*Numerical Functional Analysis and Optimization*, 32(7):768–805, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0163-0563 print/1532-2467 online
DOI: 10.1080/01630563.2011.580876

# AVERAGE PERFORMANCE OF THE SPARSEST APPROXIMATION USING A GENERAL DICTIONARY

**François Malgouyres[1] and Mila Nikolova[2]**

[1] *IMT, CNRS UMR 5219, Université Paul Sabatier, Toulouse, France*
[2] *CMLA, CNRS UMR 8536, ENS Cachan, UniverSud, France*

□ *Let $A$ be a matrix of size $N \times M$ (a dictionary) and let $\|\cdot\|$ be a norm on $\mathbb{R}^N$. For any data $d \in \mathbb{R}^N$, we consider the sparsest vector (i.e., the one with the smallest number of nonzero entries) $u \in \mathbb{R}^M$ such that $\|Au - d\| \le \tau$, for a parameter $\tau > 0$. We say that $u$ is a $K$-sparse solution if it has less than $K \in \mathbb{N}$ nonzero entries. In this article, we give a precise geometrical description of the data sets $d \in \mathbb{R}^N$ yielding a $K$-sparse solution. We parameterize and measure these sets. More precisely, we measure their intersection with a ball defined by any given norm $\delta$ and a radius $\theta$. These measures are expressed in terms of the constituents of the optimization problem—namely $A$, $\|\cdot\|$, $\tau$, $\delta$ and $\theta$—and they enable these constituents to be rated. This is the core of a new methodology, called Average Performance in Approximation (APA), inaugurated in this work. By way of application, we give the probability of obtaining a $K$-sparse solution, when $d$ is uniformly distributed in the $\delta$-ball of radius $\theta$. Analyzing the obtained formulas reveals what are the most important features of the dictionary and the norm defining the data fidelity, to obtain sparse solutions. This crucial question is largely discussed. We also provide an example when both $\|\cdot\|$ and $\delta$ are the Euclidian norm. Some among the wide-ranging perspectives raised by the new APA methodology are described as well.*

**Keywords** Approximation; Best K-term approximation; Compression; Constrained minimization; Dictionary; Estimation; Frames; $\ell_0$ norm; Measure theory; Nonconvex nonsmooth functions; Sparse representations.

**AMS Subject Classification** Primary 41-02, 41R25, 65D15, 49K30; Secondary 68U10, 90C27, 90C26.

## 1. INTRODUCTION

### 1.1. The Sparsest Approximation

The model under study is the sparsest approximation of a datum $d \in \mathbb{R}^N$ using a given $N \times M$ real-valued matrix $A$, with $M \ge N$ and

Received 9 December 2010; Revised 9 April 2011; Accepted 11 April 2011.
Address correspondence to Mila Nikolova, Mathematiques, 61 av. du Président Wilson, Cachan Cedex 94235, France; E-mail: nikolova@cmla.ens-cachan.fr

$\operatorname{rank}(A) = N$. We call the *sparsest approximation of $d$* any solution of the problem[1]

$$(\mathscr{P}_d): \qquad \begin{cases} \text{minimize } \ell_0(u) \\ \text{under the constraint: } \|Au - d\| \leq \tau, \end{cases} \tag{1}$$

where $\|\cdot\|$ is a given norm on $\mathbb{R}^N$, $\tau > 0$ is a tolerance parameter and for any $u = (u_1, \ldots, u_M) \in \mathbb{R}^M$

$$\ell_0(u) \overset{\text{def}}{=} \#\{i \in \{1, \ldots, M\} : u_i \neq 0\},$$

where # means cardinality. By a slight abuse of language, the function $\ell_0$ is commonly called the $\ell_0$-norm.

The $\ell_0$-norm can equivalently be written as

$$\ell_0(u) = \sum_{i=1}^{M} \varphi(u_i) \quad \text{where } \varphi(t) = \begin{cases} 0 & \text{if } t = 0, \\ 1 & \text{if } t \neq 0. \end{cases}$$

This function $\varphi$ has a long history, especially in image restoration using Markovian priors, see for example, [1, 2, 15–17, 24] and in hard thresholding of noisy wavelet coefficients, see [12].

Let us enumerate a few fields where problems similar to $(\mathscr{P}_d)$ arise.

In signal and image processing, approximation is related to the compression problem. A natural way to compress an image is to quantize and encode a solution of an instance of $(\mathscr{P}_d)$. In this context, the codelength is bounded from above and from below by the value of the minimum of $(\mathscr{P}_d)$ multiplied by constants.

The problem $(\mathscr{P}_d)$ is also very successful in the *compressed sensing* context (see [4, 7] for an extensive list of related papers). In this context, solving $(\mathscr{P}_d)$ permits to recover the unknown coordinates of a sparse vector $u \in \mathbb{R}^M$ from the measurements $d \in \mathbb{R}^N$.

The main drawback of $(\mathscr{P}_d)$ is to be NP-hard in general [10]. It is approximatively solved using various heuristics, see [6, 8, 19, 22] for the main historical examples. Most recently proposed algorithms are developed and analyzed with regard to their performance in a compressed sensing setting (see [3, 9, 20]). The performance of these algorithms concerning their approximation abilities is usually not analyzed. In this regard, one important consequence of the theorems obtained in the compressed sensing literature is to guarantee that, under strong hypotheses on the matrix $A$ and the sparsity of the original $u$, these

[1] Notice that, because $\operatorname{rank}(A) = N$, the constraint in $(\mathscr{P}_d)$ is nonempty whatever $d \in \mathbb{R}^N$. Therefore, the minimum is reached since $\ell_0$ takes its values in the finite set $\{0, \ldots, M\}$. Also, the problem $(\mathscr{P}_d)$ may have infinitely many solutions.

heuristics give exact solutions to $(P_d)$. The performances of typical compressed models in terms of the "best $K$-term approximation" are analyzed in [5].

## 1.2. Average Performance in Approximation

Let us first underline that evaluating the performances of sparsity promoting models/algorithms is the key to important tasks such as

- discriminate between the different models/algorithms approaching the sparsest approximation,
- tune a model/algorithm to improve its performances (i.e., design the matrix $A$ and, when possible, choose properly the norm $\|\cdot\|$).

The particular case of the sparsest approximation is an important step in this direction because by construction $(\mathscr{P}_d)$ is the model that provides the sparsest approximation of the data for any given accuracy. The obtained performance is, therefore, a limit which cannot be improved.[2]

Let us now sketch the proposed methodology for estimating the performances of $(\mathscr{P}_d)$. We denote by $\text{val}(\mathscr{P}_d)$ the value of the minimum in $(\mathscr{P}_d)$. We know that, since $d$ is a random variable, $\text{val}(\mathscr{P}_d)$ is also a random variable. Moreover, it takes its values[3] in $\{0,\ldots,N\}$. The purpose of the approach we inaugurate, namely the Average Performance in Approximation (APA), is to estimate the distribution law of $\text{val}(\mathscr{P}_d)$, given a distribution law for $d$. More precisely, we estimate the probability to obtain a $K$-sparse solution:

$$\mathbb{P}(\text{val}(\mathscr{P}_d) \leq K) \quad \text{for all } K \in \{0,\ldots,N\}. \tag{2}$$

The quantities in (2) depend on $\tau$, $A$, $\|\cdot\|$, and the assumed distribution for $d$. The latter ingredients could then be chosen in order to maximize $\mathbb{P}(\text{val}(\mathscr{P}_d) \leq K)$, for $K$ small. Such an approach promises a reasonable way to build appropriate models that favor sparsity.

## 1.3. Worst Case Analysis in Nonlinear Approximation

Evaluating the performance of a sparsity promoting model/algorithm for the purpose of realizing nonlinear approximation is a very active field of research. For a survey, we refer to [11]. Let us sketch the typical results

[2] More precisely, it cannot be improved as long as we consider constraints defined by a norm.

[3] It is easy to see that the columns of $A$ corresponding to the non zero coordinates of a minimizer of $(\mathscr{P}_d)$ must be independent in $\mathbb{R}^N$. Therefore, we obviously have $\text{val}(\mathscr{P}_d) \leq N$.

which are obtained in this field on an analogue of ($\mathscr{P}_d$) named the "best $K$-term approximation."

The best K-term approximation looks for the best possible approximation of data $d \in \mathbb{R}^N$ when using an expansion along $K$ columns of $A$. In formulae, it searches for a minimizer of

$$\inf_{\ell_0(u) \leq K} \|d - Au\| \cdot$$

The performance for a given $d \in \mathbb{R}^N$ is measured by

$$\sigma_K(d) = \inf_{\ell_0(u) \leq K} \|d - Au\| \cdot$$

When estimating performances, an hypothesis is made on the data. It takes the form $d \in B$, where the data domain $B$ is usually the unit ball for a given norm. When the analysis can be conducted, it proves that

$$C_1 K^{-r} \leq \sup_{d \in B} \sigma_K(d) \leq C_2 K^{-r}, \quad \forall K, \tag{3}$$

where $r > 0$ depends on $\| \cdot \|$ and $B$. Note that $C_1 > 0$ and $C_2 > 0$ are absolute constants in the sense that they are independent of $K$ and $N$ (see, e.g., [11]).

Let us emphasize that, because the formula (3) contains a supremum, the upper and the lower bounds do not have the same meaning. Indeed, the upper bound is true whatever $d \in B$ while the lower bound only holds for few $d \in B$, which are the worst elements of $B$. The role of the lower bound is to guarantee that the upper bound cannot be significantly improved.

The clear advantage of these results over the APA is that they apply even if one only has a vague knowledge of the data distribution. Indeed, any data distribution whose support is included in $B$ does enjoy the decay $C_2 K^{-r}$. The counterpart of this advantage is that the constants $C_2$ and $r$ may be too pessimistic when confronted to real data. The worst-case data may be indeed rare. Another important limitation of this methodology is that very little can be said about the performances of the best $K$-term approximation when $M > N$. Let us illustrate this by an example. If one considers an orthonormal basis for $A$ and $\ell_p$ norms for $\| \cdot \|$ and $B$, the worst data are not unique (and they even depend on $K$). Therefore, adding a few new columns to $A$ will not improve the performance observed on $\sup_{d \in B} \sigma_K(d)$. The vector spaces generated using these new columns generally do not permit to improve the performances for all the worst data. As a consequence, nonlinear approximation does not provide satisfactory results when $M > N$ (see [11]). We will see that APA reflects the change induced by these new columns.

### 1.4. Framework of the Results and Notations

In order to estimate (2), we need to describe and measure some geometrical sets related to $(\mathcal{P}_d)$. Let us give our notations and define these sets. These notations and hypotheses hold throughout the article.

For any integer $k$, $u \in \mathbb{R}^k$ is a column vector with $k$ entries $u_i$, for $i = 1, \ldots, k$. The Lebesgue measure in $\mathbb{R}^k$ is denoted by $\mathbb{L}^k(\cdot)$, whereas $\mathrm{I}_k$ stands for the $k \times k$ identity matrix. The Euclidean norm of any $u \in \mathbb{R}^k$ is systematically denoted by $\|u\|_2$.

We consider that the integers $M$ and $N$ are fixed. The notation $d$ either refers to a datum $d \in \mathbb{R}^N$ or a random variable taking values in $\mathbb{R}^N$. It will not be ambiguous, once in context. The norm $\|\cdot\|$ on $\mathbb{R}^N$ always denotes the norm used to define the data fidelity term in $(\mathcal{P}_d)$.

Note that the fixed integer $M$ meets $M \geq N$. Because $N$ and $M$ are constant, we do not express the dependence with regard to these values. Also we consider throughout the article a fixed $N \times M$ matrix $A$ such that $\mathrm{rank}(A) = N$. Beyond the latter, no other assumptions on $A$ are adopted.

Notice that we do not make any assumption relating $N$, $M$, $A$, and $\delta$. We are aware that the benefit of such an assumption would be to allow $N$ and $M$ to evolve and to get asymptotical results when they become infinitely large. The benefit of avoiding this hypothesis is of course generality.

We denote the columns of $A$ by $a_i$, for $i = 1, \ldots, M$. Of course, we have $a_i \in \mathbb{R}^N$. To simplify the notations, we denote $I \overset{\text{def}}{=} \{1, \ldots, M\}$. For any subset (also called support) $J \subset I$, we denote the vector subspace spanned by the columns of $A$ whose indices are in $J$ by:

$$\mathcal{A}_J \overset{\text{def}}{=} \mathrm{span}\{a_j : j \in J\}. \tag{4}$$

We also use the convention $\mathrm{span}\{\emptyset\} \overset{\text{def}}{=} \{0\}$. For any vector subspace $V$ of $\mathbb{R}^N$, we denote by $P_V$ the orthogonal projection onto $V$ and by $V^\perp$ the orthogonal complement of $V$ in $\mathbb{R}^N$. To specify the dimension of $V$, we write $\dim(V)$.

For any function $f : \mathbb{R}^k \to \mathbb{R}$ and any $\tau \in \mathbb{R}$, the $\tau$-level set of $f$ is denoted by

$$B_f(\tau) \overset{\text{def}}{=} \{w \in \mathbb{R}^k, f(w) \leq \tau\}. \tag{5}$$

If $f$ is a norm, then $B_f(\tau)$ is the relevant ball of radius $\tau$ centered at the origin.

Given an arbitrary $\tau > 0$, we introduce the subset of $\mathbb{R}^N$

$$\mathcal{A}_J^\tau \overset{\text{def}}{=} \mathcal{A}_J + P_{\mathcal{A}_J^\perp} B_{\|\cdot\|}(\tau), \tag{6}$$

where the sum is the direct sum between the two sets. Geometrically, $\mathcal{A}_J^\tau$ is a cylinder in $\mathbb{R}^N$: like a $\tau$-thick coat wrapping the subspace $\mathcal{A}_J$.

