, 60(12):7928–7946, Dec 2014.
8. A. Bourrier, R. Gribonval, and P. Pérez. Compressive gaussian mixture estimation. In *Proc. IEEE Conf. Acoustics, Speech and Signal Processing*, pages 6024–6028, 2013.
9. T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *Information Theory, IEEE Transactions on*, 60(1):122–132, Jan 2014.
10. T. T. Cai, A. Zhang, et al. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

11. E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, May 2008.
12. E. J. Candès and C. Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.
13. E. J. Candes and Y. Plan. Matrix Completion With Noise. *Proceedings of the IEEE*, 98(6):925–936, June 2010.
14. E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.
15. E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, Feb. 2006.
16. Y. D. Castro, F. Gamboa, D. Henrion, and J. B. Lasserre. Exact solutions to super resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630, Jan 2017.
17. V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.
18. A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc*, 22(1), 2009.
19. M. E. Davies and R. Gribonval. Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. *Information Theory, IEEE Transactions on*, 55(5):2203–2214, 2009.
20. S. Dirksen. Dimensionality reduction with subgaussian matrices: A unified theory. *Foundations of Computational Mathematics*, 16(5):1367–1396, 2016.
21. D. L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, 59(6):797–829, June 2006.
22. V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.
23. A. Eftekhari and M. B. Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *Applied and Computational Harmonic Analysis*, 39(1):67–109, 2015.
24. Y. C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *Signal Processing, IEEE Transactions on*, 58(6):3042–3054, 2010.
25. S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing.* Springer, 2013.
26. A. Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 2013.
27. R. Gribonval and M. Nielsen. Beyond sparsity: Recovering structured representations by ℓ^1 minimization and greedy algorithms. *Advances in Computational Mathematics*, 28(1):23–41, 2008.
28. C. J. Hillar and L.-H. Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.
29. N. Keriven, A. Bourrier, R. Gribonval, and P. Pérèz. Sketching for Large-Scale Learning of Mixture Models. In *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2015.
30. N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez. Sketching for Large-Scale Learning of Mixture Models. *Preprint https://hal.inria.fr/hal-01329195*, June 2016.
31. N. Keriven, N. Tremblay, Y. Traonmilin, and R. Gribonval. Compressive k-means. *To appear at ICASSP 2017, url : https://hal.archives-ouvertes.fr/hal-01386077*, 2016.
32. G. Kutyniok. Clustered sparsity and separation of cartoon and texture. *SIAM Journal on Imaging Sciences*, 6(2):848–874, 2013.

33. G. Puy, M. E. Davies, and R. Gribonval. Recipes for stable linear embeddings from hilbert spaces to $\mathbb{R}^m$. *To appear in Transactions on Information Theory, Preprint https://hal.inria.fr/hal-01203614*, 2017.
34. B. Recht, M. Fazel, and P. Parrilo. Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization. *SIAM Review*, 52(3):471–501, 2010.
35. J. C. Robinson. *Dimensions, embeddings, and attractors*, volume 186. Cambridge University Press, 2010.
36. C. Studer and R. G. Baraniuk. Stable restoration and separation of approximately sparse signals. *Applied and Computational Harmonic Analysis*, Sept. 2013.
37. C. Sutour, C.-A. Deledalle, and J.-F. Aujol. Estimation of the noise level function based on a nonparametric detection of homogeneous image regions. *SIAM Journal on Imaging Sciences*, 8(4):2622–2661, 2015.
38. Y. Traonmilin and R. Gribonval. Stable recovery of low-dimensional cones in hilbert spaces: One rip to rule them all. *Applied And Computational Harmonic Analysis, In Press*, 2016.
39. Y. Traonmilin, S. Ladjal, and A. Almansa. Robust multi-image processing with optimal sparse regularization. *Journal of Mathematical Imaging and Vision*, 51(3):413–429, 2015.
40. M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Transactions on Signal Processing*, 50(6):1417–1428, Jun 2002.
41. M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.