



**HAL**  
open science

# Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems

Anthony Bourrier, Mike E. Davies, Tomer Peleg, Patrick Pérez, Rémi Gribonval

► **To cite this version:**

Anthony Bourrier, Mike E. Davies, Tomer Peleg, Patrick Pérez, Rémi Gribonval. Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems. 2013. hal-00908358v1

**HAL Id: hal-00908358**

**<https://hal.science/hal-00908358v1>**

Preprint submitted on 22 Nov 2013 (v1), last revised 20 Nov 2014 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems

Anthony Bourrier      Mike E. Davies      Tomer Peleg      Patrick Pérez  
Rémi Gribonval

November 20, 2013

## Abstract

The primary challenge in linear inverse problems is to design stable and robust “decoders” to reconstruct high-dimensional vectors from a low-dimensional observation through a linear operator. Sparsity, low-rank, and related assumptions are typically exploited to design decoders which performance is then bounded based on some measure of deviation from the idealized model, typically using a norm.

This paper focuses on characterizing the fundamental performance limits that can be expected from an ideal decoder given a general model, *i.e.*, a general subset of “simple” vectors of interest. First, we extend the so-called notion of instance optimality of a decoder to settings where one only wishes to reconstruct some part of the original high dimensional vector from a low-dimensional observation. This covers practical settings such as medical imaging of a region of interest, or audio source separation when one is only interested in estimating the contribution of a specific instrument to a musical recording. We define instance optimality relatively to a model much beyond the traditional framework of sparse recovery, and characterize the existence of an instance optimal decoder in terms of joint properties of the model and the considered linear operator. Noiseless and noise-robust settings are both considered. We show somewhat surprisingly that the existence of *noise-aware* instance optimal decoders for all noise levels implies the existence of a *noise-blind* decoder.

A consequence of our results is that for models that are rich enough to contain an orthonormal basis, the existence of an  $\ell_2/\ell_2$  instance optimal decoder is only possible when the linear operator is not substantially dimension-reducing. This covers well-known cases (sparse vectors, low-rank matrices) as well as a number of seemingly new situations (structured sparsity and sparse inverse covariance matrices for instance). We exhibit an operator-dependent norm which, under a model-specific generalization of the Restricted Isometry Property (RIP), always yields a feasible instance optimality and implies instance optimality with certain familiar atomic norms such as the  $\ell_1$  norm.

## 1 Introduction

In linear inverse problems, one considers a linear measurement operator  $\mathbf{M}$  mapping the signal space  $\mathbb{R}^n$  to a measurement space  $\mathbb{R}^m$ , where typically  $\mathbf{M}$  is either ill-conditioned or dimensionality reducing. The reconstruction of  $\mathbf{x}$  from  $\mathbf{M}\mathbf{x}$  is thus a hopeless task unless one can exploit prior knowledge on  $\mathbf{x}$  to complete the incomplete observation  $\mathbf{M}\mathbf{x}$ .

Sparsity is a well-known enabling model for this type of inverse problems: it has been proven that one can expect to recover the signal  $\mathbf{x}$  from its measure  $\mathbf{M}\mathbf{x}$  provided that  $\mathbf{x}$  is sufficiently sparse, *i.e.*, it has few nonzero components [15]. If the set of  $k$ -sparse signals is denoted  $\Sigma_k = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 \leq k\}$ , where  $\|\cdot\|_0$  is the pseudo-norm counting the number of nonzero components, then this recovery property can be interpreted as the existence of a decoder  $\Delta : \mathbb{R}^m \rightarrow \mathbb{R}^n$  such that  $\forall \mathbf{x} \in \Sigma_k, \Delta(\mathbf{M}\mathbf{x}) = \mathbf{x}$ , thus making  $\mathbf{M}$  a linear encoder associated to the (typically nonlinear) decoder  $\Delta$ .

Further, the body of theoretical work around sparse recovery in linear inverse problems has given rise to the notion of compressive sensing (CS) [22], where the focus is on *choosing* a dimensionality reducing  $\mathbf{M}$  to which a decoder can be associated. It is now well-established that this can be achieved in scenarios where  $m \ll n$ , showing that a whole class of seemingly high-dimensional signals can thus be reconstructed from far lower dimensional linear measurements than their apparent dimension.

## 1.1 Instance optimal sparse decoders

A good decoder  $\Delta$  is certainly expected to have nicer properties than simply reconstructing  $\Sigma_k$ . Indeed, the signal  $\mathbf{x}$  to be reconstructed may not belong exactly in  $\Sigma_k$  but “live near”  $\Sigma_k$  under a distance  $d$ , meaning that  $d(\mathbf{x}, \Sigma_k)$  is “small” in a certain sense. In this case, one wants to be able to build a sufficiently precise estimate of  $\mathbf{x}$  from  $\mathbf{M}\mathbf{x}$ , that is a quantity  $\Delta(\mathbf{M}\mathbf{x})$  such that  $\|\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\|$  is “small” for a certain norm  $\|\cdot\|$ . This stability to the model has been formalized into the so-called *Instance Optimality* assumption on  $\Delta$ .  $\Delta$  is said to be instance optimal if:

$$\forall \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\| \leq Cd(\mathbf{x}, \Sigma_k), \quad (1)$$

for a certain choice of norm  $\|\cdot\|$  and distance  $d$ . For this property to be meaningful, the constant  $C$  must not scale with  $n$  and typically “good” instance optimal decoders are decoders which involve a constant which is the same for all  $n$  (note that this implicitly relies on the fact that a sparse set  $\Sigma_k \subset \mathbb{R}^n$  can be defined for any  $n$ ). When the norm is  $\ell_2$  or  $\ell_1$  and the distance is  $\ell_1$ , such good instance optimal decoders exist and can be implemented as the minimization of a convex objective [22, 17, 10] under assumptions on  $\mathbf{M}$  such as the Restricted Isometry Property (RIP). Note that instance optimality is a uniform upper bound on the reconstruction error, and that other types of bounds on decoders can be studied, particularly from a probabilistic point of view [18]. Other early work include upper bounds on the reconstruction error from noisy measurements with a regularizing function when the signal belongs exactly to the model [25].

In [19], the authors considered the following question: *Given the encoder  $\mathbf{M}$ , is there a simple characterization of the existence of an instance optimal decoder?* Their goal was not to find implementable decoders that would have this property, but rather to identify conditions on  $\mathbf{M}$  and  $\Sigma_k$  under which the reconstruction problem is ill-posed if one aims at finding an instance optimal decoder with small constant. The existence of a decoder  $\Delta$  which satisfies (1) will be called the *Instance Optimality Property* (IOP). The authors proved that this IOP is closely related to a property of the kernel of  $\mathbf{M}$  with respect to  $\Sigma_{2k}$ , called the *Null Space Property* (NSP). This relation allowed them to study the existence of stable decoders under several choices of norm  $\|\cdot\|$  and distance  $d(\cdot, \cdot)$ .

A related question addressed in [19] is that of the fundamental limits of dimension reduction: *Given the target dimension  $m$  and desired constant  $C$ , is there an encoder  $\mathbf{M}$  with an associated instance optimal decoder?* They particularly showed that there is a fundamental trade-off between

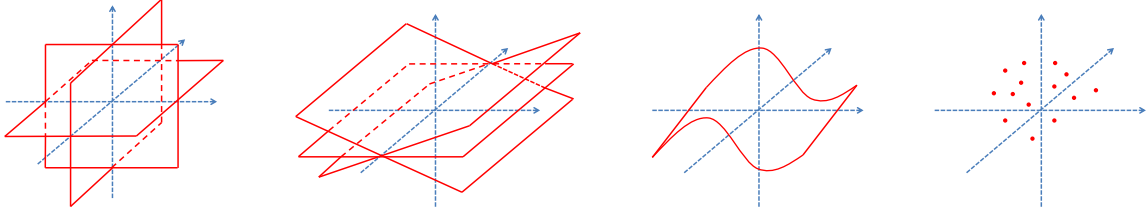


Figure 1: Illustration of several CS models. *From left to right:  $k$ -sparse vectors, union of subspaces, smooth manifold and point cloud.*

the size of the constant  $C$  in (1) (with  $\ell_2$  norm and  $\ell_2$  distance) and the dimension reduction ratio  $m/n$ .

## 1.2 Low-dimensional models beyond sparsity

Beyond the sparse model, many other low-dimensional models have been considered in the context of linear inverse problems and CS [5]. In these generalized models, the signals of interest typically live in or close to a subset  $\Sigma$  of the space, which typically contains far fewer vectors than the whole space. Such models encompass sets of elements as various as block-sparse signals [24], unions of subspaces, whether finite [8] or possibly infinite [7], signals sparse in a redundant dictionary [37], cospase signals [32], approximately low-rank matrices [38, 14], low-rank and sparse matrices [43, 12], symmetric matrices with sparse inverse [42, 41] or manifolds [6, 23]. An old result which can also be interpreted as generalized CS is the low-dimensional embedding of a point cloud [31, 1]. Some of these models are pictured in Figure 1.

Since these models generalize the sparse model, the following question arises: can they be considered under a general framework, sharing common reconstruction properties? In this work, we are particularly interested in the extension of the results of [19] to these general models, allowing to further investigate the well-posedness of such problems.

In [36], the theoretical results of [19] are generalized in the case where one aims at stably decoding a vector living near a finite union of subspaces (UoS). They also show in this case the impossibility of getting a good  $\ell_2/\ell_2$  instance optimal decoder with substantial dimensionality reduction. Their extension also covers the case where the quantity one wants to decode is not the signal itself but a linear measure of the signal.

In this work, we further extend the study of the IOP to general models of signals: we consider signals of interest living in or near a subset  $\Sigma$  of a vector space  $E$ , without further restriction, and show that instance optimality can be generalized for such models. In fact, we consider the following generalizations:

- **Robustness to noise:** noise-robust instance optimality is characterized, showing somewhat surprisingly the equivalence between the existence of two flavors of noise-robust decoders (*noise-aware* and *noise-blind*);
- **Infinite dimension:** signal spaces  $E$  that may be infinite dimensional are considered. For example  $E$  may be a Banach space such as an  $L^p$  space, the space of signed measures, etc. This is motivated by recent work on infinite dimensional compressed sensing [4, 30] or compressive density estimation [9];

- **Task-oriented decoders:** the decoder is not constrained to approximate the signal  $\mathbf{x}$  itself but rather a linear feature derived from the signal,  $\mathbf{Ax}$ , as in [36]; in the usual inverse problem framework,  $\mathbf{A}$  is the identity. Examples of problems where  $\mathbf{A} \neq \mathbf{I}$  include:
  - Medical imaging of a particular region of the body: as in Magnetic Resonance Imaging, one may acquire Fourier coefficients of a function defined on the body, but only want to reconstruct properly a particular region. In this case,  $\mathbf{A}$  would be the orthogonal projection on this region.
  - Partial source separation: given an audio signal mixed from several sources whose positions are known, as well as the microphone filters, the task of isolating one of the sources from the mixed signal is a reconstruction task where  $E$  is the space of concatenated sources, and  $\mathbf{A}$  orthogonally projects such a signal in a single source signal space.

### 1.3 This work

We summarize below our main contributions which fall under two categories: the characterization of the existence of an instance optimal decoder given a linear measurement operator  $\mathbf{M}$  and a model  $\Sigma$ ; the limits of dimensionality reduction.

#### 1.3.1 Instance optimality for inverse problems with general models

In the noiseless case, we express a concept of instance optimality which does not necessarily involve homogeneous norms and distances but some pseudo-norms instead. Such a generalized instance optimality can be expressed as follows:

$$\forall \mathbf{x} \in E, \|\mathbf{Ax} - \Delta(\mathbf{Mx})\|_G \leq C d_E(\mathbf{x}, \Sigma), \quad (2)$$

where  $\|\cdot\|_G$  is a pseudo-norm and  $d_E$  is a distance the properties of which will be specified in due time, and  $\mathbf{A}$  is a linear operator representing the feature one wants to estimate from  $\mathbf{Mx}$ . Our first contribution is to prove that the existence of a decoder  $\Delta$  satisfying (2), which is a generalized IOP, can be linked with a generalized NSP, similarly to the sparse case. This generalized NSP can be stated as:

$$\forall \mathbf{h} \in \ker(\mathbf{M}), \|\mathbf{Ah}\|_G \leq D d_E(\mathbf{h}, \Sigma - \Sigma), \quad (3)$$

where the set  $\Sigma - \Sigma$  is comprised of all differences of elements in  $\Sigma$ , that is  $\Sigma - \Sigma = \{\mathbf{z}_1 - \mathbf{z}_2 | \mathbf{z}_1, \mathbf{z}_2 \in \Sigma\}$ . The constants  $C$  and  $D$  are related by a factor no more than 2, as will be stated in Theorems 1 and 2 characterizing the relationships between these two properties. In particular, all previously mentioned low-dimensional models can fit in this generalized framework.

#### 1.3.2 Noise-robust instance optimality

Our second contribution (Theorems 3 and 4) is to provide a noise-robust extension of instance optimality and link it to a property called the Robust NSP. Section 2 regroups these noiseless and noise-robust results after a review of the initial IOP/NSP results of [19]. We show somewhat surprisingly that the existence of *noise-aware* instance optimal decoders for all noise levels implies the existence of a *noise-blind* decoder (Theorem 5).

### 1.3.3 Infinite-dimensional inverse problems

The generalization to arbitrary vector spaces allows us to consider infinite dimensional inverse problems. Here things are not always as straightforward as in the finite dimensional setting. For example, in the theory of generalized sampling [3], even when the signal model  $\Sigma$  is simply a finite dimensional subspace, it can be necessary to oversample by some factor in order to guarantee stable recovery. In fact Theorem 4.1 of [2] can be read as a statement of  $\ell_2/\ell_2$  instance optimality for a specific (linear) decoder given in terms of the NSP constant of the measurement operator. The results presented here therefore provide an extension of generalized sampling for linear models beyond  $\ell_2$ .

In [2] the authors also combine generalized sampling theory with compressed sensing. However, rather than developing a uniform instance optimality theory which seems the most natural extension of generalized sampling, they adopt a non-uniform approach based on [13]. Our results should enable the development of a uniform infinite dimensional CS theory.

### 1.3.4 Limits of dimensionality reduction with generalized models

The reformulation of IOP as an NSP allows us to consider the  $\ell_2/\ell_2$  instance optimality for general models in Section 3. In this case, the NSP can be interpreted in terms of scalar product and we precise the necessity of the NSP for the existence of an instance optimal decoder. This leads to the proof of Theorem 6 stating that, just as in the sparse case, *one cannot expect to build an  $\ell_2/\ell_2$  instance optimal decoder if  $\mathbf{M}$  reduces substantially the dimension and the model is “too large”* in a precise sense. In particular, we will see that the model is “too large” when the set  $\Sigma - \Sigma$  contains an orthonormal basis. This encompasses a wide range of standard models where a consequence of our results is that  $\ell_2/\ell_2$  IOP with dimensionality reduction is impossible:

- **$k$ -sparse vectors.** In the case where  $\Sigma = \Sigma_k$  is the set of  $k$ -sparse vectors,  $\Sigma$  contains the null vector and the canonical basis, so that  $\Sigma - \Sigma$  contains the canonical basis. Note that the impossibility of good  $\ell_2/\ell_2$  IOP has been proved in [19].
- **Block-sparse vectors** [24]. The same argument as above applies in this case as well, implying that imposing a block structure on sparsity does not improve  $\ell_2/\ell_2$  feasibility.
- **Low-rank matrices** [38, 14]. In the case where  $E = \mathcal{M}_n(\mathbb{R})$  and  $\Sigma$  is the set of matrices of rank  $\leq k$ ,  $\Sigma$  also contains the null matrix and the canonical basis.
- **Low-rank + sparse matrices** [43, 12]. The same argument applies to the case where the model contains all matrices that are jointly low-rank *and* sparse, which appear in phase retrieval [35, 34, 16].
- **Low-rank matrices with non-sparsity constraints.** In order to reduce the ambivalence of the low-rank + sparse decomposition of a matrix, [12] introduced non-sparsity constraints on the low-rank matrix in order to enforce its entries to have approximately the same magnitude. However, as shown in Lemma 3, an orthonormal Fourier basis of the matrix space can be written as differences of matrices which belong to this model.
- **Reduced union of subspace models** [5] obtained by pruning out the combinatorial collection of  $k$ -dimensional subspaces associated to  $k$ -sparse vectors. This covers block-sparse vectors [24], tree-structured sparse vectors, and more. Despite the fact that these unions of

subspaces may contain much fewer  $k$ -dimensional subspaces than the combinatorial number of subspaces of the standard  $k$ -sparse model, the same argument as in the  $k$ -sparse model applies to these signal models, provided they contain the basis collection of 1-sparse signals. This contradicts the naive intuition that  $\ell_2/\ell_2$  IOP could be achievable at the price of substantially reducing the richness of the model through a drastic pruning of its subspaces.

