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

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

Anthony Bourrier, Mike E. Davies, *Senior Member, IEEE,* Tomer Peleg, *Student Member, IEEE,* Patrick Pérez and Rémi Gribonval, *Fellow, IEEE*

***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 ℓ^2/ℓ^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 property. This norm can be upper bounded by an atomic norm relative to the considered model.

***Index Terms*—Linear inverse problems, instance optimality, null space property, restricted isometry property.**

I. 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 $\mathbf{Mx}$ is thus a hopeless task unless one can exploit prior knowledge on $\mathbf{x}$ to complete the incomplete observation $\mathbf{Mx}$.

Sparsity is a well-known enabling model for this type of inverse problems: it has been proven that for certain such operators $\mathbf{M}$, one can expect to recover the signal $\mathbf{x}$ from its measure $\mathbf{Mx}$ provided that $\mathbf{x}$ is sufficiently sparse, *i.e.*, it has few nonzero components [1]. If the set of k-sparse signals is denoted $\Sigma_k = \{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|_0 \leq k\}$, where $\|.\|_0$ is the pseudo-norm counting the number of nonzero components, then this recovery property can be interpreted as the existence of a decoder $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ such that $\forall \mathbf{x} \in \Sigma_k,\ \Delta(\mathbf{Mx}) = \mathbf{x}$, thus making $\mathbf{M}$ a linear encoder associated to the (typically nonlinear) decoder Δ.

Further, the body of theoretical work around sparse recovery in linear inverse problems has given rise to the notion of compressive sensing (CS) [2], where the focus is on *choosing* – among a more or less constrained set of operators – a dimensionality reducing $\mathbf{M}$ to which a decoder can be associated[1]. It is now well-established that this can be achieved in scenarii 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.

A. Instance optimal sparse decoders

A good decoder Δ is certainly expected to have nicer properties than simply reconstructing Σ_k. Indeed, the signal $\mathbf{x}$ to be reconstructed may not belong exactly in Σ_k but "live near" Σ_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 $\mathbf{Mx}$, that is a quantity $\Delta(\mathbf{Mx})$ such that $\|\mathbf{x} - \Delta(\mathbf{Mx})\|$ is "small" for a certain norm $\|.\|$. This stability to the model has been formalized into the so-called *Instance Optimality* assumption on Δ. Decoder Δ is said to be instance optimal if:

$$\forall \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x} - \Delta(\mathbf{Mx})\| \leq C d(\mathbf{x}, \Sigma_k), \tag{1}$$

for a certain choice of norm $\|.\|$ 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 ℓ^2 or ℓ^1 and the distance is ℓ^1, such good instance optimal decoders exist and can be implemented as the minimization

A. Bourrier is with Gipsa-Lab. M.E. Davies is with University of Edimburgh. T. Peleg is with Israel Institute of Technology. P. Pérez is with Technicolor. R. Gribonval is with INRIA.

[1]By contrast, linear inverse problems usually refer to a setup where one aims at reconstructing a signal from its measurements by a given operator (*e.g.* imposed by the underlying physics), which may be dimensionality-reducing.

of a convex objective [2]–[4] 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 [5]. 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 [6].

In [7], 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 Σ_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 Δ 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 Σ_{2k}, called the *Null Space Property* (NSP). This relation allowed them to study the existence of stable decoders under several choices of norm $\|.\|$ and distance $d(.,.)$.

A related question addressed in [7] 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 the size of the constant C in (1) (with ℓ^2 norm and ℓ^2 distance) and the dimension reduction ratio m/n.

B. 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 [8]. In these generalized models, the signals of interest typically live in or close to a subset Σ 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 [9], unions of subspaces, whether finite [10] or possibly infinite [11], signals sparse in a redundant dictionary [12], cosparse signals [13], approximately low-rank matrices [14], [15], low-rank and sparse matrices [16], [17], symmetric matrices with sparse inverse [18], [19] or manifolds [20], [21]. An old result which can also be interpreted as generalized CS is the low-dimensional embedding of a point cloud [22], [23]. 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 [7] to these general models, allowing to further investigate the well-posedness of such problems.

In [24], the theoretical results of [7] 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 ℓ^2/ℓ^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 Σ 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 of instance optimality as considered in [7]:

- **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 or a space a signed measures. This is motivated by recent work on infinite dimensional compressed sensing [25] or compressive density estimation [26];
- **Task-oriented decoders:** the decoder is not constrained to approximate the signal $\mathbf{x}$ itself but rather a linear feature derived from the signal, $\mathbf{A}\mathbf{x}$, as in [24]; 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.
- **Pseudo-norms:** Instead of considering instance optimality involving norms, we use pseudo-norms with fewer constraints, allowing us to characterize a wider range of instance optimality properties. As we will see in Section II-C, this flexibility on the pseudo-norms has a relationship with the previous point: it essentially allows one to suppose $\mathbf{A} = \mathbf{I}$ in every case, up to a change in the pseudo-norm considered for the approximation error.

C. Contributions of this work

We summarize below our main contributions.

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:

$$\forall \mathbf{x} \in E, \|\mathbf{A}\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\|_G \leq C d_E(\mathbf{x}, \Sigma), \tag{2}$$

where $\|.\|_G$ 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

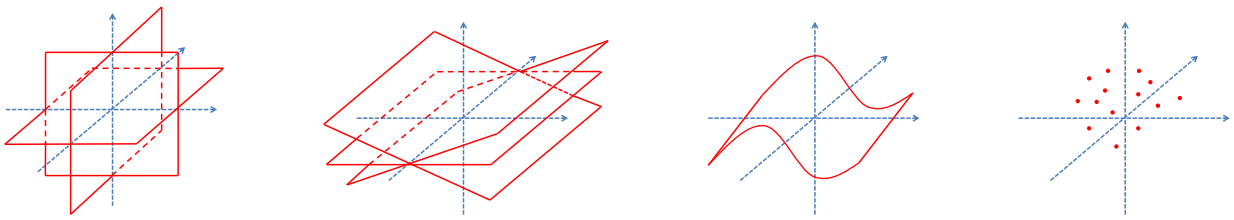


Fig. 1. Illustration of several CS models. *From left to right:* k-sparse vectors, union of subspaces, smooth manifold and point cloud.

from $\mathbf{Mx}$. Our first contribution is to prove that the existence of a decoder Δ 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:

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

where the set $\Sigma - \Sigma$ is comprised of all differences of elements in Σ, 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.

2) Noise-robust instance optimality: Our second contribution (Theorems 3 and 4) is to link a noise-robust extension of instance optimality to a property called the Robust NSP. Section II regroups these noiseless and noise-robust results after a review of the initial IOP/NSP results of [7]. 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).

If a Robust NSP is satisfied, an instance optimal decoder can be defined as:

$$\Delta(\mathbf{y}) = \underset{\mathbf{u} \in E}{\operatorname{argmin}} \quad D_1 d_E(\mathbf{u}, \Sigma) + D_2 d_F(\mathbf{Mu}, \mathbf{y}), \tag{4}$$

where the constants D_1, D_2 and distances d_E, d_F are those which appear in the Robust NSP. The objective function is the sum of two terms: a distance to the model and a distance to the measurements. Also note that by fixing D_2 to infinity, one defines an instance optimal noise-free decoder provided the corresponding NSP is satisfied.

3) Limits of dimensionality reduction with generalized models: The reformulation of IOP as an NSP allows us to consider the ℓ^2/ℓ^2 instance optimality for general models in Section III. 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 ℓ^2/ℓ^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 ℓ^2/ℓ^2 IOP with dimensionality reduction is impossible:

- **k-sparse vectors**. In the case where $\Sigma = \Sigma_k$ is the set of k-sparse vectors, Σ contains the null vector and the canonical basis, so that $\Sigma - \Sigma$ contains the canonical basis. Note that the impossibility of good ℓ^2/ℓ^2 IOP has been proved in [7].
- **Block-sparse vectors** [9]. The same argument as above applies in this case as well, implying that imposing a block structure on sparsity does not improve ℓ^2/ℓ^2 feasibility.
- **Low-rank matrices** [14], [15]. In the case where $E = \mathcal{M}_n(\mathbb{R})$ and Σ is the set of matrices of rank $\leq k$, Σ also contains the null matrix and the canonical basis.
- **Low-rank + sparse matrices** [16], [17]. 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 [27]–[29].
- **Low-rank matrices with non-sparsity constraints**. In order to reduce the ambivalence of the low-rank + sparse decomposition of a matrix, [17] 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** [8] obtained by pruning out the combinatorial collection of k-dimensional subspaces associated to k-sparse vectors. This covers block-sparse vectors [9], 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 ℓ^2/ℓ^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** [12]. 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** [13], [24]. As shown in [24], 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) [18], [19]. In very high-dimensional cases, compressive solutions to this problem would be appealing.

- Johnson-Lindenstrauss embedding of **point clouds** [23]. Given a set $\mathcal{X}$ of L vectors in $\mathbb{R}^n$ and $\epsilon > 0$, there exists a linear mapping $f : \mathbb{R}^n \to \mathbb{R}^m$, with $m = \mathcal{O}(\ln(L)/\epsilon^2)$ and

$$(1-\epsilon)\|\mathbf{x}-\mathbf{y}\|_2 \leq \|f(\mathbf{x})-f(\mathbf{y})\|_2 \leq (1+\epsilon)\|\mathbf{x}-\mathbf{y}\|_2 \quad (5)$$

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 ℓ^2-sense, except for a very limited neighborhood of the point cloud. This is further discussed in Section V.

4) Generalized Restricted Isometry Property: Our last contribution, in Section IV, is to study the relations between the NSP and a generalized version of the Restricted Isometry Property (RIP). This generalized RIP bounds $\|\mathbf{Mx}\|_F$ from below and/or above on a certain set V, and can be decomposed in:

$$\text{Lower} - \text{RIP} : \forall \mathbf{x} \in V, \alpha\|\mathbf{x}\|_G \leq \|\mathbf{Mx}\|_F \quad (6)$$

$$\text{Upper} - \text{RIP} : \forall \mathbf{x} \in V, \|\mathbf{Mx}\|_F \leq \beta\|\mathbf{x}\|_G, \quad (7)$$

where $\|.\|_G$ and $\|.\|_F$ are pseudo-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"[2].