Let us define as well

$$G_K \overset{\text{def}}{=} \{J \subset I : \dim(\mathcal{A}_J) \leq K\}. \tag{7}$$

In general, there may exist two subsets $J_1$ and $J_2 \subseteq I$, such that $\mathcal{A}_{J_1} = \mathcal{A}_{J_2}$ and $J_1 \neq J_2$. A non-redundant listing of all the different subspaces that can be generated by the elements of $G_K$ is obtained as described next. For any $K = 0, \ldots, N$, define $\mathcal{J}(K)$ by the following three properties:

$$\begin{cases} (a) & \mathcal{J}(K) \subset \{J \subset I : \dim(\mathcal{A}_J) = K\}; \\ (b) & \text{if } J_1, J_2 \in \mathcal{J}(K) \text{ and } J_1 \neq J_2, \text{ then} \\ & \mathcal{A}_{J_1} \neq \mathcal{A}_{J_2}; \\ (c) & \mathcal{J}(K) \text{ is maximal:} \\ & \text{if } J_1 \subset I \text{ yields } \dim(\mathcal{A}_{J_1}) = K \text{ then} \\ & \exists J \in \mathcal{J}(K) \text{ such that } \mathcal{A}_J = \mathcal{A}_{J_1}. \end{cases} \tag{8}$$

Notice that in particular, $\mathcal{J}(0) = \{\emptyset\}$ and $\#\mathcal{J}(N) = 1$ because for any $J \subset I$ such that $\dim(\mathcal{A}_J) = N$ we have $\mathcal{A}_J = \mathbb{R}^N$. Also, unless some columns of $A$ are aligned,[4] we have $\#\mathcal{J}(K) = \frac{M!}{K!(M-K)!}$, for $K = 1, \ldots, N-1$.

Observe that $G_K$, as defined in (7), satisfies

$$G_K \supset \bigcup_{k=0}^{K} \mathcal{J}(k)$$

and

$$\{\mathcal{A}_J : J \in G_K\} = \{\mathcal{A}_J : J \in \mathcal{J}(k) \text{ for } k \in \{0, \ldots, K\}\}. \tag{9}$$

For any $d \in \mathbb{R}^N$, any solution $u^*$ of $(\mathcal{P}_d)$ (see (1)) has the same $\ell_0$ norm, so we systematically denote

$$\text{val}(\mathcal{P}_d) \overset{\text{def}}{=} \ell_0(u^*).$$

For any given $K \in \{0, \ldots, N\}$ and $\tau > 0$, the subset $\Sigma_K^\tau$ below

$$\Sigma_K^\tau \overset{\text{def}}{=} \{d \in \mathbb{R}^N : \text{val}(\mathcal{P}_d) \leq K\} \tag{10}$$

contains all data subsets from $\mathbb{R}^N$ leading to a $K$-sparse solution.

[4] This is the usual condition imposed when using $(\mathcal{P}_d)$ in a compressed sensing framework. This condition is much weaker than the RIP, *spark*, and incoherence [4, 7].

Throughout this article, we consider a norm $\delta$ on $\mathbb{R}^N$ and a positive real $\theta$. We assume that the random variable $d$ is uniformly distributed in $B_\delta(\theta)$. Typically, compressible data are modeled using $\delta$ equal to the $l^1$ norm, the Minkowski function defined by a polytope, an so on.

As is usual, we write $o(t)$ for a function satisfying $\lim_{t\to 0}\frac{o(t)}{t} = 0$.

### 1.5. Our Contribution

In this article, we inaugurate the APA in order to estimate the ability of the model $(\mathscr{P}_d)$ to provide a sparse representation of data which are uniformly distributed in $B_\delta(\theta)$. The main steps of the proposed methodology are sketched below.

- We establish in Theorem 1 that, for $K = 0, \ldots, N$ and $\tau > 0$,

$$\Sigma^\tau_K = \bigcup_{J\in\mathscr{J}(K)} \mathscr{A}^\tau_J.$$

The geometry behind this formula is illustrated in Figure 1.

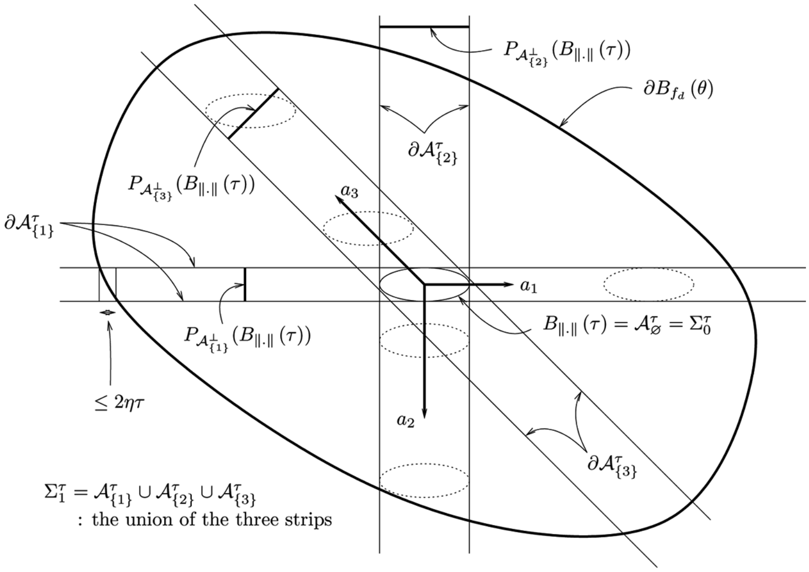

$P_{\mathcal{A}^\perp_{\{2\}}}(B_{\|\cdot\|}(\tau))$
$\partial B_{f_d}(\theta)$
$\partial\mathcal{A}^\tau_{\{2\}}$
$P_{\mathcal{A}^\perp_{\{3\}}}(B_{\|\cdot\|}(\tau))$
$a_3$
$\partial\mathcal{A}^\tau_{\{1\}}$
$a_1$
$B_{\|\cdot\|}(\tau) = \mathcal{A}^\tau_\varnothing = \Sigma^\tau_0$
$P_{\mathcal{A}^\perp_{\{1\}}}(B_{\|\cdot\|}(\tau))$
$\leq 2\eta\tau$
$a_2$
$\partial\mathcal{A}^\tau_{\{3\}}$
$\Sigma^\tau_1 = \mathcal{A}^\tau_{\{1\}} \cup \mathcal{A}^\tau_{\{2\}} \cup \mathcal{A}^\tau_{\{3\}}$
: the union of the three strips

**FIGURE 1** Example in dimension 2. Let $\{a_1, a_2, a_3\}$ be a dictionary on $\mathbb{R}^2$. On the drawing, the sets $P_{\mathscr{A}^\perp_{\{i\}}}(B_{\|\cdot\|}(\tau))$, for $i = 1, 2, 3$, are shifted by an element of $\mathscr{A}_{\{i\}}$. The dotted sets represent translations of $B_{\|\cdot\|}(\tau)$. The set-valued function $\Sigma^\tau_\cdot$, as presented in (10) and Proposition 1, gives rise to the following situations: $\Sigma^\tau_0 = B_{\|\cdot\|}(\tau) = \mathscr{A}^\tau_\emptyset$, $\Sigma^\tau_1 = \mathscr{A}^\tau_{\{1\}} \cup \mathscr{A}^\tau_{\{2\}} \cup \mathscr{A}^\tau_{\{3\}}$ and $\Sigma^\tau_2 = \mathbb{R}^2 = \mathscr{A}^\tau_{\{1,2\}} = \mathscr{A}^\tau_{\{2,3\}} = \ldots$ The symbol $\partial$ is used to denote the boundaries of the sets.

- Then, after some intermediate calculations, we bound both from above and from below
$$\mathbb{L}^N\Big(\Sigma_K^\tau \bigcap B_\delta(\theta)\Big).$$
The precise result is given in Theorem 2. The upper and the lower bounds are both of the form
$$\mathbf{C}_K\left(\frac{\tau}{\theta}\right)^{N-K}\theta^N + o\left(\frac{\tau}{\theta}\right)^{N-K}\theta^N, \qquad (11)$$
with
$$\mathbf{C}_K = \sum_{J\in\mathcal{J}(K)} \mathbb{L}^{N-K}(P_{\mathcal{A}_J^\perp} B_{\|\cdot\|}(1))\mathbb{L}^K(\mathcal{A}_J \cap B_\delta(1)).$$
- When $d$ is uniformly distributed in $B_\delta(\theta)$, we have
$$\mathbb{P}(\text{val}(\mathcal{P}_d) \le K) = \frac{\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta))}{\mathbb{L}^N(B_\delta(\theta))}.$$
So, we only need to divide (11) by $\mathbb{L}^N(B_\delta(\theta))$ in order to obtain bounds on $\mathbb{P}(\text{val}(\mathcal{P}_d) \le K)$, see section 5. The simplified statement says that
$$\mathbb{P}(\text{val}(\mathcal{P}_d) \le K) = \frac{\mathbf{C}_K}{\mathbb{L}^N(B_\delta(1))}\left(\frac{\tau}{\theta}\right)^{N-K} + o\left(\frac{\tau}{\theta}\right)^{N-K}. \qquad (12)$$
This shows that asymptotically, the performance of $(\mathcal{P}_d)$ reads in the constants
$$\frac{\mathbf{C}_K}{\mathbb{L}^N(B_\delta(1))} = \sum_{J\in\mathcal{J}(K)} \frac{\mathbb{L}^{N-K}\big(P_{\mathcal{A}_J^\perp} B_{\|\cdot\|}(1)\big)\mathbb{L}^K(\mathcal{A}_J \cap B_\delta(1))}{\mathbb{L}^N(B_\delta(1))}. \qquad (13)$$
Increasing these constants improves the performance of the model.
- We discuss alternative statements in section 6 and illustrate our results in an Euclidean context in section 7.

The main limitation of (11) and (12) is that the $o\left(\frac{\tau}{\theta}\right)^{N-K}$ terms depend on all the ingredients of the model and might require strong constraints on $\frac{\tau}{\theta}$. A precise and dedicated study of simple situations need to be performed to better understand how the $o(\cdot)$ term behaves when $N$ varies. This requires to define $A$, $\|\cdot\|$ and $\delta$ as a function of $N$ and is out of the scope of this article. Said differently, the constants exhibited in this article depend on $A$, $\|\cdot\|$, $\delta$, but are independent of $\tau/\theta$. The dependence with regard to $K$ is specified by a subscript. This limitation being clarified, we will see in the next paragraph that the interpretations of (13) are very reasonable and agree with the common intuition on $(\mathscr{P}_d)$.

Let us analyze the meaning of (13).

- The sum in (13) involves *all* the possible vector subspaces of dimension $K$ spanned by the columns of $A$. Moreover, all the summands are positive. Therefore, all these subspaces contribute to the success of the model. This is well known to the readers familiar with the subject.
  A trivial consequence of the above remark is that, when adding a new column to the matrix $A$, $\mathscr{J}(K)$ grows. All the former subvector spaces are still available and we add more. As a consequence, the constant $\mathbf{C}_K$ in (13) increases, obtaining sparse solutions is more likelly and therefore the model is improved. This simple and intuitive statement is not visible in the nonlinear approximation context explained in Section 1.3.
- The term $\mathbb{L}^K(\mathscr{A}_J \cap B_\delta(1))$ represents the measure of the *whole* set $\mathscr{A}_J \cap B_\delta(1)$. This completes the previous remark and is again known to be an important property of $(\mathscr{P}_d)$.
- The term $\mathbb{L}^N(B_\delta(1))$ in the denominator makes an impact on the results for all $K$. Its consequence is that data living in a small set are easier to capture if we manage to keep $\mathbb{L}^K(\mathscr{A}_J \cap B_\delta(1))$ fixed.
- The data fidelity term is contained in the term $\mathbb{L}^{N-K}(P_{\mathscr{A}_J^\perp} B_{\|\cdot\|}(1))$. Maximizing the probability with regard to $\|\cdot\|$ means maximizing a weighted sum of the form (13). Even in simplified settings, it is unlikely that the Euclidean norm maximizes it. Therefore, assuming that $\|\cdot\|$ is the Euclidean norm, as is common in the compressed sensing framework (see [14, 25, 26]), appears like a strong limitation in the approximation framework.

## 2. SETS OF DATA YIELDING *K*-SPARSE SOLUTIONS

**Proposition 1.** *Given the notations of section* 1.4, *we have for any* $K \in \{0, \ldots, N\}$ *and any* $\tau > 0$

$$\Sigma_K^\tau = \bigcup_{J \in G_K} \mathscr{A}_J + B_{\|\cdot\|}(\tau).$$

Some sets $\Sigma_K^\tau$, as defined in (10) and considered in the last proposition, are illustrated on Figure 1.

***Proof.*** The case $K = 0$ is trivial $(G_0 = \{\emptyset\})$ and we assume in the following that $K \geq 1$.

Let $d \in \Sigma_K^\tau$ for $\Sigma_K^\tau$ as given in (10). This means there is $u^*$—a solution of $(\mathscr{P}_d)$—that satisfies $\ell_0(u^*) \leq K$. Hence, $d = \sum_{i \in J} u_i^* a_i + w$, with $w \in B_{\|\cdot\|}(\tau)$, and $\#J \leq K$ for $J = \{i \in I : u_i^* \neq 0\}$.

Consequently, $\dim(\mathscr{A}_J) \leq \#J \leq K$, which implies that $d \in \cup_{J\in G_K}\mathscr{A}_J + B_{\|\cdot\|}(\tau)$.

Conversely, let $d \in \cup_{J\in G_K}\mathscr{A}_J + B_{\|\cdot\|}(\tau)$, then $d = v + w$ where $v \in \cup_{J\in G_K}\mathscr{A}_J$ and $w \in B_{\|\cdot\|}(\tau)$. Then:

- $\exists\, J \subset I$ such that $v \in \mathscr{A}_J$ and the latter satisfies $\dim(\mathscr{A}_J) \leq K$;
- there exist $\{u_i \in \mathbb{R} : i \in J\}$ involving at most $\dim(\mathscr{A}_J)$ non-zero components (hence, $\ell_0(u) \leq \dim(\mathscr{A}_J) \leq K$) such that $v = \sum_{i\in J} u_i a_i$.
- $w \in B_{\|\cdot\|}(\tau)$ means that $\|w\| \leq \tau$.

It follows that $d = \sum_{i\in J} u_i a_i + w \in \Sigma_K^\tau$. □

Next we establish that each component in the right term in the equation of Proposition 1 is of the form $\mathscr{A}_J^\tau$.

