



**HAL**  
open science

## Average performance of the approximation in a dictionary using an l0 objective

François Malgouyres, Mila Nikolova

► **To cite this version:**

François Malgouyres, Mila Nikolova. Average performance of the approximation in a dictionary using an l0 objective. *Comptes Rendus. Mathématique*, 2009, 347 (9-10), pp.565-570. 10.1016/j.crma.2009.02.026 . hal-01486812

**HAL Id: hal-01486812**

**<https://hal.science/hal-01486812>**

Submitted on 14 Mar 2017

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

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

# Average performance of the approximation in a dictionary using an $\ell_0$ objective

François Malgouyres<sup>a</sup>, Mila Nikolova<sup>b</sup>

<sup>a</sup>LAGA - CNRS UMR 7539, Université Paris 13, 99 avenue J.B. Clément, F-93430 Villetaneuse, France

<sup>b</sup>CMLA, ENS Cachan, CNRS, PRES UniverSud, 61 Av. Président Wilson, F-94230 Cachan, France

Received \*\*\*\*\*, accepted after revision +++++

Presented by

---

## Abstract

We consider the minimization of the number of non-zero coefficients (the  $\ell_0$  “norm”) of the representation of a data set in a general dictionary under a fidelity constraint. This (nonconvex) optimization problem leads to the sparsest approximation. The average performance of the model consists in the probability (on the data) to obtain a  $K$ -sparse solution—involving at most  $K$  nonzero components—from data uniformly distributed on a domain. These probabilities are expressed in terms of the parameters of the model and the accuracy of the approximation. We comment the obtained formulas and give a simulation. *To cite this article: F. Malgouyres, M. Nikolova, C. R. Acad. Sci. Paris, Ser. I XXX (2009).*

## Résumé

**Performance moyenne de l’approximation la plus parcimonieuse avec un dictionnaire** Nous tudions la minimisation du nombre de coefficients non-nuls (la “norme”  $\ell_0$ ) de la représentation d’un ensemble de données dans un dictionnaire arbitraire sous une contrainte de fidélité. Ce problème d’optimisation (non-convexe) mne naturellement aux représentations les plus parcimonieuses. La performance moyenne du modèle est décrite par la probabilité que les données donnent une solution  $K$ -parcimonieuse - contenant pas plus de  $K$  composantes non-nulles - en supposant que les données sont uniformément distribuées sur un domaine. Ces probabilités s’expriment en fonction des paramètres du modèle et de la précision de l’approximation. Nous commentons les formules obtenues et fournissons une illustration. *Pour citer cet article : F. Malgouyres, M. Nikolova, , C. R. Acad. Sci. Paris, Ser. I XXX (2009).*

---

## Version française abrégée

Étant donné un dictionnaire  $A = \{a_1, \dots, a_M\} \in \mathbb{R}^{N \times M}$ ,  $\text{rank}(A) = N$ , nous considérons l’approximation la plus parcimonieuse  $u^* \in \mathbb{R}^M$  des données observées  $d \in \mathbb{R}^N$  telle que  $Au^* \approx d$ . Celle-ci est une solution du problème d’optimisation sous contrainte  $(\mathcal{P}_d)$ , cf. (2), où  $\#$  dénote cardinalité,  $\|\cdot\|$  est une norme et  $\tau > 0$  un paramètre. Pour chaque  $d \in \mathbb{R}^N$ , la contrainte dans (2) est non vide et  $u^*$  vérifie (3).

$(\mathcal{P}_d)$  n’est qu’une réécriture de la bien connue *meilleure approximation à  $K$  termes* (BKTA), cf. [3], qui résout (4). Sous des hypothèses sur le support de la distribution des données, BKTA fournit des bornes de la

---

*Email addresses:* malgouy@math.univ-paris13.fr (François Malgouyres), nikolova@cmla.ens-cachan.fr (Mila Nikolova).

forme (5), où  $u_K^*$  est une solution de (4) pour  $K \rightarrow \infty$ . En d'autres mots, l'évaluation des performances est gouvernée par les pires données  $d$ , même si c'est une situation invraisemblable. En outre, les résultats sont vagues quand  $M > N$  car les performances sont déduites *via* différentes approximations de (4).

