Average performance of the sparsest approximation in a dictionary

François Malgouyres, Mila Nikolova

Average performance of the sparsest approximation in a dictionary

François Malgouyres⋆
and Mila Nikolova•

***Abstract*—Given data $d \in \mathbb{R}^N$, we consider its representation u^* involving the least number of non-zero elements (denoted by $\ell_0(u^*)$) using a dictionary A (represented by a matrix) under the constraint $\|Au-d\| \leq \tau$, for $\tau > 0$ and a norm $\|.\|$. This (nonconvex) optimization problem leads to the sparsest approximation of d.**

We assume that data d are uniformly distributed in $\theta B_{f_d}(1)$ where $\theta > 0$ and $B_{f_d}(1)$ is the unit ball for a norm f_d. Our main result is to estimate the probability that the data d give rise to a $K-$sparse solution u^*: we prove that

$$\mathbb{P}\left(\ell_0(u^*) \leq K\right) = C_K(\frac{\tau}{\theta})^{(N-K)} + o((\frac{\tau}{\theta})^{(N-K)}),$$

where u^* is the sparsest approximation of the data d and $C_K > 0$. The constants C_K are an explicit function of $\|.\|$, A, f_d and K which allows us to analyze the role of these parameters for the obtention of a sparsest $K-$sparse approximation. Consequently, given f_d and θ, we have a tool to build A and $\|.\|$ in such a way that C_K (and hence $\mathbb{P}\left(\ell_0(u^*) \leq K\right)$) are as large as possible for K small.

In order to obtain the above estimate, we give a precise characterization of the set Σ_K^τ of all data leading to a $K-$sparse result. The main difficulty is to estimate accurately the Lebesgue measure of the sets $\{\Sigma_K^\tau \cap B_{f_d}(\theta)\}$.

We sketch a comparative analysis between our Average Performance in Approximation (APA) methodology and the well known Nonlinear Approximation (NA) which also assess the performance in approximation.

***Index Terms*—sparsest approximation, ℓ_0 minimization, average performance in approximation, nonlinear approximation.**

⋆ Université Paris 13, CNRS UMR 7539 LAGA, 99 avenue J.B. Clément, F-93430 Villetaneuse, France; malgouy@math.univ-paris13.fr, Phone: (33/0) 1 49 40 35 83

• CMLA, ENS Cachan, CNRS, PRES UniverSud, 61 Avenue du Président Wilson, 94230 Cachan, France; nikolova@cmla.ens-cachan.fr, Phone: (33/0) 1 40 39 07 34

I. The Average Performance in Approximation (APA) methodology

We consider the *sparsest* approximation of observed data $d \in \mathbb{R}^N$ using a dictionary $A = \{a_1, \ldots, a_M\} \in \mathbb{R}^{N \times M}$ with rank$(A) = N$ under a tolerance constraint given by a norm $\|.\|$ and $\tau \gtrapprox 0$. This approximation is a solution of the non-convex constrained optimization problem $(\mathcal{P}_d)$ given below:

$$(\mathcal{P}_d): \left\{ \begin{array}{l} \text{minimize}_{u \in \mathbb{R}^M} \ \ell_0(u), \\ \text{under the constraint :} \quad \|Au-d\| \leq \tau, \end{array} \right.$$

with

$$\ell_0(u) \stackrel{\text{def}}{=} \#\big\{1 \leq i \leq M : u_i \neq 0\big\},$$

where $\#$ denotes cardinality.

For $\theta > 0$ and a norm f_d, we assume that data d are uniformly distributed on the $\theta-$level set of f_d:

$$B_{f_d}(\theta) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^N, f_d(v) \leq \theta\} = \theta B_{f_d}(1).$$

For sparse data, f_d is typically the ℓ_1 norm.

We inaugurate an Average Performance in Approximation (APA) methodology to evaluate the performance of $(\mathcal{P}_d)$. Denoting the minimum $\ell_0(u^*)$ in $(\mathcal{P}_d)$ by $\text{val}(\mathcal{P}_d)$, APA consists in estimating for every $K = 0, \ldots, N$ the value of

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

The larger $\mathbb{P}\left(\text{val}(\mathcal{P}_d) \leq K\right)$ for K small, the better the model involved $(\mathcal{P}_d)$. The set of all data d leading to $\text{val}(\mathcal{P}_d) \leq K$ reads:

$$\Sigma_K^\tau \stackrel{\text{def}}{=} \{d \in \mathbb{R}^N, \text{val}(\mathcal{P}_d) \leq K\}. \qquad (1)$$

Following our APA methodology—see the Reports [7], [6]—we accurately describe the geometry of Σ_K^τ. Combining this with the model for the data distribution, we derive a lower and an upper bound on $\mathbb{P}\left(\text{val}(\mathcal{P}_d) \leq K\right)$. We show that these bounds share the same asymptotical behavior as $\tau/\theta \to 0$. The formulae describing this behavior clarify the role of the ingredients of the model (i.e. $A, \|.\|, \tau$) whit respect to the ingredients of the data distribution (i.e. f_d and θ).