**Lemma 1.** *Using the notations of section* 1.4, *for any $J \subset I$ (including $J = \emptyset$) and any $\tau > 0$, the set $\mathscr{A}_J^\tau$ in* (6) *satisfies*

$$\mathscr{A}_J^\tau = \mathscr{A}_J + B_{\|\cdot\|}(\tau). \tag{14}$$

*As a consequence, for any $J_1 \subset J \subset I$,*

$$\mathscr{A}_{J_1}^\tau \subset \mathscr{A}_J^\tau. \tag{15}$$

The proof of the lemma is outlined in Appendix A.1. We can anticipate that the form of $\mathscr{A}_J^\tau$ in (6) is better adapted for the goal of measuring subsets.

Using the above notations and the non-redundant listing $\mathscr{J}(K)$ in (8), we provide a sparser formulation of $\Sigma_K^\tau$ than the one given in Proposition 1.

**Theorem 1.** *Given the notations of section* 1.4, *for any $K \in \{0, \ldots, N\}$ and any $\tau > 0$, we have*

$$\Sigma_K^\tau = \bigcup_{J\in\mathscr{J}(K)} \mathscr{A}_J^\tau.$$

*Moreover, $\Sigma_K^\tau$ is closed and measurable.*

*Proof.* The case $J = \emptyset$ (i.e., $K = 0$) is trivial because of the convention $\text{span}(\emptyset) = \{0\}$ and $\mathscr{J}(0) = \{\emptyset\}$.

Let us first prove that $\Sigma_K^\tau = \bigcup_{J\in G_K} \mathscr{A}_J^\tau$. Using Proposition 1,

$$\Sigma_K^\tau = \left(\bigcup_{J\in G_K} \mathscr{A}_J\right) + B_{\|\cdot\|}(\tau) = \bigcup_{J\in G_K} \big(\mathscr{A}_J + B_{\|\cdot\|}(\tau)\big).$$

The last equality above is a trivial observation. Using (14) in Lemma 1, this summarizes as

$$\Sigma_K^\tau = \bigcup_{J\in G_K} \mathcal{A}_J^\tau.$$

Using (9), we deduce that

$$\{\mathcal{A}_J^\tau : J \in G_K\} = \{\mathcal{A}_J^\tau : J \in \mathcal{J}(k) \text{ for } k \in \{0,\ldots,K\}\},$$

and, therefore,

$$\Sigma_K^\tau = \bigcup_{k=0}^{K} \bigcup_{J\in\mathcal{J}(k)} \mathcal{A}_J^\tau.$$

Furthermore, for any $J_1 \in \mathcal{J}(k)$, for $k \in \{0,\ldots,K-1\}$, there exists $J \in \mathcal{J}(K)$ such that $J_1 \subset J$. By (15) in Lemma 1, we have $\mathcal{A}_{J_1}^\tau \subset \mathcal{A}_J^\tau$ and, hence,

$$\Sigma_K^\tau = \bigcup_{J\in\mathcal{J}(K)} \mathcal{A}_J^\tau.$$

The sets $\mathcal{A}_J$ and $P_{\mathcal{A}_J^\perp} B_{\|\cdot\|}(\tau)$ are closed and mutually orthogonal. Therefore, $\mathcal{A}_J^\tau$ is closed. As a consequence $\mathcal{A}_J^\tau$ is a Borel set and is Lebesgue measurable. Since $\Sigma_K^\tau$ is a finite union of closed measurable sets, $\Sigma_K^\tau$ is closed and measurable as well. □

## 3. PRELIMINARIES TO MEASURE DATA SUBSETS

### 3.1. Motivation

Data $d \in \mathbb{R}^N$ being uniformly distributed in $B_\delta(\theta)$ by assumption, the following identity holds for all $K = 1,\ldots,N$:

$$\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le K) = \frac{\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta))}{\mathbb{L}^N(B_\delta(\theta))}.$$

In order to evaluate $\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le K)$, we need to calculate $\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta))$. The latter calculation presents the main difficulty in the APA methodology developed in this work.

Using Theorem 1, it is straightforward that

$$\Sigma_K^\tau \cap B_\delta(\theta) = \bigcup_{J\in\mathcal{J}(K)} \left(\mathcal{A}_J^\tau \cap B_\delta(\theta)\right) \tag{16}$$

and that

$$\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta)) = \mathbb{L}^N\Big(\bigcup_{J\in\mathcal{J}(K)} \big(\mathcal{A}_J^\tau \cap B_\delta(\theta)\big)\Big). \tag{17}$$

Considering the right side of the latter equality (17), we will evaluate the measure of subsets of the form $\mathcal{A}_J^\tau \cap B_\delta(\theta)$. We indeed know that we can sum those measures to obtain an estimate of the measure of the union if the intersections of the form $(\mathcal{A}_{J_1}^\tau \cap B_\delta(\theta)) \cap (\mathcal{A}_{J_2}^\tau \cap B_\delta(\theta))$ remain controlled whenever $J_1 \neq J_2$ in $\mathcal{J}(K)$.

These problems are addressed in detail in sections 3.2 and 3.3.

### 3.2. Measuring Bounded Cylinder-Like Subsets of $\mathbb{R}^N$

Here, we characterize subsets of the form $P_{V^\perp}B_{\|\cdot\|}(\tau)$, see (6), for general vector subspaces $V \subset \mathbb{R}^N$.

**Lemma 2.** *For any vector subspace $V \subset \mathbb{R}^N$ and any norm $\|\cdot\|$ on $\mathbb{R}^N$, define the function $h$ on $V^\perp$ by*

$$h(u) \overset{\text{def}}{=} \inf\Big\{t \geq 0 : \frac{u}{t} \in P_{V^\perp}B_{\|\cdot\|}(1)\Big\}, \tag{18}$$

*for all $u \in V^\perp$.*

*Then the following statements hold:*

(i) *For any $\tau \geq 0$, we have*

$$B_h(\tau) = P_{V^\perp}B_{\|\cdot\|}(\tau). \tag{19}$$

(ii) *The function $h$ in (18) is a norm on $V^\perp$.*

(iii) *There exist constants $\eta$ and $\eta_2 > 0$ which only depend on $\delta$ and $\|\cdot\|$ (and are independent of $V$) such that*

$$\delta(u) \leq \eta\ h(u), \quad \forall u \in V^\perp, \tag{20}$$

$$\|u\|_2 \leq \eta_2\ h(u), \quad \forall u \in V^\perp. \tag{21}$$

The proof of this lemma is outlined in Appendix A.2. Note that $h : V^\perp \to \mathbb{R}$ in (18) is the usual Minkowski functional of $P_{V^\perp}B_{\|\cdot\|}(1)$, see, for example, [18, p. 131].

The next proposition, proven in Appendix A.3 is a key result that will be used several times in what follows.

**Proposition 2.** *For any vector subspace $V$ of $\mathbb{R}^N$, any norm $\|\cdot\|$ on $\mathbb{R}^N$ and any $\tau > 0$, define*

$$V^\tau = V + P_{V^\perp} B_{\|\cdot\|}(\tau). \tag{22}$$

*Then the following hold:*

(i) *$V^\tau$ is closed and measurable in $\mathbb{R}^N$;*
(ii) *Let $\delta$ be any norm on $\mathbb{R}^N$, $h : V^\perp \to \mathbb{R}$ the norm defined in Lemma 2, $K = \dim(V)$ and $\lambda_V$ be any constant such that*

$$\delta(u) \leq \lambda_V h(u), \quad \forall u \in V^\perp. \tag{23}$$

*If $\theta \geq \tau\lambda_V$, then*

$$C\tau^{N-K}(\theta - \tau\lambda_V)^K \leq \mathbb{L}^N(V^\tau \cap B_\delta(\theta)) \leq C\tau^{N-K}(\theta + \tau\lambda_V)^K, \tag{24}$$

*where*

$$C = \mathbb{L}^{N-K}(P_{V^\perp} B_{\|\cdot\|}(1))\ \mathbb{L}^K(V \cap B_\delta(1)) > 0 \tag{25}$$

*is finite.*[5]

**Remark 1.** Using Lemma 2, the condition in (23) holds for any $\lambda_V \geq \lambda_V^*$ for some optimal $\lambda_V^* \in [0, \eta]$, where $\eta$ is given in Lemma 2. Let us emphasize that $\lambda_V$ and $\lambda_V^*$ may depend on $V$ (which explains the letter "V" in the index). However, the proposition clearly holds if we take $\lambda_V = \eta$, where $\eta$ is the constant of Lemma 2, assertion (iii). In this case, the constant is independent of $V$.

Constant $C$ depends only on the norms $\|\cdot\|$ and $\delta$, and potentially on $V$.

**Remark 2.** An important consequence of Proposition 2 is that asymptotically we get

$$\mathbb{L}^N(V^\tau \cap B_\delta(\theta)) = C\theta^N\left(\frac{\tau}{\theta}\right)^{N-K} + \theta^N o\left(\left(\frac{\tau}{\theta}\right)^{N-K}\right) \quad \text{as } \frac{\tau}{\theta} \to 0.$$

### 3.3. Measuring $\mathscr{A}_J^\tau \cap B_\delta(\theta)$ and the Intersection of Two Such Sets

Using the results of section 3.2, we can address subsets of the form $\mathscr{A}_J^\tau \cap B_\delta(\theta)$ as in (16), as well as intersections of such subsets relevant to different $J$s. These are obtained as consequences of Proposition 2.

[5] Remind that for $V = \{0\}$, $K = 0$ and we have $\mathbb{L}^K(V \cap B_\delta(1)) = 1$.

**Proposition 3.** *Using the notations of section* 1.4*, for any* $J \subset I$ *(including* $J = \emptyset$*) and any* $\tau > 0$*, put* $K \overset{\text{def}}{=} \dim(\mathscr{A}_J)$*. Then there exists* $\lambda_J \in [0, \eta]$*, for* $\eta$ *as given in Lemma* 2*(iii), such that for* $\theta \geq \lambda_J \tau$ *we have*

$$C_J \tau^{N-K}(\theta - \lambda_J \tau)^K \leq \mathbb{L}^N\big(\mathscr{A}_J^\tau \cap B_\delta(\theta)\big) \leq C_J \tau^{N-K}(\theta + \lambda_J \tau)^K, \tag{26}$$

*where*

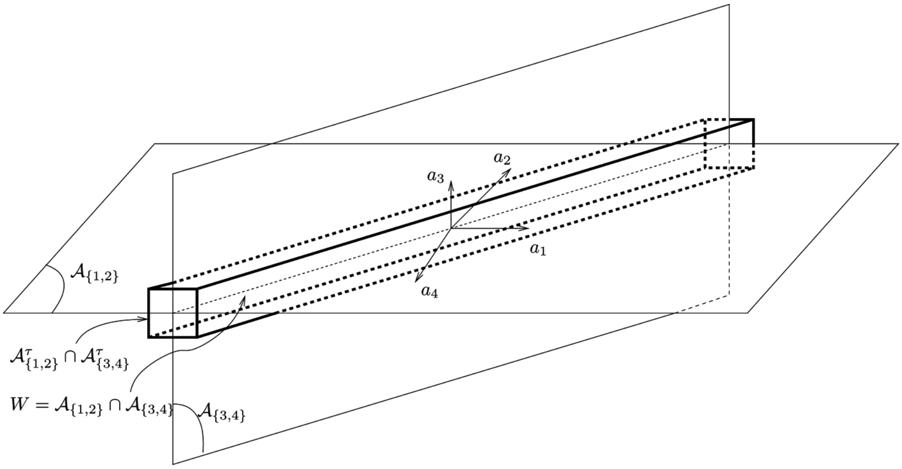
$$C_J = \mathbb{L}^{N-K}\big(P_{\mathscr{A}_J^\perp} B_{\|\cdot\|}(1)\big)\mathbb{L}^K(\mathscr{A}_J \cap B_\delta(1)) > 0 \tag{27}$$

*is finite.*

*Proof.* The proposition is a direct consequence of Proposition 2. Notice that we now write $\lambda_J$ for the constant $\lambda_V$ with $V = \mathscr{A}_J$ in Proposition 2. □

Remind that the constant $\eta$, defined in Lemma 2, depends only on the norms $\| \cdot \|$ and $\delta$.

Next we focus on the intersection of two different subsets of the form $\mathscr{A}_J^\tau \cap B_\delta(\theta)$. The proposition below confirms the intuition that the measure of this kind of subsection is small and almost negligible when compared with the volume of $\mathscr{A}_J^\tau \cap B_\delta(\theta)$. This is illustrated on Figure 2.



$a_2$
$a_3$
$a_1$
$a_4$
$\mathcal{A}_{\{1,2\}}$
$\mathcal{A}^\tau_{\{1,2\}} \cap \mathcal{A}^\tau_{\{3,4\}}$
$W = \mathcal{A}_{\{1,2\}} \cap \mathcal{A}_{\{3,4\}}$
$\mathcal{A}_{\{3,4\}}$

**FIGURE 2** Example of an intersection in dimension 3. $\mathscr{A}^\tau_{\{1,2\}}$ is in between two planes, parallel to $\mathscr{A}_{\{1,2\}}$. Same remark for $\mathscr{A}^\tau_{\{3,4\}}$. The set $\mathscr{A}^\tau_{\{1,2\}} \cap \mathscr{A}^\tau_{\{3,4\}}$ is of the form $W + P_{W^\perp} B_{\tilde{g}}(\tau)$, where $\tilde{g}$ is a norm for $W = \mathscr{A}_{\{1,2\}} \cap \mathscr{A}_{\{3,4\}}$. We also have $\dim(\mathscr{A}_{\{1,2\}} \cap \mathscr{A}_{\{3,4\}}) < \dim(\mathscr{A}_{\{1,2\}}) = \dim(\mathscr{A}_{\{3,4\}})$.