- **$k$ -sparse expansions in a dictionary model** [37]. More generally, if the model is the set of vectors which a linear combination of at most  $k$  elements of a dictionary  $\mathbf{D}$  which contains an orthogonal family or a tight frame, then Theorem 6 applies.
- **Cosparse vectors with respect to the finite difference operator** [32, 36]. As shown in [36], the canonical basis is highly cosparse with respect to the finite difference operator, hence it is contained in the corresponding union of subspaces.
- As shown in Lemma 2, this is also the case for **symmetric definite positive square matrices with  $k$ -sparse inverse**. The covariance matrix of high-dimensional Gaussian graphical models is of this type: the numerous pairwise conditional independences that characterize the structure of such models, and make them tractable, translate into zeros entries of the inverse covariance matrix (the concentration matrix). Combining sparsity prior on the concentration matrix with maximum likelihood estimation of covariance from data, permits to learn jointly the structure and the parameters of Gaussian graphical models (so called “covariance selection” problem) [42, 41]. In very high-dimensional cases, compressive solutions to this problem would be appealing.
- Johnson-Lindenstrauss embedding of **point clouds** [1]. Given a set  $\mathcal{X}$  of  $L$  vectors in  $\mathbb{R}^n$  and  $\epsilon > 0$ , there exists a linear mapping  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , with  $m = \mathcal{O}(\ln(L)/\epsilon^2)$  and

$$(1 - \epsilon)\|\mathbf{x} - \mathbf{y}\|_2 \leq \|f(\mathbf{x}) - f(\mathbf{y})\|_2 \leq (1 + \epsilon)\|\mathbf{x} - \mathbf{y}\|_2 \quad (4)$$

holds for all  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ . The fact that the point cloud contains a tight frame is satisfied if it “spreads” in a number of directions which span the space. In this case, *one cannot guarantee precise out-of-sample reconstruction* of the points in  $\mathbb{R}^n$  in the  $\ell_2$ -sense, except for a very limited neighborhood of the point cloud. This is further discussed in Section 5.

### 1.3.5 Generalized Restricted Isometry Property

Our last contribution, in Section 4, is to study the relations between the NSP and a generalized version of the Restricted Isometry Property (RIP), which encompasses classical or recent RIP formulations, such as

- the **D-RIP** [11] for the dictionary model;
- the RIP for low-rank matrices [14];
- the  **$\Omega$ -RIP** [28] for the cosparse model.

This generalized RIP bounds  $\|\mathbf{M}\mathbf{x}\|_F$  from below and/or above on a certain set  $V$ , and can be decomposed in:

$$\text{Lower - RIP : } \forall \mathbf{x} \in V, \alpha \|\mathbf{x}\|_G \leq \|\mathbf{M}\mathbf{x}\|_F \quad (5)$$

$$\text{Upper - RIP : } \forall \mathbf{x} \in V, \|\mathbf{M}\mathbf{x}\|_F \leq \beta \|\mathbf{x}\|_G, \quad (6)$$

where  $\|\cdot\|_G$  and  $\|\cdot\|_F$  are norms defined respectively on the signal space and on the measure space, and  $0 < \alpha \leq \beta < +\infty$ . We prove particularly in Theorem 7 that a generalized lower-RIP on  $\Sigma - \Sigma$  implies the existence of instance optimal decoders in the noiseless and the noisy cases for a certain norm  $\|\cdot\|_E$  we call the “ $M$ -norm”<sup>1</sup>.

Furthermore, we prove that under an upper-RIP assumption on  $\Sigma$ , this  $M$ -norm can be upper bounded by an atomic norm [18] defined using  $\Sigma$  and denoted  $\|\cdot\|_\Sigma$ . This norm is easier to interpret than the  $M$ -norm: it can in particular be upper bounded by usual norms for the  $k$ -sparse vectors and low-rank matrices models. We have the following general result relating generalized RIP and IOP (Theorem 8): if  $\mathbf{M}$  satisfies a lower-RIP (5) for  $V = \Sigma - \Sigma$  and an upper-RIP (6) for  $V = \Sigma$ , then for all  $\delta > 0$ , there exists a decoder  $\Delta_\delta$  satisfying  $\forall \mathbf{x} \in E, \forall \mathbf{e} \in F$ ,

$$\|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq 2 \left(1 + \frac{\beta}{\alpha}\right) d_\Sigma(\mathbf{x}, \Sigma) + \frac{2}{\alpha} \|\mathbf{e}\|_E + \delta, \quad (7)$$

which is a particular case of Robust instance optimality, as described in Section 2.

## 1.4 Structure of the paper

We will now describe the layout of the paper. Section 2 first contains a quick review of the relationship between IOP and NSP in the usual sparse case, then exposes the more general setting considered in this paper, for which these properties and their relationship are extended, both in noiseless and noisy settings. Section 3 then focuses on the particular case of  $\ell_2/\ell_2$  IOP, proving the impossibility for a certain class of models to achieve such IOP with decent precision in dimension reducing scenarii. In particular, we show that this encompasses a wide range of usual models. Finally, in Section 4, we get back to the problem of IO with general norms and prove that a generalized version of the lower-RIP implies the existence of an instance optimal decoder for a certain norm we call the “ $M$ -norm”. We propose an upper-bound on this norm under a generalized upper-RIP assumption to get an IOP with simpler norms, illustrating the result in standard cases.

## 2 Generalized IOP and NSP equivalences

In this section, we review the initial IOP/NSP relationship before extending it in several ways.

### 2.1 The initial result

In [19], the authors consider two norms  $\|\cdot\|_G$  and  $\|\cdot\|_E$  defined on the signal space  $\mathbb{R}^n$ . The distance derived from  $\|\cdot\|_E$  will be denoted  $d_E$ . Given a vector  $\mathbf{x} \in \mathbb{R}^n$  and a subset  $A \subset \mathbb{R}^n$ , the distance from  $\mathbf{x}$  to  $A$  is defined as  $d_E(\mathbf{x}, A) = \inf_{\mathbf{y} \in A} \|\mathbf{x} - \mathbf{y}\|_E$ . These two norms allow the definition of instance optimality: a decoder  $\Delta : \mathbb{R}^m \rightarrow \mathbb{R}^n$  is said to be instance optimal for  $k$ -sparse signals if

$$\forall \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\|_G \leq C d_E(\mathbf{x}, \Sigma_k), \quad (8)$$

for some constant  $C > 0$ .

This property on  $\Delta$  upper bounds the reconstruction error of a vector, measured by  $\|\cdot\|_G$ , by the distance from the vector to the model, measured by  $d_E$ . The authors prove that the existence

---

<sup>1</sup>The prefix “ $M$ -” should be thought as “Measurement-related norm” since in other works the measurement matrix may be denoted by other letters.



of an instance optimal decoder, called IOP, is closely related to the NSP of  $\mathbf{M}$  with respect to the set  $\Sigma_{2k}$  of  $2k$ -sparse vectors. Noting  $\mathcal{N} = \ker(\mathbf{M})$ , this NSP states

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_G \leq D d_E(\mathbf{h}, \Sigma_{2k}) \quad (9)$$

for some constant  $D$ .

The relationship between the IOP and the NSP is the following: if there exists an instance optimal decoder  $\Delta$  satisfying (8), then (9) holds with  $D = C$ . Conversely, if (9) holds, then there exists a decoder  $\Delta$  such that (8) holds with  $C = 2D$ . Such a decoder can be defined as follows, supposing  $\mathbf{M}$  is onto:

$$\Delta(\mathbf{M}\mathbf{x}) = \underset{\mathbf{z} \in (\mathbf{x} + \mathcal{N})}{\operatorname{argmin}} d_E(\mathbf{z}, \Sigma_k), \quad (10)$$

$\mathbf{x} + \mathcal{N}$  denoting the set  $\{\mathbf{x} + \mathbf{h}, \mathbf{h} \in \mathcal{N}\}$ . The well-posedness of this definition is discussed in Appendix A, in the more general setting where the model is a finite union of subspaces in finite dimension. Note that for generalized models, such a decoder may not necessarily exist since the infimum of  $d_E(\mathbf{z}, \Sigma)$  may not be achieved, as we will discuss in the next section.

This result can be seen as an “equivalence” between the IOP and the NSP, with similar constants.

## 2.2 Proposed extensions

The framework we consider is more general. The signal space is a vector space  $E$ , possibly infinite-dimensional. In particular,  $E$  may be a Banach space such as an  $L^p$  space, the space of signed measures, etc. On this space is defined a linear operator  $\mathbf{M} : E \rightarrow F$ , where  $F$  is the measurement space, which will most likely be finite-dimensional in practice. We assume that  $\mathbf{M}$  is onto. We further define a signal model  $\Sigma \subset E$  comprising the signals which we want to be able to “reconstruct” from their images by  $\mathbf{M}$ . In the framework we consider, this “reconstruction” is not necessarily an inverse problem where we want to recover  $\mathbf{x}$  from  $\mathbf{M}\mathbf{x}$ . More precisely, as in [36], we consider a case where we want to recover from  $\mathbf{M}\mathbf{x}$  a quantity  $\mathbf{A}\mathbf{x}$ , where  $\mathbf{A}$  is a linear operator mapping  $E$  into a space  $G$ . When  $G = E$  and  $\mathbf{A} = \mathbf{I}$ , we are brought back to the usual case where we want to reconstruct  $\mathbf{x}$ . This generalized framework is illustrated in Figure 2.

In this generalized framework, we are now interested in the concepts of IOP and NSP, as well as their relationship. A decoder  $\Delta : F \rightarrow G$  will aim at approximating  $\mathbf{A}\mathbf{x}$  from  $\mathbf{M}\mathbf{x}$ .

The approximation error will be measured by a function  $\|\cdot\|_G : G \rightarrow \mathbb{R}_+$ . This function needs not be a norm in order to state the following results. It still must satisfy the following properties:

$$\mathbf{Symmetry} : \|\mathbf{x}\|_G = \|-\mathbf{x}\|_G \quad (11)$$

$$\mathbf{Triangle\ inequality} : \|\mathbf{x} + \mathbf{y}\|_G \leq \|\mathbf{x}\|_G + \|\mathbf{y}\|_G. \quad (12)$$

The differences with a regular norm is that neither definiteness nor homogeneity is required:  $\|\mathbf{x}\|_G = 0$  needs not imply  $\mathbf{x} = 0$  and  $\|\lambda\mathbf{x}\|_G$  needs not equal  $|\lambda|\|\mathbf{x}\|_G$ . We provide two examples of such pseudo-norms in the case where  $G = \mathbb{R}^n$ :

- $\|\cdot\|_G$  can be defined as a “non-normalized”  $\ell_p$ -quasinorm for  $0 \leq p \leq 1$ , that is  $\|\mathbf{x}\|_G = \sum_{i=1}^n |x_i|^p$ . In this case,  $\|\lambda\mathbf{x}\|_G = |\lambda|^p \|\mathbf{x}\|_G$ .
- More generally, if  $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  is a concave function such that  $f(x) = 0 \Leftrightarrow x = 0$ , then  $\|\cdot\|_G$  can be defined as the  $f$ -(pseudo-)norm  $\|\mathbf{x}\|_f = \sum_{i=1}^n f(|x_i|)$ , see [29].

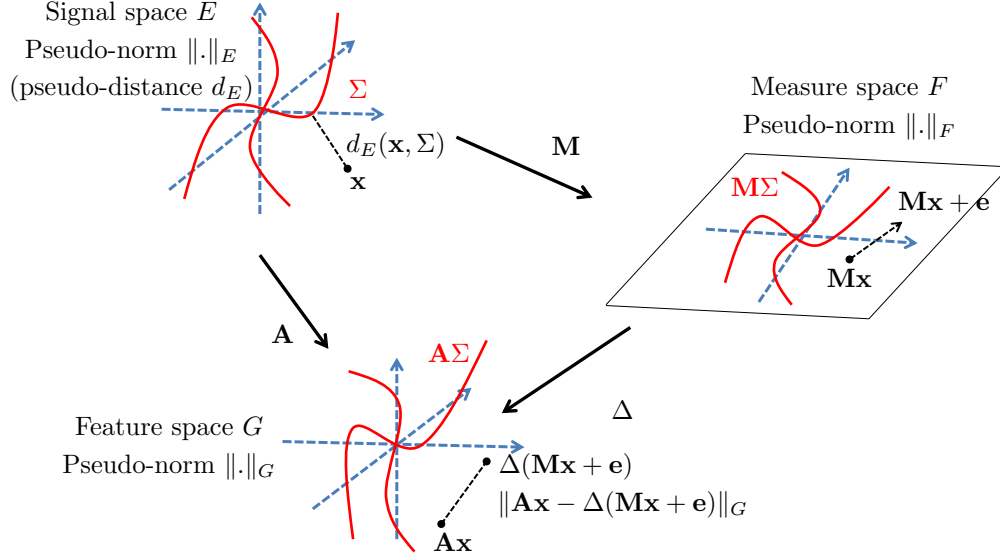


Figure 2: Illustration of the proposed generalized setting. The signals belong to the space  $E$ , supplied with a pseudo-norm  $\|\cdot\|_E$  used to measure the distance from a vector to the model  $\Sigma$  containing the signals of interest.  $E$  is mapped in the measure space  $F$  by the operator  $\mathbf{M}$  and the measure is perturbed by an additive noise  $\mathbf{e}$ . The space  $F$  is supplied with a pseudo-norm  $\|\cdot\|_F$ . The feature space  $G$ , supplied with a norm  $\|\cdot\|_G$ , is composed of vectors obtained by applying a linear operator  $\mathbf{A}$  to the signals in  $E$ . These feature vectors are the vectors one wants to reconstruct from the measures in  $\mathbf{M}$  by applying a decoder  $\Delta$ . The reconstruction error for the vector  $\mathbf{x}$  and noise  $\mathbf{e}$  is therefore  $\|\mathbf{A}\mathbf{x} - \Delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G$ . Note that in the case where  $E = G$  and  $\mathbf{A} = \mathbf{I}$ , the decoder is aimed at reconstructing exactly the signals.

In order to measure the *distance from a vector to the model*, we also endow  $E$  with a pseudo-norm  $\|\cdot\|_E : E \rightarrow \mathbb{R}_+$  which satisfies the same properties as  $\|\cdot\|_G$  with the additional requirement that  $\|0\|_E = 0$ . The pseudo-distance  $d_E$  is defined on  $E^2$  by  $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_E$ . Yet again,  $\|\cdot\|_E$  can be defined as a non-normalized  $\ell_p$ -norm or an  $f$ -norm.

We will also consider a noisy framework where the measure  $\mathbf{M}\mathbf{x}$  is perturbed by an additive noise term  $\mathbf{e}$ . To consider IOP and NSP in this context, we *measure the amount of noise* with a pseudo-norm in the measurement space  $F$ , which we will denote by  $\|\cdot\|_F$ . The assumptions we make on  $\|\cdot\|_F$  are the same as the assumptions on  $\|\cdot\|_E$ .