Afin d'évaluer les performances de (2), nous proposons la méthodologie dite Moyenne Performance en Approximation : elle consiste en l'estimation de  $\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K)$ ,  $\forall K \in \{0, \dots, N\}$  en fonction de  $\tau, \|\cdot\|$  et  $A$ , où la probabilité porte sur la variable aléatoire  $d$  dont la loi est connue. Par exemple, on peut modéliser des données parcimonieuses comme uniformément distribués dans une  $\ell_1$  boule. Plus  $\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K)$  est grande pour  $K$  petit, meilleur est le modèle. Cette note résume les résultats obtenus dans [6]. Nous espérons qu'elle ouvre une nouvelle voie pour de futures recherches.

### Principaux résultats

Pour  $\theta > 0$  et toute fonction  $f : \mathbb{R}^N \rightarrow \mathbb{R}$ , les  $\theta$  sous-ensembles de niveau sont définis par (6). Pour tout  $J \subset \{1, \dots, M\}$ , le sous espace  $\mathcal{A}_J$  est défini dans (7), son complément orthogonal est désigné par  $\mathcal{A}_J^\perp$  et la projection orthogonale sur  $\mathcal{A}_J^\perp$  par  $P_{\mathcal{A}_J^\perp}$ . Le sous ensemble  $\mathcal{A}_J^\tau$  est défini dans (8). Pour tout  $K \leq N$ , nous posons (cf [6] pour les détails)  $\mathcal{J}(K)$  par (9), avec la convention que  $\mathcal{J}(0) = \{\emptyset\}$ . Le théorème 0.1 est la clé pour les résultats qui suivent.

**Théorème 0.1** *Pour toute matrice  $A \in \mathbb{R}^{N \times M}$ ,  $\text{rang}(A) = N$  et pour toute norme  $\|\cdot\|$ , on a*

$$K \in \{0, \dots, N\} \Rightarrow \{d \in \mathbb{R}^N, \text{val}(\mathcal{P}_d \leq K)\} = \bigcup_{J \in \mathcal{J}(K)} \mathcal{A}_J^\tau.$$

Pour tout  $n \in \mathbb{N}$ , la mesure de Lebesgue de  $C \subset \mathbb{R}^n$  est notée par  $\mathbb{L}^n(C)$ . Nous définissons les constantes  $C_J$  comme dans (10) où  $f_d$  est une norme telle que  $d$  est uniforme sur  $\theta B_{f_d}(1)$ . Pour tout  $K = 0, \dots, N$ , les constantes  $\overline{C}_K$  sont définies dans (11). A l'aide de ces notations, nous présentons le résultat principal de [6].

**Théorème 0.2** *Soit  $f_d$  et  $\|\cdot\|$  deux normes, et  $A \in \mathbb{R}^{N \times M}$  avec  $\text{rang}(A) = N$ . Pour tout  $\theta > 0$ , soit  $d$  une variable aléatoire uniformément distribuée sur  $B_{f_d}(\theta)$ . Alors,  $\forall K \in \{0, \dots, N\}$*

$$\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K) = \overline{C}_K (\tau/\theta)^{N-K} + o\left((\tau/\theta)^{N-K}\right) \text{ pour } \tau/\theta \rightarrow 0. \quad (1)$$

### Commentaires sur les résultats

**(a) Une limitation.** D'après le théorème 0.2, la précision de  $\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K)$  augmente lorsque  $\tau/\theta \rightarrow 0$ . Notons que la fonction  $o(\cdot)$  dépend de tous les ingrédients du modèle ( $f_d, A, \|\cdot\|, K$  et  $N$ ). Afin d'améliorer les bornes pour des  $\tau/\theta$  raisonnables, il nous faut des mesures plus précises des ensembles  $\mathcal{A}_J^\tau$  et leurs intersection. Dans des cas très généraux, on se heurte des problèmes ouverts en mathématiques fondamentales, cf. [7].