**Proposition 4.** *Let $J_1 \subset I$ and $J_2 \subset I$ be such that $\mathscr{A}_{J_1} \neq \mathscr{A}_{J_2}$ and $\dim(\mathscr{A}_{J_1}) = \dim(\mathscr{A}_{J_2}) \overset{\text{def}}{=} K$. Let $\tau > 0$ and $\theta > 0$. Then we have the following results:*

(i) *The set given below*

$$\mathscr{A}_{J_1}^{\tau} \cap \mathscr{A}_{J_2}^{\tau} \cap B_{\delta}(\theta) \tag{28}$$

*is closed and measurable.*

(ii) *Define $k = \dim\left(\mathscr{A}_{J_1} \cap \mathscr{A}_{J_2}\right)$. For $\eta$ as given in Lemma 2(iii), there is a constant $\lambda_{J_1,J_2} \in [0, 3\eta]$ such that for $\theta \geq \lambda_{J_1,J_2}\tau$ we have*

$$\mathbb{L}^N(\mathscr{A}_{J_1}^{\tau} \cap \mathscr{A}_{J_2}^{\tau} \cap B_{\delta}(\theta)) \leq Q_{J_1,J_2}\tau^{N-k}(\theta + \lambda_{J_1,J_2}\tau)^k,$$

*where $Q_{J_1,J_2}$ reads*

$$Q_{J_1,J_2} \overset{\text{def}}{=} \mathbb{L}^{N-k}(W^{\perp} \cap B_{\|\cdot\|_2}(2\eta_2))\mathbb{L}^k(W \cap B_{\delta}(1)) \tag{29}$$

*for $W \overset{\text{def}}{=} \mathscr{A}_{J_1} \cap \mathscr{A}_{J_2}$ and $\eta_2$ defined in Lemma 2(iii).*

Notice that $Q_{J_1,J_2}$ depends only on $\{a_j : j \in J_1\}$ and $\{a_j : j \in J_2\}$, and the norms $\|\cdot\|$ and $\delta$. A tighter bound can be found in the proof of the proposition (see equation (80) in Appendix A.4). The bound is expressed in terms of a norm $\tilde{g}$ constructed there. The proof of the proposition is presented in Appendix A.4.

**Remark 3.** We have the following asymptotical result:

$$\begin{aligned}
\mathbb{L}^N(\mathscr{A}_{J_1}^{\tau} \cap \mathscr{A}_{J_2}^{\tau} \cap B_{\delta}(\theta)) &\leq Q_{J_1,J_2}\theta^N\left(\frac{\tau}{\theta}\right)^{N-k}\left(1 + \lambda_{J_1,J_2}\frac{\tau}{\theta}\right)^k \\
&= Q_{J_1,J_2}\theta^N\left(\frac{\tau}{\theta}\right)^{N-k} + o\left(\left(\frac{\tau}{\theta}\right)^{N-k}\right) \quad \text{as } \frac{\tau}{\theta} \to 0 \\
&= \theta^N o\left(\left(\frac{\tau}{\theta}\right)^{N-K}\right) \quad \text{as } \frac{\tau}{\theta} \to 0,
\end{aligned}$$

where the last inequality holds because $Q_{J_1,J_2}$ is a constant and $k = \dim(W) < K$.

## 4. MEASURING DATA SETS YIELDING $K$-SPARSE SOLUTIONS

Using the results presented in section 3.3, we now derive upper and lower bounds on $\mathbb{L}^N(\Sigma_K^{\tau} \cap B_{\delta}(\theta))$ as given in (17), namely

$$\mathbb{L}^N(\Sigma_K^{\tau} \cap B_{\delta}(\theta)) = \mathbb{L}^N\left(\bigcup_{J \in \mathcal{J}(K)} \left(\mathscr{A}_J^{\tau} \cap B_{\delta}(\theta)\right)\right). \tag{30}$$

To this end, we introduce several constants based on those obtained in section 3.3.

For any $K = 0, \ldots, N$, define the constants $\hat{\lambda}_K$ and $\mathbf{C}_K$ as it follows:

$$\hat{\lambda}_K \overset{\text{def}}{=} \max_{J \in \mathcal{J}(K)} \lambda_J, \tag{31}$$

$$\mathbf{C}_K \overset{\text{def}}{=} \sum_{J \in \mathcal{J}(K)} C_J, \tag{32}$$

where $\lambda_J \in [0, \eta]$ and $C_J$ are the constants exhibited in Proposition 3. Clearly,

$$0 \leq \hat{\lambda}_K \leq \eta. \tag{33}$$

In particular,

$$\mathbf{C}_0 = \mathbb{L}^N(B_{\|\cdot\|}(1)) \quad \text{and} \quad \mathbf{C}_N = \mathbb{L}^N(B_\delta(1)). \tag{34}$$

With $\mathcal{J}(K)$, let us associate the family of subsets:

$$\mathcal{H}(K, k) \overset{\text{def}}{=} \{(J_1, J_2) \in \mathcal{J}(K)^2 \text{ such that } \dim(\mathcal{A}_{J_1} \cap \mathcal{A}_{J_2}) = k\}, \tag{35}$$

where $K = 1, 2, \ldots, N$ and $k = 0, 1, \ldots, K-1$.

Notice that $\mathcal{H}(K, k)$ may be empty for some $k$. Consider $(J_1, J_2) \in \mathcal{J}(K)^2$ and the classical decomposition

$$\mathcal{A}_{J_1} + \mathcal{A}_{J_2} = \left(\mathcal{A}_{J_1} \cap \mathcal{A}_{J_2}\right) \oplus \left(\mathcal{A}_{J_1} \cap \mathcal{A}_{J_2}^{\perp}\right) \oplus \left(\mathcal{A}_{J_2} \cap \mathcal{A}_{J_1}^{\perp}\right) \subset \mathbb{R}^N.$$

The dimension of the above vector subspace satisfies

$$\dim(\mathcal{A}_{J_1} + \mathcal{A}_{J_2}) = k + (K - k) + (K - k) \leq N$$

and, therefore, $k \geq 2K - N$. We see that

$$\mathcal{H}(K, k) \neq \emptyset \Rightarrow k \geq 2K - N.$$

Conversely,

$$k < k_K \overset{\text{def}}{=} \max\{0,\ 2K - N\} \Rightarrow \mathcal{H}(K, k) = \emptyset. \tag{36}$$

Notice that $\mathcal{H}(N, k) = \emptyset$ for all $k = 0, \ldots, N-1$. Moreover, for any $K \leq N - 1$, we have $-N \leq -K - 1$ and therefore

$$2K - N \leq 2K - K - 1 = K - 1.$$

As a consequence, for $K = 1, \dots, N-1$, we have $0 \leq k_K \leq K-1$. We conclude that $\mathscr{H}(K,k)$ is non-empty (at most) for the indexes $K = 1, \dots, N-1$ and $k = k_K, \dots, K-1$.

For $K \in \{1, \dots, N-1\}$ and $k \in \{k_K, \dots, K-1\}$ let us define

$$\widehat{\chi}_{K,k} \overset{\text{def}}{=} \max\left\{0, \max_{(J_1,J_2)\in\mathscr{H}(K,k)} \lambda_{J_1,J_2}\right\}, \tag{37}$$

$$\mathbf{Q}_{K,k} \overset{\text{def}}{=} \sum_{(J_1,J_2)\in\mathscr{H}(K,k)} Q_{J_1,J_2}, \tag{38}$$

where $Q_{J_1,J_2}$ and $\lambda_{J_1,J_2} \in [0, 3\eta]$ are as in Proposition 4. It follows that for any $K = 1, \dots, N-1$ and any $k = k_K, \dots, K-1$

$$0 \leq \widehat{\chi}_{K,k} \leq 3\eta. \tag{39}$$

It is also clear that if $\mathscr{H}(K,k) = \emptyset$ then we find $\mathbf{Q}_{K,k} = 0$ and $\widehat{\chi}_{K,k} = 0$. In particular for any $k$,

$$\widehat{\chi}_{0,k} = \widehat{\chi}_{N,k} = 0 \quad \text{and} \quad \boldsymbol{Q}_{0,k} = \boldsymbol{Q}_{N,k} = 0. \tag{40}$$

Last, define recursively $\Lambda_0 = \hat{\lambda}_0$ and

$$\Lambda_K = \max\left\{\Lambda_{K-1}, \hat{\lambda}_K, \max_{k_K \leq k \leq K-1} \widehat{\chi}_{K,k}\right\} \tag{41}$$

if $0 < K < N$ and

$$\Lambda_N = \max\{\Lambda_{N-1}, \hat{\lambda}_N\}. \tag{42}$$

Using (33) and (39),

$$0 \leq \Lambda_K \leq 3\eta. \tag{43}$$

**Remark 4.** All the constants introduced between (31) and (41)–(42), namely $\hat{\lambda}_K$ in (31), $\mathbf{C}_K$ in (32), $\widehat{\chi}_{K,k}$ in (37), $\mathbf{Q}_{K,k}$ in (38) and $\Lambda_K$ in (41)–(42), depend only on the dictionary $A$, the norms $\|\cdot\|$ and $\delta$, $K$ and $k$. The latter dependence is explicitly denoted by subscripts $k$ and $K$. The upper bounds on $\Lambda_K$ and $\widehat{\chi}_{K,k}$ use just $\eta$, as defined in Lemma 2(iii). The constant $\eta$ depends only on $\|\cdot\|$ and $\delta$. These constants are involved in the theorem below which provides a crucial result in this work.

The proof of the theorem below in provided in Appendix A.5.

**Theorem 2.** *Let $K \in \{0, \dots, N\}$, the norms $\|\cdot\|$ and $\delta$, and the dictionary $A$, be any. Let $\tau > 0$ and $\theta \geq \tau\Lambda_K$ where $\Lambda_K$ is defined in* (41)–(42). *The Lebesgue measure in $\mathbb{R}^N$ of the set $\Sigma_K^\tau \cap B_\delta(\theta)$ satisfies*

$$\begin{aligned}\mathbf{C}_K \tau^{N-K}(\theta - \hat{\lambda}_K\tau)^K - \theta^N \varepsilon_0(K,\tau,\theta) &\leq \mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta)) \\ &\leq \mathbf{C}_K \tau^{N-K}(\theta + \hat{\lambda}_K\tau)^K, \qquad (44)\end{aligned}$$

*where $\varepsilon_0(K,\tau,\theta) = 0$ if $K \in \{0, N\}$ while for $1 \leq K \leq N-1$,*

$$\varepsilon_0(K,\tau,\theta) = \sum_{k=k_K}^{K-1} \mathbf{Q}_{K,k}\left(\frac{\tau}{\theta}\right)^{N-k}\left(1 + \widehat{\chi}_{K,k}\frac{\tau}{\theta}\right)^k. \qquad (45)$$

*Here $\hat{\lambda}_K$, $\mathbf{C}_K$, $k_K$, $\mathbf{Q}_{K,k}$, and $\widehat{\chi}_{K,k}$ defined by* (31), (32), (36), (38), *and* (37), *respectively. Moreover,* (33), (39), *and* (43) *provide bounds on $\hat{\lambda}_K$, $\widehat{\chi}_{K,k}$, and $\Lambda_K$, respectively, which depend only on the norms $\|\cdot\|$ and $\delta$, via $\eta$ (see Lemma 2(iii)).*

**Remark 5.** We posit the assumptions of Theorem 2. Considering the asymptotic of (44), we have

$$\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta)) = \mathbf{C}_K\theta^N\left(\frac{\tau}{\theta}\right)^{N-K} + \theta^N o\left(\left(\frac{\tau}{\theta}\right)^{N-K}\right) \quad \text{as } \frac{\tau}{\theta} \to 0.$$

**Remark 6.** In the proof of this theorem we notice (see (84), (86), and (82)) that

$$\begin{aligned}&\sum_{J \in \mathcal{J}(K)} \mathbb{L}^N(\mathcal{A}_J^\tau \cap B_\delta(\theta)) - \sum_{k=k_K}^{K-1} \sum_{(J_1,J_2)\in\mathcal{H}(K,k)} \mathbb{L}^N(B_\delta(\theta) \cap \mathcal{A}_{J_1}^\tau \cap \mathcal{A}_{J_2}^\tau) \\ &\quad \leq \mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta)) \leq \sum_{J\in\mathcal{J}(K)} \mathbb{L}^N(\mathcal{A}_J^\tau \cap B_\delta(\theta)). \qquad (46)\end{aligned}$$

These are the main approximations in the evaluation of $\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta))$ involved in Theorem 2. The precision of the bounds given in the theorem could be more accurate by improving the above inequalities. The loss of accuracy due to (46) has however the same order of magnitude as the accuracy in the calculus of $\mathbb{L}^N(\mathcal{A}_J^\tau \cap B_\delta(\theta))$.

In order to give a simplified version of (44), we introduce some additional notations. For any integer $n > 0$, we denote the volume of unit ball for the Euclidian norm $\|\cdot\|_2$ in $\mathbb{R}^n$ by $\alpha(n)$. It reads

$$\alpha(n) = \frac{\pi^{n/2}}{\Gamma(n/2+1)} \quad \text{for } \Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx, \qquad (47)$$

where $\Gamma$ is the usual Gamma function. For any $k = 0, \ldots, N$, let us define

$$\mathscr{C}(k) \overset{\text{def}}{=} \alpha(k)\alpha(N-k). \tag{48}$$

We can now simplify (44). The constants $\Lambda_K$, $\hat{\lambda}_K$ and $\widehat{\chi}_{K,k}$, depend on $A$ and $K$, as well as on the norms $\|\cdot\|$ and $\delta$. Using the uniform bound $\eta$ exhibited in Lemma 2(ii) in place of $\hat{\lambda}_K$ and $\widehat{\chi}_{K,k}$ leads to the following result.