To sum up, here are the extensions we propose compared to the framework of [19, 36] :

- The measure  $\mathbf{M}\mathbf{x}$  can be perturbed by an additive noise  $\mathbf{e}$ .
- The model set  $\Sigma$  can be any subset of  $E$ .
- $E$  is not necessarily  $\mathbb{R}^n$  but can be any vector space, possibly infinite-dimensional.
- The reconstruction of  $\mathbf{A}\mathbf{x}$  is targeted rather than that of  $\mathbf{x}$ .

	Triangle Inequality	Symmetry	$\ 0\  = 0$	Definiteness	Homogeneity
$\ \cdot\ _E$	X	X	X	-	-
$\ \cdot\ _F$	X	X	X	-	-
$\ \cdot\ _G$	X	X	-	-	-

Table 1: Summary of the hypotheses on the pseudo-norms  $\|\cdot\|_E$ ,  $\|\cdot\|_F$  and  $\|\cdot\|_G$ . A cross means the property is required, a horizontal bar means it is not.

- The functions  $\|\cdot\|_E$ ,  $\|\cdot\|_F$  and  $\|\cdot\|_G$  need not be norms but can be pseudo-norms with relaxed hypotheses. In particular, Table 1 summarizes the requirements on these functions.

### 2.2.1 The noiseless case

We first consider the same framework as [19, 36], where one measures  $\mathbf{M}\mathbf{x}$  with infinite precision. In our generalized framework, instance optimality for a decoder  $\Delta$  reads:

$$\forall \mathbf{x} \in E, \|\mathbf{A}\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\|_G \leq C d_E(\mathbf{x}, \Sigma).$$

We will prove that if IOP holds, *i.e.*, if the above holds for a certain decoder  $\Delta$ , then a generalized NSP is satisfied, that is:

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{A}\mathbf{h}\|_G \leq D d_E(\mathbf{h}, \Sigma - \Sigma),$$

with  $D = C$ . Note that the set  $\Sigma_{2k}$  has been replaced by  $\Sigma - \Sigma = \{\mathbf{x} - \mathbf{y}, \mathbf{x} \in \Sigma, \mathbf{y} \in \Sigma\}$ . When  $\Sigma = \Sigma_k$ , we have indeed  $\Sigma - \Sigma = \Sigma_{2k}$ .

The construction of an instance optimal decoder from the NSP is more complicated and the form of the instance optimality we get depends on additional assumptions on  $\Sigma$  and  $\mathbf{M}$ . Let's first suppose that for all  $\mathbf{x} \in E$ , there exists  $\mathbf{z} \in (\mathbf{x} + \mathcal{N})$  such that  $d_E(\mathbf{z}, \Sigma) = d_E(\mathbf{x} + \mathcal{N}, \Sigma)$ . Then the NSP (3) implies the existence of an instance optimal decoder satisfying (2) with  $C = 2D$ . If this assumption is not true anymore, then the NSP implies a slightly modified IOP, which states, for any  $\delta > 0$ , the existence of a decoder  $\Delta_\delta$  such that:

$$\forall \mathbf{x} \in E, \|\mathbf{A}\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x})\|_G \leq C d_E(\mathbf{x}, \Sigma) + \delta, \quad (13)$$

reflecting the fact that one cannot necessarily consider the exact quantity

$$\operatorname{argmin}_{\mathbf{z} \in (\mathbf{x} + \mathcal{N})} d_E(\mathbf{z}, \Sigma)$$

but rather a certain vector  $\mathbf{z} \in (\mathbf{x} + \mathcal{N})$  satisfying  $d_E(\mathbf{z}, \Sigma) \leq d_E(\mathbf{x} + \mathcal{N}, \Sigma) + \delta$ . A similar positive “projection error” appears in [7].

**Remark 1.** *To understand the necessity of such an additive error term when  $\Sigma$  is a general set, we can consider the following toy example depicted in Figure 3 where  $E = \mathbb{R}^2$ ,  $\mathcal{N} = \mathbb{R} \times \{0\}$ ,  $\Sigma = \{(x_1, x_2) \in (\mathbb{R}_+)^2 : x_2 = \frac{1}{x_1}\}$  and  $\|\cdot\|_G / \|\cdot\|_E$  are the  $\ell_2$  norm. In this case, the minimal distance between  $\mathbf{x} + \mathcal{N}$  and  $\Sigma$  is not reached at any point, making it necessary to add the  $\delta$  term for the decoder to be well-defined.*

In this setting, the NSP (3) implies the existence of instance optimal decoders in the sense of (13) for all  $\delta > 0$ . Moreover, this weak IOP formulation still implies the regular NSP with  $D = C$ . This is summarized in Theorems 1 and 2.

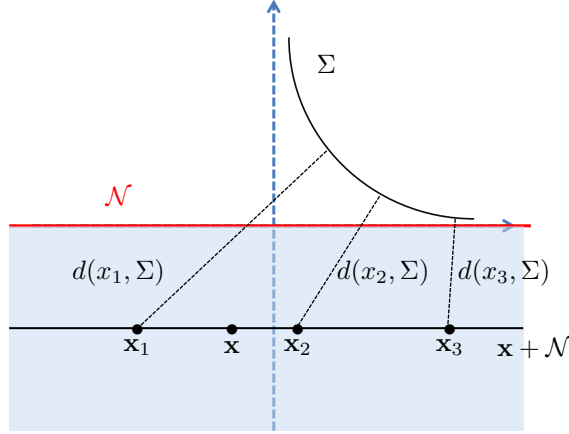


Figure 3: Necessity of the additive term  $\delta$  in a simple case. For each  $\mathbf{x}$  in the blue half-plane, the distance  $d_E(x + \mathcal{N}, \Sigma)$  is never reached at a particular point of  $\mathbf{x} + \mathcal{N}$ : the distance strictly decreases as one goes right along the affine plane  $\mathbf{x} + \mathcal{N}$  ( $d(\mathbf{x}_1, \Sigma) < d(\mathbf{x}_2, \Sigma) < d(\mathbf{x}_3, \Sigma)$ ), so that the minimal distance is reached “at infinity”.

**Theorem 1.** Suppose  $\forall \delta > 0$ , there exists a decoder  $\Delta_\delta$  satisfying (13):

$$\forall \mathbf{x} \in E, \|\mathbf{Ax} - \Delta_\delta(\mathbf{Mx})\|_G \leq C d_E(\mathbf{x}, \Sigma) + \delta.$$

Then  $\mathbf{M}$  satisfies the NSP (3):

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{Ah}\|_G \leq D d_E(\mathbf{h}, \Sigma - \Sigma),$$

with constant  $D = C$ .

**Theorem 2.** Suppose that  $\mathbf{M}$  satisfies the NSP (3):

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{Ah}\|_G \leq D d_E(\mathbf{h}, \Sigma - \Sigma).$$

Then  $\forall \delta > 0$ , there exists a decoder  $\Delta_\delta$  satisfying (13):

$$\forall \mathbf{x} \in E, \|\mathbf{Ax} - \Delta_\delta(\mathbf{Mx})\|_G \leq C d_E(\mathbf{x}, \Sigma) + \delta,$$

with  $C = 2D$ .

If we further assume that

$$\forall \mathbf{x} \in E, \exists \mathbf{z} \in (\mathbf{x} + \mathcal{N}), d_E(\mathbf{z}, \Sigma) = d_E(\mathbf{x} + \mathcal{N}, \Sigma), \quad (14)$$

then there exists a decoder  $\Delta$  satisfying (2):

$$\forall \mathbf{x} \in E, \|\mathbf{Ax} - \Delta(\mathbf{Mx})\|_G \leq C d_E(\mathbf{x}, \Sigma) \quad (15)$$

with  $C = 2D$ .

Note that this result is similar to the result proven in [36], which was stated in the case where  $\Sigma$  is a finite union of subspaces in finite dimension. In this framework, condition (14) is always satisfied as soon as  $\|\cdot\|_E$  is a norm, by the same argument as in usual CS (see Appendix A).

Let's also note the following property: if  $\|\cdot\|_E$  is definite, that is  $\|\mathbf{x}\|_E = 0 \Rightarrow \mathbf{x} = 0$ , then  $d_E$  is a distance. In the following proposition, we prove that if we further suppose that the set  $\Sigma + \mathcal{N}$  is a closed set with respect to  $d_E$ , then the NSP (3) implies for any  $\delta > 0$  the existence of a decoder  $\Delta_\delta$  satisfying (2) with  $C = (2 + \delta)D$ . This assumption therefore allows us to suppress the additive constant in (13) and replace it by an arbitrarily small increase in the multiplicative constant of (2).

**Proposition 1.** *Suppose that  $\mathbf{M}$  satisfies the NSP (3), that  $d_E$  is a distance and that  $\Sigma + \mathcal{N}$  is a closed set with respect to  $d_E$ . Then  $\forall \delta > 0$ , there exists a decoder  $\Delta_\delta$  satisfying:*

$$\forall \mathbf{x} \in E, \|\mathbf{Ax} - \Delta_\delta(\mathbf{Mx})\|_G \leq (2 + \delta)Dd_E(\mathbf{x}, \Sigma). \quad (16)$$

### 2.2.2 The noisy case

In practice, it is not likely that one can measure with infinite precision the quantity  $\mathbf{Mx}$ . This measure is likely to be contaminated with some noise, which will be considered in the following as an additive term  $\mathbf{e} \in F$ , so that the measure one gets is  $\mathbf{y} = \mathbf{Mx} + \mathbf{e}$ . In this case, a good decoder should be robust to noise, so that moderate values of  $\mathbf{e}$  should not have a severe impact on the approximation error. We are interested in the existence of similar results as before in this noisy setting.

We first need to define a noise-robust version of instance optimality. The robustness to noise of practical decoders is in fact a problem that has been considered by many authors. A first type of result considers *noise-aware decoders*, where given the noise level  $\epsilon \geq 0$  a decoder  $\Delta$  fulfills the following property

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{e}\|_F \leq \epsilon \Rightarrow \|\mathbf{Ax} - \Delta(\mathbf{Mx} + \mathbf{e})\|_G \leq C_1d_E(\mathbf{x}, \Sigma) + C_2\epsilon. \quad (17)$$

Here, the upper bound on the approximation error gets a new term measuring the amplitude of the noise. For example, this noise-robust instance optimality holds for a noise-aware  $\ell_1$  decoder in the sparse case with bounded noise [10] for  $\|\cdot\|_G = \|\cdot\|_2$  and  $\|\cdot\|_E = \|\cdot\|_1/\sqrt{k}$ , provided  $\mathbf{M}$  satisfies the RIP on  $\Sigma_{2k}$ .

In practical settings, it is hard to assume that one knows precisely the noise level. To exploit the above guarantee with a noise-aware decoder, one typically needs to overestimate the noise level. This loosens the effective performance guarantee and potentially degrades the actual performance of the decoder. An apparently stronger property for a decoder is to be robust *even without knowledge of the noise level*:

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{Ax} - \Delta(\mathbf{Mx} + \mathbf{e})\|_G \leq C_1d_E(\mathbf{x}, \Sigma) + C_2\|\mathbf{e}\|_F. \quad (18)$$

Further on, such decoders will be referred to as *noise-blind*. Guarantees of this type have been obtained under a RIP assumption for practical decoders such as iterative hard thresholding, CoSAMP, or hard thresholding pursuit, see e.g. [26, Corollary 3.9].

Of course, the existence of a *noise-blind* noise-robust decoder in the sense of (18) implies the existence of a *noise-aware* noise-robust decoder in the sense of (17) for any noise level  $\epsilon$ . We will see that, somewhat surprisingly, the converse is true in a sense, for both are equivalent to a noise-robust NSP.

Just as in the noiseless case, dealing with an arbitrary model  $\Sigma$  and possibly infinite dimensional  $E$  requires some caution. For  $\delta > 0$ , the noise-robust (and noise-blind) instance optimality of a decoder  $\Delta_\delta$  is defined as:

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{Ax} - \Delta_\delta(\mathbf{Mx} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2 \|\mathbf{e}\|_F + \delta. \quad (19)$$

One can see that  $\Delta_\delta$  necessarily also satisfies the noiseless instance optimality (13) by setting  $\mathbf{e} = 0$ .

As we show below, if for every  $\delta > 0$  there exists a noise-robust instance optimal decoder  $\Delta_\delta$  satisfying (19), then a generalized NSP for  $\mathbf{M}$  relatively to  $\Sigma - \Sigma$ , referred to as Robust NSP, must hold:

$$\forall \mathbf{h} \in E, \|\mathbf{Ah}\|_G \leq D_1 d_E(\mathbf{h}, \Sigma - \Sigma) + D_2 \|\mathbf{Mh}\|_F, \quad (20)$$

with  $D_1 = C_1$  and  $D_2 = C_2$ . This property appears e.g. in [27] (Chap. 4) with  $\|\cdot\|_G = \|\cdot\|_E = \|\cdot\|_1$  and  $\|\cdot\|_F$  any norm. Note that this Robust NSP concerns every vector of  $E$  and not just the vectors of the null space  $\mathcal{N} = \ker(\mathbf{M})^2$ . In the case where  $\mathbf{h} \in \mathcal{N}$ , one retrieves the regular NSP. For other vectors  $\mathbf{h}$ , another additive term, measuring the “size” of  $\mathbf{Mh}$ , appears in the upper bound.

Conversely, the Robust NSP implies the existence of noise-robust instance optimal decoders  $\Delta_\delta$  satisfying (19) with  $C_1 = 2D_1$  and  $C_2 = 2D_2$  for all  $\delta > 0$ . These results are summarized in Theorems 3 and 4.

**Theorem 3.** *Suppose  $\forall \delta > 0$ , there exists a decoder  $\Delta_\delta$  satisfying (19):*

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \Rightarrow \|\mathbf{Ax} - \Delta_\delta(\mathbf{Mx} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2 \|\mathbf{e}\|_F + \delta.$$

*Then  $\mathbf{M}$  satisfies the Robust NSP (20):*

$$\forall \mathbf{h} \in E, \|\mathbf{Ah}\|_G \leq D_1 d_E(\mathbf{h}, \Sigma - \Sigma) + D_2 \|\mathbf{Mh}\|_F,$$

*with constants  $D_1 = C_1$  and  $D_2 = C_2$ .*

**Theorem 4.** *Suppose that  $\mathbf{M}$  satisfies the Robust NSP (20):*

$$\forall \mathbf{h} \in E, \|\mathbf{Ah}\|_G \leq D_1 d_E(\mathbf{h}, \Sigma - \Sigma) + D_2 \|\mathbf{Mh}\|_F.$$

*Then  $\forall \delta > 0$ , there exists a decoder  $\Delta_\delta$  satisfying (19):*

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{Ax} - \Delta_\delta(\mathbf{Mx} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2 \|\mathbf{e}\|_F + \delta,$$

*with constants  $C_1 = 2D_1$  and  $C_2 = 2D_2$ .*

We conclude this section by discussing the relation between noise-aware and noise-blind decoders. A noise-aware version of noise-robust instance optimality can be defined where for  $\epsilon \geq 0, \delta > 0$  we require

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{e}\|_F \leq \epsilon \Rightarrow \|\mathbf{Ax} - \Delta_{\delta, \epsilon}(\mathbf{Mx} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2 \epsilon + \delta. \quad (21)$$

Of course, the existence of a noise-blind instance optimal decoder implies that of noise-aware decoders for every  $\epsilon \geq 0$ . The converse is indeed essentially true, up to the value of the constants  $C_i$ :

---

<sup>2</sup>In fact, unlike the NSP (3), (20) is not purely a property of the null space  $\mathcal{N}$  even though it implies the NSP. The name Robust NSP is thus somewhat improper, but has become a standard for this type of property.