Furthermore, we prove that under an upper-RIP assumption on Σ, this M-norm can be upper bounded by an atomic norm [5] defined using Σ and denoted $\|.\|_\Sigma$. 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 9): if $\mathbf{M}$ satisfies a lower-RIP (6) for $V = \Sigma - \Sigma$ and an upper-RIP (7) for $V = \Sigma$, then for all $\delta > 0$, there exists a decoder Δ_δ satisfying $\forall \mathbf{x} \in E, \forall \mathbf{e} \in F$,

$$\|\mathbf{x} - \Delta_\delta(\mathbf{Mx}+\mathbf{e})\|_G \leq 2\left(1+\frac{\beta}{\alpha}\right) d_\Sigma(\mathbf{x},\Sigma) + \frac{2}{\alpha}\|\mathbf{e}\|_E + \delta, \quad (8)$$

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

[2]The prefix "M-" should be thought as "Measurement-related norm" since in other works the measurement matrix may be denoted by other letters.

In particular, this generalized RIP encompasses classical or recent RIP formulations, such as

- The **standard RIP** [4] with V as the set of k-sparse vectors, $\|\cdot\|_G$ and $\|\cdot\|_F$ being ℓ^2 norms.
- The **Union of Subspaces RIP** [11] with V as a union of subspaces, $\|\cdot\|_G$ and $\|\cdot\|_F$ being ℓ^2 norms.
- The **RIP for low-rank matrices** [15] with V as the set of matrices of rank $\leq r$, $\|\cdot\|_G$ as the Frobenius norm and $\|\cdot\|_F$ as the ℓ^2 norm;
- The **D-RIP** [30] for the dictionary model with V as the set of vectors spanned by k columns of a dictionary matrix, $\|\cdot\|_G$ and $\|\cdot\|_F$ being ℓ^2 norms;
- The **$\mathbf{\Omega}$-RIP** [31] for the cosparse model with V as the set of vectors $\mathbf{x}$ such that $\mathbf{\Omega x}$ is k-sparse, where Ω is the cosparse operator, $\|\cdot\|_G$ and $\|\cdot\|_F$ being ℓ^2 norms;
- Similarly, the **task-RIP** can be defined given a linear operator $\mathbf{A}$ such that one aims at reconstructing the quantity $\mathbf{Ax}$ (instead of $\mathbf{x}$) to perform a particular task. As we will see in Section II-C, in terms of IOP, this is essentially equivalent to reconstructing $\mathbf{x}$ in terms of the norm $\|\mathbf{A}\cdot\|_G$. In this case, the corresponding lower task-RIP reads:

$$\alpha\|\mathbf{Ax}\|_G \leq \|\mathbf{Mx}\|_F. \quad (9)$$

5) Infinite-dimensional inverse problems: The generalization of the relationship between the IOP, the NSP and the RIP to arbitrary vector spaces allows us to consider recovery results for infinite dimensional inverse problems. Such problems have mainly been considered in separable Hilbert spaces [25], [32], where the signals of interest are sparse with respect to a Hilbert basis and the measurement operators subsample along another Hilbert basis. In the theory of generalized sampling [32], even when the signal model Σ 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 [33] can be read as a statement of ℓ^2/ℓ^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 ℓ^2.

However, as mentioned in Section IV-C1, *one cannot hope to get uniform instance optimality in this setting for a standard sparsity model*. This is mentioned in [25] when the authors state that no RIP can be satisfied in this case. In section IV-C2, we nevertheless discuss the possibility of uniform instance optimality results in infinite dimensions *with a proper choice of model Σ and pseudo-norms*. In particular, a non-constructive topological result ensures that a generalized RIP is satisfied for a model of finite box-counting dimension [34]. This generalized RIP leads to an IOP, according to Theorem 7.

Hopefully, our results will therefore help characterizing conditions under which infinite-dimensional uniform IOP is possible.

D. Structure of the paper

We will now describe the layout of the paper. Section II 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 III then focuses on the particular case of ℓ^2/ℓ^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 IV, 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". Using a topological result, we illustrate that this implication may be exploited for certain models and norms, even in infinite dimensions. 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.

II. Generalized IOP and NSP equivalences

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

A. The initial result

In [7], the authors consider two norms $\|.\|_G$ and $\|.\|_E$ defined on the signal space $\mathbb{R}^n$. The distance derived from $\|.\|_E$ 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 $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ is said to be instance optimal for k-sparse signals if

$$\forall \mathbf{x} \in \mathbb{R}^n, \|\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\|_G \leq C d_E(\mathbf{x}, \Sigma_k), \tag{10}$$

for some constant $C > 0$.

This property on Δ upper bounds the reconstruction error of a vector, measured by $\|.\|_G$, by the distance from the vector to the model, measured by d_E. The authors prove that the existence of an instance optimal decoder, called IOP, is closely related to the NSP of $\mathbf{M}$ with respect to the set Σ_{2k} of $2k$-sparse vectors. Noting $\mathcal{N} = \ker(\mathbf{M})$, this NSP states

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_G \leq D d_E(\mathbf{h}, \Sigma_{2k}) \tag{11}$$

for some constant D.

The relationship between the IOP and the NSP is the following: if there exists an instance optimal decoder Δ satisfying (10), then (11) holds with $D = C$. Conversely, if (11) holds, then there exists a decoder Δ such that (10) holds with $C = 2D$. Such a decoder can be defined as follows, supposing $\mathbf{M}$ is onto:

$$\Delta(\mathbf{M}\mathbf{x}) = \underset{\mathbf{z}\in(\mathbf{x}+\mathcal{N})}{\operatorname{argmin}} \quad d_E(\mathbf{z}, \Sigma_k), \tag{12}$$

$\mathbf{x}+\mathcal{N}$ 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.

B. 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 or a space of signed measures. On this space is defined a linear operator $\mathbf{M} : E \to F$, 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 [24], 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 $\Delta : F \to G$ will aim at approximating $\mathbf{A}\mathbf{x}$ from $\mathbf{M}\mathbf{x}$.

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

$$\textbf{Symmetry} : \|\mathbf{x}\|_G = \|-\mathbf{x}\|_G \tag{13}$$

$$\textbf{Triangle inequality} : \|\mathbf{x}+\mathbf{y}\|_G \leq \|\mathbf{x}\|_G + \|\mathbf{y}\|_G. \tag{14}$$

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 $G = \mathbb{R}^n$:

- $\|.\|_G$ can be defined as a "non-normalized" ℓ^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 $f : \mathbb{R}_+ \to \mathbb{R}_+$ is a concave function such that $f(x) = 0 \Leftrightarrow x = 0$, then $\|.\|_G$ can be defined as the f-(pseudo-)norm $\|\mathbf{x}\|_f = \sum_{i=1}^n f(|x_i|)$, see [35].

In order to measure the *distance from a vector to the model*, we also endow E with a pseudo-norm $\|.\|_E : E \to \mathbb{R}_+$ which satisfies the same properties as $\|.\|_G$ 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, $\|.\|_E$ can be defined as a non-normalized ℓ^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 $\|.\|_F$. The assumptions we make on $\|.\|_F$ are the same as the assumptions on $\|.\|_E$.

To sum up, here are the extensions we propose compared to the framework of [7], [24] :

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

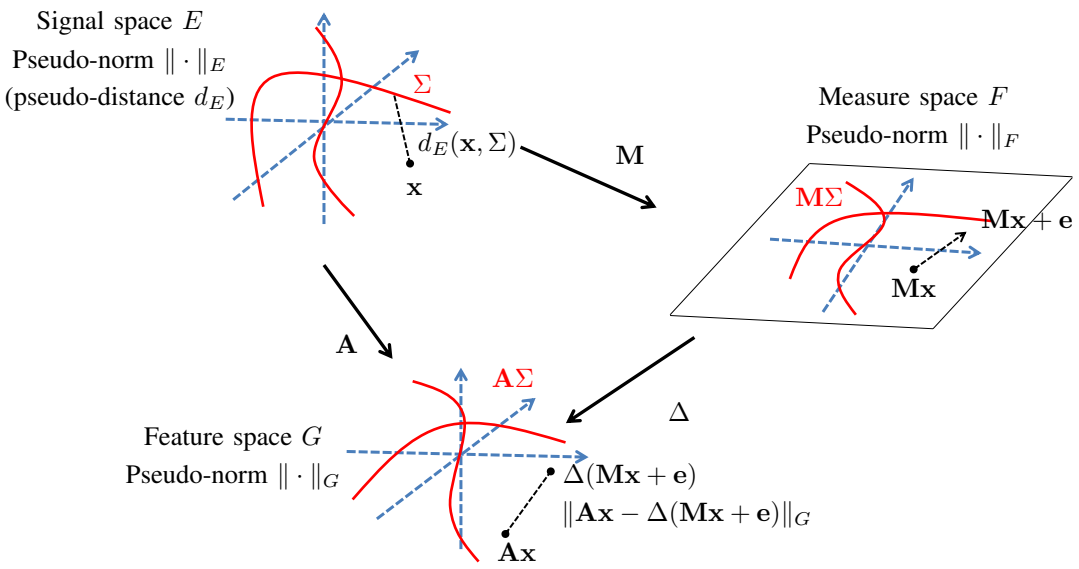


Fig. 2. Illustration of the proposed generalized setting. The signals belong to the space E, supplied with a pseudo-norm $\|.\|_E$ used to measure the distance from a vector to the model Σ containing the signals of interest. E is mapped in the measure space F by the operator $\mathbf{M}$ and the measure is perturbed by an additive noise $\mathbf{e}$. The space F is supplied with a pseudo-norm $\|.\|_F$. The feature space G, supplied with a norm $\|.\|_G$, 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 Δ. The reconstruction error for the vector $\mathbf{x}$ and noise $\mathbf{e}$ is therefore $\|\mathbf{Ax}-\Delta(\mathbf{Mx}+\mathbf{e})\|_G$. Note that in the case where $E=G$ and $\mathbf{A}=\mathbf{I}$, the decoder is aimed at reconstructing exactly the signals.

- The reconstruction of $\mathbf{Ax}$ is targeted rather than that of $\mathbf{x}$.
- The functions $\|.\|_E$, $\|.\|_F$ and $\|.\|_G$ need not be norms but can be pseudo-norms with relaxed hypotheses. In particular, Table I summarizes the requirements on these functions.

Once we have derived the generalized IOP and NSP equivalences, we see that one can essentially be brought back to the case where $\mathbf{A}=\mathbf{I}$ with a proper choice of $\|\cdot\|_G$. This will be discussed in Section II-C.