**(b) Amélioration cruciale en approximation.** Si l'on enrichit le dictionnaire avec un *nouvel* élément  $a_{M+1} \in \mathbb{R}^N$  (non collinéaire avec aucune de colonnes de  $A$ ), alors la valeur de  $\overline{C}_K$  dans (11) augmente et en conséquence la probabilité dans (12) augmente. *Donc le modèle s'améliore.* Cependant, cela ne changera rien dans l'évaluation proposé par BKTA.

**(c) Le rôle de  $\|\cdot\|$ .** Dans (2) et (4), le choix de  $\|\cdot\|$  est d'une importance cruciale. Les équations (10), (11) et (12) suggèrent d'optimiser une somme pondérée des mesures de  $P_{\mathcal{A}_J^\perp}(B_{\|\cdot\|}(1))$ .

**(d) Approximation versus compressed-sensing (CS).** Les hypothèses typiques en CS mènent à des matrices spéciales  $A$  telles que les sous espaces vectoriels dans  $\mathcal{J}(K)$  sont presque orthogonaux quand  $K$  est assez petit. Alors  $\mathbb{L}^N(\mathcal{A}_J^\tau \cap \mathcal{A}_{J'}^\tau \cap B_{f_d}(\theta))$ , pour  $(J, J') \in (\mathcal{J}(K))^2$  diminue, mais de toutes faons ces termes sont asymptotiquement négligeables. l'impact d'une telle hypothèse sur les  $\overline{C}_K$  est une question ouverte.

Avant d'appliquer les résultats en CS de [8,4] en approximation, il est important d'évaluer le hiatus entre la performance de  $(\mathcal{P}_d)$  sous les hypothèses de base en CS (voir [1]), sans restrictions supplémentaires.

Enfin, si l'on utilise  $(\mathcal{P}_d)$  en approximation, l'ajout d'une nouvelle colonne  $A$  améliore toujours la performance de  $(\mathcal{P}_d)$  (voir (b)). Si  $(\mathcal{P}_d)$  est utilisé en CS, la nouvelle colonne peut dégrader sa performance.

(e) **Distribution des donnes.** Une analyse de (10)-(11) montre que le dictionnaire  $A$  doit être construit de telle sorte que  $\mathbb{L}^K(\mathcal{A}_J \cap B_{f_d}(1)) / \mathbb{L}^N(B_{f_d}(1))$  est le plus grand possible en particulier quand  $\#J$  est petit.

## 1. The $\ell_0$ approximation and the evaluation of its performance

We deal with sparse approximation of observed data  $d \in \mathbb{R}^N$  using a dictionary  $A = \{a_1, \dots, a_M\}$ —an  $N \times M$  matrix with  $M \geq N$  and  $\text{rank}(A) = N$ . The *sparsest* vector  $u^* \in \mathbb{R}^M$  such that  $Au^* \approx d$  is a solution of the constraint optimization problem  $(\mathcal{P}_d)$  below:

$$(\mathcal{P}_d) : \begin{cases} \text{minimize}_{u \in \mathbb{R}^M} \ell_0(u), & \text{where } \ell_0(u) \stackrel{\text{def}}{=} \#\{1 \leq i \leq M : u_i \neq 0\}, \\ \text{under the constraint : } & \|Au - d\| \leq \tau, \end{cases} \quad (2)$$

where  $\#$  means cardinality,  $\|\cdot\|$  is a norm (e.g. the  $\ell_2$  norm) and  $\tau > 0$  is a parameter. For any  $d \in \mathbb{R}^N$ , the constraint in  $(\mathcal{P}_d)$  is nonempty and the minimum is reached for an  $u^*$  such that

$$\text{val}(\mathcal{P}_d) \stackrel{\text{def}}{=} \ell_0(u^*) \leq N. \quad (3)$$

Problem  $(\mathcal{P}_d)$  is simply a different way to parameterize the so called *best  $K$ -term approximation* (BKTA), defined as a solution to

$$\begin{cases} \text{minimize}_{u \in \mathbb{R}^M} \|Au - d\|, \\ \text{under the constraint : } \ell_0(u) \leq K, \text{ where } 0 \leq K \leq N. \end{cases} \quad (4)$$

The evaluation of the performances of (4) is a well developed field, see [3]. It is named *non-linear approximation* if  $M = N$  and *highly non-linear approximation* is  $M > N$ . It is mostly developed for infinite dimensional vector spaces. Under hypotheses on the support of the data distribution, it gives bounds of the form

$$\|Au_K^* - d\| \leq C/K^\alpha, \text{ for } C > 0, \alpha > 0, \quad (5)$$

where  $u_K^*$  is the BKTA of  $d$ . Doing so, it gives the asymptotic behavior of  $\|Au^* - d\|$  when  $K$  goes to infinity. The performance of the BKTA is governed by the worst data  $d$ , even if it is an unlikely scenario.