**Corollary 1.** *Let $K \in \{0, \ldots, N\}$, the norms $\|\cdot\|$ and $\delta$, and $A$, be any. Let $\tau > 0$ and $\theta \geq 3\tau\eta$ where $\eta$ is derived in Lemma 2(ii) and depends only on $\delta$ and $\|\cdot\|$. The set $\Sigma_K^\tau$ defined by (10) satisfies*

$$\begin{aligned}\mathbf{C}_K \tau^{N-K}(\theta - \eta\tau)^K - \theta^N \varepsilon_0^u(K, \tau, \theta) &\leq \mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta)) \\ &\leq \mathbf{C}_K \tau^{N-K}(\theta + \eta\tau)^K,\end{aligned} \tag{49}$$

*where $\varepsilon_0^u(K, \tau, \theta) = 0$, for $K = 0$ or $K = N$ and*

$$\varepsilon_0^u(K, \tau, \theta) = \sum_{k=k_K}^{K-1} \mathbf{Q}_{K,k} \left(\frac{\tau}{\theta}\right)^{N-k} \left(1 + 3\eta\frac{\tau}{\theta}\right)^k,$$

*for $K = 1, \ldots, N-1$.*

*Moreover, for $K = 1, \ldots, N-1$ and $k = k_K, \ldots, K-1$, we have*

$$\mathbf{Q}_{K,k} \leq \#\mathcal{J}(K)(\#\mathcal{J}(K) - 1)\mathscr{C}(k)(2\eta_2)^{N-k}\eta_3^k, \tag{50}$$

*where $\mathscr{C}(k)$ is given in (48) and*

$$\#\mathcal{J}(K) \leq \frac{M!}{K!(M-K)!}. \tag{51}$$

*The constant $\eta_2$ is defined in Lemma 2 and $\eta_3$ is such that*

$$\|w\|_2 \leq \eta_3 \delta(w), \quad \forall w \in \mathbb{R}^N.$$

*Proof.* Equation (49) is obtained by inserting in (44) in Theorem 2 the uniform bounds on $\hat{\lambda}_K$, $\widehat{\chi}_{K,k}$ and $\Lambda_K$ given in (33), (39), and (43), respectively.

The upper bound for $\mathbf{Q}_{K,k}$ is calculated as follows. Using (38) and (29), we obtain

$$\mathbf{Q}_{K,k} = \sum_{(J_1,J_2)\in\mathscr{H}(K,k)} \mathbb{L}^{N-k}((\mathscr{A}_{J_1} \cap \mathscr{A}_{J_2})^\perp \cap B_{\|\cdot\|_2}(2\eta_2))\mathbb{L}^k(\mathscr{A}_{J_1} \cap \mathscr{A}_{J_2} \cap B_\delta(1)).$$

Moreover, as is standard, (see, e.g., [23])

$$\begin{aligned}\mathbb{L}^{N-k}((\mathscr{A}_{J_1}\cap\mathscr{A}_{J_2})^{\perp}\cap B_{\|\cdot\|_2}(2\eta_2)) &= \alpha(N-k)(2\eta_2)^{N-k},\\ \mathbb{L}^{k}(\mathscr{A}_{J_1}\cap\mathscr{A}_{J_2}\cap B_{\delta}(1)) &\le \mathbb{L}^{k}(\mathscr{A}_{J_1}\cap\mathscr{A}_{J_2}\cap B_{\|\cdot\|_2}(\eta_3))\\ &= \alpha(k)(\eta_3)^{k},\end{aligned}$$

and we obviously have

$$\#\mathscr{H}(K,k)\le \#\mathscr{J}(K)(\#\mathscr{J}(K)-1).$$

□

The above corollary shows that the "quality" of the asymptotic as $\frac{\tau}{\theta}\to 0$ depends on $\|\cdot\|$, $\delta$ and on the dictionary through the terms $\mathbf{Q}_{K,k}$. The latter terms are bounded from above using (50) and (51) and they are clearly overestimated. Even though the bounds we provide are very pessimistic, they depend only on $\|\cdot\|$, $\delta$ and can be computed.

**Remark 7.** Let us emphasize that "uniform" bounds in the spirit of Corollary 1 can be derived from Propositions 3 and 5 as well. We leave this task to interested readers that need to compute easily the relevant bounds.

## 5. STATISTICAL MEANING OF THE RESULTS

In this section, we give a statistical interpretation of our main results, namely Theorem 2.

**Proposition 5.** *Let $\delta$ and $\|\cdot\|$ be any two norms and $A$ be a dictionary in $\mathbb{R}^N$. For any $K\in\{0,\dots,N\}$, let $\tau>0$ and $\theta$ be such that $\theta\ge\tau\Lambda_K$ where $\Lambda_K$ is defined in* (41)–(42). *Consider a random variable $d$ with uniform distribution on $B_\delta(\theta)$. Then*

$$\begin{aligned}&\frac{\mathbf{C}_K}{\mathbb{L}^N(B_\delta(1))}\left(\frac{\tau}{\theta}\right)^{N-K}\left(1-\hat{\lambda}_K\frac{\tau}{\theta}\right)^{K}-\frac{\varepsilon_0(K,\tau,\theta)}{\mathbb{L}^N(B_\delta(1))}\\ &\quad\le \mathbb{P}(\mathrm{val}(\mathscr{P}_d)\le K)\le \frac{\mathbf{C}_K}{\mathbb{L}^N(B_\delta(1))}\left(\frac{\tau}{\theta}\right)^{N-K}\left(1+\hat{\lambda}_K\frac{\tau}{\theta}\right)^{K},\end{aligned}$$

*where $\varepsilon_0(K,\tau,\theta)$ is given in Theorem 2, equation* (45). *Moreover we have the following asymptotical result:*

$$\mathbb{P}(\mathrm{val}(\mathscr{P}_d)\le K)=\frac{\mathbf{C}_K}{\mathbb{L}^N(B_\delta(1))}\left(\frac{\tau}{\theta}\right)^{N-K}+o\left(\left(\frac{\tau}{\theta}\right)^{N-K}\right)\quad as\ \frac{\tau}{\theta}\to 0.$$

*Proof.* Consider the set $\Sigma_K^\tau$ defined by (10). We have

$$\begin{aligned}\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le K) &= \mathbb{P}(d \in \Sigma_K^\tau \cap B_\delta(\theta)) \\ &= \frac{\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta))}{\mathbb{L}^N B_\delta(\theta)},\end{aligned}$$

since $d$ is uniformly distributed on $B_\delta(\theta)$. The inequality result follows from Theorem 2, equation (44) and uses the observation that $\mathbb{L}^N(B_\delta(\theta)) = \theta^N \mathbb{L}^N(B_\delta(1))$.

The asymptotical result is a direct consequence of Remark 5. □

**Remark 8.** Notice that, as already noticed in (34), $\mathbf{C}_N = \mathbb{L}^N(B_\delta(1))$ and the asymptotic in Proposition 5 reads for $K = N$

$$\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le N) = 1 + o(1) \quad \text{as } \frac{\tau}{\theta} \to 0.$$

In fact a better estimate is easy to obtain in this particular case. We know indeed that for all $d \in \mathbb{R}^N$, any solution of $\mathcal{P}_d$ involves an independent system of elements of $A$. (A sparser decomposition would otherwise exist.) Therefore, we know that for all $d \in \mathbb{R}^N$, $\mathrm{val}(\mathcal{P}_d) \le N$. This yields

$$\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le N) = 1.$$

Yet, again, this clearly shows that the bounds exhibited in this article are pessimistic.

## 6. VARIANTS OF THE AVERAGE PERFORMANCE IN APPROXIMATION

### 6.1. A Result on the Expectation

From our estimates of the law of $\mathrm{val}(\mathcal{P}_d)$, we can estimate its expectation. The expectation $\mathbb{E}(\mathrm{val}(\mathcal{P}_d))$ of $\mathrm{val}(\mathcal{P}_d)$ is defined by

$$\mathbb{E}(\mathrm{val}(\mathcal{P}_d)) = \sum_{K=1}^{N} K \mathbb{P}(\mathrm{val}(\mathcal{P}_d) = K).$$

Using that

$$\mathbb{P}(\mathrm{val}(\mathcal{P}_d) = K) = \mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le K) - \mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le K - 1)$$

and that $\mathbb{P}(\mathrm{val}(\mathcal{P}_d) = N) = 1$, we obtain

$$\mathbb{E}(\mathrm{val}(\mathcal{P}_d)) = N - \sum_{K=0}^{N-1} \mathbb{P}(\mathrm{val}(\mathcal{P}_d) \le K).$$

This yields the following theorem.

**Theorem 3.** *Let $\delta$ and $\|\cdot\|$ be any two norms and $A$ be a dictionary in $\mathbb{R}^N$. Let $\tau > 0$ and $\theta$ be such that $\theta \geq \tau \max_{0\leq K\leq N} \Lambda_K$ where $\Lambda_K$ is defined in* (41)–(42). *Consider a random variable $d$ with uniform distribution on $B_\delta(\theta)$. Then*

$$N - \sum_{K=0}^{N-1} \frac{\mathbf{C}_K}{\mathbb{L}^N(B_\delta(1))}\left(\frac{\tau}{\theta}\right)^{N-K}\left(1+\hat{\lambda}_K\frac{\tau}{\theta}\right)^K$$
$$\leq \mathbb{E}(val(\mathscr{P}_d)) \leq N - \sum_{K=0}^{N-1} \frac{\mathbf{C}_K}{\mathbb{L}^N(B_\delta(1))}\left(\frac{\tau}{\theta}\right)^{N-K}\left(1-\hat{\lambda}_K\frac{\tau}{\theta}\right)^K - \frac{\varepsilon_0(K,\tau,\theta)}{\mathbb{L}^N(B_\delta(1))}$$

*where $\varepsilon_0(K,\tau,\theta)$ is given in Theorem 2, equation* (45). *Moreover we have the following asymptotical result:*

$$\mathbb{E}(\mathrm{val}(\mathscr{P}_d)) = N - \frac{\mathbf{C}_{N-1}}{\mathbb{L}^N(B_\delta(1))}\frac{\tau}{\theta} + o\left(\frac{\tau}{\theta}\right) \quad as\ \frac{\tau}{\theta} \to 0.$$

### 6.2. A Result on Exactly $K$-Sparse Solutions

For any $K \in \{0,\ldots,N\}$ and $\tau > 0$, all data in $\mathbb{R}^N$ that lead to exactly $K$-sparse solutions read

$$\mathscr{D}_K^\tau \stackrel{\text{def}}{=} \{d \in \mathbb{R}^N : \mathrm{val}(\mathscr{P}_d) = K\}.$$

From the definition of $\Sigma_K^\tau$ in (10), it is straightforward that

$$\mathscr{D}_K^\tau = \Sigma_K^\tau \backslash \Sigma_{K-1}^\tau, \quad \forall K \in \{0,\ldots,N\},$$

where we extend the definition of $\Sigma_K^\tau$ with $\Sigma_{-1}^\tau = \emptyset$. Being the difference of two measurable closed sets, $\mathscr{D}_K^\tau$ is clearly measurable. Noticing also that

$$\Sigma_{K-1}^\tau \subset \Sigma_K^\tau,$$

we get

$$\mathbb{L}^N(\mathscr{D}_K^\tau \cap B_\delta(\theta)) = \mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta)) - \mathbb{L}^N(\Sigma_{K-1}^\tau \cap B_\delta(\theta)).$$

Said differently,

$$\mathbb{P}(\mathrm{val}(\mathscr{P}_d) = K) = \mathbb{P}(\mathrm{val}(\mathscr{P}_d) \leq K) - \mathbb{P}(\mathrm{val}(\mathscr{P}_d) \leq K-1).$$

All these equalities permit to derive statements analogue to those given in this article but for $\mathscr{D}_K^\tau$ and the event $\mathrm{val}(\mathscr{P}_d) = K$.

## 7. ILLUSTRATION: EUCLIDEAN NORMS FOR $\|\cdot\|$ AND $\delta$

In the example presented below, all constants derived in the previous sections admit an explicit form.

Consider the situation when both $\|\cdot\|$ and $\delta$ are the Euclidean norm on $\mathbb{R}^N$:

$$\|\cdot\| = \delta = \|\cdot\|_2 \quad \text{where } \|u\|_2 = \sqrt{\langle u, u\rangle}$$

with

$$\langle u, v\rangle = \sum_{i=1}^{N} u_i v_i.$$

Noticing that the Euclidean norm is rotation invariant, for any vector subspace $V \subseteq \mathbb{R}^N$ we have

$$\begin{aligned} P_{V^\perp} B_{\|\cdot\|_2}(\tau) &= V^\perp \cap B_{\|\cdot\|_2}(\tau) && (52)\\ &= \{u \in V^\perp : \|u\|_2 \le \tau\}. && (53) \end{aligned}$$

The equivalent norm $h$ and the constant $\eta$ derived in Lemma 2 are simply

$$\begin{aligned} h(u) &= \|u\|_2, \quad \forall u \in V^\perp,\\ \eta &= 1. \end{aligned}$$

The constant $\lambda_V$ in assertion (ii) of Proposition 2, defined by (68), reads $\lambda_V = 1$. Then the inequality condition on $\theta$ and $\tau$ is simplified to $\theta \ge \tau$.

The constant $C$ in (25) in the same proposition depends on $K$ (the dimension of the subspace $V$) and reads (see [13, p. 60] for details)

$$C = \mathscr{C}(K),$$

where $\mathscr{C}(K)$ is given by (48). Let us remind that using that $\Gamma(n+1) = n\Gamma(n)$, it comes

$$\mathscr{C}(K) = \frac{4\pi^{\frac{N}{2}}}{K(N-K)\Gamma\left(\frac{N-K}{2}\right)\Gamma\left(\frac{K}{2}\right)}. \tag{54}$$

From the preceding, the constants $\lambda_J$ and $C_J$ in Proposition 3 read

$$\begin{aligned} \lambda_J &= 1, \quad \forall J \subset I, && (55)\\ C_J &= \mathscr{C}(K), && (56) \end{aligned}$$

where the expression of $\mathscr{C}(K)$ is given in (54).

The norm $g$ arising in (75) in Proposition 4 reads

$$g(u) = \sup\{\|u_1\|_2 + \|u_2\|_2, \|u_1\|_2 + \|u_3\|_2\}$$
$$= \|u_1\|_2 + \sup\{\|u_2\|_2, \|u_3\|_2\}$$

where $u = u_1 + u_2 + u_3$ is decomposed according to (74). Then

$$\delta(u) = \|u\|_2 = \|u_1\|_2 + \|u_2\|_2 + \|u_3\|_2 \leq \lambda_{J_1,J_2} g(u), \quad \forall u \in W^\perp \quad \text{if } \lambda_{J_1,J_2} = 2.$$

The constants $\lambda_{J_1,J_2}$ and $Q_{J_1,J_2}$ in Proposition 4 read

$$\lambda_{J_1,J_2} = 2, \tag{57}$$
$$Q_{J_1,J_2} = \mathscr{C}(k)\ , \tag{58}$$

where we remind again that $\mathscr{C}(k)$ is defined according to (48).