II. Main results already obtained

For $J \subset \{1,\dots,M\}$, let us denote

$$\mathcal{A}_J = \operatorname{span}\big\{a_j : j \in J\big\},$$

where a_j is the j^{th} column of the matrix A. 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

$$\mathcal{A}_J^\tau \stackrel{\text{def}}{=} \mathcal{A}_J + B_{\|.\|}(\tau)$$

and show that

$$\mathcal{A}_J^\tau = \mathcal{A}_J + P_{\mathcal{A}_J^\perp}\left(B_{\|.\|}(\tau)\right). \tag{2}$$

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

For any given dimension $1 \le K \le N$, we define (see [7] for the details):

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

We always have $\#\mathcal{J}(N) = 1$ and set $\mathcal{J}(0) = \{\emptyset\}$. Theorem 1 is the *key* for the obtention of the sought-after results. It provides an easy geometrical vision of the problem.

Theorem 1: For any $N \times M$ matrix A with rank$(A) = N$, any norm $\|.\|$, any $\tau > 0$ and any $K \in \{0,\dots,N\}$, the set Σ_K^τ in (1) *reads*

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

A 2D example in Fig. 1 illustrates sets Σ_K^τ for several values of K.

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

$$C_J = \mathbb{L}^{N-K}\left(P_{\mathcal{A}_J^\perp}\left(B_{\|.\|}(1)\right)\right) \times \mathbb{L}^K\left(\mathcal{A}_J \cap B_{f_d}(1)\right) \tag{3}$$

where $K = \dim(\mathcal{A}_J)$ and for any $K = 0,\dots,N$,

$$\boldsymbol{C}_K = \frac{\sum_{J \in \mathcal{J}(K)} C_J}{\mathbb{L}^N\left(B_{f_d}(1)\right)}. \tag{4}$$

Using these notations, we state next the main result of [7].

Theorem 2: Let f_d and $\|.\|$ be any two norms and A an $N \times M$ matrix with rank$(A) = N$. For $\theta > 0$, consider a random variable d with uniform distribution in $B_{f_d}(\theta)$. Then, for any $K = 0,\dots,N$, we have

$$\mathbb{P}\left(\operatorname{val}(\mathcal{P}_d) \le K\right) = \boldsymbol{C}_K \left(\frac{\tau}{\theta}\right)^{N-K} + o\left(\left(\frac{\tau}{\theta}\right)^{N-K}\right) \text{ as } \frac{\tau}{\theta} \to 0. \tag{5}$$

The proof of the Theorem involves four main steps. These steps are illustrated on Figure 1.

(i) For $J \in \mathcal{J}(K)$, estimate

$$\mathbb{L}^N\left(\mathcal{A}_J^\tau \cap B_{f_d}(\theta)\right).$$

It reads $\theta^N C_J (\frac{\tau}{\theta})^{N-K} + \theta^N o\big((\frac{\tau}{\theta})^{N-K}\big)$ when $\tau/\theta \to 0$.

(ii) 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

$$\mathbb{L}^N\left(\mathcal{A}_J^\tau \cap \mathcal{A}_{J'}^\tau \cap B_{f_d}(\theta)\right) = \theta^N o\big((\frac{\tau}{\theta})^{N-K}\big)$$

when $\tau/\theta \to 0$.

(iii) Using (i)-(ii), estimate the volume of

$$\bigcup_{J \in \mathcal{J}(K)} \mathcal{A}_J^\tau \cap B_{f_d}(\theta).$$

It reads

$$\theta^N \left(\sum_{J \in \mathcal{J}(K)} C_J\right) \left(\frac{\tau}{\theta}\right)^{N-K} + \theta^N o\left(\left(\frac{\tau}{\theta}\right)^{N-K}\right)$$

when $\tau/\theta \to 0$.

(iv) Divide the result of (iii) by $\theta^N \mathbb{L}^N\left(B_{f_d}(1)\right)$ to obtain the probabilities given in (5).