Unfortunately the results when  $M > N$  are vague: the bounds are typically derived from a heuristic approximation of the solution of (4) (e.g. the Orthogonal Matching Pursuit (OMP)). The resultant performances are heavily biased by the heuristic approximation. Moreover, since the BKTA is NP-hard (see [2]), another critical point is to discriminate between different heuristics. In particular, the non linear approximation theory fails to discriminate between  $\ell_1$  approximation (where  $\ell_0$  is replaced by  $\ell_1$ ) and greedy approaches such as OMP. The latter open question is a reasonable perspective for the Average Performance in Approximation (APA).

To evaluate the performance of the sparsest approximation (i.e.  $(\mathcal{P}_d)$ ), we propose the Average Performance in Approximation methodology which consists in estimating for every  $K = 0, \dots, N$

$$\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K) \text{ as a function of } \tau, \|\cdot\| \text{ and } A,$$

where the probability is on  $d$  and  $d$  is a random variable of known distribution law. For example, sparse data are typically modeled as uniformly distributed in an  $\ell_1$  ball. The larger  $\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K)$ , for  $K$  small, the better the model. This note summarizes the results obtained in [6] and emphasizes their meaning. We hope it also provides a focus for future research.

## 2. Main results

For  $\theta > 0$  and any real function  $f$  defined on  $\mathbb{R}^N$ , we denote the  $\theta$ -level set of  $f$  by

$$B_f(\theta) = \{w \in \mathbb{R}^N, f(w) \leq \theta\}, \quad \theta > 0. \quad (6)$$

Let us denote, for  $J \subset \{1, \dots, M\}$ ,

$$\mathcal{A}_J = \text{span}\{a_j : j \in J\}, \text{ where } a_j \text{ is the } j^{\text{th}} \text{ column of the matrix } A. \quad (7)$$

We systematically denote by  $\mathcal{A}_J^\perp$  the orthogonal complement of  $\mathcal{A}_J$  in  $\mathbb{R}^N$  and by  $P_{\mathcal{A}_J^\perp}$  the orthogonal projector onto  $\mathcal{A}_J^\perp$ . We set

$$\mathcal{A}_J^\tau \stackrel{\text{def}}{=} \mathcal{A}_J + B_{\|\cdot\|}(\tau) \text{ and we have } \mathcal{A}_J^\tau = \mathcal{A}_J + P_{\mathcal{A}_J^\perp} (B_{\|\cdot\|}(\tau)). \quad (8)$$

Geometrically,  $\mathcal{A}_J^\tau$  is an infinite cylinder in  $\mathbb{R}^N$ : like a  $\tau$ -thick coat wrapping the subspace  $\mathcal{A}_J$ .

For any given dimension  $K \leq N$ , we set that (see [6] for the precise definition)

$$\mathcal{J}(K) \text{ is a maximal non-redundant listing of all subspaces } \mathcal{A}_J \subset \mathbb{R}^N \text{ of dimension } K; \quad (9)$$

We always have  $\#\mathcal{J}(N) = 1$  and set by convention  $\mathcal{J}(0) = \{\emptyset\}$ . Theorem 2.1 is the key for the results that follow. It provides an easy geometrical vision of the problem.