**Theorem 5.** *Suppose  $\forall \epsilon, \delta > 0$ , there exists a noise-aware decoder  $\Delta_{\delta, \epsilon}$  satisfying (21):*

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{e}\|_F \leq \epsilon \Rightarrow \|\mathbf{Ax} - \Delta_{\delta, \epsilon}(\mathbf{Mx} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2 \epsilon + \delta.$$

*Then  $\mathbf{M}$  satisfies the Robust NSP (20) with constants  $D_1 = C_1$  and  $D_2 = 2C_2$ . Therefore, by Theorem 4, there exists an instance optimal noise-blind decoder satisfying:*

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{Ax} - \Delta_\delta(\mathbf{Mx} + \mathbf{e})\|_G \leq 2C_1 d_E(\mathbf{x}, \Sigma) + 4C_2 \|\mathbf{e}\|_F + \delta.$$

### 3 $\ell_2/\ell_2$ Instance Optimality

In this section, we suppose that  $E$  is a Hilbert space equipped with the norm  $\|\cdot\|_2$  and scalar product  $\langle \cdot, \cdot \rangle$ , that  $F = \mathbb{R}^m$  and we consider a finite-dimensional subspace  $V$  of dimension  $n$ , on which we define the measure operator  $\mathbf{M}$ . We are interested in the following question in the noiseless framework: *Is it possible to have a “good” noiseless instance optimal decoder with  $\|\cdot\|_G = \|\cdot\|_E = \|\cdot\|_2$  in a dimensionality reducing context where  $m \ll n$ ?*

A result of [19] states that in the usual sparse setting, one cannot expect to get a good instance optimal decoder if  $\mathbf{M}$  performs a substantial dimensionality reduction, the best corresponding constant being  $\sqrt{\frac{n}{m}}$ . In [36], the authors prove that this lower bound on the constant holds in the case where  $\Sigma$  is a finite union of subspaces in finite dimension. Here, we are interested in a version of this result for the general case where  $\Sigma$  can be a more general subset of  $E$ . More precisely, we will give a sufficient condition on  $\Sigma$  under which the optimal  $\ell_2/\ell_2$  instance optimality constant is of the order of  $\sqrt{\frac{n}{m}}$ , thus preventing the existence of a  $\ell_2/\ell_2$  instance optimal decoder with small constant if  $m \ll n$ .

#### 3.1 Homogeneity of the NSP

In the case where  $\|\cdot\|_G$ ,  $\|\cdot\|_E$  and  $\|\cdot\|_F$  are actual norms, the general NSP can be rewritten as an NSP holding on the cone  $\mathbb{R}(\Sigma - \Sigma)$  generated by  $\Sigma - \Sigma$ , i.e., the set  $\{\lambda \mathbf{z} | \lambda \in \mathbb{R}, \mathbf{z} \in \Sigma - \Sigma\}$ .

**Lemma 1.** *If  $\|\cdot\|_G$  and  $\|\cdot\|_E$  are norms, we have an equivalence between the NSP on  $\Sigma - \Sigma$ :*

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{Ah}\|_G \leq D d_E(\mathbf{h}, \Sigma - \Sigma), \quad (22)$$

*and the NSP on  $\mathbb{R}(\Sigma - \Sigma)$ :*

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{Ah}\|_G \leq D d_E(\mathbf{h}, \mathbb{R}(\Sigma - \Sigma)). \quad (23)$$

*Similarly, if  $\|\cdot\|_G$ ,  $\|\cdot\|_E$  and  $\|\cdot\|_F$  are norms, we have an equivalence between the robust NSP on  $\Sigma - \Sigma$ :*

$$\forall \mathbf{h} \in E, \|\mathbf{Ah}\|_G \leq D_1 d_E(\mathbf{h}, \Sigma - \Sigma) + D_2 \|\mathbf{Mh}\|_F, \quad (24)$$

*and the robust NSP on  $\mathbb{R}(\Sigma - \Sigma)$ :*

$$\forall \mathbf{h} \in E, \|\mathbf{Ah}\|_G \leq D_1 d_E(\mathbf{h}, \mathbb{R}(\Sigma - \Sigma)) + D_2 \|\mathbf{Mh}\|_F. \quad (25)$$

This lemma, which is valid even in the case where  $\mathbf{A}$  is not the identity, shows that the NSP imposes a constraint on the whole linear cone spanned by the elements of  $\Sigma - \Sigma$  and not only on the elements themselves. Note that this equivalence is trivial in the case where  $\Sigma$  is a union of subspaces since  $\Sigma - \Sigma$  is already a cone in this case.

### 3.2 The optimal $\ell_2/\ell_2$ NSP constant

**Remark 2.** In the subsequent sections of the paper, we will assume that  $\mathbf{A} = \mathbf{I}$  (this implies  $G = E$ ), so that one aims at reconstructing the actual signal.

In the  $\ell_2/\ell_2$  case, one can give a simple definition of the optimal NSP constant  $D_*$ , that is the minimal real positive number  $D$  such that the  $\ell_2/\ell_2$  NSP is satisfied with constant  $D$ :

$$D_* = \inf \{D \in \mathbb{R}_+ | \forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_2 \leq D d_2(\mathbf{h}, \Sigma - \Sigma)\}. \quad (26)$$

This definition assumes that there exists some constant so that the NSP is satisfied. Using the NSP definition and Lemma 1, we get that

$$D_* = \sup_{\mathbf{h} \in \mathcal{N}} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma)} \frac{\|\mathbf{h}\|_2}{\|\mathbf{h} - \mathbf{z}\|_2} = \sup_{\mathbf{h} \in \mathcal{N} \setminus \{0\}} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma)} \frac{1}{\left\| \frac{\mathbf{h}}{\|\mathbf{h}\|_2} - \frac{\mathbf{z}}{\|\mathbf{h}\|_2} \right\|_2}. \quad (27)$$

Denoting  $\mathcal{B}_2$  the unit ball for the  $\ell_2$  norm, we can rewrite this last expression as:

$$D_* = \sup_{\mathbf{h} \in \mathcal{N} \cap \mathcal{B}_2} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma)} \frac{1}{\|\mathbf{h} - \mathbf{z}\|_2} = \sup_{\mathbf{h} \in \mathcal{N} \cap \mathcal{B}_2} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma) \cap \mathcal{B}_2} \sup_{\lambda \in \mathbb{R}} \frac{1}{\|\mathbf{h} - \lambda \mathbf{z}\|_2}. \quad (28)$$

A simple study gives that if  $\|\mathbf{h}\|_2 = \|\mathbf{z}\|_2 = 1$ , then  $\sup_{\lambda \in \mathbb{R}} \frac{1}{\|\mathbf{h} - \lambda \mathbf{z}\|_2} = \frac{1}{\sqrt{1 - \langle \mathbf{h}, \mathbf{z} \rangle^2}}$ , so that:

$$D_* = \sup_{\mathbf{h} \in \mathcal{N} \cap \mathcal{B}_2} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma) \cap \mathcal{B}_2} \frac{1}{\sqrt{1 - \langle \mathbf{h}, \mathbf{z} \rangle^2}}. \quad (29)$$

The contraposition of Theorem 1 gives the following result : if the NSP (3) is not satisfied for a certain constant  $D$ , then no decoder  $\Delta_\delta$  can satisfy instance optimality (13) with constant  $D$ . In the  $\ell_2/\ell_2$  case, considering  $D < D_*$ ,  $\mathbf{h} \in \mathcal{N} \cap \mathcal{B}_2$  and  $\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma) \cap \mathcal{B}_2$  such that  $\langle \mathbf{h}, \mathbf{z} \rangle^2 \geq 1 - \frac{1}{D^2}$ , we can construct two vectors such that for any decoder, instance optimality with constant  $< \sqrt{D^2 - 1}$  can only be satisfied for at most one of them. This will shed light on the link between NSP and IOP. We have  $\mathbf{z} = \frac{\mathbf{z}_1 - \mathbf{z}_2}{\|\mathbf{z}_1 - \mathbf{z}_2\|_2}$  for some  $\mathbf{z}_1, \mathbf{z}_2 \in \Sigma$ . Let  $\Delta$  be a decoder. If  $\Delta(\mathbf{M}\mathbf{z}_1) \neq \mathbf{z}_1$ , then this vector prevents  $\Delta$  from being instance optimal. The same goes for  $\mathbf{z}_2$  if  $\Delta(\mathbf{M}\mathbf{z}_2) \neq \mathbf{z}_2$ . Now, let's suppose that  $\mathbf{z}_1$  and  $\mathbf{z}_2$  are correctly decoded. In this case,  $(\mathbf{z}_1 + \mathbf{z}_2)/2$  is decoded with a constant worse than  $\sqrt{D^2 - 1}$ , as depicted in Figure 4. Indeed, noting  $\mathbf{p} = (\mathbf{z}_1 + \mathbf{z}_2)/2$  and defining the vectors  $\mathbf{p}_1$  and  $\mathbf{p}_2$  respectively as the orthogonal projections of  $\mathbf{z}_1$  and  $\mathbf{z}_2$  on the affine plane  $\mathbf{p} + \mathcal{N}$ , we must have  $\Delta(\mathbf{M}\mathbf{p}_1) = \Delta(\mathbf{M}\mathbf{p}_2)$ . Denoting as  $p_{\mathcal{N}^\perp}$  the orthogonal projection on  $\mathcal{N}^\perp$ , we have  $d_2(\mathbf{p}_1, \Sigma) \leq d_2(\mathbf{p}_1, \mathbf{z}_1) = \|p_{\mathcal{N}^\perp}(\mathbf{z}_2 - \mathbf{z}_1)\|_2/2$ . Similarly,  $d_2(\mathbf{p}_2, \Sigma) \leq \|p_{\mathcal{N}^\perp}(\mathbf{z}_2 - \mathbf{z}_1)\|_2/2$ . The fact that  $\Delta(\mathbf{M}\mathbf{p}_1) = \Delta(\mathbf{M}\mathbf{p}_2)$  implies that there exists  $i \in \{1, 2\}$  such that  $\|\mathbf{p}_i - \Delta(\mathbf{M}\mathbf{p}_i)\|_2 \geq \|\mathbf{p}_1 - \mathbf{p}_2\|_2/2 = \|p_{\mathcal{N}}(\mathbf{z}_1 - \mathbf{z}_2)\|_2/2$ . Therefore,  $\frac{\|\mathbf{p}_i - \Delta(\mathbf{M}\mathbf{p}_i)\|_2}{d_2(\mathbf{p}_i, \Sigma)} \geq \frac{\|p_{\mathcal{N}}(\mathbf{z}_2 - \mathbf{z}_1)\|_2}{\|p_{\mathcal{N}^\perp}(\mathbf{z}_2 - \mathbf{z}_1)\|_2} \geq D \sqrt{1 - \frac{1}{D^2}} = \sqrt{D^2 - 1}$ . This illustrates the closeness between NSP and IOP: a vector of  $\mathbb{R}(\Sigma - \Sigma)$  which is correlated with  $\mathcal{N}$  can be used to define a couple of vectors such that for any decoder, one of the vectors will not be well decoded.

### 3.3 $\ell_2/\ell_2$ IO with dimensionality reduction

#### 3.3.1 Main theorem

Let's now exploit the expression of  $D_*$  to state the main result of this section: if  $\mathbb{R}(\Sigma - \Sigma)$  contains an orthonormal basis of the finite-dimensional subspace  $V \subset E$  (or even a family of vectors that



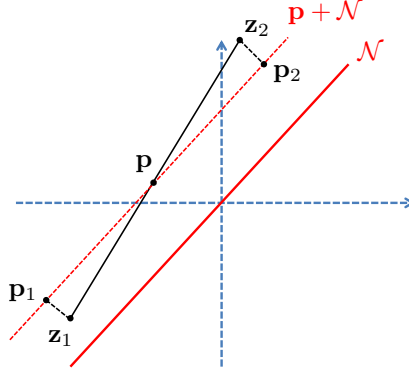


Figure 4: Illustration of the impact of the correlation between  $\mathcal{N}$  and  $\Sigma - \Sigma$  on instance optimality. Here,  $\mathbf{z}_1$  and  $\mathbf{z}_2$  are two vectors in  $\Sigma$  such that  $\mathbf{z}_1 - \mathbf{z}_2$  is well correlated with  $\mathcal{N}$ , implying that at least one of the two vectors  $\mathbf{p}_1$  and  $\mathbf{p}_2$ , which are close to  $\Sigma$  but far from one another, will not be well decoded.

is sufficiently correlated with every vector of  $V$ ), then one cannot expect to get a  $\ell_2/\ell_2$  instance optimal decoder with a small constant while  $\mathbf{M}$  substantially reduces the dimension of  $V$ . The fact that  $\mathbb{R}(\Sigma - \Sigma)$  contains such a tight frame implies that the dimension of  $\mathcal{N}$  cannot be too big without  $\mathcal{N}$  being strongly correlated with  $\Sigma - \Sigma$ , thus yielding the impossibility of a good instance optimal decoder.

Before showing examples where this theorem applies, let's first state it and prove it.

**Theorem 6.** *Suppose  $V$  is of dimension  $n$  and  $\Sigma - \Sigma$  contains a family  $\mathbf{z}_1, \dots, \mathbf{z}_n$  of unit-norm vectors of  $E$  satisfying  $\forall \mathbf{x} \in V, \sum_{i=1}^n \langle \mathbf{z}_i, \mathbf{x} \rangle^2 \geq K \|\mathbf{x}\|_2^2$ . Then to satisfy the NSP on  $V$ ,  $\mathbf{M}$  must map  $V$  into a space of dimension at least  $\left(1 - \frac{1}{K} \left(1 - \frac{1}{D_*^2}\right)\right) n$ .*

*If the number of measurements  $m$  is fixed, then an  $\ell_2/\ell_2$  IO decoder must have a constant at least  $\frac{1}{\sqrt{1 - K(1 - \frac{m}{n})}}$ .*

In particular, if  $\Sigma - \Sigma$  contains an orthonormal basis of  $V$ , then  $K = 1$  and the minimal number of measures to achieve NSP with constant  $D_*$  is  $n/D_*^2$ . Similarly, if  $m$  is fixed so that  $m \ll n$ , then a  $\ell_2/\ell_2$  instance optimal decoder has constant at least  $\sqrt{\frac{n}{m}}$ .

### 3.3.2 Examples

As discussed in the introduction, there is a wide range of standard models where  $\Sigma - \Sigma$  contains an orthonormal basis, and so where  $\ell_2/\ell_2$  IOP with dimensionality reduction is impossible. We provide here less trivial examples, where  $E = V$  is finite-dimensional.

#### Symmetric definite positive matrices with sparse inverse.

**Lemma 2.** *Consider  $E$  is the space of symmetric  $n$ -dimensional matrices, and  $\Sigma \subset E$  the subset of symmetric positive-definite matrices with sparse inverse and with sparsity constant  $k \geq n + 2$  (note*

that  $k \geq n$  is necessary for the matrix to be invertible). The set  $\Sigma - \Sigma$  contains an orthonormal basis of  $E$ .

*Proof.* This orthonormal basis we consider is made of the  $n(n+1)/2$  matrices:  $\mathbf{E}_{i,i}$  and  $\frac{1}{\sqrt{2}}(\mathbf{E}_{i,j} + \mathbf{E}_{j,i})_{i \neq j}$ , where  $\mathbf{E}_{i,j}$  is the matrix where the only nonzero entry is the  $(i, j)$  entry which has value 1.

First, consider  $\mathbf{B}_i = \mathbf{I} + \mathbf{E}_{i,i}$ , where  $\mathbf{I}$  is the identity matrix. Since  $\mathbf{B}_i^{-1} = \mathbf{I} - \frac{1}{2}\mathbf{E}_{i,i}$  is  $n$ -sparse, we have  $\mathbf{B}_i \in \Sigma$ . Since,  $\mathbf{I} \in \Sigma$ , we have  $\mathbf{E}_{i,i} = \mathbf{B}_i - \mathbf{I} \in \Sigma - \Sigma$ .