Remark 1: The main difficulty at this stage of our research to obtain more precise, non asymptotic bounds for the measures evoked in (i), (ii) and (iii) comes up against fundamental mathematical problems in Geometry of Banach Spaces, see [8].

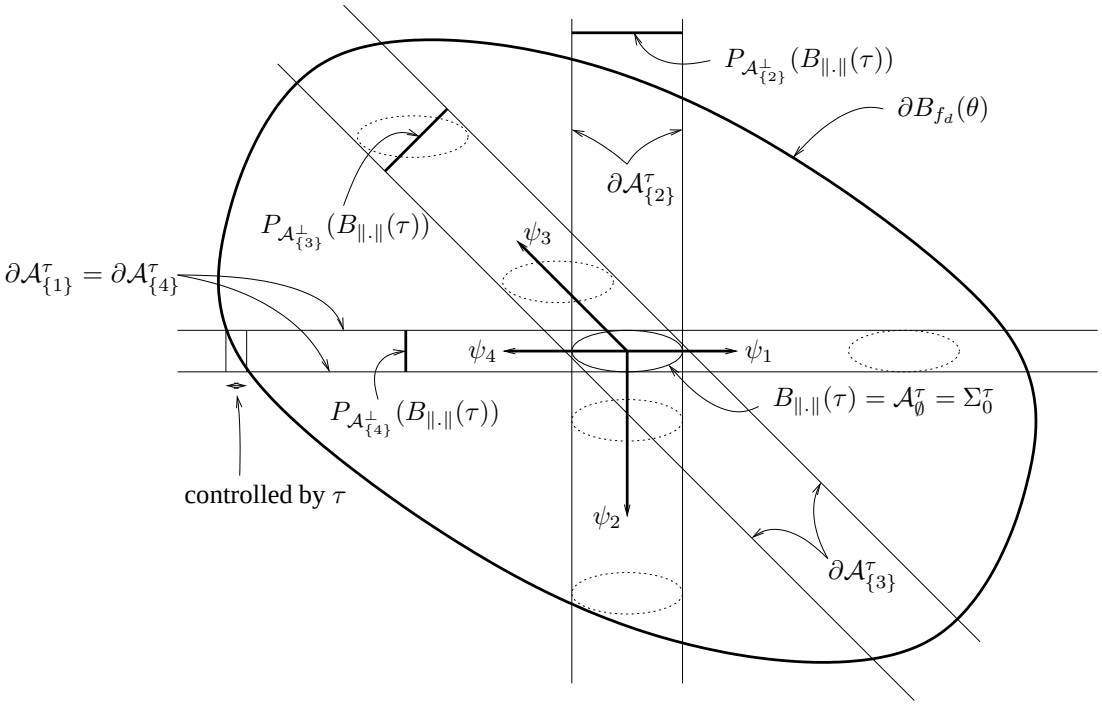

Fig. 1. Example in dimension 2. Let the dictionary read $\{\psi_1, \psi_2, \psi_3, \psi_4\}$. On the drawing, the sets $\mathcal{A}^\tau_{\{i\}}$ are obtained as the direct sum of $P_{\mathcal{A}^\perp_{\{i\}}}(B_{\|.\|}(\tau))$ and $\mathcal{A}_{\{i\}}$, for $i = 2, 3, 4$. The dotted ellipses represent translations of $B_{\|.\|}(\tau)$. The set-valued function $\Sigma^\tau_\bullet$, as presented in (1) and Theorem 1, gives rise to the following situations: $\Sigma^\tau_0 = B_{\|.\|}(\tau) = \mathcal{A}^\tau_\emptyset$, $\Sigma^\tau_1 = \mathcal{A}^\tau_{\{1\}} \cup \mathcal{A}^\tau_{\{2\}} \cup \mathcal{A}^\tau_{\{3\}}$ and $\Sigma^\tau_2 = \mathbb{R}^2 = \mathcal{A}^\tau_{\{1,2\}} = \mathcal{A}^\tau_{\{2,3\}} = \ldots$ The symbol ∂ is used to denote the boundary of a set.

III. APA and Nonlinear Approximation

A. Reminder on Nonlinear Approximation

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

$$\begin{cases} \text{minimize}_{u\in\mathbb{R}^M} \|Au - d\|, \\ \text{under the constraint :} \quad \ell_0(u) \leq K, \end{cases} \tag{6}$$

where $0 \leq K \leq N$.