**Theorem 2.1** *For any  $A \in \mathbb{R}^{N \times M}$  with  $\text{rank}(A) = N$ , any norm  $\|\cdot\|$ ,  $\forall \tau > 0$  and  $\forall K \in \{0, \dots, N\}$ , we have*

$$\{d \in \mathbb{R}^N, \text{val}(\mathcal{P}_d \leq K)\} = \bigcup_{J \in \mathcal{J}(K)} \mathcal{A}_J^\tau.$$

For any  $n \in \mathbb{N}$ , the Lebesgue measure of  $C \subset \mathbb{R}^n$  is denoted by  $\mathbb{L}^n(C)$ . Let us now define

$$C_J = \mathbb{L}^{N-K} (P_{\mathcal{A}_J^\perp} (B_{\|\cdot\|}(1))) \mathbb{L}^K (\mathcal{A}_J \cap B_{f_d}(1)), \text{ where } K = \dim(\mathcal{A}_J) \quad (10)$$

and  $f_d$  is a norm such that  $d$  is uniform on  $\theta B_{f_d}(1)$  (typically  $f_d$  is the  $\ell_1$  norm for sparse data).

For any  $K = 0, \dots, N$ , define the constants  $\overline{C}_K$  as it follows:

$$\overline{C}_K = \frac{\sum_{J \in \mathcal{J}(K)} C_J}{\mathbb{L}^N (B_{f_d}(1))}. \quad (11)$$

Using these notations, we can state the main result of [6].

**Theorem 2.2** *Let  $f_d$  and  $\|\cdot\|$  be any two norms and  $A$  an  $N \times M$  matrix with  $\text{rank}(A) = N$ . For  $\theta > 0$ , consider a random variable  $d$  with uniform distribution on  $B_{f_d}(\theta)$ . Then, for any  $K = 0, \dots, N$*

$$\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K) = \overline{C}_K (\tau/\theta)^{N-K} + o((\tau/\theta)^{N-K}) \text{ as } \tau/\theta \rightarrow 0. \quad (12)$$

The proof involves three main steps:

- (i) For  $J \in \mathcal{J}(K)$ , estimate  $\mathbb{L}^N (\mathcal{A}_J^\tau \cap B_{f_d}(\theta))$ . It has the form  $\theta^N C_J (\tau/\theta)^{N-K} + \theta^N o((\tau/\theta)^{N-K})$ .
- (ii) Estimate the volume of  $\bigcup_{J \in \mathcal{J}(K)} \mathcal{A}_J^\tau \cap B_{f_d}(\theta)$ . An important intermediate result says that if  $(J, J') \subset \{1, \dots, N\}^2$  with  $\mathcal{A}_J \neq \mathcal{A}_{J'}$  and  $\dim(\mathcal{A}_J) = \dim(\mathcal{A}_{J'}) = K$ , then  $\mathbb{L}^N (\mathcal{A}_J^\tau \cap \mathcal{A}_{J'}^\tau \cap B_{f_d}(\theta)) = \theta^N o((\tau/\theta)^{N-K})$ . It becomes negligible when  $\tau/\theta \rightarrow 0$ .
- (iii) Divide the result of (ii) by  $\theta^N \mathbb{L}^N (B_{f_d}(1))$  to obtain the sought-after probabilities.

### 3. Meaning of the results

**(a) A limitation.** Theorem 2.2 describes the behavior of  $\mathbb{P}(\text{val}(\mathcal{P}_d) \leq K)$  with increasing precision when  $\tau/\theta$  goes to 0. However, the  $o()$  function depends on all the ingredients of the model, namely  $f_d$ ,  $A$ ,  $\|\cdot\|$ ,  $K$  and  $N$ . For most interesting scenarios, the current estimates (see [6]) are only valid for a range of values  $\tau/\theta$  which are too small, when compared to the values  $\tau/\theta$  which are met in applications. In order to provide accurate bounds for adequate  $\tau/\theta$ , we need better measures for the sets  $\mathcal{A}_J^\tau$  and for their intersections of different orders. One can notice that, in the very general setting of  $(\mathcal{P}_d)$ , this comes up against fundamental mathematical problems relevant to Geometry of Banach Spaces that remain open—see [7].