Let's note that even though [7] does not consider the noisy case, some other works have studied noisy instance optimality for the standard sparse model and with ℓ^p-norms ([36], Chapter 11 of [37]). They mainly study conditions under which standard ℓ^1 decoders are instance optimal. Here, we adopt a more conceptual approach by considering conditions for the *existence* of an instance optimal decoder, without restriction on its practical tractability. This has the advantage of providing fairly simple equivalences and also to identify fundamental performance limits in a certain framework.

In these works, the underlined relationships between ℓ^1 instance optimality and NSP are somewhat different than ours since they usually take advantage of the particular geometry of the sparse problem with ℓ^1 decoder. An interesting open question is to what extent we can bridge the gap between this particular setup and a more general setup.

1) The noiseless case: We first consider the same framework as [7], [24], where one measures $\mathbf{Mx}$ with infinite precision. In our generalized framework, instance optimality for a decoder Δ reads:

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

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

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

with $D=C$. Note that the set Σ_{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 Σ 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 Δ_δ such that:

$$\forall \mathbf{x}\in E, \|\mathbf{Ax}-\Delta_\delta(\mathbf{Mx})\|_G \leq C d_E(\mathbf{x},\Sigma)+\delta, \qquad (15)$$

reflecting the fact that one cannot necessarily consider the exact quantity

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

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 [11].

Remark 1. *To understand the necessity of such an additive error term when Σ 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 $\|.\|_G/\|.\|_E$ are the ℓ^2 norm. In this case, the minimal distance between $\mathbf{x}+\mathcal{N}$ and Σ is not reached at any point, making it necessary to add the δ term for the decoder to be well-defined.*

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

TABLE I
SUMMARY OF THE HYPOTHESES ON THE PSEUDO-NORMS $\|.\|_E$, $\|.\|_F$ AND $\|.\|_G$. A CROSS MEANS THE PROPERTY IS REQUIRED, A HORIZONTAL BAR MEANS IT IS NOT.

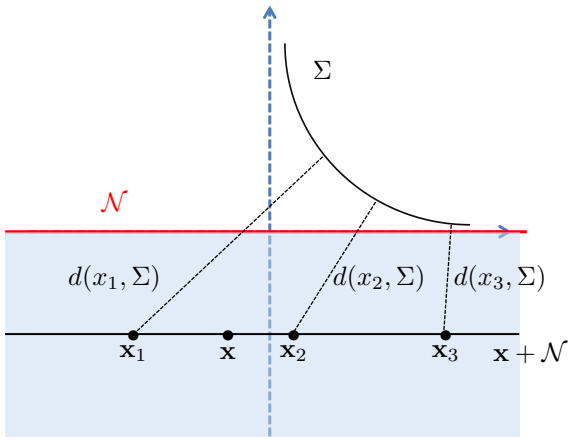


Fig. 3. Necessity of the additive term δ in a simple case. For each $\mathbf{x}$ in the blue half-plane, the distance $d_E(x + \mathcal{N}, \Sigma)$ is never reached at a particular point of $\mathbf{x} + \mathcal{N}$: the distance strictly decreases as one goes right along the affine plane $\mathbf{x}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

In this setting, the NSP (3) implies the existence of instance optimal decoders in the sense of (15) 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.

Theorem 1. *Suppose $\forall \delta > 0$, there exists a decoder Δ_δ satisfying* (15)*:*

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

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

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

with constant $D = C$.

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

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

Then $\forall \delta > 0$, there exists a decoder Δ_δ satisfying (15)*:*

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

with $C = 2D$.

If we further assume that

$$\forall \mathbf{x} \in E, \exists \mathbf{z} \in (\mathbf{x} + \mathcal{N}), d_E(\mathbf{z}, \Sigma) = d_E(\mathbf{x} + \mathcal{N}, \Sigma), \quad (16)$$

then there exists a decoder Δ satisfying (2)*:*

$$\forall \mathbf{x} \in E, \|\mathbf{A}\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\|_G \leq C d_E(\mathbf{x}, \Sigma) \quad (17)$$

with $C = 2D$.

Note that this result is similar to the result proven in [24], which was stated in the case where Σ is a finite union of subspaces in finite dimension. In this framework, condition (16) is always satisfied as soon as $\|.\|_E$ is a norm, by the same argument as in usual CS (see Appendix A).

Let's also note the following property: if $\|.\|_E$ 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 Δ_δ satisfying (2) with $C = (2 + \delta)D$. This assumption therefore allows us to suppress the additive constant in (15) 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 Δ_δ satisfying:*

$$\forall \mathbf{x} \in E, \|\mathbf{A}\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x})\|_G \leq (2 + \delta) D d_E(\mathbf{x}, \Sigma). \quad (18)$$

2) The noisy case: In practice, it is not likely that one can measure with infinite precision the quantity $\mathbf{M}\mathbf{x}$. 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 $\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{e}$. 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 Δ fulfills the following

property:

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F,\ \|\mathbf{e}\|_F \leq \epsilon \\ \Rightarrow \|\mathbf{A}\mathbf{x} - \Delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2\epsilon. \quad (19)$$

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 ℓ^1 decoder in the sparse case with bounded noise [4] for $\|.\|_G = \|.\|_2$ and $\|.\|_E = \|.\|_1/\sqrt{k}$, provided $\mathbf{M}$ satisfies the RIP on Σ_{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*:

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \\ \|\mathbf{A}\mathbf{x} - \Delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2\|\mathbf{e}\|_F. \quad (20)$$

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. [38, Corollary 3.9].

Of course, the existence of a *noise-blind* noise-robust decoder in the sense of (20) implies the existence of a *noise-aware* noise-robust decoder in the sense of (19) for any noise level ϵ. 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 Σ and possibly infinite dimensional E requires some caution. For $\delta > 0$, the noise-robust (and noise-blind) instance optimality of a decoder Δ_δ is defined as:

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

One can see that Δ_δ necessarily also satisfies the noiseless instance optimality (15) by setting $\mathbf{e} = 0$.

As we show below, if for every $\delta > 0$ there exists a noise-robust instance optimal decoder Δ_δ satisfying (21), then a generalized NSP for $\mathbf{M}$ relatively to $\Sigma - \Sigma$, referred to as Robust NSP, must hold:

$$\forall \mathbf{h} \in E, \|\mathbf{A}\mathbf{h}\|_G \leq D_1 d_E(\mathbf{h}, \Sigma - \Sigma) + D_2\|\mathbf{M}\mathbf{h}\|_F, \quad (22)$$

with $D_1 = C_1$ and $D_2 = C_2$. This property appears e.g. in [37] (Chap. 4) with $\|.\|_G = \|.\|_E = \|.\|_1$ and $\|.\|_F$ any norm. Note that this Robust NSP concerns every vector of E and not just the vectors of the null space $\mathcal{N} = \ker(\mathbf{M})$[3]. 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 $\mathbf{M}\mathbf{h}$, appears in the upper bound.

[3]In fact, unlike the NSP (3), (22) 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.

Conversely, the Robust NSP implies the existence of noise-robust instance optimal decoders Δ_δ satisfying (21) 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 Δ_δ satisfying* (21)*:*

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

Then $\mathbf{M}$ satisfies the Robust NSP (22)*:*

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

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

Theorem 4. *Suppose that $\mathbf{M}$ satisfies the Robust NSP* (22)*:*

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

Then $\forall \delta > 0$, there exists a decoder Δ_δ satisfying (21)*:*

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

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

$$\forall \mathbf{x} \in E, \forall \mathbf{e} \in F,\ \|\mathbf{e}\|_F \leq \epsilon \\ \Rightarrow \|\mathbf{A}\mathbf{x} - \Delta_{\delta,\epsilon}(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq C_1 d_E(\mathbf{x}, \Sigma) + C_2\epsilon + \delta. \quad (23)$$

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:

Theorem 5. *Suppose $\forall \epsilon, \delta > 0$, there exists a* noise-aware *decoder $\Delta_{\delta,\epsilon}$ satisfying* (23)*:*

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

Then $\mathbf{M}$ satisfies the Robust NSP (22) *with constants $D_1 = C_1$ and $D_2 = 2C_2$. Therefore, by Theorem 4, there exists an instance optimal* noise-blind *decoder satisfying:*

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

C. Task-oriented instance optimality

In this section, we show that the generalized instance optimality as stated in (15) is essentially equivalent to the same property with $\mathbf{A} = \mathbf{I}$ and a different choice for the pseudo-norm $\|.\|_G$.

Indeed, let's consider that one aims at reconstructing a certain feature $\mathbf{A}\mathbf{x}$ from the measurements $\mathbf{M}\mathbf{x}$. If for any $\delta > 0$ there exists an instance optimal decoder Δ_δ such that (15) is satisfied, then Theorem 1 ensures that NSP (3)

is satisfied. Let's define the following pseudo-norm for any signal $\mathbf{x} \in E$:

$$\|\mathbf{x}\| = \|\mathbf{A}\mathbf{x}\|_G. \tag{24}$$

The following NSP is satisfied:

$$\forall \mathbf{h} \in \ker(\mathbf{M}),\ \|\mathbf{h}\| \leq C d_E(\mathbf{h}, \Sigma - \Sigma). \tag{25}$$

Therefore, Theorem 2 ensures that there exists decoders $\Delta'_\delta : F \to G$ instance optimal in the following sense:

$$\|\mathbf{x} - \Delta'_\delta(\mathbf{M}\mathbf{x})\| \leq 2C d_E(\mathbf{x}, \Sigma) + \delta. \tag{26}$$

This means that if a family of decoders Δ_δ aimed at decoding a feature $\mathbf{A}\mathbf{x}$ is instance optimal for the pseudo-norm $\|.\|_G$, then there exists a family of decoders Δ'_δ aimed at decoding the *signal* $\mathbf{x}$ which is instance optimal for the pseudo-norm $\|.\|$ with a similar constant (up to a factor 2). Conversely, if there exists a family of decoders Δ'_δ aimed at decoding $\mathbf{x}$ which is instance optimal for the pseudo-norm $\|.\|$, then a simple rewriting of the IOP gives that the decoders $\Delta_\delta = \mathbf{A}\Delta'_\delta$ are instance optimal for the pseudo-norm $\|.\|_G$.