The evaluation of the performances of the latter is a well developed field, see [2]. It is named *Nonlinear Approximation* (NA) if $M = N$ and *Highly Nonlinear Approximation* (HNA) if $M > N$. Under some hypotheses on the support S of the data distribution, for a given $\alpha > 0$ it provides bounds of the form

$$\|Au^*_K - d\| \leq cK^{-\alpha}, \forall d \in S, \tag{7}$$

where u^*_K is the BK-TA of d and $c > 0$. Doing so, it gives the asymptotical behavior of $\|Au^*_K - d\|$ when $K \to \infty$. Note that NA considers approximation in general Hilbert spaces, Besov spaces and so on.

In spite of the different level of achievement, we resume in the following sections the main similarity and differences between NA and APA.

B. APA and NA are complementary

The goals of APA and NA are completely different, which is a mean to say that they are complementary:

- APA looks for the average performance of an approximation;
- NA exhibits the worst case in this approximation.

If we assume that data live in $B_{f_d}(\theta)$ (with no assumption on their distribution), the inequality (7), transposed in our context, amounts to

$$B_{f_d}(\theta) \subset \Sigma_K^{c\theta K^{-\alpha}}, \ \forall K.$$

In words, for $\frac{\tau}{\theta} = cK^{-\alpha}$ we get

$$\begin{aligned}\mathbb{P}\left(\mathrm{val}(\mathcal{P}_d) \leq K\right) &= \frac{\mathbb{L}^N\left(\Sigma_K^{c\theta K^{-\alpha}} \cap B_{f_d}(\theta)\right)}{\mathbb{L}^N\left(B_{f_d}(\theta)\right)} \\ &= 1, \ \forall K,\end{aligned}$$

whatever the data distribution as long as it is supported inside $B_{f_d}(\theta)$. Notice that APA considers cases where $B_{f_d}(\theta) \not\subset \Sigma_K^{\tau}$ which means that $\frac{\tau}{\theta} < cK^{-\alpha}$, see e.g. Fig. 1.

We display on Figure 2 the typical curve obtained when plotting, for a fixed K, the function $\tau \rightarrow \mathbb{P}\left(\mathrm{val}(\mathcal{P}_d) \leq K\right)$, in a "loglog" plot (both x and y-axes use a logarithmic scale). We also display the typical information obtained by both NA and APA. These curves emphasize that NA and APA do not provide information in the same range of values for τ/θ.

The current results in APA are relevant for small values of τ/θ. For instance, $\mathbb{P}\left(\mathrm{val}(\mathcal{P}_d) \leq K\right)$ according to (7) is correctly approximated by $C_K\left(\frac{\tau}{\theta}\right)^{(N-K)}$ only for $\frac{\tau}{\theta} < cK^{-\alpha}$. This suggest that APA is better suited when the value $cK^{-\alpha}$ is larger than the set of values $\frac{\tau}{\theta}$ which are relevant to the applicative context for the interesting K.

The other situations where APA is interesting is (obviously) when NA is not sufficiently accurate.

C. The hypotheses on the data distribution

NA needs a weaker hypothesis on the data distribution, which is obviously a good point. However, the counterpart of this advantage is that the estimated performances correspond to the worst possible data in the support. Notice that these data form a set which is not necessarily connected and may be very small with respect to the support. (E.g., construct a simple example of data distributed in the ℓ_1 ball, when $\|.\|$ is the Euclidean norm.)

The main consequence of the above remark is that very little can be said in HNA. In practice, the performance when using a redundant dictionary A (i.e. $M > N$) are identical to the performances when $M = N$ (see [2]).

A quick look at (4) shows that if a new vector a_{M+1}, which is collinear with none of the $a_i, \ i \in \{1, \cdots, M\}$, is concatenated to A, the value of C_K is increased, since the cardinality of $\mathcal{J}(K)$ goes up. Thus the probability in (5) increases.

The difference between HNA and APA, with the regard to the hypothesis on the data distribution can be sketched as it follows:

HNA only makes an hypothesis on the support of the data distribution but the results when $M > N$ are vague;

APA needs assumptions on the data distribution but gives a full description of the chance of getting a K-sparse solution when $M \geq N$.

D. The sparsest approximation and its heuristics

As a matter of fact, the sparsest approximation $(\mathcal{P}_d)$, or equivalently the BK-TA, is an essentially theoretical problem. The reason is that it is NP-hard in general (see [1]). The determination of assumptions which enable its solution by feasible numerical schemes is currently a very active field of research, see e.g. [9], [3], [10]). The typical assumptions are threefold:

- There must exist a (very) sparse decomposition for representing data d;
- The matrix A must be incoherent, in some way;
- The norm $\|.\|$ must be the ℓ_2 norm.