**(b) Crucial improvement in Approximation.** If we enrich the dictionary with a new element  $a_{M+1} \in \mathbb{R}^N$  (i.e. which is collinear to none of the columns of  $A$ ), then  $A$  is  $N \times (M+1)$  and  $\mathcal{J}(K)$  contains more elements (except if  $K \in \{0, N\}$ ). As a consequence, the value of  $\overline{C}_K$  in (11) increases and then the probability in (12) increases. *Hence the model is improved.* In words, adding an element to the dictionary improves the average performance. However, the performance might not change for the worst case analysis, since we can hardly expect the new element to improve the approximation of all the worst possible data. The possibility to draw such a conclusion emphasizes the benefit of the APA over a worst-case point of view.

(c) **The role of  $\|\cdot\|$ .** In  $(\mathcal{P}_d)$  and in the BKTA (4), the choice of  $\|\cdot\|$  is of critical importance. Equations (10), (11) and (12) suggest to optimize a weighted sum of the measures of  $P_{\mathcal{A}_J^\perp}(B_{\|\cdot\|}(1))$ . In particular, the optimal  $\|\cdot\|$  depends on the data distribution ( $f_d$ ) and on the dictionary ( $A$ ). We can hardly expect that there is a *universal* choice for  $\|\cdot\|$  (e.g.  $\|\cdot\| = \|\cdot\|_2$ , the Euclidian norm) such that  $\overline{\mathcal{C}}_K$  reaches its optimal value for *all*  $K \in \{1, \dots, N\}$ . With this regard, we anticipate that in the derivation of similar results for the  $\ell_1$  minimization, the corresponding term is different (see [5]), since it is governed by the Kuhn-Tucker optimality conditions.

(d) **Approximation versus compressed-sensing (CS).** In CS, it has been shown that under suitable hypotheses on the matrix  $A$ , the data  $d$  and when  $\|\cdot\|$  is the  $\ell_2$ -norm, some of the heuristics that approximate  $(\mathcal{P}_d)$  (or BKTA) provide solutions with the correct support (see [8,4]). These results are aggregated with methods for constructing suitable matrices  $A$  and the remark that, for these matrices  $A$ , the sparsest decomposition is unique for sparse signals. Altogether, this forms the core of CS (see [1] for a typical CS statement). In order to apply the results of [8,4] in the context of approximation it is important to assess the gap between the performance of the sparsest approximation model under the CS assumptions and without any restrictive assumptions.

The typical CS hypothesis lead to special matrices  $A$  such that the vector spaces in  $\mathcal{J}(K)$  are almost orthogonal when  $K$  is small enough. Then  $\mathbb{L}^N(\mathcal{A}_J^\tau \cap \mathcal{A}_{J'}^\tau \cap B_{f_d}(\theta))$ , for  $(J, J') \in (\mathcal{J}(K))^2$  is reduced—see (ii) in the sketch of the proof of Theorem 2.2—but in any case these terms are asymptotically negligible when  $\tau/\theta$  is small. The impact of such an hypothesis on  $\overline{\mathcal{C}}_K$  is an open question, see paragraph (e). The discussion on the hypothesis “ $\|\cdot\|$  is the Euclidean norm” is in paragraph (c).

Last, if  $(\mathcal{P}_d)$  is used for approximation, adding a new column to  $A$  always improves the performance of  $(\mathcal{P}_d)$ —see (b). However, if  $(\mathcal{P}_d)$  is used for CS, the new column may degrade its performance. It is e.g. the case if the new column belongs to an element  $\mathcal{J}(2)$ .

(e) **Distribution of the data.** Analyzing (10)-(11) shows that the distribution of the data (defined using  $f_d$ ) occurs in coefficients  $\overline{\mathcal{C}}_K$  at first in the numerator—but restricted to a subspace of dimension  $K$ —and at second in their denominator as defined on  $\mathbb{R}^N$ . As a consequence, the matrix  $A$  should be designed so that the ratio  $\mathbb{L}^K(\mathcal{A}_J \cap B_{f_d}(1)) / \mathbb{L}^N(B_{f_d}(1))$  is as large as possible, in particular when  $\#J$  is small.