Therefore, IOP with a task-oriented decoder is essentially equivalent to IOP with a standard decoder provided a suitable change in the pseudo-norm is performed. The same reasoning can be applied to deduce this equivalence for Robust IOP. As a consequence, we will only consider the case $\mathbf{A} = \mathbf{I}$ in the remainder of the paper.

III. ℓ^2/ℓ^2 Instance Optimality

In this section, we suppose that E is a Hilbert space equipped with the norm $\|.\|_2$ and scalar product $\langle .,. \rangle$, 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 $\|.\|_G = \|.\|_E = \|.\|_2$ in a dimensionality reducing context where $m \ll n$?*

A result of [7] 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 [24], the authors prove that this lower bound on the constant holds in the case where Σ is a finite union of subspaces in finite dimension. Here, we are interested in a version of this result for the general case where Σ can be a more general subset of E. More precisely, we will give a sufficient condition on Σ under which the optimal ℓ^2/ℓ^2 instance optimalityconstant is of the order of $\sqrt{\frac{n}{m}}$, thus preventing the existence of a ℓ^2/ℓ^2 instance optimal decoder with small constant if $m \ll n$.

A. Homogeneity of the NSP

In the case where $\|.\|_G$, $\|.\|_E$ and $\|.\|_F$ are homogeneous with the same degree, 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 $\|.\|_G$ and $\|.\|_E$ are homogeneous with the same degree, we have an equivalence between the NSP on $\Sigma - \Sigma$:*

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_G \leq D d_E(\mathbf{h}, \Sigma - \Sigma), \tag{27}$$

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

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_G \leq D d_E(\mathbf{h}, \mathbb{R}(\Sigma - \Sigma)). \tag{28}$$

Similarly, if $\|.\|_G$, $\|.\|_E$ and $\|.\|_F$ are homogeneous with the same degree, we have an equivalence between the robust NSP on $\Sigma - \Sigma$:

$$\forall \mathbf{h} \in E, \|\mathbf{h}\|_G \leq D_1 d_E(\mathbf{h}, \Sigma - \Sigma) + D_2 \|\mathbf{M}\mathbf{h}\|_F, \tag{29}$$

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

$$\forall \mathbf{h} \in E, \|\mathbf{h}\|_G \leq D_1 d_E(\mathbf{h}, \mathbb{R}(\Sigma - \Sigma)) + D_2 \|\mathbf{M}\mathbf{h}\|_F. \tag{30}$$

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 Σ is a union of subspaces since $\Sigma - \Sigma$ is already a cone in this case.

B. The optimal ℓ^2/ℓ^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 ℓ^2/ℓ^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 ℓ^2/ℓ^2 NSP is satisfied with constant D:

$$D_* = \inf \quad \{D \in \mathbb{R}_+ | \forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_2 \leq D d_2(\mathbf{h}, \Sigma - \Sigma)\}. \tag{31}$$

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

$$\begin{aligned} D_* &= \sup_{\mathbf{h} \in \mathcal{N}} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma)} \frac{\|\mathbf{h}\|_2}{\|\mathbf{h} - \mathbf{z}\|_2} \\ &= \sup_{\mathbf{h} \in \mathcal{N} \setminus \{0\}} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma)} \frac{1}{\|\frac{\mathbf{h}}{\|\mathbf{h}\|_2} - \frac{\mathbf{z}}{\|\mathbf{h}\|_2}\|_2}. \end{aligned} \tag{32}$$

Denoting $\mathcal{B}_2$ the unit ball for the ℓ^2 norm, we can rewrite this last expression as:

$$\begin{aligned} D_* &= \sup_{\mathbf{h} \in \mathcal{N} \cap \mathcal{B}_2} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma)} \frac{1}{\|\mathbf{h} - \mathbf{z}\|_2} \\ &= \sup_{\mathbf{h} \in \mathcal{N} \cap \mathcal{B}_2} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma) \cap \mathcal{B}_2} \sup_{\lambda \in \mathbb{R}} \frac{1}{\|\mathbf{h} - \lambda \mathbf{z}\|_2}. \end{aligned} \tag{33}$$

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:

$$D_* = \sup_{\mathbf{h} \in \mathcal{N} \cap \mathcal{B}_2} \sup_{\mathbf{z} \in \mathbb{R}(\Sigma - \Sigma) \cap \mathcal{B}_2} \frac{1}{\sqrt{1 - \langle \mathbf{h}, \mathbf{z} \rangle^2}}. \tag{34}$$

The contraposition of Theorem 1 gives the following result : if the NSP (3) is not satisfied for a certain constant D, then no decoder Δ_δ can satisfy instance optimality (15) with constant D. In the ℓ^2/ℓ^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

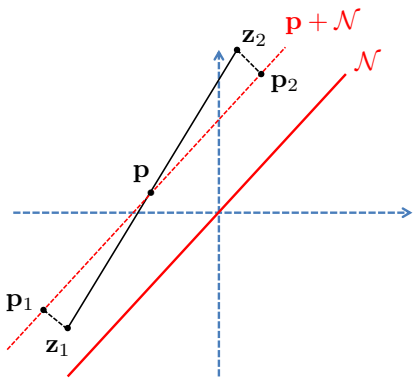


Fig. 4. Illustration of the impact of the correlation between $\mathcal{N}$ and $\Sigma-\Sigma$ on instance optimality. Here, $\mathbf{z}_1$ and $\mathbf{z}_2$ are two vectors in Σ such that $\mathbf{z}_1-\mathbf{z}_2$ is well correlated with $\mathcal{N}$, implying that at least one of the two vectors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

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 Δ be a decoder. If $\Delta(\mathbf{M}\mathbf{z}_1) \neq \mathbf{z}_1$, then this vector prevents Δ 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,

$$\frac{\|\mathbf{p}_i - \Delta(\mathbf{M}\mathbf{p}_i)\|_2}{d_2(\mathbf{p}_i,\Sigma)} \geq \frac{\|p_{\mathcal{N}}(\mathbf{z}_2-\mathbf{z}_1)\|_2}{\|p_{\mathcal{N}^\perp}(\mathbf{z}_2-\mathbf{z}_1)\|_2} \geq D\sqrt{1-\frac{1}{D^2}} = \sqrt{D^2-1}. \quad (35)$$

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.

C. ℓ^2/ℓ^2 IO with dimensionality reduction

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 is sufficiently correlated with every vector of V), then one cannot expect to get a ℓ^2/ℓ^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 ℓ^2/ℓ^2 IO decoder must have a constant at least $\frac{1}{\sqrt{1-K\left(1-\frac{m}{n}\right)}}$.

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 ℓ^2/ℓ^2 instance optimal decoder has constant at least $\sqrt{\frac{n}{m}}$.

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 ℓ^2/ℓ^2 IOP with dimensionality reduction is impossible. We provide here less trivial examples, where $E=V$ is finite-dimensional.

a) 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 $\mathbf{x}^T\mathbf{C}_{i,j}\mathbf{x} = 2(\|\mathbf{x}\|_2^2 - x_ix_j) \geq 0$, 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 Σ is a positive cone, multiplying this equality by $\frac{1}{\sqrt{2}}$ yields the desired result. □

b) Low-rank and nonsparse matrices: In [17], 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

$$\mathbf{e}_k = \frac{1}{\sqrt{n_1}}\left[1, e^{2i\pi k/n_1}, \ldots, e^{2i\pi(n_1-1)k/n_1}\right]^T$$
$$\text{and} \quad \mathbf{f}_\ell = \frac{1}{\sqrt{n_2}}\left[1, e^{2i\pi \ell/n_2}, \ldots, e^{2i\pi(n_2-1)\ell/n_2}\right]^T.$$

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 $\langle .,. \rangle$ the Hermitian scalar product on E,

$$\begin{aligned}
&\langle \mathbf{e}_k\mathbf{f}_\ell^*, \mathbf{e}_{k'}\mathbf{f}_{\ell'}^* \rangle \\
&= \sum_{u=0}^{n_1-1} \exp\left(2i\pi u \frac{k-k'}{n_1}\right) \sum_{v=0}^{n_2-1} \exp\left(2i\pi v \frac{\ell-\ell'}{n_2}\right) \\
&= \delta_k^{k'}\delta_\ell^{\ell'}, \qquad (36)
\end{aligned}$$

proving that these matrices form an orthonormal basis of E. □

IV. The NSP and its relationship with the RIP

As we have seen in the previous section, one cannot expect to get ℓ^2/ℓ^2 instance optimality in a dimensionality reduction context. This raises the following question: given pseudo-norms $\|.\|_G$ and $\|.\|_F$ defined respectively on E and F, is there a pseudo-norm $\|.\|_E$ such that IOP holds? We will see that this property is closely related to the RIP on $\mathbf{M}$.

A. 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 Σ_{2k}. The norms $\|.\|_G$ and $\|.\|_F$ are usually both taken as the ℓ^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 α, β such that

$$\forall \mathbf{z} \in \Sigma - \Sigma, \alpha\|\mathbf{z}\|_G \leq \|\mathbf{M}\mathbf{z}\|_F \leq \beta\|\mathbf{z}\|_G. \qquad (37)$$

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, as mentioned in Section I-C4.

Let's now suppose the existence of decoders robust to noise, that is for all $\delta > 0$, (21) is satisfied for a certain Δ_δ. This property implies the Robust NSP (22) with the same constants according to Theorem 3. By considering $\mathbf{h} \in \Sigma - \Sigma$, the Robust NSP reads:

$$\forall \mathbf{h} \in \Sigma - \Sigma, \|\mathbf{h}\|_G \leq D_2\|\mathbf{M}\mathbf{h}\|_F. \qquad (38)$$

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 Σ, which is therefore necessary if one seeks the existence of a decoder robust to noise.

B. 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 α, 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 $\|.\|_G$ and $\|.\|_M$, the latter being called "M-norm"[4] and involving $\|.\|_G$ and $\|.\|_F$.