For any $k = 1, \ldots, N$, the constants $\hat{\lambda}_k$ and $\mathbf{C}_k$ in (31)–(32) read

$$\hat{\lambda}_k = 1,$$
$$\mathbf{C}_k = \mathscr{C}(k)\ \#\mathscr{J}(k).$$

Clearly, $\#\mathscr{J}(K)$ depends on the dictionary $A$.

The constants $\widehat{\chi}_{K,k}$ and $\mathbf{Q}_{K,k}$, introduced in (37) and (38), respectively, are

$$\widehat{\chi}_{K,k} = 2, \tag{59}$$
$$\mathbf{Q}_{K,k} = \mathscr{C}(k)\#\mathscr{H}(K,k). \tag{60}$$

Here, again, $\#\mathscr{H}(K,k)$ depends on the choice of dictionary and in any case, $\#\mathscr{H}(K,k) = 0$ for $k < k_K$ (where $k_K$ is defined in (36)). The constant in (41)–(42) is $\Lambda_K = 2$ and the inequality (43) is satisfied.

The main inequality in Theorem 2 now reads

$$\mathscr{C}(K)\#\{\mathscr{J}(K)\}\tau^{N-K}(\theta-\tau)^K - \varepsilon_0(K,\tau,\theta) \leq \mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta))$$
$$\leq \mathscr{C}(K)\#\{\mathscr{J}(K)\}\tau^{N-K}(\theta+\tau)^K,$$

where $\mathscr{C}(K)$ is defined by (48) and the error term $\varepsilon_0(K,\tau,\theta)$ is

$$\varepsilon_0(K,\tau,\theta) = \frac{1}{2}\sum_{k=k_0}^{K-1} \mathscr{C}(k)\#\{\mathscr{H}(K,k)\}\tau^{N-k}(\theta+2\tau)^k.$$

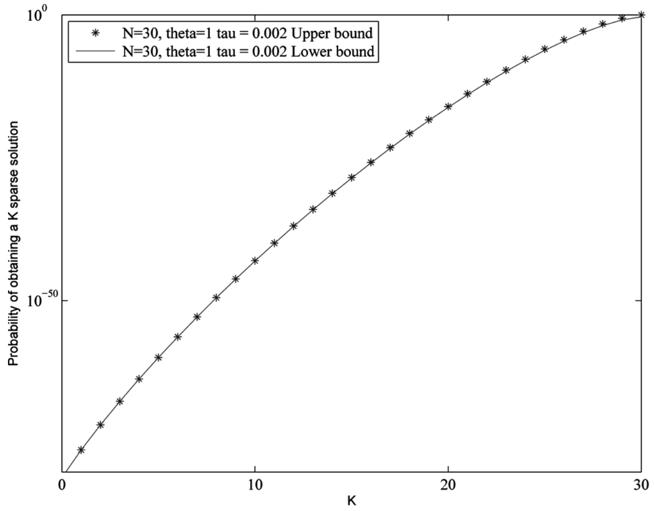


**FIGURE 3** Upper and lower bounds of $\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \leq K)$ in the context of Euclidean norms.

In order to provide the statistical interpretation in section 5, we notice that $\mathbb{L}^N(B_\delta(1)) = \alpha(N)$ for $\alpha(\cdot)$ as given in (47) and, hence,

$$\mathbb{L}^N(B_\delta(1)) = \frac{\pi^{N/2}}{\Gamma(N/2+1)}.$$

We display on Figure 3, a curve representing the upper and lower bounds of $\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \leq K)$ in the context of this illustration.

## 8. CONCLUSION AND PERSPECTIVES

In this article, we derive lower and upper bounds for different quantities concerning a model $(\mathcal{P}_d)$ of the form as defined in (1).

The intuitive interpretation of these bounds is that adding more columns to $A$ improves the approximation performances of $(\mathcal{P}_d)$. More precisely, the columns of $A$ should be such that

$$\mathbb{L}^K(\mathcal{A}_J \cap B_\delta(1))$$

are as large as possible for all $J$ such that $\dim(\mathcal{A}_J) = K$. Moreover, the data fidelity term should be such that

$$\mathbb{L}^{N-K}(P_{\mathcal{A}_J^\perp} B_{\|\cdot\|}(1))$$

is large. The performance is improved since these terms are in the numerator of the expression for $\mathbb{P}(\mathrm{val}(\mathcal{P}_d) \leq K)$, as seen in (12)–(13).

Typically, the difference between the upper and the lower bounds derived in this article has an order of magnitude $(\frac{\tau}{\theta})^{N-K+1}$, $0 < \frac{\tau}{\theta} \ll \Lambda_K$ (where $\Lambda_K$ is defined in (41)–(42)) while the quantities which are estimated are propositional to $\left(\frac{\tau}{\theta}\right)^{N-K}$. The difference between the upper and lower bounds is made of the quantities listed next.

- The terms $\theta \pm \lambda_V \tau$ which come from the inclusions $S_0 \subseteq V^\tau \cap B_\delta(\theta) \subseteq S_1$, in the proof Proposition 2. This approximation is of the order $\left(\frac{\tau}{\theta}\right)^{N-K+1}$. It may be possible to reach a larger order of magnitude (e.g., $\left(\frac{\tau}{\theta}\right)^{N-K+2}$) under the assumption that $\delta$ is regular away from 0 (e.g. twice differentiable). This would permit to improve Proposition 2 and the theorems that use its conclusions.
- A term of the form $-\theta^N \varepsilon_0(K, \tau, \theta)$ could be added to the upper bound in (44). This term is not present because of the approximation made in (82) in A.2. Such a term "$-\theta^N \varepsilon_0(K, \tau, \theta)$" could be obtained by computing the size of the intersection of more than two cylinder-like sets in Proposition 4 (doing so we would also avoid the approximation in (86)) and by improving this proposition by bounding $\mathbb{L}^N(\mathscr{A}^\tau_{J_1} \cap \mathscr{A}^\tau_{J_2} \cap B_\delta(\theta))$ from below. This may be a straightforward refinement of the current proof of Proposition 4.
  This improvement is possible but not necessary in this article since (again) this approximation yields an error whose order of magnitude is $\left(\frac{\tau}{\theta}\right)^{N-K+1}$. Anyway, we cannot get a better order of magnitude unless the approximation mentioned in the previous item is not improved (i.e., more regularity is assumed for $\delta$).

Besides those aspects, several future developments of this work can be envisaged:

- An important improvement would be to assume a more specialized form for the data distribution. One first step would be a distribution of the shape $\propto e^{-\beta\delta(d)}$, $\beta > 0$, which is continuous. In our opinion, one possible goal is to deal with a data distribution defined by a kernel. This is indeed one of the standard techniques used in machine learning theory to approximate data distributions.
- Another way of improvement is to adapt those results to the context of infinite dimensional spaces. This adaptation might not be trivial since (for instance) there is no Lebesgue measure in those spaces.
- A similar analysis can be performed for the Basis Pursuit Denoising (i.e., $\ell_1$ regularization; see, e.g., [8]) with the same asymptotic. It will clearly show what is in common and what are the differences between $\ell_0$ and $\ell_1$ penalization.
- Performing a similar analysis for the Orthogonal Matching Pursuit (see, e.g., [21, 22, 27]) would, of course, be an interesting and complementary result.

## APPENDIX

### A.1. Proof of Lemma 1

The case $J = \emptyset$ is trivial because of the convention $\mathrm{span}(\emptyset) = \{0\}$. Consider next that $J$ is nonempty. Let $w \in \mathcal{A}_J^\tau$ where $\mathcal{A}_J^\tau$ is defined in (6). Then $w$ admits a unique decomposition as

$$w = v + u \quad \text{where } v \in \mathcal{A}_J \text{ and } u \in \mathcal{A}_J^\perp.$$

From the definition of $\mathcal{A}_J^\tau$, there exists $w_u \in B_{\|\cdot\|}(\tau)$ such that $P_{\mathcal{A}_J^\perp} w_u = u$. Noticing that $u - w_u = P_{\mathcal{A}_J^\perp} w_u - w_u \in \mathcal{A}_J$ and that $v + u - w_u \in \mathcal{A}_J$, we can see that

$$\begin{aligned} w &= (v + u - w_u) + w_u \\ &\in \mathcal{A}_J + B_{\|\cdot\|}(\tau). \end{aligned}$$

Hence, $w \in \mathcal{A}_J^\tau$ as given in (14) in the lemma.

Conversely, let $w \in \mathcal{A}_J + B_{\|\cdot\|}(\tau)$. Then

$$w = v_1 + v, \quad \text{where } v_1 \in \mathcal{A}_J \text{ and } v \in B_{\|\cdot\|}(\tau).$$

Furthermore, $v$ has a unique decomposition of the form

$$v = v_2 + u \quad \text{where } v_2 \in \mathcal{A}_J \text{ and } u \in \mathcal{A}_J^\perp.$$

In particular,

$$u = P_{\mathcal{A}_J^\perp} v \in P_{\mathcal{A}_J^\perp} B_{\|\cdot\|}(\tau)$$

Combining this with the fact that $v_1 + v_2 \in \mathcal{A}_J$ shows that $w = (v_1 + v_2) + u \in \mathcal{A}_J^\tau$ as defined in (6).

### A.2. Proof of Lemma 2

The case $V = \{0\}$ is trivial (we obtain $h = \|\cdot\|$). Further we assume that $\dim(V) \geq 1$.

*Assertion (i).* The set $P_{V^\perp} B_{\|\cdot\|}(1)$ is convex since $\|\cdot\|$ is a norm and $P_{V^\perp}$ is linear. Moreover, the origin 0 belongs to its interior with respect to $V^\perp$. Indeed, there is $\varepsilon > 0$ such that if $w \in \mathbb{R}^N$ satisfies $\|w\|_2 < \varepsilon$, then $\|w\| < 1$. Consequently $0 \in \mathrm{Int}\big(B_{\|\cdot\|_2}(\varepsilon)\big) \subset B_{\|\cdot\|}(1)$. Using that $\|\cdot\|_2$ is rotationally invariant and that $P_{V^\perp}$ is a contraction, we deduce that $0 \in \mathrm{Int}\big(P_{V^\perp} B_{\|\cdot\|_2}(\varepsilon)\big) \subset P_{V^\perp} B_{\|\cdot\|}(1)$. Then the application $h : V^\perp \to \mathbb{R}$ in (18) is the usual Minkowski functional of $P_{V^\perp}B_{\|\cdot\|}(1)$, as defined and commented in [18, p. 131]. Since $P_{V^\perp}B_{\|\cdot\|}(1)$ is closed, we have

$$P_{V^\perp}B_{\|\cdot\|}(1) = \{u \in V^\perp : h(u) \leq 1\}.$$

Using that the Minkowski functional is positively homogeneous, that is,

$$h(cu) = ch(u), \quad \forall c > 0,$$

lead to the equality in (19).

*Assertion (ii).* For $h$ to be a norm, we have to show that

$$h(cu) = |c|h(u), \quad \forall c \in \mathbb{R}$$

(i.e., that $h$ is symmetric with respect to the origin). First, for any $c \in \mathbb{R}$

$$\begin{aligned} h(cu) &= \inf\{t \geq 0 : cu \in P_{V^\perp}B_{\|\cdot\|}(t)\} \\ &= \inf\left\{t \geq 0 : u \in P_{V^\perp}B_{\|\cdot\|}\left(\frac{t}{|c|}\right)\right\} \\ &= |c| \inf\{t \geq 0 : u \in P_{V^\perp}B_{\|\cdot\|}(t)\} \\ &= |c|h(u), \end{aligned} \tag{61}$$

where in (61) we use the facts that $P_{V^\perp}$ is linear and that $\|\cdot\|$ is a norm. Second, it is well known that the Minkowski functional is non negative, finite and satisfies[6] $h(u+v) \leq h(u) + h(v)$ for any $u, v \in V^\perp$.

Finally, since $B_h(0) = P_{V^\perp}B_{\|\cdot\|}(0) = \{0\}$, we have

$$h(u) = 0 \Leftrightarrow u = 0.$$

Consequently, $h$ defines a norm on $V^\perp$.

*Assertion (iii).* For any two norms $\delta$ and $\|\cdot\|$ on $\mathbb{R}^N$, there exist constants $\eta_1 > 0$ and $\eta_2 > 0$ satisfying

$$v \in \mathbb{R}^N \Rightarrow \delta(v) \leq \eta_1 \|v\|_2 \quad \text{and} \quad \|v\|_2 \leq \eta_2 \|v\| \cdot \tag{62}$$

[6] For completeness, we give the details:

$$\begin{aligned} h(u+v) &= \inf\{t \geq 0 : (u+v) \in P_{V^\perp}B_{\|\cdot\|}(t)\} \\ &\leq \inf\{t \geq 0 : u \in P_{V^\perp}B_{\|\cdot\|}(t)\} + \inf\{t \geq 0 : v \in P_{V^\perp}B_{\|\cdot\|}(t)\} \\ &= h(u) + h(v). \end{aligned}$$

Put

$$\eta \overset{\text{def}}{=} \eta_1\eta_2. \tag{63}$$

Let us first remark that

$$B_{\|\cdot\|}(1) \subset B_{\|\cdot\|_2}(\eta_2) \subset B_\delta(\eta_1\eta_2) = B_\delta(\eta).$$

Combining this with (19) and the fact that $\| \cdot \|_2$ is rotationally invariant, we have

$$\begin{aligned} B_h(1) = P_{V^\perp}B_{\|\cdot\|}(1) &\subset P_{V^\perp}B_{\|\cdot\|_2}(\eta_2) \\ &= B_{\|\cdot\|_2}(\eta_2) \cap V^\perp \qquad (64) \\ &\subset B_\delta(\eta_1\eta_2) \cap V^\perp \\ &= B_\delta(\eta) \cap V^\perp. \qquad (65) \end{aligned}$$

We will prove (20) and (21) jointly. To this end let us consider a norm $g$ on $\mathbb{R}^N$ and $\rho > 0$ such that

$$P_{V^\perp}B_{\|\cdot\|}(1) \subset B_g(\rho) \cap V^\perp. \tag{66}$$

Using that each norm can be expressed as a Minkowski functional, for any $u \in V^\perp$ we can write down the following:

$$\begin{aligned} g(u) &= \inf\left\{t \geq 0 : g\left(\frac{u}{t}\right) \leq 1\right\} \\ &= \inf\left\{t \geq 0 : g\left(\frac{\rho}{t}u\right) \leq \rho\right\} \\ &= \rho \inf\left\{t \geq 0 : g\left(\frac{u}{t}\right) \leq \rho\right\} \\ &= \rho \inf\left\{t \geq 0 : \frac{u}{t} \in B_g(\rho)\right\} \\ &\leq \rho \inf\left\{t \geq 0 : \frac{u}{t} \in P_{V^\perp}B_{\|\cdot\|}(1)\right\} \\ &= \rho h(u), \qquad (67) \end{aligned}$$

where the inequality in (67) comes from (66).