## 4. Two examples in $\ell_p(\mathbb{R}^N)$ spaces

For any  $1 \leq p < \infty$ , the  $\ell_p$  norm of a vector  $u \in \mathbb{R}^n$ ,  $\forall n \in \mathbb{N}$ , is denoted  $\|u\|_p = (\sum_{i=1}^n |u_i|^p)^{\frac{1}{p}}$ . It may be useful to remind that the volume of the unit ball of  $\ell_p(\mathbb{R}^n)$  reads [7]

$$V_n(p) = (2\Gamma(1 + 1/p))^n / \Gamma(1 + n/p), \quad (13)$$

where  $\Gamma$  is the Gamma function. Remind that for  $1 \leq p < \infty$ ,  $V_n(p)$  rapidly decay to 0 as far as  $n$  increases.

### 4.1. Case $\|\cdot\| = \ell_2$ , $f_d = \ell_1$ and $A = \text{Id}$

In this case we write  $\overline{\mathcal{C}}_K^1$  in place of  $\overline{\mathcal{C}}_K$ . Since  $A = \text{Id}$ ,  $\forall J \subset \{1, \dots, N\}$  we have  $\dim \mathcal{A}_J = \#J$ , hence

$$\mathbb{L}^{N-\#J}(P_{\mathcal{A}_J^\perp}(B_{\|\cdot\|}(1))) = V_{N-\#J}(2) \quad \text{and} \quad \mathbb{L}^{\#J}(\mathcal{A}_J \cap B_{f_d}(1)) = V_{\#J}(1),$$

where  $V_n(p)$  is given in (13). By (10)-(11) we obtain that  $\overline{\mathcal{C}}_K^1 = \frac{N!}{K!(N-K)!} \frac{V_{N-K}(2)V_K(1)}{V_N(1)}$ ,  $\forall K \in \{0, \dots, N\}$ .

### 4.2. Case $\|\cdot\| = \ell_2$ and $f_d = \ell_2$

We label this case by writing  $\overline{\mathcal{C}}_K^{2,M}$  in place of  $\overline{\mathcal{C}}_K$ , where  $A$  is of size  $N \times M$ , with  $N \leq M$ . Assume also that  $A$  is such that for any  $J_1 \subset \{1, \dots, M\}$  with  $\#J_1 < N$ , we have  $\dim(\mathcal{A}_{J_1}) = \#J_1$ , and for any  $J_2 \subset \{1, \dots, M\}$  satisfying  $\#J_1 = \#J_2$  and  $J_1 \neq J_2$ , we have  $\mathcal{A}_{J_1} \neq \mathcal{A}_{J_2}$ . Hence  $\#\mathcal{J}(N) = 1$  and  $\#\mathcal{J}(K) = \frac{M!}{K!(M-K)!}$ , for  $K = 0, \dots, N-1$ . Moreover, for all  $J \subset \{1, \dots, M\}$  with  $\#J \leq N$

we have  $\mathbb{L}^{N-\#J}(P_{\mathcal{A}_J^\perp}(B_{\|\cdot\|}(1))) = V_{N-\#J}(2)$  and  $\mathbb{L}^{\#J}(\mathcal{A}_J \cap B_{f_d}(1)) = V_{\#J}(2)$ . Using (10)-(11) yet again,  $\overline{C}_K^{2,M} = \#\mathcal{J}(K) \frac{V_{N-K}(2)V_K(2)}{V_N(2)}$ , for all  $K \in \{0, \dots, N\}$ . Observe that  $\overline{C}_K^{2,M}$  depends on  $A$  only via  $M$ .

#### 4.3. Simulation in problem comparison

By way of illustration, we consider the following question: How much redundancy (i.e.  $\frac{M}{N}$ ) is needed to capture data uniformly distributed on  $B_{\|\cdot\|_2}(\beta\theta)$ , for  $\beta > 0$ , as accurately as data uniformly distributed on  $B_{\|\cdot\|_1}(\theta)$  without redundancy? Denoting  $(\mathcal{P}_d^1)$  the problem described in Section 4.1 and  $(\mathcal{P}_{d,\beta}^{2,M})$  the problem of Section 4.2, for given  $M$  and  $\beta$ , we have

$$\mathbb{P}(\text{val}(\mathcal{P}_{d,\beta}^{2,M}) \leq K) / \mathbb{P}(\text{val}(\mathcal{P}_d^1) \leq K) = \overline{C}_K^{2,M} \beta^{K-N} / \overline{C}_K^1 + o(1) \quad , \text{ as } \tau/\theta \rightarrow 0.$$