Let's define the M-norm on E as the following quantity, extending its definition for ℓ^2 norms in [24] and its implicit appearance in the proof of early results of the field [4]:

$$\forall \mathbf{x} \in E, \|\mathbf{x}\|_M = \|\mathbf{x}\|_G + \frac{1}{\alpha}\|\mathbf{M}\mathbf{x}\|_F. \qquad (39)$$

Note that the term M-norm should be understood as M-pseudo-norm in the general case: if $\|.\|_F$ and $\|.\|_G$ satisfy the properties listed in Table I, then $\|.\|_M$ satisfies the same properties as $\|.\|_G$. However, when $\|.\|_G$ and $\|.\|_F$ are norms, $\|.\|_M$ is also a norm. We will note $d_M(.,.)$ its associated (pseudo-)distance. The following theorem states that this $\|.\|_M$ 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 α (left inequality of (37)). Then the following Robust NSP is satisfied:*

$$\forall \mathbf{h} \in E, \|\mathbf{h}\|_G \leq d_M(\mathbf{h}, \Sigma - \Sigma) + \frac{1}{\alpha}\|\mathbf{M}\mathbf{h}\|_F. \qquad (40)$$

In particular, the following regular NSP is satisfied:

$$\forall \mathbf{h} \in \mathcal{N}, \|\mathbf{h}\|_G \leq d_M(\mathbf{h}, \Sigma - \Sigma). \qquad (41)$$

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

$$\begin{aligned}
&\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \\
&\|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq 2d_M(\mathbf{x}, \Sigma) + \frac{2}{\alpha}\|\mathbf{e}\|_F + \delta. \qquad (42)
\end{aligned}$$

[4] to highlight its dependency on the Measurement operator

Note that in [11], the author explored the implication of a lower-RIP on $\Sigma - \Sigma$ for the case where Σ is an arbitrary UoS and $\|.\|_G/\|.\|_F$ are the ℓ^2 norm. He proved that this generalized lower-RIP implies the following IOP: for all $\delta > 0$, there exists a decoder Δ_δ such that $\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \forall \mathbf{z} \in \Sigma$,

$$\begin{aligned} &\|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_2 \\ &\leq \|\mathbf{x} - \mathbf{z}\|_2 + \frac{2}{\alpha}\|\mathbf{M}(\mathbf{x} - \mathbf{z}) + \mathbf{e}\|_2 + \delta. \end{aligned} \tag{43}$$

In this set-up, the instance optimality in equation (42) can be reformulated as $\forall \mathbf{x} \in E, \forall \mathbf{e} \in F, \forall \mathbf{z} \in \Sigma$,

$$\begin{aligned} &\|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_2 \\ &\leq 2\|\mathbf{x} - \mathbf{z}\|_2 + \frac{2}{\alpha}\|\mathbf{M}(\mathbf{x} - \mathbf{z})\|_2 + \frac{2}{\alpha}\|\mathbf{e}\|_2 + \delta. \end{aligned} \tag{44}$$

Comparing these two instance optimality results, we can remark that the one in [11] 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 [11] can also be derived in our case with the same proof layout if we suppose the lower-RIP. Compared to [11], our theory deals with general (pseudo-)norms and sets Σ beyond Union of Subspaces.

C. Infinite-dimensional examples

As mentioned in the introduction, we do not constrain the signal space to be finite-dimensional, so that we can apply our results in an infinite-dimensional framework.

1) Negative example: Sparse model in a separable Hilbert space: If $E = L^2([0,1])$ and $\{\varphi_n\}_{n\in\mathbb{N}}$ is an orthonormal basis of E, a typical measurement process of a signal $\mathbf{x}$ is to subsample along another orthonormal basis $\{\Psi_n\}_{n\in\mathbb{N}}$. Typically, $\{\varphi_n\}$ is a wavelet basis and $\{\Psi_n\}$ a Fourier basis. In [25], the authors argue that standard sparsity does not represent well natural signals, which are rather *asymptotically sparse*, that is more sparse at fine levels than at coarse levels. They propose an asymptotic sparsity model with different levels of sparsity on different scales.

Indeed, as the authors mention in their paper, a standard sparsity model Σ with respect to basis $\{\varphi_n\}$ cannot yield uniform recovery for the L^2 norm: this is an obvious consequence of Theorem 6 if d_E is the L^2 distance and it is actually true for any other distance. Indeed, the NSP can never be satisfied since one has $\|\varphi_n + \varphi_{n+1}\|_2 = \sqrt{2}$ for all n (since the family $\{\varphi_n\}$ is orthonormal) while the right hand side term of the NSP for $\mathbf{h}_n = \varphi_n + \varphi_{n+1}$ is equal to

$$d(\mathbf{h}_n, \Sigma - \Sigma) + \|\mathbf{M}\mathbf{h}_n\|_2 = \|\mathbf{M}\mathbf{h}_n\|_2, \tag{45}$$

which goes to 0 when $n \to \infty$ (since $\mathbf{M}$ is continuous).

2) Positive example: Topological RIP result for Σ of finite box-counting dimension: Even though the IOP cannot be satisfied for the standard sparse model in a Hilbert space, it does not mean IOP is impossible for all models in an infinite-dimensional space. Let's mention the following topological result, which is Theorem 8.1 in [34] and ensures that a RIP is satisfied in some settings:

Theorem 8. *Let Σ be a compact subset of a Banach space $\mathcal{B}$ supplied with norm $\|.\|_\mathcal{B}$. Suppose Σ has finite (upper) box-counting dimension d. Then for any $m > 2d$, any norm $\|.\|$ on $\mathbb{R}^m$, and any θ satisfying*

$$0 < \theta < \frac{m - 2d}{m(1 + d)},$$

there exists a prevalent set[5] of continuous linear operators $\mathbf{M} : \mathcal{B} \to \mathbb{R}^m$ such that for any $\mathbf{x}, \mathbf{y} \in \Sigma$,

$$C_\mathbf{M}\|\mathbf{x} - \mathbf{y}\|_\mathcal{B} \leq \|\mathbf{M}\mathbf{x} - \mathbf{M}\mathbf{y}\|^\theta. \tag{46}$$

This theorem essentially says that if Σ has finite upper box-counting dimension[6], then a lower-RIP is satisfied for a prevalent set of operators $\mathbf{M}$ with the pseudo-norms $\|.\|_\mathcal{B}$ and $\|.\|^\theta$ – this last one being a pseudo-norm since $\theta \leq 1$. According to the previous section, this implies an IOP with the corresponding M-norm, that is the existence of a family of decoders Δ_δ such that for all $\mathbf{x} \in E$, $\mathbf{e} \in F$ and $\mathbf{z} \in \Sigma$:

$$\|\mathbf{x} - \Delta(\mathbf{M}\mathbf{x})\|_\mathcal{B} \leq 2d_M(\mathbf{x}, \Sigma) + \frac{2}{C_\mathbf{M}}\|\mathbf{e}\|^\theta + \delta. \tag{47}$$

We necessarily have $\theta \leq \frac{1}{d}$, meaning the exponent drops to 0 as d grows, and therefore the corresponding IOP becomes much less powerful with a high-dimensional set Σ (in the sense of the upper box-counting dimension). However, this essentially proves that uniform instance optimality is possible even in infinite dimensions with appropriate Σ and pseudo-norms. Furthermore, weakening the existence of a *prevalent* set of operators $\mathbf{M}$ to the existence of *an* operator $\mathbf{M}$ or a *certain class* of such operators satisfying a Robust IOP has the potential to yield IOP with better pseudo-norms.

As an example of an infinite-dimensional model with finite upper box-counting dimension, let's consider the problem experimented in [26]: we consider $E = L^1(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ and aim at decoding a probability density $p \in E$ from a linear measurement $\mathbf{M}p$. The *a priori* on p is that it can be expressed as a linear combination of a few densities taken in a set $\mathcal{P}$. In [26], the authors considered $\mathcal{P}$ as a set of isotropic Gaussians, that is

$$\mathcal{P} = \left\{p_{\boldsymbol{\mu}} : \mathbf{x} \to \exp\left(-\|\mathbf{x} - \boldsymbol{\mu}\|_2^2\right) | \boldsymbol{\mu} \in \mathbb{R}^n\right\}. \tag{48}$$

Denoting $\Sigma_k(\mathcal{P})$ the compact set of convex linear combinations of k elements in $\mathcal{P}$ with $\|\boldsymbol{\mu}\|_2 \leq C$, the upper-box counting dimension of $\Sigma_k(\mathcal{P})$ is upper bounded by $k(n+1)$, so that a prevalent set of linear operators satisfies the IOP (47) as soon as the number of measurements satisfies $m > 2k(n+1)$.

This example shows that one can obtain uniform IOP for an infinite-dimensional model which "spans in an infinite number of directions", such as the aforementioned model $\Sigma_k(\mathcal{P})$. We hope that more precise characterizations on this kind of IOP can be obtained in this general framework.

[5] A prevalent set being a set which complementary is negligible in a certain sense. Definition is given in [34].

[6] This is a notion of dimension defined by asymptotic behavior of ϵ-covers.

D. 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 $\|.\|_E$. 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 Σ' (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 [5]. 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 $\|.\|_G$ is a norm.

1) The atomic norm $\|.\|_{\Sigma'}$: Let Σ' be a subset of E and let E' be the closure of $\mathrm{span}(\Sigma')$ with respect to the norm $\|.\|_G$. For $\mathbf{x} \in E'$, one can define the "norm" $\|\mathbf{x}\|_{\Sigma'}$ by:

$$\|\mathbf{x}\|_{\Sigma'} := \inf \left\{ \sum_{k=0}^{+\infty} \|\mathbf{x}_k\|_G : \forall k, \mathbf{x}_k \in \mathbb{R}\Sigma' \right. \\ \left. \text{and } \|\mathbf{x} - \sum_{k=0}^{K} \mathbf{x}_k\|_G \to_{K\to+\infty} 0 \right\}. \quad (49)$$

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 $V = \{\mathbf{x} \in E | \|\mathbf{x}\|_{\Sigma'} < +\infty\}$ is a normed subspace of E which contains Σ' [39]. In the following, we assume that $V = E$. Note that this norm can be linked to atomic norms defined in [5] by considering $\mathcal{A}$ as the set of normalized elements of Σ' with respect to $\|.\|_G$.