Now, consider the matrix  $\mathbf{C}_{i,j} = 2\mathbf{I} + \mathbf{E}_{i,j} + \mathbf{E}_{j,i}$ . This matrix is symmetric and for  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ , we have  $\mathbf{x}^T \mathbf{C}_{i,j} \mathbf{x} = 2(\|\mathbf{x}\|_2^2 - x_i x_j) \geq 0$ , so that  $\mathbf{C}_{i,j}$  is semi-definite positive. We can remark that  $\mathbf{C}_{i,j}$  is invertible and that its inverse is  $\frac{1}{2}\mathbf{I} + \frac{1}{6}(\mathbf{E}_{i,i} + \mathbf{E}_{j,j}) - \frac{1}{3}(\mathbf{E}_{i,j} + \mathbf{E}_{j,i})$ , which is  $n+2$ -sparse. The fact that  $\mathbf{C}_{i,j}$  is invertible implies that it is definite, so that  $\mathbf{C}_{i,j} \in \Sigma$ . Therefore, we can write  $\mathbf{E}_{i,i} + \mathbf{E}_{j,j} = \mathbf{C}_{i,j} - 2\mathbf{I} \in \Sigma - \Sigma$ . Since  $\Sigma$  is a positive cone, multiplying this equality by  $\frac{1}{\sqrt{2}}$  yields the desired result.  $\square$

**Low-rank and nonsparse matrices.** In [12], the authors consider a matrix decomposition of the form  $\mathbf{L} + \mathbf{S}$ , where  $\mathbf{L}$  is low-rank and  $\mathbf{S}$  is sparse. In order to give meaning to this decomposition, one must avoid  $\mathbf{L}$  to be sparse. To this end, a ‘‘nonsparsity model’’ for low-rank matrices was introduced.

Let  $E$  be the space of complex matrices of size  $n_1 \times n_2$ . Given  $\mu \geq 1$  and  $r \leq \min(n_1, n_2)$ , let  $\Sigma_{\mu,r}$  be the set of matrices of  $E$  of rank  $\leq r$  satisfying the two following conditions (denoting the SVD of such a matrix by  $\sum_{k=1}^r \sigma_k \mathbf{u}_k \mathbf{v}_k^*$ , where  $\sigma_k > 0$  and the  $\mathbf{u}_k$  and  $\mathbf{v}_k$  are unit-norm vectors) :

1.  $\forall k, \|\mathbf{u}_k\|_\infty \leq \sqrt{\frac{\mu r}{n_1}}$  and  $\|\mathbf{v}_k\|_\infty \leq \sqrt{\frac{\mu r}{n_2}}$ .
2. Denoting  $\mathbf{U}$  and  $\mathbf{V}$  the matrices obtained by concatenating the vectors  $\mathbf{u}_k$  and  $\mathbf{v}_k$ ,  $\|\mathbf{U}\mathbf{V}^*\|_\infty \leq \sqrt{\frac{\mu r}{n_1 n_2}}$ .

These two conditions aim at ‘‘homogenizing’’ the entries of  $\mathbf{U}$  and  $\mathbf{V}$ . Note that we necessarily have  $\mu \geq 1$ .

**Lemma 3.** *Let  $E = \mathcal{M}_{n_1, n_2}(\mathbb{C})$  and  $\Sigma_{\mu,r}$  be the subset of  $E$  containing the matrices satisfying the two above conditions (with  $\mu \geq 1$  and  $r \geq 1$ ). Then  $\Sigma_{\mu,r} - \Sigma_{\mu,r}$  contains an orthonormal basis.*

*Proof.* Since  $\Sigma_{\mu,r}$  contains the null matrix, it is sufficient to prove that  $\Sigma_{\mu,r}$  contains an orthonormal basis. Let  $\{\mathbf{e}_k\}_{k=1}^{n_1}$  and  $\{\mathbf{f}_\ell\}_{\ell=1}^{n_2}$  be the discrete Fourier bases of  $\mathbb{C}^{n_1}$  and  $\mathbb{C}^{n_2}$ , that is

$$\mathbf{e}_k = \frac{1}{\sqrt{n_1}} [1, \exp(2i\pi k/n_1), \dots, \exp(2i\pi(n_1-1)k/n_1)]^T$$

$$\text{and } \mathbf{f}_\ell = \frac{1}{\sqrt{n_2}} [1, \exp(2i\pi \ell/n_2), \dots, \exp(2i\pi(n_2-1)\ell/n_2)]^T.$$

Then the  $n_1 n_2$  rank-1 matrices of the form  $\mathbf{e}_k \mathbf{f}_\ell^*$  are elements of  $\Sigma_{\mu,r}$  since they obviously satisfy the two above conditions. But they also form an orthonormal basis of  $E$ , since each entry of  $\mathbf{e}_k \mathbf{f}_\ell^*$  is of module  $\frac{1}{\sqrt{n_1 n_2}}$  and that, denoting  $\langle \cdot, \cdot \rangle$  the Hermitian scalar product on  $E$ ,

$$\langle \mathbf{e}_k \mathbf{f}_\ell^*, \mathbf{e}_{k'} \mathbf{f}_{\ell'}^* \rangle = \sum_{u=0}^{n_1-1} \exp\left(2i\pi u \frac{k-k'}{n_1}\right) \sum_{v=0}^{n_2-1} \exp\left(2i\pi v \frac{\ell-\ell'}{n_2}\right) = \delta_k^{k'} \delta_\ell^{\ell'}, \quad (30)$$

proving that these matrices form an orthonormal basis of  $E$ .  $\square$

## 4 The NSP and its relationship with the RIP

In the following section, we consider  $E = G$  (and  $\mathbf{A} = \mathbf{I}$ ). As we have seen, one cannot expect to get  $\ell_2/\ell_2$  instance optimality in a dimensionality reduction context. This raises the following question: given pseudo-norms  $\|\cdot\|_G$  and  $\|\cdot\|_F$  defined respectively on  $E$  and  $F$ , is there a pseudo-norm  $\|\cdot\|_E$  such that IOP holds? We will see that this property is closely related to the RIP on  $\mathbf{M}$ .

### 4.1 Generalized RIP and its necessity for robustness

The Restricted Isometry Property is a widely-used property on the operator  $\mathbf{M}$  which yields nice stability and robustness results on the recovery of vectors from their compressive measurements. In the usual CS framework, the RIP provides a relation of the form  $(1 - \delta)\|\mathbf{x}\|_G \leq \|\mathbf{M}\mathbf{x}\|_F \leq (1 + \delta)\|\mathbf{x}\|_G$  for any vector  $\mathbf{x}$  in  $\Sigma_{2k}$ . The norms  $\|\cdot\|_G$  and  $\|\cdot\|_F$  are usually both taken as the  $\ell_2$ -norm. A form of RIP can easily be stated in a generalized framework: we will say that  $\mathbf{M}$  satisfies the RIP on  $\Sigma - \Sigma$  if there exists positive constants  $\alpha, \beta$  such that

$$\forall \mathbf{z} \in \Sigma - \Sigma, \alpha\|\mathbf{z}\|_G \leq \|\mathbf{M}\mathbf{z}\|_F \leq \beta\|\mathbf{z}\|_G. \quad (31)$$

Similarly to the sparse case, it is possible to make a distinction between *lower-RIP* (left inequality) and *upper-RIP* (right inequality). Let's remark that this definition has been stated for vectors of  $\Sigma - \Sigma$ : this choice is justified by the links between this formulation and the NSP, which will be discussed later in this section. Let's also note that this form of RIP encompasses several generalized RIP previously proposed: the  $\Omega$ -RIP [28], the  $\mathbf{D}$ -RIP [11] and the Union of Subspaces RIP [7].

Let's now suppose the existence of decoders robust to noise, that is for all  $\delta > 0$ , (19) is satisfied for a certain  $\Delta_\delta$ . This property implies the Robust NSP (20) with the same constants according to Theorem 3. By considering  $\mathbf{h} \in \Sigma - \Sigma$ , the Robust NSP reads:

$$\forall \mathbf{h} \in \Sigma - \Sigma, \|\mathbf{h}\|_G \leq D_2\|\mathbf{M}\mathbf{h}\|_F. \quad (32)$$

This is the lower-RIP on  $\Sigma - \Sigma$ , with constant  $1/D_2$ . The stability to noise therefore implies the lower-RIP on the set of differences of vectors of  $\Sigma$ , which is therefore necessary if one seeks the existence of a decoder robust to noise.

### 4.2 $M$ -norm instance optimality with the RIP

The lower-RIP is necessary for the existence of a Robust instance optimal decoder, but what can we say this time if we suppose that  $\mathbf{M}$  satisfies the lower-RIP on  $\Sigma - \Sigma$  with constant  $\alpha$ , that is  $\forall \mathbf{z} \in \Sigma - \Sigma, \alpha\|\mathbf{z}\|_G \leq \|\mathbf{M}\mathbf{z}\|_F$ ? We will prove that in both the noiseless and the noisy cases, this implies the IOP with norms  $\|\cdot\|_G$  and  $\|\cdot\|_M$ , the latter being called " $M$ -norm"<sup>3</sup> and involving  $\|\cdot\|_G$  and  $\|\cdot\|_F$ .

Let's define the  $M$ -norm on  $E$  as the following quantity, extending its definition for  $\ell_2$  norms in [36] and its implicit appearance in the proof of early results of the field [10]:

$$\forall \mathbf{x} \in E, \|\mathbf{x}\|_M = \|\mathbf{x}\|_G + \frac{1}{\alpha}\|\mathbf{M}\mathbf{x}\|_F. \quad (33)$$

Note that the term  $M$ -norm should be understood as  $M$ -pseudo-norm in the general case: if  $\|\cdot\|_F$  and  $\|\cdot\|_G$  satisfy the properties listed in Table 1, then  $\|\cdot\|_M$  satisfies the same properties as

<sup>3</sup>to highlight its dependency on the Measurement operator

$\|\cdot\|_G$ . However, when  $\|\cdot\|_G$  and  $\|\cdot\|_F$  are norms,  $\|\cdot\|_M$  is also a norm. We will note  $d_M(\cdot, \cdot)$  its associated (pseudo-)distance. The following theorem states that this  $\|\cdot\|_M$  allows one to derive an NSP from the lower-RIP on  $\Sigma - \Sigma$ .

**Theorem 7.** *Let's suppose that  $\mathbf{M}$  satisfies the lower-RIP on  $\Sigma - \Sigma$  with constant  $\alpha$  (left inequality of (31)). Then the following Robust NSP is satisfied:*

$$\forall \mathbf{h} \in E, \|\mathbf{h}\|_G \leq d_M(\mathbf{h}, \Sigma - \Sigma) + \frac{1}{\alpha} \|\mathbf{M}\mathbf{h}\|_F. \quad (34)$$

In particular, the following regular NSP is satisfied:

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_G \leq d_M(\mathbf{h}, \Sigma - \Sigma). \quad (35)$$

Therefore, if  $\mathbf{M}$  satisfies the lower-RIP on  $\Sigma - \Sigma$  with constant  $\alpha$ , then for all  $\delta > 0$ , there exists a noise-robust instance optimal decoder  $\Delta_\delta$  satisfying the following property (Theorem 4):

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq 2d_M(\mathbf{x}, \Sigma) + \frac{2}{\alpha} \|\mathbf{e}\|_F + \delta. \quad (36)$$

Note that in [7], the author explored the implication of a lower-RIP on  $\Sigma - \Sigma$  for the case where  $\Sigma$  is an arbitrary UoS and  $\|\cdot\|_G/\|\cdot\|_F$  are the  $\ell_2$  norm. He proved that this generalized lower-RIP implies the following IOP: for all  $\delta > 0$ , there exists a decoder  $\Delta_\delta$  such that

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \forall \mathbf{z} \in \Sigma, \|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_2 \leq \|\mathbf{x} - \mathbf{z}\|_2 + \frac{2}{\alpha} \|\mathbf{M}(\mathbf{x} - \mathbf{z}) + \mathbf{e}\|_2 + \delta. \quad (37)$$

In this set-up, the instance optimality in equation (36) can be reformulated as

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \forall \mathbf{z} \in \Sigma, \|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_2 \leq 2\|\mathbf{x} - \mathbf{z}\|_2 + \frac{2}{\alpha} \|\mathbf{M}(\mathbf{x} - \mathbf{z})\|_2 + \frac{2}{\alpha} \|\mathbf{e}\|_2 + \delta. \quad (38)$$

Comparing these two instance optimality results, we can remark that the one in [7] is slightly tighter. This is merely a consequence of the difference in our method of proof, as we add the NSP as an intermediate result to prove instance optimality. The upper bound in [7] can also be derived in our case with the same proof layout if we suppose the lower-RIP. Compared to [7], our theory deals with general (pseudo-)norms and sets  $\Sigma$  beyond Union of Subspaces.

### 4.3 Upper-bound on the $M$ -norm by an atomic norm

As we have seen, provided a *lower-RIP* on  $\Sigma - \Sigma$ , an NSP can be derived with the  $M$ -norm as  $\|\cdot\|_E$ . However, this may look like a tautology since the  $M$ -norm explicitly depends on  $\mathbf{M}$ . Hence, one may wonder if this NSP is of any use. We will prove in the following that provided an *upper-RIP* on a certain cone  $\Sigma'$  (which can be taken as  $\mathbb{R}\Sigma$ ), a more natural upper bound can be derived by bounding the  $M$ -norm with an atomic norm [18]. In particular, this type of inequality applied to the usual  $k$ -sparse vectors and low-rank matrices models give, under standard RIP conditions, instance optimality upper bounds with typical norms.

We will suppose in this section that  $\|\cdot\|_G$  is a norm.

### 4.3.1 The atomic norm $\|\cdot\|_{\Sigma'}$

Let  $\Sigma'$  be a subset of  $E$  and let  $E'$  be the closure of  $\text{span}(\Sigma')$  with respect to the norm  $\|\cdot\|_G$ . For  $\mathbf{x} \in E'$ , one can define the “norm”  $\|\mathbf{x}\|_{\Sigma'}$  by:

$$\|\mathbf{x}\|_{\Sigma'} := \inf \left\{ \sum_{k=0}^{+\infty} \|\mathbf{x}_k\|_G : \forall k, \mathbf{x}_k \in \mathbb{R}\Sigma' \text{ and } \|\mathbf{x} - \sum_{k=0}^K \mathbf{x}_k\|_G \xrightarrow{K \rightarrow +\infty} 0 \right\}. \quad (39)$$

Remark that there may be some vectors  $\mathbf{x}$  for which  $\|\mathbf{x}\|_{\Sigma'} = +\infty$ , if  $\sum_{k=0}^{+\infty} \|\mathbf{x}_k\|_G = +\infty$  for any decomposition of  $\mathbf{x}$  as an infinite sum of elements of  $\mathbb{R}\Sigma'$ . However, the set  $V = \{\mathbf{x} \in E \mid \|\mathbf{x}\|_{\Sigma'} < +\infty\}$  is a normed subspace of  $E$  which contains  $\Sigma'$  [21]. In the following, we assume that  $V = E$ . Note that this norm can be linked to atomic norms defined in [18] by considering  $\mathcal{A}$  as the set of normalized elements of  $\Sigma'$  with respect to  $\|\cdot\|_G$ .