We display on Fig. 1  $K \rightarrow \overline{C}_K^{2,M} \beta^{N-K} / \overline{C}_K^1$  for  $N = 10^2$  and: (1)  $M \in \{10^2, 10^3, 10^4\}$  with  $\beta = 1$ ; (2)  $M = 10^2$  with  $\beta = \left(\frac{V_N(1)}{V_N(2)}\right)^{1/N}$ . Experimentally, redundancy does not help when  $\tau/\theta$  and  $K$  are small. However, the performances for  $B_{\|\cdot\|_2}(\beta\theta)$  seem very close to those for  $B_{\|\cdot\|_1}(\theta)$ . Notice that  $0 < c < \beta = \left(\frac{V_N(1)}{V_N(2)}\right)^{1/N} \leq 1$  where  $c$  is a universal constant, see [7].

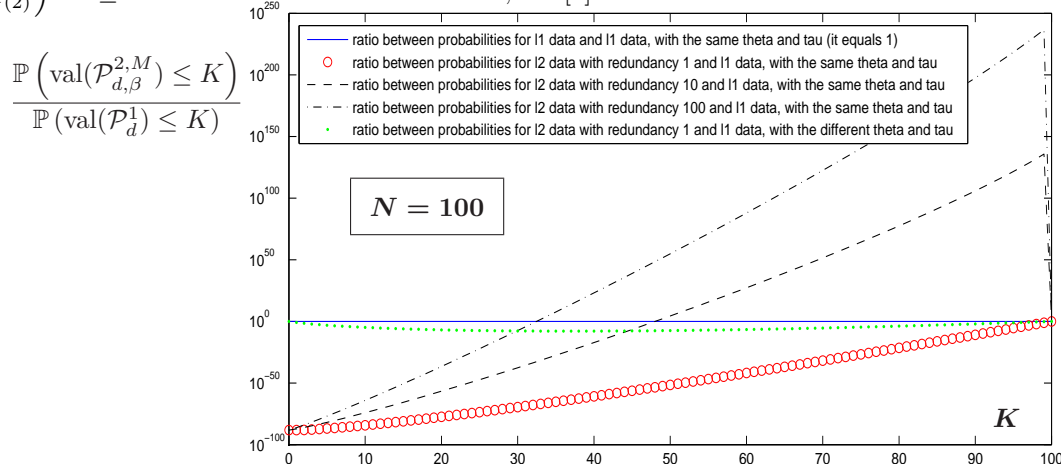


Figure 1. All curves correspond to various values of  $M$  and  $\beta$  where  $\tau/\theta$  is small.

## References

- [1] E. Candes, J. Romberg, T. Tao, Robust uncertainty principles : Exact Signal reconstruction from highly incomplete frequency information *IEEE, Trans. on Information Theory*, 52(2):489–509, 2006.
- [2] G. Davis, S. Mallat, M. Avellaneda, Adaptive greedy approximations *Constructive approximation*, 13(1):57–98, 1997.
- [3] R.A. DeVore, Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.
- [4] J.J. Fuchs, Recovery of exact sparse representations in the presence of bounded noise. *IEEE, trans. on Information Theory*, 51(10):3601–3608, Oct. 2005.
- [5] F. Malgouyres, Rank related properties for basis pursuit and total variation regularization. *Signal Processing*, 87(11):2695–2707, Nov. 2007.
- [6] F. Malgouyres, M. Nikolova, Average performance of the sparsest approximation using a general dictionary. Report HAL-00260707 and CMLA n.2008-08.
- [7] G. Pisier, The volume of convex bodies and Banach space geometry *Cambridge University Press*, 1989.
- [8] J.A. Tropp, Greed is good: algorithmic results for sparse approximation *IEEE, trans. on Information Theory*, 50(10):2231–2242, Oct. 2004.
- [9] J.A. Tropp, Just relax: convex programming methods for identifying sparse signals in noise *IEEE, trans. on Information Theory*, 52(3):1030–1051, March 2006.