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

$$\forall \mathbf{x}' \in \Sigma', \|\mathbf{M}\mathbf{x}'\|_F \leq \beta \|\mathbf{x}'\|_G. \quad (50)$$

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,

$$\forall \mathbf{x} \in E, \|\mathbf{x}\|_M \leq \|\mathbf{x}\|_G + \frac{\beta}{\alpha}\|\mathbf{x}\|_{\Sigma'} \leq \left(1 + \frac{\beta}{\alpha}\right) \|\mathbf{x}\|_{\Sigma'}. \quad (51)$$

In particular, we have the following result:

Theorem 9. *Suppose $\mathbf{M}$ satisfies the lower-RIP on $\Sigma - \Sigma$ with constant α and the upper-RIP on Σ with constant β, that is*

$$\forall \mathbf{x} \in \Sigma - \Sigma, \alpha \|\mathbf{x}\|_G \leq \|\mathbf{M}\mathbf{x}\|_F \quad (52)$$

and

$$\forall \mathbf{x} \in \Sigma, \|\mathbf{M}\mathbf{x}\|_F \leq \beta \|\mathbf{x}\|_G. \quad (53)$$

Then for all $\delta > 0$, there exists a decoder Δ_δ satisfying $\forall \mathbf{x} \in E, \forall \mathbf{e} \in F$,

$$\|\mathbf{x} - \Delta_\delta(\mathbf{M}\mathbf{x} + \mathbf{e})\|_G \leq 2\left(1 + \frac{\beta}{\alpha}\right) d_\Sigma(\mathbf{x}, \Sigma) + \frac{2}{\alpha}\|\mathbf{e}\|_E + \delta, \quad (54)$$

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

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

2) Study of $\|.\|_\Sigma$ in two usual cases: We now provide a more thorough analysis of the norm $\|.\|_\Sigma$ 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 $\|.\|_G = \|.\|_2$ (for matrices, this is the Frobenius norm).

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

- In the case where Σ 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 Σ 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 $\|.\|_*$ is the trace norm, defined as the sum of singular values. We can therefore, in both cases, upper bound the norm $\|.\|_\Sigma$. In the case of k-sparse vectors, this gives:

$$\|\mathbf{x}\|_\Sigma \leq \|\mathbf{x}_1\|_2 + \sum_{j\geq 1} \frac{\|\mathbf{x}_j\|_1}{\sqrt{k}} \leq \|\mathbf{x}\|_2 + \frac{\|\mathbf{x}\|_1}{\sqrt{k}}. \quad (55)$$

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

$$\|\mathbf{N}\|_\Sigma \leq \|\mathbf{N}_1\|_2 + \sum_{j\geq 1} \frac{\|\mathbf{N}_j\|_1}{\sqrt{k}} \leq \|\mathbf{N}\|_F + \frac{\|\mathbf{N}\|_*}{\sqrt{k}}. \quad (56)$$

We can also upper bound the right hand side of these equations by $\mathcal{O}(\|.\|_\Sigma)$ 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

$$\frac{\|\mathbf{x}\|_1}{\sqrt{k}} \leq \sum_{j\geq 1} \frac{\|\mathbf{x}_j\|_1}{\sqrt{k}} \leq \sum_{j\geq 1} \|\mathbf{x}_j\|_2 \quad (57)$$

(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,

$$\frac{\|\mathbf{N}\|_*}{\sqrt{k}} \leq \sum_{j\geq 1} \|\mathbf{N}_j\|_F. \quad (58)$$

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

$$\|\mathbf{x}\|_2 + \frac{\|\mathbf{x}\|_1}{\sqrt{k}} \leq 2\|\mathbf{x}\|_\Sigma \quad \text{and} \quad \|\mathbf{N}\|_F + \frac{\|\mathbf{N}\|_*}{\sqrt{k}} \leq 2\|\mathbf{N}\|_\Sigma. \quad (59)$$

We have thus shown:

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

$$\|\cdot\|_\Sigma \leq \|\cdot\|_2 + \frac{\|\cdot\|_1}{\sqrt{k}} \leq 2\|\cdot\|_\Sigma. \tag{60}$$

When Σ is the set of rank-k matrices, it satisfies

$$\|\cdot\|_\Sigma \leq \|\cdot\|_F + \frac{\|\cdot\|_*}{\sqrt{k}} \leq 2\|\cdot\|_\Sigma. \tag{61}$$

We can thus remark that for these two standard models, the norm $\|.\|_\Sigma$ 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 [4], the author proves that under upper and lower-RIP on $\Sigma - \Sigma$ with Σ being the set of k-sparse vectors, an instance optimal decoder can be defined as the minimization of a convex objective: the ℓ^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 $\|\cdot\|_\Sigma$?

V. 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.

a) Condition for the well-posedness of the "optimal" decoder: We have seen that for general models Σ, an additionnal term δ appears in the right hand side term of the instance optimality inequality ((15),(21)), reflecting the fact that the minimal distance of the "optimal" decoder (69) 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 δ and get a more usual instance optimality result.

b) Compressed graphical models: As has been mentioned in Section I-C3, the case where Σ 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 ℓ^2/ℓ^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?

c) 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 [40] 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 [41], 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 ∇ is the gradient operator, Σ_k the k-sparse union of subspaces (in the gradient space) and d_1 is the ℓ^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?

d) 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 [42]. 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 $\mathbf{Ax}$, 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 $\mathbf{Mx}$, 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 Δ?

Note that this classification-oriented set-up might call for a specific norm $\|.\|_G$ 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 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 Δ, hence in a privacy-preserving fashion?

e) 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 Σ is a point cloud mostly concentrated along the line $\mathcal{D}$ 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 [7]: given Ω a probability space with probability measure P, and considering $\mathbf{M}$ as a random variable on Ω, is there a decoder $\Delta(.|\mathbf{M})$ (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

$$\|\mathbf{x} - \Delta(\mathbf{Mx}|\mathbf{M})\|_G \leq C d_E(\mathbf{x}, \Sigma) \tag{62}$$

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 [43].

APPENDIX A
WELL-POSEDNESS OF THE FINITE UOS DECODER

In this section, we will prove that if Σ is a finite union of subspaces in $\mathbb{R}^n$ and $\|.\|$ 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 $\|.\|$, 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 $\|.\|$ 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 $\|.\|$.*

Proof. Let Φ be defined on $V+W$ by $\Phi(\mathbf{u}) = \|\mathbf{u}-\mathbf{x}\|$. Since $\Phi(\mathbf{u}) \geq \|\mathbf{u}\| - \|\mathbf{x}\|$, we have $\lim_{\|\mathbf{u}\|\to+\infty} \Phi(\mathbf{u}) = +\infty$, 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 Φ is continuous, Φ 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 Φ.

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. □

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, $\Delta(\mathbf{Mx})$ can be defined as $\underset{\{\mathbf{x}_i, i\in[\![1,p]\!]\}}{\operatorname{argmin}} d_E(\mathbf{x}_i, V_i)$ and satisfies $d_E(\Delta(\mathbf{Mx}), \Sigma) = d_E(\mathbf{x}+\mathcal{N}, \Sigma)$, so that the decoder $\Delta(\mathbf{Mx}) = \underset{\mathbf{z}\in(\mathbf{x}+\mathcal{N})}{\operatorname{argmin}} d(\mathbf{z}, \Sigma)$ is properly defined. In particular, this applies to the decoder (12).

APPENDIX B
PROOF OF THEOREM 1

Let $\delta > 0$ and Δ_δ and C be such that (15) holds $\forall \mathbf{x} \in E$. Let $\mathbf{h} \in \mathcal{N}$. Then $\exists \mathbf{h}_0 \in \Sigma - \Sigma$ such that $d_E(\mathbf{h}, \mathbf{h}_0) \leq$

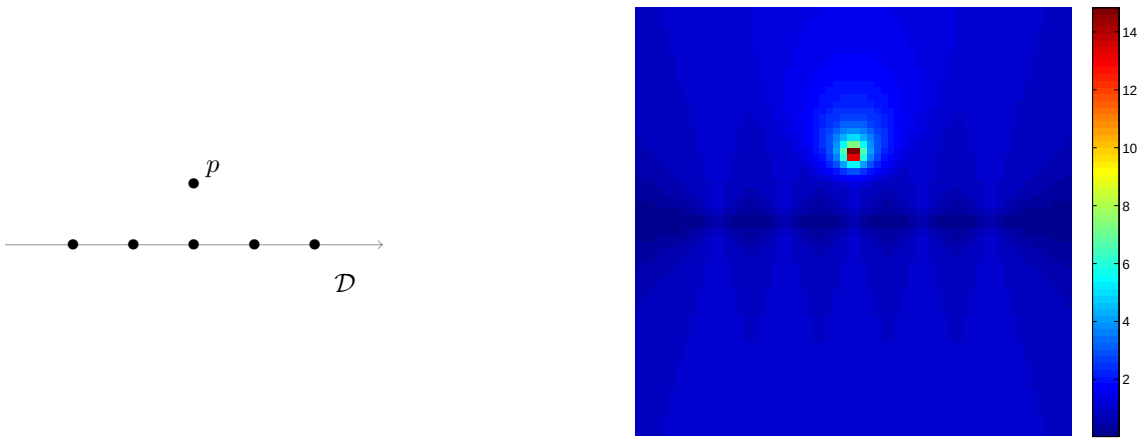


Fig. 5. Drawback of uniform instance optimality in a simple case: the model Σ (*Left*) is the set of black points including those on D and the point p and the operator $\mathbf{M}$ is the 1-dimensional orthogonal projection on the horizontal axis. If we choose Δ as the pseudo-inverse of $\mathbf{M}$,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

$d_E(\mathbf{h}, \Sigma - \Sigma) + \delta$. 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:

$$\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3) = \mathbf{M}\mathbf{h}_2. \tag{63}$$

Applying (15) to $\mathbf{x} = \mathbf{h}_2 \in \Sigma$ and using the fact that $\|0\|_E = 0$, we get:

$$\|\mathbf{A}\mathbf{h}_2 - \Delta_\delta(\mathbf{M}\mathbf{h}_2)\|_G \leq \delta. \tag{64}$$

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

$$\begin{aligned}
&\|\mathbf{A}\mathbf{h}\|_G = \|\mathbf{A}(\mathbf{h}_1 - \mathbf{h}_2 + \mathbf{h}_3)\|_G \\
&= \|\mathbf{A}(\mathbf{h}_1 + \mathbf{h}_3) - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3)) \\
&\quad - \mathbf{A}\mathbf{h}_2 + \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G \\
&\leq \|\mathbf{A}(\mathbf{h}_1 + \mathbf{h}_3) - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G \\
&\quad + \|\mathbf{A}\mathbf{h}_2 - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G,
\end{aligned} \tag{65}$$

where we have used (13) and (14) for the last inequality. Combining (63) and (64), we get that:

$$\|\mathbf{A}\mathbf{h}_2 - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G \leq \delta. \tag{66}$$