If we identify $g$ with $\delta$ and $\rho$ with $\eta$, (66) is satisfied according to (19) (for $\tau = 1$) and (65), and we obtain (20). Similarly, identifying $g$ with $\| \cdot \|_2$ and $\rho$ with $\eta_2$, (66) holds yet again by (19) (for $\tau = 1$) and (64), and this yields (21). This concludes the proof.

## A.3. Proof of Proposition 2

*Assertion (i).* The sets $V$ and $P_{V^\perp} B_{\|\cdot\|}(\tau)$ are closed. Moreover, $V$ and $P_{V^\perp} B_{\|\cdot\|}(\tau)$ are orthogonal. Therefore $V^\tau$ is closed. As a consequence $V^\tau$ is a Borel set and is Lebesgue measurable.

*Assertion (ii).* Since the restriction of $\delta$ to $V^\perp$ is a norm on $V^\perp$ and the function $h$ defined in (18) is also a norm (see Lemma 2(ii)) there exists $\lambda_V$ such that

$$\delta(u) \leq \lambda_V h(u), \quad \forall u \in V^\perp. \tag{68}$$

By (20) in Lemma 2, such a $\lambda_V$ belongs to $[0, \eta]$. To simplify the notations, in the rest of the proof we will write $\lambda$ for $\lambda_V$.

For any $u \in V^\perp$ and $v \in V$, using (68) we have

$$\delta(v) - \lambda h(u) \leq \delta(v) - \delta(u) \leq \delta(u+v) \leq \delta(v) + \delta(u) \leq \delta(v) + \lambda h(u).$$

In particular, for $h(u) \leq \tau$, we get

$$\delta(v) - \lambda\tau \leq \delta(u+v) \leq \delta(v) + \lambda\tau. \tag{69}$$

As required in assertion (ii), we have $\theta - \lambda\tau \geq 0$. If in addition $v \in V$ is such that $\delta(v) \leq \theta - \lambda\tau$, then

$$\delta(u+v) \leq \theta. \tag{70}$$

Noticing that

$$B_\delta(\theta) = \{u + v : (u,v) \in (V^\perp \times V), \delta(u+v) \leq \theta\},$$

this implies that

$$\begin{aligned} S_0 &\overset{\text{def}}{=} \{u + v : (u,v) \in (V^\perp \times V), h(u) \leq \tau, \delta(v) \leq \theta - \lambda\tau\} \\ &\subseteq V^\tau \cap B_\delta(\theta). \end{aligned}$$

Combining the left-hand side of (69) and (70) shows that $\delta(v) \leq \theta + \lambda\tau$, hence,

$$\begin{aligned} S_1 &\overset{\text{def}}{=} \{u + v : (u,v) \in (V^\perp \times V), h(u) \leq \tau, \delta(v) \leq \theta + \lambda\tau\} \\ &\supseteq V^\tau \cap B_\delta(\theta). \end{aligned}$$

Consider the pair of applications

$$\begin{aligned} \varphi_0 : B_h(1) \times (V \cap B_\delta(1)) &\to \mathbb{R}^N \\ (u,v) &\to \tau u + (\theta - \lambda\tau) v \end{aligned}$$

and

$$\varphi_1 : B_h(1) \times (V \cap B_\delta(1)) \to \mathbb{R}^N$$
$$(u, v) \to \tau u + (\theta + \lambda\tau)v$$

Clearly, whatever $i \in \{0, 1\}$, $\varphi_i$ is a Lipschitz homeomorphism satisfying $\varphi_i\big(B_h(1) \times \big(V \cap B_\delta(1)\big)\big) = S_i$. Moreover, their derivatives $D\varphi_i$ take the form:

$$D\varphi_0 = \begin{bmatrix} \tau \mathrm{I}_{N-K} & 0 \\ 0 & (\theta - \lambda\tau)\mathrm{I}_K \end{bmatrix}$$

and

$$D\varphi_1 = \begin{bmatrix} \tau \mathrm{I}_{N-K} & 0 \\ 0 & (\theta + \lambda\tau)\mathrm{I}_K \end{bmatrix}.$$

Then $\mathbb{L}^N(S_i)$ can be computed using (see [13] for details)

$$\mathbb{L}^N(S_i) = \int_{u \in B_h(1)} \int_{v \in V \cap B_\delta(1)} [\![\varphi_i]\!] dv\, du, \tag{71}$$

where $[\![\varphi_i]\!]$ is the Jacobian of $\varphi_i$, for $i = 0$ or $i = 1$. In particular,

$$[\![\varphi_0]\!] = \det(D\varphi_0) = \tau^{N-K}(\theta - \lambda\tau)^K,$$
$$[\![\varphi_1]\!] = \det(D\varphi_1) = \tau^{N-K}(\theta + \lambda\tau)^K.$$

It follows that

$$\mathbb{L}^N(S_0) = C\tau^{N-K}(\theta - \lambda\tau)^K$$

and

$$\mathbb{L}^N(S_1) = C\tau^{N-K}(\theta + \lambda\tau)^K,$$

where the constant $C$ (see (71)) reads

$$C = \int_{B_h(1)} du \int_{V \cap B_\delta(1)} dv$$
$$= \mathbb{L}^{N-K}(P_{V^\perp} B_{\|\cdot\|}(1))\mathbb{L}^K(V \cap B_\delta(1)).$$

Clearly, $C$ is positive and finite. Using the inclusion $S_0 \subseteq V^\tau \cap B_\delta(\theta) \subseteq S_1$ shows that

$$C\tau^{N-K}(\theta - \lambda\tau)^K \leq \mathbb{L}^N(V^\tau \cap B_\delta(\theta)) \leq C\tau^{N-K}(\theta + \lambda\tau)^K.$$

The proof is complete.

### A.4. Proof of Proposition 4

The subset in (28) is closed and measurable, as being a finite intersection of closed measurable sets (see Proposition 2).

Let

$$h_1 : \mathscr{A}_{J_1}^{\perp} \to \mathbb{R} \quad \text{and} \quad h_2 : \mathscr{A}_{J_2}^{\perp} \to \mathbb{R}$$

be the norms exhibited in Lemma 2. Then by statement (i) of the same lemma, for any $\tau \geq 0$ we have

$$B_{h_1}(\tau) = P_{\mathscr{A}_{J_1}^{\perp}} B_{\|\cdot\|}(\tau) \quad \text{and} \quad B_{h_2}(\tau) = P_{\mathscr{A}_{J_2}^{\perp}} B_{\|\cdot\|}(\tau).$$

Recall that by definition

$$W = \mathscr{A}_{J_1} \cap \mathscr{A}_{J_2}.$$

By De Morgan's law,

$$W^{\perp} = \mathscr{A}_{J_1}^{\perp} + \mathscr{A}_{J_2}^{\perp}.$$

Using that

$$\begin{aligned} \mathscr{A}_{J_1}^{\perp} &= \big(\mathscr{A}_{J_1}^{\perp} \cap \mathscr{A}_{J_2}^{\perp}\big) \oplus \big(\mathscr{A}_{J_1}^{\perp} \cap \mathscr{A}_{J_2}\big), \\ \mathscr{A}_{J_2}^{\perp} &= \big(\mathscr{A}_{J_1}^{\perp} \cap \mathscr{A}_{J_2}^{\perp}\big) \oplus \big(\mathscr{A}_{J_1} \cap \mathscr{A}_{J_2}^{\perp}\big), \end{aligned} \tag{72}$$

we can express $W^{\perp}$ as a direct sum of subspaces:

$$W^{\perp} = \big(\mathscr{A}_{J_1}^{\perp} \cap \mathscr{A}_{J_2}^{\perp}\big) \oplus \big(\mathscr{A}_{J_1}^{\perp} \cap \mathscr{A}_{J_2}\big) \oplus \big(\mathscr{A}_{J_1} \cap \mathscr{A}_{J_2}^{\perp}\big). \tag{73}$$

From (73), any $u \in W^{\perp}$ has a unique decomposition as

$$u = u_1 + u_2 + u_3 \quad \text{where} \begin{array}{l} u_1 \in \mathscr{A}_{J_1}^{\perp} \cap \mathscr{A}_{J_2}^{\perp} \\ u_2 \in \mathscr{A}_{J_1}^{\perp} \cap \mathscr{A}_{J_2} \\ u_3 \in \mathscr{A}_{J_1} \cap \mathscr{A}_{J_2}^{\perp} \end{array} \tag{74}$$

Let us introduce the function $g$, defined for all $u \in W^{\perp}$ by

$$g(u) = \sup\{h_1(u_1 + u_2), h_2(u_1 + u_3)\}, \tag{75}$$

where $u$ is decomposed according to (74). In the next lines, we show that $g$ is a norm on $W^{\perp}$:

- $h_1$ and $h_2$ being norms, $g(\lambda u) = |\lambda| g(u)$, for all $\lambda \in \mathbb{R}$;

- if $g(u) = 0$ then $u_1 + u_2 = u_1 + u_3 = 0$; noticing that $u_1 \perp u_2$ and that $u_1 \perp u_3$ yields $u = 0$;
- for $u \in W^\perp$ and $v \in W^\perp$ (both decomposed according to (74)),

$$\begin{aligned} g(u+v) &= \sup\{h_1(u_1+u_2+v_1+v_2), h_2(u_1+u_3+v_1+v_3)\} \\ &\le \sup\{h_1(u_1+u_2)+h_1(v_1+v_2),\ h_2(u_1+u_3)+h_2(v_1+v_3)\} \\ &\le \sup\{h_1(u_1+u_2), h_2(u_1+u_3)\} + \sup\{h_1(v_1+v_2), h_2(v_1+v_3)\} \\ &= g(u)+g(v). \end{aligned}$$

Furthermore, the norm $g$ on $W^\perp$ can be extended to a norm $\tilde{g}$ on $\mathbb{R}^N$ such that $\forall u \in W^\perp$, we have $\tilde{g}(u) = g(u)$ and

$$B_g(\tau) = P_{W^\perp} B_{\tilde{g}}(\tau), \quad \forall \tau > 0. \tag{76}$$

Let us then define

$$\begin{aligned} W^\tau &= W + P_{W^\perp} B_{\tilde{g}}(\tau) \\ &= \{w + u : (u, w) \in (W^\perp \times W), g(u) \le \tau\}. \end{aligned} \tag{77}$$

We are going to show that $(\mathscr{A}^\tau_{J_1} \cap \mathscr{A}^\tau_{J_2}) \subset W^\tau$. In order to do so, we consider an arbitrary

$$v \in \mathscr{A}^\tau_{J_1} \cap \mathscr{A}^\tau_{J_2}. \tag{78}$$

It admits a unique decomposition of the form

$$v = w + u_1 + u_2 + u_3,$$

where $w \in W$, and $u_1, u_2$, and $u_3$ are the components exhibited in (74). The latter, combined with (72) and

$$\begin{aligned} \mathscr{A}_{J_1} &= W \oplus \left(\mathscr{A}_{J_1} \cap \mathscr{A}^\perp_{J_2}\right), \\ \mathscr{A}_{J_2} &= W \oplus \left(\mathscr{A}^\perp_{J_1} \cap \mathscr{A}_{J_2}\right), \end{aligned}$$

shows that

$$\begin{aligned} u_1 + u_2 \in \mathscr{A}^\perp_{J_1} \quad &\text{and} \quad w + u_3 \in \mathscr{A}_{J_1}, \\ u_1 + u_3 \in \mathscr{A}^\perp_{J_2} \quad &\text{and} \quad w + u_2 \in \mathscr{A}_{J_2}. \end{aligned}$$

The inclusions given above, combined with (78), show that

$$h_1(u_1 + u_2) \le \tau \quad \text{and} \quad h_2(u_1 + u_3) \le \tau.$$

By the definition of $g$ in (74)–(75), the inequalities given above imply that $g(u) \leq \tau$. Combining this with the definition of $W^\tau$ in (77) entails that $v \in W^\tau$. Consequently,

$$(\mathscr{A}_{J_1}^\tau \cap \mathscr{A}_{J_2}^\tau) \subset W^\tau$$

and

$$\left(\mathscr{A}_{J_1}^\tau \cap \mathscr{A}_{J_2}^\tau \cap B_\delta(\theta)\right) \subset \left(W^\tau \cap B_\delta(\theta)\right).$$

It follows that

$$\mathbb{L}^N(\mathscr{A}_{J_1}^\tau \cap \mathscr{A}_{J_2}^\tau \cap B_\delta(\theta)) \leq \mathbb{L}^N(W^\tau \cap B_\delta(\theta)).$$

Let us choose $\lambda_{J_1,J_2} \geq 0$ such that

$$\delta(u) \leq \lambda_{J_1,J_2} g(u), \quad \forall u \in W^\perp, \tag{79}$$

Applying now the right-hand side of (24) in Proposition 2 with $W^\tau$ in place of $V^\tau$, $g$ in place of $h$ and $\lambda_{J_1,J_2}$ in place of $\eta$ leads to

$$\mathbb{L}^N(W^\tau \cap B_\delta(\theta))N \leq Q'_{J_1,J_2}\tau^{N-k}(\theta + \lambda_{J_1,J_2}\tau)^k,$$

where it is easy to see, using (25) in the same proposition, that

$$\begin{aligned} Q'_{J_1,J_2} &= \mathbb{L}^{N-k}(P_{W^\perp}B_{\tilde{g}}(\tau))\mathbb{L}^k(W \cap B_\delta(1)) \\ &= \mathbb{L}^{N-k}(B_g(1))\mathbb{L}^k(W \cap B_\delta(1)). \end{aligned} \tag{80}$$

In order to obtain (29), we are going to show that $B_g(1) \subset (B_{\|\cdot\|_2}(2\eta_2) \cap W^\perp)$. Using Lemma 2 (ii), if $u \in W^\perp$ is decomposed according to (74), we obtain

$$\begin{aligned} \|u\|_2 &= \left(\|u_1\|_2^2 + \|u_2\|_2^2 + \|u_3\|_2^2\right)^{\frac{1}{2}} \\ &\leq \|2u_1 + u_2 + u_3\|_2 \\ &\leq \|u_1 + u_2\|_2 + \|u_1 + u_3\|_2 \\ &\leq \eta_2 h_1(u_1 + u_2) + \eta_2 h_2(u_1 + u_3) \\ &\leq 2\eta_2 g(u). \end{aligned}$$

So $B_g(1) \subset (B_{\|\cdot\|_2}(2\eta_2) \cap W^\perp)$ and $Q'_{J_1,J_2} \leq Q_{J_1,J_2}$, for $Q_{J_1,J_2}$ as given in the proposition.