Now suppose  $\mathbf{M}$  satisfies an upper-RIP on  $\Sigma'$ , so that

$$\forall \mathbf{x}' \in \Sigma', \|\mathbf{M}\mathbf{x}'\|_F \leq \beta \|\mathbf{x}'\|_G. \quad (40)$$

For  $\mathbf{x} \in E$  admitting a decomposition  $\sum_{k=0}^{+\infty} \mathbf{x}_k$  on  $\mathbb{R}\Sigma'$ , we can therefore upper bound  $\|\mathbf{M}\mathbf{x}\|_F$  by  $\sum_{k=0}^{+\infty} \|\mathbf{M}\mathbf{x}_k\|_F \leq \beta \sum_{k=0}^{+\infty} \|\mathbf{x}_k\|_G$ . This inequality is valid for any decomposition of  $\mathbf{x}$  as a sum of elements of  $\mathbb{R}\Sigma'$ , so that  $\|\mathbf{M}\mathbf{x}\|_F \leq \beta \|\mathbf{x}\|_{\Sigma'}$ . Therefore, under these hypotheses,

$$\forall \mathbf{x} \in E, \|\mathbf{x}\|_M \leq \|\mathbf{x}\|_G + \frac{\beta}{\alpha} \|\mathbf{x}\|_{\Sigma'} \leq \left(1 + \frac{\beta}{\alpha}\right) \|\mathbf{x}\|_{\Sigma'}. \quad (41)$$

In particular, we have the following result:

**Theorem 8.** *Suppose  $\mathbf{M}$  satisfies the lower-RIP on  $\Sigma - \Sigma$  with constant  $\alpha$  and the upper-RIP on  $\Sigma$  with constant  $\beta$ , that is*

$$\forall \mathbf{x} \in \Sigma - \Sigma, \alpha \|\mathbf{x}\|_G \leq \|\mathbf{M}\mathbf{x}\|_F \quad (42)$$

and

$$\forall \mathbf{x} \in \Sigma, \|\mathbf{M}\mathbf{x}\|_F \leq \beta \|\mathbf{x}\|_G. \quad (43)$$

Then for all  $\delta > 0$ , there exists a decoder  $\Delta_\delta$  satisfying

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq 2 \left(1 + \frac{\beta}{\alpha}\right) d_\Sigma(\mathbf{x}, \Sigma) + \frac{2}{\alpha} \|\mathbf{e}\|_E + \delta, \quad (44)$$

where  $d_\Sigma$  is the distance associated to the norm  $\|\cdot\|_\Sigma$ .

**Remark 3.** *Note that these results can be extended with relative ease to the case where  $\|\cdot\|_G$  is not necessarily homogeneous but  $p$ -homogeneous, that is  $\|\lambda\mathbf{x}\|_G = |\lambda|^p \|\mathbf{x}\|_G$ .*

### 4.3.2 Study of $\|\cdot\|_\Sigma$ in two usual cases

We now provide a more thorough analysis of the norm  $\|\cdot\|_\Sigma$  for usual models which are sparse vectors and low-rank matrices. In particular, we give a simple equivalent of this norm involving usual norms in the case where  $\|\cdot\|_G = \|\cdot\|_2$  (for matrices, this is the Frobenius norm).

The norm  $\|\cdot\|_\Sigma$  relies on the decomposition of a vector as a sum of elements of  $\mathbb{R}\Sigma$ . When  $\Sigma$  is the set of  $k$ -sparse vectors or the set of matrices of rank  $k$ , there are particular decompositions of this type:

- In the case where  $\Sigma$  is the set of  $k$ -sparse vectors, a vector  $\mathbf{x}$  can be decomposed as  $\sum_{j=1}^{\infty} \mathbf{x}_j$ , where all  $\mathbf{x}_j$  are  $k$ -sparse vectors with disjoint supports, which are eventually zero, and such that any entry of  $\mathbf{x}_j$  does not exceed any entry of  $\mathbf{x}_{j-1}$  (in magnitude). This is a decomposition of  $\mathbf{x}$  into disjoint supports of size  $k$  with a nonincreasing constraint on the coefficients.
- Similarly, in the case where  $\Sigma$  is the set of matrices of rank  $k$  and  $\mathbf{N}$  is a matrix, the SVD of  $\mathbf{N}$  gives a decomposition of the form  $\mathbf{N} = \sum_{j=1}^{\infty} \mathbf{N}_j$ , where the  $\mathbf{N}_j$  are rank  $k$ , eventually zero matrices such that any singular value of  $\mathbf{N}_j$  does not exceed any singular value of  $\mathbf{N}_{j-1}$ .

For  $j \geq 2$ , we can upper bound the quantity  $\|\mathbf{x}_j\|_2$  using the assumption on the particular decomposition:  $\|\mathbf{x}_j\|_2 \leq \sqrt{k}\|\mathbf{x}_j\|_{\infty} \leq \sqrt{k} \frac{\|\mathbf{x}_{j-1}\|_1}{k} = \frac{\|\mathbf{x}_{j-1}\|_1}{\sqrt{k}}$ . Similarly,  $\|\mathbf{N}_j\|_2 \leq \frac{\|\mathbf{N}_{j-1}\|_*}{\sqrt{k}}$ , where  $\|\cdot\|_*$  is the trace norm, defined as the sum of singular values. We can therefore, in both cases, upper bound the norm  $\|\cdot\|_{\Sigma}$ . In the case of  $k$ -sparse vectors, this gives:

$$\|\mathbf{x}\|_{\Sigma} \leq \|\mathbf{x}_1\|_2 + \sum_{j \geq 1} \frac{\|\mathbf{x}_j\|_1}{\sqrt{k}} \leq \|\mathbf{x}\|_2 + \frac{\|\mathbf{x}\|_1}{\sqrt{k}}. \quad (45)$$

In the case of matrices of rank  $k$ , this gives:

$$\|\mathbf{N}\|_{\Sigma} \leq \|\mathbf{N}_1\|_2 + \sum_{j \geq 1} \frac{\|\mathbf{N}_j\|_1}{\sqrt{k}} \leq \|\mathbf{N}\|_F + \frac{\|\mathbf{N}\|_*}{\sqrt{k}}. \quad (46)$$

We can also upper bound the right hand side of these equations by  $\mathcal{O}(\|\cdot\|_{\Sigma})$  with a small constant, which will prove that the norms defined in these equations are of the same order. Indeed, a simple application of the triangle inequality gives us first that  $\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_{\Sigma}$  and  $\|\mathbf{N}\|_F \leq \|\mathbf{N}\|_{\Sigma}$ . Then, considering a decomposition of  $\mathbf{x}$  as a sum of  $k$ -sparse vectors  $\sum_{j \geq 1} \mathbf{x}_j$ , we get

$$\frac{\|\mathbf{x}\|_1}{\sqrt{k}} \leq \sum_{j \geq 1} \frac{\|\mathbf{x}_j\|_1}{\sqrt{k}} \leq \sum_{j \geq 1} \|\mathbf{x}_j\|_2 \quad (47)$$

(indeed, each  $\mathbf{x}_j$  can be viewed as a  $k$ -dimensional vector and we have for such a vector  $\|\mathbf{x}_j\|_1 \leq \sqrt{k}\|\mathbf{x}_j\|_2$ ). Similarly,

$$\frac{\|\mathbf{N}\|_*}{\sqrt{k}} \leq \sum_{j \geq 1} \|\mathbf{N}_j\|_F. \quad (48)$$

Since these upper bounds are satisfied for any decomposition, they can be replaced respectively by  $\|\mathbf{x}\|_{\Sigma}$  and  $\|\mathbf{N}\|_{\Sigma}$ . Finally, we have

$$\|\mathbf{x}\|_2 + \frac{\|\mathbf{x}\|_1}{\sqrt{k}} \leq 2\|\mathbf{x}\|_{\Sigma} \quad \text{and} \quad \|\mathbf{N}\|_F + \frac{\|\mathbf{N}\|_*}{\sqrt{k}} \leq 2\|\mathbf{N}\|_{\Sigma}. \quad (49)$$

We have thus shown:

**Lemma 4.** *When  $\Sigma$  is the set of  $k$ -sparse vectors, the norm  $\|\cdot\|_{\Sigma}$  satisfies*

$$\|\cdot\|_{\Sigma} \leq \|\cdot\|_2 + \frac{\|\cdot\|_1}{\sqrt{k}} \leq 2\|\cdot\|_{\Sigma}. \quad (50)$$

*When  $\Sigma$  is the set of rank- $k$  matrices, it satisfies*

$$\|\cdot\|_{\Sigma} \leq \|\cdot\|_F + \frac{\|\cdot\|_*}{\sqrt{k}} \leq 2\|\cdot\|_{\Sigma}. \quad (51)$$

We can thus remark that for these two standard models, the norm  $\|\cdot\|_{\Sigma}$  can easily be upper bounded by usual norms under RIP conditions, yielding an IOP with a usual upper bound. We can also note that stronger RIP conditions can yield a stronger result: in [10], the author proves that under upper and lower-RIP on  $\Sigma - \Sigma$  with  $\Sigma$  being the set of  $k$ -sparse vectors, an instance optimal decoder can be defined as the minimization of a convex objective: the  $\ell_1$  norm, which appears as strongly connected to the norm  $\|\cdot\|_{\Sigma}$ . One may then wonder if a generalization of such a result is possible: when can an instance optimal decoder be obtained by solving a convex minimization problem with a norm related to  $\|\cdot\|_{\Sigma}$  ?

## 5 Discussion and outlooks on Instance Optimality

Let's now summarize the results and give some insights on interesting future work. As has been detailed throughout the paper, Instance Optimality is a property presenting several benefits:

- It can be defined in a very general framework, for any signal space, signal model and pseudo-norms, as well as for both noiseless and noisy settings.
- It is a nice uniform formulation of the “good behavior” of a decoder and thus of the well-posedness of an inverse problem.
- It can be linked to Null Space Property and Restricted Isometry Property, which provide necessary and/or sufficient conditions for the existence of an Instance Optimal decoder.

We now present some immediate outlooks and interesting open questions related to instance optimality and to the results presented in this paper.

**Condition for the well-posedness of the “optimal” decoder.** We have seen that for general models  $\Sigma$ , an additional term  $\delta$  appears in the right hand side term of the instance optimality inequality ((13),(19)), reflecting the fact that the minimal distance of the “optimal” decoder (59) may not be reached at a specific point. However, as mentioned in Property 1, this additive constant can be dropped in the noiseless case provided  $\Sigma + \mathcal{N}$  is a closed set. One can then wonder if there exists a similar condition (e.g., a sort of local compactness property) in the noisy case for which one can drop the constant  $\delta$  and get a more usual instance optimality result.

**Compressed graphical models.** As has been mentioned in Section 1.3.4, the case where  $\Sigma$  is the set of symmetric definite positive square matrices with sparse inverse is related to high-dimensional Gaussian graphical models. In Lemma 2, we showed this type of models fits in our theory since we could apply Theorem 6 in this case, proving the impossibility of  $\ell_2/\ell_2$  IOP in a dimension-reduction case. Yet, as for other signal models, can Gaussian graphical models satisfy some IOP/NSP with different norms in a compressive framework?

**Guarantees for signal-space reconstructions and more.** When  $\mathbf{D}$  is a redundant dictionary of size  $d \times n$  and the signals of interest are vectors of the form  $\mathbf{z} = \mathbf{D}\mathbf{x}$ , where  $\mathbf{x}$  is a sparse vector, traditional reconstruction guarantees from  $\mathbf{y} = \mathbf{M}\mathbf{z}$  assume the RIP on the matrix  $\mathbf{M}\mathbf{D}$ . This is often too restrictive: for example when  $\mathbf{D}$  has strongly correlated columns, failure to identify  $\mathbf{x}$  from  $\mathbf{y}$  does not necessarily prevent one from correctly estimating  $\mathbf{z}$ . Recent work on *signal-space* algorithms [20] has shown that the  $\mathbf{D}$ -RIP assumption on  $\mathbf{M}$  is in fact sufficient.

The framework presented in this paper offers two ways to approach this setting:

- Considering  $\Sigma = \Sigma_k$  as the set of  $k$ -sparse vectors of dimension  $n$  and  $\mathbf{A} = \mathbf{D}$ , the upper bound on the reconstruction error is of the form  $d_E(\mathbf{x}, \Sigma_k)$ . Signal-space guarantees can be envisioned by choosing a metric  $\|\cdot\|_E = \|\mathbf{D}\cdot\|$ .
- Considering  $\Sigma = \mathbf{D}\Sigma_k$  as the set of  $d$ -dimensional vectors that have a  $k$ -sparse representation in the dictionary  $\mathbf{D}$  and  $\mathbf{A} = \mathbf{I}$ , the upper bound is of the form  $d'(\mathbf{z}, \mathbf{D}\Sigma_k)$ .

In [33], the authors propose a result similar to instance optimality by upper bounding, for a Total Variation decoder, the reconstruction error of an image  $\mathbf{X}$  from compressive measurement by a quantity involving  $d_1(\nabla\mathbf{X}, \Sigma_k)$ , where  $\nabla$  is the gradient operator,  $\Sigma_k$  the  $k$ -sparse union of subspaces (in the gradient space) and  $d_1$  is the  $\ell_1$  distance. This quantity is therefore the distance between the gradient of the image and the  $k$ -sparse vectors model. Can such a bound be interpreted in our framework, and possibly be generalized to other types of signals?

**Task-oriented decoders versus general purpose decoders.** We already mentioned two very different application set-ups, in medical imaging and audio source separation, where only parts of the original signals need to be recovered. One can think of other, more dramatic, cases where only task-oriented linear features should be reconstructed. One such situation is met in current image classification work-flows. Indeed, most recent state-of-art image classification methods rely on very high-dimensional image representation (e.g., so called Fisher vectors, of dimension ranging from 10,000 to 200,000) and conduct supervised learning on such labeled signals by means of linear SVMs [39]. Not only this approach yields top-ranking performance in terms of classification accuracy on challenging image classification benchmarks, but it also permits very large scale learning thanks to the low complexity of linear SVM training and its efficient implementations, e.g., with stochastic gradient descent. For each visual category to recognize, a linear classifier  $\mathbf{w}$  is learned, which associates to an input image with representation  $\mathbf{x}$  the score  $\mathbf{w}^T\mathbf{x}$ . The single or multiple labels that are finally assigned to  $\mathbf{x}$  by the system depend on the scores provided by all trained classifiers (typically from 10 to 100), hence on a vector of the form  $\mathbf{A}\mathbf{x}$ , where each row of  $\mathbf{A}$  is one one-vs-all linear SVM. In this set-up, the operator  $\mathbf{A}$  implies a dramatic dimension reduction. For very large scale problems of this type, storing and manipulating original image signatures in the database can become intractable. The theoretical framework proposed in this paper might help designing new solutions to this problem in the future. In particular, it will provide tools to answer the following questions:

- $\mathbf{A}$  being given (learned on a labeled subset of the database): can one design a compressive measurement operator  $\mathbf{M}$  such that the “classifiers” scores can be recovered directly from the compressed image signature  $\mathbf{M}\mathbf{x}$ , hence avoiding the prior reconstruction of the high-dimensional signal  $\mathbf{x}$ ?
- $\mathbf{M}$  being given (“legacy” compressed storing of image signatures): what are the linear classifier collections that can be precisely emulated in the compressed domain thanks to a good decoder  $\Delta$ ?

Note that this classification-oriented set-up might call for a specific norm  $\|\cdot\|_G$  on the output of a linear score bank.