Applying (15) to $\mathbf{x} = \mathbf{h}_1 + \mathbf{h}_3$, we get:

$$\begin{aligned}
&\|\mathbf{A}(\mathbf{h}_1 + \mathbf{h}_3) - \Delta_\delta(\mathbf{M}(\mathbf{h}_1 + \mathbf{h}_3))\|_G \\
&\leq C d_E(\mathbf{h}_1 + \mathbf{h}_3, \Sigma) + \delta \leq C\|\mathbf{h}_3\|_E + \delta \\
&= C d_E(\mathbf{h}, \mathbf{h}_0) + \delta \leq C d_E(\mathbf{h}, \Sigma - \Sigma) + (C+1)\delta.
\end{aligned} \tag{67}$$

Combining (65), (66) and (67) gives:

$$\|\mathbf{A}\mathbf{h}\|_G \leq C d_E(\mathbf{h}, \Sigma - \Sigma) + (C+2)\delta. \tag{68}$$

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

Appendix C
Proof of Theorem 2

Let's first assume that (16) holds and define the following decoder on F:

$$\Delta'(\mathbf{M}\mathbf{x}) = \operatorname*{argmin}_{\mathbf{z} \in (\mathbf{x} + \mathcal{N})} d_E(\mathbf{z}, \Sigma). \tag{69}$$

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:

$$\begin{aligned}
\|\mathbf{A}\mathbf{x} - \mathbf{A}\Delta'(\mathbf{M}\mathbf{x})\|_G &\leq D d_E(\mathbf{x} - \Delta'(\mathbf{M}\mathbf{x}), \Sigma - \Sigma) \\
&\leq D d_E(\mathbf{x}, \Sigma) + D d_E(\Delta'(\mathbf{M}\mathbf{x}), \Sigma) \\
&\leq 2D d_E(\mathbf{x}, \Sigma),
\end{aligned} \tag{70}$$

where we have used (14) for the second inequality. The last inequality comes from (69), 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:

$$\Delta'_\nu(\mathbf{M}\mathbf{x}) \in \{\mathbf{u} \in (\mathbf{x} + \mathcal{N}) | d_E(\mathbf{u}, \Sigma) \leq d_E(\mathbf{x} + \mathcal{N}, \Sigma) + \nu\}. \tag{71}$$

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):

$$\begin{aligned}
\|\mathbf{A}\mathbf{x} - \mathbf{A}\Delta'_\nu(\mathbf{M}\mathbf{x})\|_G &\leq D d_E(\mathbf{x} - \Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma - \Sigma) \\
&\leq D d_E(\mathbf{x}, \Sigma) + D d_E(\Delta'_\nu(\mathbf{M}\mathbf{x}), \Sigma) \\
&\leq 2D d_E(\mathbf{x}, \Sigma) + D\nu,
\end{aligned} \tag{72}$$

where we have used (14) again for the second inequality. The last inequality comes from (71), 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 (15).

Appendix 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 $\Delta'_\nu(\mathbf{M}\mathbf{x}) \in \{\mathbf{u} \in (\mathbf{x} + \mathcal{N}) | d_E(\mathbf{u}, \Sigma) \leq (1+\nu) d_E(\mathbf{x} + \mathcal{N}, \Sigma)\}$. This provides a consistent definition of Δ'_ν.

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

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

$$\begin{aligned}\|\mathbf{Ax} - \mathbf{A}\Delta'_\nu(\mathbf{Mx})\|_G &\leq D d_E(\mathbf{x} - \Delta'_\nu(\mathbf{Mx}), \Sigma - \Sigma) \\ &\leq D d_E(\mathbf{x}, \Sigma) + D d_E(\Delta'_\nu(\mathbf{Mx}), \Sigma) \\ &\leq (2+\nu) D d_E(\mathbf{x}, \Sigma). \end{aligned} \tag{73}$$

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

Appendix E
Proof of Theorem 3 and Theorem 5

Let's first remark that applying (21) (resp. (23)) 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 (21) (resp. (23)) with $\mathbf{x} = \mathbf{z} - \mathbf{h}$, $\mathbf{e} = \mathbf{Mh}$, and $\epsilon = \|\mathbf{Mh}\|_F$, which yields:

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

Let's remark that (13) and (14) 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:

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

This last inequality is valid for all $\mathbf{z} \in \Sigma$, therefore (21) implies:

$$\begin{aligned}\|\mathbf{Ah}\|_G &\leq C_1 \inf_{\mathbf{z}\in\Sigma} \; d_E(\mathbf{z}-\mathbf{h}, \Sigma) + C_2\|\mathbf{Mh}\|_F + 2\delta \\ &= C_1 \inf_{\mathbf{z}\in\Sigma} \; \inf_{\mathbf{u}\in\Sigma} \; \|\mathbf{z}-\mathbf{h}-\mathbf{u}\|_E + C_2\|\mathbf{Mh}\|_F + 2\delta \\ &= C_1 d_E(\mathbf{h}, \Sigma - \Sigma) + C_2\|\mathbf{Mh}\|_F + 2\delta, \end{aligned} \tag{74}$$

where we have used (13) for the last inequality. Similarly, (23) implies

$$\|\mathbf{Ah}\|_G \leq C_1 d_E(\mathbf{h}, \Sigma - \Sigma) + 2C_2\|\mathbf{Mh}\|_F + 2\delta. \tag{75}$$

We conclude by using the fact that (74) and (75) hold for all $\delta > 0$.

Appendix F
Proof of Theorem 4

Let's suppose (22) and define for $\delta > 0$ the decoder $\Delta'_\delta : F \to E$ such that $\forall \mathbf{y} \in F$:

$$\begin{aligned} &D_1 d_E(\Delta'_\delta(\mathbf{y}), \Sigma) + D_2 d_F(\mathbf{M}\Delta'_\delta(\mathbf{y}), \mathbf{y}) \\ &\leq \inf_{\mathbf{u}\in E} \; [D_1 d_E(\mathbf{u}, \Sigma) + D_2 d_F(\mathbf{Mu}, \mathbf{y})] + \delta. \end{aligned} \tag{76}$$

Let's prove that this decoder meets property (21).

Let $\mathbf{x} \in E$ and $\mathbf{e} \in F$. Applying (22) with $\mathbf{h} = \mathbf{x} - \Delta'_\delta(\mathbf{Mx} + \mathbf{e})$, we get:

$$\begin{aligned} &\|\mathbf{A}(\mathbf{x} - \Delta'_\delta(\mathbf{Mx}+\mathbf{e}))\|_G \\ &\leq D_1 d_E(\mathbf{x} - \Delta'_\delta(\mathbf{Mx}+\mathbf{e}), \Sigma - \Sigma) \\ &\quad + D_2\|\mathbf{M}(\mathbf{x} - \Delta'_\delta(\mathbf{Mx}+\mathbf{e}))\|_F \\ &\leq D_1 d_E(\mathbf{x}, \Sigma) + D_1 d_E(\Delta'_\delta(\mathbf{Mx}+\mathbf{e}), \Sigma) \\ &\quad + D_2 d_F(\mathbf{M}\Delta'_\delta(\mathbf{Mx}+\mathbf{e}), \mathbf{Mx}+\mathbf{e}) + D_2\|\mathbf{e}\|_F \\ &\leq 2D_1 d_E(\mathbf{x}, \Sigma) + 2D_2\|\mathbf{e}\|_F + \delta, \end{aligned} \tag{77}$$

where we have used (13) and (14) for the second inequality and the last inequality is a consequence of (76).

Posing $\Delta_\delta = \mathbf{A}\Delta'_\delta$ proves (21) with $C_1 = 2D_1$ and $C_2 = 2D_2$.

Appendix G
Proof of Lemma 1

The two equivalences are very similar to prove, so that we will only prove the first. (28) $\Rightarrow$ (27) is obvious. Let's now suppose (27), so that:

$$\forall \mathbf{h} \in \mathcal{N}, \forall \mathbf{z} \in \Sigma - \Sigma, \|\mathbf{h}\|_G \leq D\|\mathbf{h}-\mathbf{z}\|_E. \tag{78}$$

By homogeneity, we also have:

$$\forall \lambda \in \mathbb{R}^*, \forall \mathbf{h} \in \mathcal{N}, \forall \mathbf{z} \in \Sigma - \Sigma, \|\lambda\mathbf{h}\|_G \leq D\|\lambda\mathbf{h} - \mathbf{z}\|_E, \tag{79}$$

so that:

$$\forall \lambda \in \mathbb{R}^*, \forall \mathbf{h} \in \mathcal{N}, \forall \mathbf{z} \in \Sigma - \Sigma, \|\mathbf{h}\|_G \leq D\|\mathbf{h} - \mathbf{z}/\lambda\|_E. \tag{80}$$

This last inequality yields (28).

Appendix 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 $\mathbf{h}_1, \ldots, \mathbf{h}_{n-m}$ be an orthonormal basis of $\widetilde{\mathcal{N}}$. We have:

$$n - m = \sum_{j=1}^{n-m} \|\mathbf{h}_j\|_2^2 \leq \frac{1}{K} \sum_{j=1}^{n-m} \sum_{i=1}^{n} \langle \mathbf{h}_j, \mathbf{z}_i \rangle^2. \tag{81}$$

Using (34), 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 (81), we get:

$$n - m \leq \frac{n}{K}\left(1 - \frac{1}{D_*^2}\right), \tag{82}$$

so that:

$$m \geq n\left(1 - \frac{1}{K}\left(1 - \frac{1}{D_*^2}\right)\right). \tag{83}$$

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

APPENDIX I
PROOF OF THEOREM 7

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

$$\|\mathbf{h}\|_G \leq \|\mathbf{h}-\mathbf{z}\|_G + \|\mathbf{z}\|_G \leq \|\mathbf{h}-\mathbf{z}\|_G + \frac{1}{\alpha}\|\mathbf{M}\mathbf{z}\|_F, \quad (84)$$

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

A similar consideration on $\mathbf{M}\mathbf{z}$ yields:

$$\|\mathbf{M}\mathbf{z}\|_F \leq \|\mathbf{M}(\mathbf{z}-\mathbf{h})\|_F + \|\mathbf{M}\mathbf{h}\|_F. \quad (85)$$

Substituting (85) into (84), we get:

$$\begin{aligned}\|\mathbf{h}\|_G &\leq \|\mathbf{h}-\mathbf{z}\|_G + \frac{1}{\alpha}\|\mathbf{M}(\mathbf{h}-\mathbf{z})\|_F + \frac{1}{\alpha}\|\mathbf{M}\mathbf{h}\|_F \\ &= \|\mathbf{h}-\mathbf{z}\|_M + \frac{1}{\alpha}\|\mathbf{M}\mathbf{h}\|_F. \quad (86)\end{aligned}$$

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

$$\|\mathbf{h}\|_G \leq d_M(\mathbf{h}, \Sigma - \Sigma) + \frac{1}{\alpha}\|\mathbf{M}\mathbf{h}\|_F. \quad (87)$$

ACKNOWLEDGMENT

This work was supported in part by the European Research Council, PLEASE project (ERC-StG-2011-277906). The authors also wants to thank the anonymous reviewers for their remarks about instance optimality in infinite dimensions and for providing the example given in Section IV-C1.