When using $(\mathcal{P}_d)$ in an approximation context, it is therefore critically important to clarify the following questions:

- what is the gap in performance when enforcing the constraints on the matrix A and on $\|.\|$, mentioned above;
- when these hypotheses cannot be met, what is the gap in performance between the sparsest approximation and its heuristics.

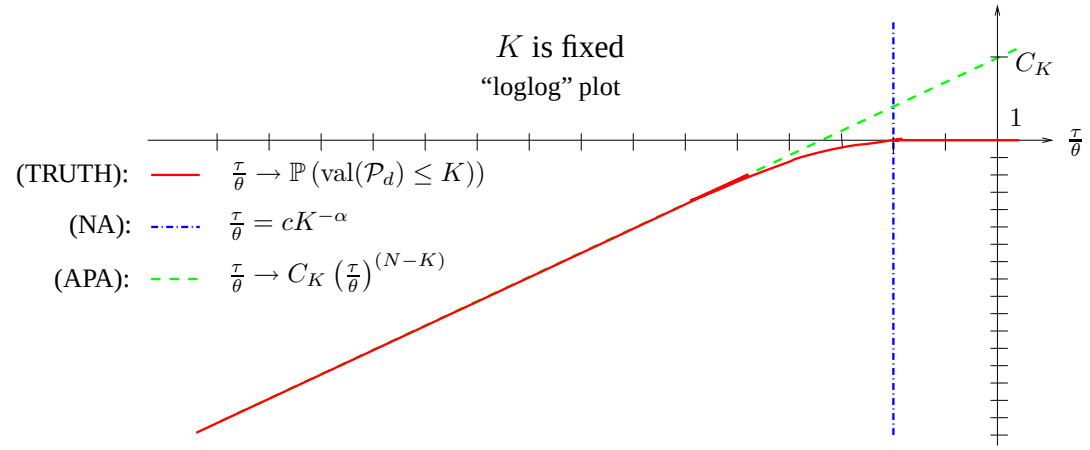

Fig. 2. The value of C_K is fixed by the model.
In red : the curve $\frac{\tau}{\theta} \to \mathbb{P}\left(\mathrm{val}(\mathcal{P}_d) \leq K\right)$.
In dotted blue : the results of Non Linear Approximation tell us that on the right of the blue line, the red curve equals 1.
In dashed green : the results of APA tell us that the red curve and the dashed green line are asymptotically equal when $\frac{\tau}{\theta}$ goes to 0. The slope of the dashed green line is always $N - K$.

In order to answer (even approximatively) these questions, we need a methodology to describe the performances of the different models when $M > N$. As far as ℓ_1 approximation is concerned (in ℓ_1 approximation, the ℓ_0 "norm" is replaced by the ℓ_1 norm), it is a reasonable perspective for APA (see [4], [5]). It is an open question for greedy algorithms such as the Orthogonal Matching Pursuit as analyzed in [9].

IV. Conclusion

The proposed APA methodology is morally a reasonable approach to assess the performances in sparse approximation. It is a very young field so there are numerous open questions to explore. It can provide a good complement to the well developed NA approach and can deal with situations such as HNA.

References

[1] G. Davis and S. Mallat and M. Avellaneda. Adaptive greedy approximations *Constructive approximation*, 13(1):57–98, 1997.

[2] R.A. Devore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[3] 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.

[4] F. Malgouyres. Rank related properties for basis pursuit and total variation regularization. *Signal Processing*, 87(11):2695–2707, Nov. 2007.

[5] F. Malgouyres. Projecting onto a polytope simplifies data distributions. Report—University Paris 13, 2006-1, Jan. 2006.

[6] F. Malgouyres and M. Nikolova. Average performance of the sparsest approximation using a general dictionary. CMLA Report n.2008-13, accepted with minor modifications in *Comptes Rendus de l'Académie des Sciences, série mathématiques.*

[7] F. Malgouyres and M. Nikolova. Average performance of the approximation in a dictionary using an ℓ_0 objective. Report HAL-00260707 and CMLA n.2008-08.

[8] G. Pisier. The volume of convex bodies and Banach space geometry. *Cambridge University Press*, 1989.

[9] J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE, trans. on Information Theory*, 50(10):2231–2242, Oct. 2004.

[10] 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.