Another important domain of application that might benefit from both aspects (general purpose and task-oriented) of our work is data analysis under privacy constraints. Two scenarii can be



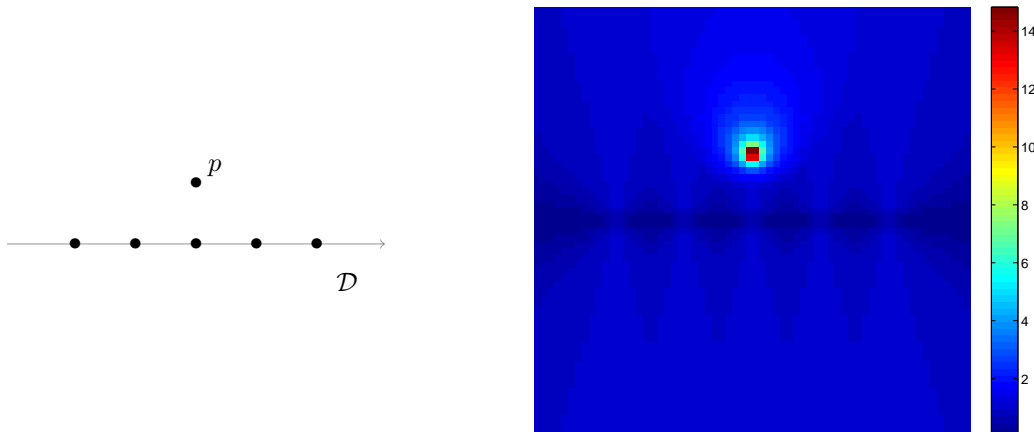


Figure 5: Drawback of uniform instance optimality in a simple case: the model  $\Sigma$  (*Left*) is the set of black points including those on  $D$  and the point  $p$  and the operator  $\mathbf{M}$  is the 1-dimensional orthogonal projection on the horizontal axis. If we choose  $\Delta$  as the pseudo-inverse of  $\mathbf{M}$ , the depicted IO ratio (*Right*) is low on most of the space, but the uniform constant is infinite.

envisioned, where our framework could help decide whether or not such constraints are compatible with the analysis of interest:

- *General purpose scenario*: given a linear measurement operator  $\mathbf{M}$  of interest for further analysis, can one guarantee that there is no decoder permitting good enough recovery of original signals?
- *Task-oriented scenario*: the operator  $\mathbf{M}$  serving as a means to obfuscate original signals such that critical information can't be recovered, let's consider a specific analysis task on original signals requiring the application of the feature extractor  $\mathbf{A}$ . Can this task be implemented on obfuscated signals instead, via a good decoder  $\Delta$ , hence in a privacy-preserving fashion?

**Worst case versus average case instance optimality.** The raw concept of Instance Optimality has a major drawback: the uniformity of the bound may impose, in some settings, a large global instance optimality constant whereas the inverse problem is well posed for the vast majority of signals. Let's consider the example depicted in Figure 5, where the signal space  $E$  is of dimension 2, the signal model  $\Sigma$  is a point cloud mostly concentrated along the line  $D$  and the measurement operator  $\mathbf{M}$  is the orthogonal projection on  $D$ . The figure depicts the ratio (approximation error)/(distance to model) for each  $\mathbf{x} \in \mathbb{R}^2$ . The optimal constant, which is the supremum of these ratios, is infinite: the ratio actually goes to infinity in the vicinity of the point  $p$ . However, for the vast majority of vectors, the ratio is rather low (the blue section covers most of the space).

An interesting outlook to circumvent this pessimistic “worst-case” phenomenon is to consider a probabilistic formulation of instance optimality, as in [19]: given  $\Omega$  a probability space with probability measure  $P$ , and considering  $\mathbf{M}$  as a random variable on  $\Omega$ , is there a decoder  $\Delta(\cdot|\mathbf{M})$  (which computes an estimate given the observation *and* the particular draw of the measurement

operator  $\mathbf{M}$ ) such that for any  $\mathbf{x} \in E$ , the instance optimality inequality

$$\|\mathbf{x} - \Delta(\mathbf{M}\mathbf{x}|\mathbf{M})\|_G \leq Cd_E(\mathbf{x}, \Sigma) \quad (52)$$

holds with high probability on the drawing of  $\mathbf{M}$ ? A particular challenge would be to understand in which dimension reduction scenarii there exists both a probability measure and a decoder with the above property. Another possible formulation of probabilistic instance optimality is to define a probability distribution on the signal space and to upper bound the average reconstruction error of the vectors, as in [40].

## References

- [1] D. Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281, 2001.
- [2] B. Adcock and A. C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. *DAMTP Tech. Rep.*, 2011.
- [3] B. Adcock and A. C. Hansen. A generalized sampling theorem for stable reconstructions in arbitrary bases. *J. Fourier Anal. Appl.*, 18(4):685–716, November 2012.
- [4] B. Adcock, A. C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: asymptotic incoherence and asymptotic sparsity in compressed sensing. *arXiv*, pages 1–44, February 2013.
- [5] R. G. Baraniuk, V. Cevher, and M. B. Wakin. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proceedings of the IEEE*, 98(6):959–971, 2010.
- [6] R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. In *Foundations of Computational Mathematics*, pages 941–944, 2006.
- [7] T. Blumensath. Sampling and reconstructing signals from a union of linear subspaces. *IEEE Transactions on Information Theory*, 57(7):4660–4671, 2011.
- [8] T. Blumensath and M. E. Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Transactions on Information Theory*, 55(4):1872–1882, 2009.
- [9] A. Bourrier, R. Gribonval, and P. Pérez. Compressive Gaussian Mixture Estimation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2013.
- [10] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. Paris S'ér. I Math.*, 346:589–592, 2008.
- [11] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Applied and Computational Harmonic Analysis*, 31(1):59–73, 2011.
- [12] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

- [13] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235 – 7254, November 2011.
- [14] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theor.*, 57(4):2342–2359, April 2011.
- [15] E. J. Candès, J. K. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.
- [16] E. J. Candès, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *CoRR*, abs/1109.4499, 2011.
- [17] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.
- [18] V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.
- [19] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, pages 211–231, 2009.
- [20] M. A. Davenport, D. Needell, and M. B. Wakin. Signal space cosamp for sparse recovery with redundant dictionaries. *IEEE Transactions on Information Theory*, 59(10):6820–6829, 2013.
- [21] Ronald A DeVore and V N Temlyakov. Some remarks on greedy algorithms. *Adv. Comp. Math.*, 5(2-3):173–187, 1996.
- [22] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52:1289–1306, 2006.
- [23] A. Eftekhari and M. B. Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *CoRR*, abs/1306.4748, 2013.
- [24] Y. C. Eldar, P. Kuppinger, and H. Bölcskei. Block-sparse signals: uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010.
- [25] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Springer, 1996.
- [26] S. Foucart. Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J. Numerical Analysis*, 49:2543–2563, 2011.
- [27] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Springer, 2013.
- [28] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies. Greedy-Like Algorithms for the Cospase Analysis Model. partially funded by the ERC, PLEASE project, ERC-2011-StG-277906, January 2013.
- [29] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Appl. Comp. Harm. Anal.*, 22(3):335–355, 2007.

- [30] A. C. Hansen and B. Adcock. Generalized sampling and infinite dimensional compressed sensing. *Magnetic Resonance Imaging*, 2011.
- [31] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.
- [32] S. Nam, M. E. Davies, M. Elad, and R. Gribonval. The Cospase Analysis Model and Algorithms. *Applied and Computational Harmonic Analysis*, 34(1):30–56, 2013.
- [33] D. Needell and R. Ward. Stable image reconstruction using total variation minimization. *SIAM J. Imaging Sciences*, 6(2):1035–1058, 2013.
- [34] H. Ohlsson, A. Y. Yang, and S. S. Sastry. Compressive phase retrieval from squared output measurements via semidefinite programming. *CoRR*, abs/1111.6323, 2011.
- [35] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *CoRR*, abs/1212.3753, 2012.
- [36] T. Peleg, R. Gribonval, and M.E. Davies. Compressed sensing and best approximation from union of subspaces: Beyond dictionaries. In *EUSIPCO*, 2013.
- [37] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Transactions on Information Theory*, 54(5):2210–2219, 2008.
- [38] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [39] J. Sanchez, F. Perronnin, Th. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *Int. J. Computer Vision*, 105(3):22–245, 2013.
- [40] G. Yu and G. Sapiro. Statistical compressed sensing of gaussian mixture models. *IEEE Transactions on Signal Processing*, 59(12):5842–5858, 2011.
- [41] M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.
- [42] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 2007.
- [43] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma. Stable principal component pursuit. In *ISIT*, pages 1518–1522, 2010.

## A Well-posedness of the finite UoS decoder

In this section, we will prove that if  $\Sigma$  is a finite union of subspaces in  $\mathbb{R}^n$  and  $\|\cdot\|$  a norm on  $\mathbb{R}^n$ , then the quantity  $\arg \min_{\mathbf{z} \in (\mathbf{x} + \mathcal{N})} d(\mathbf{z}, \Sigma)$ , where  $d$  is the distance relative to  $\|\cdot\|$ , is defined for all  $\mathbf{x} \in \mathbb{R}^n$ .

Let’s first prove the following lemma:

**Lemma 5.** *Let  $V$  and  $W$  be two subspaces of  $\mathbb{R}^n$  and  $\|\cdot\|$  a norm on  $\mathbb{R}^n$ . Then  $\forall \mathbf{x} \in \mathbb{R}^n, \exists \mathbf{y} \in (\mathbf{x}+V)$  such that  $d(\mathbf{y}, W) = d(\mathbf{x} + V, W)$ , where  $d$  is the distance derived from  $\|\cdot\|$ .*

*Proof.* Let  $\Phi$  be defined on  $V + W$  by  $\Phi(\mathbf{u}) = \|\mathbf{u} - \mathbf{x}\|$ . Since  $\Phi(\mathbf{u}) \geq \|\mathbf{u}\| - \|\mathbf{x}\|$ , we have  $\lim_{\|\mathbf{u}\| \rightarrow +\infty} \Phi(\mathbf{u}) = +\infty$ , so that  $\exists M > 0$  such that  $\|\mathbf{u}\| > M \Rightarrow \Phi(\mathbf{u}) \geq \|\mathbf{x}\|$ . The set  $B = \{\mathbf{u} \in V + W, \|\mathbf{u}\| \leq M\}$  is a closed ball of  $V + W$  and is thus a compact. Since  $\Phi$  is continuous,  $\Phi$  has a minimizer  $\mathbf{v}$  on  $B$ .  $0 \in B$ , so that  $\Phi(0) = \|\mathbf{x}\| \geq \Phi(\mathbf{v})$ . For all  $\mathbf{u}$  such that  $\|\mathbf{u}\| > M$ , we have  $\Phi(\mathbf{u}) \geq \|\mathbf{x}\| \geq \Phi(\mathbf{v})$ , so that  $\mathbf{v}$  is a global minimizer of  $\Phi$ .

We therefore have  $\forall (\mathbf{u}, \mathbf{w}) \in V \times W, \|\mathbf{x} - \mathbf{v}\| \leq \|\mathbf{x} - (\mathbf{u} + \mathbf{w})\|$ . The vector  $\mathbf{v}$  can be written  $\mathbf{f} + \mathbf{g}$  with  $\mathbf{f} \in V$  and  $\mathbf{g} \in W$ , so that the vector  $\mathbf{y} = \mathbf{x} - \mathbf{f}$ , which belongs to  $\mathbf{x} + V$ , satisfies  $d(\mathbf{x} - \mathbf{f}, W) = \|(\mathbf{x} - \mathbf{f}) - \mathbf{g}\| = d(\mathbf{x}, V + W) = d(\mathbf{x} + V, W)$ , which proves the result.  $\square$

Let  $\Sigma = \bigcup_{i \in [1, p]} V_i$ , where  $V_i$  are subspaces of  $\mathbb{R}^n$ . Lemma 5 applied to  $V = \mathcal{N}$  and  $W = V_i$  ensures the existence of  $\mathbf{x}_i \in (\mathbf{x} + \mathcal{N})$  such that  $d_E(\mathbf{x}_i, V_i) = d_E(\mathbf{x} + \mathcal{N}, V_i)$ . Therefore,  $\Delta(\mathbf{M}\mathbf{x})$  can be defined as  $\operatorname{argmin}_{\{\mathbf{x}_i, i \in [1, p]\}} d_E(\mathbf{x}_i, V_i)$  and satisfies  $d_E(\Delta(\mathbf{M}\mathbf{x}), \Sigma) = d_E(\mathbf{x} + \mathcal{N}, \Sigma)$ , so that the decoder  $\Delta(\mathbf{M}\mathbf{x}) = \operatorname{argmin}_{\mathbf{z} \in (\mathbf{x} + \mathcal{N})} d(\mathbf{z}, \Sigma)$  is properly defined. In particular, this applies to the decoder (10).

## B Proof of Theorem 1

Let  $\delta > 0$  and  $\Delta_\delta$  and  $C$  be such that (13) holds  $\forall \mathbf{x} \in E$ . Let  $\mathbf{h} \in \mathcal{N}$ . Then  $\exists \mathbf{h}_0 \in \Sigma - \Sigma$  such that  $d_E(\mathbf{h}, \mathbf{h}_0) \leq d_E(\mathbf{h}, \Sigma - \Sigma) + \delta$ . Let  $\mathbf{h}_0 = \mathbf{h}_1 - \mathbf{h}_2$  with  $\mathbf{h}_1, \mathbf{h}_2 \in \Sigma$ , and  $\mathbf{h}_3 = \mathbf{h} - \mathbf{h}_0$ . Since  $\mathbf{h} \in \mathcal{N}$ , we have:

$$\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3) = \mathbf{M}\mathbf{h}_2. \quad (53)$$

Applying (13) to  $\mathbf{x} = \mathbf{h}_2 \in \Sigma$  and using the fact that  $\|0\|_E = 0$ , we get:

$$\|\mathbf{A}\mathbf{h}_2 - \Delta_\delta(\mathbf{M}\mathbf{h}_2)\|_G \leq \delta. \quad (54)$$

Let's now find an upper bound for  $\|\mathbf{A}\mathbf{h}\|_G$ :

$$\begin{aligned} \|\mathbf{A}\mathbf{h}\|_G &= \|\mathbf{A}(\mathbf{h}_1 - \mathbf{h}_2 + \mathbf{h}_3)\|_G \\ &= \|\mathbf{A}(\mathbf{h}_1 + \mathbf{h}_3) - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3)) - \mathbf{A}\mathbf{h}_2 + \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G \\ &\leq \|\mathbf{A}(\mathbf{h}_1 + \mathbf{h}_3) - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G + \|\mathbf{A}\mathbf{h}_2 - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G, \end{aligned} \quad (55)$$

where we have used (11) and (12) for the last inequality. Combining (53) and (54), we get that:

$$\|\mathbf{A}\mathbf{h}_2 - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G \leq \delta. \quad (56)$$

Applying (13) to  $\mathbf{x} = \mathbf{h}_1 + \mathbf{h}_3$ , we get:

$$\begin{aligned} \|\mathbf{A}(\mathbf{h}_1 + \mathbf{h}_3) - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G &\leq C d_E(\mathbf{h}_1 + \mathbf{h}_3, \Sigma) + \delta \leq C \|\mathbf{h}_3\|_E + \delta \\ &= C d_E(\mathbf{h}, \mathbf{h}_0) + \delta \leq C d_E(\mathbf{h}, \Sigma - \Sigma) + (C + 1)\delta. \end{aligned} \quad (57)$$

Combining (55), (56) and (57) gives:

$$\|\mathbf{A}\mathbf{h}\|_G \leq C d_E(\mathbf{h}, \Sigma - \Sigma) + (C + 2)\delta. \quad (58)$$

(58) is valid for all  $\delta > 0$ , so it is valid for  $\delta = 0$ . This gives us the property (3) with  $D = C$ .

## C Proof of Theorem 2

Let's first assume that (14) holds and define the following decoder on  $F$ :

$$\Delta'(\mathbf{M}\mathbf{x}) = \underset{\mathbf{z} \in (\mathbf{x} + \mathcal{N})}{\operatorname{argmin}} d_E(\mathbf{z}, \Sigma). \quad (59)$$

Note that the decoder is well defined, since  $\mathbf{M}\mathbf{x}_1 = \mathbf{M}\mathbf{x}_2 \Rightarrow \mathbf{x}_1 + \mathcal{N} = \mathbf{x}_2 + \mathcal{N}$ .