REFERENCES

[1] 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*, vol. 52, no. 2, pp. 489–509, 2006.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, 2006.

[3] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[4] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Acad. Sci. Paris S'er. I Math.*, vol. 346, pp. 589–592, 2008.

[5] V. Chandrasekaran, B. Recht, P. Parrilo, and A. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[6] H. W. Engl, M. Hanke, and A. Neubbauer, *Regularization of Inverse Problems*. Springer, 1996.

[7] A. Cohen, W. Dahmen, and R. Devore, "Compressed sensing and best k-term approximation," *J. Amer. Math. Soc*, pp. 211–231, 2009.

[8] 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*, vol. 98, no. 6, pp. 959–971, 2010.

[9] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 3042–3054, 2010.

[10] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[11] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4660–4671, 2011.

[12] H. Rauhut, K. Schnass, and P. Vandergheynst, "Compressed sensing and redundant dictionaries," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2210–2219, 2008.

[13] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, "The Cosparse Analysis Model and Algorithms," *Applied and Computational Harmonic Analysis*, vol. 34, no. 1, pp. 30–56, 2013.

[14] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.

[15] 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.*, vol. 57, no. 4, pp. 2342–2359, April 2011. [Online]. Available: http://dx.doi.org/10.1109/TIT.2011.2111771

[16] Z. Zhou, X. Li, J. Wright, E. J. Candès, and Y. Ma, "Stable principal component pursuit," in *ISIT*, 2010, pp. 1518–1522.

[17] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, p. 11, 2011.

[18] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, 2007.

[19] M. Yuan, "High dimensional inverse covariance matrix estimation via linear programming," *Journal of Machine Learning Research*, vol. 11, pp. 2261–2286, 2010.

[20] R. G. Baraniuk and M. B. Wakin, "Random projections of smooth manifolds," in *Foundations of Computational Mathematics*, 2006, pp. 941–944.

[21] A. Eftekhari and M. B. Wakin, "New analysis of manifold embeddings and signal recovery from compressive measurements," *CoRR*, vol. abs/1306.4748, 2013.

[22] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," in *Conference in modern analysis and probability (New Haven, Conn., 1982)*, ser. Contemporary Mathematics. American Mathematical Society, 1984, vol. 26, pp. 189–206.

[23] D. Achlioptas, "Database-friendly random projections," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2001, pp. 274–281.

[24] T. Peleg, R. Gribonval, and M. Davies, "Compressed sensing and best approximation from union of subspaces: Beyond dictionaries," in *EUSIPCO*, 2013.

[25] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: A new theory for compressed sensing," *arXiv*, pp. 1–44, February 2014.

[26] A. Bourrier, R. Gribonval, and P. Pérez, "Compressive Gaussian Mixture Estimation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2013.

[27] S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi, "Simultaneously structured models with application to sparse and low-rank matrices," *CoRR*, vol. abs/1212.3753, 2012.

[28] H. Ohlsson, A. Y. Yang, and S. S. Sastry, "Compressive phase retrieval from squared output measurements via semidefinite programming," *CoRR*, vol. abs/1111.6323, 2011.

[29] E. J. Candès, T. Strohmer, and V. Voroninski, "Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming," *CoRR*, vol. abs/1109.4499, 2011.

[30] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Applied and Computational Harmonic Analysis*, vol. 31, no. 1, pp. 59–73, 2011.

[31] R. Giryes, S. Nam, M. Elad, R. Gribonval, and M. E. Davies, "Greedy-Like Algorithms for the Cosparse Analysis Model," January 2013, partially funded by the ERC, PLEASE project, ERC-2011-StG-277906. [Online]. Available: http://hal.inria.fr/hal-00716593

[32] B. Adcock and A. C. Hansen, "A generalized sampling theorem for stable reconstructions in arbitrary bases," *J. Fourier Anal. Appl.*, vol. 18, no. 4, pp. 685–716, November 2012.

[33] ——, "Generalized sampling and infinite-dimensional compressed sensing," *DAMTP Tech. Rep.*, 2011.

[34] J. C. Robinson, *Dimensions, Embeddings, and Attractors*, ser. Cambridge Tracts in Mathematics. Leiden: Cambridge University Press, 2010.

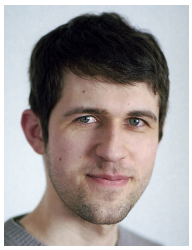
[35] R. Gribonval and M. Nielsen, "Highly sparse representations from dictionaries are unique and independent of the sparseness measure," *Appl. Comp. Harm. Anal.*, vol. 22, no. 3, pp. 335–355, 2007.

[36] P. Wojtaszczyk, "Stability and instance optimality for gaussian measurements in compressed sensing," *Foundations of Computational Mathematics*, vol. 10, no. 1, pp. 1–13, 2010.

[37] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*, ser. Applied and Numerical Harmonic Analysis. Springer, 2013.

[38] S. Foucart, "Hard thresholding pursuit: An algorithm for compressive sensing," *SIAM J. Numerical Analysis*, vol. 49, pp. 2543–2563, 2011.

[39] R. A. DeVore and V. N. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comp. Math.*, vol. 5, no. 2-3, pp. 173–187, 1996.
[40] M. A. Davenport, D. Needell, and M. B. Wakin, "Signal space cosamp for sparse recovery with redundant dictionaries," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6820–6829, 2013.
[41] D. Needell and R. Ward, "Stable image reconstruction using total variation minimization," *SIAM J. Imaging Sciences*, vol. 6, no. 2, pp. 1035–1058, 2013.
[42] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Computer Vision*, vol. 105, no. 3, pp. 22–245, 2013.
[43] G. Yu and G. Sapiro, "Statistical compressed sensing of gaussian mixture models," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5842–5858, 2011.



Anthony Bourrier received the engineering degree from École Centrale (Paris, France) with a specialization in applied mathematics and the M.Sc. in Mathematics, Computer Vision and Machine Learning from École Normale Supérieure (Cachan, France) in 2010. He received the Ph.D. degree in signal processing from University of Rennes I (Rennes, France) in 2014. Since 2014, he is a postdoctoral fellow in Gipsa-Lab (Grenoble, France). His main research interests include compressed sensing, linear inverse problems and information theory.

Mike Davies (M'00-SM'11) holds the Jeffrey Collins Chair in Signal and Image Processing at University of Edinburgh, where he is Head if the Institute of Digital Communicatiosn and Director of the Joint Research Institute in Signal and Image Processing, a the collaborative research venture between the University of Edinburgh and Heriot-Watt University. He received an M.A. in engineering from Cambridge University in 1989 where he was awarded a Foundation Scholarship (1987), and a Ph.D. degree in nonlinear dynamics and signal processing from University College London (UCL) in 1993. He was awarded a Royal Society University Research Fellowship in 1993 and was appointed a Texas Instruments Distinguished Visiting Professor at Rice University in 2012. He acted as an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING during 2003–2007. Currently he leads the University Defence Research Collaboration (UDRC), a UK programme of signal processing research in defence in collaboration with the UK Defence Science and Technology Laboratory (DSTL).

His research has focused on nonlinear time series, source separation, sparse representations and compressed sensing. He has made key contributions to these fields including: the development of signal embedding theorems for nonlinear dynamical time series and the development of new theoretical and algorithmic results in Independent Component Analysis (ICA). Most recently he has pioneered the use of sparse representations as a fundamental tool in signal processing, source separation and compressed sensing. This work includes: the proposal and analysis of the highly popular Iterative Hard Thresholding algorithm for sparse reconstruction. He has further applied these ideas to advanced medical imaging and RF based sensing applications. He also has an active interest in the related topics of machine learning, high-dimensional statistics and information theory.

Tomer Peleg received the B.Sc. degree in electrical engineering (*summa cum laude*) and the B.Sc. degree in physics (*summa cum laude*) both from the Technion, Haifa, Israel, in 2009. He is currently pursuing the Ph.D. degree in electrical engineering at the Technion. From 2007 to 2009, he worked at RAFAEL Research Laboratories, Israel Ministry of Defense. His research interests include statistical signal processing, sparse representations, image processing, inverse problems and graphical models. Since 2012 he holds a fellowship in the Azrieli program for outstanding Israeli graduate students.

Patrick Pérez received the Ph.D. degree from the University of Rennes in 1993 and joined INRIA in 1994 as a full time researcher. From 2000 to 2004, he was with Microsoft Research Cambridge. In 2009, he joined Technicolor as a Distinguished Scientist and Fellow. He is currently a member of the editorial board of the International Journal of Computer Vision. His research interests include image description, search and analysis, as well as photo/video editing and computational imaging.

Rémi Gribonval (FM'14) is a Senior Researcher with Inria (Rennes, France), and the scientific leader of the PANAMA research group on sparse audio processing. A former student at École Normale Supérieure (Paris, France), he received the Ph.D. degree in applied mathematics from Université de Paris-IX Dauphine (Paris, France) in 1999, and the Habilitation à Diriger des Recherches in applied mathematics from Université de Rennes I (Rennes, France) in 2007. His research focuses on mathematical signal processing, machine learning, approximation theory and statistics, with an emphasis on sparse approximation, audio source separation, dictionary learning and compressed sensing. He founded the series of international workshops SPARS on Signal Processing with Adaptive/Sparse Representations. In 2011, he was awarded the Blaise Pascal Award in Applied Mathematics and Scientific Engineering from the SMAI by the French National Academy of Sciences, and a starting investigator grant from the European Research Council.