At last, we have to show that $\lambda_{J_1,J_2} \in [0, 3\eta]$. Using Lemma 2(ii), if $u \in W^\perp$ is decomposed according to (74), we obtain

$$\begin{aligned} \delta(u) &= \delta(u_1 + u_2 + u_3) \\ &\leq \delta(2u_1 + u_2 + u_3) + \delta(u_1) \end{aligned}$$

$$\begin{aligned}&\leq \delta(u_1+u_2)+\delta(u_1+u_3)+\delta(u_1)\\ &\leq \eta h_1(u_1+u_2)+\eta h_2(u_1+u_3)+\eta_1\|u_1\|_2, \qquad (81)\end{aligned}$$

where $\eta_1$ is defined in (62), in the proof of Lemma 2.

Using (21) in Lemma 2, $\|u_1\|_2$ satisfies the following two inequalities

$$\begin{aligned}\|u_1\|_2 \leq \|u_1+u_2\|_2 \leq \eta_2 h_1(u_1+u_2),\\ \|u_1\|_2 \leq \|u_1+u_3\|_2 \leq \eta_2 h_2(u_1+u_3).\end{aligned}$$

Adding these inequalities and using (62)–(63), we obtain

$$\eta_1\|u_1\|_2 \leq \frac{\eta}{2}(h_1(u_1+u_2)+h_2(u_1+u_3)).$$

Using (81), we finally conclude that, for $u \in W^{\perp}$

$$\begin{aligned}\delta(u) &\leq \frac{3\eta}{2}\ (h_1(u_1+u_2)+h_2(u_1+u_3))\\ &\leq 3\eta\ g(u).\end{aligned}$$

The proof is complete.

### A.5. Proof of Theorem 2

When $K=0$ or $K=N$, we have $\#\mathcal{J}(K)=1$. Using (40), we have $\varepsilon_0(K,\tau,\theta)=0$. By (17) and the assumption $\theta > \tau\Lambda_K$, we have $\Sigma_0^\tau \cap B_\delta(\theta) = B_{\|\cdot\|}(\tau)$ and $\Sigma_N^\tau \cap B_\delta(\theta) = B_\delta(\theta)$. Note that we can take $\Lambda_N = 0$. Combining these facts with (34) shows that (44) holds true and that it is an equality.

The remaining of the proof is to find relevant bounds for the right-hand side of (17) under the assumption that $1 \leq K \leq N-1$.

*Upper bound.* By (17) and the definition of a measure, and using Proposition 3, it is found that

$$\begin{aligned}\mathbb{L}^N(\Sigma_K^\tau \cap B_\delta(\theta)) &\leq \sum_{J\in\mathcal{J}(K)} \mathbb{L}^N(\mathscr{A}_J^\tau \cap B_\delta(\theta))\\ &\leq \tau^{N-K} \sum_{J\in\mathcal{J}(K)} C_J(\theta+\lambda_J\tau)^K\\ &\leq \tau^{N-K}(\theta+\hat{\lambda}_K\tau)^K \sum_{J\in\mathcal{J}(K)} C_J\\ &= \mathbf{C}_K \tau^{N-K}(\theta+\hat{\lambda}_K\tau)^K, \qquad (82)\end{aligned}$$

where the constants $\hat{\lambda}_K$ and $\mathbf{C}_K$ are defined in (31) and (32), respectively. This establishes the right hand side inequality in (44).

*Lower bound.* First we represent the right side of (16) as a union of disjoint subsets. Since $\mathcal{J}(K)$ is finite, let us enumerate its elements as

$$\mathcal{J}(K) = \{J_1, \ldots, J_L\} \quad \text{where } L = \#(\mathcal{J}(K)).$$

To simplify the expressions that follow, for any $J$ we denote

$$B_J \overset{\text{def}}{=} \mathscr{A}_J^\tau \cap B_\delta(\theta). \tag{83}$$

Then

$$\bigcup_{J \in \mathcal{J}(K)} (\mathscr{A}_J^\tau \cap B_\delta(\theta)) = \bigcup_{i=1}^{L} B_{J_i}.$$

Consider the following decomposition:

$$\bigcup_{i=1}^{L} B_{J_i} = (B_{J_1}) \cup \bigcup_{i=2}^{L} \left( B_{J_i} \Bigg\backslash \left( \bigcup_{j=1}^{i-1} (B_{J_j} \cap B_{J_i}) \right) \right).$$

Since the last row is a union of disjoint sets, we have

$$\mathbb{L}^N \left( \bigcup_{i=1}^{L} B_{J_i} \right) = \mathbb{L}^N (B_{J_1}) + \sum_{i=2}^{L} \mathbb{L}^L \left( B_{J_i} \Bigg\backslash \left( \bigcup_{j=1}^{i-1} (B_{J_j} \cap B_{J_i}) \right) \right).$$

Noticing that $\left( \bigcup_{j=1}^{i-1} (B_{J_j} \cap B_{J_i}) \right) \subset B_{J_i}$ entails that for all $i = 2, \ldots, L$, we have

$$\mathbb{L}^N \left( B_{J_i} \Bigg\backslash \left( \bigcup_{j=1}^{i-1} (B_{J_j} \cap B_{J_i}) \right) \right) = \mathbb{L}^N (B_{J_i}) - \mathbb{L}^N \left( \bigcup_{j=1}^{i-1} (B_{J_j} \cap B_{J_i}) \right).$$

Hence,

$$\mathbb{L}^N \left( \bigcup_{i=1}^{L} B_{J_i} \right) = \sum_{i=1}^{L} \mathbb{L}^N (B_{J_i}) - \sum_{i=2}^{L} \mathbb{L}^N \left( \bigcup_{j=1}^{i-1} (B_{J_j} \cap B_{J_i}) \right). \tag{84}$$

Using successively (83), Proposition 3, the definitions of $\hat{\lambda}_K$ and $\mathbf{C}_K$ in (31) and (32), respectively, and the assumption that $\theta \geq \tau \Lambda_K$, shows that

$$\sum_{i=1}^{L} \mathbb{L}^N (B_{J_i}) = \sum_{J \in \mathcal{J}(K)} \mathbb{L}^N (\mathscr{A}_J^\tau \cap B_\delta(\theta))$$

$$\begin{aligned}&\geq \sum_{J\in\mathcal{J}(K)} C_J\tau^{N-K}(\theta-\lambda_J\tau)^K\\ &\geq \mathbf{C}_K\tau^{N-K}(\theta-\hat{\lambda}_K\tau)^K. \qquad (85)\end{aligned}$$

Note that by the definition of $\Lambda_K$ in (41) we have $\hat{\lambda}_K \leq \Lambda_K$.

Using the original notation (83), each term, for $i = 2, \ldots, L$, in the last sum in (84) satisfies

$$\mathbb{L}^N\left(\bigcup_{j=1}^{i-1}(B_{J_j}\cap B_{J_i})\right) \leq \sum_{j=1}^{i-1}\mathbb{L}^N(B_{J_j}\cap B_{J_i}) = \sum_{j=1}^{i-1}\mathbb{L}^N(B_\delta(\theta)\cap\mathscr{A}^\tau_{J_j}\cap\mathscr{A}^\tau_{J_i}). \qquad (86)$$

Let us remind that $\dim(\mathscr{A}_{J_i}) = K$ for every $i = 1, \ldots, L$ and that by the definition of $\mathcal{J}(K)$ (see (8)) we have $\mathscr{A}_{J_j} \neq \mathscr{A}_{J_i}$ if $i \neq j$. Proposition 4 can, hence, be applied to each term of the last sum:

$$\mathbb{L}^N(B_\delta(\theta)\cap\mathscr{A}^\tau_{J_j}\cap\mathscr{A}^\tau_{J_i}) \leq Q_{J_i,J_j}\tau^{N-k_{i,j}}(\theta+\lambda_{J_i,J_j})\tau)^{k_{i,j}}$$
$$\text{where } k_{i,j} = \dim(\mathscr{A}_{J_j}\cap\mathscr{A}_{J_i}).$$

Then (86) leads to

$$\mathbb{L}^N\left(\bigcup_{j=1}^{i-1}(B_{J_j}\cap B_{J_i})\right) \leq \sum_{j=1}^{i-1} Q_{J_j,J_i}\tau^{N-k_{i,j}}(\theta+\lambda_{J_j,J_i}\tau)^{k_{i,j}}.$$

By rearranging the last sum in (84) and taking into account (36), we obtain

$$\sum_{i=2}^{L}\mathbb{L}^N\left(\bigcup_{j=1}^{i-1}(B_{J_j}\cap B_{J_i})\right) \leq \sum_{k=k_K}^{K-1}\mathbf{Q}_{K,k}\tau^{N-k}(\theta+\widehat{\chi}_{K,k}\tau)^k, \qquad (87)$$

where $\widehat{\chi}_{K,k}$ and $\mathbf{Q}_{K,k}$ are given in (37) and (38), respectively.

Combining (17) along with the original notations (83) and then (84), (85), and (87) yield

$$\begin{aligned}\mathbb{L}^N(\Sigma^\tau_K\cap B_\delta(\theta)) &= \mathbb{L}^N\left(\bigcup_{i=1}^{L}B_{J_i}\right)\\ &\geq \mathbf{C}_K\tau^{N-K}(\theta-\hat{\lambda}_K\tau)^K - \varepsilon_0(K,\tau,\theta), \qquad (88)\end{aligned}$$

where $\varepsilon_0(\cdot)$ is as in the proposition. This finishes the proof.

## REFERENCES

1. J.E. Besag (1986). On the statistical analysis of dirty pictures (with discussion). *J. R. Stat. Soc. B* 48:259–302.
2. J.E. Besag (1989). Digital image processing: Towards Bayesian image analysis. *J. Appl. Stat.* 16:395–407.
3. T. Blumensath and M. Davies (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmonic Anal.* 27:265–274.
4. A.M. Bruckstein, D.L. Donoho, and M. Elad (2009). From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.* 51:34–81.
5. A. Cohen, W. Dahmien, and R.A. DeVore (2009). Compressed sensing and best k-term approximation. *J. Am. Math. Soc.* 22:211–231.
6. R.R. Coifman and M.V. Wickerhauser (1992). Entropy-based algorithms for best basis selection. *IEEE Trans. Information Theory* 38:713–718.
7. Compressive Sensing Resources. [Online]. Available at http://dsp.rice.edu/cs
8. S.S. Chen, D.L. Donoho, and M.A. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* 43:129–159.
9. W. Dai and O. Milenkovic (2009). Subspace pursuit for compressive sensing signal reconstruction. *IEEE IT* 55:2230–2249.
10. G. Davis, S. Mallat, and M. Avellaneda (1997). Adaptive greedy approximations. *Constructive Approximation* 13:57–98.
11. R.A. DeVore (1998). Nonlinear approximation. *Acta Numerica* 7:51–150.
12. D. Donoho, I. Johnstone, J. Hoch, and A. Stern (1992). Maximum entropy and the nearly black object. *J. R. Stat. Soci. B* 54:41–81.
13. L.C. Evans and R.F. Gariepy (1992). *Measure Theory and Fine Properties of Functions.* Studies in Advanced Mathematics, CRC Press, Roca Baton, FL.
14. J.-J. Fuchs (2005). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Trans. Information Theory* 51:3601–3608.
15. S. Geman and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE PAMI* 6:721–741.
16. Y. Leclerc (1989). Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision* 3:73–102.
17. S. Li (1995). *Markov Random Field Modeling in Computer Vision* (1st ed.). Springer-Verlag, London.
18. D. Luenberger (1969). *Optimization by Vector Space Methods* (1st edn.). John Wiley.
19. S. Mallat and Z. Zhang (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Sign. Process.* 41:3397–3415.
20. D. Needell and J.A. Tropp (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmonic Anal.* 26:301–321.
21. D. Needell and R. Vershynin (2009). Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Compu. Math.* 9:317–334.
22. Y. Pati, R. Rezaiifar, and P. Krishnaprasad (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *The 27th Annual Asilomar Conference on Signals, Systems, and Computers,* Vol. 1, pp. 40–44. Available at: http://www.asilomarssc.org
23. G. Pisier (1989). *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press, Cambridge, UK.
24. M. Robini, A. Lachal, and I. Magnin (2007). A stochastic continuation approach to piecewise constant reconstruction. *IEEE Trans. Image Process.* 16:2576–2589.
25. J. Tropp (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Information Theory* 50:2231–2242.
26. J. Tropp (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Tran. Information Theory* 52:1030–1051.
27. J. Tropp and A. Gilbert (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Information Theory* 53:4655–4666.