For  $\mathbf{x} \in E$ , we have  $\mathbf{x} - \Delta'(\mathbf{M}\mathbf{x}) \in \mathcal{N}$ , so that (3) yields:

$$\begin{aligned} \|\mathbf{A}\mathbf{x} - \mathbf{A}\Delta'(\mathbf{M}\mathbf{x})\|_G &\leq Dd_E(\mathbf{x} - \Delta'(\mathbf{M}\mathbf{x}), \Sigma - \Sigma) \\ &\leq Dd_E(\mathbf{x}, \Sigma) + Dd_E(\Delta'(\mathbf{M}\mathbf{x}), \Sigma) \\ &\leq 2Dd_E(\mathbf{x}, \Sigma), \end{aligned} \quad (60)$$

where we have used (12) for the second inequality. The last inequality comes from (59), which yields  $d_E(\Delta'(\mathbf{M}\mathbf{x}), \Sigma) \leq d_E(\mathbf{x}, \Sigma)$ . Therefore, by posing  $\Delta = \mathbf{A}\Delta'$ , we get (2).

Let's return to the general case, and consider  $\nu > 0$ . We define the following decoder on  $F$ :

$$\Delta'_\nu(\mathbf{M}\mathbf{x}) \in \{\mathbf{u} \in (\mathbf{x} + \mathcal{N}) \mid d_E(\mathbf{u}, \Sigma) \leq d_E(\mathbf{x} + \mathcal{N}, \Sigma) + \nu\}. \quad (61)$$

Note that this set may not contain a unique element and thus this definition relies on the axiom of choice.

For  $\mathbf{x} \in E$ , we have again  $\mathbf{x} - \Delta'_\nu(\mathbf{M}\mathbf{x}) \in \mathcal{N}$ , so that by (3):

$$\begin{aligned} \|\mathbf{A}\mathbf{x} - \mathbf{A}\Delta'_\nu(\mathbf{M}\mathbf{x})\|_G &\leq Dd_E(\mathbf{x} - \Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma - \Sigma) \\ &\leq Dd_E(\mathbf{x}, \Sigma) + Dd_E(\Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma) \\ &\leq 2Dd_E(\mathbf{x}, \Sigma) + D\nu, \end{aligned} \quad (62)$$

where we have used (12) again for the second inequality. The last inequality comes from (61), which yields  $d_E(\Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma) \leq d_E(\mathbf{x}, \Sigma) + \nu$ . Therefore, by posing  $\Delta_\delta = \mathbf{A}\Delta'_{\delta/D}$ , we get (13).

## D Proof of Proposition 1

Let  $\mathbf{x} \in E$  and  $\nu > 0$ . If  $0 = d_E(\mathbf{x} + \mathcal{N}, \Sigma) = d_E(\mathbf{x}, \Sigma + \mathcal{N})$ , then since  $\Sigma + \mathcal{N}$  is a closed set,  $\mathbf{x} \in \Sigma + \mathcal{N}$ , and therefore  $(\mathbf{x} + \mathcal{N}) \cap \Sigma \neq \emptyset$ . In this case, we define  $\Delta'_\nu(\mathbf{M}\mathbf{x})$  as any element of  $(\mathbf{x} + \mathcal{N}) \cap \Sigma$ .

If  $d_E(\mathbf{x} + \mathcal{N}, \Sigma) > 0$ , then we define  $\Delta'_\nu(\mathbf{M}\mathbf{x}) \in \{\mathbf{u} \in (\mathbf{x} + \mathcal{N}) \mid d_E(\mathbf{u}, \Sigma) \leq (1 + \nu)d_E(\mathbf{x} + \mathcal{N}, \Sigma)\}$ . This provides a consistent definition of  $\Delta'_\nu$ .

Let's remark that for all  $\mathbf{x} \in E$ ,  $d_E(\Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma) \leq (1 + \nu)d_E(\mathbf{x} + \mathcal{N}, \Sigma) \leq (1 + \nu)d_E(\mathbf{x}, \Sigma)$ .

For  $\mathbf{x} \in E$ ,  $\mathbf{x} - \Delta'_\nu(\mathbf{M}\mathbf{x}) \in \mathcal{N}$ , so that (3) gives:

$$\begin{aligned} \|\mathbf{A}\mathbf{x} - \mathbf{A}\Delta'_\nu(\mathbf{M}\mathbf{x})\|_G &\leq Dd_E(\mathbf{x} - \Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma - \Sigma) \\ &\leq Dd_E(\mathbf{x}, \Sigma) + Dd_E(\Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma) \\ &\leq (2 + \nu)Dd_E(\mathbf{x}, \Sigma). \end{aligned} \quad (63)$$

Defining  $\Delta_\delta = \mathbf{A}\Delta'_\nu$ , we get the desired result.

## E Proof of Theorem 3 and Theorem 5

Let's first remark that applying (19) (resp. (21)) with  $\mathbf{x} = \mathbf{z} \in \Sigma$  and  $\mathbf{e} = 0$  yields  $\|\mathbf{Az} - \Delta_\delta(\mathbf{Mz})\|_G \leq \delta$  and  $\|\mathbf{Az} - \Delta_{\delta,\epsilon}(\mathbf{Mz})\|_G \leq C_2\epsilon + \delta$  for any  $\mathbf{z} \in \Sigma$ ,  $\epsilon \geq 0$ , where we have used the fact that  $\|0\|_F = 0$ .

Let  $\mathbf{h} \in E$  and  $\mathbf{z} \in \Sigma$ . We apply (19) (resp. (21)) with  $\mathbf{x} = \mathbf{z} - \mathbf{h}$ ,  $\mathbf{e} = \mathbf{Mh}$ , and  $\epsilon = \|\mathbf{Mh}\|_F$ , which yields:

$$\begin{aligned} \|\mathbf{Az} - \mathbf{Ah} - \Delta_\delta(\mathbf{Mz})\|_G &\leq C_1 d_E(\mathbf{z} - \mathbf{h}, \Sigma) + C_2 \|\mathbf{Mh}\|_F + \delta. \\ \|\mathbf{Az} - \mathbf{Ah} - \Delta_{\delta,\epsilon}(\mathbf{Mz})\|_G &\leq C_1 d_E(\mathbf{z} - \mathbf{h}, \Sigma) + C_2 \|\mathbf{Mh}\|_F + \delta. \end{aligned}$$

Let's remark that (11) and (12) imply  $\|\mathbf{y}\|_G \leq \|\mathbf{x} - \mathbf{y}\|_G + \|\mathbf{x}\|_G$  for all  $\mathbf{x}, \mathbf{y} \in G$ . Therefore, since  $\|\mathbf{Az} - \Delta_\delta(\mathbf{Mz})\|_G \leq \delta$  (resp.  $\|\mathbf{Az} - \Delta_{\delta,\epsilon}(\mathbf{Mz})\|_G \leq C_2 \|\mathbf{Mh}\|_F + \delta$ ), we have:

$$\begin{aligned} \|\mathbf{Ah}\|_G &\leq C_1 d_E(\mathbf{z} - \mathbf{h}, \Sigma) + C_2 \|\mathbf{Mh}\|_F + 2\delta. \\ (\text{respectively } \|\mathbf{Ah}\|_G &\leq C_1 d_E(\mathbf{z} - \mathbf{h}, \Sigma) + 2C_2 \|\mathbf{Mh}\|_F + 2\delta.) \end{aligned}$$

This last inequality is valid for all  $\mathbf{z} \in \Sigma$ , therefore (19) implies:

$$\begin{aligned} \|\mathbf{Ah}\|_G &\leq C_1 \inf_{\mathbf{z} \in \Sigma} d_E(\mathbf{z} - \mathbf{h}, \Sigma) + C_2 \|\mathbf{Mh}\|_F + 2\delta \\ &= C_1 \inf_{\mathbf{z} \in \Sigma} \inf_{\mathbf{u} \in \Sigma} \|\mathbf{z} - \mathbf{h} - \mathbf{u}\|_E + C_2 \|\mathbf{Mh}\|_F + 2\delta \\ &= C_1 d_E(\mathbf{h}, \Sigma - \Sigma) + C_2 \|\mathbf{Mh}\|_F + 2\delta, \end{aligned} \tag{64}$$

where we have used (11) for the last inequality. Similarly, (21) implies

$$\|\mathbf{Ah}\|_G \leq C_1 d_E(\mathbf{h}, \Sigma - \Sigma) + 2C_2 \|\mathbf{Mh}\|_F + 2\delta. \tag{65}$$

We conclude by using the fact that (64) and (65) hold for all  $\delta > 0$ .

## F Proof of Theorem 4

Let's suppose (20) and define for  $\delta > 0$  the decoder  $\Delta'_\delta : F \rightarrow E$  such that  $\forall \mathbf{y} \in F$ :

$$D_1 d_E(\Delta'_\delta(\mathbf{y}), \Sigma) + D_2 d_F(\mathbf{M}\Delta'_\delta(\mathbf{y}), \mathbf{y}) \leq \inf_{\mathbf{u} \in E} [D_1 d_E(\mathbf{u}, \Sigma) + D_2 d_F(\mathbf{Mu}, \mathbf{y})] + \delta. \tag{66}$$

Let's prove that this decoder meets property (19).

Let  $\mathbf{x} \in E$  and  $\mathbf{e} \in F$ . Applying (20) with  $\mathbf{h} = \mathbf{x} - \Delta'_\delta(\mathbf{Mx} + \mathbf{e})$ , we get:

$$\begin{aligned} \|\mathbf{A}(\mathbf{x} - \Delta'_\delta(\mathbf{Mx} + \mathbf{e}))\|_G &\leq D_1 d_E(\mathbf{x} - \Delta'_\delta(\mathbf{Mx} + \mathbf{e}), \Sigma - \Sigma) + D_2 \|\mathbf{M}(\mathbf{x} - \Delta'_\delta(\mathbf{Mx} + \mathbf{e}))\|_F \\ &\leq D_1 d_E(\mathbf{x}, \Sigma) + D_1 d_E(\Delta'_\delta(\mathbf{Mx} + \mathbf{e}), \Sigma) \\ &\quad + D_2 d_F(\mathbf{M}\Delta'_\delta(\mathbf{Mx} + \mathbf{e}), \mathbf{Mx} + \mathbf{e}) + D_2 \|\mathbf{e}\|_F \\ &\leq 2D_1 d_E(\mathbf{x}, \Sigma) + 2D_2 \|\mathbf{e}\|_F + \delta, \end{aligned} \tag{67}$$

where we have used (11) and (12) for the second inequality and the last inequality is a consequence of (66).

Posing  $\Delta_\delta = \mathbf{A}\Delta'_\delta$  proves (19) with  $C_1 = 2D_1$  and  $C_2 = 2D_2$ .

## G Proof of Lemma 1

The two equivalences are very similar to prove, so that we will only prove the first. (23)  $\Rightarrow$  (22) is obvious. Let's now suppose (22), so that:

$$\forall \mathbf{h} \in \mathcal{N}, \forall \mathbf{z} \in \Sigma - \Sigma, \|\mathbf{A}\mathbf{h}\|_G \leq D\|\mathbf{h} - \mathbf{z}\|_E. \quad (68)$$

We also have:

$$\forall \lambda \in \mathbb{R}^*, \forall \mathbf{h} \in \mathcal{N}, \forall \mathbf{z} \in \Sigma - \Sigma, \|\lambda\mathbf{A}\mathbf{h}\|_G \leq D\|\lambda\mathbf{h} - \mathbf{z}\|_E, \quad (69)$$

so that:

$$\forall \lambda \in \mathbb{R}^*, \forall \mathbf{h} \in \mathcal{N}, \forall \mathbf{z} \in \Sigma - \Sigma, \|\mathbf{A}\mathbf{h}\|_G \leq D\|\mathbf{h} - \mathbf{z}/\lambda\|_E. \quad (70)$$

This last inequality yields (23).

## H Proof of Theorem 6

Let's note  $\widetilde{\mathbf{M}} = \mathbf{M}|_V$  and  $\widetilde{\mathcal{N}} = \mathcal{N} \cap V$ . Let  $m$  be the dimension of the range of  $\widetilde{\mathbf{M}}$ , so that  $\widetilde{\mathcal{N}}$  is of dimension  $n - m$ . Let  $\mathbf{h}_1, \dots, \mathbf{h}_{n-m}$  be an orthonormal basis of  $\widetilde{\mathcal{N}}$ . We have:

$$n - m = \sum_{j=1}^{n-m} \|\mathbf{h}_j\|_2^2 \leq \frac{1}{K} \sum_{j=1}^{n-m} \sum_{i=1}^n \langle \mathbf{h}_j, \mathbf{z}_i \rangle^2. \quad (71)$$

Using (29), we get that, for all  $\mathbf{h} \in \mathcal{N}$  and unit-norm vector  $\mathbf{z} \in \Sigma - \Sigma$ ,  $\langle \mathbf{h}, \mathbf{z} \rangle^2 \leq \left(1 - \frac{1}{D_*^2}\right) \|\mathbf{h}\|_2^2$ . If we denote by  $p_{\widetilde{\mathcal{N}}}$  the orthogonal projection on  $\widetilde{\mathcal{N}}$  and apply this inequality with  $\mathbf{h} = p_{\widetilde{\mathcal{N}}}(\mathbf{z}_i) = \sum_{j=1}^{n-m} \langle \mathbf{h}_j, \mathbf{z}_i \rangle \mathbf{h}_j$  and  $\mathbf{z} = \mathbf{z}_i$ , we get that  $\|p_{\widetilde{\mathcal{N}}}(\mathbf{z}_i)\|_2^4 \leq \left(1 - \frac{1}{D_*^2}\right) \|p_{\widetilde{\mathcal{N}}}(\mathbf{z}_i)\|_2^2$ , which can be simplified to  $\|p_{\widetilde{\mathcal{N}}}(\mathbf{z}_i)\|_2^2 = \sum_{j=1}^{n-m} \langle \mathbf{h}_j, \mathbf{z}_i \rangle^2 \leq \left(1 - \frac{1}{D_*^2}\right)$  even if  $\|p_{\widetilde{\mathcal{N}}}(\mathbf{z}_i)\|_2 = 0$ .

Using this relation in (71), we get:

$$n - m \leq \frac{n}{K} \left(1 - \frac{1}{D_*^2}\right), \quad (72)$$

so that:

$$m \geq n \left(1 - \frac{1}{K} \left(1 - \frac{1}{D_*^2}\right)\right). \quad (73)$$

We get the lower bound on  $D_*^2$  by isolating it in the inequality.

## I Proof of Theorem 7

Let  $\mathbf{h} \in E$  and  $\mathbf{z} \in \Sigma - \Sigma$ . We have the following inequalities:

$$\|\mathbf{h}\|_G \leq \|\mathbf{h} - \mathbf{z}\|_G + \|\mathbf{z}\|_G \leq \|\mathbf{h} - \mathbf{z}\|_G + \frac{1}{\alpha} \|\mathbf{M}\mathbf{z}\|_F, \quad (74)$$

where we have used the lower-RIP for the second inequality.



A similar consideration on  $\mathbf{Mz}$  yields:

$$\|\mathbf{Mz}\|_F \leq \|\mathbf{M}(\mathbf{z} - \mathbf{h})\|_F + \|\mathbf{Mh}\|_F. \quad (75)$$

Substituting (75) into (74), we get:

$$\begin{aligned} \|\mathbf{h}\|_G &\leq \|\mathbf{h} - \mathbf{z}\|_G + \frac{1}{\alpha} \|\mathbf{M}(\mathbf{h} - \mathbf{z})\|_F + \frac{1}{\alpha} \|\mathbf{Mh}\|_F \\ &= \|\mathbf{h} - \mathbf{z}\|_M + \frac{1}{\alpha} \|\mathbf{Mh}\|_F. \end{aligned} \quad (76)$$

Taking the infimum of the right hand-side quantity over all  $\mathbf{z} \in \Sigma - \Sigma$ , one gets the desired Robust NSP:

$$\|\mathbf{h}\|_G \leq d_M(\mathbf{h}, \Sigma - \Sigma) + \frac{1}{\alpha} \|\mathbf{Mh}\|_F. \quad (77)$$